SENSE — Sensory Engagement Network for Shared Experiences: State-Of-The-Art

The SENSE project team

January 5, 2026

Abstract

The rapid evolution of audio understanding and generation technologies, driven by advances in deep learning and large-scale pretraining, has fundamentally reshaped the landscape of interactive and multisensory digital systems. Recent progress in music source separation, audio tagging, music information retrieval, automatic speech recognition, and multimodal generative modeling enables increasingly rich forms of human–computer interaction that extend beyond purely auditory experiences. In this context, the *Sensory Engagement Network for Shared Experiences (SENSE)* project aims to explore the scientific and technological foundations required for transforming audio content into coherent multisensory representations, including visual, tactile, and haptic feedback.

This article presents a comprehensive state-of-the-art analysis of the core research directions underpinning the SENSE platform. We systematically review modern approaches to audio source separation, highlighting the transition from traditional spectrogram-based pipelines to end-to-end waveform and transformer-based architectures capable of achieving high perceptual quality and real-time performance. We further examine large-scale audio tagging and representation learning frameworks, with particular emphasis on convolutional, transformer-based, and hybrid models pretrained on weakly-labeled datasets, which have become essential building blocks for semantic audio understanding and transfer learning.

In addition, the study surveys key developments in music information retrieval, including transcription, harmonic analysis, and rhythmic modeling, illustrating the shift from handcrafted features toward self-supervised and generative paradigms. Advances in automatic speech recognition are discussed in relation to multilingual, self-supervised, and foundation-model-based systems that enable robust speech understanding across diverse languages and acoustic conditions. The analysis is extended to audio–visual and multimodal content generation, where diffusion models and large-scale generative architectures enable controllable synthesis and cross-modal alignment between sound, language, and visual representations.

A dedicated section addresses the emergence of audio and multimodal foundation models, such as contrastively trained audio–text systems and unified audio–language architectures, which provide a scalable substrate for zero-shot inference, multimodal reasoning, and downstream task adaptation. Particular attention is given to recent research on haptic and vibrotactile generation models, underscoring their relevance for closing the sensory loop in multisensory systems and enabling embodied forms of audio perception.

By integrating these research strands, this state-of-the-art review establishes a unified conceptual and technological baseline for the design of the SENSE platform. The analysis identifies key challenges related to scalability, real-time constraints, semantic alignment, and multisensory coherence, while outlining promising directions for future research. Overall, the article provides a structured foundation for the subsequent development and validation of multisensory audio-centric systems, positioning SENSE within the broader trajectory of human-centered and multimodal artificial intelligence.

Contents

1 Introduction

The rapid evolution of audio understanding and generation technologies, driven by advances in deep learning and large-scale pretraining, has fundamentally reshaped the landscape of interactive and multisensory digital systems. In this context, the Sensory Engagement Network for Shared Experiences (SENSE) project aims to explore the scientific and technological foundations required for transforming audio content into coherent multisensory representations, including visual, tactile, and haptic feedback.

1.1 The Paradigm Shift: From Signal Processing to Deep Learning

The field of audio processing has undergone a radical transformation over the past decade. Traditionally, Music Information Retrieval (MIR) and audio analysis relied heavily on manually engineered features—such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and spectral centroids—which were then processed through classical machine learning algorithms like Support Vector Machines (SVMs). While effective for specific tasks like genre classification, these methods often struggled to generalize across diverse acoustic environments and required significant domain expertise.

The introduction of deep learning reoriented the field toward end-to-end models capable of learning features directly from raw waveforms or spectrograms. Early breakthroughs utilized Convolutional Neural Networks (CNNs) for audio tagging and tagging tasks, and Recurrent Neural Networks (RNNs/LSTMs) to capture the temporal dependencies inherent in sequential audio data. More recently, Transformer-based architectures have surpassed these models in discriminating among genres and managing long-range dependencies, particularly in multimodal settings.

1.2 The Era of Foundation Models and Self-Supervision

The current state-of-the-art is increasingly defined by the rise of "foundation models"—large-scale pretrained models that unify audio, language, and multimodal understanding. Self-supervised learning (SSL) frameworks, such as wav2vec 2.0 and Contrastive Predictive Coding (CPC), have enabled models to learn high-level representations from massive, unlabeled audio collections.

These foundation models, such as CLAP (Contrastive Language-Audio Pretraining) and AudioCLIP, encode sound into a shared latent space with text and images, allowing for zero-shot capabilities in audio-text retrieval and complex query-based tasks. Furthermore, the integration of audio capabilities into Large Language Models (LLMs) has enabled systems to follow open-ended natural language instructions for both understanding and controllable generation of speech and audio.

1.3 Multimodal Synergy and Multisensory Engagement

Despite these advances in auditory intelligence, humans naturally experience the world through a combination of senses, including vision, touch, and proprioception. Multimodal interaction represents a transformative shift in human-computer interaction (HCI), making technology more akin to human communication patterns where multiple sensory inputs are processed simultaneously to convey intent and emotion.

The SENSE project addresses a critical gap in this multisensory integration: the transition from purely auditory experiences to "audio-centered" design. While vision often dominates VR and AR applications, including tactile and haptic information can significantly enhance the sense of presence and realism, as the sense of touch operates with higher temporal resolution than vision or hearing.

1.4 Challenges in Real-Time Alignment and Perception

Transforming audio into tactile or visual feedback in real-time presents significant technical hurdles. Universal sound separation, a core component of the SENSE pipeline, often faces a fundamental misalignment where models optimized for low-level signal metrics (like SDR) fail to produce semantically meaningful results that match human preference. This is particularly evident in the “Cocktail Party Problem”, where the ability to selectively focus on a single auditory source amid background noise is required.

Achieving precise temporal alignment between audio events and their corresponding multisensory outputs (e.g., a visual motion or a haptic pulse) is essential for maintaining perceptual coherence. Current research is pushing toward "semantically-aligned" systems that use reinforcement learning and contrastive objectives to ensure that the generated multisensory feedback is not only synchronized but also perceptually and semantically consistent with the original audio source.

1.5 Societal Importance and Applicability

The importance of the SENSE framework extends beyond entertainment into critical domains like accessibility and robotics:

- **Accessibility:** By converting audio into tactile and visual signals, SENSE-like systems can empower the hearing impaired to experience music and environmental cues in intuitive ways.
- **Robotics:** Foundation models that unify sight, sound, and touch enable robots to better understand their environments, navigate complex unscripted scenes, and interact more naturally with humans.
- **Immersive Environments:** In VR and AR, audio-first design ensures that digital experiences are as plausible and convincing as the real world, reducing the cognitive strain associated with sensory-mismatched interfaces.

  By systematically reviewing modern approaches to audio source separation, tagging, and generative modeling, this report provides the technical roadmap for the SENSE platform to bridge the gap between auditory data and rich, multisensory digital experiences.

We structure our state-of-the-art study into the following subject-related sections:

- **Source Separation** (Section 2): We review modern approaches for isolating musical stems from mixtures, from classical matrix factorization pipelines to deep learning architectures and end-to-end waveform-domain models such as Demucs, which established a strong benchmark for music source separation [1].
- **Music Tagging** (Section 3): We examine methods for large-scale multi-label classification of music audio, focusing on convolutional and transformer-based architectures that learn semantic representations directly from raw or time–frequency features, following influential end-to-end tagging frameworks such as [2].
- **Music Information Retrieval** (Section 4): This section surveys core MIR tasks—including beat tracking, pitch estimation, onset detection, and melodic analysis—highlighting the transition from handcrafted features to learned audio representations and self-supervised frameworks such as contrastive predictive coding, which have reshaped modern MIR pipelines [3].
- **Automatic Speech Recognition** (Section 5): We review recent advances in end-to-end ASR, emphasizing the shift toward self-supervised representation learning frameworks such as wav2vec 2.0, which dramatically improved low-resource and large-scale speech recognition performance [4].
- **Audio-Visual Content Generation** (Section 6): We describe generative systems capable of producing coherent audio from visual or multimodal cues, with audio captioning and cross-modal generative modeling playing a key role in building aligned multimodal representations, as demonstrated in the AudioCaps framework [5].
- **Foundation Models** (Section 7): We discuss large-scale pretrained models that unify audio, language, and multimodal understanding, including contrastively trained audio–text models such as CLAP, which offer strong zero-shot capabilities and form an emerging foundation for general-purpose audio intelligence [6].

2 Source Separation

2.1 Moises-Light: Resource-efficient Band-split U-Net For Music Source Separation

This paper [7] presents a lighweight model, namely, Moises-Light, that achieves competitive results in Music Source Separation (MSS) for MUSDB-HQ dataset [8], with a reduced number of parameters by up to 13 times lower. Building on the foundation of DTTNet [9], they integrated insights from previous research to enhance performance. More specifically, they incorporated band-splitting techniques inspired by BSRNN [10] and BS-RoFormer [11], integrated RoPE transformer blocks for sequence modeling, adopted the encoder-decoder design from SCNet [12], and implemented training strategies informed by various prior studies. By carefully optimizing architectural design and training strategies, this lightweight model, with around 5 million parameters of a single stem model, achieves competitive results on the MUSDB-HQ benchmark dataset.

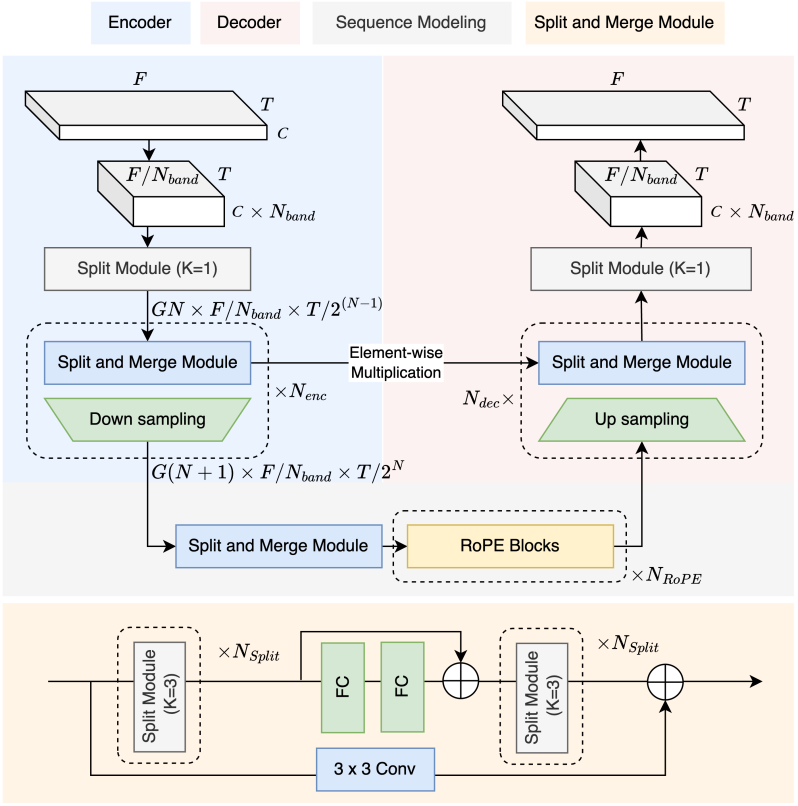


Figure 1: The overall architecture of the Moises-Light model.

Moises-Light is build upon DTTNet [9], but incorporated RoPE transformer blocks and multi-band mask estimation instead of dual-band RNN. Using insights from SCNet [12], they developed an asymetrical encoder-decoder architecture, where the encoder is heavier than the decoder, enabling more effective feature extraction from the input. They also replaced the original TFC-TDF V3 blocks with their proposed split and merge modules. Moise-Light architecture is represented in Fig. 1. Additionally, they improved the training pipeline by incorporating data augmentation techniques and employing a multi-resolution complex spectrogram mean absolute error (MAE) [13] loss function.

Although not reflected on the SDRs, this proposed band-splitting sometimes introduces perceivable artifacts. Another aspect to consider is the fact that their proposed model is single stem, meaning that they require a model for each instrument that must be separated, resulting in 4 models for the general MSS task using MUSDB-HQ dataset. One remarkable thing to note is the fact that their model is much smaller, with only 5 million parameters, compared to other state of the art architectures. With their new lightweight model, they managed to surpass the previous models and define the new state of the art results.

2.2 SCNet: Sparse Compression Network for Music Source separation

The authors propose a method called SCNet [12], a novel frequency-domain network to explicitly split the spectrogram of the mixture into several subbands and introduce a sparsity-based encoder to model different frequency bands. Unlike existing methods, which incompletely address frequency differences or lose information in the process of generating the subband, SCNet splits the spectrogram into three distinct subbands (low, medium, high) and applies differential compression to reflect the information density of each. The low subband, rich in relevant details, is modeled with more parameters, while the medium and high subbands receive stronger compression for efficiency.

The architecture combines a sparse compression encoder, a dual-path RNN-based separation network, and a symmetric decoder, using top-down, bottom-up, and skip connections for efficient information integration. This approach significantly reduces model complexity while preserving essential signal details.

Evaluations on MUSDB18-HQ show that SCNet achieves 9.0 dB SDR, outperforming state-of-the-art models without using additional data. In addition, CPU inference requires only 48% of the time required by HT-Demucs, making SCNet a powerful and efficient solution for resource-constrained applications.

The proposed method, SCNet, aims to separate audio sources in the frequency domain using an architecture composed of three main stages: encoder, separation network and decoder. The input signal is first transformed into a complex spectrogram via STFT, which serves as a representation for the entire processing flow.

The encoder applies three sparse down-sampling (SD) blocks, specifically designed for non-uniform frequency compression. Each SD block uses three parallel convolutions, with different layers, which produces stronger compression in the high-frequency areas and preserves more detail in the low frequencies. The blocks include Conformer-inspired convolutional modules, with GroupNorm and additional layers for low frequencies, maximizing modeling fidelity where information is denser. The design of the convolution module is inspired by the Conformer [14].

In the separation stage, after sufficient spectrogram compression and uniform information density, a dual-path RNN architecture, derived from BSRNN, is used. To avoid the diminishing utility of layer accumulation, transformations are integrated between layers using a mechanism similar to the space conversions in TFCNet. This alternate projection between the time domain and the frequency domain improves the learning of global dependencies.

The decoder reconstructs the separated spectrogram through a series of sparse up-sampling (SU) layers, symmetric to the encoder. The skip connections include a fusion module based on 2D convolutions and GLU, which improve the integration of multi-level information. The last layer adapts the dimensionality for the number of targeted sources.

The model training uses an RMSE loss function applied directly to complex spectrograms, evaluating the real and imaginary components separately. This choice aligns the optimization objective with the spectral representation processed by the architecture, replacing the waveform criteria frequently used in previous works.

The evaluations were performed on MUSDB18-HQ [8], the most widely used dataset for binaural music separation, which contains 150 complete tracks along with individual stems (drums, bass, vocals, etc.). To evaluate generalization and performance gain, MoisesDB, an extended set of 240 multitrack tracks from 12 genres, was also used, allowing testing of the model for both additional performance and adaptability.

In the training setup, the tracks are divided into 11-second segments with 1-second overlap, and the remix and scale augmentations, taken from Demucs, increase the diversity of the data by dynamically combining stems and volume variations. The STFT uses a 4096-point window (92 ms) and 23 ms hop, generating 2049 frequency bands. The encoder applies three SD layers with feature sizes of 32, 64, and 128, and the separation uses 6 dual-path RNN modules, alternating between 128 and 256-unit BiLSTMs. All experiments are run on 8 Nvidia V100 GPUs. Training on MUSDB18-HQ requires 130 epochs, with Adam optimizer and initial learning rate of 5e-4, adjusted to 3e-4 when MoisesDB is added to avoid gradient instability.

The performance is evaluated using the Signal-to-Distortion Ratio (SDR) [15], calculated with museval, and the efficiency is measured by the Real-Time Factor (RTF), which indicates the ratio of processing time to the input signal duration. The study investigates the impact of sparse compression and modeling parameters on the performance of SCNet in separating musical sources.

Experiments show that reducing the global compression ratio (GCR) significantly improves performance, but increases computational cost. The choice of the low-band ratio is crucial: maximum

performance occurs around 17.5%. Below this value, critical information is lost, and above it, the mid-band is reduced too much. The optimal configuration selected is GCR = 70% and low-band = 17.5

SCNet achieves 9.0 dB SDR average without additional data, outperforming all previous models, especially in separating drums and bass, and remaining competitive on vocals. The extended SCNet-large model increases the performance to 9.69 dB. With additional training on MoisesDB, the performance increases further. In addition, SCNet is a very efficient model, with only 10M parameters, about four times less than HT-Demucs.

Generalization test on MoisesDB. Without fine-tuning and without additional data, SCNet outperforms HT Demucs trained with 800 extra songs, achieving 10.33 dB SDR, indicating superior generalization ability to novel material.

The evaluations confirm the critical role of sparse compression: its removal leads to drastic performance drops (from 8.52 dB to 6.20 dB SDR on average), highlighting the benefits of the proposed strategy. Overall, SCNet offers a remarkable balance between performance, computational efficiency, and generalization ability, demonstrating superiority over reference methods.

In conclusion, experimental results show that SCNet outperforms existing reference methods on MUSDB18-HQ while maintaining low computational cost—both with and without additional training data. These results highlight the potential of the architecture for fast and accurate music separation applications.

2.3 Music source separation with band-split rope transformer

In this work [11] it is proposed a new architecture, Band-Split RoPE Transformer (BS- RoFormer), for separation sources (Music Source Separation – MSS), which combines frequency band processing with the advantages Transformers Modern and introduces Rotary Position Embedding (RoPE) as a mechanism essential for stabilizing learning in a frame time – frequency. In Background in which MSS performance stagnated in the last few years, and the models dominant (for example Demucs or BSRNN) is either based on processing in field of time, either on architectures recurring limited in capturing long -term addictions, the work demonstrate that careful design calibrated Transformers maybe brings progress significant.

BS- RoFormer architecture operates fully in field frequency, starting from the observation that instruments musical busy distributions spectral different and that treatment uniform throughout spectrograms — approach use frequent in the models conventional does not capitalize structure intrinsic sound musical. To exploit these particularities, spectrogram complexity is divided in N non- overlapping frequency bands, with granularity adapted density spectral: bands finer at frequencies low, where information is complexity and criticism, and bands may wide at frequencies high. Each band is then project in a latent space through MLPs independent, which allows model saddle learn representations specialized and contextualised for each region spectral.

Processing future is accomplished through a hierarchy of Transformers that alternate between modeling time and inter- band modeling. In first row, a time-Transformer treated each individual tape, capturing dependencies local and temporal improving coherence time a estimates. In the second row, a subband -Transformer shape interactions spectral from lanes, allowing information exchange between frequency regions that contribute together in training stamp of an instrument or when superimposed sound of more several sources. Alternating these two types of processing constitutes a mechanism hierarchical efficient for capturing addictions complexity in area time – frequency. An innovation critical architecture is integration Rotary Position Embedding (RoPE), which replaces the absolute positional embeddings used traditionally in Transformers. The authors show that the model equivalent without RoPE — BS-Transformer — suffers from instability severe and fails saddle converge in training, suggesting that RoPE is the only one mechanism compatible with positioning structure dimensional implied in the processes alternative time and spectral. RoPE introduces rotations position dependent in attention projections, preserving the necessary relative relationships for processing coherence of segments spectral and temporal. This finding is important methodolog major, indicating that the design positional is a key factor in Transformers for audio.

After processing, each band is passed through a dedicated MLP to generate a mask complex (cIRM), including both part real, how much and the one imaginary line of the separate spectrograms. The bands are then concatenated to form the spectrogram complete output, which is converted back in audio signal through inverse Fourier transform (iSTFT). The authors adopt a

deframing method by "overlap and average", preferred concatenation methods direct, because it produces transitions smoother between windows and avoid the artifacts perceptual.

For training, model use both the standard set MUSDB18HQ, how much and a body additional approximately 500 internal songs, which grow diversity stylistic. The large size of the model — especially the complete with 12 Transformer blocks — requires extensive hardware resources (16 A100 GPUs for four weeks). To manage costs computational, techniques such as checkpointing, FlashAttention are integrated and drive in precision mixed. It is also adopted a scheme augmentation complex based on a dynamic pool, in which individual stems are mixed randomly to generate examples us at each era. Experimental results are remarkable. BS- RoFormer obtained place first in competition International SDX'23, recording an average SDR of 9.97 dB, with a lead of +0.71 dB over the next system — a difference significantly in MSS. For sources such as voice, bass and drums, winnings Average exceed +0.9 dB, indicating a quality perceptual superior. Version extended version of the model, trained with additional data, achieves 11.99 dB SDR, a new absolute record reported for MSS in field frequency.

Ablation experiments confirm role essential of RoPE and of strip architecture. A variant reduced by just 6 Transformer blocks exceeds performance patterns established even and without additional data, demonstrating strength design. Also, the authors notice that the model produces spectrograms may clear and with more few residues, although sometimes they can appear "sharp" artifacts, suggesting potential directions for improvement by overlap the bands.

In conclusion, BS- RoFormer represents a contribution major in segmentation sources musical, both by performance obtained, as and by demonstration importance design positional and tape processing in Transformer architectures for audio. The model sets a new standard in MSS and provides a framework that can be extended to a range wide range of time–frequency tasks from processing audio signals.

2.4 Band-SCNet: A Causal, Lightweight Model for High-Performance Real-Time Music Source Separation

Music Source Separation (MSS) [16] is a central problem in audio signal processing, aiming at the decomposition of a musical mixture into its distinct components, such as voice, drums, bass, or instrumental accompaniment. Although recent advances have led to considerable performance gains in offline systems, the application of these methods in real time still presents significant challenges, mainly due to latency constraints and computational limitations. As a result, real-time models tend to be considerably less accurate than their non-real-time counterparts.

In this context, the paper introduces *Band-SCNet*, a lightweight model (2.59 million parameters) specifically designed for real-time music source separation. The model is based on the causal version of the SCNet architecture–referred to as *Online SCNet*–but integrates several architectural optimizations aimed at improving both efficiency and accuracy. In particular, Band-SCNet combines sparse spectral compression (Sparse Compression) with specialized cross-band and narrow-band processing modules, as well as a novel fusion mechanism inspired by multi-head attention, named the *CSA Fusion Module*.

Existing real-time architectures exhibit three major limitations. First, most systems employ uniform spectral compression across the entire frequency range, neglecting the fact that low frequencies contain more stable and informative musical content, which is crucial for accurate separation. Second, the use of relatively short STFT windows prevents adequate capture of low-frequency dynamics. Third, many networks operate directly on the full spectrogram, without exploiting inter-band relationships or introducing adaptive multi-band modeling strategies.

SCNet, the core model underlying Band-SCNet, addresses part of these challenges through Sparse Compression, a technique that applies non-uniform downsampling depending on the spectral region. It preserves a dense representation at low frequencies while applying more aggressive compression at mid and high frequencies. However, adapting SCNet into a fully causal architecture requires several modifications—including the replacement of standard convolutions with causal variants, the use of cumulative normalization, and the removal of operations relying on full-signal context—all of which lead to a noticeable drop in performance.

Band-SCNet retains the encoder–decoder structure of Online SCNet, including the Sparse Downsampling and Sparse Upsampling mechanisms, which ensure high resolution in critical spectral regions. The major innovation appears within the *Separation Network*, which alternates two types of specialized processing blocks:

- **Cross-band Blocks**, designed to capture relationships between adjacent frequency bands. These blocks employ frequency-domain convolutional modules (F-Conv) together with a full-band linear transformation, enabling the model to detect harmonic interactions essential for musical instrument separation.
- **Narrow-band Blocks**, focused on modeling each band independently. They integrate a Multi-Head Self-Attention (MHSA) module and a temporal feed-forward network, offering a finer representation of temporal characteristics.

The combination of these two module types is made possible by the underlying non-uniform spectral compression, which reduces the dimensionality of the input in less informative frequency regions while preserving accuracy where it matters most.

In the decoding stage, the authors replace the original SCNet fusion module with the *CSA Fusion Module*, which aims to improve the integration of skip-connection information with the representations produced by the separation network. The module employs a compressed variant of multi-head attention (CMHSA), complemented by a linear gating mechanism and a GLU unit. This approach significantly reduces the parameter count while simultaneously yielding more accurate spectrogram reconstruction.

The model was evaluated on the MUSDB18-HQ dataset, where it achieved an SDR of 7.79 dB, thereby establishing a new benchmark for real-time music source separation methods. The system latency of 92 ms and a real-time factor (RTF) of 0.478 further demonstrate its suitability for resource-constrained real-time applications. Compared to other real-time models, such as Wave-U-Net or HS-TasNet, Band-SCNet stands out for both its superior accuracy and improved computational efficiency. Relative to Online SCNet, the approximately 0.7 dB performance gain highlights the contribution of each architectural component introduced.

In conclusion, Band-SCNet represents a notable contribution to the field of real-time music source separation. By integrating a hybrid multi-band architecture and an efficient compressed-attention mechanism, the model significantly reduces the performance gap between real-time and offline approaches. This work opens promising perspectives for the future development of interactive audio systems, where achieving a balance between low latency and high accuracy remains essential.

2.5 Conv-TasNet: Time-Domain Speech Separation Beyond Ideal T–F Masking

The paper by Luo and Mesgarani [17] introduces *Conv-TasNet*, a fully convolutional, end-to-end time-domain architecture that surpasses the long-standing ideal time–frequency (T–F) magnitude masking limits in single-channel speech separation. Unlike prior deep separation models reliant on the short-time Fourier transform (STFT), Conv-TasNet eliminates the phase–magnitude decoupling inherent to T–F representations and avoids the high-latency constraints imposed by spectrogram windows, offering a fundamentally different paradigm for speech separation. The model leverages a learnable 1-D convolutional encoder–decoder to construct an overcomplete representation of the mixture waveform, which is then processed by a temporal convolutional network (TCN) composed of stacked dilated convolutional blocks. Depthwise separable convolutions significantly reduce model size while enabling large receptive fields, allowing the system to model long-range temporal dependencies essential for speech structure.

The authors evaluate Conv-TasNet on the standard WSJ0-2mix and WSJ0-3mix benchmarks [18] :contentReference[oaicite:1]index=1, using mixtures generated from the WSJ0 corpus at varying SNR levels. The system is trained with utterance-level permutation invariant training (uPIT) and optimized using the scale-invariant signal-to-noise ratio (SI-SNR). Objective evaluation uses SI-SNR improvement (SI-SNRi) and SDR improvement (SDRi), and subjective perceptual quality is measured via PESQ and human mean opinion scores (MOS).

Experimental results show that noncausal Conv-TasNet achieves dramatic improvements over all prior STFT-based methods, including surpassing ideal binary, ideal ratio, and Wiener filter-like magnitude masks. On WSJ0-2mix, Conv-TasNet reaches 15.3 dB SI-SNRi and 15.6 dB SDRi—substantially higher than BLSTM-TasNet and all T–F masking baselines—while using only 5.1M parameters. A causal version maintains real-time feasibility with sub-frame latency (0.4 ms TPF on CPU), outperforming previous causal LSTM systems in both speed and robustness to mixture start-time shifts. Human listening tests further reveal that Conv-TasNet's MOS exceeds that

of IRM-separated audio, highlighting its superior perceptual fidelity despite PESQ's bias against time-domain methods.

The analysis of encoder and decoder basis functions reveals that the learned filters disproportionately emphasize low-frequency components and encode explicit phase diversity—properties difficult to represent in STFT magnitude-only systems. This suggests that phase information and fine-grained low-frequency cues such as pitch are central to the model's success. The authors note limitations including handling long-term speaker tracking and robustness in highly reverberant or noisy environments, pointing toward future extensions such as multi-channel Conv-TasNet architectures.

Overall, Conv-TasNet represents a major conceptual and practical advance in speech separation: a low-latency, small-footprint, high-accuracy time-domain model that redefines what is achievable without spectrograms. Its impact has been broad, influencing subsequent research in speech enhancement, multi-talker ASR front-ends, and real-time applications on embedded and wearable devices.

2.6 Demucs: Music Source Separation in the Waveform Domain

The paper [1] investigates music source separation directly in the waveform domain, challenging the long-standing dominance of spectrogram-based masking approaches. The task consists of decomposing a musical mixture into four constituent stems—*vocals, drums, bass, and other*—given fully supervised training data. While prior state-of-the-art systems primarily rely on short-time Fourier transform (STFT) representations and magnitude masking, the authors explore whether end-to-end waveform-to-waveform modeling can achieve competitive or superior performance while avoiding artifacts induced by phase reuse and spectrogram masking.

As a first step, the paper adapts *Conv-TasNet*, originally designed for monophonic speech separation, to stereophonic music source separation at 44.1 kHz. Although this adapted Conv-TasNet surpasses many spectrogram-domain baselines in terms of signal-to-distortion ratio (SDR), extensive listening tests reveal significant artifacts, including broadband noise, hollow transients, and missing notes, particularly for drums and bass. These observations motivate the introduction of a novel architecture, *Demucs* (Deep Extractor for Music Sources), explicitly designed for music separation in the waveform domain.

Demucs is a waveform-to-waveform encoder–decoder model with a U-Net structure. It employs a deep convolutional encoder with progressively increasing channel widths, a central bidirectional LSTM to capture long-range temporal dependencies, and a convolutional decoder with transposed convolutions and skip connections. Gated Linear Units (GLUs) are used to increase expressivity, while skip connections allow the model to preserve fine temporal details and implicitly reuse phase information from the input signal. The authors further introduce several important design choices, including resampling within the network, a tailored weight rescaling initialization scheme, and a *shift trick* to improve time-shift equivariance at inference.

Training and evaluation are conducted on the MUSDB18 benchmark, consisting of 150 stereo tracks with official train, validation, and test splits. Models are trained using simple waveform-level regression losses (primarily L1), avoiding spectrogram-based objectives. Performance is evaluated using the SDR metric from the BSS Eval framework, following the SiSEC 2018 protocol. Demucs achieves an average SDR of 6.28 dB on the MUSDB test set without extra data, outperforming all prior waveform-domain methods and surpassing the best spectrogram-based systems. When augmented with an additional 150 training songs, Demucs reaches up to 6.79 dB SDR and, notably, exceeds the Ideal Ratio Mask (IRM) oracle for the bass source.

Beyond objective metrics, the paper places strong emphasis on perceptual evaluation. Mean Opinion Score (MOS) listening tests demonstrate that Demucs produces more natural-sounding separations with fewer audible artifacts than Conv-TasNet, although it exhibits slightly higher source leakage in some cases. An extensive ablation study highlights the critical role of the bidirectional LSTM, pitch/tempo data augmentation, GLU activations, and the proposed initialization strategy in achieving state-of-the-art performance.

The authors acknowledge several limitations. Demucs models are substantially larger than many competing approaches, leading to high memory usage, although quantization techniques can reduce model size significantly without degrading performance. Additionally, waveform-domain models still exhibit trade-offs between artifact suppression and source leakage, particularly for harmonically rich sources such as vocals. Overall, the paper establishes Demucs as a foundational

architecture for music source separation, demonstrating for the first time that carefully designed waveform-domain models can outperform spectrogram-based systems both objectively and perceptually, and laying the groundwork for subsequent hybrid and transformer-based extensions.

2.7 Multi-Source Diffusion Models for Simultaneous Music Generation and Separation

The paper [19] introduces a unified generative framework for music source separation and music generation based on score-based diffusion models. Departing from the traditional separation between discriminative source separators and unconditional music generators, the authors propose a single probabilistic model capable of performing three tasks at inference time: *(i)* total music generation, *(ii)* partial generation or source imputation (accompaniment generation), and *(iii)* music source separation. The central idea is to explicitly model the joint distribution of musical sources belonging to the same mixture, thereby capturing the contextual dependencies that naturally arise between instrumental stems.

As illustrated in Figure 2, the Multi-Source Diffusion Model learns the joint distribution of contextual audio sources, enabling both generation and separation within a unified diffusion framework.

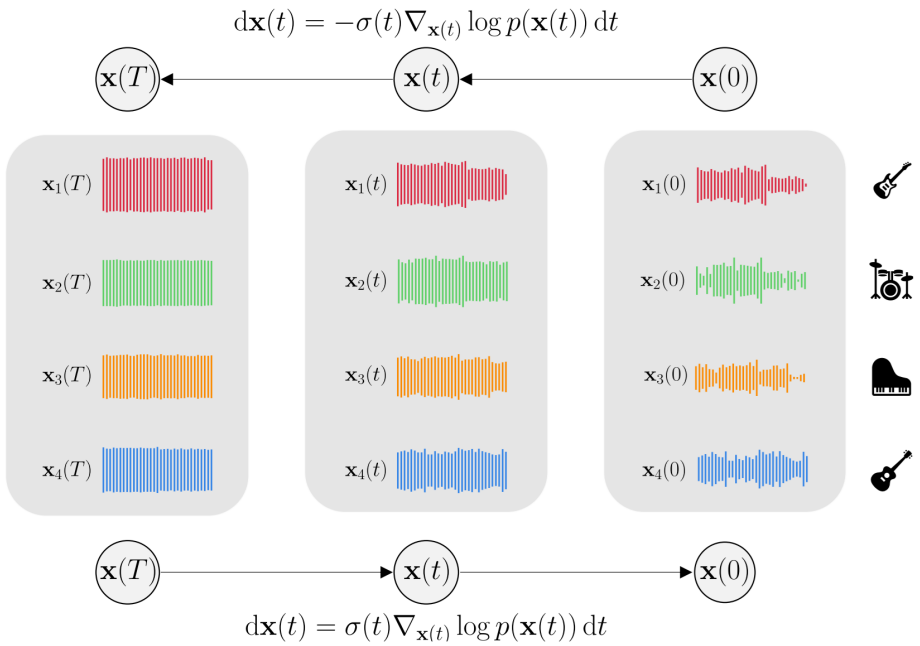


Figure 2: Overview of the Multi-Source Diffusion Model (MSDM). The forward diffusion process progressively corrupts multiple audio sources with Gaussian noise, while the reverse process jointly denoises all sources, enabling total generation, partial generation, and source separation within a single score-based framework.

Formally, the method targets the joint prior distribution $p(x_1, \ldots, x_N)$ over N source waveforms, where the mixture is given by their sum. Unlike conventional generative approaches that directly model the mixture distribution $p(y)$, or separation models that condition on the mixture by design, the proposed Multi-Source Diffusion Model (MSDM) learns an unconditional prior over source sets using denoising score matching. This design choice enables unconditional sampling for music generation, while also allowing conditioning on observed information during inference without architectural constraints.

The model is implemented as a score-based diffusion process operating directly on raw audio waveforms. Gaussian noise is progressively added to the stacked source signals during the forward diffusion process, and a neural score network is trained to approximate the gradient of the log-density of the perturbed joint distribution. The score network follows a U-Net–like convolutional

architecture inspired by recent diffusion models for audio, but is adapted to handle multiple sources jointly. By reversing the diffusion process through numerical integration of the corresponding probability flow ODE, the model can generate coherent sets of instrumental stems or sample from conditional distributions defined at inference time.

A key technical contribution of the paper is a novel inference procedure for source separation based on a Dirac delta likelihood. Instead of modeling the mixture likelihood with a Gaussian distribution, as done in prior generative separators, the authors exploit the deterministic relationship between sources and mixture by enforcing a hard constraint that their sum equals the observed mixture. This *Dirac likelihood* leads to a modified posterior score that improves separation performance and yields results competitive with state-of-the-art discriminative models. The authors also show that the same formulation naturally supports partial generation, where a subset of sources is fixed and the remaining stems are generated consistently.

Experiments are conducted primarily on the Slakh2100 dataset, a large-scale multitrack music dataset containing approximately 145 hours of synthesized music with aligned instrumental stems. The model is evaluated on music generation using both objective metrics, such as Fréchet Audio Distance (FAD) and a proposed sub-FAD metric, and subjective listening tests assessing quality, coherence, and density. For source separation, performance is measured using the scale-invariant signal-to-distortion ratio improvement (SI-SDRi), following established evaluation protocols. Results demonstrate that MSDM achieves separation quality comparable to strong regressor-based baselines while simultaneously supporting generation and accompaniment tasks within the same model.

The paper highlights several limitations and directions for future work. The quality of the learned joint prior is constrained by the availability of large-scale contextual multitrack data, and waveform-based diffusion models remain computationally expensive at inference time. Nevertheless, the proposed framework represents the first demonstration of a single diffusion-based model that unifies music generation, source separation, and source imputation, positioning MSDM as an important step toward general-purpose generative audio models with compositional control.

2.8 Open-Unmix: A Reference Implementation for Music Source Separation

The paper [20] presents *Open-Unmix*, an open-source reference implementation for music source separation based on deep neural networks. The addressed task consists of decomposing a musical mixture into its constituent stems—typically *vocals, drums, bass, and other*—under a fully supervised learning setting. While deep learning had already demonstrated state-of-the-art performance in music source separation, the authors identify a critical gap in the field: the absence of a freely available, well-engineered, and reproducible open-source system that matches the performance of the best proprietary or unpublished methods. Open-Unmix is explicitly designed to close this gap and to serve as a strong and transparent baseline for future research.

The system operates in the frequency domain and relies on the magnitude of the short-time Fourier transform (STFT) as input representation. In contrast to waveform-domain approaches, Open-Unmix follows the dominant paradigm of spectrogram-based separation, focusing on stable and interpretable design choices that have been validated by the community. The core model architecture is based on a bidirectional long short-term memory (BiLSTM) network, inspired by prior work on deep neural network-based instrument separation. Fully connected layers are used to map time–frequency representations to source-specific magnitude estimates, which are subsequently combined with the mixture phase to reconstruct time-domain signals.

A central design goal of Open-Unmix is to balance competitive performance with conceptual simplicity and extensibility. The authors emphasize modularity and clarity over architectural novelty, allowing researchers to easily modify or replace individual components such as data loading, preprocessing, or the neural network architecture itself. The reference implementation is provided primarily in PyTorch, with additional ports planned or available for other deep learning frameworks, reinforcing the role of Open-Unmix as a framework-agnostic baseline rather than a tightly coupled software package.

Training and evaluation are conducted on the MUSDB18 dataset, which comprises 150 full-length stereo music tracks and represents the largest freely available benchmark for music source separation at the time of publication. Open-Unmix is trained exclusively on MUSDB18 and does not rely on external proprietary data. The authors also provide pre-trained model weights, enabling

immediate use of the system or optional fine-tuning on user-provided datasets. Performance is evaluated using the signal-to-distortion ratio (SDR) metric from the BSS Eval framework, following the established evaluation protocols used in the Signal Separation Evaluation Campaigns (SiSEC).

Empirical results demonstrate that Open-Unmix achieves state-of-the-art performance among open-source systems on MUSDB18. In particular, the reported SDR scores are statistically indistinguishable from those of the best-performing system submitted to the most recent SiSEC evaluation, while surpassing all other publicly available methods. This result establishes Open-Unmix as the first open-source implementation to reach parity with top-performing but unreleased or proprietary systems, validating its role as a strong baseline for the research community.

Beyond raw performance, the paper places strong emphasis on reproducibility and community adoption. The release includes pre-trained models, automated tests, and detailed documentation, and is embedded within a broader ecosystem of open datasets, evaluation tools, and community-driven benchmarks maintained by the authors. Rather than pursuing aggressive architectural optimization, the authors deliberately scope Open-Unmix as a stable reference system, encouraging researchers to fork and extend the codebase when exploring novel representations or learning architectures.

Overall, the paper positions Open-Unmix not as a final solution, but as an essential infrastructural contribution to the field of music source separation. By providing a transparent, reproducible, and state-of-the-art open-source baseline, Open-Unmix enables meaningful comparison between methods and supports cumulative scientific progress in audio source separation research.

2.9 Band-split RNN (BSRNN)

This paper introduces Band-Split RNN (BSRNN) [10], a frequency-domain architecture designed for high-sample-rate music source separation (MSS) that addresses the limitations of applying speech-centric models to musical signals. The research emphasizes integrating a priori knowledge of musical instrument characteristics into model design through flexible frequency-band modeling.

The BSRNN architecture is composed of three primary stages:

- Band Split Module: Unlike standard models that treat all frequency bins equally, BSRNN splits the complex-valued spectrogram into K subbands with varying bandwidths based on the target instrument3333. Each subband is normalized and transformed into a unified feature dimension N.
- Band and Sequence Modeling: This module utilizes interleaved residual RNN layers. A sequence-level RNN captures temporal dependencies across frames (shared across subbands to save parameters), while a band-level RNN models intra-band dependencies across the frequency dimension.
- Mask Estimation: The system estimates a complex-valued time-frequency mask for each subband using a multilayer perceptron (MLP). These masks are merged and applied to the original mixture spectrogram to extract the target source.

A core contribution is the finding that fine-grained band-splitting at lower frequencies significantly improves performance. Splitting frequencies below 1 kHz into narrow 100 Hz bands allows the model to better capture fundamental frequencies and harmonics for vocals stems. Different schemes were developed for Bass (30 subbands) and Drums (55 subbands) to match their unique spectral patterns.

To overcome the scarcity of high-quality labeled data, the authors propose a self-boosting semi-supervised pipeline:

- Dual-Role Model: A pre-trained model serves as both a pseudo-label generator and a source activity detector.
- Filtering: An energy-based method filters unlabeled data; segments with high energy differences (e.g., >30 dB) are treated as clean target or residual samples, while others provide pseudo-labels.
- Iterative Improvement: The model is updated continuously, replacing the "teacher" model whenever a new "student" achieves superior validation performance.

Evaluated on the MUSDB18-HQ dataset using uSDR and cSDR metrics, BSRNN demonstrated great results, illustrating the performance of Band-Split methods.

2.10 Hybrid Transformer Demucs (HTDemucs)

This paper introduces Hybrid Transformer Demucs (HT Demucs) [21], an architecture that evolves the bi-U-Net structure of Hybrid Demucs by integrating Transformer layers to better capture long-range contextual information in music signals.

HT Demucs utilizes a dual-domain approach, processing signals in both the temporal (waveform) and spectral (spectrogram) domains simultaneously.

- Hybrid Structure: The model retains the outermost four layers of the original Hybrid Demucs bi-U-Net but replaces the two innermost layers with a Cross-domain Transformer Encoder.
- Dual Attention: Within the Transformer block, the model interleaves self-attention (to model dependencies within a single domain) and cross-attention (to integrate information across the temporal and spectral domains).
- Flexibility: Unlike convolutional layers that require strict alignment between representations, the Transformer encoder can process heterogeneous data shapes, making it more flexible.

Transformers are traditionally memory-intensive, especially for high-resolution 44.1 kHz audio. To address this, the authors used:

- Sparse Attention: The authors leverage sparse attention kernels and a Locally Sensitive Hashing (LSH) scheme.
- Efficiency: This scheme dynamically determines a sparsity pattern (removing 90% of elements in the softmax), allowing the model to process input lengths of up to 12.2 seconds during training without exceeding memory limits.

The study emphasizes that Transformer-based architectures are "data-hungry" and perform poorly when trained solely on the standard MUSDB18 dataset.

- Dataset Expansion: The researchers used an internal dataset of 800 curated songs on top of MUSDB18-HQ to reach competitive performance.
- Data Augmentation: Techniques such as remixing (shuffling stems within a batch) and repitching remained critical; removing the remixing augmentation resulted in a significant 0.7 dB SDR loss.

HT Demucs achieved good results on the MUSDB18-HQ test set. While it surpasses previous versions of Demucs and Spleeter, it competes closely with the Band-Split RNN, which achieved 8.97 dB using a different augmentation approach. Compared to other newer methods, this algorithm offers lower performance than state-of-the-art models like BS-Roformer and SCNET.

2.11 Conclusion

The review of current music source separation (MSS) methodologies reveals a field in rapid transition, moving from classical signal processing techniques toward highly sophisticated, data-driven neural architectures. This chapter has examined a diverse array of models, ranging from resource-efficient designs to large-scale foundation architectures, highlighting several key trends that define the current state-of-the-art.

A primary shift identified in this research is the move away from traditional time-frequency masking on spectrograms toward end-to-end waveform-domain processing. While early successful models like Open-Unmix relied on frequency-domain representations, newer benchmarks established by architectures like Demucs and Conv-TasNet demonstrate the advantages of modeling audio directly in the time domain. This approach preserves phase information more effectively, leading to higher perceptual quality and reduced artifacting in isolated stems.

The integration of Transformers and attention mechanisms has set new performance benchmarks. The Hybrid Transformer Demucs (HTDemucs) represents a significant milestone by combining the local feature extraction capabilities of Convolutional Neural Networks (CNNs) with the global dependency modeling of Transformers. Despite their high computational and memory demands, the adoption of sparse attention kernels and Locally Sensitive Hashing (LSH) has enabled these models to process high-resolution audio (44.1 kHz) over extended temporal windows.

Furthermore, band-splitting techniques, as seen in BSRNN and BS-RoFormer, have proven essential for managing the high dimensionality of audio signals by processing frequency sub-bands independently before sequence modeling.

The analysis highlights an emerging divergence between "heavyweight" models optimized for maximum Signal-to-Distortion Ratio (SDR) and "lightweight" models optimized for real-time deployment. Models such as Moises-Light and Band-SCNet demonstrate that it is possible to achieve competitive results on benchmarks like MUSDB18-HQ while using up to 13 times fewer parameters than their predecessors. This efficiency is achieved through clever architectural optimizations, such as the Sparse Compression Network (SCNet) design and causal processing, making high-quality separation accessible for consumer-grade hardware and low-latency applications.

Finally, this chapter underscores that architectural innovation alone is insufficient. State-of-the-art performance is heavily contingent on large-scale dataset expansion and robust data augmentation strategies. Techniques such as stem remixing and repitching remain critical; as seen in the evaluation of HTDemucs, the absence of remixing can lead to a significant drop in SDR (e.g., 0.7 dB). The transition toward diffusion-based models, such as Multi-Source Diffusion Models, further suggests a future where separation and generation are treated as dual tasks, allowing models to “hallucinate” missing spectral components to achieve unprecedented clarity.

In summary, the field is converging toward hybrid models that balance the efficiency of sub-band processing with the expressive power of Transformers, supported by massive, curated datasets and sophisticated temporal modeling.

3 Music Tagging

3.1 PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition

The paper [22] presents *PANNs* (Pretrained Audio Neural Networks), a comprehensive family of deep architectures trained on the large-scale AudioSet dataset [23] for general-purpose audio tagging and transfer learning. Motivated by the growing impact of large-scale pretraining in machine learning, the authors investigate whether weakly labeled audio at scale can serve as an effective supervisory signal for learning transferable acoustic representations. AudioSet, containing 1.9 million ten-second audio clips spanning 527 sound classes, provides a diverse and challenging benchmark due to its long-tailed label distribution and significant label noise.

The authors explore a wide range of architectures to identify effective design principles for audio tagging. These include convolutional networks of varying depths (CNN6, CNN10, CNN14), deeper residual architectures (ResNet22, ResNet38, ResNet54), efficient MobileNet variants, and several one-dimensional raw-waveform models such as DaiNet [24] and LeeNet [25]. The key finding is that deeper architectures significantly outperform shallow ones on large-scale data—CNN14 achieves 0.431 mAP compared to CNN6's 0.343 mAP—contrary to results on smaller datasets where shallow networks avoid overfitting. In addition to evaluating existing architectures, the paper introduces two novel models: *Wavegram-CNN*, which learns hierarchical time-domain features using stacked one-dimensional convolutional filters, and *Wavegram-Logmel-CNN*, which integrates both learned waveform-based representations and conventional log-mel spectrogram inputs. These hybrid models enable the network to benefit from both data-driven feature extraction and established time-frequency representations.

PANNs achieve strong audio tagging performance on AudioSet, with the best model, Wavegram-Logmel-CNN, reaching an mAP of 0.439, outperforming Google's VGGish baseline (0.314 mAP) and the previous state-of-the-art DeepRes system (0.392 mAP). The authors conduct detailed ablation studies examining the effects of architectural depth, receptive field size, embedding dimensionality, sampling and hop parameters, mixup-based data augmentation, and balanced sampling strategies. These experiments demonstrate that deeper architectures, larger embedding dimensions, mixup regularization, and careful handling of the class imbalance substantially improve performance. The paper further analyzes class-wise metrics, revealing large variations in average precision due to differences in label quality and dataset imbalance, emphasizing the challenges inherent in AudioSet's annotation process.

Beyond audio tagging, the authors systematically evaluate the transferability of PANNs across multiple downstream tasks. Fine-tuning or using PANN embeddings leads to strong performance on environmental sound classification (ESC-50), acoustic scene classification (DCASE 2019 Task 1), general-purpose audio tagging (DCASE 2018 Task 2), music/speech/other classification (MSoS), and music genre classification (GTZAN). In many cases, pretrained PANN models outperform counterparts trained from scratch, highlighting the benefits of large-scale weakly supervised pretraining. Notably, transfer learning experiments consistently show that deep pretrained networks provide robust representations across both environmental and music-related tasks.

The authors also reflect on several limitations. AudioSet's weak labels and highly imbalanced distribution introduce significant uncertainty in class-wise metrics. PANNs, particularly the deeper CNN and Wavegram models, remain computationally intensive, requiring tens of billions of multiply-add operations per inference. Furthermore, because AudioSet consists of YouTube-sourced audio, the pretrained models inherit dataset-specific biases that may not generalize uniformly across domains. Despite these challenges, the work establishes PANNs as an early and influential foundation for large-scale audio representation learning, demonstrating both strong tagging performance and broad applicability to downstream audio tasks. However, the reliance on convolutional architectures with local receptive fields requiring careful hierarchical design to capture global context motivated the subsequent shift to transformer architectures capable of attending to all time-frequency regions simultaneously—a paradigm shift exemplified by the Audio Spectrogram Transformer (AST) introduced shortly after.

3.2 AST: Audio Spectrogram Transformer

This paper [26] introduces the Audio Spectrogram Transformer (AST), the first convolution-free, purely attention-based model for audio classification. Unlike previous approaches that combine

CNNs with attention mechanisms, AST directly applies a standard Transformer encoder to audio spectrograms, demonstrating that CNNs are not essential for achieving state-of-the-art performance in audio tasks. Building on the foundation established by PANNs [22], which achieved 0.439 mAP on AudioSet using deep CNNs, AST pushes performance further to 0.459 mAP (single model) and 0.485 mAP (ensemble) by replacing convolutional hierarchies with global self-attention from the first layer. The model is inspired by the Vision Transformer (ViT) [27] but adapted to handle variable-length audio inputs ranging from 1 to 10 seconds.

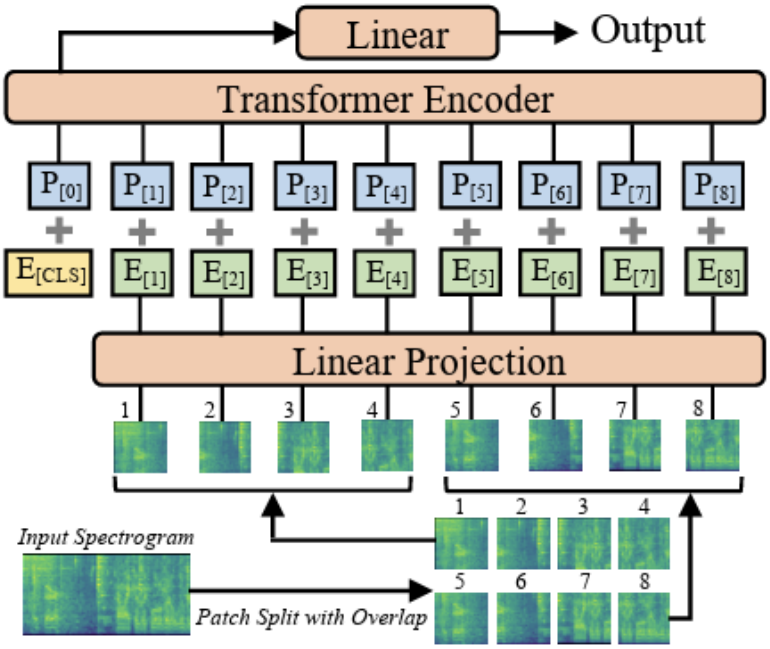


Figure 3: The Audio Spectrogram Transformer (AST) architecture. The 2D spectrogram is split into 16×16 patches with overlap, linearly projected to embeddings, and processed by a standard Transformer encoder.

The AST architecture (illustrated in Fig. 3) splits the input log-mel spectrogram into a sequence of 16×16 patches with an overlap of 6 in both time and frequency dimensions. Each patch is flattened and linearly projected to a 768-dimensional embedding, with learnable positional embeddings added to preserve spatial structure. A [CLS] token is prepended to the sequence, and the entire sequence is processed by a standard 12-layer Transformer encoder with 768 embedding dimensions and 12 attention heads. The output of the [CLS] token serves as the audio representation for classification.

A key innovation is the cross-modality transfer learning approach from ImageNet-pretrained Vision Transformers. The authors adapt pretrained ViT weights through careful handling of input channel differences (averaging RGB channels to single-channel spectrogram weights) and positional embedding interpolation. Specifically, they cut and bi-linearly interpolate the 2D positional embeddings from the fixed-size ImageNet images (e.g., 24×24 patches for 384×384 images) to match the variable-length audio spectrograms (e.g., 12×100 patches for 10-second audio). This transfer learning strategy significantly improves performance, especially when training data is limited, as the authors demonstrate through linear probing experiments (training only a simple linear classifier on top of frozen representations) showing that ImageNet pretraining substantially reduces the need for in-domain audio data.

AST achieves state-of-the-art results across multiple benchmarks: 0.485 mAP on AudioSet (full training set with ensemble), 95.6% accuracy on ESC-50 (with AudioSet pretraining), and 98.1% accuracy on Speech Commands V2. The model demonstrates strong performance even on the smaller AudioSet balanced set (22K samples), achieving 0.347 mAP with a single model, outperforming previous CNN-attention hybrid architectures. Notably, AST uses the same architecture across all tasks despite varying input lengths (1s for Speech Commands, 5s for ESC-50, 10s for AudioSet), requiring only 5 training epochs on AudioSet compared to 30 epochs for previous CNN-

based models. However, AST introduces significant computational challenges. The model requires more training data than CNNs when training from scratch due to Transformers lacking inductive biases like spatial locality and translation equivariance—though ImageNet pretraining mitigates this. More critically, the standard self-attention mechanism has quadratic complexity in sequence length, making training on long audio sequences (up to 1,212 patches for 10-second audio) computationally expensive and memory-intensive. While AST achieves superior performance, training requires approximately 3 days on 4 NVIDIA Tesla-V100 GPUs even with the relatively short 5-epoch schedule. These computational demands motivated subsequent work on efficient training strategies, particularly PaSST's Patchout technique [28], which addresses the memory bottleneck by randomly dropping patches during training to enable processing of longer sequences with limited GPU memory.

3.3 Efficient Training of Audio Transformers with Patchout

This paper [28] addresses the computational limitations of the Audio Spectrogram Transformer (AST) [26] through two key innovations: Patchout, a training-time method that randomly drops patches to reduce sequence length, and disentangled positional encodings that separate time and frequency dimensions. While AST achieved 0.459 mAP on AudioSet, its quadratic attention complexity makes training expensive—requiring approximately 3 days on 4 Tesla-V100 GPUs and 2.33 GB memory per sample. The long sequences (up to 1,212 patches for 10-second audio with overlapping patches) create a memory bottleneck that limits accessibility for researchers without institutional GPU resources.

The paper *Efficient Training of Audio Transformers with Patchout* directly addresses these inefficiencies. The authors introduce **Patchout**, a method for reducing sequence length during training by randomly dropping a portion of the input patches, and propose the **Patchout faSt Spectrogram Transformer (PaSST)** architecture (illustrated in Fig. 4). Patchout reduces training complexity, acts as an effective regularizer, and enables transformer models to be trained on consumer-grade hardware. The resulting PaSST models achieve state-of-the-art performance on Audioset and demonstrate strong generalization when fine-tuned on several downstream audio classification tasks.

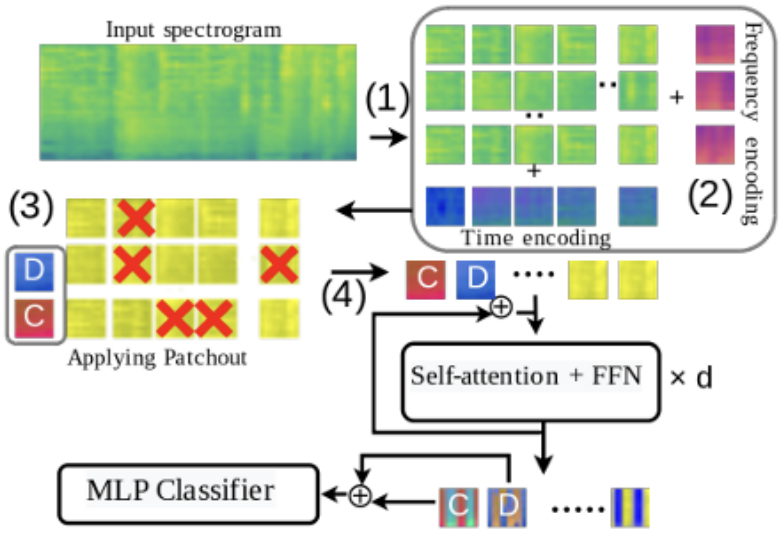


Figure 4: The PaSST architecture showing (1) patch extraction and linear projection, (2) disentangled time and frequency positional encodings, (3) Patchout application where patches are randomly dropped during training, and (4) transformer layers with self-attention. The classification token C (and distillation token D for DeiT-based models) are processed to produce final predictions.

The core contribution of the work is the Patchout technique. Instead of feeding all extracted spectrogram patches to the transformer, Patchout *removes* a subset of them during training. This approach offers three main benefits:

1. **Reduced computational cost.** Since self-attention has $O(n^2)$ complexity with respect to sequence length n, even a modest reduction in n yields a substantial reduction in both runtime and memory usage.

2. **Regularization effect.** By forcing the model to classify audio clips based on incomplete sequences, Patchout improves robustness and prevents overfitting, analogous to techniques such as Dropout or SpecAugment.

3. **Compatibility with inference.** Patchout is applied only during training; the full sequence is restored at inference time, ensuring no loss in representational capacity for prediction.

The paper introduces two Patchout variants:

- **Unstructured Patchout (PaSST-U):** randomly drops individual patches across the spectrogram.

- **Structured Patchout (PaSST-S):** removes complete time steps or frequency bands, akin to SpecAugment-style masking.

Structured Patchout aligns well with the 2D structure of spectrograms and generally yields the best trade-off between speed and predictive accuracy.

Unlike AST, which uses 2D grid positional encodings derived from image transformers, PaSST adopts *disentangled* positional encodings: separate embeddings for the time axis and the frequency axis. This design offers several advantages:

- it reflects the distinct semantic roles of time and frequency in audio data;

- it simplifies fine-tuning on shorter audio clips, as only the time-axis encoding must be cropped;

- it avoids the need for interpolation of positional embeddings when adapting to new input lengths.

This positional encoding refinement contributes to improved performance even in the absence of Patchout.

Evaluations are based on the Audioset dataset, containing approximately 2 million ten-second audio clips annotated with 527 sound event classes. The authors compute 128-band Mel spectrograms using 25 ms windows and 10 ms hops. To address significant class imbalance, importance sampling is applied, assigning higher sampling probability to clips with rarer labels.

All models are initialized from ImageNet-pretrained DeiT-B models, following prior observations that visual pretraining greatly improves audio transformer performance on Audioset.

Given the propensity of transformers to overfit, the authors employ extensive data augmentation: Mixup on both raw waveforms and spectrograms, SpecAugment masking, random time-shifts (rolling), and random gain perturbations. These augmentations play a crucial role in stabilizing training and improving generalization.

PaSST variants consistently outperform AST in mean average precision (mAP). Notably:

- **PaSST-S** achieves the highest single-model performance.

- Training with Patchout is dramatically faster: up to **4× faster** with **75% less** GPU memory than AST.

- **PaSST-S-N** (structured Patchout + non-overlapping patches) maintains strong accuracy while achieving an **8× speedup** and requiring **less than 10%** of the GPU memory used by AST.

Even the baseline PaSST (without Patchout) outperforms AST due to the improved positional encoding scheme.

Ensembling PaSST models trained with varying patch strides (i.e., overlap amounts) yields additional gains. The best ensemble achieves **mAP = 0.496**, surpassing ensembles of AST models and establishing a new state of the art on Audioset.

PaSST demonstrates strong transfer learning across diverse tasks with minimal fine-tuning time. On OpenMIC polyphonic instrument recognition, PaSST-S achieves 0.843 mAP (previous SOTA: 0.831) in under 30 minutes. On ESC-50 environmental sound classification, PaSST-S reaches 96.8%

accuracy in under 5 minutes of fine-tuning. On FSD50K sound event tagging, PaSST-S achieves 0.653 mAP, significantly outperforming the previous state-of-the-art of 0.558 mAP, with fine-tuning completed in under 2 hours. These results demonstrate that Patchout not only accelerates training but also improves generalization through its regularization effect.

Structured Patchout generally yields the best results, although non-overlapping variants offer even greater speed with competitive performance.

Patchout serves both as a computational optimization and an effective regularizer. Given that audio events tend to span contiguous time–frequency regions, removing subsets of patches has limited impact on semantic content and encourages the model to learn more robust representations. Structured Patchout, in particular, aligns naturally with spectrogram structure and SpecAugment-style masking.

Additionally, disentangled positional encodings improve adaptability to variable-length audio inputs and eliminate the need for positional interpolation. This design choice, combined with visual pretraining and aggressive data augmentation, enables PaSST models to be trained on consumer GPUs while achieving state-of-the-art accuracy.

The paper introduces Patchout, a simple yet powerful technique for improving the efficiency and generalization of audio transformers. Integrated into the PaSST architecture, Patchout enables:

- state-of-the-art performance on Audioset,
- massive reductions in training time and memory (up to 8× speedups),
- strong and efficient transfer to multiple downstream audio tasks,
- practical training of transformer models on consumer hardware.

PaSST establishes that transformer efficiency and performance are not mutually exclusive for audio tasks. Through Patchout's structured dropout and disentangled positional encodings, the work makes transformers practical for researchers with limited computational resources while achieving state-of-the-art results. However, PaSST still relies on ImageNet pretraining from vision models and extensive data augmentation (mixup, SpecAugment, rolling, random gain) to prevent overfitting. This dependency on cross-modal transfer and handcrafted augmentation motivated subsequent research into more data-efficient pretraining strategies, such as contrastive language-audio pretraining (CLAP) that learns from paired audio-text data, and self-supervised methods like BEATS that learn acoustic representations from unlabeled audio without requiring vision model initialization.

3.4 Efficient Large-Scale Audio Tagging via Transformer-to-CNN Knowledge Distillation

While PaSST [28] demonstrated that transformers could be trained efficiently through Patchout and achieve state-of-the-art performance (0.471 mAP for PaSST-S), the computational demands at inference remained substantial. Even the most efficient PaSST variant (PaSST-S-N) requires hundreds of millions of multiply-accumulate operations per audio clip, limiting deployment on edge devices and real-time applications. This paper [29] addresses the inference efficiency challenge through a complementary approach: using knowledge distillation to transfer the performance advantages of transformer ensembles to lightweight convolutional neural networks.

This paper [29] addresses the inference efficiency gap left by transformer-based audio taggers. While PaSST-S achieves 0.471 mAP on AudioSet with 87 million parameters, and AST reaches 0.459 mAP with similar complexity, their quadratic attention mechanisms make deployment on resource-constrained devices impractical. The authors propose transferring transformer ensemble performance to efficient CNNs through offline knowledge distillation, achieving comparable accuracy with 10× fewer parameters and 100× fewer multiply-accumulate operations.

The work focuses on the task of multi-label audio tagging, where one or more semantic sound event labels are assigned to a ten-second audio clip. Experiments are conducted on AudioSet, a large-scale, weakly labeled dataset comprising over two million audio clips spanning 527 sound classes. AudioSet poses significant challenges due to its long-tailed class distribution, label noise, and reliance on weak supervision. While transformer-based models such as AST, PaSST, and HTS-AT currently dominate the AudioSet leaderboard, their quadratic attention complexity and large model sizes motivate the authors' exploration of more efficient CNN-based solutions.

As a student architecture, the paper adopts MobileNetV3, an efficient CNN design originally developed for mobile vision applications. The authors systematically enhance this baseline using offline knowledge distillation from a high-performing teacher ensemble composed of multiple PaSST audio transformers with varying patch sizes and strides. Teacher predictions are precomputed on AudioSet and used as soft targets during student training. The distillation objective combines a standard binary cross-entropy loss with a distillation loss computed from the teacher's sigmoid-activated logits, enabling the student model to exploit inter-class similarity information captured by the transformer ensemble.

A comprehensive experimental study demonstrates that knowledge distillation dramatically improves the performance of MobileNetV3-based models. Without distillation, the CNN baseline lags behind existing CNN and transformer approaches. With KD, however, the distilled models match or surpass the performance of single transformer models while requiring an order of magnitude fewer parameters and up to two orders of magnitude fewer multiply-accumulate operations. By scaling network width (multiplying channel counts by factor α) and adjusting spectrogram resolution, the authors create a model family spanning from sub-million parameter edge models to a 68-million parameter variant achieving 0.483 mAP (illustrated in Fig. 5) a new single-model state-of-the-art on AudioSet.

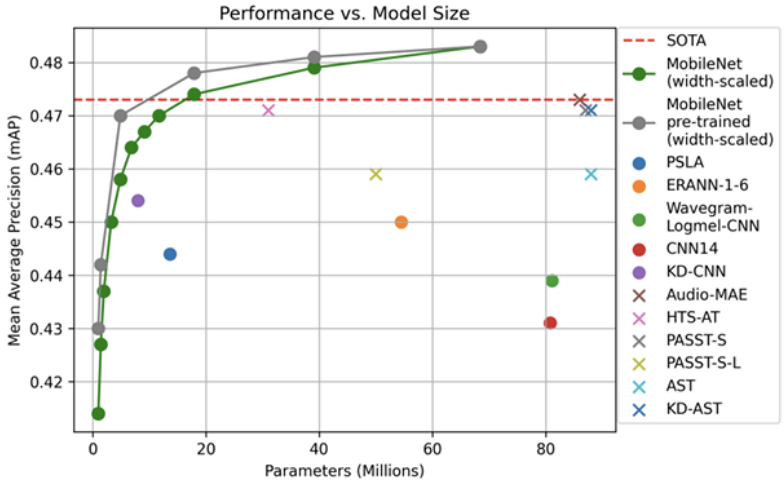


Figure 5: Performance versus parameter count for audio tagging models on AudioSet. Crosses denote transformer architectures while circles denote CNNs. Gray and green curves show width-scaled MobileNetV3 models without and with ImageNet pretraining, respectively. The proposed distilled models achieve transformer-level performance with an order of magnitude fewer parameters.

The paper further includes detailed ablation studies analyzing the impact of key design choices. These cover distillation hyperparameters, data augmentation strategies, squeeze-and-excitation mechanisms, and different classification heads. Results show that high-weight distillation losses with low temperature settings are most effective, that mixup-based augmentation provides modest gains, and that channel-wise squeeze-and-excitation layers significantly improve performance despite their parameter cost. Simpler MLP-based classification heads offer the best performance–complexity trade-off compared to attention-based or fully convolutional alternatives.

Computational efficiency gains are substantial: the default MobileNetV3 model (α=1.0, achieving 0.470 mAP) requires only 0.6 billion MACs compared to PaSST-S's 60+ billion MACs—a 100× reduction. GPU benchmarking on an Nvidia A100 confirms practical speedups: the distilled MobileNet achieves 4,767 clips/second throughput versus PaSST's 78 clips/second, representing a 61× real-world acceleration while maintaining comparable accuracy.

The work demonstrates that knowledge distillation from transformer ensembles enables efficient CNNs to approach transformer-level performance while maintaining the computational advantages of convolutional architectures. This establishes transformer-to-CNN distillation as a practical deployment strategy when inference efficiency is critical. However, the approach still relies on expensive transformer teacher training and ImageNet cross-modal pretraining. These dependencies motivated parallel research directions exploring alternative pretraining paradigms: contrastive language-audio pretraining (CLAP) [30] learns audio-text alignments for zero-shot capabilities

without ImageNet initialization, while self-supervised methods like BEATS [31] learn acoustic representations directly from unlabeled audio through discrete tokenization, eliminating cross-domain transfer altogether.

3.5 Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation

This paper [30] introduces CLAP (Contrastive Language-Audio Pretraining), adapting CLIP's [32] image-text contrastive learning framework to the audio domain. Rather than training on curated AudioSet labels, CLAP learns joint embeddings between audio and natural language descriptions, enabling zero-shot audio classification and text-to-audio retrieval. The work addresses audio's historical dependence on expensive labeled datasets by leveraging large-scale weakly supervised audio-text pairs and contrastive objectives.

A central objective is to overcome the historical dependence on curated labels for audio tasks, which severely limits scalability. By leveraging large-scale, weakly supervised data and contrastive objectives, the authors aim to develop audio representations that generalize well across a variety of downstream tasks. The contributions include the release of the LAION-Audio-630K dataset, methodology for converting audio labels to natural language captions, a feature fusion mechanism for handling variable-length audio, and a contrastively trained model that achieves state-of-the-art performance across retrieval and classification benchmarks.

The work introduces **LAION-Audio-630K**, containing 633,526 audio-caption pairs totaling 4,325 hours—an order of magnitude larger than previous datasets (AudioCaps: 53K pairs/145 hours, Clotho: 6K pairs/37 hours, SoundDescs: 33K pairs/1,060 hours). The dataset aggregates audio from a wide range of public sources, including sound effects libraries, user-generated recordings, natural ambient sound collections, and creative commons repositories. The captions include human-written descriptions and text derived from filenames. Compared to AudioCaps, Clotho, or SoundDescs, LAION-Audio-630K is orders of magnitude larger, enabling effective large-scale contrastive pretraining.

To further increase the amount of paired audio–text data, the authors apply **keyword-to-caption augmentation**. Many audio datasets, such as AudioSet, provide labels or short keyword descriptions but lack full-sentence captions. Using a pretrained T5 language model, the authors automatically generate natural language captions from label sets. They also apply post-processing steps to reduce demographic and gender bias by replacing gendered terms with neutral descriptors. Through this augmentation, the effective training corpus grows to approximately 2.5 million captioned audio samples.

Audio is standardized to 48 kHz mono FLAC format, while text is tokenized with a maximum length of 77 tokens. For datasets lacking captions, either templated text or keyword-to-caption augmented text is used. Strict filtering and overlap removal ensure that training data does not contaminate evaluation sets used in downstream experiments.

The proposed model closely follows the CLIP architecture, utilizing separate encoders for audio and text. Each encoder output passes through a projection multilayer perceptron that maps it into a shared latent space (illustrated in Fig. 6). Training uses a contrastive objective that maximizes similarity between paired audio–text samples while minimizing similarity to mismatched pairs. A learnable temperature parameter controls the sharpness of similarity distributions.

The study explores combinations of two audio encoders and three text encoders:

- **Audio encoders:** PANN (CNN-based) and HTSAT (transformer-based).
- **Text encoders:** CLIP text transformer, BERT, and RoBERTa.

Extensive comparison reveals that the combination of **HTSAT + RoBERTa** consistently yields the best performance for retrieval tasks. The CLIP text transformer suffers from overfitting and poor generalization in the audio domain.

A significant architectural innovation is the **feature fusion mechanism**. Natural audio varies widely in length, creating difficulties for transformer-based audio encoders. The authors introduce a dual-path method combining:

1. A **global representation** obtained from a downsampled audio signal.
2. Multiple **local representations** sampled from distinct temporal regions.

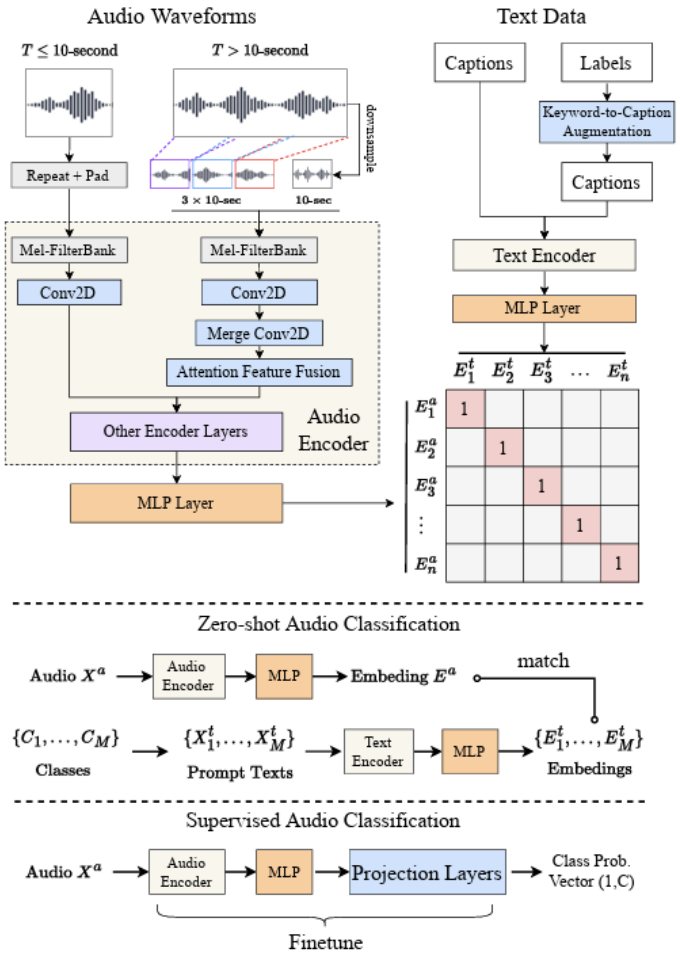


Figure 6: The CLAP architecture showing audio and text encoders with projection layers into shared embedding space. The system incorporates attentional feature fusion for variable-length audio (combining downsampled global and sampled local representations) and keyword-to-caption augmentation to expand training data from labels to natural language descriptions. The learned embeddings enable both zero-shot classification (via text-audio matching) and supervised fine-tuning.

An attentional feature fusion (AFF) module learns to combine these representations into a single embedding. This method outperforms traditional slice-and-average strategies, particularly for long audio recordings, and reduces computational inefficiency.

The authors evaluate performance on the AudioCaps and Clotho datasets using recall metrics and mean average precision. Results show that:

- HTSAT–RoBERTa significantly outperforms all other encoder combinations.
- Scaling training data increases generalization but may reduce performance on datasets whose distribution is close to pretrained audio encoders.
- Feature fusion improves retrieval performance, especially for datasets with longer audio.
- Keyword-to-caption augmentation improves almost all evaluation metrics.

The best configuration surpasses previous state-of-the-art methods such as CLAP-HTSAT and MMT.

Zero-shot classification is evaluated on ESC-50, UrbanSound8K, and VGGSound. Using prompt-based retrieval ("the sound of <class>"), the model achieves new state-of-the-art results across all three datasets. The keyword-to-caption augmentation improves performance particularly on VGGSound and UrbanSound8K by providing richer text supervision.

By fine-tuning the audio encoder, the model achieves:

- State-of-the-art classification accuracy on VGGSound.
- Near state-of-the-art performance on FSD50K.

These results demonstrate the strong generalizability of the pretrained audio representation.

CLAP demonstrates that large-scale contrastive language-audio pretraining produces robust, generalizable audio representations competitive with supervised methods while enabling zero-shot capabilities through natural language. The LAION-Audio-630K dataset, keyword-to-caption augmentation for scaling to 2.5 million pairs, and attentional feature fusion for variable-length audio collectively advance multimodal audio understanding. However, CLAP's reliance on paired audio-text data and contrastive learning from existing captions limits its ability to discover acoustic patterns beyond what natural language describes. This motivated self-supervised approaches like BEATS [31], which learn acoustic representations directly from raw audio through discrete tokenization, discovering patterns that may not have corresponding text descriptions and eliminating the dependency on paired multimodal data.

3.6 Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers

The paper [33] presents Whisper-AT, a unified model that performs automatic speech recognition (ASR) and audio event tagging simultaneously. The key finding is that while Whisper [32] is highly robust against background noise, its audio representations are actually noise-variant rather than noise-invariant, encoding rich information about non-speech background sounds. This counter-intuitive discovery challenges the conventional assumption that robust ASR models should learn to ignore background noise. Instead, Whisper encodes the type of background sound present and performs speech recognition conditioned on the noise type—recognizing not just what is said, but also the acoustic context in which it is said.

The authors demonstrate that Whisper's noise robustness correlates positively with the amount of general audio event information encoded in its intermediate representations. Through linear probing experiments (testing frozen Whisper representations with a simple linear classifier) on ESC-50 environmental sounds, they show that Whisper achieves the best sound classification accuracy among ASR models, with representations from deeper layers maintaining strong audio event

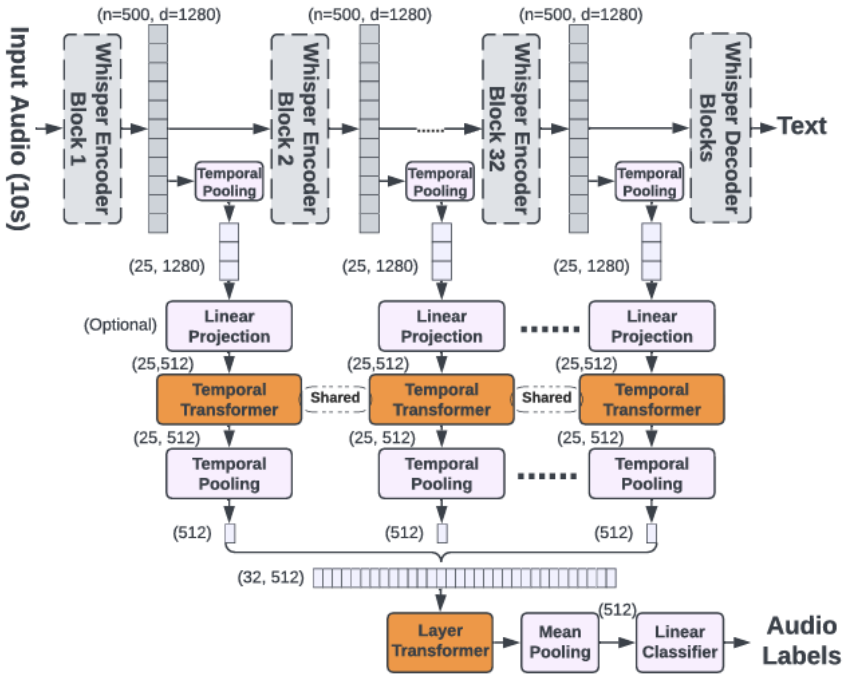


Figure 7: The Time and Layer-wise Transformer (TL-Tr) architecture for Whisper-AT. The model adds lightweight audio tagging layers on top of frozen Whisper encoder representations.

encoding capabilities. This is in stark contrast to other ASR models like wav2vec2[34] and HuBERT[35], where deeper layers progressively ignore background sound information. The class-wise analysis reveals that Whisper's robustness against a specific background sound type (e.g., music, dog barking, car horn) positively correlates with its ability to recognize that sound type, confirming that noise-awareness rather than noise-invariance is the mechanism behind Whisper's robustness.

Building on this finding, Whisper-AT freezes the original Whisper backbone and trains a lightweight Time and Layer-wise Transformer (TL-Tr) module on top. The TL-Tr architecture (illustrated in Fig. 7) applies temporal pooling to reduce sequence length, optional linear projection to reduce dimensionality (from 1280 to 512), and separate temporal and layer-wise Transformers to aggregate information across both dimensions. This design enables different sound classes to leverage representations from different Whisper layers, as the authors found that optimal layer representations vary by sound class.

The model achieves competitive audio tagging performance on AudioSet (41.5 mAP on AS-2M, 32.8 mAP on AS-20K) and ESC-50 (91.7% accuracy) while being over 40× faster than standalone audio tagging models like AST [26]. With only 7 million parameters for the audio tagging component (TL-Tr512) and less than 1% additional computational cost compared to running Whisper alone, Whisper-AT provides a highly efficient solution for applications requiring both speech transcription and acoustic scene analysis, such as video transcribing, voice assistants, and hearing aid systems. The model outputs audio event labels from the 527-class AudioSet ontology alongside speech transcripts in a single forward pass.

3.7 BEATS: Audio Pre-Training with Acoustic Tokenizers

While CLAP [30] demonstrated that contrastive language-audio pretraining enables zero-shot capabilities and strong generalization, its reliance on paired audio-text data limits discovery of acoustic patterns beyond textual descriptions. This paper [31] introduces BEATS (Bidirectional Encoder representation from Audio Transformers), which takes a fundamentally different approach: self-supervised learning through discrete label prediction rather than either reconstruction (Audio-MAE) or language supervision (CLAP). BEATS addresses a core challenge—obtaining semantic-rich discrete tokens for continuous audio signals without relying on phoneme sequences (as in speech) or text descriptions (as in CLAP).

This paper [31] introduces BEATS, a self-supervised audio pretraining framework that replaces reconstruction-based objectives with discrete label prediction. While previous SOTA audio SSL models (Audio-MAE: 47.3 mAP, MaskSpec: 47.1 mAP) employ spectrogram reconstruction, BEATS argues that discrete, semantic-rich targets better encourage abstraction and discard redundant details, mimicking human perception. The central challenge is obtaining semantically meaningful discrete tokens for general audio, which lacks the natural tokenization of language (words) or speech (phonemes).

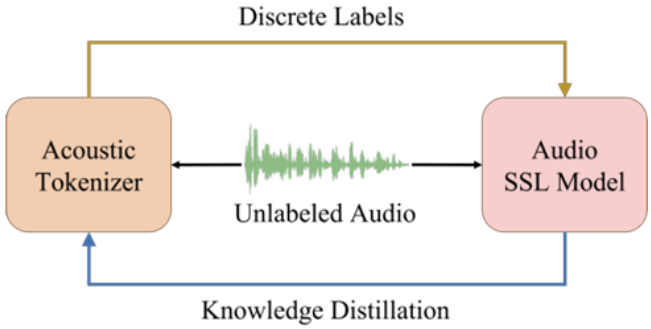


Figure 8: BEATS iterative pretraining framework. The acoustic tokenizer generates discrete labels for SSL model training via masked prediction. After convergence, the SSL model serves as teacher for tokenizer refinement through knowledge distillation. This alternating optimization repeats across iterations, with optional fine-tuning to incorporate supervised knowledge.

BEATS proposes an **iterative pretraining framework** where an acoustic tokenizer and an audio SSL model are jointly optimized through alternating updates (illustrated in Fig. 8). In each

iteration: (1) the acoustic tokenizer converts continuous audio into patch-level discrete labels, (2) these labels serve as targets for masked prediction training, (3) the trained SSL model then acts as teacher to refine the tokenizer via knowledge distillation. This mutual improvement process repeats until convergence, typically stabilizing after 2-3 iterations.

The audio SSL model in BEATS adopts a Vision Transformer (ViT) architecture adapted for audio. Input waveforms are converted to log-Mel filter bank features, split into fixed-size time–frequency patches, and embedded via a linear projection before being processed by a stack of Transformer encoder layers. During pre-training, BEATS employs a masked audio modeling objective in which 75% of the input patches are masked, and the model is trained to predict the corresponding discrete acoustic labels using a cross-entropy loss. Unlike reconstruction-based approaches, the model predicts token identities rather than continuous spectrogram values, aligning the training objective with semantic abstraction.

The acoustic tokenizer plays a crucial role in enabling discrete label prediction. In the first iteration, BEATS uses a random-projection tokenizer that assigns discrete labels by nearest-neighbor lookup in a randomly initialized codebook, providing a cold start without supervision. In subsequent iterations, a self-distilled tokenizer is trained using a Transformer-based encoder and a learnable codebook. This tokenizer is supervised by the representations produced by the previous iteration's audio SSL model through a combination of cosine similarity and vector-quantization losses. The resulting discrete labels are shown to be more robust to noise and better aligned with semantic content than raw acoustic features.

BEATS is evaluated on a broad set of downstream audio and speech classification tasks. Pretraining is performed on the AudioSet dataset, which contains over two million 10-second audio clips spanning 527 sound event classes. The pretrained models are evaluated on AudioSet-2M and AudioSet-20K for audio tagging, ESC-50 for environmental sound classification, Speech Commands V1 and V2 for keyword spotting, and IEMOCAP for speech emotion recognition. Experimental results demonstrate that BEATS achieves state-of-the-art or competitive performance across all tasks, including a new state-of-the-art mean average precision on AudioSet-2M for audio-only models and high accuracy on ESC-50, despite using fewer parameters and no external supervised data.

The authors further analyze the behavior of BEATS through ablation studies and visualizations. Comparisons with reconstruction-based SSL models show that the discrete targets learned by BEATS are more invariant to random disturbances and better clustered according to semantic categories. The iterative refinement of the tokenizer is shown to yield consistent improvements over successive iterations, while convergence is typically achieved after only a few rounds. Ensemble experiments further demonstrate that BEATS scales effectively, achieving additional gains when multiple pretrained models are combined.

Several limitations are discussed. Although BEATS improves semantic abstraction, the framework still relies on large-scale pre-training data and significant computational resources. The tokenizer training procedure introduces additional complexity compared to standard SSL pipelines, and performance remains sensitive to the quality of the teacher model used for distillation. Moreover, while BEATS unifies audio and speech pre-training objectives, it is primarily evaluated on classification tasks and does not directly address generative audio modeling. Despite these limitations, the paper establishes BEATS as a strong alternative to reconstruction-based audio SSL, demonstrating that iterative discrete label prediction is a powerful paradigm for learning transferable audio representations.

BEATS establishes discrete label prediction as a powerful alternative to reconstruction-based audio SSL, unifying pretraining objectives across language, vision, speech, and audio modalities. The iterative tokenizer-model co-training successfully learns semantic-rich discrete representations from continuous audio without external supervision. However, BEATS still requires large-scale pretraining data (AudioSet's 2M clips) and substantial compute, and the iterative framework adds complexity compared to single-stage SSL. While BEATS excels at classification, it focuses on discriminative tasks and does not address generative audio modeling. These results nonetheless demonstrate that semantic abstraction through discrete tokenization—whether from text (CLAP) or self-distilled labels (BEATS)—consistently outperforms low-level reconstruction, suggesting that future audio foundation models should prioritize high-level semantic learning over spectral reconstruction.

3.8 Streaming Audio Transformers for Online Audio Tagging

While BEATS [31] demonstrated that discrete label prediction achieves superior performance for audio tagging (48.6% mAP on AudioSet), existing transformer-based methods including BEATS share a critical limitation: they are optimized for offline inference with full 10-second context, resulting in high memory consumption, quadratic attention complexity, and minimum 10-second delays. This paper [36] addresses the deployment gap by introducing Streaming Audio Transformers (SAT), which combine ViT architectures with chunk-wise recurrence mechanisms to enable real-time audio tagging with 1-2 second delays while maintaining competitive performance.

This paper [36] addresses transformer-based audio tagging models' incompatibility with low-latency deployment. While ViT-based models (AST: 45.9% mAP, BEATS: 48.6% mAP, HTS-AT: 47.1% mAP) achieve state-of-the-art performance on AudioSet, they require full 10-second context with quadratic attention complexity ($O(n^2)$), consuming 2.2 GB memory (AST) and incurring minimum 10-second delays. The authors target streaming inference with strict constraints: 1-2 second latency, low memory footprint, and robust performance—requirements unmet by existing methods.

The work considers the task of multi-label audio tagging, where an input audio signal is mapped to a fixed set of semantic sound event classes. Experiments are conducted primarily on AudioSet, a large-scale weakly labeled dataset containing over two million ten-second audio clips spanning hundreds of sound categories. AudioSet presents several challenges, including noisy labels, long-tailed class distributions, and the need for robust aggregation of predictions over time. In contrast to prior work that assumes access to full-context inputs, the authors explicitly target streaming inference with short delays of one or two seconds.

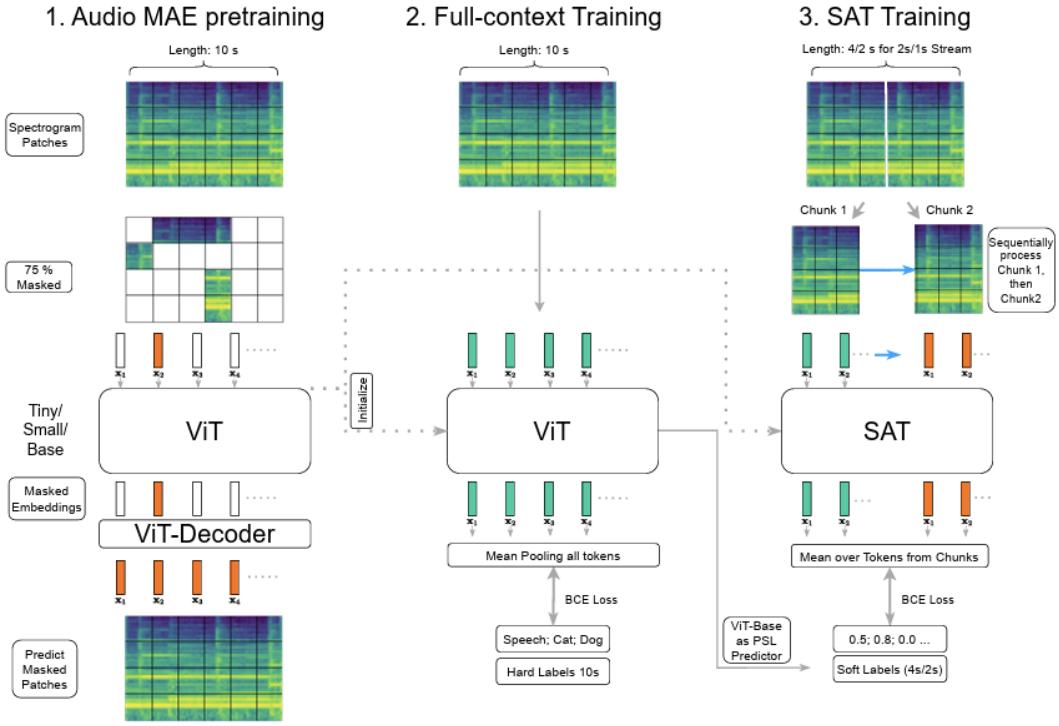


Figure 9: SAT training pipeline: (1) MAE pretraining with 75% patch masking, (2) full-context fine-tuning on 10-second AudioSet clips, (3) streaming training using pseudo strong labels (PSL) generated by ViT-B at 1-2 second resolution. The chunk-wise processing with cached key-value pairs enables linear memory scaling for streaming inference.

To this end, the paper introduces *Streaming Audio Transformers* (SAT), a family of ViT-based models adapted for online audio tagging. The proposed architecture combines the patch-based spectrogram processing of Vision Transformers with a Transformer-XL–like chunk-wise recurrence mechanism. Incoming audio is processed in fixed-length temporal chunks, and self-attention within each chunk is augmented with cached key–value representations from previous chunks. This de-

sign enables linear memory growth with respect to time while preserving access to past context, thereby avoiding expensive recomputation and reducing inference latency. Three model variants are studied—SAT-Tiny (SAT-T), SAT-Small (SAT-S), and SAT-Base (SAT-B)—corresponding to different embedding dimensions and computational budgets.

The training pipeline consists of three stages (illustrated in Fig. 9). First, the authors pretrain ViT encoders using a masked autoencoder (MAE) objective on log-mel spectrograms, following recent advances in self-supervised audio representation learning. Second, the pretrained models are fine-tuned on AudioSet using full-context (ten-second) clips with a binary cross-entropy objective. Third, streaming variants are trained using pseudo strong labels (PSL), where a high-performing offline model generates fine-grained soft labels at one- or two-second resolution. These soft targets allow SAT models to learn chunk-level predictions while maintaining consistency with full-context supervision.

Extensive experiments demonstrate that standard transformer models suffer substantial performance degradation when naively evaluated under short-delay conditions. In contrast, the proposed SAT models consistently outperform both their non-streaming counterparts and existing transformer baselines when evaluated with one- or two-second delays. The best-performing model, SAT-B, achieves an mAP of 45.1 on AudioSet with a two-second delay, approaching full-context transformer performance while using a fraction of the memory and computational resources. Additional evaluations on the strongly labeled AudioSet subset show that SAT models also achieve superior segment-level and onset-level F1 scores, indicating improved temporal localization of sound events.

SAT establishes streaming transformers as a practical solution for real-time audio tagging, demonstrating that transformer architectures can be adapted to meet strict latency and memory constraints without sacrificing competitive performance. The chunk-wise recurrence mechanism with cached attention enables linear memory scaling while preserving temporal context—a critical advancement for deployment on resource-constrained devices. However, SAT still requires three-stage training (MAE pretraining, full-context fine-tuning, PSL-based streaming training) and relies on a strong offline teacher for optimal performance, adding complexity compared to end-to-end approaches. The work demonstrates a fundamental trade-off between context and latency: while streaming enables real-world deployment, performance inherently lags behind full-context models that access complete 10-second clips. This motivates exploring efficient audio tagging (EAT) approaches that further optimize this performance-efficiency trade-off through architectural innovations and training strategies specifically designed for the audio tagging task.

3.9 EAT: Self-Supervised Pre-Training with Efficient Audio Transformer

While Streaming Audio Transformers [36] addressed real-time deployment constraints, existing audio SSL models still face a critical limitation: expensive computational costs during pretraining. BEATS requires 342 epochs and Audio-MAE requires 32 epochs to achieve strong performance, making experimentation and model development prohibitively slow. This paper [37] introduces EAT (Efficient Audio Transformer), achieving state-of-the-art performance with only 10 epochs of pretraining—a 15× speedup over BEATS and 10× over Audio-MAE—through a novel combination of utterance-level and frame-level learning objectives.

EAT adopts a bootstrap self-supervised framework combining masked modeling with teacher-student learning (illustrated in Fig. 10). Input log-mel spectrograms are downsampled into non-overlapping patches via a CNN encoder with stride 16. A student Transformer processes heavily masked inputs (80% masking ratio), while a teacher Transformer with identical architecture receives the full unmasked spectrogram and is updated via exponential moving average of student parameters. Unlike reconstruction-based objectives that recover raw patches, EAT predicts latent representations generated by the teacher network—specifically, the average across all transformer layers rather than just the final layer, capturing both shallow acoustic features and deep semantic representations.

A central contribution of the paper is the proposed *Utterance–Frame Objective* (UFO), which combines global and local learning signals during pre-training. At the frame level, the student model predicts teacher latent representations at masked patch positions using a lightweight CNN decoder, optimized with a mean squared error loss. At the utterance level, a learnable CLS token in the student encoder aggregates global information across visible patches and is trained to regress

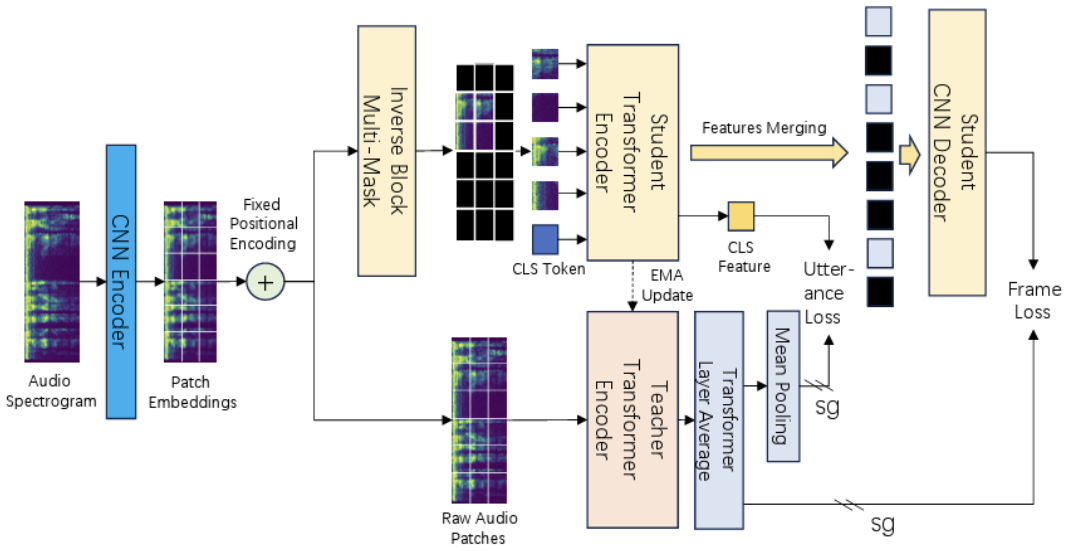


Figure 10: EAT architecture showing bootstrap framework with teacher-student learning. The student processes inverse block-masked patches (80% masking) while the teacher receives complete input. Utterance-Frame Objective combines CLS token-based global prediction with CNN decoder frame-level reconstruction. Multi-mask strategy creates 16 variants per spectrogram for parallel student processing, dramatically improving data utilization.

the mean-pooled teacher representations. By jointly optimizing frame-level and utterance-level losses, EAT explicitly encourages the learning of both local acoustic patterns and holistic clip-level semantics without introducing additional projection heads or complex decoders.

Pre-training efficiency is further enhanced through an inverse block masking strategy with a high masking ratio of up to 80%. Instead of random patch masking, EAT preserves contiguous blocks in the time–frequency plane, increasing task difficulty while maintaining meaningful local structure. A multi-mask strategy is additionally employed, whereby multiple masked variants of the same spectrogram are processed in parallel by the student model, improving data utilization and accelerating convergence. The asymmetric architecture—a Transformer encoder paired with a lightweight CNN decoder—significantly reduces computational overhead compared to Transformer-based decoders used in prior work.

The model is pre-trained on the unbalanced AudioSet-2M dataset and evaluated via fine-tuning on several downstream tasks. EAT achieves state-of-the-art performance among base-size audio SSL models on AudioSet-2M and AudioSet-20K, reaching an mAP of 48.6 and 40.2, respectively. It also performs strongly on environmental sound classification (ESC-50), achieving 95.9% accuracy, and on speech command recognition (Speech Commands V2), with an accuracy of 98.3%. Notably, these results are obtained with only 10 epochs of pre-training, yielding up to a 15× speedup compared to BEATs and Audio-MAE while matching or surpassing their performance.

The authors conduct extensive ablation studies demonstrating the importance of utterance-level learning, the balance between frame and utterance losses, the use of a CLS token for downstream prediction, and the choice of inverse block masking size. These analyses show that incorporating global supervision during pre-training and using structured masking are key to both performance and efficiency gains.

Despite its strong results, EAT is evaluated primarily on classification-oriented downstream tasks, leaving open questions regarding its effectiveness for generation or low-level audio synthesis tasks. Furthermore, like other AudioSet-based models, EAT inherits biases from weakly labeled, web-sourced audio data. Nevertheless, the work establishes EAT as an efficient and effective SSL framework, demonstrating that carefully designed objectives and masking strategies can significantly reduce pre-training cost without sacrificing representation quality.

3.10 Conclusion

The evolution of large-scale audio tagging has transitioned from deep convolutional architectures to purely attention-based models and efficient self-supervised frameworks. Early work with **PANNs** established that deep convolutional networks like CNN14 and Wavegram-Logmel-CNN could effectively learn from the 527 sound classes of AudioSet, outperforming previous baselines by leveraging hierarchical features and waveform-based representations [22]. The **Audio Spectrogram Transformer (AST)** represented a paradigm shift as the first convolution-free model, demonstrating that global self-attention and cross-modality transfer learning from ImageNet could achieve superior performance over CNN-attention hybrids [26].

To address the high computational costs and memory bottlenecks of standard transformers, **PaSST** introduced the "Patchout" technique, which randomly drops spectrogram patches during training to reduce quadratic complexity, enabling state-of-the-art results on consumer-grade hardware [28]. Conversely, researchers explored the efficiency of convolutional inductive biases through **Knowledge Distillation (KD)**, successfully transferring the performance of transformer ensembles into lightweight **MobileNetV3** models, which achieved comparable accuracy with 10× fewer parameters and significantly higher inference throughput [29].

Further advancements focused on multimodal pretraining and perception-based learning. **CLAP** introduced multimodal contrastive learning, utilizing the massive LAION-Audio-630K dataset to align audio with natural language, thereby enabling zero-shot classification and text-to-audio retrieval [30]. **BEATS** improved self-supervised learning by employing an iterative framework with acoustic tokenizers to predict discrete semantic labels, mimicking human auditory perception more closely than traditional reconstruction-based methods [31]. For online applications, **SAT** adapted transformer architectures for streaming inference with low latency (1–2 seconds) using chunk-wise recurrence [36]. Meanwhile, **Whisper-AT** demonstrated that robust ASR models like Whisper could serve as strong general-purpose taggers with minimal extra computational cost [33]. Most recently, **EAT** optimized pretraining efficiency through the Utterance-Frame Objective (UFO), reaching state-of-the-art results in only 10 epochs [37].

The performance metrics for these representative models on the AudioSet-2M benchmark are summarized in Table 1.

Table 1: Audio tagging models ranked by AudioSet-2M performance (mAP). Only models with AS-2M results shown.

Rank	Model	AS-2M mAP	AS-20K mAP	ESC-50 Acc	Params (M)
1	PANN (CNN14)	43.1	27.8	83.3	81
2	MBT	44.3	31.3	–	86
3	PSLA	44.4	31.9	–	14
4	ERANN	45.0	–	89.2	55
5	AST	45.9	34.7	88.7	86
6	PaSST	47.1	–	96.8	86
6	HTS-AT	47.1	–	97.0	31
6	MaskSpec	47.1	32.3	89.6	86
9	Audio-MAE	47.3	37.1	94.1	86
10	Audio-MAE Large	47.4	37.6	–	304
11	BEATS_{iter1}	47.9	36.0	94.0	90
12	BEATS_{iter2}	48.1	38.3	95.1	90
13	BEATS_{iter3}	48.0	38.3	95.6	90
14	T2CNN-KD	48.3	–	–	4.88
1	**BEATS_{iter3+}**	**48.6**	38.9	**98.1**	90
1	**EAT**	**48.6**	**40.2**	95.9	88

Although the models discussed have significantly advanced the detection of general sound events, the specific nuances of musical content require even more specialized approaches. Understanding complex polyphony, genre characteristics, and instrument-specific features often requires moving beyond simple tagging toward more granular retrieval systems. The following section explores **Music Retrieval Information**, detailing how these large-scale representations are adapted to handle the unique temporal and harmonic complexities of music data.

4 Music Information Retrieval (MIR)

4.1 End-to-End Musical Key Estimation Using a Convolutional Neural Network

Korzeniowski and Widmer [38] propose an end-to-end convolutional neural network for global musical key estimation. Traditional approaches rely on multi-stage pipelines with hand-crafted chroma features and template matching, requiring expert knowledge and genre-specific tuning. This work replaces the fragmented pipeline with a single trainable model optimized directly from data.

The primary objective is to eliminate manual feature engineering while achieving competitive or superior performance on key estimation. By learning harmonic representations directly from spectrograms, the model aims to generalize across musical genres without requiring domain-specific pre-processing steps such as tuning correction or chroma extraction.

The proposed system operates directly on audio-derived time-frequency representations. As input, the authors compute a logarithmically filtered log-magnitude spectrogram, obtained by applying a bank of logarithmically spaced triangular filters to the magnitude spectrogram and subsequently compressing the dynamic range via a logarithmic transform. This representation is similar in spirit to the constant-Q transform but is computationally cheaper, while still preserving pitch-related structure that is crucial for harmonic analysis. Importantly, no chroma features, tuning correction, or beat synchronization are applied, allowing the network to learn relevant harmonic representations autonomously.

Architecturally the model, presented in 11, consists of five convolutional layers with small two-dimensional kernels, followed by a frame-wise dense projection layer. The resulting frame-level representations are aggregated over time using global average pooling, producing a fixed-length embedding for an entire musical piece. A final softmax layer performs classification over 24 classes, corresponding to the 12 chromatic tonics combined with major and minor modes. All layers except the output use exponential linear unit (ELU) activations. The design intentionally avoids recurrent or sequential modeling; preliminary experiments indicated that more complex temporal architectures did not yield improved performance for the global key estimation task.

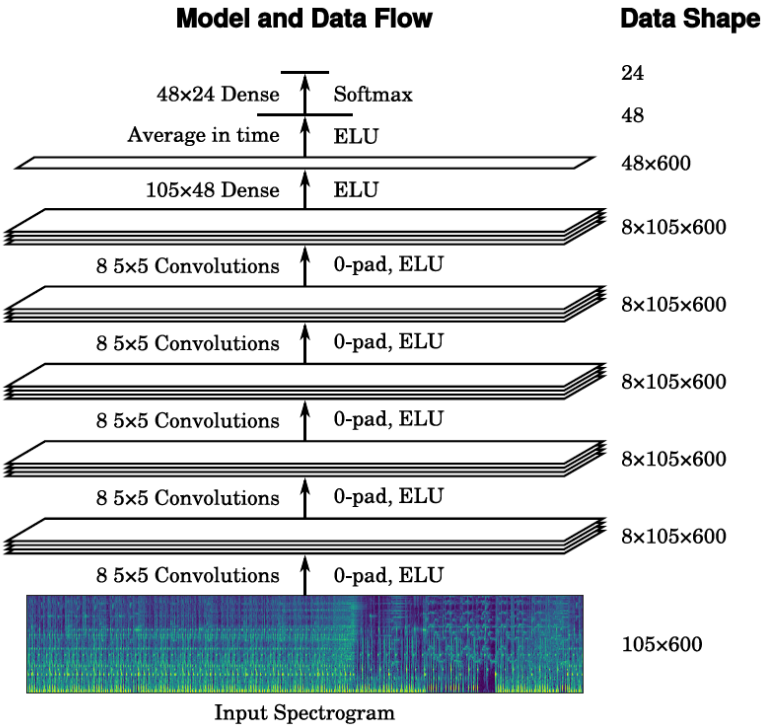


Figure 11: CNN architecture for end-to-end key estimation. Five convolutional layers extract frame-level features, which are temporally aggregated via global average pooling before classification into 24 key classes.

The network is trained using stochastic gradient descent with momentum and categorical cross-entropy loss. Given the limited size of available annotated datasets, the authors employ extensive data augmentation through pitch shifting. Each training example is transposed across a range of semitone shifts, with corresponding adjustments to the target key labels, increasing the effective training set size by a factor of twelve. This augmentation strategy leverages the transpositional invariance of musical key and is particularly well suited to harmony-related tasks.

Evaluation is conducted on three datasets: GiantSteps (GS, electronic music), GiantSteps-MTG (GSMTG, training), and McGill Billboard (pop/rock). Performance is assessed using the MIREX weighted score protocol, which distinguishes between error types (fifth relations, relative/parallel major/minor confusions) and assigns partial credit.

Table 2 presents the performance of different training configurations compared to template-based reference systems. The proposed CNN models are denoted as CK1 (trained on GSMTG), CK2 (trained on Billboard), and CK3 (trained on both datasets).

Table 2: Key estimation results on GiantSteps and Billboard test sets. Weighted scores follow MIREX protocol. Bold indicates best performance per dataset.

Method	**Training Data**	**GiantSteps (GS)**	**Billboard (BBTE)**
Proposed CNN Models			
CK1	GSMTG (EDM)	**74.3**	72.8
CK2	Billboard (Pop/Rock)	57.3	**83.9**
CK3	GSMTG + Billboard	69.2	79.7
Template-Based Reference Systems			
EDMA	Auto-derived templates	65.6	78.7
EDMM	Manual-tuned templates	70.1	28.9
EDMT	Classical templates	44.6	75.4
QM	Bach-based templates	50.4	60.9

When trained and evaluated within the same genre (CK1 on GS, CK2 on BBTE), the CNN achieves state-of-the-art performance, outperforming the best template-based systems (74.3% vs. 70.1% on GiantSteps; 83.9% vs. 78.7% on Billboard). Cross-genre scenarios reveal substantial performance degradation—CK2 trained on pop/rock achieves only 57.3% on electronic music. The multi-genre model (CK3) provides competitive compromise performance (69.2% on GS, 79.7% on BBTE) but does not match specialized models.

These results demonstrate that end-to-end learning can effectively replace hand-crafted feature extraction for key estimation, particularly within well-defined genres. The cross-genre performance degradation reveals an important limitation: while the model learns fundamental tonal relationships (evidenced by low severe error rates), it also captures genre-specific characteristics that hinder generalization. This suggests that deeper networks or more diverse training data may be necessary for truly genre-agnostic key estimation. The reliance on global average pooling, while computationally efficient, fundamentally limits the model to single-key predictions and discards potentially useful temporal dynamics.

The authors identify several limitations of their approach. Most notably, the model estimates only a single global key per piece and cannot capture key modulations that are common in certain musical styles, such as classical music. Temporal structure is simplified through global averaging, potentially discarding informative sequential patterns. Furthermore, despite augmentation, the available training data remain limited, which may constrain generalization. Nonetheless, the work demonstrates that end-to-end learning can effectively replace hand-crafted pipelines for harmonic analysis, establishing an important early example of fully data-driven key estimation in MIR.

Nonetheless, this work establishes an important precedent for data-driven harmonic analysis in MIR. While key estimation addresses tonal center identification at the piece level, a more detailed understanding of musical harmony requires frame-level chord recognition—a related but distinct chall enge that demands different architectural considerations, as explored in the following section.

4.2 ChordFormer: A Conformer-Based Architecture for Large-Vocabulary Audio Chord Recognition

The paper [39] introduces *ChordFormer*, a Conformer-based deep learning architecture designed for large-vocabulary automatic chord recognition (ACR) from audio. Motivated by the limitations of existing convolutional, recurrent, and transformer-based approaches in modeling long-term harmonic dependencies and handling severe class imbalance, the authors address the challenge of recognizing complex and rare chord types in real-world music. Large-vocabulary chord recognition is particularly demanding due to the long-tailed distribution of chord classes, overlapping harmonic structures among extended chords, and the need to jointly capture local spectral patterns and global temporal context.

The primary objective is to overcome limitations of existing CNN, RNN, and Transformer approaches in modeling long-term harmonic dependencies while handling severe class imbalance inherent in chord vocabularies. By combining local spectral modeling with global temporal context, ChordFormer aims to achieve robust recognition of both common and rare chord types, addressing the long-tailed distribution where extended chords are significantly underrepresented.

ChordFormer builds upon the Conformer architecture, ilustrated in 12, which integrates convolutional neural networks and self-attention mechanisms within a unified block. This hybrid design enables the model to effectively capture both local time–frequency patterns and long-range dependencies in chord sequences. As input, raw audio signals are transformed into Constant-Q Transform (CQT) spectrograms, which provide a musically meaningful time–frequency representation with logarithmic frequency spacing. These features are processed by stacked Conformer blocks composed of feed-forward modules, multi-headed self-attention with relative positional encoding, and depthwise separable convolutional layers. To ensure temporal coherence and reduce spurious chord transitions, the frame-level predictions are further decoded using a linear Conditional Random Field (CRF).

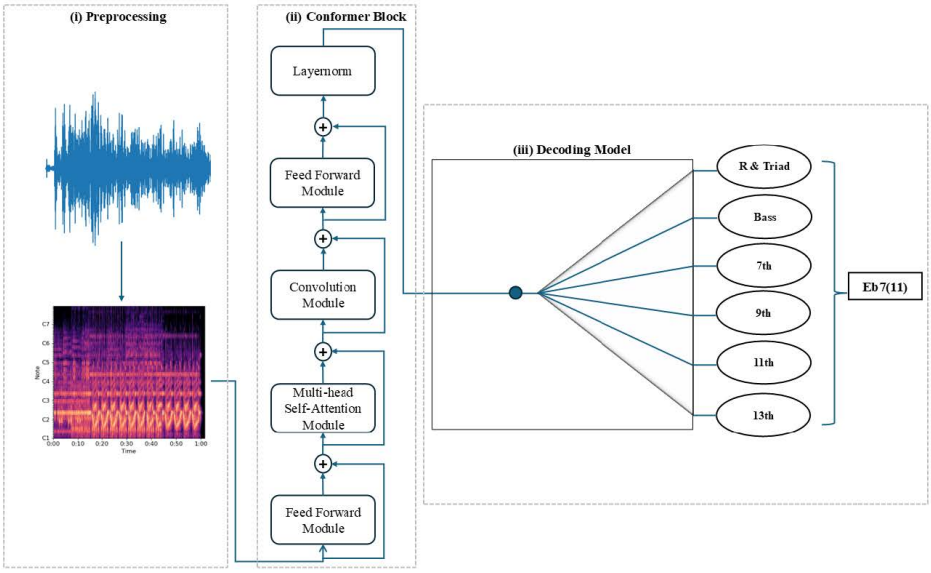


Figure 12: ChordFormer architecture. Audio is transformed to CQT spectrograms, processed by stacked Conformer blocks (combining self-attention and convolution), and decoded to six chord components via independent softmax heads followed by CRF temporal smoothing.

A central contribution of the work is the use of a *structured chord representation* inspired by prior large-vocabulary chord transcription research. Instead of treating each chord as a flat class, chords are decomposed into six musically meaningful components: root+triad, bass, seventh, ninth, eleventh, and thirteenth extensions. Each component is predicted independently using a softmax output, enabling the model to generalize better to rare and complex chords while maintaining interpretability aligned with music theory. To address the pervasive class imbalance across chord components, the authors employ a reweighted cross-entropy loss function that increases the contribution of underrepresented chord classes during training.

The approach is evaluated on 1,217 songs from Isophonics, McGill Billboard, and MARL

collections using five-fold cross-validation. Performance is assessed via Weighted Chord Symbol Recall (WCSR) across multiple granularities and mean frame-wise/class-wise accuracy for large-vocabulary scenarios. Pitch-shift augmentation is applied to improve robustness.

Table 3 presents comparative results against state-of-the-art baselines. ChordFormer achieves 84.69% Root accuracy, 84.09% Major/Minor accuracy, and 83.62% MIREX score, consistently outperforming CNN-, BLSTM-, and Transformer-based models. Notably, the Conformer architecture surpasses pure Transformer approaches (83.62% vs. 72.23% MIREX), demonstrating the value of integrating convolutional and attention mechanisms.

Table 3: Chord recognition performance comparison. WCSR metrics across harmonic granularities. Best results in bold.

Model	Root	Thirds	MajMin	Triads	Sevenths	Tetrads	MIREX
BTC+CNN	54.28	47.94	49.00	44.67	37.99	34.01	47.94
Transformer	78.55	72.91	74.24	67.75	57.42	51.46	72.23
CNN	81.76	78.69	80.88	74.13	67.27	60.48	79.42
Transformer+CNN	82.40	79.40	81.67	74.88	67.84	61.04	80.22
CNN+BLSTM	83.39	80.04	82.62	75.91	69.78	62.87	81.52
ChordFormer-R	83.87	80.54	81.86	76.02	69.80	63.48	82.98
ChordFormer	**84.69**	**81.75**	**84.09**	**77.55**	**72.28**	**65.32**	**83.62**

For large-vocabulary evaluation (301 chord classes), ChordFormer achieves frame-wise accuracy of 0.7877 and class-wise accuracy of 0.3884 without reweighting. With the reweighted loss function (ChordFormer-R at $\gamma = 0.7$, $w_{\max} = 20.0$), class-wise accuracy improves to 0.4471 while frame-wise accuracy decreases to 0.7416, demonstrating an explicit trade-off between overall correctness and balanced recognition across rare chord classes.

These results establish ChordFormer as state-of-the-art for large-vocabulary chord recognition, with particularly strong improvements on extended chords (72.28% Sevenths vs. 69.78% for CNN+BLSTM). The Conformer architecture's superiority over pure Transformers validates the hypothesis that chord recognition benefits from explicit local-global modeling. However, the class imbalance problem remains partially unsolved—even with reweighting, class-wise accuracy (0.4471) lags significantly behind frame-wise accuracy (0.7877), indicating that rare chord types remain challenging.

The structured chord representation proves effective for generalization, but the 6-component decomposition assumes Western tonal harmony conventions. The reliance on supervised learning and evaluation on a single benchmark corpus (despite its size) limits conclusions about true generalization to diverse musical styles. The CRF temporal smoothing, while reducing spurious transitions, may over-smooth genuine rapid chord changes in rhythmically complex music.

The authors conclude that ChordFormer effectively bridges the gap between local spectral modeling and global harmonic context, offering a robust solution for large-vocabulary chord recognition. While ChordFormer advances frame-level harmonic analysis through its hybrid architecture, understanding musical structure also requires accurate temporal analysis of rhythmic patterns. Beat tracking, which identifies the temporal locations of beats and downbeats, represents a complementary challenge that demands different architectural considerations for modeling periodicities and metrical hierarchies, as explored in the following section.

4.3 Transformer-Based Beat Tracking with Multi-Resolution Architecture

This paper [40] presents a novel Transformer-based architecture for musical beat tracking that addresses fundamental challenges in the field through a dual-resolution approach combining a low-resolution encoder with a high-resolution decoder. The primary innovation lies in the model's ability to simultaneously capture global musical structure while maintaining precise temporal localization of beat positions, effectively handling the competing demands of long-range context and fine-grained prediction accuracy.

The primary objective is to overcome two fundamental challenges in beat tracking: processing long sequential inputs efficiently while handling severe data imbalance between beat and non-beat frames. Previous approaches sacrificed either global temporal coherence or local prediction

precision. This work aims to achieve both simultaneously through architectural separation of global feature learning and precise beat localization.

The architecture comprises three main components: initial feature extraction, low-resolution encoder, and high-resolution decoder. Input Mel-spectrograms are first processed by two-dimensional convolutional layers that extract spectral-temporal features. The low-resolution encoder operates on temporally downsampled sequences, creating shorter representations that facilitate global pattern learning without computational burden. This encoder incorporates one-dimensional convolutional layers within the Transformer framework to enhance local feature extraction.

The high-resolution decoder operates at full temporal resolution to predict precise beat locations. A critical innovation replaces traditional positional embeddings with upsampled features from the encoder, enabling the decoder to leverage global context while maintaining temporal precision. Post-processing employs a Dynamic Bayesian Network (DBN) from the Madmom library, which refines raw predictions into final beat sequences by incorporating musical constraints and temporal continuity.

Figure 13 illustrates the complete pipeline, showing the flow from multi-resolution Mel-spectrograms through encoder and decoder to final beat predictions.

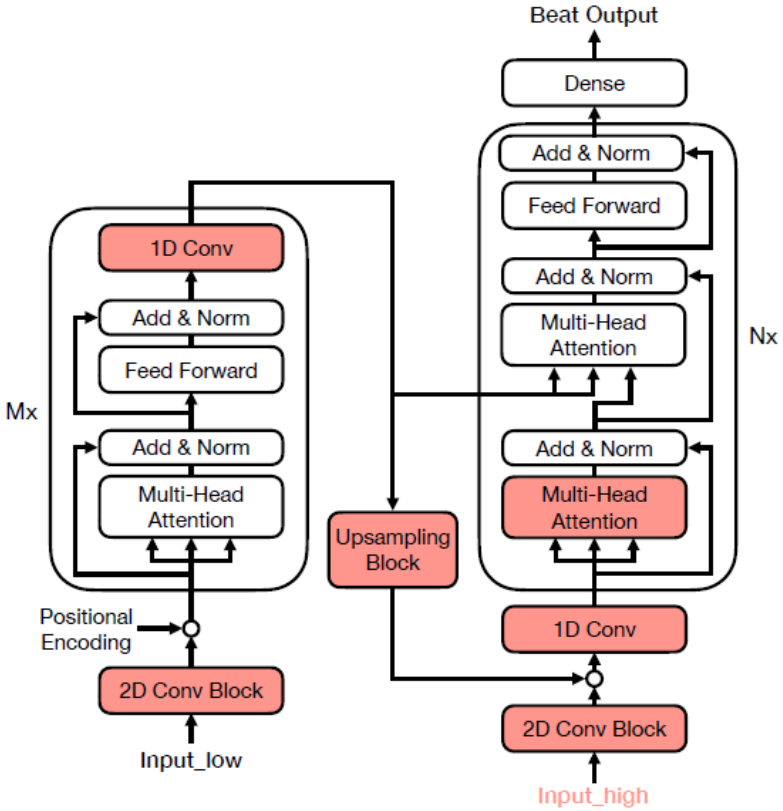


Figure 13: Multi-resolution Transformer architecture for beat tracking. Low-resolution encoder processes downsampled input to capture global structure; high-resolution decoder predicts precise beat times using upsampled encoder features. DBN post-processing ensures temporal coherence.

The training methodology explored three distinct strategies to optimize model performance. The authors compared training the entire model end-to-end, training only the decoder with a frozen encoder, and a two-stage approach involving encoder pre-training followed by joint training of both components. The two-stage training strategy proved most effective, suggesting that establishing robust global feature representations before fine-tuning for precise prediction yields superior results. Data augmentation played a crucial role in improving generalization, with techniques including variation of the hop size parameter during spectrogram computation and application of Harmonic Percussive Source Separation to create additional training examples with different spectral characteristics.

Evaluation is conducted on eight standard datasets representing diverse musical styles: Beatles, Harmonix, RWC, tapcorrect, Ballroom, Hainsworth, SMC, and GTZAN. Performance is assessed

using F-measure (70ms tolerance), CMLt (Continuity-based Metrical Level tracking), and AMLt (Accuracy-based Metrical Level with tolerance).

Table 4 presents results on four representative datasets. The proposed method achieves strong F-measure performance: 95.0% on Ballroom, 87.0% on Hainsworth, 55.4% on SMC, and 88.4% on GTZAN. These results are competitive with or exceed state-of-the-art methods including TCN and Beat Transformer baselines.

Table 4: Beat tracking performance comparison. F-measure with 70ms tolerance, CMLt (continuity), and AMLt (accuracy with alternate levels). Best results in bold.

Dataset	**Method**	**F-measure**	**CMLt**	**AMLt**
Ballroom	Encoder only	93.0	87.4	96.1
	Decoder (Proposed)	**95.0**	91.1	96.4
	Beat Transformer	96.8	**95.4**	**96.6**
	TCN	96.2	94.7	96.1
Hainsworth	Encoder only	88.2	81.0	93.4
	Decoder (Proposed)	87.0	76.2	**93.6**
	Beat Transformer	**90.2**	**84.2**	91.8
	TCN	90.4	85.1	93.7
SMC	Encoder only	55.0	45.8	64.1
	Decoder (Proposed)	55.4	45.1	**65.6**
	Beat Transformer	**59.6**	**45.6**	63.5
	TCN	55.2	46.5	64.3
GTZAN*	Encoder only	87.8	78.5	93.7
	Decoder (Proposed)	**88.4**	**80.8**	**94.0**
	Beat Transformer	88.5	80.0	92.2
	TCN	88.5	81.3	93.1

*GTZAN held out for testing only; others used in 8-fold cross-validation

Ablation studies confirm that the high-resolution decoder significantly improves precision, recovering missed beats and filtering spurious predictions. The two-stage training strategy (encoder pre-training followed by joint fine-tuning) outperforms end-to-end training, validating the importance of establishing robust global representations before precise prediction.

While the model achieves competitive F-measure scores, performance on continuity metrics (CMLt) reveals limitations in global tempo consistency. On Hainsworth, the proposed method scores 76.2% CMLt compared to 84.2% for Beat Transformer, indicating higher rates of phase errors (correct period but misaligned phase) and octave errors (half/double tempo predictions). This discrepancy suggests the architecture excels at local beat prediction but occasionally fails to maintain consistent global interpretation.

The multi-resolution design successfully addresses the input length and data imbalance challenges, as evidenced by strong performance on diverse datasets. However, the reliance on DBN post-processing indicates that learned representations alone may be insufficient for temporal coherence. The model's competitive but not dominant performance (often second-best) suggests that pure Transformer architectures may not optimally capture the periodic structure inherent in musical rhythm compared to specialized architectures like TCN that explicitly model temporal hierarchies.

The strong AMLt scores (often best or near-best) indicate the model successfully captures metrically valid beat placements even when phase-shifted, suggesting the continuity issues stem from local ambiguities rather than fundamental misunderstanding of musical meter.

This work demonstrates that Transformer architectures can be effectively adapted for beat tracking through careful multi-scale design, though specialized temporal modeling remains competitive. The authors suggest future work on multi-task learning jointly optimizing beat tracking, downbeat detection, and tempo estimation to address phase consistency issues.

While beat tracking identifies the temporal grid of music, understanding rhythmic content also requires transcribing the specific drum patterns that drive this pulse. Automatic drum transcription presents complementary challenges involving multi-label classification and onset detection,

requiring architectures capable of handling polyphonic percussion events, as explored in the following section.

4.4 Dual-Path Beat Tracking: Combining Temporal Convolutional Networks and Transformers in Parallel

The paper [41] propose a hybrid architecture combining Temporal Convolutional Networks (TCNs) and Transformers in parallel for beat tracking. TCNs capture fine-grained local temporal dependencies while Transformers model global rhythmic patterns, with outputs fused via learnable parameters before Dynamic Bayesian Network post-processing. This design achieves state-of-the-art performance with significantly fewer parameters (3.68M) than competing methods.

The primary objective is to combine TCN and Transformer strengths in a unified architecture that improves beat tracking accuracy across diverse genres while reducing model complexity. Previous approaches faced trade-offs: TCNs with dilated convolutions capture local patterns efficiently but struggle with complex rhythms; Transformers model long-range dependencies but require substantial training data. This work investigates whether parallel processing can overcome these limitations while maintaining parameter efficiency and improving interpretability.

The architecture comprises four components processing log-magnitude spectrograms (44.1 kHz audio, 4096-sample window, 1024-sample hop). A lightweight convolutional frontend applies three layers with progressively increasing filters (32→64→128), each followed by max-pooling and dropout (0.2), reducing frequency dimensions while preserving temporal resolution.

The parallel processing branches operate independently on 128-dimensional features. The TCN branch employs eleven dilated 1D convolutional layers with exponentially increasing dilation rates (2^0 to 2^{11}), capturing local temporal dependencies efficiently. The Transformer branch consists of six stacked RoFormer layers with rotary positional embeddings, using 16 attention heads (size 8) and pointwise feedforward networks (expansion factor 4), modeling global temporal patterns without explicit frequency modeling.

Feature aggregation combines TCN and Transformer activations using a learnable weighting parameter, preserving unique features from each branch. A fully connected layer projects the combined 128-dimensional representation to beat probability scores. DBN post-processing with Viterbi decoding ensures optimal temporal alignment.

Figure 14 illustrates the complete parallel architecture from spectrogram input through independent TCN/Transformer processing to fused beat predictions.

Figure 14: Dual-path architecture combining TCN and Transformer in parallel. Both branches process the same continuous input independently; outputs are fused via learnable parameters before DBN post-processing.

This design differs from prior work (SpecTNT-TCN) by using parallel rather than sequential processing, sophisticated fusion instead of simple averaging, and processing entire sequences without chunking.

Training employs eight-fold cross-validation on four datasets (Ballroom, Hainsworth, Harmonix, SMC) with GTZAN held out for testing. The model processes entire sequences of varying lengths, with beat locations quantized to spectrogram frame rate. Target regions are expanded by ± 2 frames with decaying weights (0.5 for adjacent, 0.25 for next) to account for temporal uncertainty.

Binary Cross-Entropy loss with AdamW optimizer is used, with initial learning rate 1×10^{-3} reduced by factor of 5 when validation loss plateaus. No data augmentation or specialized loss functions are employed. Training converges in approximately 20 hours per fold on RTX 4090 GPU.

Evaluation uses standard metrics across five datasets: F-Measure (70ms tolerance), CMLt (continuity-based), and AMLt (accuracy-based with loose tolerance). Table 5 presents comparative results.

Table 5: Dual-path beat tracking performance across datasets. Beat This achieves highest F-Measure but CMLt/AMLt not reported. Best results per metric in bold.

Dataset	Method	F-Measure	CMLt	AMLt
Ballroom	TCN	0.933	0.864	0.881
	SpecTNT-TCN	0.962	0.939	0.967
	Demixed DSA	0.968	**0.954**	0.966
	Beat This	**0.975**	–	–
	Dual-Path (Ours)	0.961	0.937	0.962
Hainsworth	TCN	0.874	0.755	0.795
	SpecTNT-TCN	0.877	0.862	0.915
	Demixed DSA	0.902	0.842	**0.918**
	Beat This	**0.919**	–	–
	Dual-Path (Ours)	0.883	**0.916**	0.906
Harmonix	SpecTNT-TCN	0.953	**0.939**	0.959
	Demixed DSA	0.954	0.905	0.957
	Beat This	**0.958**	–	–
	Dual-Path (Ours)	0.954	0.904	**0.964**
GTZAN*	TCN	0.843	0.695	0.715
	SpecTNT-TCN	0.887	0.812	**0.920**
	Demixed DSA	0.885	0.800	0.922
	Dual-Path (Ours)	0.867	0.774	0.905

*GTZAN held out for testing; others used in 8-fold cross-validation
Parameters: Dual-Path 3.68M, SpecTNT-TCN 4.64M, Demixed DSA 9.29M

The model achieves competitive F-Measure scores across all datasets, with particularly strong continuity performance on Hainsworth (0.916 CMLt, exceeding all baselines) and highest AMLt on Harmonix (0.964). Remarkably, these results are achieved with 3.68M parameters—21% fewer than SpecTNT-TCN and 60% fewer than Demixed DSA—without data augmentation or specialized loss functions.

The parallel architecture achieves competitive but not state-of-the-art F-Measure performance—Beat This consistently achieves highest F-Measure scores (e.g., 0.975 vs. 0.961 on Ballroom). However, direct comparison is challenging: Beat This does not report CMLt/AMLt metrics, suggesting different evaluation protocols or post-processing strategies. The Dual-Path model's strength lies in balanced performance across all three metrics while maintaining parameter efficiency.

The model demonstrates particularly strong continuity on Hainsworth (0.916 CMLt, exceeding SpecTNT-TCN's 0.862 and Demixed DSA's 0.842), suggesting effective capture of metrical structure. The parameter efficiency (3.68M vs. 9.29M for Demixed DSA) while maintaining competitive performance validates the hypothesis that parallel local-global modeling can match specialized architectures without extensive augmentation or complex training procedures.

Grad-CAM visualization reveals that mid-range frequencies (drums, snare) correlate strongly with beat detection, with periodic low/high-frequency activations at 4-5 beat intervals corresponding to downbeats. However, framewise classification limits ability to capture temporal dependencies, and Grad-CAM struggles to reveal cross-frequency interactions essential for rhythmic understanding. This highlights an ongoing challenge in interpretability for sequential music analysis tasks.

This work demonstrates that parallel combination of TCN and Transformer architectures achieves state-of-the-art beat tracking with substantially reduced model complexity (3.68M parameters). The architecture successfully balances local and global temporal modeling without requiring extensive data augmentation or complex loss functions, representing an efficient solution for beat tracking across diverse musical genres.

While beat tracking identifies the temporal grid underlying musical rhythm, understanding the specific instrumental events that create this rhythm requires detailed transcription. Automatic drum transcription, which localizes and classifies individual percussion onsets, presents complementary challenges involving multi-label frame-level prediction and handling instrument-specific timbral variations, as explored in the following section.

4.5 Vocal Melody Extraction via HRNet-Based Singing Voice Separation and Encoder-Decoder-Based F0 Estimation

The paper [42] presents a two-stage system for vocal melody extraction (VME) that addresses the fundamental challenge of accurately estimating the fundamental frequency (F0) of singing voices in polyphonic music. The core contribution lies in combining a High-Resolution Network (HRNet) for singing voice separation with a custom encoder-decoder network for F0 estimation, enabling each task to be optimized independently with precise annotations while avoiding the limitations of joint training approaches.

Vocal melody extraction represents a critical task in music information retrieval with applications spanning query by humming, cover song identification, and music transcription. The primary difficulty stems from the harmonic mixing of multiple instruments and singing voices in polyphonic music, which obscures the identification of individual F0 values. Even when F0 values can be correctly identified, determining whether they belong to the leading melody requires substantial effort. Prior source separation-based methods have shown promise in overcoming these challenges, yet existing approaches face significant limitations. Joint training methods that simultaneously optimize singing voice separation and melody extraction are constrained by limited datasets containing both pure vocal tracks and corresponding F0 annotations, often relying on automatically generated annotations that contain errors. Furthermore, such systems demonstrate poor robustness to backing vocals, which belong to the vocal category in separation tasks but should not be included in melody extraction. The authors propose a two-stage approach to circumvent these limitations, allowing each task to leverage datasets with precise annotations while remaining unaffected by backing vocals.

The methodology comprises two distinct stages, each addressing a specific aspect of the vocal melody extraction problem. The first stage employs a High-Resolution Network for singing voice separation, taking magnitude spectrograms of mixed audio as input and outputting soft masks for vocal signals. Unlike conventional architectures such as U-Net and Stacked Hourglass Networks that recover high resolution from lower resolutions, the HRNet maintains high-resolution representations throughout the entire process by connecting multi-resolution branches in parallel and performing multi-resolution fusion. The architecture begins with two convolutional layers that preserve input resolution, followed by four main stages. The first stage contains four residual units with bottleneck architecture, while subsequent stages comprise modules that perform parallel multi-resolution convolutions across branches with different resolutions (1, 1/2, 1/4, and 1/8) and fuse these representations through bilinear upsampling for low-to-high transitions and strided convolutions for high-to-low transitions. The final output concatenates rescaled feature maps from all resolution levels and applies convolutional layers to produce the target vocal mask.

The second stage utilizes an encoder-decoder network specifically designed to estimate F0 values from the separated vocal spectrograms. The architecture begins with two convolutional layers, followed by three encoder blocks containing max-pooling layers applied only to the frequency axis to preserve temporal resolution, along with convolutional layers incorporating batch normalization, leaky ReLU activation, and dropout regularization. Two decoder blocks follow, each consisting of deconvolutional and convolutional layers with skip connections linking encoder and decoder layers at matching resolutions. The output feature maps are flattened along the frequency axis, producing 1024-dimensional feature vectors for each frame, which are then fed into a fully connected layer with softmax activation to classify frames into 362 pitch classes (including an unvoiced class) spanning from 73.41 Hz to 987.77 Hz with 1/8 semitone resolution.

The HRNet for singing voice separation was trained on a dataset of 400 songs (approximately 29 hours) constructed by randomly mixing vocals and accompaniment from the MUSDB18 training subset. The network processes magnitude spectrograms derived from Short-Time Fourier Transform with a window size of 1024 and hop size of 512, with audio downsampled to 16 kHz. Training employed the ADAM optimizer with a learning rate of 0.0001 and batch size of 5, minimizing an L1 norm loss function between estimated and target vocal spectrograms. The encoder-decoder network for F0 estimation was trained on a combination of iKala (225 songs), RWC Popular Music (85 songs), and MedleyDB (49 songs) datasets. Input spectrograms were calculated using STFT with a window size of 1024 and hop size of 80 at 8 kHz sampling rate. Training utilized the ADAM optimizer with an initial learning rate of 0.0015 and batch size of 16, minimizing cross-entropy loss with the learning rate decaying to 98 percent of its previous value after each epoch.

The evaluation employed three public test datasets: ADC2004 (12 vocal-dominated songs), MIREX05 (9 vocal-dominated excerpts), and MIR1k (1000 Chinese song clips). Five standard

metrics were calculated using the mir_eval toolbox: voicing recall rate (VR), voicing false alarm rate (VFA), raw pitch accuracy (RPA), raw chroma accuracy (RCA), and overall accuracy (OA). An estimated pitch was considered correct when falling within 50 cents of the ground truth. The experimental design included two primary evaluations. The first assessed the effectiveness of HRNet-based singing voice separation by comparing four melody extraction systems formed by pairing different SVS methods (HRNet, SHNet, U-Net, WaveUnet) with the proposed encoder-decoder network. The second evaluation compared the complete two-stage system against four state-of-the-art melody extraction methods (DSM, SEG, SED, JDC) and included an ablation study contrasting the full system (HR-ED) with a baseline (EDNet) that directly processes mixture audio without prior separation.

The experimental results demonstrated that the HRNet-based singing voice separation method significantly outperformed other SVS approaches when combined with the encoder-decoder network for melody extraction. On the ADC2004 dataset, the HR-ED system achieved gains of 0.4 percent in VR, 0.7 percent in RPA, 0.1 percent in RCA, and 2.0 percent in OA compared to the best existing methods, though VFA was 1.8 percent higher. Similar patterns emerged on MIREX05 (gains of 0.6 percent VR, 0.7 percent RPA, 0.6 percent RCA, 1.2 percent OA) and MIR1k (gains of 5.3 percent VR, 7.4 percent RPA, 5.8 percent RCA, 8.4 percent OA). The ablation study revealed the critical importance of the separation stage: the EDNet baseline achieved high RPA and RCA values but suffered from elevated VFA rates (26.2 percent on ADC2004, 19.8 percent on MIREX05, 29.8 percent on MIR1k), indicating poor discrimination between vocal and non-vocal segments. The complete HR-ED system dramatically reduced VFA to 9.4 percent, 4.9 percent, and 7.2 percent on the respective datasets while simultaneously improving all other metrics. Case studies illustrated both the system's strengths and limitations: successful examples showed effective removal of accompaniment with preserved vocal harmonics, correcting false-negative melodic frames and accompaniment-induced errors, while failure cases revealed instances where vocal segments were mistakenly removed during separation, resulting in missing F0 sequences in the final output.

The authors acknowledge several limitations requiring further investigation. The primary shortcoming involves the occasional removal of genuine vocal segments during the singing voice separation stage, as demonstrated in failure cases where correctly estimated F0 sequences from mixture audio were lost after separation. This suggests that the HRNet-based SVS, while effective at reducing accompaniment interference, may sometimes be overly aggressive in its filtering. The relatively higher VFA values compared to some competing methods indicate that separated vocals may retain more instrument sounds than desired, though this does not prevent the system from achieving superior overall accuracy. The authors propose vocal enhancement as a potential solution, suggesting that adding separated vocals back to the original mixture rather than solely removing accompaniment might preserve more vocal information while still reducing interference. The limited size of public datasets containing precise annotations for both singing voice separation and vocal melody extraction constrains the potential for more sophisticated training approaches. The paper does not extensively discuss computational efficiency or real-time processing capabilities, nor does it address performance on diverse musical genres beyond those represented in the evaluation datasets.

This work makes significant contributions to vocal melody extraction by demonstrating that a carefully designed two-stage approach can substantially outperform both end-to-end methods and traditional melody extraction algorithms. The key insight that maintaining high-resolution representations throughout the separation process leads to more spatially precise outputs translates directly into improved melody extraction accuracy. By decoupling the singing voice separation and F0 estimation tasks, the authors enable each component to be trained with appropriate datasets and precise annotations, avoiding the compromises inherent in joint training approaches. The comprehensive evaluation across multiple datasets and comparison with numerous state-of-the-art methods provides strong evidence for the effectiveness of this approach. The work advances the state of the art in music information retrieval by showing that architectural innovations from computer vision, specifically the HRNet's parallel multi-resolution processing, can be successfully adapted to audio processing tasks with substantial performance gains. The identified limitations regarding occasional vocal removal and the proposed direction toward vocal enhancement rather than pure separation suggest promising avenues for future research that could further improve the robustness and applicability of melody extraction systems.

4.6 Mel-RoFormer: Vocal Separation and Vocal Melody Transcription

The paper [43] introduces *Mel-RoFormer*, a spectrogram-based Transformer architecture designed for modeling complex music audio signals, with a particular focus on two core music information retrieval (MIR) tasks: vocal separation and vocal melody transcription. The work is motivated by the observation that music signals exhibit rich and structured information along both time and frequency dimensions, and that conventional approaches often fail to explicitly model frequency as a sequential dimension. Building upon prior advances in time–frequency Transformers, notably BS-RoFormer, the authors propose architectural modifications that improve generalization and performance by aligning frequency modeling with perceptual principles.

At the core of Mel-RoFormer lies a novel *Mel-band Projection* front-end. Instead of relying on convolutional layers or empirically defined non-overlapping subbands, the model employs a Mel-scale-based band-division scheme that produces overlapping frequency subbands. Each Mel-band is processed independently by a small multi-layer perceptron (MLP) that projects raw complex spectrogram subbands into latent embeddings. This design can be interpreted as a learnable Mel filter-bank, combining perceptually motivated frequency partitioning with flexible data-driven feature learning. The resulting band-wise embeddings are stacked and passed to the main Transformer backbone.

The Transformer component consists of multiple interleaved RoPE Transformer encoders, referred to as RoFormer blocks. These blocks alternately model the data as sequences over time and sequences over frequency bands, explicitly disentangling and recombining temporal and spectral dependencies. Rotary Position Embeddings (RoPE) are used to preserve positional information under repeated reshaping operations, enabling effective interaction between time-indexed and band-indexed representations. Following the Transformer stack, an *Embedding Projection* module composed of band-specific MLPs generates task-dependent outputs.

For vocal separation, Mel-RoFormer estimates complex ideal ratio masks (cIRMs) directly on the complex spectrogram. Overlapping Mel-band predictions are averaged to produce a full-resolution mask, which is applied to the input spectrogram and inverted via iSTFT to reconstruct the separated vocal signal. Training employs a combination of time-domain mean absolute error and multi-resolution complex spectrogram losses. Experiments are conducted under multiple training scenarios using MUSDB18-HQ, MoisesDB, and a large proprietary in-house dataset, demonstrating that Mel-RoFormer consistently outperforms BS-RoFormer and other strong baselines in terms of signal-to-distortion ratio (SDR).

For vocal melody transcription, the authors adopt a two-stage training strategy. A Mel-RoFormer model pretrained on vocal separation is fine-tuned for melody transcription by replacing the projection head and using uniform band-wise embeddings. The downstream transcription system follows an *onsets and frames* paradigm, employing separate predictors for note onsets and note activity. Evaluations on MIR-ST500 and POP909 show that Mel-RoFormer achieves state-of-the-art performance across all reported metrics, including Correct Onset, Correct Onset and Pitch, and the more challenging Correct Onset, Pitch, and Offset. The results highlight particularly strong robustness in note offset detection, a known difficulty in singing transcription.

The paper further analyzes the effects of model size, training data scale, and pretraining strategies. Smaller Mel-RoFormer variants achieve competitive results with substantially fewer parameters, making them suitable for resource-constrained settings. The authors also discuss dataset-related limitations, noting annotation inconsistencies in POP909 and the impact of label quality on evaluation outcomes. Overall, Mel-RoFormer is presented as an effective and versatile foundation model for music audio processing, demonstrating that perceptually grounded frequency modeling combined with interleaved time–frequency Transformers yields strong performance across both signal-level and symbolic MIR tasks.

4.7 Real-Time Automatic Drum Transcription Using Dynamic Few-Shot Learning

The paper [44] investigates the application of *dynamic few-shot learning* (FSL) to real-time automatic drum transcription (ADT), a challenging subproblem of automatic music transcription in polyphonic audio. ADT aims to detect and classify drum sound onsets belonging to multiple instrument classes in the presence of melodic accompaniment. While most existing approaches focus on offline processing and fixed class vocabularies, this work addresses two key limitations:

(i) the lack of real-time capable ADT systems with low latency, and (ii) the difficulty of adapting pretrained models to new or poorly represented drum classes with only a few examples. The authors adapt a dynamic FSL framework originally proposed for visual recognition to the audio domain, demonstrating that it can be effectively integrated into a real-time ADT pipeline.

The proposed system consists of a convolutional embedding network, a prototype-based classifier, and a dynamic few-shot prototype generator. Input audio is represented as log-mel spectrograms with 96 frequency bands, computed from 48 kHz audio using a 1024-point STFT. The embedding network comprises five convolutional layers followed by two fully connected layers, producing 384-dimensional latent representations. Classification is performed by computing cosine similarities between embeddings and class prototypes. Unlike standard softmax-based multi-class classifiers, the authors introduce a learned *negative class prototype* representing the absence of drum sounds, enabling independent probability estimation for each drum class and improving robustness to superimposed events. This design allows multiple drum classes to be active simultaneously, which is essential for realistic drum performances.

Dynamic few-shot learning is employed to handle novel or underrepresented drum classes. Base class prototypes are learned during an initial training stage without few-shot learning, while a second training stage optimizes a prototype generator that can synthesize novel class prototypes from a small number of examples. The generator combines an average embedding of the few-shot examples with an attention-based component that leverages similarities to existing base class prototypes. This mechanism allows the model to learn new drum classes or adapt to new timbral characteristics at inference time without retraining the embedding network, thereby avoiding catastrophic forgetting.

A further contribution of the paper is the introduction of the *Separate-Tracks-Annotate-Resynthesize* (STAR) dataset, a derived training dataset designed to bridge the gap between synthetic MIDI-based data and real music recordings. STAR is constructed by separating drum and non-drum stems, automatically annotating drum onsets, resynthesizing drum tracks using virtual instruments, and remixing them with recorded melodic instruments and vocals. This process yields 1200 tracks (80.7 h) with perfect temporal alignment between audio and annotations, without requiring manual labeling. STAR is used for training, while evaluation is conducted on three publicly available datasets: MDB Drums, ENST Drums, and RBMA13.

The system is evaluated using event-based, onset-level micro-averaged F-measure computed with `mir_eval`. Offline evaluation uses a ± 30 ms onset tolerance, while real-time evaluation adopts a forward tolerance window up to 60 ms to account for causal processing constraints. Experimental results show that the proposed real-time system achieves performance comparable to state-of-the-art offline ADT methods, with an average detection delay of approximately 43 ms. Increasing the temporal context in offline settings further improves performance. The authors also demonstrate successful few-shot learning of novel drum classes and effective fine-tuning for electronic drum sounds using only a handful of examples.

The paper discusses several limitations. The STAR dataset is not publicly released, which limits reproducibility. Performance on electronic drum sounds remains lower when training data is dominated by acoustic drums, and a trade-off persists between temporal context and latency in real-time operation. Nevertheless, the work provides a strong demonstration that dynamic few-shot learning can be integrated into low-latency ADT systems, enabling adaptable, real-time drum transcription with competitive accuracy.

4.8 Noise-to-Notes: A Diffusion-Based Model for Drum Transcription

Noise-to-Notes (N2N) [45] is a generative framework that redefines Automatic Drum Transcription (ADT) by shifting from the traditional discriminative formulation to a diffusion-based generative modeling paradigm. Instead of directly classifying onsets from spectrograms, the system learns to transform audio-conditioned Gaussian noise into structured symbolic drum events, including both onset timing and velocity. This reframing introduces a controllable speed–accuracy trade-off, enables inpainting when audio is partially missing, and supports unconditional generation, capabilities not present in prior ADT systems.

At the core of N2N is an audio-conditioned diffusion model built upon a transformer-based decoder architecture adapted from EDGE. The system operates on frame-level drum representations, where each frame encodes binary onset activity and continuous velocity values for multiple drum components. A key methodological innovation is the introduction of the *Annealed Pseudo-*

Huber loss, designed to jointly optimize onset and velocity despite their different statistical properties. The loss smoothly interpolates between mean-squared and mean-absolute formulations during training, addressing imbalance between sparse onset events and dense velocity values.

The conditioning pipeline integrates two complementary sources of acoustic information: log mel-spectrograms and intermediate representations extracted from the MERT music foundation model. These higher-level embeddings improve robustness to domain shifts in timbre and recording conditions, particularly for external datasets such as MedleyDB Drums and IDMT-SMT Drums. To support generative behaviors, the model applies both partial dropout (masking contiguous audio segments) and full dropout (removing all audio features), allowing N2N to learn reconstruction from incomplete evidence and to synthesize plausible drum sequences in the absence of audio.

During inference, N2N progressively denoises sampled latent variables across a variable number of diffusion steps. Even with as few as five sampling steps, the model achieves state-of-the-art performance on the E-GMD benchmark, surpassing previous discriminative systems including CRNN-based models and transformer baselines. Increasing the number of sampling steps yields further gains, with strong results also observed for out-of-domain evaluation. Component-wise analysis shows notable improvements for challenging drum classes such as hi-hat and cymbals, reflecting the advantage of combining spectrogram and music-foundation-model features.

Beyond transcription accuracy, N2N demonstrates compelling generative behavior. Through inpainting, the system can reconstruct drum activity in masked temporal spans while maintaining musical coherence with the surrounding audio. Unconditional generation produces structured rhythmic patterns, illustrating the capacity of the diffusion prior to internalize statistical regularities of drum performance. These capabilities highlight a broader shift in MIR research toward generative formulations of traditionally discriminative tasks.

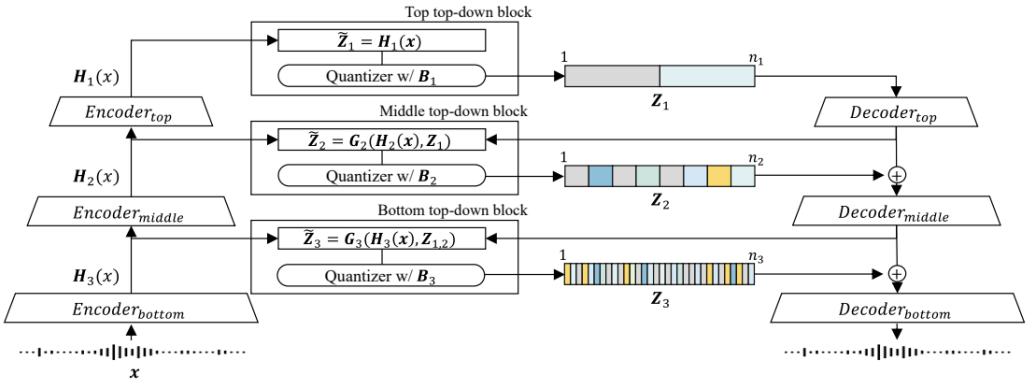
Overall, *Noise-to-Notes* positions diffusion modeling as a powerful alternative to conventional ADT pipelines. By unifying transcription, refinement, and generation within a single framework, the work expands the functional scope of drum transcription systems and establishes the first generative model to outperform discriminative baselines on multiple ADT benchmarks. Its architectural design, loss formulation, and integration of music foundation models provide a foundation for future research on hybrid generative–discriminative MIR approaches.

4.9 SoniDo: A Music Foundation Model for Hierarchical Feature Extraction

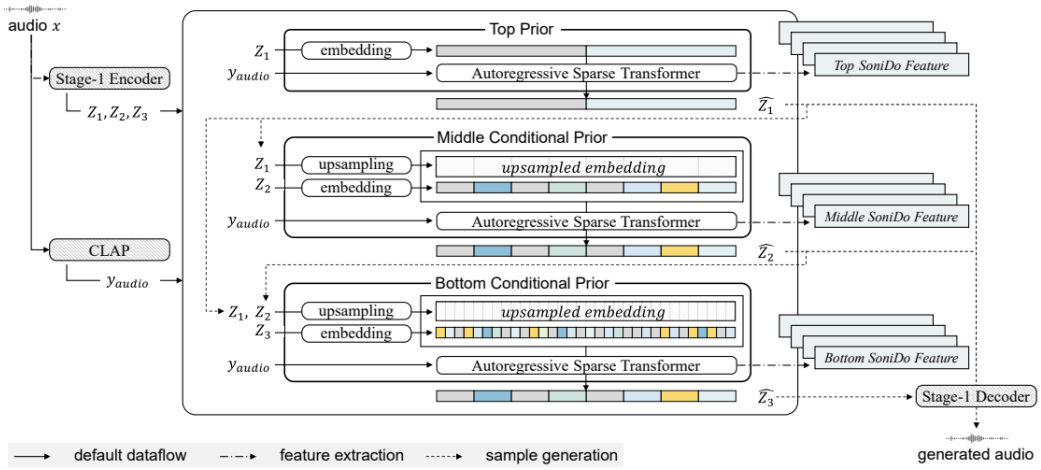
The paper [46] presented SoniDo, a novel music foundation model designed to extract hierarchical features that enhance performance across diverse music downstream tasks, including both understanding and generative applications. The authors addressed a critical gap in music information retrieval and production, noting that while natural language processing has benefited from powerful foundation models, the music domain lacked a comparable model capable of handling the complexity and diversity of music production tasks. The research was motivated by the need for a unified framework that could provide transferable representations to improve multiple downstream applications simultaneously, particularly in scenarios where training data or computational resources are limited.

The technical approach employed by SoniDo consisted of a two-stage generative architecture. In the first stage, the authors implemented a Hierarchically Quantized Variational Autoencoder (HQ-VAE) to learn hierarchical representations of music at multiple levels of granularity. This design choice enabled the model to capture both coarse-grained structural information and fine-grained acoustic details within a unified framework. The second stage utilized sparse transformers to model the token streams generated by the HQ-VAE, allowing for efficient processing of the hierarchical representations. The key innovation lay in the extraction of intermediate representations, termed "SoniDo features," from the trained model. These features encoded rich musical information at different hierarchical levels and could be transferred to downstream tasks.

The methodology for applying SoniDo features to downstream tasks involved preprocessing the extracted features and injecting them into task-specific models. The authors demonstrated the versatility of this approach across four distinct music processing domains: music tagging, music transcription, source separation, and audio mixing. For each task, the SoniDo features were integrated into existing state-of-the-art architectures, serving as additional input channels or conditioning signals that augmented the task-specific models' capabilities. This design allowed the foundation model to act as a generic feature extractor without requiring extensive task-specific



(a) Stage 1 model architecture



(b) Stage-2 model architecture

Figure 15: Architecture overview of SoniDo

modifications.

The evaluation framework encompassed multiple benchmarks specific to each downstream task. For music tagging, the authors assessed performance on standard classification datasets. Music transcription experiments focused on automatic music transcription accuracy metrics. Source separation tasks were evaluated using signal-to-distortion ratio and related metrics, while mixing tasks employed perceptual quality measures. The experimental design included ablation studies to investigate the contribution of features from different hierarchical levels of the SoniDo model, providing insights into which representations were most beneficial for specific tasks.

The results demonstrated that SoniDo features consistently improved the performance of downstream models across all evaluated tasks. The improvements were particularly pronounced in low-data regimes, where models augmented with SoniDo features achieved substantially better performance than baseline models trained from scratch. In several cases, the integration of SoniDo features led to new state-of-the-art results on established benchmarks. The ablation studies revealed that different tasks benefited from different combinations of hierarchical features, with some tasks requiring primarily coarse-grained information while others benefited from fine-grained acoustic details. The authors also observed significant improvements in training efficiency, with models incorporating SoniDo features converging faster and requiring fewer training iterations to achieve competitive performance.

The paper acknowledged several limitations of the proposed approach. A notable finding was that incorporating features from all three hierarchical levels of SoniDo did not universally improve performance across all tasks. In some cases, including fine-grained features introduced noise or irrelevant information that degraded performance, suggesting the need for task-specific feature selection mechanisms to filter out non-contributory information. The authors recognized that

determining the optimal combination of hierarchical features for each downstream task remained an open challenge requiring further investigation. Additionally, while the model demonstrated strong performance across multiple tasks, the paper did not extensively explore potential biases in the training data or discuss computational costs associated with feature extraction at inference time.

The concluding assessment positioned SoniDo as a significant advancement in music foundation models, demonstrating that hierarchical feature extraction from a generative model could serve as a generic booster for diverse music downstream tasks. The work's primary contribution lay in showing that a single foundation model could provide transferable representations beneficial across understanding and generative tasks simultaneously, addressing a long-standing challenge in music information retrieval. The demonstrated improvements in both performance and training efficiency, particularly in data-scarce scenarios, suggested that SoniDo features could facilitate more accessible and efficient development of music processing systems. The research opened avenues for future work in adaptive feature selection and task-specific optimization of hierarchical representations, while establishing a new paradigm for leveraging foundation models in music production and analysis applications.

5 Automatic Speech Recognition (ASR)

5.1 wav2vec 2.0: Self-Supervised Learning of Speech Representations

wav2vec 2.0 [34] is a self-supervised framework for learning powerful speech representations directly from raw audio. The core insight is that large quantities of unlabeled speech can be leveraged by masking latent representations of the input waveform and training a Transformer encoder to solve a contrastive task. After pre-training, the model can be fine-tuned with a Connectionist Temporal Classification (CTC) objective for downstream automatic speech recognition (ASR), achieving state-of-the-art performance with limited labeled data.

The architecture consists of three main components: a multi-layer convolutional *feature encoder* that transforms the waveform into latent speech representations; a Transformer-based *context network* that integrates information across long temporal spans; and a *quantization module* that discretizes latent features into codebook entries used as prediction targets. Unlike earlier approaches such as vq-wav2vec, wav2vec 2.0 learns both contextualized representations and discrete speech units jointly, enabling end-to-end optimization and improved performance.

During self-supervised pre-training, spans of latent representations are masked, and the context network must distinguish the true quantized representation of each masked timestep from a set of distractors sampled from within the same utterance. The contrastive objective encourages the model to capture informative acoustic structure, while an auxiliary diversity loss promotes balanced utilization of codebook entries. This formulation allows the Transformer to learn robust and generalizable speech representations without requiring transcriptions.

Fine-tuning adds a randomly initialized linear projection on top of the context network to map representations to character or phoneme vocabularies. Using the CTC loss and modest amounts of labeled data, wav2vec 2.0 attains remarkable results: with only 10 minutes of transcribed speech, the model achieves word error rates (WER) of 4.8/8.2 on the Librispeech test-clean/other sets. With an hour of labeled data, it surpasses previous semi-supervised approaches while using two orders of magnitude fewer annotations. When fine-tuned on all 960 hours of Librispeech, the model reaches 1.8/3.3 WER, competitive with or exceeding more complex semi-supervised pipelines.

The authors also evaluate phoneme recognition on TIMIT, where wav2vec 2.0 achieves a phoneme error rate (PER) of 8.3, establishing a new state of the art. Additional analyses illustrate how discrete latent units correlate with phonetic categories and how model size, masking strategy, and quantization choices affect performance.

wav2vec 2.0 demonstrates that self-supervised pre-training on raw audio can dramatically reduce the need for labeled speech data while improving recognition accuracy across resource settings. Its combination of contrastive learning, Transformer-based context modeling, and learned discrete units provides a scalable foundation for future research in speech processing and representation learning.

5.2 XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale

XLS-R [47] is a large-scale model based on wav2vec 2.0 [34], trained on significantly more data compared to previous works, covering a wide range of tasks and languages. Trained on 436k hours of annotated speech data from 128 languages, the wav2vec 2.0 Transformer is further fine-tuned on several multilingual speech tasks.

The authors pre-train large wav2vec 2.0-based Transformer models (ranging from 0.3 B to 2 B parameters) the corpus of unlabeled multilingual formed by several public corpora (VoxPopuli, MLS, CommonVoice, VoxLingua107, BABEL), using self-supervised contrastive learning where spans of feature encoder outputs are masked and predicted with distractors; models are optimized with Adam, with a 32k step warm-up followed by polynomial decay and trained for 1 M updates with long audio crops (up to 20 s), large effective batch sizes using activation checkpointing and fully sharded training to scale to many GPUs, and the multilingual data is balanced via upsampling languages and corpora in the batches; downstream finetuning for tasks like speech translation and recognition uses Adam with task-specific schedules and hyperparameters, often with large batch sizes and warmed-up learning rates.

The XLS-R models achieve state-of-the-art performance across a diverse set of multilingual speech tasks: on the CoVoST-2 speech translation benchmark they improve average BLEU by about 7.4 points over prior best systems across 21 directions into English, with particularly large

gains on mid- and low-resource languages; for automatic speech recognition (ASR), XLS-R substantially outperforms previous methods on datasets such as BABEL, Multilingual Librispeech (MLS), CommonVoice, and VoxPopuli, typically lowering error rates by roughly 14–34 relative to prior best results; on language identification (e.g., VoxLingua107) the model sets a new top benchmark with lower error rates than earlier approaches; and across these evaluations larger model sizes and cross-lingual pretraining consistently yield stronger performance, outperforming English-only pretraining for some translation directions.

5.3 Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages

Th work in [48] introduces a multilingual automatic speech recognition (ASR) system developed by Google, capable of functioning in over 100 languages and based on a 2-billion-parameter Conformer model. The work explores how speech recognition can be extended to "long-tail" languages, where transcribed data is scarce, emphasizing the massive use of unannotated data, both audio and text. The overall objective is to demonstrate a scalable framework for pre-training and supervised training of large speech models capable of generalizing to new languages, new domains, and auxiliary tasks such as automatic speech translation.

Speech recognition has gradually evolved from monolingual models to universal models that can cover multiple languages, domains, and tasks simultaneously. The major challenge is the lack of transcribed data for many languages: manual transcription is expensive, and resources are limited. In contrast, vast volumes of unannotated audio and monolingual text are available online. The paper "Scaling Automatic Speech Recognition Beyond 100 Languages" introduces the Universal Speech Model (USM), a large-scale multilingual automatic speech recognition system capable of operating in over 100 languages, including extremely low-resource languages. The model is based on a Conformer architecture with up to 2 billion parameters, trained through a three-stage pipeline that combines both supervised and unsupervised audio and text data.

In the first stage, the model's encoder is pre-trained using BEST-RQ, a self-supervised method that learns robust acoustic representations from a massive dataset of 12 million hours of unannotated audio from over 300 languages. The second stage introduces MOST (Multi-Objective Supervised Pre-Training), a multimodal scheme that combines speech and text objectives, aligning representations across the two modalities and enabling the efficient integration of 28 billion sentences from the Web-NTL corpus. Finally, the model is fine-tuned for specific tasks, such as ASR or automatic speech translation (AST), using CTC, LAS, or RNN-T transducers.

A key technical contribution is chunk-wise attention, a mechanism that prevents performance degradation on very long audio recordings. This approach surpasses strategies used by existing models such as Whisper and allows stable transcription on recordings lasting minutes or even hours. The authors also explore the use of residual adapters, which add only 2

USM achieves state-of-the-art results across multiple benchmarks: FLEURS (102 languages), CoVoST 2 (AST), SpeechStew, and CORAAL, significantly outperforming Whisper despite using far less transcribed data. Remarkably, USM generalizes well to "unseen" languages through pseudo-labeling, achieving 7–30

The paper demonstrates that, by combining massive pre-training on unannotated data, multimodal learning, and scalable architectures, it is possible to build universal speech recognition systems capable of effectively covering the global linguistic diversity. USM represents a major step toward a universal speech understanding system, capable of high performance across the world's languages, including those with extremely limited resources.

5.4 Robust Speech Recognition via Large-Scale Weak Supervision

This paper [32] investigates whether scaling weakly supervised learning can produce speech recognition systems that generalize robustly across domains, languages, and tasks without the need for dataset-specific fine-tuning. The authors introduce Whisper, a family of sequence-to-sequence Transformer models trained on approximately 680,000 hours of multilingual, multitask speech data collected from the internet. The central claim of the work is that large-scale weak supervision, when combined with model and data diversity, can yield speech recognition systems that approach human-level robustness in zero-shot settings. The motivation for this work arises from limitations in prior speech recognition paradigms. Although self-supervised approaches such as wav2vec

2.0 have demonstrated strong performance when fine-tuned on benchmark datasets, they rely on dataset-specific supervision and often exhibit poor out-of-distribution generalization. The authors argue that such evaluation protocols conflate in-distribution optimization with genuine robustness. In contrast, humans typically perform speech recognition tasks without exposure to the specific training distribution, making zero-shot evaluation a more appropriate benchmark for real-world reliability. Whisper is explicitly designed to address this gap by functioning "out of the box" across diverse conditions.

To construct the training corpus, the authors collect audio–text pairs available online and apply automated filtering techniques to mitigate transcript noise and remove machine-generated captions. The dataset spans 99 languages, includes speech transcription, speech-to-English translation, language identification, and voice activity detection, and preserves naturalistic text formatting rather than relying on normalized transcripts. Audio is segmented into 30-second chunks, enabling scalable training while supporting multitask learning. Although the dataset is noisy compared to curated benchmarks, its scale and diversity are central to Whisper's robustness.

The Whisper architecture is a standard encoder–decoder Transformer, deliberately chosen to avoid confounding architectural novelty with data-scale effects. Audio inputs are converted into log-Mel spectrograms, processed by convolutional layers, and encoded via Transformer blocks. The decoder functions as an audio-conditional language model and predicts sequences of tokens that encode both task instructions and outputs. A unified token-based multitask format allows the model to handle transcription, translation, language identification, timestamp prediction, and no-speech detection within a single framework. This design simplifies traditional speech-processing pipelines, which typically rely on multiple specialized components.

A key contribution of the paper is the zero-shot evaluation methodology. Instead of fine-tuning on benchmark training splits, the authors evaluate Whisper directly on a wide range of speech recognition datasets. Results demonstrate that although Whisper's in-distribution performance on datasets such as LibriSpeech is not state-of-the-art, it substantially outperforms supervised models when evaluated across diverse, out-of-distribution datasets. Compared to a strong supervised wav2vec 2.0 baseline matched for LibriSpeech accuracy, Whisper achieves an average 55

The authors further examine multilingual speech recognition and speech translation. Whisper performs competitively on Multilingual LibriSpeech and achieves strong zero-shot results on low-resource languages, though it underperforms some supervised models on VoxPopuli, a high-resource benchmark. Analysis across the Fleurs dataset reveals a strong correlation between the amount of pre-training data per language and downstream performance, indicating that data imbalance remains a key limitation. For speech translation, Whisper sets a new zero-shot state of the art on CoVoST2, particularly excelling in low-resource settings due to the scale of its weakly supervised translation data. Robustness is further evaluated under additive noise and long-form transcription scenarios. Whisper degrades more gracefully than LibriSpeech-trained models when exposed to both synthetic and real-world noise, especially at low signal-to-noise ratios. In long-form transcription tasks involving recordings lasting minutes to hours, Whisper matches or exceeds the performance of several commercial ASR systems. The authors attribute this success to decoding heuristics such as beam search, temperature fallback, and timestamp-aware window shifting, although they acknowledge persistent failure modes such as hallucination and repetition. Scaling analyses demonstrate consistent improvements with both model size and dataset size across most tasks, though diminishing returns are observed for English speech recognition, likely due to saturation near human performance levels. Multitask and multilingual training initially introduces negative transfer for smaller models, but becomes beneficial at scale, ultimately outperforming English-only training regimes. These findings suggest that large models are better able to exploit shared structure across languages and tasks.

The paper concludes by emphasizing that weakly supervised scaling is a viable and underexplored alternative to self-supervised pretraining for speech recognition. While Whisper has limitations—particularly in low-resource languages, language identification, and decoding reliability—the work establishes a strong foundation for robust, general-purpose speech processing. By releasing models and inference code, the authors aim to encourage further research into scalable, zero-shot speech recognition systems that better reflect real-world deployment conditions.

5.5 Scaling speech technology to 1,000+ languages

This study [49] presents the first large-scale effort to extend modern speech technologies—automatic speech recognition (ASR), language identification (LID), and text-to-speech (TTS)—to over one thousand languages. Motivated by the pressing need to support the world's linguistic diversity while addressing the data scarcity faced by most languages, the authors introduce the *Massively Multilingual Speech* (MMS) project. This initiative integrates novel data collection pipelines, self-supervised speech representation learning, and scalable modeling techniques to reach unprecedented multilingual coverage. The study represents a major advance over existing multilingual systems, which generally support only around one hundred languages.

Most modern speech models rely heavily on labeled speech data—an obstacle for the thousands of low-resource languages for which such data is scarce or nonexistent. The emergence of self-supervised learning techniques (e.g., wav2vec 2.0) has reduced the need for labeled resources, but existing multilingual models still remain limited in linguistic breadth. The MMS project confronts this limitation by constructing new datasets and models that significantly expand coverage. The authors emphasize that linguistic diversity is endangered globally and that broadening speech technology support may help language preservation efforts.

A central contribution of the paper is the creation of two large multilingual datasets.

MMS-lab: 1,107 languages with paired audio–text data. Using readings of the New Testament sourced from various online repositories, the authors build a labeled dataset by aligning long-form audio recordings with multilingual text. This pipeline includes preprocessing, scalable forced alignment, iterative alignment model training, and quality filtering procedures. The final result is a dataset covering 1,107 languages with transcribed speech, ranging widely in duration across languages.

MMS-lab-U and MMS-unlab: 3,809+ languages with unlabeled audio. To support self-supervised pretraining and LID tasks, the authors also collect vastly expanded unlabeled datasets. MMS-lab-U includes speech from 1,362 languages, while MMS-unlab incorporates even more diverse sources, making the total unlabeled coverage exceed 3,809 languages.

The approach extends earlier work such as CMU Wilderness and Common Voice. In direct comparisons, MMS-lab yields lower character error rates (CER) than CMU Wilderness and Common Voice for multiple languages—improving CER by 2.1–4.7% in some languages. The improved forced-alignment mechanism retains significantly more training data, strengthening downstream models.

Self-Supervised Pretraining on 1,406 Languages. Building on these datasets, the authors train large cross-lingual wav2vec 2.0 models of 300 M and 1 B parameters, supporting 1,406 languages using approximately 491,000 hours of speech—several times more than in any prior cross-lingual speech model.

Two key data balancing strategies guide pretraining:

1. **Language balancing** through a sampling distribution regulated by the parameter β_L, preventing high-resource languages from dominating training.

2. **Dataset balancing** across diverse corpora using another exponent β_D.

Training runs for one million updates on clusters of A100 GPUs, employing techniques such as FairScale, Fully Sharded Data Parallel (FSDP), and activation checkpointing to manage memory constraints. This scale of self-supervision enables robust multilingual representations that generalize effectively even for low-resource languages.

The Multilingual Automatic Speech Recognition (ASR) component fine-tunes the pretrained MMS models on the MMS-lab dataset, producing the first ASR system that supports 1,107 languages. The authors highlight that existing ASR systems typically support no more than 100 languages. By leveraging self-supervised representations, MMS ASR achieves high performance while requiring significantly less labeled data per language.

Additional comparisons demonstrate that MMS-lab data leads to better monolingual ASR performance than existing datasets when models are evaluated out-of-domain (e.g., on FLEURS), confirming the value of the new alignment pipeline.

Another major result is the scaling of spoken language identification to 4,017 languages, roughly 40 times more than typical 100-language coverage in prior work.

The authors fine-tune the 1B-parameter MMS model by attaching a linear classifier and training across combined labeled and unlabeled datasets (MMS-lab-U+unlab). They balance across

languages and datasets using the sampling exponents β_L and β_D. Experiments consider varying learning rates, update counts, and sampling strategies.

Although LID models trained solely on existing datasets such as FLEURS and VoxLingua-107 achieve slightly higher in-domain accuracy, MMS-lab-U+unlab provides competitive performance across 72 shared languages, trailing in-domain models by only 1.6–2.1%. Crucially, MMS enables scaling to thousands of languages with only minimal performance degradation, as demonstrated in Table 7 of the paper.

The paper also details Multilingual Text-to-Speech (TTS) models for 1,107 languages. Although specific model architectures are not included in the provided excerpts, the general approach involves training multilingual TTS systems on the aligned MMS-lab dataset. Results demonstrate intelligible and natural-sounding speech in a wide range of languages, including many with extremely limited digital resources.

The MMS project represents a significant step toward democratizing speech technology for the world's languages. By building datasets and models at unprecedented scale, the authors unlock new possibilities for research and practical applications in low-resource settings. They outline key directions for future work:

- Scaling to additional languages and dialects, noting the existence of more than 7,000 global languages, many underrepresented even within high-resource languages such as English.
- Developing multitask models capable of performing various speech tasks jointly rather than relying on separate models for ASR, TTS, and LID.
- Tackling additional speech tasks, such as speech translation, keyword spotting, and intent classification, extending the benefits to more use cases.

Scaling Speech Technology to 1,000+ Languages is a landmark effort demonstrating that multilingual speech processing can be scaled far beyond previously assumed limits. Through innovative dataset construction, extensive self-supervised learning, and careful system design, the MMS project expands high-quality ASR, TTS, and LID to thousands of languages—an increase of 10–40× over prior work. The authors highlight both the technological importance and the cultural significance of supporting the long tail of the world's linguistic diversity.

5.6 Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages

Omnilingual ASR [50] introduces a recipe for designing large-scale ASR systems, allowing for easy integration of neglected languages using just a few samples. Its increased performance is mostly given by the massive training corpus, which in turn leads to enhanced zero-shot performance on unseen languages. The training corpus is formed using a wide collection of publicly available ASR datasets, combined with partner-created private data, along with commissioned ASR data from various native speakers of less-common languages.

The system uses a large multilingual speech encoder trained on an extensive unlabeled speech corpus. This corpus was assembled from all available sources at the time—before ASR fine-tuning and before the full Omnilingual ASR datasets were completed, and was further expanded with a large internal collection. In total, the pre-training data included 3.84 million hours of speech from 1,239 languages plus an additional 460,000 hours without language identification. The final ASR fine-tuning dataset totals 120,710 hours covering 1,690 languages.

Architecturally, the ASR models follow a standard encoder-decoder structure. wav2vec2.0 [34] was used during self-supervised learning of the encoder, which was afterwards scaled to 7B parameters to capture the massively multilingual characteristics of the fine-tuning dataset, compared to the original wav2vec2.0 containing only 300M parameters. The proposed ASR systems are depicted in Figure 16, illustrating a standard ASR and a zero-shot variant aimed to faithfully transcribe from languages un-*heard* during training. This is done by conditioning, during test-time, on several input-output pairs from languages observed during training, while appending the target sample at the end, hence forcing the autoregressive transformer to utilize and generalize from known information to new data. Four encoder sizes are considered for comparison, in both architectures: 300M, 1B, 3B, and 7B parameters. The Autoregressive Transformer Decoder is a 12-layer Transformer with inner dimension 4096 and eight attention heads, totaling 1.2B parameters.

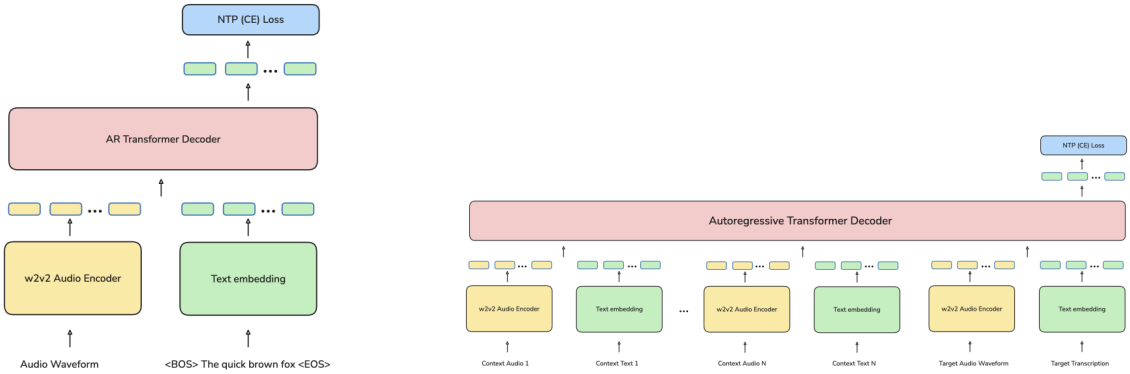


Figure 16: The two ASR variants proposed by Omnilingual ASR [50]. LLM-ASR (left) is a connectionist temporal classification (CTC) approach, which uses a pre-trained wav2vec2.0 and trains a single linear layer connecting the AR Decoder. The right architecture extends LLM-ASR with zero-shot capabilities, allowing it to perform ASR over unseen languages, by providing N context pairs from widely known languages as input, while conditioning on a single sample from the unseen language.

Character Error Rate (CER) is the sole metric used within their study to compare performance with other high-performance ASR systems such as Whisper [32] and Universal Speech Model (USM) [48]. Although trained on a much smaller coverage of languages, Omnilingual ASR outperforms them over their original set of languages, even in extreme low-resource conditions.

5.7 Samba-ASR: State-Of-The-Art Speech Recognition Leveraging Structured State-Space Models

Samba-ASR [51] takes a different turn on the architecture of ASR systems, proposing a Mamba-like encoder decoder architecture [52] using state-space models, in contrast to previous works which leveraged transformer-based models. The motivation lies in the inherent problems of audio transformers which lead to high computational costs when scaled to process long audio sequences. Given the dynamics of state-space models, Samba-ASR results in a more efficient and accurate solution for modern ASR requirements.

The high-level architecture of Samba-ASR resembles the Encoder-Decoder Transformer used by Whisper [32], replacing the initial blocks with Mamba-style blocks [52] which imply state-space models instead of self-attention blocks. This core modification enables efficient processing of long temporal sequences, eliminating the need of point-wise attention scoring. Similarly, the text decoder uses Mamba blocks to transform input tokens into high-level representations, additionally conditioned on audio features through Mamba-cross-connection mechanisms.

The model is trained on a combination of public domain-specific datasets, offering a broad coverage across speaking styles and audio quality: LibriSpeech [53], GigaSpeech [54] and SPGISpeech [55]. While dataset coverage is broad, it is still English-centric and might not reflect speech in other languages or heavily accented speech beyond those already present. WER is used as the sole evaluation metric.

In terms of performance, Samba-ASR outperforms Whisper-v3 and several other models on Gigaspeech, LS Clean, LS Other, and SPGISpeech, however the results are superficially reported and several training and architectural details are missing. This, however, is to be expected since Samba-ASR is in the end positioned as an enterprise, production-grade ASR system, remaining to this point a closed system.

5.8 Universal-1: Anatomy of Industrial Scale Multilingual ASR

Universal-1 [56] is an industrial-grade ASR system proposed by AssemblyAI[1]. The system is trained on an extensive publicly-acquired dataset, implying in the self-supervision stage a corpus of 12.5M hours of audio. For the refinement stage, a separate annotated dataset was purchased and curated

[1] https://www.assemblyai.com

in order to ensure accurate transcriptions, resulting in a corpus of 188k hours. Other pseudo-labeled datasets are also used during the supervision stage, however no details are given regarding the source of these datasets.

The system's architecture follows that of Conformer [14], with RNN-T for decoding text transcriptions, totaling 600M parameters. BERT-based Speech pre-Training with Random-projection Quantizer (BEST-RQ) [57] is used for pre-training the Conformer encoder module. More specific, speech features (e.g. log-mel filterbanks) are projected via a random matrix and assigned to the nearest vector in a randomly initialized codebook, generating discrete token indices. During pre-training, portions of the input are masked, and the model learns to predict the discrete indices of the masked frames, similar to BERT's masked token prediction. After pre-training, the encoder is fine-tuned on labeled speech data for ASR.

One key element is the usage of Sequential Transducer Loss during fine-tuning, replacing the original RNN-T loss which computes the probability of the target token sequence by summing over all possible alignments between encoder frame outputs and label sequences. This requires instantiating a full-dimensional lattice over timesteps and labels for each batch of audio data, resulting in significant memory usage. Their proposed sequential implementation mitigates this aspect by performing a step-by-step scanning over the encoder's time dimension, during both forward and backward passes. This enables training with larger batch sizes, while maintaining the same initial objective, hence increasing generalization power.

The inference pipeline is straightforward: segment the audio into shorter chunks, apply greedy decoding fo each chunk separately, and re-align transcriptions from each segment. The authors argue this scheme allows for efficient processing of long-form audios. Compared to Whisper [32], here individual audio chunks are processed in parallel to reduce latency, making each segment as long as possible without breaking it in the middle of utterances. This is performed by additionally including a VAD[2] module, setting a segment's end once VAD's detected activity drops under a certain threshold.

Evaluation performance indicates that Universal-1 outperforms several well-established systems, such as Whisper, on English and Multilingual ASR, while also being significantly more lightweight. Additionally, the model showed good generalization for code-switching experiments (i.e. the speaker switches between two or more languages in a conversation). The inference latency is significantly lower for Universal-1, quantified by the Real Time Factor (RTF). An extensive analysis regarding hallucinations is also performed, in order to identify *fabrication errors* – consecutive insertions or substitutions, and *omission errors* – consecutive deletions. When presented solely with ambient noise audio, Universal-1 resulted in the least percentage of fabrications generated in non-speech data, generating blank characters when non-speech audio was applied.

[2] https://webrtc.org

6 Audio-Visual Content Generation

6.1 Jukebox: A Generative Model for Music

Jukebox [58] is a large-scale generative framework introduced by OpenAI for producing music directly in the raw audio domain, including instrumental tracks and singing conditioned on lyrics. The central challenge addressed by the model is the extreme temporal length and high dimensionality of raw audio signals, which make conventional autoregressive modeling impractical for capturing long-range musical structure. To overcome this limitation, Jukebox combines hierarchical vector-quantized autoencoding with autoregressive Transformer-based priors, enabling coherent music generation over timescales of multiple minutes.

Unlike symbolic music generation approaches that operate on MIDI or note-level representations, Jukebox models audio waveforms sampled at 44.1 kHz. This design choice allows the system to capture fine-grained timbral details, expressive dynamics, and vocal characteristics, but requires substantial compression to make learning tractable. To this end, the authors introduce a three-level residual VQ-VAE architecture, where raw audio is encoded into discrete latent sequences at progressively coarser temporal resolutions. Each level employs its own encoder, codebook, and decoder, trained independently to prevent information collapse across scales.

The highest level of abstraction produces a heavily compressed representation that captures global musical properties such as song structure, harmonic progression, rhythm, and stylistic patterns. Two lower levels encode increasingly fine-grained details, refining the representation toward waveform-level fidelity. Reconstruction is performed by decoding the bottom-level codes, while the intermediate representations serve as conditioning signals for hierarchical generation. A spectral reconstruction loss is incorporated during VQ-VAE training to preserve mid- and high-frequency content that would otherwise be lost under purely time-domain objectives.

Music generation proceeds through a cascade of autoregressive priors. A top-level Transformer prior models the most abstract latent sequence and is responsible for establishing long-range coherence across tens of seconds. Two subsequent Transformer-based upsamplers generate the middle- and bottom-level codes, conditioned on the representations produced at higher levels. Sparse attention mechanisms are employed to keep computation feasible while maintaining long effective context lengths. Together, these components form a hierarchical probabilistic model over discrete audio representations.

Jukebox supports several forms of conditioning that allow controllable generation. Artist and genre embeddings are used to steer the model toward specific stylistic regions of the data distribution, improving both quality and diversity. Additional timing information, such as the relative position within a song, enables the model to learn structural regularities associated with beginnings, transitions, and endings. For lyric-conditioned generation, the system is trained on unaligned text, requiring the model to implicitly learn the temporal alignment between phoneme sequences and sung audio. An encoder–decoder attention mechanism allows the music model to attend to lyric representations, enabling intelligible and rhythmically plausible singing.

To generate music longer than the fixed context length of the priors, the authors employ windowed sampling with overlapping segments, ensuring continuity across generated sections. The model can also perform primed sampling, where an existing audio clip is encoded into VQ-VAE codes and used as an initial context for continuation. These mechanisms allow Jukebox to produce multi-minute musical pieces that maintain consistency in melody, harmony, instrumentation, and vocal style.

Evaluation of Jukebox primarily relies on qualitative analysis and human listening. The authors assess coherence, musicality, diversity, and novelty through extensive sampling and curated examples. Quantitative metrics are used mainly in ablation studies of the VQ-VAE component, focusing on reconstruction fidelity via spectral convergence. The results demonstrate that increasing model scale and hierarchical depth significantly improves perceptual quality, while lyric conditioning enables recognizable and controllable singing.

Overall, Jukebox represents a major step forward in generative music modeling by demonstrating that raw waveform synthesis with long-range structure and expressive vocals is feasible at scale. Its hierarchical combination of discrete audio representations and large autoregressive models establishes a foundation for future research in long-context audio generation, controllable music synthesis, and human–AI co-creation.

6.2 Simple and Controllable Music Generation

The paper [59] introduces *MUSICGEN*, a state-of-the-art framework for conditional music generation that emphasizes architectural simplicity, controllability, and high audio fidelity. The work addresses the task of text-to-music generation, with extensions to melody-conditioned generation and stereophonic audio synthesis. Motivated by the growing complexity of prior approaches—often relying on multi-stage pipelines or cascades of models—the authors investigate whether a single-stage autoregressive language model operating over discrete audio tokens can achieve competitive or superior performance while remaining computationally efficient and easier to analyze.

At the core of MUSICGEN lies an autoregressive Transformer decoder trained over compressed discrete audio representations obtained from a neural audio tokenizer based on residual vector quantization. This tokenizer converts raw waveforms into multiple parallel streams of discrete tokens, each associated with a distinct codebook. A central challenge addressed in the paper is how to model these parallel codebook streams efficiently. Rather than flattening all codebooks into a long sequence or using hierarchical upsampling as in earlier work, the authors propose a general framework of *codebook interleaving patterns*. These patterns define how tokens from different codebooks and time steps are ordered or grouped during autoregressive prediction, allowing the model to trade off exactness of the factorization against computational cost.

Several interleaving strategies are explored, including flattening, parallel prediction, coarse-first, partial delay, and delay-based patterns. Empirical results show that a simple *delay pattern*, which introduces offsets between codebook streams, achieves a favorable balance between audio quality and efficiency. Using this pattern, MUSICGEN reduces the number of autoregressive steps required for long audio sequences while preserving high perceptual quality. The same modeling framework naturally extends to stereo audio by treating left and right channels as additional codebook streams, enabling stereophonic generation without increasing inference cost.

MUSICGEN supports two main forms of conditioning. For text conditioning, the model relies on pretrained text encoders such as T5, with experiments also considering instruction-tuned and audio–text joint representations. Text descriptions are augmented with metadata such as genre, tempo, and tags, and classifier-free guidance is employed to control adherence to the conditioning signal. In addition to text, the paper introduces an unsupervised approach to melody conditioning based on chromagram representations. By extracting dominant pitch information from a reference audio track, the model can generate music that follows a given melodic structure while still allowing stylistic variation, without requiring supervised melody annotations.

The models are trained on approximately 20,000 hours of fully licensed music, combining internal datasets with licensed commercial music collections. Evaluation is conducted primarily on the MusicCaps benchmark, which provides expert-curated text–music pairs. MUSICGEN is assessed using a combination of objective and subjective metrics. Objective measures include Fréchet Audio Distance, KL divergence between label distributions of real and generated audio, and audio–text similarity scores based on joint embeddings. Subjective evaluation relies on large-scale human listening tests, reporting mean opinion scores for overall quality and text relevance. For melody-conditioned generation, the authors additionally introduce a chroma-based cosine similarity metric to quantify melodic alignment.

Experimental results demonstrate that MUSICGEN achieves strong performance relative to existing baselines, particularly in human evaluations of perceptual quality and text relevance. Ablation studies systematically analyze the impact of model size, codebook interleaving strategy, conditioning methods, and text preprocessing, revealing that moderate model scales offer the best trade-off between quality and efficiency. The paper also includes a memorization analysis, suggesting limited verbatim reproduction of training data under the evaluated settings.

The authors discuss several limitations of the proposed approach. Fine-grained control over the degree of adherence to text or melody conditioning remains limited and relies primarily on guidance mechanisms. Objective metrics do not always correlate well with human judgments at higher quality levels, and the chromagram-based melody metric captures harmonic structure but ignores timbral and rhythmic aspects. Additionally, despite licensing guarantees, the training data exhibit a bias toward Western musical styles, which may affect generalization.

Overall, MUSICGEN establishes a compelling baseline for controllable music generation with discrete audio language models. By demonstrating that a single-stage autoregressive Transformer with carefully designed token interleaving can achieve high-quality, controllable, and stereo-capable music generation, the work contributes both practical modeling insights and a unifying framework for future research in neural audio synthesis.

6.3 STEMGEN: A Music Generation Model That Listens

6.3.1 Core Problem & Motivation

Most existing music generation models (like MusicGen or MusicLM) focus on text-to-audio generation (e.g., "generate a jazz song"). These models often struggle to integrate into real-world music production workflows because they cannot effectively "listen" or react to existing musical elements. StemGen aims to bridge this gap by functioning as a virtual musician that can hear what is currently playing and improvise a matching part (e.g., adding a drum beat to a guitar riff).

6.3.2 Model Architecture

StemGen [60] utilizes a non-autoregressive, transformer-based architecture, similar to models like SoundStorm or VampNet. Key architectural features include:

- **Audio Tokenization:** It uses a pre-trained neural audio codec (Encodec) to convert audio into discrete tokens.
- **Concatenated Embeddings:** It combines multiple audio channels (context and target) into a single sequence for processing.
- **Novel Improvements:** The authors introduced two specific technical improvements to enhance performance:
  - Multi-source Classifier-Free Guidance: A technique to better control how strongly the model adheres to the input context versus the text conditioning.
  - Causal Bias during Iterative Decoding: A sampling improvement that helps generate more temporally consistent audio.

6.3.3 Training & Datasets

The model was trained on two primary datasets to ensure versatility:

- **Slakh2100:** An open-source dataset of synthesized MIDI audio, allowing for clean separation of instruments.
- **Internal Proprietary Dataset:** A larger, real-world dataset to improve the model's ability to handle realistic audio production scenarios.

6.3.4 Evaluation & Results

The researchers evaluated StemGen using both objective metrics and subjective listening tests:

- **Audio Quality:** Using Fréchet Audio Distance (FAD), the model demonstrated audio quality comparable to state-of-the-art text-conditioned models.
- **Musical Coherence:** They introduced a new metric based on Music Information Retrieval (MIR) descriptors (specifically MIRDD) to measure how well the generated stem aligns rhythmically and harmonically with the context.
- **Human Evaluation:** A Mean Opinion Score (MOS) test with musically trained participants confirmed that StemGen produces plausible and coherent musical additions.

6.3.5 Use Cases

The paper highlights several practical applications for StemGen:

- **Interactive Composition:** Users can build a track iteratively, adding one instrument at a time (e.g., start with chords, ask the AI for a bassline, then ask for drums).
- **Accompaniment Generation:** Providing a full mix or a single track and having the model generate a supporting instrument.
- **Live Performance:** The authors demonstrated a prototype "live" device where the model generates stems in real-time during a performance.

6.4 Music ControlNet: Multiple Time-Varying Controls for Music Generation

The paper introduces Music ControlNet [61], a diffusion-based music generation model designed to provide precise, time-varying controls over generated audio, addressing fundamental limitations in existing text-to-music generation systems. The primary contribution lies in enabling fine-grained, multi-faceted control over music generation beyond simple text prompts, specifically targeting dynamic musical attributes such as melody, dynamics, and rhythm that vary over time.

Current text-to-music generation models, while capable of producing high-quality audio, primarily offer control over global musical attributes like genre, mood, and tempo. These systems struggle with precise, time-varying control over attributes such as the exact positions of beats or the changing dynamics of music. The research hypothesizes that it is possible to adapt image-domain control mechanisms, specifically ControlNet, to music generation to enable precise, time-varying controls.

Music ControlNet builds upon the methodology of text-to-image generation with pixel-level controls, extending these principles to text-to-audio generation. The framework formulates the controllable audio generation task as learning a conditional generative model over audio waveforms given global text controls and a set of time-varying controls. Due to the high sampling rates of audio, the authors adopt a hierarchical approach using spectrograms as an intermediary representation. The architecture adapts the UNet structure from image diffusion models, but with critical modifications to account for the unique properties of audio. Unlike images where both dimensions are spatial, spectrograms have distinct semantic meanings—one representing time and the other frequency. To incorporate time-varying controls, an additional multi-layer perceptron transforms the control dimensions to match the frequency bins and simultaneously learns the relationship between control classes and frequency bins. A novel masking strategy enables partially-specified controls, allowing creators to specify controls for only a portion of the generation and enabling the model to improvise in the remaining segments.

The framework proposes three complementary time-varying control signals: melody, dynamics, and rhythm. The melody control is encoded using a variation of the chromagram, representing the most prominent musical tone over time. The dynamics control is derived from frame energy mapped to the decibel scale, characterizing loudness and correlating with musical intensity-related attributes. The rhythm control employs an RNN-based beat detector to predict whether a frame is situated on a beat, a downbeat, or neither. The framework supports composable controls, meaning it can generate music corresponding to any subset of the available controls.

The model was trained on approximately 1,800 hours of licensed instrumental music with genre and mood tags. The evaluation employed multiple objective metrics including melody accuracy, dynamics correlation, rhythm F1 scores, CLAP score for text control adherence, and FAD for audio realism. The evaluation of single versus multiple extracted controls demonstrated that time-varying controllability metrics are notably higher when corresponding controls are enforced. When melody control was applied, melody accuracy increased from 8.5% to 58.3%; when dynamics control was applied, dynamics correlation increased from near-zero to 88.8%; and when rhythm control was applied, rhythm F1 scores increased substantially. The model successfully learned to simultaneously respond to multiple controls, with time-varying controllability metrics remaining largely the same in multi-control scenarios.

The comparison between extracted and created controls revealed that time-varying controllability metrics actually improved when using created controls, demonstrating the model's generalizability to out-of-domain control inputs. The benchmark comparison against the 1.5 billion-parameter MusicGen revealed that Music ControlNet responds more precisely to melody control, particularly on created melodies, where the model was as much as 49% relatively more faithful to the control. Despite being much smaller (41 million parameters versus 1.5 billion) and trained on significantly less data, Music ControlNet demonstrated superior melody faithfulness while additionally accepting multiple controls and partially-specified spans.

Music ControlNet represents a significant advancement in controllable music generation by providing a framework for precise, multiple, time-varying controls that are composable, can be fully or partially specified, and generalize to creator-envisioned controls. The work overcomes fundamental limitations of previous text-to-music models by enabling fine-grained temporal control over musical elements, empowering creators with more expressive agency and better integration with existing creative workflows.

6.5 AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

AudioLDM 2 [62] introduces a unified framework for conditional audio generation that aims to bridge traditionally separated sub-domains such as sound effects, music, and speech synthesis. Motivated by the fragmentation of prior approaches—which rely on domain-specific inductive biases and representations—the authors propose a holistic perspective in which diverse audio generation tasks are addressed within a single architectural and learning paradigm. The central objective of AudioLDM 2 is to enable flexible, high-quality audio generation conditioned on multiple modalities, while leveraging large-scale self-supervised pretraining to reduce dependence on annotated data.

At the core of the proposed framework lies a continuous intermediate representation termed the *Language of Audio* (LOA). LOA is designed to capture both coarse-grained semantic information (e.g., sound identity or spoken content) and fine-grained acoustic characteristics in a form that is easier to model than raw waveforms. To construct this representation, AudioLDM 2 employs a self-supervised *Audio Masked Autoencoder* (AudioMAE), which extracts high-level features from log-mel spectrograms by reconstructing masked time–frequency patches. Unlike discrete token-based audio representations, LOA remains continuous, enabling richer expressivity and avoiding quantization artifacts that may limit generative fidelity.

The overall generation process is factorized into two stages. First, conditioning information—such as text descriptions, phoneme sequences, or audio embeddings—is translated into an estimated LOA sequence. This translation is formulated as an autoregressive sequence modeling problem and implemented using a GPT-2 Transformer. By modeling LOA sequences rather than raw audio tokens, the approach significantly reduces sequence length and mitigates error accumulation typically observed in long-horizon autoregressive audio models. Multiple conditioning encoders, including CLAP, FLAN-T5, and a phoneme encoder for speech tasks, are integrated through a unified embedding interface, allowing AudioLDM 2 to flexibly adapt to different generation scenarios.

In the second stage, AudioLDM 2 maps LOA representations to audio waveforms using a latent diffusion model (LDM). Audio signals are first compressed into a low-dimensional latent space via a variational autoencoder (VAE) operating on mel-spectrograms. The diffusion process is then learned in this latent space using a Transformer-UNet architecture with cross-attention mechanisms that condition generation on LOA sequences. Crucially, the latent diffusion model can be pretrained in a fully self-supervised manner by conditioning on ground-truth LOA extracted from audio, enabling large-scale training on unlabeled audio corpora.

The framework supports classifier-free guidance to control the trade-off between audio quality and adherence to conditioning signals during sampling. Additionally, a joint finetuning strategy is introduced, in which the GPT-2 LOA predictor and the latent diffusion model are optimized together. A probabilistic switcher alternates between ground-truth and predicted LOA during training, encouraging robustness to prediction errors at inference time and improving overall generation quality.

AudioLDM 2 is evaluated on three major tasks: text-to-audio, text-to-music, and text-to-speech. Objective metrics such as Frechet Audio Distance (FAD), KL divergence, and CLAP score are complemented by extensive subjective evaluations, including Overall Impression (OVL), Audio–Text Relevance (REL), and Mean Opinion Score (MOS) for speech. Experimental results demonstrate that AudioLDM 2 achieves state-of-the-art or competitive performance across all tasks, substantially outperforming prior text-to-audio systems in both perceptual quality and audio–text alignment, while also generating intelligible and natural speech without task-specific architectural modifications.

Overall, AudioLDM 2 advances the field of audio generation by demonstrating that a single, self-supervised, diffusion-based architecture can effectively unify audio, music, and speech synthesis. By introducing LOA as a general-purpose audio representation and combining autoregressive modeling with latent diffusion, the framework provides a scalable and versatile foundation for future research in multimodal audio generation, generalist audio models, and unified audio understanding and synthesis.

6.6 AudioCaps: Generating Captions for Audios in the Wild

AudioCaps [5] addresses the problem of *audio captioning*, defined as the automatic generation of natural language descriptions for arbitrary sounds occurring in real-world environments. While

prior research in multimedia captioning has largely focused on visual modalities—images and videos—the auditory domain has remained comparatively underexplored, with most efforts concentrated on speech recognition, sound event classification, or detection. The authors argue that such label-based or event-centric formulations are insufficient for capturing the richness of auditory scenes, as they fail to express properties such as temporal ordering, co-occurrence, intensity, or interactions between sound sources. Audio captioning is proposed as a more expressive alternative that enables a deeper semantic interpretation of sound.

The primary contribution of the work is the introduction of the *AudioCaps* dataset, the first large-scale benchmark specifically designed for captioning audios in the wild. AudioCaps consists of approximately 46,000 pairs of 10-second audio clips and human-written captions, derived from carefully curated subsets of the AudioSet collection. To ensure that captions are grounded in auditory perception rather than visual cues, the dataset construction process emphasizes sound-only descriptions and explicitly discourages visual speculation. Music-related categories are excluded due to their ambiguity for non-expert listeners, and categories requiring specialized knowledge or visual confirmation are filtered out. The final dataset is split into training, validation, and test sets and spans a wide range of everyday sounds, including human, animal, mechanical, and environmental audio events.

Beyond dataset construction, the authors investigate modeling strategies for audio captioning and analyze which audio representations and architectural components are most effective for the task. The proposed baseline architecture follows an encoder–decoder paradigm with an LSTM-based decoder, augmented by attention mechanisms. Audio inputs are represented using multiple levels of features, including MFCCs and pretrained embeddings such as VGGish and SoundNet. Experimental comparisons show that features pretrained on large-scale audio datasets—particularly VGGish, which is trained on AudioSet—provide substantial gains over raw waveform or handcrafted features, highlighting the importance of domain-aligned pretraining for audio-language tasks.

To further improve captioning performance, the paper introduces two architectural components that are designed to better capture the temporal and semantic structure of natural sounds. The first is a *top-down multi-scale encoder*, which jointly exploits mid-level and high-level audio representations by injecting global semantic context into temporally localized features. This design reflects the observation that real-world sounds are often sparse, transient, and non-stationary, requiring both fine-grained temporal sensitivity and global contextual awareness. The second contribution is *aligned semantic attention*, a mechanism that enforces consistency between temporal attention over audio features and semantic attention over attribute words derived from weak labels. By aligning these two forms of attention through an attention-flow mechanism, the model reduces mismatches between what the decoder attends to acoustically and semantically during caption generation.

The AudioCaps benchmark is evaluated using standard captioning metrics drawn from the natural language processing and vision-language literature, including BLEU, METEOR, CIDEr, ROUGE-L, and SPICE. Through extensive experiments, the authors demonstrate that audio-based captioning models perform substantially better on AudioCaps than video-based models, confirming that the collected captions are indeed faithful to auditory information rather than visual content. Among the tested approaches, models that combine pretrained VGGish features with temporal attention, multi-scale encoding, and aligned semantic attention achieve the best overall performance.

In summary, AudioCaps establishes audio captioning as a distinct and well-defined research problem at the intersection of audio understanding and natural language generation. By contributing a carefully constructed large-scale dataset and empirically validated modeling components, the work provides a foundational benchmark for future research in audio–language learning. The study also highlights the limitations of label-centric audio understanding and motivates richer, sentence-level representations as a path toward more comprehensive semantic modeling of sound in real-world environments.

6.7 MusicLM: Generating Music from Text

MusicLM [63] tackles the challenging problem of *text-to-music generation*, aiming to synthesize high-fidelity, long-form musical audio directly from natural language descriptions such as "a calming violin melody backed by a distorted guitar riff." Unlike earlier approaches that either relied on symbolic representations (e.g., MIDI) or were limited to short, acoustically simple outputs,

MusicLM targets realistic music generation at the waveform level, with coherence over several minutes and strong semantic alignment to rich textual prompts. The work positions itself at the intersection of conditional audio generation, large-scale self-supervised learning, and multimodal representation learning.

The core difficulty addressed by MusicLM lies in the inherent mismatch between text and music. Musical signals are highly structured over time, involve multiple interacting components (melody, rhythm, harmony, timbre), and often encode semantics that are difficult to describe exhaustively in language. Moreover, large-scale paired music–text datasets are scarce and expensive to construct, especially when compared to image–text corpora. To overcome these challenges, MusicLM builds upon the AudioLM framework, extending it with text conditioning through a joint music–text embedding space, thereby decoupling music generation from the need for explicit captions during training.

At a high level, MusicLM formulates music generation as a *hierarchical sequence-to-sequence modeling problem* over discrete audio tokens. Audio is first transformed into multiple levels of discrete representations using pretrained and frozen models. Acoustic details are captured using SoundStream, a neural audio codec based on residual vector quantization (RVQ), which represents 24 kHz audio at a bitrate of approximately 6 kbps. Long-term musical structure is modeled using semantic tokens derived from a pretrained w2v-BERT model, which captures higher-level temporal patterns through self-supervised masked language modeling on audio. These two token streams allow the system to separately handle global musical coherence and fine-grained audio fidelity.

Text conditioning is introduced via MuLan, a joint music–text embedding model trained with contrastive learning to map music clips and their corresponding textual descriptions into a shared semantic space. Crucially, MusicLM leverages MuLan embeddings in a novel way: during training, MuLan embeddings are computed from the audio itself, enabling training on massive audio-only corpora without requiring paired text annotations. During inference, these audio-derived embeddings are replaced by embeddings computed from a user-provided text prompt. This design choice allows MusicLM to scale to hundreds of thousands of hours of unlabeled music while still supporting text-conditioned generation at inference time.

The generation process itself is organized hierarchically. In the first stage, a semantic Transformer autoregressively predicts semantic tokens conditioned on the MuLan embedding, capturing high-level musical structure such as genre, tempo, and overall style. In the second stage, an acoustic Transformer predicts SoundStream tokens conditioned on both the semantic tokens and the MuLan embedding. Following AudioLM, the acoustic stage is further divided into coarse and fine sub-stages, which improves efficiency and stability while preserving audio quality. The final waveform is reconstructed using the SoundStream decoder. This hierarchical design is central to MusicLM's ability to generate music that is both temporally coherent and acoustically realistic over long durations.

To evaluate text-to-music generation, the authors introduce *MusicCaps*, the first dataset specifically designed for this task. MusicCaps consists of 5.5k music clips drawn from AudioSet, each paired with detailed, expert-written captions averaging multiple sentences. Unlike AudioCaps, which focuses on general environmental sounds, MusicCaps is exclusively music-focused and includes rich descriptions of genre, instrumentation, mood, rhythm, and vocal characteristics. The dataset also provides structured aspect annotations and a genre-balanced subset to support robust evaluation. MusicCaps thus fills a critical gap in benchmarking for music–language research.

Evaluation of MusicLM combines objective metrics, embedding-based measures, and human listening studies. Audio quality is assessed using the Fréchet Audio Distance (FAD), computed with both speech-oriented (TRILL) and music-oriented (VGGish) embeddings. Text adherence is evaluated using Kullback–Leibler divergence between AudioSet classifier outputs, as well as MuLan Cycle Consistency (MCC), which measures cosine similarity between embeddings of generated music and the conditioning text. In addition, large-scale human preference tests compare MusicLM against baselines such as Mubert and Riffusion. Across these metrics, MusicLM consistently outperforms prior systems, achieving lower FAD scores and substantially higher alignment with textual descriptions, while still remaining below the upper bound established by ground-truth MusicCaps audio.

Beyond basic text conditioning, the paper explores several important extensions. One notable contribution is *melody conditioning*, where MusicLM is augmented to accept an additional audio input representing a melody (e.g., humming or whistling). By learning a melody-invariant embedding and concatenating it with the text-derived MuLan tokens, the system can generate music that

follows a specified melodic contour while rendering it in the style described by the text prompt. Another extension enables long-form generation and "story mode," where the text conditioning signal is changed over time to produce smooth stylistic transitions within a single musical piece, demonstrating the flexibility of the autoregressive framework.

The authors also conduct a careful analysis of memorization risks, adapting techniques from the large language model literature to the audio domain. By comparing generated semantic token sequences with those in the training data, they find that exact memorization is rare and approximate matches are limited, suggesting that MusicLM primarily recombines learned musical patterns rather than reproducing training examples verbatim. Nonetheless, the paper explicitly acknowledges ethical and societal concerns, including cultural bias, potential misappropriation of creative content, and the broader implications of automated music generation.

In summary, MusicLM represents a significant advance in text-conditioned music generation. Its key innovations include hierarchical autoregressive modeling over discrete audio tokens, the use of joint music–text embeddings to eliminate the need for paired data during training, and the introduction of MusicCaps as a high-quality evaluation benchmark. By demonstrating long-duration, high-fidelity music generation that remains faithful to complex natural language descriptions, MusicLM establishes a strong foundation for future research in multimodal generative audio systems and creative AI applications.

6.8 NotaGen: Advancing Musicality in Symbolic Music Generation

NotaGen [64] addresses the problem of *symbolic sheet music generation*, with a particular focus on enhancing musicality, structural coherence, and controllability in the context of classical music composition. Unlike audio-based music generation systems that operate directly on waveforms, NotaGen targets score-level representations, aiming to produce high-quality, human-readable sheet music suitable for analysis, performance, and further musicological processing. The work explicitly draws inspiration from large language models (LLMs) and adapts their training paradigms—pre-training, fine-tuning, and reinforcement learning—to the domain of symbolic music.

The core challenge tackled by NotaGen lies in the gap between next-token prediction accuracy and perceived musical quality. While Transformer-based models trained on symbolic representations have shown promise, they often struggle to capture higher-level musical structure, stylistic consistency, and long-range dependencies, particularly in multi-voice classical compositions. Furthermore, high-quality annotated sheet music datasets are relatively scarce compared to large-scale audio or text corpora, limiting the effectiveness of purely supervised training approaches. NotaGen positions itself as a response to these limitations by systematically transferring LLM-style training strategies to symbolic music generation.

At the representation level, NotaGen adopts *interleaved ABC notation*, a text-based sheet music format that encodes multiple voices within aligned bars. To balance musical fidelity and computational efficiency, the model employs *bar-stream patching*, which segments tune headers and musical bars into fixed-length patches. This approach preserves bar-level structural information while enabling efficient long-context modeling. Additional preprocessing steps, such as the removal of full-rest bars and explicit bar index annotations, further improve information density and support long-form generation.

Architecturally, NotaGen is a hierarchical Transformer model based on GPT-style decoders. It consists of a patch-level decoder that models temporal dependencies across bar-stream patches and a character-level decoder that autoregressively generates the symbolic content within each patch. This two-level design allows the model to separately capture long-range musical structure and fine-grained notation details. The architecture builds upon the Tunesformer framework while extending it to handle richer classical sheet music scenarios.

Training follows a three-stage paradigm inspired by LLMs. In the pre-training stage, NotaGen is trained on a large internal corpus of approximately 1.6 million ABC notation sheets, spanning diverse genres and instrumentations. Data augmentation is performed by transposing each piece into multiple musical keys, enabling the model to learn key-invariant musical patterns. Pre-training focuses on next-token prediction and equips the model with general musical knowledge.

The second stage, fine-tuning, targets musicality and stylistic control. The authors curate a high-quality classical sheet music dataset comprising 8,948 works aggregated from several public corpora (including DCML, OpenScore, ATEPP, and KernScores) as well as internal sources. Each piece is labeled with *period*, *composer*, and *instrumentation*, and these labels are prepended as

prompts during training. This conditioning mechanism enables controlled generation across historical styles and ensemble types, ranging from Baroque keyboard works to Romantic orchestral compositions.

To further enhance musical quality and prompt alignment, NotaGen introduces a reinforcement learning stage based on *CLaMP-DPO*, a Direct Preference Optimization (DPO) method that relies on AI-generated feedback rather than human annotations. The approach leverages *CLaMP 2*, a multimodal symbolic music retrieval model, to evaluate semantic similarity between generated outputs and reference compositions. Generated samples are ranked according to their CLaMP 2 scores, forming preference pairs that drive optimization. This reinforcement learning from AI feedback (RLAIF) framework allows NotaGen to improve musicality and controllability without the high cost of human preference labeling.

Evaluation combines objective metrics and human studies. Objective measures include the Average CLaMP 2 Score (ACS) for semantic alignment, Label Accuracy (LA) for prompt controllability, Bar Alignment Error (BAE) for structural correctness in sheet music, and language-model Perplexity (PPL). In addition, extensive human A/B preference tests involving trained musicians assess perceived musicality across multiple dimensions such as melodic appeal, harmonic fluency, counterpoint, and notation quality. Results show consistent improvements after reinforcement learning and demonstrate that NotaGen outperforms baseline symbolic music models, although it still falls short of human-composed ground truth.

The authors also discuss limitations and challenges. CLaMP-based evaluation assumes syntactically valid outputs and may be unreliable for severely corrupted generations. Moreover, while NotaGen performs well on small to medium ensembles, orchestral music remains more difficult to model due to its complexity and scale. These observations highlight open challenges in symbolic music generation and motivate future work on broader genres and more expressive representations.

In summary, NotaGen represents a significant step toward transferring LLM training paradigms to symbolic music generation. Its key contributions include a hierarchical Transformer architecture for sheet music, large-scale pre-training on symbolic data, prompt-based fine-tuning for stylistic control, and reinforcement learning via AI feedback to enhance musicality. By demonstrating measurable gains in both objective metrics and human preference evaluations, NotaGen establishes a compelling framework for future research on high-quality, controllable symbolic music generation.

6.9 Stable Audio Open: Open-Weights Text-to-Audio Generation with Latent Diffusion

Stable Audio Open [65] addresses the problem of *text-conditioned audio generation* with a particular emphasis on openness, data transparency, and high-fidelity sound synthesis. In contrast to many recent text-to-audio systems that are accessible only through proprietary APIs or closed weights, Stable Audio Open is released with publicly available model weights and code, and is trained exclusively on Creative Commons licensed audio. The work targets non-speech audio generation, including environmental sounds and instrumental music, and aims to provide a competitive open baseline for research and artistic applications.

The primary challenge tackled by Stable Audio Open lies in reconciling three often competing objectives: (i) state-of-the-art audio quality at high sampling rates, (ii) long-form and variable-length generation, and (iii) full transparency with respect to training data licensing. While recent closed models have demonstrated impressive results in text-to-audio synthesis, they typically rely on undisclosed datasets and restrict downstream reuse. Conversely, existing open models often lag behind in audio quality, coherence over long durations, or inference efficiency. Stable Audio Open positions itself as a response to this gap by combining a modern latent diffusion architecture with carefully curated Creative Commons data.

At the architectural level, Stable Audio Open is a *latent diffusion model* operating on compressed audio representations. The system consists of three main components: a variational autoencoder that maps raw audio waveforms to a continuous latent space, a pretrained text encoder for conditioning, and a diffusion transformer (DiT) that performs generative modeling in the latent domain. The autoencoder operates directly on stereo waveforms sampled at 44.1 kHz and compresses them into a low-rate continuous latent representation with 64 channels at 21.5 Hz. This compression enables efficient modeling of long audio sequences while preserving perceptual fidelity.

The generative core of the system is a transformer-based diffusion model inspired by recent advances in DiT architectures. The DiT predicts noise increments in the latent space using the

v-objective formulation of diffusion training. Conditioning is provided through multiple signals, including text embeddings, temporal information for variable-length generation, and diffusion timestep embeddings. Text conditioning is implemented via a pretrained T5-base encoder, while the audio autoencoder and diffusion transformer are trained from scratch. Efficient attention mechanisms and gradient checkpointing are employed to reduce memory consumption and enable inference on consumer-grade GPUs.

A key design feature of Stable Audio Open is its support for *variable-length audio generation*. The model is trained with a fixed maximum window (up to approximately 47 seconds) and learns to fill unused portions of the window with silence. At inference time, generated audio can be trimmed to the desired length, allowing flexible output durations without retraining. This approach builds on prior work in timing-conditioned diffusion and is shown to be effective for both sound and music generation tasks.

Training data selection and governance play a central role in the contribution of Stable Audio Open. The authors construct a large-scale dataset comprising approximately 486,000 audio recordings (around 7,300 hours) drawn from Freesound and the Free Music Archive (FMA), restricted to Creative Commons licenses (CC0, CC-BY, and CC-Sampling+). Extensive filtering procedures are applied to remove copyrighted material, including automatic music detection, external copyright checks, and metadata-based screening. This process results in a dataset that prioritizes transparency and reusability, albeit at the cost of reduced coverage of commercial music styles.

Training proceeds in two main stages corresponding to the autoencoder and the diffusion transformer. The autoencoder is trained using a combination of reconstruction losses based on multi-resolution STFT, adversarial losses with multiple discriminators, and a lightly weighted KL divergence term. The diffusion transformer is then trained to predict noise increments in the latent space using AdamW optimization and large-scale GPU resources. Inference employs DPM-Solver++ sampling with classifier-free guidance, balancing generation quality and computational efficiency.

Evaluation focuses on both generative quality and reconstruction fidelity. For generative modeling, the authors employ established objective metrics including Fréchet Distance computed on OpenL3 embeddings (FDopenl3), KL divergence on semantic audio embeddings (KLpasst), and CLAPScore to assess text-audio alignment. Experiments are conducted on the AudioCaps dataset for sound generation and the Song Describer Dataset for instrumental music generation. Results show that Stable Audio Open outperforms comparable open baselines on sound generation metrics, particularly FDopenl3, while remaining less competitive than closed models for music generation.

In addition to generative evaluation, the autoencoder component is evaluated independently using reconstruction metrics such as STFT distance, MEL distance, and scale-invariant signal-to-distortion ratio (SI-SDR). Comparisons against established neural audio codecs demonstrate that the autoencoder achieves performance comparable to prior Stable Audio models despite being trained solely on Creative Commons data. Memorization analyses further indicate that the model does not reproduce training examples verbatim, addressing concerns related to data leakage and responsible model development.

The authors also discuss several limitations. Stable Audio Open is not designed for intelligible speech or vocal generation and struggles with prompts involving complex linguistic connectors or multiple simultaneous events. Music generation quality is constrained by the limited availability of high-quality Creative Commons music, and the model performs best with English-language prompts. These limitations reflect broader challenges in open audio modeling and highlight trade-offs between openness, data availability, and generative performance.

In summary, Stable Audio Open represents a significant contribution to open text-to-audio generation. Its key contributions include a high-fidelity latent diffusion architecture for stereo audio at 44.1 kHz, transparent training on Creative Commons data, competitive performance on sound generation benchmarks, and full public release of model weights and code. By emphasizing data governance and reproducibility alongside technical performance, Stable Audio Open establishes a strong open baseline and provides a foundation for future research on accountable and accessible generative audio systems.

6.10 Multimodal Representation Alignment for Image Generation: Text-Image Interleaved Control Is Easier Than You Think

The authors proposed DreamEngine [66], a multimodal framework that generates images based on a text-to-image diffusion model. This paper tackles the problem of controlling the predicted images with certain conditions, by merging visual elements or various concepts from different images during the generation process. With DreamEngine, they show that large multimodal models (LMMs) provide a shared representation space which can be used to allign text-image pairs with the purpose of coditioning other diffusion models.

DreamEngine is build upon Stable Diffursion 3.5 [67] text-to-image model, but with multiple modifications. First, they replaced the encoders with QwenVL [68] to take into account multimodal information. Second, their training paradigm is a two-stage approach, represented by joint text-image and interleaved instruction tuning with multimodal information. In Figure 17, DreamEngine framework is represented, capable of adapting LLMs as text-encoder, without changing their parameters, for diffusion models solving text-to-image tasks. Using contributions from both images and text by combining object detection with captioning, the image output is highly customizable.

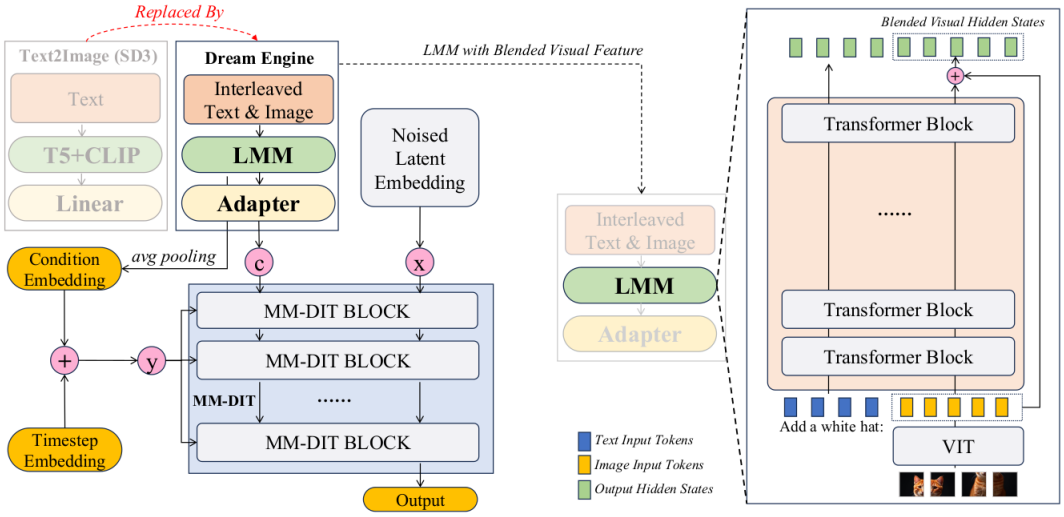


Figure 17: DreamEngine [66] architecture

DreamEngine has 2 fundamental modules: a Large Multimodal Model (LMM) and a Multimodal Diffusion Transformer (MM-DiT). The first one is composed of a Visual Transformer (ViT) encoder, as well as a large language model (LLM), followed by an alignment layer that correlates both text-image representation spaces. The second module (MM-DiT) is build upon Latent Diffusion Model (LDM) [69] and Diffusion Transformer (DiT) [70]. Its purpose it to combine textual information with noised latent embeddings and obtain an unified conditioned sequence. An adapter layer represented by a two-layer Multi-Layer-Perceptron (MLP) was added to align and adapt the LMM and MM-DiT modules to the desired input. Notably, DreamEngine can receive any sequence length as the text-image input.

Since DreamEngine relies on 2 pretrained components (LMM and MM-DiT), the authors structured the training process into 2 stages, unfreezing specific components to ensure an effective training. In stage 1, only the two-layer MLP adapter layer was trained, to align the representation spaces betweem QwenVL and SD3.5. For stage 1, the authors used public image-caption dataset from CC12M [71] and JourneyDB [72]. In stage 2, MM-DiT module as well as the adapter layer were trained to have a better control over the process of image generation. In this last stage, the framework was trained for 2 task: *(i)* Free-Form Image Editing using the UltraEdit [73] dataset and *(ii)* Objects Driven Generation, using COCO [74] dataset.

6.11 FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space

The paper introduces FLUX.1 Kontext, a unified generative model that seamlessly integrates in-context image generation and editing within a single architecture through flow matching in latent space. The core contribution lies in its ability to handle both local and global editing tasks while generating novel views by incorporating semantic context from text and reference images, addressing critical limitations in character consistency, inference speed, and iterative editing stability that plague existing approaches.

The research addresses three fundamental shortcomings of contemporary image editing methods: instruction-based approaches trained on synthetic pairs inherit limitations from their generation pipelines, restricting edit variety and realism; maintaining accurate appearance of characters and objects across multiple edits remains unsolved, hindering storytelling and brand-sensitive applications; and autoregressive editing models integrated into large multimodal systems suffer from prohibitively long runtimes incompatible with interactive use. The authors formulate their objective as learning to approximate the conditional distribution $p(x|y, c)$, where x represents the target image, y denotes a context image (or $\emptyset$), and c is a natural-language instruction, enabling the network to perform image-driven edits when context exists and create new content from scratch otherwise.

The technical approach builds upon a rectified flow transformer trained in the latent space of an image autoencoder, employing a simple yet effective sequence concatenation method to incorporate semantic context. Images are encoded into latent tokens by the frozen FLUX autoencoder, with context image tokens appended to target image tokens and fed into the visual stream. This design choice supports different input/output resolutions and aspect ratios while readily extending to multiple images, outperforming channel-wise concatenation in initial experiments. Positional information is encoded via three-dimensional Rotary Positional Embeddings (3D RoPE), where context tokens receive a constant offset treated as a virtual time step that cleanly separates context and target blocks while preserving their internal spatial structure. The model is trained using a rectified flow-matching loss

$$\mathcal{L}_\theta = \mathbb{E}_{t\sim p(t),x,y,c}\left[\|v_\theta(z_t, t, y, c) - (\epsilon - x)\|_2^2\right] \tag{1}$$

where z_t represents the linearly interpolated latent between target image x and noise $\epsilon \sim \mathcal{N}(0, 1)$. A logit normal shift schedule is employed for the time distribution, with the mode adjusted depending on data resolution during training.

To address sampling efficiency and quality, the authors implement Latent Adversarial Diffusion Distillation (LADD), which reduces the number of sampling steps while increasing sample quality through adversarial training. This innovation tackles the computational expense of multi-step sampling that typically requires 50-250 guided network evaluations and mitigates potential visual artifacts such as over-saturation introduced by guidance. The architecture itself comprises a mix of double stream and single stream blocks, where double stream blocks employ separate weights for image and text tokens with mixing performed through attention over concatenated tokens, followed by 38 single stream blocks applied to the concatenated sequence before discarding text tokens and decoding image tokens. The model leverages fused feed-forward blocks that reduce modulation parameters by a factor of two and fuse attention input-output linear layers with the MLP, leading to more efficient training and inference.

The implementation begins from a pure text-to-image checkpoint, jointly fine-tuning the model on image-to-image and text-to-image tasks. While the formulation naturally covers multiple input images, the current implementation focuses on single context images for conditioning. FLUX.1 Kontext [pro] is trained with the flow objective followed by LADD, whereas FLUX.1 Kontext [dev] is obtained through guidance-distillation into a 12B diffusion transformer, with exclusive focus on image-to-image training to optimize edit task performance. Safety training measures including classifier-based filtering and adversarial training are incorporated to prevent generation of non-consensual intimate imagery and child sexual abuse material.

The model supports several specialized applications beyond standard generation, including style reference (SREF) for style transfer from reference images while maintaining semantic control, intuitive editing through visual cues such as geometric markers, and sophisticated text editing capabilities including logo refinement and spelling corrections. Character reference (CREF) enables consistent generation of specific characters or objects across novel settings, a critical capability for

brand-sensitive applications.

For evaluation, the authors introduce KontextBench, a comprehensive benchmark comprising 1026 unique image-prompt pairs derived from 108 base images including personal photos, CC-licensed art, public domain images, and AI-generated content. The benchmark spans five core tasks: local instruction editing (416 examples), global instruction editing (262), text editing (92), style reference (63), and character reference (193), addressing gaps in existing benchmarks that rely on synthetic data or lack comprehensive coverage of real-world applications. The evaluation framework employs both objective metrics and subjective human evaluations across multiple dimensions, with quantitative assessment for character reference using AuraFace embeddings to measure facial characteristic preservation. For text-to-image evaluation, the authors decompose assessment into five distinct dimensions: prompt following, aesthetics, realism, typography accuracy, and inference speed, addressing limitations of broad preference evaluations that favor a characteristic "AI aesthetic."

The experimental results demonstrate that FLUX.1 Kontext achieves competitive performance with state-of-the-art systems while delivering significantly faster generation times of 3-5 seconds for 1024×1024 images, outperforming related models by up to an order of magnitude in speed. In image-to-image tasks, FLUX.1 Kontext [max] and [pro] achieve top scores in local and text editing categories and for general character reference, with quantitative CREF scores outperforming all other models. For global editing and style reference, the model ranks second only to GPT-Image-1 and Gen-4 References, respectively. In text-to-image synthesis, FLUX.1 Kontext demonstrates balanced performance across evaluation categories, with consistent improvement over its predecessor FLUX1.1 [pro] and progressive gains from [pro] to [max] versions. Critically, the model exhibits significantly improved preservation of objects and characters across multiple editing turns, with cosine similarity measurements of AuraFace embeddings demonstrating slower character identity drift compared to competing methods.

The authors acknowledge several limitations in the current implementation. Excessive multi-turn editing can introduce visual artifacts that degrade image quality, and the model occasionally fails to follow instructions accurately, ignoring specific prompt requirements. Additionally, the distillation process can introduce visual artifacts that impact output fidelity. Future work should focus on extending to multiple image inputs, further scaling, reducing inference latency for real-time applications, and most importantly, reducing degradation during multi-turn editing to enable infinitely fluid content creation. A natural extension involves incorporating edits in the video domain.

FLUX.1 Kontext represents a significant advancement in unified image processing models by addressing key limitations of prior work while achieving state-of-the-art performance in character consistency and editing capabilities. The model's ability to deliver interactive speeds while maintaining superior multi-turn consistency makes it particularly valuable for applications requiring stable brand characters, asset continuity in media production, and preservation of product details in e-commerce. The introduction of KontextBench provides a comprehensive evaluation framework for future research in in-context image generation and editing, establishing new standards for the field. By successfully unifying generation and editing tasks within a single architecture while overcoming the speed-quality trade-off that limited previous approaches, FLUX.1 Kontext enables practical interactive workflows and rapid prototyping that were previously infeasible.

6.12 SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers

The paper [75] introduces *SANA*, an efficient text-to-image generation framework designed to synthesize high-quality images at resolutions up to 4096×4096 while significantly reducing computational cost and inference latency. The central contribution of SANA lies in demonstrating that state-of-the-art image quality and text–image alignment can be achieved without scaling model size to tens of billions of parameters. Instead, the authors propose a carefully co-designed pipeline that combines aggressive latent-space compression, linearized diffusion transformers, and modern decoder-only language models for text encoding.

As illustrated in Figure 18, SANA integrates aggressive latent compression, linearized diffusion transformers, and a decoder-only LLM-based text encoder into a unified and highly efficient text-to-image generation pipeline.

The work [67] is motivated by three key limitations observed in contemporary high-resolution

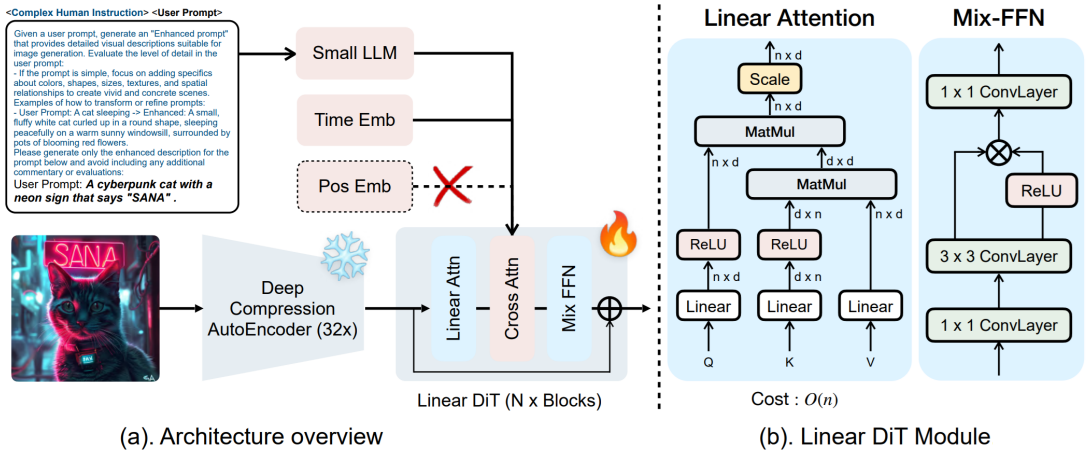


Figure 18: Architecture overview of SANA. The framework combines a deep compression autoencoder (32×), a Linear Diffusion Transformer (Linear DiT) with linear attention and Mix-FFN blocks, and a decoder-only LLM for text encoding with complex human instructions. Positional embeddings are omitted without loss of performance.

diffusion models. First, existing approaches rely on quadratic self-attention, resulting in prohibitive memory and computation costs as image resolution increases. Second, most large-scale models pursue performance gains primarily through parameter scaling, which restricts deployment to high-end cloud infrastructure. Third, traditional text encoders such as CLIP or T5 exhibit limited instruction-following and reasoning capabilities, constraining text–image alignment, especially for complex prompts. SANA addresses these challenges by jointly optimizing architectural design, training strategy, and inference procedures.

At the core of the proposed system is a *deep compression autoencoder* that compresses images by a factor of 32×, substantially reducing the number of latent tokens processed by the diffusion model. Unlike prior latent diffusion systems that rely on moderate compression and larger patch sizes, SANA assigns the full responsibility of compression to the autoencoder and performs diffusion with a patch size of one. This design reduces token counts by up to 16× compared to standard configurations, enabling efficient training and inference for ultra-high-resolution images without a significant loss in reconstruction fidelity.

Building on this compressed latent space, the authors introduce a *Linear Diffusion Transformer (Linear DiT)* that replaces all quadratic self-attention operations with ReLU-based linear attention, reducing computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. To compensate for the reduced expressiveness of linear attention, the model incorporates a *Mix-FFN* module that augments feed-forward layers with depth-wise convolutions, improving local information aggregation. Notably, the architecture eliminates positional embeddings entirely, relying on convolutional inductive biases to encode spatial information without degrading generation quality.

For text conditioning, SANA departs from encoder-based language models and adopts a *decoder-only small LLM* (Gemma-2) as the text encoder. To stabilize training and exploit the instruction-following capabilities of modern LLMs, the authors introduce *Complex Human Instructions (CHI)* that guide prompt enhancement through in-context learning. This approach improves text–image alignment and prompt robustness, particularly for short or underspecified user inputs. Empirical ablations demonstrate that CHI consistently improves alignment metrics with minimal additional training cost.

The training and inference pipeline further incorporates a *flow-based formulation* inspired by Rectified Flow and EDM models. The authors propose a customized *Flow-DPM-Solver* that reduces the number of sampling steps required for convergence, achieving high-quality results with as few as 14–20 steps. Combined with cascade resolution training and CLIP-score-based caption sampling, this strategy accelerates convergence while preserving semantic fidelity.

SANA is evaluated using a comprehensive set of established benchmarks and metrics, including FID and CLIP Score on the MJHQ-30K dataset, as well as GenEval, DPG-Bench, and ImageReward for assessing text–image alignment and human preference. Experimental results show that

SANA-0.6B and SANA-1.6B achieve performance competitive with or superior to much larger models such as FLUX and SD3, while delivering speedups exceeding 100× for 4K image generation and enabling deployment on consumer-grade GPUs.

The authors also demonstrate on-device deployment through 8-bit quantization, achieving subsecond generation times for 1024×1024 images on laptop GPUs with minimal quality degradation. Despite these advances, the paper acknowledges limitations related to safety guarantees, controllability, and challenging visual cases such as faces, hands, and complex text rendering.

Overall, SANA presents a compelling case for efficiency-driven design in text-to-image diffusion models. By co-optimizing compression, attention mechanisms, text encoding, and flow-based sampling, the framework establishes a practical alternative to large-scale diffusion systems and highlights a viable path toward accessible, high-resolution generative models without extreme parameter scaling.

6.13 Scaling Rectified Flow Transformers for High-Resolution Image Synthesis

The research presented in this paper introduces an advanced framework for high-resolution text-to-image synthesis by scaling rectified flow models using a novel transformer-based architecture. The core contribution is twofold: the improvement of noise sampling techniques for training rectified flows by biasing them towards perceptually relevant scales, and the development of a scalable, multimodal transformer backbone (MM-DiT) that enables bidirectional information flow between text and image modalities.

Diffusion models have become the standard for generating high-dimensional perceptual data, but they often suffer from computational inefficiency due to curved forward paths that require numerous integration steps during sampling. While Rectified Flow (RF) formulations offer theoretical advantages by connecting data and noise via straight lines—potentially allowing for single-step generation—they had not yet been decisively established as superior in large-scale practical applications. The authors identified that standard uniform noise sampling in RF fails to prioritize the intermediate timesteps where the learning of the velocity field is most challenging. Furthermore, existing architectures often treat text as a fixed condition (e.g., via cross-attention), which limits the model's ability to achieve deep multimodal integration and precise prompt following.

The methodology centers on the simulation-free training of flows by regressing a vector field that generates a probability path between noise and data distributions. The authors parameterized the velocity of the flow through a neural network, utilizing a straight-path formulation.

A key technical innovation is the introduction of tailored Signal-to-Noise Ratio (SNR) samplers. Specifically, the authors proposed Logit-Normal sampling, which uses a logit-normal distribution to bias training timesteps towards the middle of the trajectory or towards data/noise endpoints. This approach re-weights the loss function to provide a stronger optimization signal where the velocity prediction is most difficult.

The proposed architecture, MM-DiT, improves upon standard Diffusion Transformers (DiT). Unlike models that use fixed text representations, MM-DiT incorporates separate learnable streams for image and text tokens. By joining these sequences for attention operations while maintaining independent weights for each modality, the architecture allows both representations to evolve in their own space while simultaneously accounting for the other.

The models were trained on large-scale datasets, including ImageNet and CC12M, with the largest iterations reaching 8 billion parameters. The evaluation framework utilized a combination of objective metrics—such as FID on CLIP features, CLIP scores, T2I-CompBench, and GenEval—alongside extensive human preference ratings. Results from a large-scale study of 61 different formulations demonstrated that the improved RF with logit-normal sampling consistently outperformed established diffusion variants like EDM and standard linear/cosine schedules. The scaling study revealed a predictable trend where lower validation loss strongly correlated with improved human preference and prompt adherence.

The authors acknowledged that despite significant improvements in typography and spatial reasoning, gaps remain between automatic metrics and nuanced human aesthetic judgment. Additionally, the study noted that while straight paths reduce error accumulation, the computational cost of training 8B parameter models remains a significant constraint.

In conclusion, this work advances the state of the art by proving that Rectified Flow, when combined with proper noise sampling and a truly multimodal transformer architecture, surpasses

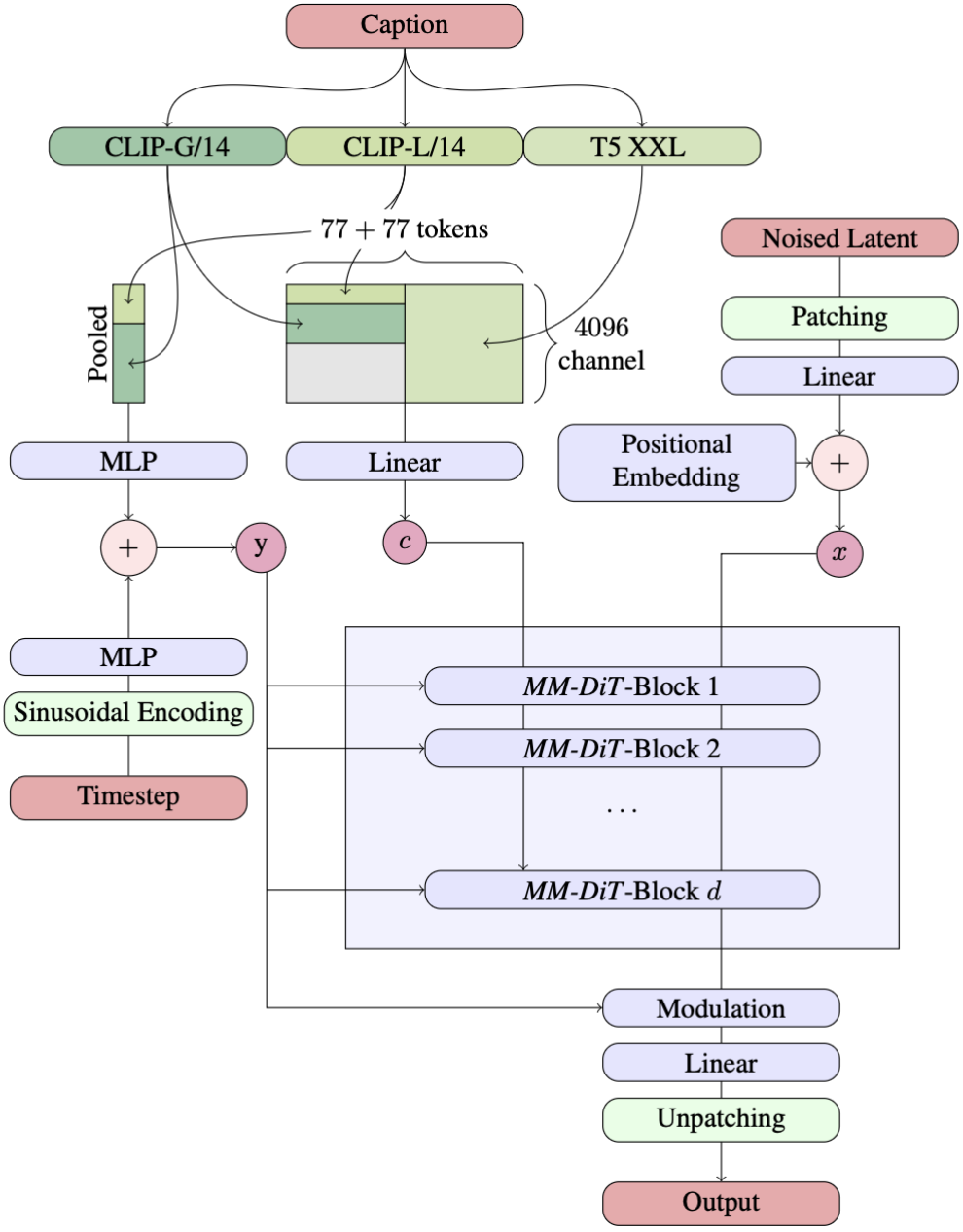


Figure 19: Model architecture. The RMS-Norm for Q and K can be added to stabilize training runs.

existing diffusion standards. The findings suggest that scaling such models leads to superior high-resolution synthesis, making it a viable and efficient alternative for future generative modeling research.

6.14 MusicRL: Aligning Music Generation to Human Preferences

The paper [76] introduces **MusicRL**, the first large-scale text-to-music framework explicitly fine-tuned using Reinforcement Learning from Human Feedback (RLHF). Built upon the autoregressive MusicLM model, MusicRL addresses a critical limitation of existing systems: the misalignment between high-fidelity audio production and subjective human musical preferences, which automatic metrics fail to capture.

The core methodology reframes music generation as a policy optimization problem, maximizing sequence-level rewards. The authors simplify the original MusicLM pipeline into a single autoregressive stage that jointly models semantic and coarse acoustic tokens using a **hierarchical**

Transformer architecture. Fine acoustic details are subsequently handled by the SoundStorm parallel decoding model during inference.

The reinforcement learning procedure employs a KL-regularized policy gradient formulation, utilizing three distinct reward signals:

- **Text Adherence:** Measured via cosine similarity between MuLan text–audio embeddings.
- **Acoustic Quality:** Estimated by a reference-free predictor trained to approximate human Mean Opinion Scores (MOS).
- **User Preferences:** Derived from over 300,000 pairwise comparisons collected via a web interface and modeled using a Bradley–Terry formulation.

The study evaluates three progressively refined models: **MusicRL-R** (automated rewards), **MusicRL-U** (user preference rewards), and **MusicRL-RU** (a hybrid sequential fine-tuning). Results demonstrate that all RL-tuned variants significantly outperform the MusicLM baseline, with MusicRL-RU achieving the highest human preference scores across diverse genres.

A key insight from the analysis is that neither text adherence nor acoustic quality alone explains human musical appeal. These findings emphasize that human judgment relies on high-level musical structures that transcend current automatic metrics, positioning human-in-the-loop learning as a vital paradigm for future advances in generative audio.

7 Foundation Models for Music-Sensorial Systems

7.1 UniAudio: An Audio Foundation Model Toward Universal Audio Generation

UniAudio [77] presents a unified large language model (LLM) framework that advances the field of generative audio by moving away from task-specific designs toward a universal generation paradigm. Unlike prior approaches that rely on separate models for speech, music, and sound effects, UniAudio tokenizes all input conditions—such as phonemes, textual descriptions, and MIDI—and target audio into a single discrete sequence[cite: 8, 9]. By concatenating source–target pairs and employing a next-token prediction objective, the system is capable of performing 11 distinct audio generation tasks, including Text-to-Speech (TTS), Voice Conversion (VC), Singing Voice Synthesis (SVS), and Speech Enhancement (SE), within a single model.

The architectural backbone of UniAudio is a novel *multi-scale Transformer* designed to address the computational challenges posed by the long sequences typical of neural audio codecs. Standard flattening of residual vector quantization (RVQ) codes results in prohibitive sequence lengths; therefore, UniAudio utilizes a hierarchical structure comprising a global Transformer and a local Transformer[cite: 32]. The global module models inter-frame correlations at a semantic level, while the local module handles intra-frame acoustic details[cite: 33]. This design effectively reduces the sequence length processed by the global model from $T \times n_q$ to T (where n_q is the number of quantizers), allowing for efficient scaling and the use of larger quantization depths without quadratic complexity explosion.

To achieve the capabilities of a foundation model, training is conducted on a massive scale involving 165,000 hours of public audio data and approximately 1 billion parameters. The training process is bifurcated into two stages: a pre-training stage where the model learns from 7 diverse tasks jointly to capture intrinsic audio properties and cross-modal relationships, and a fine-tuning stage where the model adapts to unseen tasks such as audio editing and speech dereverberation. Experimental results demonstrate that multi-task learning is mutually beneficial; the joint-trained UniAudio consistently outperforms single-task baselines trained on identical data subsets.

Evaluations show that UniAudio achieves state-of-the-art or competitive performance across the supported tasks. In zero-shot TTS and Voice Conversion, the model demonstrates superior speaker similarity and lower word error rates compared to strong baselines like VALL-E and YourTTS. In sound and music generation, it achieves results comparable to specialized systems, although it does not yet surpass models trained on significantly larger private datasets for music[cite: 601, 606]. Furthermore, ablation studies on the multi-scale architecture confirm that it provides a better trade-off between generation quality and inference efficiency compared to flattening, parallel, or coarse-first prediction strategies.

Overall, *UniAudio* establishes that building a universal audio generation model is both feasible and advantageous. By formulating audio generation as a unified sequence modeling problem, the framework simplifies the development of generative audio systems and demonstrates strong potential for scalability and generalization to new, unseen auditory tasks through simple fine-tuning.

7.2 CLAP: Learning Audio Concepts from Natural Language Supervision

CLAP [6] introduces a general-purpose audio representation learning framework that departs from traditional class-label–centric audio analytics by leveraging natural language supervision. Conventional audio models are typically trained for narrowly defined tasks under fixed label spaces, which constrains their flexibility and limits their ability to generalize to unseen categories. CLAP addresses this limitation by learning audio concepts directly from paired audio–text descriptions, enabling zero-shot inference and flexible semantic querying without requiring task-specific retraining. Inspired by the success of contrastive vision–language models such as CLIP, CLAP aims to bridge acoustic and linguistic semantics within a unified multimodal embedding space.

The core of CLAP consists of a dual-encoder architecture composed of an audio encoder and a text encoder, each followed by a learnable linear projection that maps modality-specific representations into a shared embedding space. Audio inputs are represented as log-Mel spectrograms, while textual inputs are natural language captions describing acoustic events, scenes, or actions. The model is trained using a symmetric contrastive learning objective that maximizes the similarity

between matched audio–text pairs while minimizing similarity to mismatched pairs within a batch. This formulation allows CLAP to jointly optimize both encoders end-to-end, aligning auditory and linguistic representations through a scalable and modality-agnostic objective.

Training is performed on a relatively modest dataset of 128,010 audio–text pairs constructed from four public audio captioning and tagging datasets: FSD50K, ClothoV2, AudioCaps, and MACS. Despite the limited scale of paired data compared to large vision–language models, CLAP demonstrates strong generalization capabilities. The audio encoder is instantiated as CNN14 pretrained on AudioSet, while the text encoder is based on BERT-base. Both encoders are unfrozen during training, a design choice shown empirically to yield the strongest zero-shot performance across downstream tasks.

A key capability enabled by CLAP is zero-shot linear classification. At inference time, both audio samples and textual class descriptions are embedded into the same multimodal space, and predictions are obtained by computing cosine similarity between audio embeddings and text embeddings. This approach removes the need for predefined label sets and allows users to specify arbitrary class descriptions via natural language prompts. Prompt engineering is shown to significantly affect performance, with templated descriptions such as "This is a sound of [class label]" yielding consistent improvements over raw label tokens.

CLAP is evaluated on 16 downstream datasets spanning eight domains, including sound event classification, music understanding, acoustic scene classification, speech emotion recognition, keyword spotting, vocal sound classification, and speaker counting. In zero-shot settings, CLAP establishes state-of-the-art performance on several benchmarks, notably outperforming prior zero-shot methods on ESC50, UrbanSound8K, and FSD50K. In supervised and fine-tuning setups, CLAP achieves competitive or state-of-the-art results on multiple tasks, demonstrating that the learned representations are also effective as transferable audio features.

Analysis of the results highlights both strengths and limitations of the approach. CLAP performs particularly well on sound event–centric tasks, which aligns with the nature of its training captions, but shows weaker performance on speech-centric tasks such as emotion recognition and keyword spotting, where semantic descriptions are less explicit in the training data. The authors attribute this gap to the scarcity and limited expressiveness of speech-related captions and suggest that richer and more diverse audio–text pairings could further enhance performance.

Overall, *CLAP* demonstrates that natural language supervision provides a viable and powerful alternative to class-label–based training for audio understanding. By learning aligned audio–text representations through contrastive pretraining, CLAP enables zero-shot prediction, flexible semantic querying, and broad task generalization, positioning it as an important step toward audio foundation models capable of scalable and adaptable multimodal reasoning.

7.3 ACE-Step: A Foundation Model for Fast and Controllable Music Generation

ACE-Step [78] introduces an open-source foundation model for text-to-music generation designed to bridge the gap between the speed and controllability of diffusion-based approaches and the long-range musical coherence traditionally associated with autoregressive models. The work addresses several limitations of prior open-source music generation systems, such as slow inference, weak lyric–audio alignment, limited controllability, and poor scalability to long-form musical structures. The authors position ACE-Step as a general-purpose music foundation model, aiming to play a role analogous to Stable Diffusion in the image domain.

The model is trained on a large-scale curated music corpus comprising approximately 1.8 million unique musical pieces, corresponding to roughly 100,000 hours of audio spanning 19 languages. This dataset includes both instrumental and vocal music and is enriched with multiple conditioning signals, such as descriptive audio captions, transcribed and aligned lyrics, and musical attributes including tempo (BPM), key, and stylistic tags. Data quality is enforced through automated filtering using the Audiobox aesthetics toolkit, which removes low-fidelity recordings and live performances. Training follows a two-stage strategy, consisting of large-scale pre-training on the full dataset and subsequent fine-tuning on a higher-quality subset of approximately 20,000 hours.

As illustrated in Figure 20, ACE-Step adopts a diffusion-based generation paradigm operating in a compressed latent space. Audio is represented using mel-spectrograms and encoded by a Deep Compression AutoEncoder (DCAE), adapted from prior work to achieve a favorable trade-off between compression ratio and reconstruction fidelity. The generative backbone is a Linear Diffusion

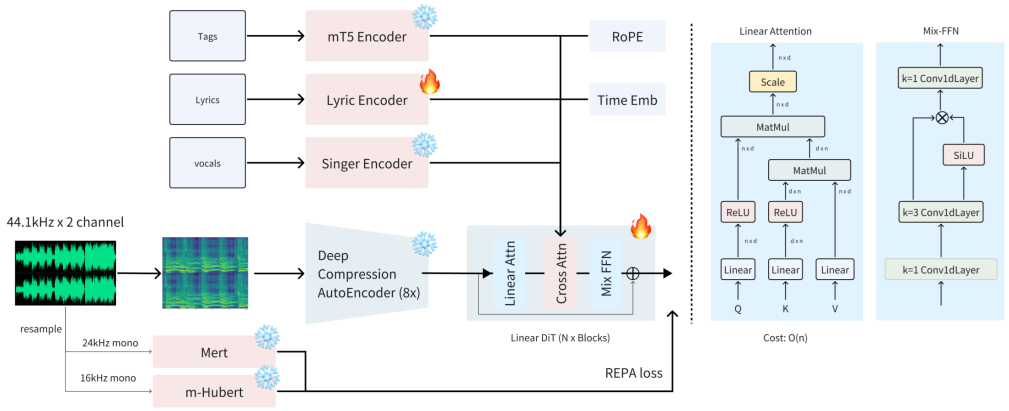


Figure 20: Overview of the ACE-Step architecture [78]. A Deep Compression AutoEncoder (DCAE) encodes mel-spectrograms into a compact latent space, which is modeled by a Linear Diffusion Transformer trained with flow matching and guided by semantic alignment (REPA) using MERT and mHuBERT.

Transformer (Linear DiT), inspired by the Sana architecture, which replaces conventional U-Net structures with a transformer-based model employing linear attention for improved scalability on long sequences. Several architectural simplifications, including shared adaptive layer normalization and one-dimensional convolutional feed-forward layers, are introduced to reduce memory consumption and computational cost.

Generation is formulated using a flow matching objective, which directly learns a continuous vector field that transports samples from a noise distribution to the data distribution. Compared to score-based diffusion, flow matching enables simpler training objectives and faster convergence. To further improve semantic coherence and lyric intelligibility, ACE-Step incorporates Representation Alignment (REPA) as an auxiliary training signal. Intermediate representations from the diffusion transformer are aligned with embeddings from pre-trained self-supervised models, namely MERT for music understanding and mHuBERT for multilingual speech representation. This alignment enforces low-frequency semantic constraints during training, preventing the model from prematurely overfitting to acoustic details at the expense of lyrical correctness.

Conditioning is achieved through multiple encoders, including a frozen multilingual text encoder for prompts, a trainable lyric encoder adapted from SongGen, and a speaker encoder used for timbre control and voice cloning. The model supports variable-length generation and robust handling of diverse input formats, including missing lyrics or purely instrumental tracks. Beyond text-to-music synthesis, ACE-Step demonstrates extensibility to specialized tasks via fine-tuning, such as lyric-to-vocal generation, text-to-sample synthesis, stem generation, singing-to-accompaniment, and style-specific LoRA adaptations.

Evaluation combines objective, perceptual, and subjective metrics. Audio fidelity is assessed using Fréchet Audio Distance (FAD) and Audiobox aesthetics scores. Musicality is evaluated using SongEval, a benchmark designed to correlate strongly with human judgments. Style alignment is measured using CLAP and the music-specific Mulan contrastive model, while lyric alignment is quantified through Whisper-based forced alignment confidence. Human listening tests with blind Likert-scale ratings further assess musicality, emotional expression, innovativeness, and sound quality. In addition, generation speed is measured using the Real-Time Factor (RTF), highlighting ACE-Step's ability to synthesize up to four minutes of music in approximately twenty seconds on modern GPUs.

Experimental results show that ACE-Step achieves state-of-the-art performance among open-source music generation models on SongEval and generation speed, while remaining competitive with commercial systems such as Suno v3 in terms of coherence and lyric alignment. The authors emphasize, however, the persistent discrepancy between automatic metrics and human perception, and identify limitations related to the mel-spectrogram bottleneck, variability in output quality, and imperfect style adherence.

Overall, ACE-Step demonstrates that combining efficient diffusion transformers, large-scale curated data, and semantic alignment with pre-trained representations enables the construction

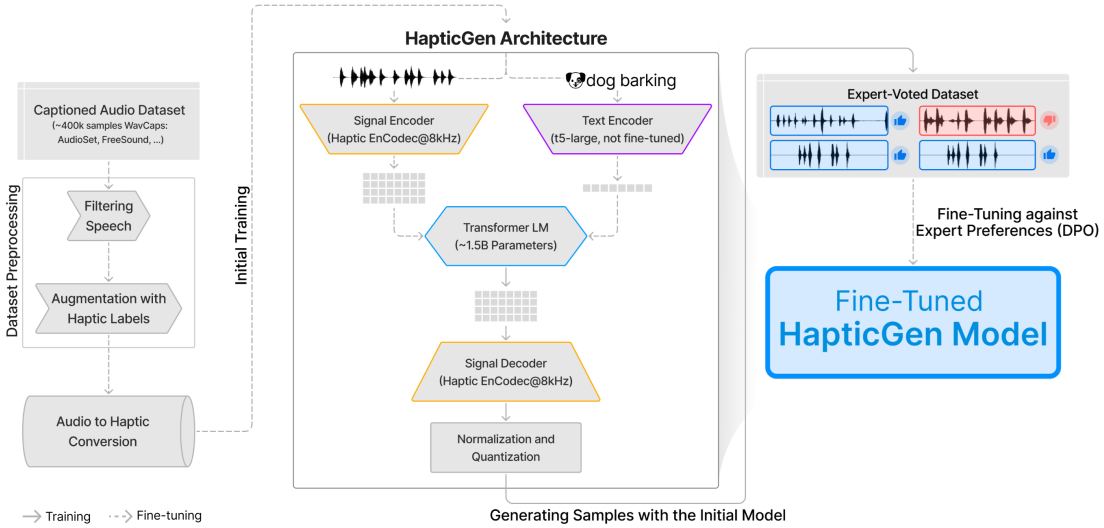


Figure 21: Overview of the HapticGen [79] architecture and training pipeline. A large captioned audio dataset (WavCaps) is converted into vibrotactile signals and augmented with tactile-oriented captions. An autoregressive transformer adapted from MusicGen/AudioGen is trained on the converted haptic data and subsequently fine-tuned using expert preference signals via Direct Preference Optimization (DPO).

of fast, controllable, and extensible music generation foundation models. The work represents a significant step toward open, general-purpose music generation systems capable of supporting diverse creative workflows.

7.4 HapticGen: Generative Text-to-Vibration Model for Streamlining Haptic Design

HapticGen [79] introduces the first generative foundation model capable of producing vibrotactile haptic signals directly from natural language descriptions. The work addresses a longstanding challenge in haptics research and practice: the lack of scalable, data-driven tools for ideating and authoring vibration-based feedback. Unlike text-to-audio or text-to-image generation, haptic design suffers from limited datasets, the absence of standardized vocabularies, and a heavy reliance on expert intuition and iterative trial-and-error. HapticGen is positioned as a generative design assistant that lowers the barrier to haptic creation, particularly in extended reality (XR) contexts, by enabling designers to specify desired tactile experiences using textual prompts.

Due to the absence of large-scale labeled haptic datasets, the authors adopt a bootstrapping strategy based on the affinity between audio and vibration signals. An initial training corpus is constructed by converting the WavCaps audio captioning dataset, which contains approximately 400,000 ten-second audio clips paired with textual descriptions, into vibrotactile signals using an automated audio-to-haptic conversion pipeline. After filtering out speech-related samples, the resulting haptic-converted dataset comprises roughly 335,000 vibration signals. To better align the textual conditioning with haptic design intent, the original audio captions are augmented with tactile-oriented descriptions generated using a large language model, yielding multiple caption variants per sample.

As illustrated in Figure 21, HapticGen is built upon an autoregressive transformer-based architecture adapted from the MusicGen and AudioGen v2 frameworks. The model operates on discrete tokens produced by an EnCodec compression model retrained specifically for vibrotactile signals at an 8 kHz sampling rate and 8-bit resolution. Textual prompts are processed using a frozen T5-large text encoder, whose embeddings condition the transformer language model. The generative backbone employs a medium-sized configuration with approximately 1.5 billion parameters, selected to balance expressive capacity and computational feasibility. Compared to post-hoc audio-to-haptic conversion at inference time, training directly on haptic-converted signals reduces latency

and allows the incorporation of expert-designed or manually edited vibrations during training and fine-tuning.

Beyond large-scale pre-training, a central contribution of HapticGen lies in its alignment with human haptic preferences. The authors conduct a design study with expert haptic designers, who use an interactive interface to prompt the initial model, evaluate generated vibrations on VR controllers, and provide binary preference feedback (thumbs up or down). This process results in an expert-voted preference dataset consisting of 1,297 samples, including paired preferences for contrastive optimization. The model is subsequently fine-tuned using Direct Preference Optimization (DPO), enabling it to better capture qualitative aspects of haptic experience that are difficult to encode through automated objectives alone. Additional interface-level enhancements, such as automatic prompt variation and signal normalization, are introduced to address issues observed during expert use, including low-intensity or imperceptible outputs.

Evaluation of HapticGen is performed through a comprehensive human-centered study rather than relying on objective perceptual metrics, which are largely absent in the haptics domain. A controlled within-subject A/B testing experiment is conducted with 32 participants, comparing the fine-tuned HapticGen model against a baseline pipeline consisting of a text-to-audio model (AudioGen v2) followed by audio-to-haptic conversion. Participants assess generated vibrations using validated Factors of Haptic Experience (HX), including Autotelics, Realism, and Expressivity, alongside system-level usability measures such as Workload, Future Use, Goal achievement, and Iteration support. Statistical analysis using paired t-tests and Wilcoxon signed-rank tests reveals significant improvements for HapticGen across multiple dimensions, including perceived realism, reduced workload, and increased willingness for future use.

Qualitative feedback further indicates that HapticGen streamlines the ideation process for both novice and expert designers, enabling the rapid exploration of diverse and nuanced vibrotactile patterns. Participants report that the model better reflects their intended scenarios and supports creative exploration, particularly for dynamic or emotionally expressive interactions. At the same time, the authors acknowledge limitations related to hardware constraints, the difficulty of generating repeated or multi-phase effects, and the inherent subjectivity of haptic evaluation. They also emphasize the lack of standardized objective metrics for haptics, which necessitates continued reliance on human studies for model assessment.

Overall, HapticGen demonstrates that adapting large-scale generative audio models to the haptic domain, combined with automated dataset construction and preference-based alignment, is a viable path toward general-purpose generative haptic systems. The work establishes a foundation for future research on text-driven haptic generation and highlights the importance of human-in-the-loop evaluation for modalities where perceptual quality cannot yet be reliably captured by automatic metrics.

7.5 MuMu-LLaMA: Multi-modal Music Understanding and Generation via Large Language Models

Multimodal Music Understanding and Generation using LLaMA (MuMu-LLaMA) [80] is a foundational model which leverages multi-modal adapters and domain-specific encoders to create a system able to capture information from music, text, images and video modalities. Figure 22 illustrates some of MuMu-LLaMA's capabilities, encompassing tasks such as music understanding, text-to-music generation, prompt-based music editing, and multi-modal music generation.

The framework's architecture is based on 4 core components:

1. Pre-trained feature encoders, specifically tailored for each modality, transforming user input information into rich abstract representations.

2. Understanding adapters – modules acting as intermediaries between modality-specific encoders and a shared feature space.

3. A LLaMA language model, acting as a bridge between all modalities, is used to ground information from different sources and interpret it based on the user's request.

4. Projection layer acting as a translator from contextualized information to generated music content, as an answer to the user's request – whenever necessary.

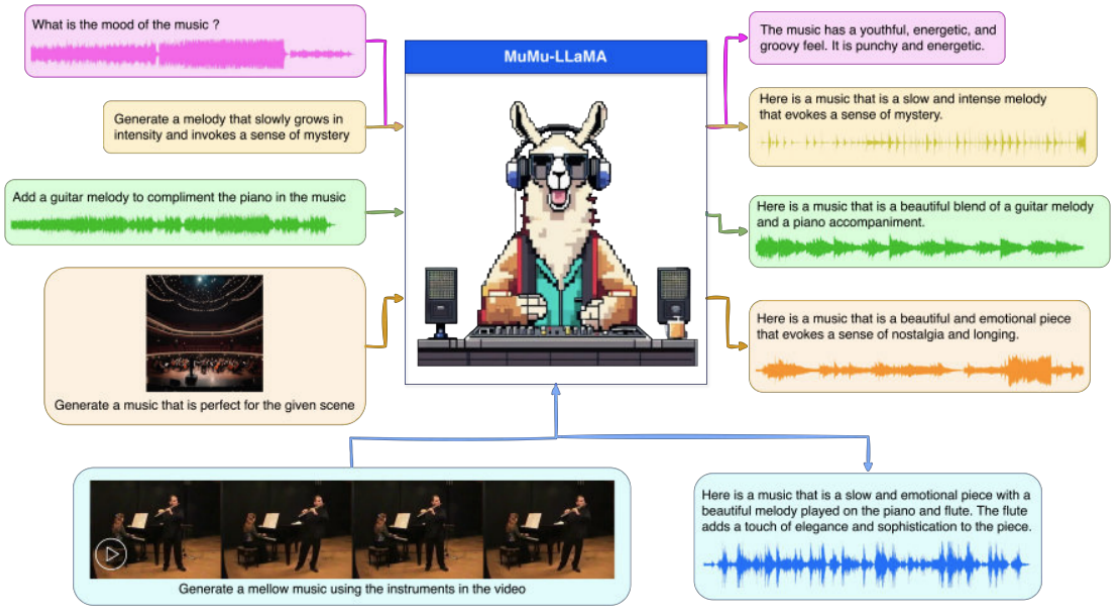


Figure 22: The multi-modal capabilities of MuMu-LLaMA [80].

A MERT-based encoder [81] is used for input audio modality, which was previously shown to excell in music tagging scenarios. A simple ViT is used to handle image-based prompts, while for video inputs an extended variant of ViViT [82] extracts spatio-temporal tokens.

The multi-modal understanding adapters consist of a 1D convolutional layer, linear projection, and a dense network. Temporal modalities are processed through an attention-enhanced RNN to flatten the temporal dimension and project audio and video features in a shared non-temporal space.

To effectively train large multi-modal large language models (MLLMs) for music understanding and generation, the authors note a key challenge: existing music datasets are insufficient for instruction tuning that covers multi-modal tasks such as text $\rightarrow$ music, image $\rightarrow$ music, and video $\rightarrow$ music. To address this gap, the authors construct a comprehensive instruction dataset tailored to music tasks by leveraging existing public datasets and large pre-trained models to annotate and generate paired multi-modal data. Afterwards, this curation results in four specialized task-related datasets, totaling $\approx$ 168 hours:

- MUCaps – for text-music understanding and generation, containing 10-second music clips paired with detailed captions and describing aspects such as instruments, tempo, mood.
- MUEdit – instruction-based music editing, starting from an original audio, modifying its Speed/Pitch or Add/Delete/Replace on some instruments, and describing this modification in a textual form. This triplet of original audio, modified audio, and textual description of modification is then used to train the system to apply a desired edit to any input audio.
- MUImage – image $\rightarrow$ music tasks, where corresponding image-audio-caption pairs are used to define a shared common space to enable associations between the two modalities.
- MUVideo – video $\rightarrow$ music tasks, with corresponding video-music pairs sampled from Balanced-AudioSet, generating captions for all acquired music files and their corresponding videos, and training

MuMu-LLaMA generally outperforms baselines (LTU or LLaMA adapters trained only on MusicQA) on established language-based evaluation metrics, showing better alignment with reference captions and correct responses to music-grounded questions. For text/image/video-to-music tasks, MuMu-LLaMA yields higher fidelity and more semantically aligned music outputs compared to previous methods. For visuals, it incorporates modality cues to guide generation more effectively than traditional pipelines that treat music as generic audio. The model showcases the capability to edit existing music guided by natural-language instructions, and performs competitively or

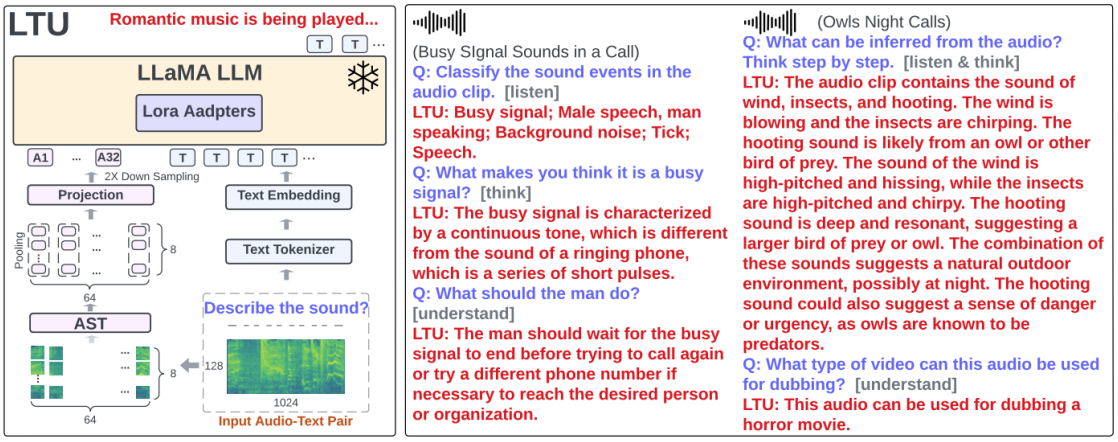


Figure 23: Architecture of LTU and several prompting examples [83].

better than external music editing baselines like InstructME and AUDIT. Subjective evaluation for text/image/video-to-music tasks indicates users' preference to content generated by MuMu-LLaMA, compared to other well-established frameworks such as AudioLDM 2 [62] or MusicGen [59], with over 80% of users showing preference towards this work's image-to-music and video-to-music generated content.

7.6 Listen, Think, and Understand (LTU)

The authors of LTU [83] propose a new foundation model trained for audio perception and reasoning. Its main characteristic is the ability to efficiently provide audio-based reasoning, with several examples being shown in Figure 23.

LTU's architecture, depicted in Figure 23, is composed of: an audio encoder based on a pre-trained Audio Spectrogram Transformer (AST), a pre-trained LLaMA-TB LLM and LoRA adapters, allowing for efficient instruction-based fine-tuning. Text (user's question) and audio tokens are concatenated and the final LLM is fine-tuned for next token prediction.

One of the main reasons for LTU's wide range of capabilities for diverse audio tasks is the development of a new audio dataset called OpenAQA-5M, consisting of (audio, question, answer) triplets, unifying nearly all audio tasks into a single dataset. This dataset is based on already well-established public dataset and does not contain any new additional samples, totaling 845K audio samples.

On conventional audio classification tasks LTU outperforms CLAP [6] by a wide margin. For open-ended audio experiments, LTU is capable of answering follow-up questions about various details in the audio, explaining its prediction in a chain-of-thought manner, while also showcasing an increased understanding of the scene and the connections between different sound sources. In human evaluation experiments, LTU's output is preferred in more than 70% cases, compared to human answers or GPT-4-generated answers. Both the instruction following rate and the factual correctness rate are significantly high, indicating LTU performs consistently well for open-ended reasoning.

7.7 DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism

DiffSinger [84] is a novel acoustic model for Singing Voice Synthesis that employs diffusion probabilistic models to address fundamental limitations in existing approaches. The work represents a significant advancement in the field by introducing a shallow diffusion mechanism that balances synthesis quality with computational efficiency, while also demonstrating generalizability to text-to-speech applications through the DiffSpeech variant.

The motivation for this research arose from persistent challenges in contemporary SVS acoustic models. Traditional approaches utilizing simple loss functions such as L1 or L2 norms suffered from over-smoothing effects that diminished the naturalness of synthesized singing voices. Alternative methods based on Generative Adversarial Networks, while capable of producing sharper outputs,

exhibited training instability that hindered their practical deployment. These limitations created a clear need for a more robust modeling framework that could generate high-quality, natural-sounding singing voices while maintaining stable training dynamics.

The technical approach centered on adapting diffusion probabilistic models to the SVS domain by modeling the conditional distribution of mel-spectrograms given musical scores. The authors introduced a shallow diffusion mechanism as their primary innovation, which fundamentally reconsidered the standard diffusion process. Rather than initiating the reverse diffusion from pure Gaussian noise, this mechanism commenced from an intermediate diffusion step denoted as k. This step was strategically identified as the intersection point between diffusion trajectories of ground-truth mel-spectrograms and those generated by a simple auxiliary decoder. By leveraging prior knowledge from simpler models, this approach reduced the computational burden on the reverse diffusion process while preserving synthesis quality.

To operationalize the shallow diffusion mechanism, the authors developed a boundary prediction network that adaptively determined the optimal starting step k for each input. This network was trained as a binary classifier to distinguish whether a mel-spectrogram at a given diffusion step originated from ground-truth data or the auxiliary decoder output. The classification boundary naturally corresponded to the intersection point of the two diffusion trajectories, enabling automatic identification of the optimal initialization point for the shallow diffusion process.

The model architecture comprised several interconnected components working in concert. An encoder processed musical score information, transforming symbolic notation into continuous representations. A step embedding module incorporated temporal information about the diffusion process. The auxiliary decoder, implemented as a simpler model, provided initial mel-spectrogram estimates that served as starting points for the shallow diffusion. The denoiser, built upon a non-causal WaveNet architecture, performed the iterative refinement process characteristic of diffusion models. This architectural design enabled the model to effectively combine the efficiency of simpler predictive models with the quality advantages of diffusion-based generation.

The training procedure followed a carefully designed two-stage protocol. During the warmup stage, the auxiliary decoder and boundary predictor were trained to establish reliable initialization capabilities. The main training stage then optimized the full DiffSinger model using standard diffusion model objectives, with the shallow diffusion mechanism integrated throughout. This staged approach ensured that each component reached appropriate performance levels before being incorporated into the complete system.

For empirical validation, the authors constructed PopCS, a Chinese singing dataset specifically designed for SVS research. The evaluation framework employed Mean Opinion Score assessments, a standard subjective metric for voice synthesis quality, comparing DiffSinger against state-of-the-art baselines including FFT-Singer and GAN-Singer. Additionally, the authors conducted ablation studies to isolate the contributions of the shallow diffusion mechanism and performed efficiency analyses measuring real-time factors to quantify computational performance.

The experimental results demonstrated DiffSinger's superiority across multiple dimensions. In subjective quality assessments, DiffSinger achieved a MOS of 3.85 ± 0.11, surpassing FFT-Singer (3.67 ± 0.11) and GAN-Singer (3.74 ± 0.12). The shallow diffusion mechanism proved instrumental in these improvements, simultaneously enhancing synthesis quality and accelerating inference by 45.1 percent, reducing the real-time factor from 0.348 to 0.191. These results validated both the effectiveness of the diffusion modeling approach and the practical utility of the shallow diffusion innovation.

The generalizability of the approach was further demonstrated through DiffSpeech, an adaptation of DiffSinger to text-to-speech synthesis. Evaluations on the LJSpeech benchmark showed DiffSpeech outperforming established TTS models including FastSpeech 2 and Glow-TTS. The shallow diffusion mechanism maintained its benefits in the TTS domain, contributing to a 29.2 percent speedup while preserving synthesis quality. This cross-task success indicated that the core innovations transcended domain-specific characteristics and represented broadly applicable advances in neural acoustic modeling.

While the paper achieved substantial progress, certain limitations remained inherent to the approach. The system relied on a separate vocoder for final waveform generation, following standard practice in mel-spectrogram-based synthesis but introducing an additional component in the synthesis pipeline. The paper did not extensively discuss potential failure modes or edge cases where the shallow diffusion mechanism might prove less effective. Additionally, the evaluation focused primarily on a single language dataset for SVS, leaving questions about cross-linguistic

generalization partially unexplored.

In conclusion, DiffSinger established diffusion probabilistic models as powerful tools for singing voice synthesis and demonstrated their broader applicability to speech synthesis tasks. The shallow diffusion mechanism represented a significant methodological contribution that addressed the quality-efficiency trade-off inherent in diffusion models. By intelligently initializing the reverse diffusion process from intermediate states informed by simpler models, the work achieved superior synthesis quality with substantially reduced computational requirements. These advances positioned diffusion-based approaches as compelling alternatives to existing paradigms in neural acoustic modeling, opening new directions for future research in voice synthesis technologies.

7.8 Movie Gen: A Cast of Media Foundation Models

This study [85] presents *Movie Gen*, a comprehensive suite of foundation models designed to generate high-quality 1080p HD videos with synchronized audio. The authors' core contribution is the development of a 30B parameter transformer-based model capable of text-to-video synthesis, video personalization, precise video editing, and video-to-audio generation, establishing a new state-of-the-art across these media generation tasks.

The research addresses the challenge of endowing AI systems with the ability to generate, compose, and predict complex media content, analogous to the capabilities Large Language Models (LLMs) have achieved in text. Prior work in video generation often faced limitations in consistency, resolution, and the seamless integration of audio. The authors aimed to overcome these hurdles by investigating the effects of scaling pre-training data, model size, and compute resources, hypothesizing that a unified, scalable architecture could solve distinct media generation problems such as personalization and editing within a single framework.

The methodology centered on training large-scale transformer models using a Flow Matching objective. The core architecture, *Movie Gen Video*, is a 30B parameter model built upon a LLaMa3 backbone, utilizing a Temporal Autoencoder (TAE) to compress videos into a spatio-temporally efficient latent space. This model was trained jointly on text-to-image and text-to-video tasks, treating images as single-frame videos to leverage diverse image datasets. To achieve high-definition output efficiently, the system generates videos at a lower resolution (e.g., 768px) and subsequently employs a dedicated Spatial Upsampler—a video-to-video generation model—to upscale content to 1080p. For audio, the authors developed *Movie Gen Audio*, a 13B parameter model that generates 48kHz audio synchronized with visual inputs, utilizing a DAC-VAE for high-fidelity audio representation.

The framework extends beyond basic generation to support advanced capabilities. For personalization, the authors introduced a post-training procedure where the model conditions on both a text prompt and a reference face image, employing a vision encoder to inject identity information and generating videos that maintain character identity. For editing, the authors developed *Movie Gen Edit*, which supports precise, instruction-guided modifications. This was achieved through a novel "Generative Instruction-Guided Video Segmentation" task, allowing the model to isolate and alter specific objects or backgrounds based on natural language instructions without requiring large-scale supervised editing data.

The models were pre-trained on an internet-scale dataset comprising approximately 100 million videos and 1 billion images for the video model, and 1 million hours of audio for the audio model. The training protocol involved rigorous data curation, including motion filtering to remove static or low-quality clips, content filtering to ensure diversity, and the use of synthetic captions generated by LLaMa3-Video to provide detailed text descriptions. Supervised Fine-Tuning (SFT) was subsequently performed on a smaller, curated set of high-quality, aesthetic videos to refine the model's generation capabilities.

To assess performance, the authors introduced two new benchmarks: *Movie Gen Video Bench* and *Movie Gen Audio Bench*. The evaluation employed both objective metrics (e.g., FVD, audio quality metrics) and subjective human evaluations comparing the model's outputs against commercial systems like Runway Gen3, LumaLabs, and OpenAI Sora. Ablation studies focused on verifying design choices such as the impact of joint image-video training, the efficacy of the spatial upsampler, and the necessity of trainable vision encoders for personalization.

Experimental results demonstrated that *Movie Gen* outperforms existing state-of-the-art models in text-to-video synthesis, video personalization, and video editing. Human evaluators consistently preferred *Movie Gen* for overall video quality and text alignment. The study confirmed that

scaling model parameters and data volume significantly improves generation fidelity and motion naturalness. Furthermore, the audio model successfully generated synchronized sound effects and music that matched the visual context, including diegetic and non-diegetic elements.

Despite these advancements, the authors acknowledged certain limitations. The model occasionally exhibits "motion completeness" issues, where generated videos may lack sufficient motion when prompted with out-of-distribution subjects or unusual activities. Additionally, while the editing capabilities are robust, maintaining perfect structural consistency during complex edits remains a challenge. The computational cost of generating high-resolution content necessitated the use of a separate upsampler, highlighting a trade-off between direct high-resolution generation and inference efficiency.

In conclusion, the paper establishes *Movie Gen* as a significant advancement in media generation, proving that simple, scalable transformer architectures trained with Flow Matching can achieve photorealistic and coherent video and audio. By unifying generation, editing, and personalization into a cohesive "cast" of models, the work sets a new benchmark for the field and provides a foundation for future research in multimodal generative AI.

7.9 LLARK: A Multimodal Instruction-Following Language Model for Music

LLARK [86] is a large-scale multimodal instruction-following language model specifically designed for music understanding. Unlike prior approaches that focus on narrowly defined Music Information Retrieval (MIR) tasks, LLARK aims to provide a unified framework capable of addressing a broad spectrum of music-related queries—ranging from low-level attribute estimation to high-level semantic reasoning—through natural language interaction. The work positions LLARK as an important step toward music-centric multimodal foundation models, analogous to recent advances in vision-language and audio-language modeling.

The central motivation of the paper lies in the observation that music exhibits unique structural properties, such as harmony, rhythm, key, and instrumentation, which are not adequately captured by existing multimodal models trained predominantly on speech or general audio. To address these limitations, the authors propose a model that jointly processes raw audio and textual instructions, producing natural language responses conditioned on both modalities. This formulation allows a single model to flexibly perform multiple tasks without task-specific heads or fine-tuning.

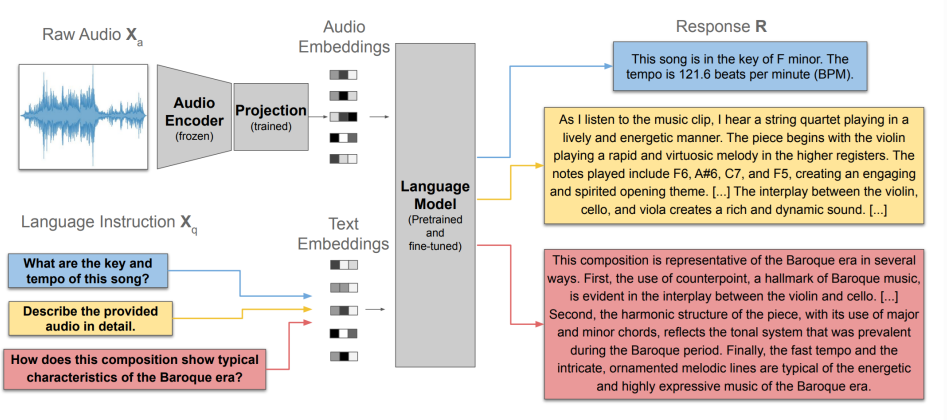


Figure 24: Overview of the LLARK architecture. A pretrained audio encoder extracts representations from raw audio, which are projected into the language model embedding space and jointly processed with textual instructions by a pretrained LLM.

Model Architecture. The LLARK architecture, illustrated in Fig. 24, consists of three principal components: a pretrained generative audio encoder, a projection module, and a pretrained large language model (LLM). The audio encoder is instantiated using Jukebox [58], a generative model shown to produce rich musical representations. Rather than collapsing temporal information via global averaging, LLARK preserves temporal structure by mean-pooling embeddings over short windows, enabling the model to reason about musical progression.

The projection module is implemented as a single linear layer that maps high-dimensional audio embeddings into the embedding space of the LLM. This design choice follows insights from recent multimodal models, demonstrating that simple projection layers can be sufficient when strong pretrained components are employed. The language model itself is based on LLaMA 2 7B [87], fine-tuned to generate responses conditioned on both textual instructions and projected audio features. During training, the audio encoder is frozen, while the projection module and the LLM are optimized jointly.

Instruction-Tuning Data Pipeline. A key contribution of the paper is the construction of a large-scale instruction-tuning dataset for music, generated entirely from open-source resources. The pipeline, depicted in Fig. 25, transforms heterogeneous music datasets into a unified instruction-following format. Given the scarcity of richly annotated music corpora, the authors augment existing metadata by automatically extracting musically salient features such as tempo, global key, chord progressions, and beat grids using pretrained signal processing models.

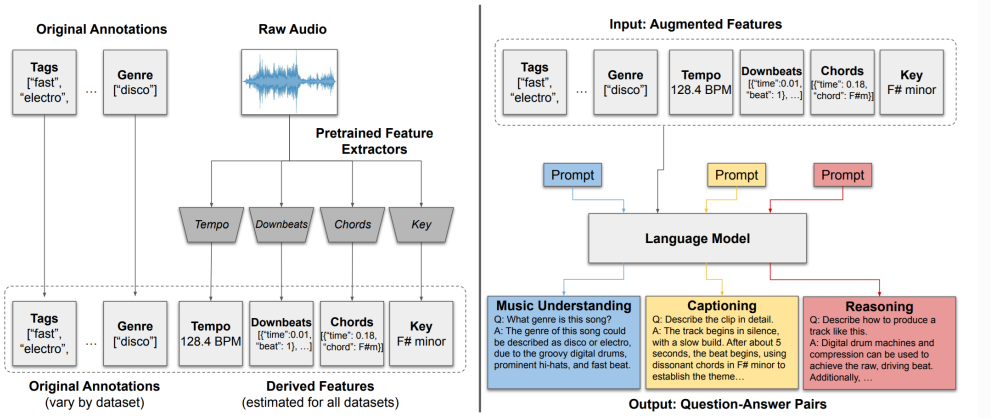


Figure 25: LLARK data generation pipeline. Left: metadata augmentation via pretrained music analysis models. Right: instruction–response generation across three task families using a language model.

These augmented annotations are then converted into natural language question–answer pairs using large language models. The resulting dataset comprises approximately 1.2 million instruction pairs derived from around 164,000 music tracks, spanning three task families: *music understanding* (e.g., tempo or key estimation), *captioning* (descriptive summaries of music clips), and *reasoning* (higher-level semantic and stylistic inference). This approach enables LLARK to learn a broad and flexible mapping between musical content and linguistic concepts.

Evaluation and Results. The authors conduct an extensive evaluation covering classification, regression, captioning, and reasoning tasks. On music understanding benchmarks, LLARK consistently outperforms existing multimodal baselines and approaches the performance of task-specific state-of-the-art models, despite operating in a zero-shot or weakly supervised setting. For captioning and reasoning tasks, human evaluations demonstrate a strong preference for LLARK's outputs, which are judged to be more musically grounded and informative than competing models.

Ablation studies further highlight the importance of the generative audio encoder and the instruction-tuning paradigm. Replacing Jukebox with contrastively trained audio encoders results in significant performance degradation, suggesting that generative audio representations capture musical structure more effectively for language-based reasoning. Dataset scaling experiments indicate diminishing returns when increasing data volume within a fixed distribution, emphasizing the importance of diversity and annotation richness over sheer dataset size.

Limitations and Discussion. Despite its strong performance, LLARK has several limitations. The model is constrained by the fixed input duration of the audio encoder, limiting long-term musical context. Additionally, evaluations for reasoning tasks rely on non-expert human annotators and LLM-based judges, which may introduce biases. The reliance on automatically extracted

musical features also introduces potential noise, although the authors argue that such augmentation acts as a useful inductive bias rather than a strict supervision signal.

Overall, LLARK represents a significant advancement toward general-purpose music understanding systems. By unifying diverse music tasks under an instruction-following framework and leveraging generative audio representations, the model demonstrates how multimodal LLMs can be adapted to domains with complex internal structure. The work lays a strong foundation for future research on scalable, interpretable, and interactive AI systems for music analysis and reasoning.

7.10 So-VITS-SVC: A Singing Voice Conversion System Based on VITS Architecture

So-VITS-SVC [88] is a dedicated singing voice conversion (SVC) system that enables the synthesis of singing voices by directly converting audio from source speakers to target speakers without relying on intermediate text representations. Unlike traditional text-to-speech approaches, this system extracts speaker-agnostic speech features and preserves the acoustic characteristics of the original audio, including pitch and intonation dynamics, through a neural vocoder-based synthesis pipeline.

Singing voice conversion presents distinct challenges compared to conventional speech synthesis or voice conversion tasks. The system addresses the limitation of requiring explicit phoneme-level text annotations by leveraging a content-based encoder approach. The core methodology employs the SoftVC content encoder to extract speech features from source audio, which are then directly integrated into a VITS-based synthesis architecture. The vocoder component was specifically replaced with NSF HiFiGAN to mitigate audio artifacts such as sound interruptions that may occur during synthesis.

The system comprises three primary components working in concert: a content encoder (SoftVC), a conditional variational autoencoder synthesizer (VITS), and a neural vocoder (NSF HiFiGAN). The content encoder extracts speaker-independent representations from input audio without converting to text. These feature vectors serve as conditioning information for the VITS synthesizer, which models the complex acoustic properties of singing. The NSF HiFiGAN vocoder converts the acoustic features into high-quality waveforms while preserving the fundamental frequency characteristics of the original signal.

The framework provides comprehensive preprocessing pipelines and flexible training configurations. Audio datasets are organized hierarchically by speaker with WAV format files. The preprocessing stage involves audio segmentation (recommended 5-15 seconds per clip), resampling to 44.1 kHz monophonic audio, and loudness normalization. Multiple speech encoder options are supported, including vec768l12, vec256l9, hubertsoft, whisper-ppg, and wavlm-based encoders. Fundamental frequency prediction employs various algorithms such as RMVPE, CREPE, DIO, or Harvest. The system optionally supports shallow diffusion-based post-processing through an auxiliary diffusion model trained on the same dataset.

The architecture supports multiple configuration parameters to accommodate diverse hardware constraints and quality requirements. Loudness embedding can be enabled to match input source loudness characteristics. The system offers alternative vocoders including nsf-snake-hifigan variants. Additionally, shallow diffusion mechanisms can be optionally integrated as a refinement stage, enabling iterative improvement of synthesis quality through diffusion-based post-processing.

The so-vits-svc project represents a significant contribution to singing voice conversion by decoupling content representation from speaker identity through speaker-agnostic feature extraction. By eliminating the requirement for text-based intermediate representations and preserving prosodic characteristics directly from source audio, the system enables flexible voice conversion applications particularly suited to fictional character voice synthesis. The modular architecture and multiple encoder/decoder options provide practitioners with substantial flexibility in balancing quality, computational efficiency, and training requirements for diverse applications in entertainment and creative content production.

References

[1] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019. [Online]. Available: https://arxiv.org/abs/1911.13254.

[2] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018. [Online]. Available: https://arxiv.org/abs/1711.02520.

[3] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. [Online]. Available: https://arxiv.org/abs/1807.03748.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477.

[5] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019. [Online]. Available: https://aclanthology.org/N19-1011/.

[6] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," *arXiv preprint arXiv:2206.04769*, 2022. [Online]. Available: https://arxiv.org/abs/2206.04769.

[7] I. Pereira, F. Korzeniowski, *et al.*, "Moises-light: Resource-efficient band-split u-net for music source separation," *arXiv preprint arXiv:2510.06785*, 2025.

[8] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-hq-an uncompressed version of musdb18," *(No Title)*, 2019.

[9] J. Chen, S. Vekkot, and P. Shukla, "Music source separation based on a lightweight deep learning framework (dttnet: Dual-path tfc-tdf unet)," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 656–660.

[10] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.

[11] W.-T. Lu, J.-C. Wang, Q. Kong, and Y.-N. Hung, "Music source separation with band-split rope transformer," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 481–485.

[12] W. Tong, J. Zhu, J. Chen, *et al.*, "Scnet: Sparse compression network for music source separation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 1276–1280.

[13] E. Gusó, J. Pons, S. Pascual, and J. Serrà, "On loss functions and evaluation metrics for music source separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 306–310.

[14] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[15] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[16] J. Yang, Y. Yang, W. Tu, X. Zhao, and C. Lin, "Band-scnet: A causal, lightweight model for high-performance real-time music source separation," in *Proc. Interspeech 2025*, 2025, pp. 4973–4977.

[17] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019, ISSN: 2329-9304. DOI: 10.1109/taslp.2019.2915167. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2019.2915167.

[18] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 31–35. DOI: 10.1109/ICASSP.2016.7471631.

[19] G. Mariani, I. Tallini, E. Postolache, M. Mancusi, L. Cosmo, and E. Rodolà, *Multi-source diffusion models for simultaneous music generation and separation*, 2024. arXiv: 2302.02257 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2302.02257.

[20] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix: A reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019. DOI: 10.21105/joss.01667. [Online]. Available: https://doi.org/10.21105/joss.01667.

[21] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[22] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020. DOI: 10.1109/TASLP.2020.3030497.

[23] J. F. Gemmeke, D. P. W. Ellis, A. J. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Accessed: 2025-02-15, IEEE, 2017, pp. 776–780. [Online]. Available: https://research.google.com/audioset/dataset/index.html.

[24] W. Dai, S. Dieleman, C. Johnson, and B. Recht, "Acoustic event detection based on deep neural networks using raw waveforms," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Accessed: 2025-02-15, IEEE, 2017, pp. 125–129. [Online]. Available: https://arxiv.org/abs/1610.00087.

[25] J. Lee, J. Kim, J. Park, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Accessed: 2025-02-15, IEEE, 2017, pp. 121–125. [Online]. Available: https://arxiv.org/abs/1703.01789.

[26] Y. Gong, Y.-A. Chung, and J. Glass, *Ast: Audio spectrogram transformer*, 2021. arXiv: 2104.01778 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2104.01778.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[28] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.

[29] F. Schmid, K. Koutini, and G. Widmer, *Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation*, 2023. arXiv: 2211.04772 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2211.04772.

[30] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[31] S. Chen, Y. Wu, C. Wang, *et al.*, *Beats: Audio pre-training with acoustic tokenizers*, 2022. arXiv: 2212.09058 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2212.09058.

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.

[33] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers," in *Proc. Interspeech 2023*, 2023, pp. 2798–2802. DOI: 10.21437/Interspeech.2023-2193.

[34] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[35] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[36] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, *Streaming audio transformers for online audio tagging*, 2024. arXiv: 2305.17834 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2305.17834.

[37] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, *Eat: Self-supervised pre-training with efficient audio transformer*, 2024. arXiv: 2401.03497 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2401.03497.

[38] F. Korzeniowski and G. Widmer, *End-to-end musical key estimation using a convolutional neural network*, 2017. arXiv: 1706.02921 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1706.02921.

[39] M. W. Akram, S. Dettori, V. Colla, and G. C. Buttazzo, *Chordformer: A conformer-based architecture for large-vocabulary audio chord recognition*, 2025. arXiv: 2502.11840 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2502.11840.

[40] T. Cheng, M. Mauch, E. Benetos, and S. Dixon, "Transformer-based beat tracking with low-resolution encoder and high-resolution decoder," in *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, 2023, pp. 55–62. [Online]. Available: https://archives.ismir.net/ismir2023/paper/000055.pdf.

[41] N. Thapa and J. Lee, "Dual-path beat tracking: Combining temporal convolutional networks and transformers in parallel," *Applied Sciences*, vol. 14, no. 24, p. 11 777, 2024. DOI: 10.3390/app142411777. [Online]. Available: https://www.mdpi.com/2076-3417/14/24/11777.

[42] Y. Gao, X. Zhang, and W. Li, "Vocal melody extraction via hrnet-based singing voice separation and encoder-decoder-based f0 estimation," *Electronics*, vol. 10, no. 3, p. 298, 2021. DOI: 10.3390/electronics10030298. [Online]. Available: https://www.mdpi.com/2079-9292/10/3/298.

[43] J.-C. Wang, W.-T. Lu, and J. Chen, *Mel-roformer for vocal separation and vocal melody transcription*, 2024. arXiv: 2409.04702 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2409.04702.

[44] P. Weber, C. Uhle, M. Müller, and M. Lang, "Real-time automatic drum transcription using dynamic few-shot learning," in *Proceedings of the 5th IEEE International Symposium on the Internet of Sounds (IS2)*, 2024, pp. 1–8. DOI: 10.1109/IS262782.2024.10704130.

[45] M. Yeung, K. Toyama, T. Teramoto, S. Takahashi, and T. Kojima, *Noise-to-notes: Diffusion-based generation and refinement for automatic drum transcription*, 2025. arXiv: 2509.21739 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2509.21739.

[46] W. Liao, K. Shimada, S. Takamichi, H. Saruwatari, and Y. Saito, "Music foundation model as generic booster for music downstream tasks," *arXiv preprint arXiv:2411.01135*, 2024. [Online]. Available: https://arxiv.org/abs/2411.01135.

[47] A. Babu, C. Wang, A. Tjandra, *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[48] Y. Zhang, W. Han, J. Qin, *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[49] V. Pratap, A. Tjandra, B. Shi, *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[50] A. Omnilingual, G. Keren, A. Kozhevnikov, *et al.*, "Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages," *arXiv preprint arXiv:2511.09690*, 2025.

[51] S. A. G. Shakhadri, K. KR, and K. B. Angadi, "Samba-asr: State-of-the-art speech recognition leveraging structured state-space models," *arXiv preprint arXiv:2501.02832*, 2025.

[52] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[53] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[54] G. Chen, S. Chai, G. Wang, *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[55] P. K. O'Neill, V. Lavrukhin, S. Majumdar, *et al.*, "Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," *arXiv preprint arXiv:2104.02014*, 2021.

[56] F. M. Ramirez, L. Chkhetiani, A. Ehrenberg, *et al.*, "Anatomy of industrial scale multilingual asr," *arXiv preprint arXiv:2404.09841*, 2024.

[57] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *International Conference on Machine Learning*, PMLR, 2022, pp. 3915–3924.

[58] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020. [Online]. Available: https://arxiv.org/abs/2005.00341.

[59] J. Copet, F. Kreuk, I. Gat, *et al.*, *Simple and controllable music generation*, 2024. arXiv: 2306.05284 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2306.05284.

[60] J. D. P. J. S. K. K. F. Y. B. K. J.-C. Wang, "Stemgen: A music generation model that listens," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10446088.

[61] C. D. Shih-Lun Wu, "Music controlnet: Multiple time-varying controls for music generation," in *ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 32*, IEEE, 2024, pp. 2692–2703.

[62] H. Liu, Y. Yuan, X. Liu, *et al.*, *Audioldm 2: Learning holistic audio generation with self-supervised pretraining*, 2024. arXiv: 2308.05734 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2308.05734.

[63] A. Agostinelli, T. I. Denk, Z. Borsos, *et al.*, *Musiclm: Generating music from text*, 2023. arXiv: 2301.11325 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2301.11325.

[64] Y. Wang, S. Wu, J. Hu, *et al.*, *Notagen: Advancing musicality in symbolic music generation with large language model training paradigms*, 2025. arXiv: 2502.18008 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2502.18008.

[65] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, *Stable audio open*, 2024. arXiv: 2407.14358 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2407.14358.

[66] L. Chen, S. Bai, W. Chai, *et al.*, "Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think," *arXiv preprint arXiv:2502.20172*, 2025.

[67] P. Esser, S. Kulal, A. Blattmann, *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[68] P. Wang, S. Bai, S. Tan, *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[69] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[70] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[71] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3558–3568.

[72] K. Sun, J. Pan, Y. Ge, *et al.*, "Journeydb: A benchmark for generative image understanding," *Advances in neural information processing systems*, vol. 36, pp. 49 659–49 678, 2023.

[73] H. Zhao, X. S. Ma, L. Chen, *et al.*, "Ultraedit: Instruction-based fine-grained image editing at scale," *Advances in Neural Information Processing Systems*, vol. 37, pp. 3058–3093, 2024.

[74] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[75] E. Xie, J. Chen, J. Chen, *et al.*, *Sana: Efficient high-resolution image synthesis with linear diffusion transformers*, 2024. arXiv: 2410.10629 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2410.10629.

[76] G. Cideron, S. Girgin, M. Verzetti, *et al.*, *Musicrl: Aligning music generation to human preferences*, 2024. arXiv: 2402.04229 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2402.04229.

[77] D. Yang, J. Tian, X. Tan, *et al.*, "Uniaudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023. [Online]. Available: https://arxiv.org/abs/2310.00704.

[78] J. Gong, S. Zhao, S. Wang, S. Xu, and J. Guo, *Ace-step: A step towards music generation foundation model*, 2025. arXiv: 2506.00045 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2506.00045.

[79] Y. Sung, K. John, S. H. Yoon, and H. Seifi, "Hapticgen: Generative text-to-vibration model for streamlining haptic design," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25, Association for Computing Machinery, 2025. DOI: 10.1145/3706598.3713609. [Online]. Available: https://doi.org/10.1145/3706598.3713609.

[80] S. Liu, A. S. Hussain, Q. Wu, C. Sun, and Y. Shan, "Mumu-llama: Multi-modal music understanding and generation via large language models," *arXiv preprint arXiv:2412.06660*, vol. 3, no. 5, p. 6, 2024.

[81] Y. Li, R. Yuan, G. Zhang, *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.

[82] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[83] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," *arXiv preprint arXiv:2305.10790*, 2023.

[84] J. Liu, C. Li, Y. Ren, F. Chen, P. Liu, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," *arXiv preprint arXiv:2105.02446*, 2021.

[85] The Movie Gen team @ Meta, *Movie gen: A cast of media foundation models*, 2024. arXiv: 2410.13720 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2410.13720.

[86] J. Gardner, S. Durand, D. Stoller, and R. M. Bittner, *Llark: A multimodal instruction-following language model for music*, 2024. arXiv: 2310.07160 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2310.07160.

[87] H. Touvron, L. Martin, K. Stone, *et al.*, *Llama 2: Open foundation and fine-tuned chat models*, 2023. arXiv: 2307.09288 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2307.09288.

[88] SVC-Develop-Team, *So-vits-svc: SoftVC VITS Singing Voice Conversion*, https://github.com/svc-develop-team/so-vits-svc, Accessed: Dec 2025, 2023.