Multimodal Video Generation Models with Audio: Present and Future

Anonymous authors
Paper under double-blind review

Abstract

Video generation models have advanced rapidly and are now widely used across entertainment, advertising, filmmaking, and robotics applications such as world modeling and simulation. However, visual content alone is often insufficient for realistic and engaging media experiences—audio is also a key component of immersion and semantic coherence. As AI-generated videos become increasingly prevalent in everyday content, demand has grown for systems that can generate synchronized sound alongside visuals. This trend has driven rising interest in **multimodal video generation**, which jointly models video and audio to produce more complete, coherent, and appealing outputs. Since late 2025, a wave of multimodal video generation models has emerged, with releases including Veo 3.1, Sora 2, Kling 2.6, Wan 2.6, OVI, and LTX 2. As multimodal generation technology advances, its impact expands across both daily consumer and industrial domains—revolutionizing daily entertainment while enabling more sophisticated world simulation for training embodied AI systems. In this paper, we provide a comprehensive overview of the multimodal video generation model literature covering the major topics: evolution and common architectures of multimodal video generation models; common post-training methods and evaluation; applications and active research areas of video generation; limitations and challenges of multimodal video generation.

1 Introduction

Video generation has advanced rapidly in recent years, with models now capable of producing realistic, high-definition outputs with strong temporal consistency and visual aesthetics Li et al. (2025a); Hu et al. (2025); Wan et al. (2025); Yang et al. (2024). However, a critical dimension of human perception has remained largely absent from previous generation of generated videos: ***sound***. In real-world experience, visual and auditory sensory information are deeply interdependent from footsteps accompany walking, the ripple of tides against the shore, and dialogue synchronizes with lip movement McGurk and MacDonald (1976); Jahncke et al. (2015). The absence of contextually appropriate audio diminishes immersion, causing videos to feel incomplete and less natural to viewers. This principle is well illustrated by the history of cinema itself: the earliest films were entirely silent, yet the industry rapidly evolved to incorporate synchronized sound, high-fidelity audio, and eventually immersive theater acoustics, each advancement driven by the fundamental recognition that visual storytelling is inseparable from its auditory counterpart Sergi (2013); Babbar (2024); Goncalves et al. (2024a). Multimodal video generation, the joint synthesis of video and semantically aligned audio, has recently emerged as a distinct and rapidly growing research direction, addressing this gap with video-audio fusion architectures and new available training data. Unlike traditional video generation, which only generates visual outputs Yang et al. (2024); Zheng et al. (2024); Peng et al. (2025), multimodal video generation must solve fundamentally different challenges: *cross-modal temporal alignment* Haji-Ali et al. (2025); Wang et al. (2024), *semantic coherence between audio and visual streams* Ruan et al. (2023a); Luo et al. (2023a), and *the generation of plausible soundscapes that adapt to scene dynamics* Lee et al. (2025); Chen et al. (2025a). This makes multimodal video generation not merely an extension of video synthesis, but a qualitatively different problem requiring different architecture designs and training paradigms. The release of several state-of-the-art models in late 2025 and early 2026, including Sora 2 Liu et al. (2024a);

OpenAI (2025a), Veo 3.1 Wiedemer et al. (2025), Grok 4 xAI (2025), Wan 2.6 Wan et al. (2025), Kling 2.6 Kuaishou Technology (2025), OVI Low et al. (2025), and LTX 2 HaCohen et al. (2026), has revealed this paradigm shift: video generation is increasingly expected to be multimodal by default. This trend is driven by growing demand across diverse application and active research domains, from advertisement production Hu et al. (2025); Anantrasirichai et al. (2026) and short film creation Zhang et al. (2025a); Huang et al. (2025a); Leininger et al. (2025) to audio-visual video editing Team et al. (2025); Guo et al. (2025); Liu et al. (2025a); Ishii et al. (2025) and social media entertainment Anderson and Niu (2025); Ye et al. (2025). Designers, researchers, and industry practitioners alike are actively exploring both the capabilities and the underlying architectures of these systems Ma et al. (2025a). While several existing surveys comprehensively review video generation Wang et al. (2025a); Ma et al. (2026); Hayawi and Shahriar (2025); Elmoghany et al. (2025); Bhagwatkar et al. (2020), they predominantly focus on the visual modality alone. This paper specifically addresses multimodal video generation with sound.

We provide a systematic review of the most recent multimodal video generation that includes: 1. foundations and evolution of architectures of modern multimodal video generation; 2. common post-training methods and evaluation strategies 3. applications and active research areas; 4. limitations of current multimodel video generation models . Our literature review aims to provide the most up-to-date overview of multimodal video generation.

2 Components of Multimodal Video Generation Architectures

Unlike traditional visual-only generation, multimodal video generation aims to model the joint distribution of visual frames and audio waveforms HaCohen et al. (2026); Cheng et al. (2025a); Ruan et al. (2023b). The core challenge lies in synchronizing these distinct modalities within a unified architecture. In this section, we trace the architectural components of existing open-source models, specifically highlighting how they integrate audio synthesis with video dynamics. A summary of widely-adopted multimodal video diffusion models is presented in Table 1.

2.1 Variational Autoencoder (VAE)

While the Variational Autoencoder (VAE) Kingma and Welling (2013) established a foundational architecture for probabilistic generative modeling, its primary usage has shifted. Originally it served as a standalone video generator, modern VAEs serve as robust perceptual compression stages that enable efficient training and modality fusion for multimodal video generation models, such as Latent Diffusion Models Hinton and Salakhutdinov (2006).

VAE in Video Generation. VAEs serve as a compression mechanism that transforms high-dimensional raw video data into compact latent representations. As shown in Figure 1b, a video VAE processes an input video sequence $x_{1:T}$ through a 3D encoder that incorporates both spatial and temporal convolutions to capture motion dynamics across frames. Video VAEs produce temporal parameters, temporal mean $\mu_{1:t}$ and temporal log-variance $\log \sigma^2_{1:t}$, that encode the full spatiotemporal structure of the video sequence.

The 3D encoder learns to compress the video into a probabilistic distribution in the latent space, where temporal dependencies are explicitly modeled. Following the reparameterization trick, a random noise sequence $\epsilon_{1:t}$ is sampled and combined with the temporal parameters to produce the spatiotemporal latent representation:

$$z_{1:T} = \mu_{1:t} + \sigma_{1:t} \odot \epsilon_{1:t} \tag{1}$$

where the subscript $1:t$ indicates that parameters and latents are computed for the temporal sequence rather than individual frames.

VAE as Encoder in Multimodal Video Generation. As illustrated in Figure 2, VAE is an important component in multimodal video generation architectures. In DiT-based approaches shown on the right, separate VAE encoders process video and audio modalities independently, producing modality-specific latent representations.

Model / System	Architecture	Representative Notes	Release
Proprietary Business Models			
Google Veo 3.1 Wiedemer et al. (2025)	Unknown	Support 8 Second Video, I2VA, T2VA	May 2025
OpenAI Sora 2 OpenAI (2025a)	Unknown	Max 10 Second Video, Portrait and Landscape cut, I2VA, T2VA	Sep 2025
Grok 4 xAI (2025)	Unknown	I2V	July 2025
Wan2.6 Wan et al. (2025)	Unknown	I2VA, T2VA	Dec 2025
Kling 2.6 (Kuaishou) Kuaishou Technology (2025)	Unknown	I2VA, T2VA	Dec 2025
Open-Source Models			
MM-Diffusion Ruan et al. (2023b) †	Decoupled U-Net	First Open-Sourced Multimodal Video Generation	Mar 2023
OVI † Low et al. (2025)	DiT + synchronized audio-video	Native 4K generation at 50fps; open-source foundation model	Oct 2025
LTX-2 (Lightricks)† HaCohen et al. (2026)	DiT + synchronized audio-video	New Open-Surce SoTA	Jan 2026

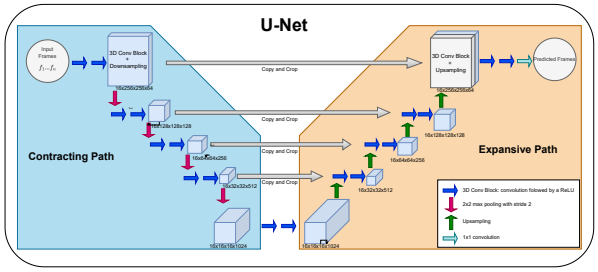
Table 1: Representative Multimodal video generation models, their architectural paradigms, and release years. Open-source models are marked with †. I2VA represents Image to Video-Audio; T2VA represents Text to Video-Audio.

For the video stream, the Video VAE Encoder compresses raw video frames into video latents that capture visual content and motion patterns. Simultaneously, the Audio VAE Encoder transforms raw audio waveforms into audio latents that encode acoustic features and temporal dynamics. These parallel encoding pathways enable the model to handle heterogeneous data types within a unified framework.

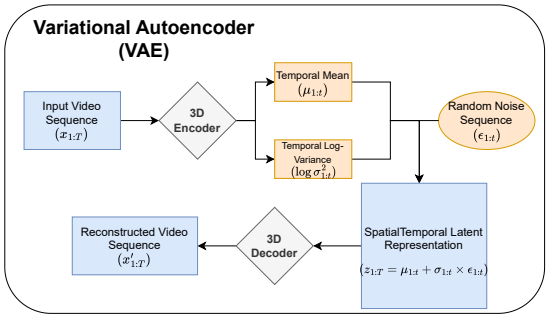
2.2 Diffusion Architecture

Diffusion models Ho et al. (2020a) become a common robust generative architecture for multimodal video generation Low et al. (2025); HaCohen et al. (2026). The core idea is to train a neural network to reverse a gradual noising process, transforming samples from a simple noise distribution back into a complex data distribution such as images or videos Ho et al. (2022a). Diffusion models learn a sequence of denoising steps that progressively remove noise, ultimately yielding a high-quality sample.

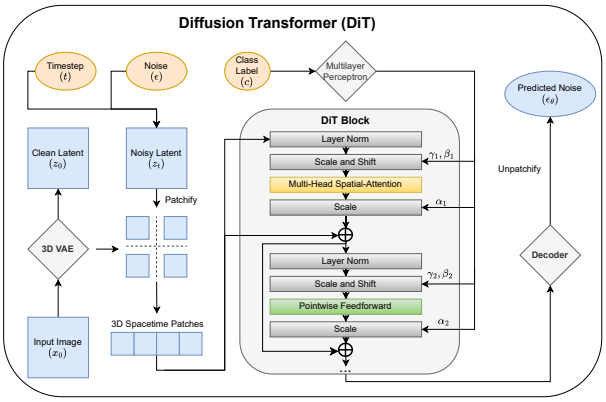
Compared to VAEs, diffusion models differ in both training objective and generative process. VAEs rely on learning a low-dimensional latent space and optimize a variational lower bound, which often leads to blurry outputs due to the imposed likelihood assumptions Kingma and Welling (2013); Rombach et al. (2022). In contrast, diffusion models directly model the data distribution through iterative denoising, avoiding explicit latent bottlenecks and shows better visual aesthetic Dhariwal and Nichol (2021); Ho et al. (2022b). While diffusion models are typically slower at inference due to their multi-step sampling process Song et al. (2022),



(a) U-Net



(b) Variational Autoencoder



(c) Diffusion Transformer (DiT)

Figure 1: Core architectures in diffusion models: (a) U-Net with skip connections for iterative denoising, (b) Variational Autoencoder (VAE) for latent space encoding and decoding, and (c) Diffusion Transformer (DiT) replacing U-Net with transformer blocks for improved scalability.

they exhibit better visual output stability during training and better coverage of complex, high-dimensional data distributions Wang et al. (2025a); Yang et al. (2023). We discuss the block architectures and components of diffusion models and their representative models below.

2.2.1 U-Net

U-Net was one of the popular backbones for diffusion models, which uses a parametric function to predict noise or denoised signals at each diffusion timestep. Originally introduced for biomedical image segmentation Ronneberger et al. (2015), its encoder–decoder structure with skip connections has proven particularly effective for generative modeling. While modern architectures have increasingly shifted toward Diffusion Transformers (DiT) for multimodal video generation, U-Net laid the foundational groundwork for video generation Ho et al. (2022a) and joint audio-video synthesis Cheng et al. (2024) and remain relevant in certain contexts.

Coupled U-Net for Joint Audio-Video Generation. MM-Diffusion Ruan et al. (2023b) introduces the first joint audio-video generation framework using a coupled U-Net architecture. Rather than using a single network, MM-Diffusion uses two parallel U-Net subnets, one for video and one for audio, that jointly denoise Gaussian noise into aligned audio-video pairs. This coupled design establishes foundations in modern architectures: (1) modality-specific processing branches that respect the distinct characteristics of audio and video signals, and (2) cross-modal attention mechanisms that enforce alignment between generated modalities. Following U-Net-based works such as MM-LDM Sun et al. (2024) extends this paradigm by operating in a shared latent space to reduce computational costs while maintaining cross-modal consistency.

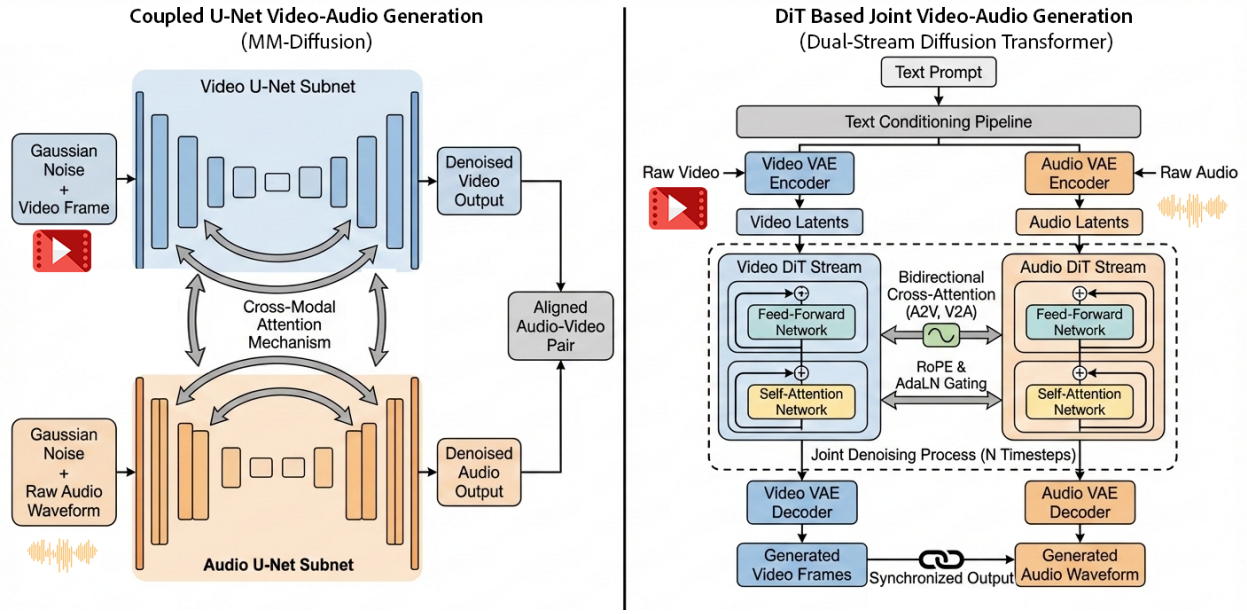


Figure 2: Evoluation of rchitectures for multimodal video generation in the open-source community.

Audio Processing Representation in Multimodal U-Nets. In joint audio-video frameworks, audio is typically represented as latent embeddings from pre-trained audio encoders, allowing the audio U-Net branch to process audio as 2D image-like representations Luo et al. (2023a); Liu et al. (2023a). MM-Diffusion processes raw audio waveforms through a dedicated U-Net subnet, while MM-LDM encodes audio into a shared latent space with video before applying the diffusion process.

Limitations and Transition to DiTs. Despite their foundational role, U-Net architectures face inherent limitations for joint audio-video generation Peebles and Xie (2023a). The locality of convolutional operations restricts long-range dependency modeling, which is crucial for maintaining semantic consistency across extended audio-video sequences Vaswani et al. (2017); Ma et al. (2024). Additionally, scaling U-Nets to higher resolutions and longer durations becomes computationally prohibitive Blattmann et al. (2023). These limitations motivate the transition toward DiT architectures, which show better scalability and more flexible *cross-modal attention mechanisms* Xu et al. (2025a).

2.2.2 Diffusion Transformer Backbones For Multimodal Video Generation

The Diffusion Transformer (DiT) is the successor to U-Net backbones, showing better scalability in model capacity, training data utilization, and generation fidelity Peebles and Xie (2023b). As illustrated in Figure 1c, DiT operates in a latent patch space: inputs are encoded by a pre-trained VAE, corrupted with noise, and processed as a sequence of tokens by a Transformer conditioned on timestep and semantic cues. The model learns to predict the noise via the standard denoising objective Ho et al. (2020b).

Unlike U-Net architectures, which rely on local convolutional inductive biases that struggle with distant dependencies Ronneberger et al. (2015); Ho et al. (2020b), Transformer backbones leverage self-attention to capture global spatiotemporal reasoning Peebles and Xie (2023b); Arnab et al. (2021). This global context is particularly vital for joint audio-video synthesis, where visual events (e.g., an explosion) must align perfectly with auditory signals (the sound) across long temporal windows. To manage the computational cost of this cross-modal alignment, architectures typically use either factorized spatial–temporal attention Ho et al. (2022c); Bertasius et al. (2021) for efficiency or full spatiotemporal attention Arnab et al. (2021); Tong et al. (2022) for maximal expressiveness.

The Shift to Native Audio-Visual Synthesis. By 2026, the industry has shifted toward native multimodal generation, driven by the realization that separate generation pipelines sever the natural correlation

between sight and sound. The release of OVI (2025) Low et al. (2025) and LTX-2 (2026) HaCohen et al. (2026) marked a turning point as the first open-source foundation models to generate synchronized video and audio using the DiT architecture, following proprietary successes like Sora 2 and Veo 3. Instead of training video and audios separately, existing open-source models use pre-trained video and audio encoders to pretrain jointly to fuse audio and video modalities to fit lip-sync (Figure 2). Below, we detail the specific architectural components these models use to unify audio and video generation.

VAE encoders. VAE encoders are generally used for training and conditional generation Diederik and Max (2019). In joint video-audio generation, there consists of a *video VAE encoder* Yu et al. (2024)and an *audio VAE encoder* Liu et al. (2024b). The Video VAE Encoder compresses raw video frames into compact video latents. A causal VAE architecture is used to to maintain temporal consistency. The Audio VAE Encoder Converts audio (via Mel Spectrogram) into audio latents. Also causal to preserve temporal structure.

Text Conditioning Pipeline. The text conditioning pipeline includes a Text Encoder, a pretrained language model or a multimodal encoder (like T5 Raffel et al. (2020) or CLIP Radford et al. (2021)) that tokenizes and encodes the text prompt; Feature Extractor refines the raw text encoder outputs into features better suited for conditioning; Text Connectors projects text embeddings to be compatible with both the audio and video streams

Dual-Stream Diffusion Transformer Fusion serves as the core mechanism enabling video and audio streams to communicate bidirectionally. Each stream contains self-attention for intra-modal coherence, text cross-attention (T2V, T2A) for semantic conditioning from the shared text embedding, and a feed-forward network. The key part is cross-modal communication: A2V cross-attention allows video queries to attend to audio keys/values, while V2A cross-attention does the reverse, enabling each modality to inform the other. This cross-attention uses Temporal 1D Rotary Positional Encoding (RoPE) Su et al. (2024) on queries and keys for temporal alignment across modalities, AdaLN (Scale, Shift) conditioned on each stream's diffusion timestep, and gating mechanisms to regulate cross-modal information flow.

Joint Inference. Both modalities start from independent Gaussian noise but denoise jointly in parallel over N shared timesteps. At each step, the dual-stream transformer processes both latents simultaneously, with bidirectional cross-attention allowing audio and video to communicate with each other. After denoising completes, separate VAE decoders reconstruct the final video frames and audio waveform.

2.2.3 Potential Future Architectural Design: Mixture of Experts and Autoregressive Generation

Mixture of Experts (MoE). As the video diffusion model component scales to billions of parameters, computational costs increase proportionally during both training and inference Wan et al. (2025); NVIDIA (2025). Mixture of Experts (MoE) architectures mitigate this challenge by introducing sparse activation, in which only a subset of model parameters is activated for a given input Shazeer et al. (2017).

While standard video-only MoE architectures route tokens based on spatial-temporal features (allocating experts to specialize in specific visual textures or motion dynamics Riquelme et al. (2021)), Audio-video MoEs must handle multimodal tokens with distinct sampling rates and semantic granularities, with the most recent works shown in Uni-MoE-2.0-omni Li et al. (2025b).

Token-level MoE Lepikhin et al. (2020); Fedus et al. (2022); Riquelme et al. (2021); Li et al. (2025c); Dai et al. (2024). In token-level MoE diffusion transformers, the standard feed-forward network (FFN) layers are replaced with multiple parallel expert networks and a learned routing function. For each input token $\mathbf{h}_i$, a router computes a distribution over experts and selects the top-k experts g_i (typically $k = 1$ or $k = 2$) to process that token:

$$g_i = \text{Top-}k\big(\text{Softmax}(W_r \mathbf{h}_i)\big),$$
$$\text{FFN}(\mathbf{h}_i) = \sum_{e \in g_i} p_{i,e} \cdot E_e(\mathbf{h}_i),$$

where E_e denotes the e-th expert and $p_{i,e}$ are the routing weights. This design enables large model capacity while keeping per-token computation tractable. Token-level MoE is particularly effective when token complexity varies spatially or temporally, such as regions with intricate textures or complex motion patterns. Representative examples include SegMoE Ortigossa and Segal (2026) and Race-DiT Yuan et al.

(2025a), which demonstrate improved sample quality and reduced Floating Point Operations (FLOPs) through learned, dynamic routing.

Timestep-level MoE Balaji et al. (2023); Zhuang et al. (2025); Cheng et al. (2025b). Beyond token-level routing, diffusion models show a natural form of multi-task structure across denoising timesteps: early timesteps (high noise phases) focus on global layout and motion planning, while later timesteps (low-noise phases) refine fine-grained appearance and temporal details. Timestep-level MoE uses this property by assigning different experts to distinct noise phases rather than routing individual tokens.

Formally, let $\epsilon_{\theta_e}(x_t, t)$ denote the denoising network of expert e at timestep t. A hard-gated timestep MoE routes computation based solely on the diffusion timestep:

$$e(t) = \begin{cases} \text{high}, & t > t_{\text{switch}}, \\ \text{low}, & t \leq t_{\text{switch}}, \end{cases}$$

$$\epsilon_\theta(x_t, t) = \epsilon_{\theta_{e(t)}}(x_t, t).$$

Only a single expert is active at each timestep, resulting in sparse activation without introducing additional routing overhead. This formulation preserves inference cost per step while enabling strong specialization across denoising stages.

The state-of-the-art video-only generative model Wan 2.2 Wan et al. (2025) adopts such a timestep-specialized MoE design, maintaining separate model weights for high-noise and low-noise phases. During sampling, early denoising steps are handled by a high-noise expert that captures global structure and motion, while later steps are processed by a low-noise expert that refines textures and temporal consistency. This new design of MoE architecture improves video generation quality and efficiency compared to the non-MoE version (Wan2.1), also showing a promising architectureal design direction for adopting to multimodal video generation with MoE.

Autoregressive Generation for Multimodal Video-Audio Synthesis. Autoregressive (AR) generation has emerged as a promising paradigm for unifying different modalities (video, image, auido, text, etc) together due to its natural alignment with the sequential nature of temporal media Kondratyuk et al. (2024); ai et al. (2025), and it demonstrate strong scaling abilities in Large Language Models Yang et al. (2025a); OpenAI (2024), Vision-language models Li et al. (2025d); Bai et al. (2025), and unified models Deng et al. (2025); Xu et al. (2025b;c). For broader multimodal capabilities, Unified-IO2Lu et al. (2023), BAGEL Deng et al. (2025), EMMA He et al. (2025a), DeepSeek Nanus Pro Chen et al. (2025b) demonstrate that a single autoregressive transformer can be trained to understand and generate across text, images, audio, and action by tokenizing all modalities into a shared semantic space, though it processes these as separate generation tasks rather than jointly synthesizing video and audio. Similarly, Large World Model (LWM)Liu et al. (2024c) extends autoregressive transformers to million-token contexts for video understanding and generation but primarily focuses on the visual modality. For truly joint video-audio generation, diffusion-based approaches currently dominate Polyak et al. (2024); HaCohen et al. (2026); Low et al. (2025); Ruan et al. (2023a). However, the development of unified autoregressive architectures that can jointly understand and generate both video and audio in a single forward pass remains an open and promising research direction, with potential benefits including more natural temporal coherence and simplified training pipelines compared to multi-stage or separate-model approaches Kondratyuk et al. (2024).

3 Post-Training and Evaluation

Post-training and evaluation play critical roles in improving multimodal video-audio generation models on user-specified downstream tasks, where pre-trained base models fall short Liu et al. (2025b;c). We review several fundamental post-training methods for adapting video generation models to produce synchronized audio, including parameter efficient fine-tuning (PEFT) Mangrulkar et al. (2022), audio-visual alignment modules, attention manipulation, and ControlNet-based conditioning. We present the post-training introductions below.

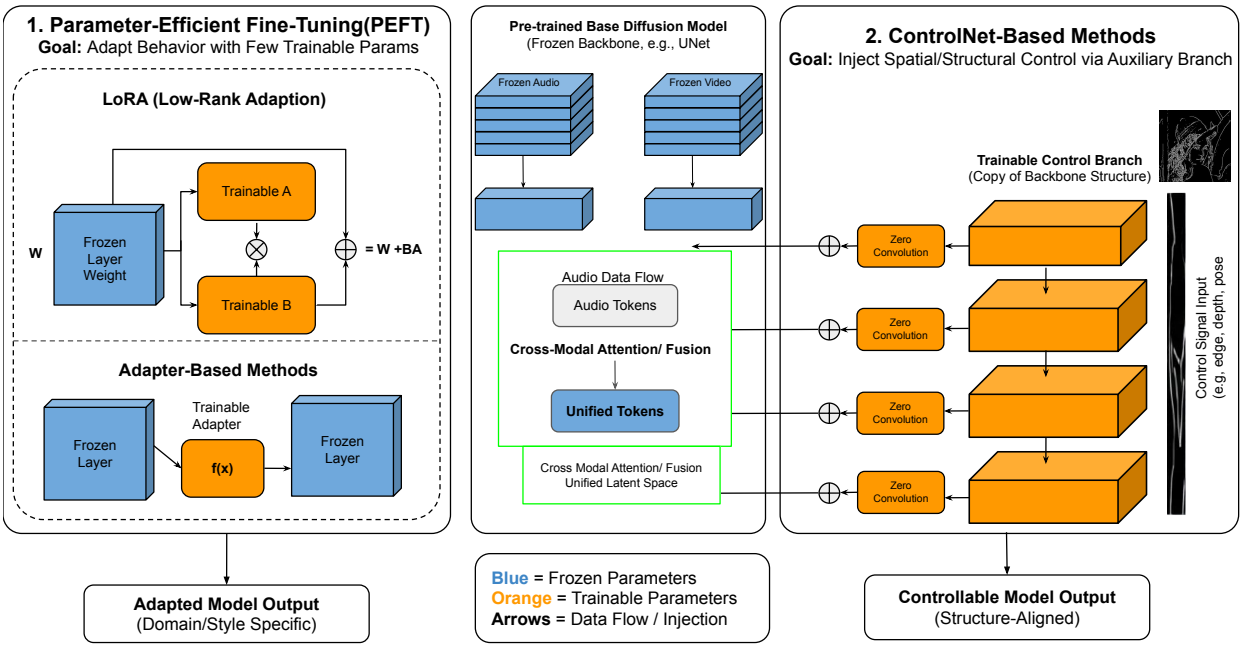


Figure 3: Exemplar post-training methods for multimodal video-audio generation.

3.1 Training Data Preparation

The post-training data preparation for multimodal video-audio generation requires carefully curated audio-video pairs with precise temporal alignment. For video-to-audio (V2A) tasks, training data consists of videos paired with their corresponding audio tracks, often with additional annotations such as sound event timestamps from datasets like AudioSet Strong Hershey et al. (2021), audio captions, or onset markers. For joint audio-video generation, synchronized multimodal data with aligned textual descriptions for both modalities is essential Cheng et al. (2025a); Zhao et al. (2025). The quantity of curated post-training data depends on the specific techniques: adapter-based methods like LoRA and semantic adapters typically require 100–1,000 video-audio pairs to generalize Zhang et al. (2024), while full-parameter fine-tuning may require millions of paired samples. Recent approaches leverage automated annotation pipelines using multimodal models to generate aligned video captions, audio captions, and speech transcriptions, ensuring temporal synchronization and semantic consistency Wang et al. (2025b).

3.2 Training-free Methods

Training-free methods operate during inference time to achieve audio-visual alignment without modifying base model weights Zhang et al. (2025b); Singer et al. (2025). These approaches enable novel multimodal capabilities on top of frozen pretrained models.

Audio-Visual Guidance. Training-free audio guidance methods steer video generation toward audio-synchronized outputs during the denoising process. By manipulating attention scores or injecting audio-derived conditioning signals at inference time, these approaches can enforce temporal alignment between generated video frames and audio events without additional training Xing et al. (2024a). Such methods are particularly useful for audio-to-video generation, where audio features guide the visual dynamics.

Synchronization Guidance. Recent work introduces synchronization guidance during sampling to strengthen audio-driven motion generation Song et al. (2025). By modifying the flow matching or diffusion loss to emphasize regions with large motion, and applying guidance that biases generation toward audio-

aligned temporal patterns, these methods improve lip synchronization and event timing without requiring model retraining.

3.3 Parameter-Efficient Fine-Tuning (PEFT)

Parameter-efficient fine-tuning (PEFT) Hu et al. (2022); Ruiz et al. (2022); Mangrulkar et al. (2022) adapts large pretrained models to multimodal audio-video tasks while introducing only a small number of trainable parameters. PEFT significantly reduces training cost and the risk of catastrophic forgetting, making it particularly attractive for adapting video or audio diffusion models to joint generation tasks.

Low-Rank Adaptation (LoRA). Low-Rank Adaptation Hu et al. (2022) injects trainable low-rank matrices into existing linear transformations, most commonly within attention layers. A weight matrix $W \in \mathbb{R}^{d\times d}$ is reparameterized as $W' = W + \Delta W$, where $\Delta W = BA$ with $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$ ($r \ll d$). For multimodal generation, LoRA has been applied to adapt text-to-audio models for video conditioning, enabling efficient V2A fine-tuning. AV-DiT Wang et al. (2025c) demonstrates this approach by inserting LoRA layers into the projection modules of frozen attention blocks to bridge the domain gap between image-pretrained DiT and audio generation. Extensions apply LoRA to cross-modal attention layers, allowing video features to modulate audio generation while preserving the base model's audio quality.

Adapter-Based Methods. Adapter-based PEFT Xing et al. (2024b) introduces lightweight modules between layers of frozen backbones. For video-to-audio generation, semantic adapters process video features through trainable cross-attention layers that condition audio generation on visual content. FoleyCrafter employs a semantic adapter that utilizes parallel cross-attention layers to condition audio generation on video features, producing realistic sound effects semantically relevant to visual content. Temporal adapters further extend this paradigm by encoding timestamp conditions to achieve precise audio-video synchronization Zhang et al. (2024). AV-DiT Wang et al. (2025d) also incorporates bottleneck MLP adapters in parallel with feed-forward networks to adapt image-based knowledge for audio modeling.

3.4 Audio-Visual Alignment Modules

A critical challenge in multimodal video-audio generation is achieving precise temporal and semantic alignment between modalities. Several specialized modules have been developed to address this challenge.

Synchronization Modules. Synchronization modules inject frame-aligned features to ensure precise audio-video temporal correspondence. MMAudio Cheng et al. (2025a) introduces a conditional synchronization module that aligns video conditions with audio latents at the frame level using aligned RoPE position embeddings and feature injection through adaptive layer normalization. The synchronization features are extracted using Synchformer Iashin et al. (2024), a self-supervised audio-visual desynchronization detector operating at 24 fps. This enables the model to capture fine-grained temporal correlations at millisecond granularity, which is critical since humans can perceive audio-visual misalignment as small as 25 milliseconds.

Onset Detection and Timestamp Conditioning. For video-to-audio generation, onset detectors predict when sound events should occur based on visual motion cues. FoleyCrafter Zhang et al. (2024) employs a timestamp detector that predicts sound and silence labels from video frames, trained with ground truth audio event timestamps from AudioSet Strong Hershey et al. (2021). The predicted timestamps are then encoded by a temporal adapter that follows the ControlNet design Zhang et al. (2023), injecting synchronization information into the audio generation backbone to ensure that generated sounds align with visual actions.

Cross-Modal Feature Fusion. Joint audio-video generation requires effective fusion of heterogeneous modality features. Approaches range from concatenating audio-video tokens in a shared latent space—UniForm Zhao et al. (2025) employs a unified single-tower DiT architecture to process concatenated audio-video tokens with task-specific noise schemes—to dual-branch architectures with cross-attention between parallel DiT streams. UniAVGen Zhang et al. (2025c) introduces asymmetric cross-modal interaction with bidirectional, temporally aligned cross-attention. Hierarchical encoders like Synchformer Iashin et al. (2024) capture fine-grained dynamic cues for temporal alignment, while semantic encoders (CLIP, SigLIP) provide global scene understanding for semantic consistency Cheng et al. (2025a); Shan et al. (2025).

Contrastive Audio-Visual Pretraining (CAVP). Diff-Foley Luo et al. (2023b) proposes contrastive audio-visual pretraining to learn temporally and semantically aligned features before training the diffusion model. CAVP uses a CLIP-like framework with both semantic contrast (maximizing audio-visual similarity within videos) and temporal contrast (emphasizing audio-visual synchronization within video segments). The CAVP-aligned visual features enable the latent diffusion model to capture subtler audio-visual correlations via cross-attention modules.

3.5 Attention Injection for Audio-Visual Control

Attention injection modifies attention mechanisms to introduce audio or video conditioning signals while preserving backbone weights. Unlike PEFT that adapts through weight updates, attention injection operates directly in the attention space Yuan et al. (2025b); Cai et al. (2025a).

Cross-Attention for Modality Conditioning. A common approach introduces cross-attention layers where audio latents attend to video features (or vice versa). FoleyCrafter Zhang et al. (2024) integrates parallel cross-attention layers alongside text-based attention, allowing audio generation to be conditioned on video features without compromising text-to-audio capabilities. For audio-to-video generation, Syncphony Song et al. (2025) injects audio features via cross-attention with RoPE to enable audio-motion alignment on top of a DiT architecture.

Joint Self-Attention. Joint audio-video generation models often employ unified self-attention over concatenated audio and video tokens. MMAudio Cheng et al. (2025c) uses multimodal transformer blocks where video, text, and audio latents jointly interact, allowing bidirectional information flow between modalities during the denoising process. UniForm Zhao et al. (2025) similarly processes audio and video tokens within a unified latent space using a shared DiT, enabling the model to learn implicit correlations between visual dynamics and audio characteristics.

Attention Feature Injection. Some methods inject external attention features—such as precomputed audio embeddings or onset-aligned temporal features—into existing attention layers. Rather than adding new layers, these approaches modify attention computation by blending or biasing query-key affinities. MMAudio Cheng et al. (2025a) uses this approach for temporal synchronization, where Synchformer-derived features are injected through adaptive layer normalization to enforce audio-video alignment at specific timestamps. AV-DiT Wang et al. (2025d) facilitates feature interaction between audio and visual modalities by pooling video tokens temporally and concatenating them with audio tokens in a shared attention block augmented with LoRA.

3.6 ControlNet-Based Methods for Audio-Visual Generation

ControlNet-based approaches Zhang et al. (2023); Chen et al. (2025c); Gu et al. (2025) extend pretrained diffusion models with conditional branches to enable fine-grained control over multimodal generation. For video-to-audio synthesis, ControlNet augments audio generation backbones with video-conditioned control networks.

Temporal ControlNet for V2A. FoleyCrafter's Zhang et al. (2024) temporal adapter follows the ControlNet design, duplicating the UNet encoder structure and introducing zero-initialized connections that inject temporal control features into the audio generation process. The adapter takes timestamp masks as input and adds residual control signals to the original UNet, enforcing precise temporal alignment between generated audio and video events. During training, only the replicated UNet blocks are updated using the same optimization objective as the diffusion model.

Multi-Stream Temporal Control. For complex audio-visual scenarios involving speech, sound effects, and music, multi-stream ControlNet architectures process different audio components separately. MTV Weng et al. (2025) introduces a Multi-Stream Temporal ControlNet with interval streams for speech and effects tracks (controlling lip motion and event timing) and holistic streams for music (controlling visual mood). The interval stream employs interval interaction blocks to understand each track individually, while the holistic stream extracts features using a holistic context encoder that serves as style embeddings applied uniformly to all frames.

Video Feature Injection via ControlNet. ControlNet can embed video characteristics into text-to-audio synthesis by processing visual features through a control branch. FoleyCrafter Zhang et al. (2024) extracts video frames using pretrained visual encoders and injects the resulting features into the audio diffusion backbone. HunyuanVideo-Foley Shan et al. (2025) extends this by using Representation Alignment (REPA) to align intermediate DiT representations with frame-level audio features from pretrained self-supervised models, enhancing both semantic and acoustic modeling.

3.7 Common Evaluation Practices

Evaluating joint-video-audio generation is hard, with two main common evaluation practices for joint video-audio generation models: *quantitative* and *qualitative*. Quantitative evaluations for text-to-video-audio (T2VA), video-to-audio (V2A), and related multimodal generation tasks use automated metrics to assess the quality of generated content across both modalities Ruan et al. (2023a); Kilgour et al. (2019a); Unterthiner et al. (2019); Wu* et al. (2023), while qualitative evaluations rely on human judgments and rubrics to rate perceptual quality, synchronization, and semantic coherence Luo et al. (2023b); Huang et al. (2024a; 2025b); Zheng et al. (2025); Liu et al. (2023b).

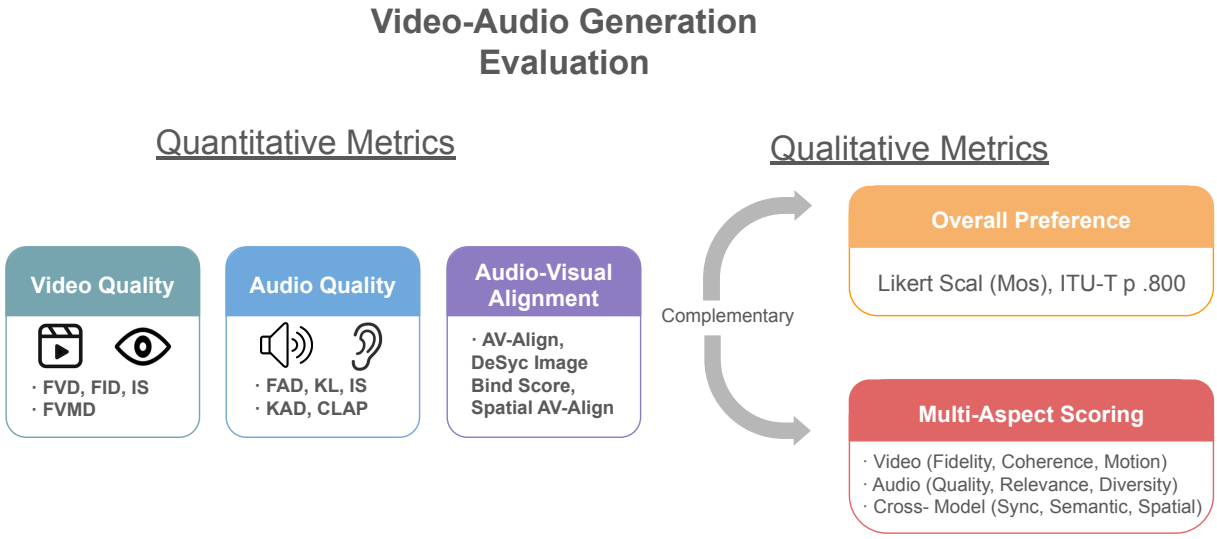


Figure 4: Multimodal Evaluation Common Practices.

3.7.1 Quantitative Evaluations

Joint video-audio evaluation requires metrics that assess each modality independently as well as their cross-modal alignment. These metrics are in three categories: video quality, audio quality, and audio-visual alignment.

Video Quality Metrics. Common metrics include *Fréchet Video Distance (FVD)*, a reference-based matric which measures the distribution distance between generated and real video features using I3D network embeddings pretrained on Kinetics Kay et al. (2017), capturing both spatial quality and temporal coherence Unterthiner et al. (2019). *CLIPScore* Hessel et al. (2021) is a reference-free metric that computes the cosine similarity between CLIP Radford et al. (2021) visual embeddings of generated video and the textual embedding of the input prompt, thereby measuring text-video semantic alignment. Originally proposed for image captioning evaluation, it has been widely adopted in video generation by averaging frame-level CLIP similarities across all frames of a generated video.

VBench series Huang et al. (2024a; 2025b); Zheng et al. (2025) provide a more fine-grained evaluation by decomposing video generation quality into many dimensions. Representative and popular dimensions

include e.g, *Dynamic Degree*, which measures the extent of motion present in the generated video; *Motion Smoothness*, measures the motion priors of a video frame interpolation model to assess whether generated motions are physically plausible; *Aesthetic Quality* measures the artistic and beauty value perceived by humans towards each video frame using the LAION aesthetic predictor Wu* et al. (2023).

VBench++Huang et al. (2025b) extends this framework to image-to-video evaluation and model trustworthiness assessment, while VBench-2.0Zheng et al. (2025) introduces 18 additional dimensions targeting intrinsic faithfulness, including physics-based realism, commonsense reasoning, and human fidelity, leveraging VLM-based evaluation pipelines to assess capabilities where earlier metrics have begun to saturate Li et al. (2025e); Guan et al. (2024).

Audio Quality Metrics. Fréchet Audio Distance (FAD) compares generated and reference audio distributions in an embedding space, typically using VGGish features Kilgour et al. (2019a). Kullback-Leibler (KL) divergence measures the distribution similarity between generated and target audio classification outputs. Inception Score (IS) adapted for audio evaluates sample diversity and quality. Kernel Audio Distance (KAD) addresses FAD's Gaussian assumption limitations through an MMD-based approach Chung et al. (2025a). For text-conditioned generation, CLAP Score measures text-audio alignment via cosine similarity in the CLAP embedding space Wu* et al. (2023). Recent advancements (2024–2025) have introduced reference-free and model-based metrics to address the reliance on ground-truth datasets. PAM leverages Audio-Language Models to score quality via text prompting Deshmukh et al. (2024), while Audiobox Aesthetics decomposes quality into specific axes like *Production Quality* and *Content Enjoyment* Vyas et al. (2023). Additionally, MAUVE Audio Divergence (MAD) has emerged as a non-Gaussian alternative to FAD for better distribution profiling Zhang et al. (2025d).

Audio-Visual Alignment Metrics. Evaluating the synchronization and semantic coherence between generated audio and video requires specialized metrics that assess whether sounds correspond to visible events, occur at appropriate times, and originate from correct spatial locations Girdhar et al. (2023a); Goncalves et al. (2024b); Cheng et al. (2025a). AV-Align Yariv et al. (2023a) measures semantic correspondence between audio and video streams, while DeSync Feng et al. (2025a) quantifies temporal misalignment in seconds using the Synchformer model. For cross-modal evaluation in a shared embedding space, ImageBind Score (IB) Girdhar et al. (2023b) computes the cosine similarity between audio and video representations projected into ImageBind's joint embedding space. FAVD Kilgour et al. (2019b) extends the Fréchet distance framework to joint audio-visual features, capturing distributional similarity of generated audio-video pairs against real data. Finally, Spatial AV-Align Yariv et al. (2023a) evaluates spatial coherence by combining object detection with sound event localization and detection (SELD) to verify that generated sounds originate from the correct locations in the visual scene.

3.7.2 Qualitative Evaluation

Human evaluation is essential for joint video-audio generation, as automatic metrics often fail to capture perceptual synchronization quality and semantic coherence across modalities Huang et al. (2024a); Liu et al. (2025c); Ruan et al. (2023a). Evaluations typically involve in two paradigms:

Overall Preference. Annotators select the better sample between model outputs or rate overall quality on a Likert scale (typically 1–5), following the Mean Opinion Score (MOS) protocol standardized in ITU-T P.800 MET. For audio-visual content, annotators assess the combined experience of seeing and hearing the generated output.

Multi-Aspect Scoring. Quality is decomposed into modality-specific and cross-modal dimensions (Table 2).

The PEAVS framework Goncalves et al. (2024b) provides a comprehensive protocol for perceptual evaluation of audio-visual synchrony, covering temporal offsets, speed variations, and content-level alignment. For stereo audio generation, evaluators additionally assess whether spatial audio positioning corresponds to object locations in the visual scene Zhou et al. (2020).

Paradigm	Category	Metrics / Aspects
Quantitative	Video Quality	FVD Unterthiner et al. (2018), CLIPScore Hessel et al. (2021), VBench series Huang et al. (2024b)
	Audio Quality	FAD Kilgour et al. (2019c), KL Divergence Kullback and Leibler (1951), KAD Chung et al. (2025b), CLAP Score Elizalde et al. (2022), PAM Deshmukh et al. (2024), Audiobox Aesthetics Tjandra et al. (2025)
	AV Alignment	AV-Align Yariv et al. (2023b), DeSync Zhou et al. (2025), ImageBind Score Girdhar et al. (2023b), FAVD Mo et al. (2024), Spatial AV-Align Shimada et al. (2024a)
Qualitative	Protocol	Overall Preference (MOS) itu (2016), Multi-Aspect Scoring
	Video Aspects	Visual Fidelity, Temporal Coherence, Motion Realism Huang et al. (2024b); Han et al. (2025)
	Audio Aspects	Audio Quality, Sound Relevance, Audio Diversity Kreuk et al. (2023); Vinay et al. (2022)
	Cross-Modal	AV Synchronization Goncalves et al. (2024c), Semantic Coherence, Spatial Consistency Shimada et al. (2024b)
	Prompt	Text Alignment Han et al. (2025)

Table 2: Evaluation Framework for Joint Video-Audio Generation

4 Applications and New Research Directions

With the emerging capabilities and scalability of joint video-audio generation models Wiedemer et al. (2025); HaCohen et al. (2026), multimodal content creation is entering a new phase where visual and auditory elements are produced synchronously rather than as separate post-production steps. Starting in 2026, the creative and commerce landscape is moving beyond the *uncanny valley* of silent or post-dubbed generated video into an era where synchronized soundscapes, dialogue, ambient audio, sound effects, and music, are generated alongside the visual stream as a unified output OpenAI (2025a); HaCohen et al. (2026); xAI (2025). The current market also reflects this shift. Proprietary models such as Sora 2 OpenAI (2025b), Veo 3.1 Google AI Studio (2025); Google DeepMind (2025), Wan2.6 AtlasCloud (2025); Higgsfield (2025), and Grok 4 xAI (2026) have begun integrating native audio generation pipelines, while creator platforms like Kling AI Kling AI (2025), Runway Gen-4 Runway (2025a; 2026), Pika Pika (2024; 2026), and Doubao Yahoo Finance (2024); WIRED (2025) on TikTok increasingly provide joint audiovisual outputs that eliminate the traditional separation between video editing and sound design. Creator-friendly interfaces such as ComfyUI ComfyUI Blog (2025; 2026); gitmylo (2026) are similarly incorporating audio-aware workflows. In this section, we summarize popular applications of joint video-audio generation and how the addition of synchronized audio transforms use cases that were previously limited to silent video for both personal users and the industry. We then survey active research and engineering directions specific to the audiovisual setting, including temporal audio-visual synchronization, spatially grounded sound generation, and multimodal coherence at scale.

4.0.1 Personal User Applications

Social Media Content Creation and Entertainment. One of the most widespread applications is short-form video generation for social media platforms. Users can now generate engaging audiovisual content directly from text prompts or reference images, without the need for cameras, actors, sound recording equipment, or complex editing software Zheng et al. (2024). OpenAI's *Sora 2* OpenAI (2025a), Tiktok's Doubao Yahoo Finance (2024) Kuaishou (Kling AI) Kuaishou (2024) allow users to create and remix stylized short clips with matched soundtracks from images or text prompts and post to their own media platform. The addition of audio-driven animation to social media platforms, where an audio clip can drive facial expressions and lip movements in the animated photograph, deepens the sense of presence and realism, while simultaneously raising ethical considerations around consent and authenticity Rosenberg (2025).

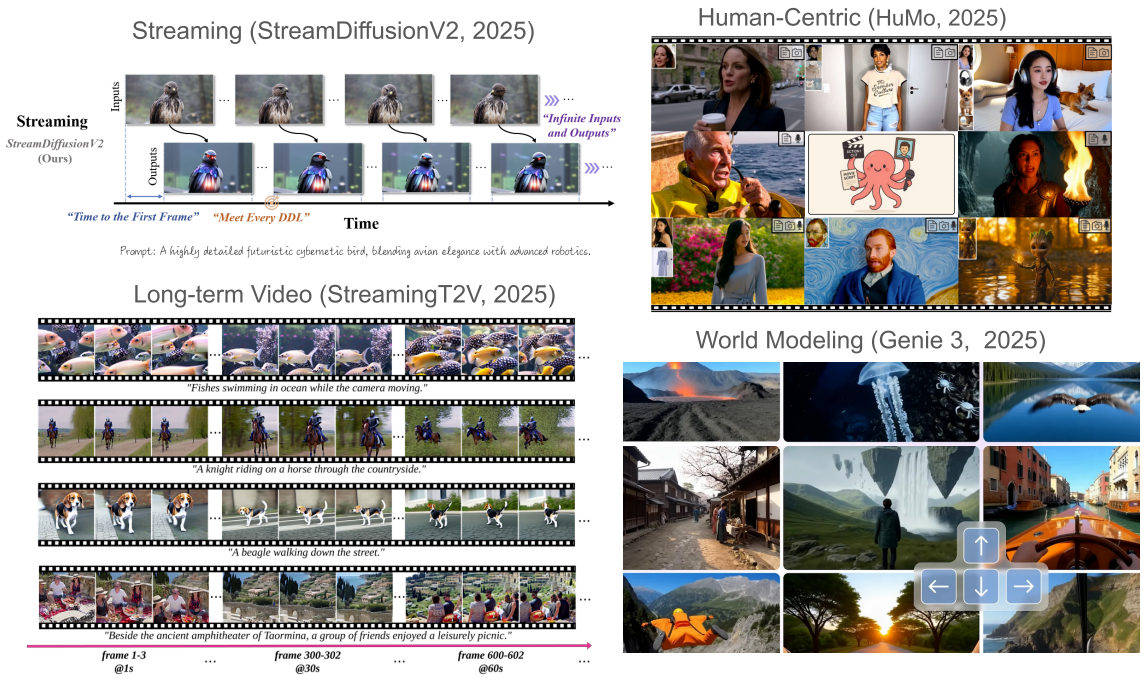


Figure 5: Main Stream Multimodal Video Generation Research Areas. Streaming video generation to generate real-time videos; Human-centric multimodal video generation that generates mainly based on human characters with voices; long-term video generation where to maintain temporal consistency for long generation; world modeling for generating simulation of real-world and sound.

Multimodal Video Personalization and Avatar Creation. Multimodal video generation has enabled sophisticated personalization through virtual avatars and digital humans. Recent advances in audio-conditioned generation allow these avatars to speak and emote with natural lip synchronization and co-speech gestures driven directly from audio input, reducing the need for manual motion capture or animation rigging. Audio-driven portrait animation methods such as *Hallo* Xu et al. (2024a), *EchoMimic* Chen et al. (2024); Meng et al. (2024), and *EMO* Tian et al. (2024; 2025) allow users to create talking avatars from a single photograph, generating synchronized lip movements and facial expressions from an audio track. ByteDance's *OmniHuman-1* Lin et al. (2025a) substantially advances this paradigm by supporting full-body audio-driven animation, including talking, singing, and gestural co-speech motion, from a single reference image and an audio signal.

4.0.2 Commercial Applications of Video Generation

Products and models are being developed, but they cannot sustain growth until they are deployed in real-world use cases, generate revenue, and receive feedback from actual users Maslej et al. (2024). As joint video-audio generation models advance, commercial adoption is accelerating across advertising, media production, and enterprise workflows, with native audio capabilities emerging as a key differentiator that reduces post-production overhead and enables fully immersive audiovisual content from a single generative pipeline.

Advertisement and Marketing. Google demonstrated this capability by creating full TV commercials with *Veo 3*, including a holiday advertisement that aired on television and in cinemas, with synchronized dialogue, sound, and music generated natively alongside the visuals Sullivan (2025).

Film and Video Production. A representative example is at Adobe MAX 2025, Adobe introduced *Firefly* as an all-in-one creative AI studio with multimodal generation spanning image, video, and audio Adobe Newsroom (2025). The platform includes *Generate Soundtrack* (powered by the commercially safe Firefly

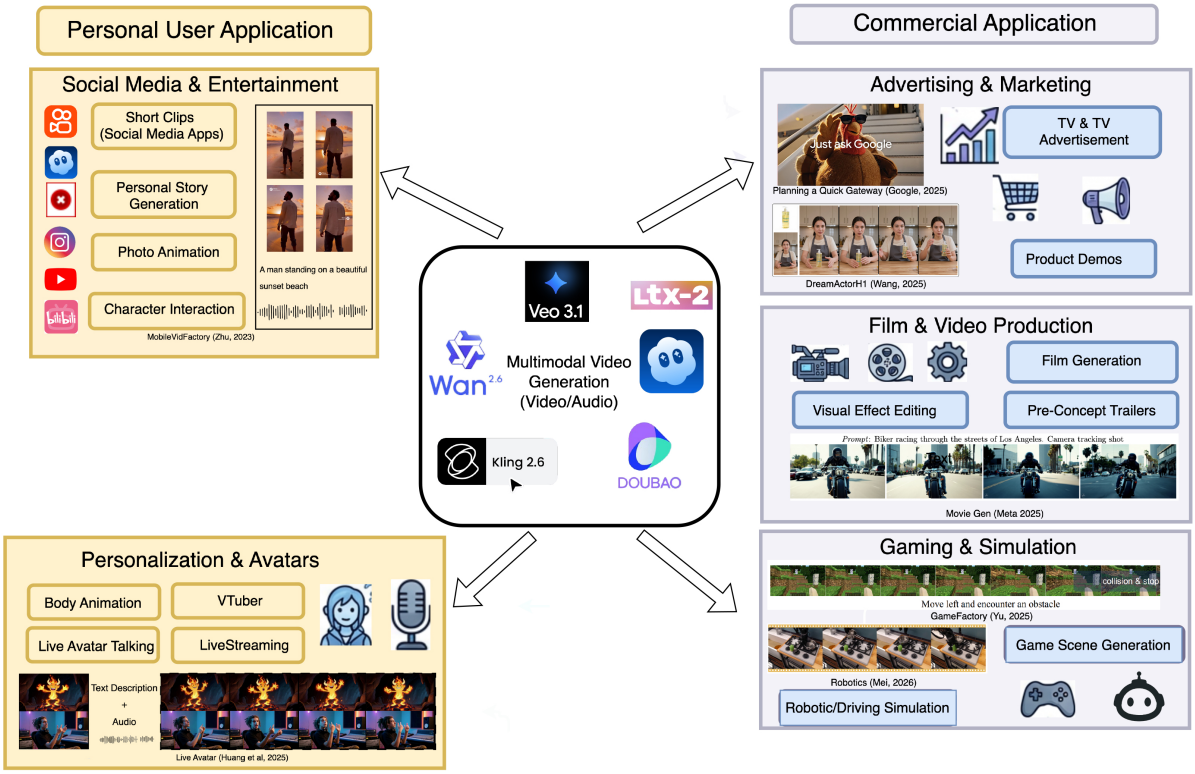


Figure 6: Widely used applications of multimodal video generations. They mainly focus on personal and business applcations.

Audio Model) for creating fully licensed, studio-quality instrumental tracks synchronized to video footage, and *Generate Speech* for crystal-clear voiceovers Chedraoui (2025).

Gaming and Interactive Entertainment. *ElevenLabs* ElevenLabs (2025) provides AI voice generation for NPCs and player characters, supporting real-time text-to-speech with low latency, voice cloning for consistent character voices, which enables games to feature adaptive, personalized dialogue without pre-recorded voice acting for every scenario.

The addition of synchronized audio to video generation pipelines marks a fundamental shift from silent visual generation to complete audiovisual content creation, reducing production overhead and enabling new creative possibilities across commercial applications.

5 Active Research Areas in Multimodal Video Generation

With the advancement of multimodal video generation, there exist many active research areas to explore. Specifically, these areas include Streaming Multimodal Video Generation, Human-Centric Multimodal Generation, and Long Multimodal Video Generation. We summarize the most recent active research directions and their implications for joint video-audio synthesis below.

5.1 Streaming Multimodal Video Generation

Streaming video generation produces frames incrementally in real time with strict latency and memory constraints Kodaira et al. (2025a). Extending this paradigm to multimodal generation requires maintaining audio-visual synchronization during incremental synthesis, which is essential for live avatars, interactive

agents, and real-time content creation Huang et al. (2025c). However, this introduces new challenges such as *causal* context degrades generation quality overtime; *memory efficiency* becomes a bottleneck for maintaining separate Key-Value (KV) caches for long-duration video and audio streams; *cross-modal alignment* without future look-ahead is difficult, where minor latency variations between audio and video streams can cause noticeable synchronization drift.

Causal temporal modeling refers to autoregressive generation where each frame at time t is predicted from past frames and latent states without access to future context, enabling on-the-fly, low-latency synthesis Yan et al. (2021). **CausVid** adapts a pretrained bidirectional video diffusion model into a causal autoregressive transformer, introducing distribution matching distillation to compress a slow bidirectional teacher into a fast causal student Yin et al. (2025). **MotionStream** extends causal autoregressive video generation to interactive, motion-controlled streaming using sliding-window causal attention with attention sinks and KV cache rolling Shin et al. (2025). **StreamDiffusion** provides a real-time diffusion pipeline optimized for interactive generation Kodaira et al. (2025b); Feng et al. (2025b). For multimodal streaming, these causal architectures can be extended to jointly generate synchronized audio tokens alongside video frames, requiring cross-modal attention mechanisms that operate under strict causality constraints.

Temporal consistency without teacher forcing. Streaming models must self-condition on their own previous outputs, leading to drift errors. **Self-Forcing** addresses this by simulating the inference process during training, conditioning each frame on its own previously generated outputs Huang et al. (2025d). **Reward Forcing** introduces reward-guided distribution matching distillation with an EMA-Sink mechanism for motion consistency Lu et al. (2025). **LongLive** aligns training with long-sequence inference to reduce temporal drift Yang et al. (2025b). **VideoGPA**distills scene-level geometric priors from a feedforward geometric foundation model to mitigate temporal drift Du et al. (2026). For multimodal generation, these self-conditioning strategies must account for both visual drift and audio-visual desynchronization over extended generation.

Interactivity and Reactivity. Interactive streaming responds to real-time user inputs. **PersonaLive** enables real-time portrait animation for live streaming Li et al. (2025f). **LiveAvatar** supports interactive, infinite-length avatar video generation conditioned on live audio or dialogue inputs Huang et al. (2025c). This existing research sshowcases the ability of integration of audio conditioning in streaming generation in the current pure video generation pipeline, pointing toward fully joint audio-video streaming synthesis.

5.2 Human-Centric Multimodal Video Generation

Human-centric video generation produces realistic imagery of human subjects conditioned on signals such as audio, pose, or text Hu et al. (2024a); Xu et al. (2023). This domain is inherently multimodal, as human video often requires synchronized speech, environmental sounds, and background music Xu et al. (2024b); Siyao et al. (2022). However, multimodal human-centric generation faces challenges such as *Fine-grained synchronization*, *temporal consistency*, and *computational efficiency*.

Face Animation. Face animation synthesizes realistic facial motion, such as lip articulation, expressions, and head movement, conditioned on driving signals including audio, text, or motion cues . Audio-driven methods such as MultiTalk Kong et al. (2025), InfiniteTalk Yang et al. (2025c), and FantasyTalking Wang et al. (2025e) map speech signals to mouth shapes and facial motion. More recent approaches including StableAvatar Tu et al. (2025), LongCat-Avatar Meituan LongCat Team (2025), and KlingAvatar 2.0 Kling Team et al. (2025) extend to infinite-length generation, addressing temporal drift and identity degradation. Streaming systems such as PersonaLive Li et al. (2025f), AnyTalker Zhong et al. (2025), and LiveAvatar Huang et al. (2025e) push toward real-time, low-latency generation with live audio input. Expressive models like OmniHuman-1 Lin et al. (2025b) and Sonic Ji et al. (2025) incorporate global audio perception and multimodal conditioning for richer facial dynamics Ji et al. (2025); Lin et al. (2025b).

Pose Animation. Pose animation generates full-body human motion videos from pose sequences or control signals. Methods such as Wan-Animate Cheng et al. (2025d), SCAIL Yan et al. (2025a), and Follow Your Pose map target poses to realistic human videos, while SteadyDancer Zhang et al. (2025e) and AnimateDiff Guo et al. (2024) enhance temporal fidelity through diffusion-based architectures. Integrating audio with pose animation enables music-driven dance generation and speech-driven gesture synthesis, requiring models to

learn correlations between audio rhythm/prosody and body movement dynamics Yi et al. (2023); Tseng et al. (2023).

Customization. Customization enables controllable identity, appearance, and motion in generated videos. Models such as HuMo Chen et al. (2025d), FFGo Wu et al. (2025), and Stand-In Xue et al. (2025) transfer target identities onto video sequences, while HunyuanCustom Hu et al. (2024b), HyperMotion Xu et al. (2025d), and Phantom Liu et al. (2025d) support fine-grained control over gestures and motion. For multimodal generation, customization extends to voice cloning and audio style/identity control, enabling consistent audio-visual identity across generated content Qiang et al. (2026); He et al. (2025b).

5.3 Long Multimodal Video Generation

Long video generation produces temporally coherent videos extending over minutes to unbounded lengths Elmoghany et al. (2025). For multimodal content, this requires maintaining audio-visual synchronization and semantic consistency over extended durations. However, scaling generation to extended durations introduces critical bottlenecks with *Temporal drift* and *computational efficiency* as primary concerns.

Single-Shot Generation. Single-shot methods generate long videos frame-by-frame in a continuous manner using latent caching, rolling attention windows, or self-conditioning. Representative works include FramePack Zhang et al. (2025f), StreamingT2V Henschel et al. (2025), Rolling Forcing Liu et al. (2025e), LongLive Yang et al. (2025b), SVI Li et al. (2025g), Self-Forcing++ Cui et al. (2025), LongVie Gao et al. (2025), LongCat-Video Meituan LongCat Team et al. (2025), FreeLong Lu et al. (2024), and Infinity-RoPE Yesiltepe et al. (2025). Extending these to multimodal generation requires joint latent caching for both video and audio streams, with synchronization mechanisms that prevent cross-modal drift over long sequences.

Multi-Shot Generation. Multi-shot methods divide videos into segments, synthesizing each with context-aware conditioning for cross-boundary coherence An et al. (2025). Examples include HoloCine Meng et al. (2025), Mixture of Contexts Cai et al. (2025b), and StoryMem Zhang et al. (2025g), which use memory mechanisms for consistent narrative and style. For multimodal content, multi-shot approaches enable scene-aware audio generation where audio characteristics (ambient sounds, music, dialogue) can shift naturally across scene boundaries while maintaining global coherence Zhang et al. (2025h).

Agentic Storytelling. Recent agentic systems improve long-horizon video alignment via iterative planning and critique: VISTA Long et al. (2025) refines prompts in a loop using temporal plans, tournament selection, and specialized critics; VideoAgent Soni et al. (2024) reduces hallucinations in video plans for robot control via self-conditioning consistency and environment feedback; AutoMV Tang et al. (2025) coordinates music analysis with script/director/verifier agents to generate coherent full-length MVs and benchmark them against human experts; and ScripterAgent Mu et al. (2026) translates dialogue into executable scripts (ScriptBench) that guide cross-scene generation, evaluated with CriticAgent and a visual–script alignment metric.

5.4 Interactive Multimodal Video Generantion and World Models

World models Ha and Schmidhuber (2018) encompass a broad class of approaches that aim to capture and simulate the underlying dynamics of the real world at varying levels of abstraction. Typically, a world model maintains an internal representation of the environment state, receives actions from decision-making agents, and predicts the subsequent state conditioned on those actions. While early world models focused primarily on visual observations, real-world perception is inherently multimodal—audio carries rich information about environmental properties that are often complementary to visual cues, including spatial location of sound sources, acoustic characteristics of physical spaces, and temporal evolution of auditory events.

Based on differences in goals and methodologies, world models can be broadly divided into two categories Huang (2025): **(a) Representation World Models**, which learn abstract, semantic-level representations to predict physical events, often embedded within LLM/VLM or agentic frameworks Zhang et al. (2025i); Bolton et al. (2025); and **(b) Generative World Models**, which represent the world state as a detailed description of the physical environment, functioning as high-fidelity simulators that explicitly generate

future world states. In this survey, we primarily focus on the latter category, with emphasis on multimodal generation capabilities.

Spatial and texture awareness in audio–visual generation. A key challenge in interactive multimodal generation is producing audio that matches physical cues—spatial layout, material/texture, and resonance. Recent work makes steady progress: ELSA Devnani et al. (2024) learns spatially grounded text–audio embeddings for open-vocabulary retrieval and language-based 3D localization; Visual Acoustic Fields Li et al. (2025h) couples 3DGS with diffusion to generate and localize impact sounds in 3D; xRIR Liu et al. (2025f) generalizes RIR prediction across rooms using depth geometry plus a few reference RIRs; AV-DAR Jin and Gao (2025) renders acoustics via differentiable beam tracing guided by multi-view visuals; and ViSAGe Kim et al. (2025) synthesizes first-order ambisonics from silent video with directional guidance, outperforming two-stage spatialization pipelines.

Audio-Visual World Models. Recent work extends world models from vision-only to synchronized audio generation. **Audio-Visual World Models (AVWMs)** target action-conditioned simulation of joint audio-visual dynamics with reward prediction; AV-CDiT Wang et al. (2025f) uses a diffusion transformer with modality experts and stagewise training for balanced multimodal forecasting, enabling coherent futures for embodied AI (e.g., audio-visual navigation). Zhang and Gienger (2025) proposes a latent flow-matching world model that predicts future audio for temporally grounded planning, improving manipulation under in-the-wild sounds and music where rhythmic dynamics matter. GWM-1 Runway (2025b) is a real-time, action-conditioned "general world model" (Gen-4.5) that generates controllable frame-by-frame simulations across explorable worlds, avatars, and robotics. MovieGen Polyak et al. (2024) combines a 30B video model with a 13B audio model to produce synchronized ambient/Foley/music, learning physical and perceptual audio-visual links for realistic generation. Veo3 Veo-Team (2024) adds audio synthesis to video generation, reflecting the broader trend toward unified audio-visual simulation.

Integration with Embodied Intelligence. For embodied agents operating in real environments, multimodal world models provide critical capabilities beyond visual prediction. Audio signals enable agents to perceive occluded objects, estimate room acoustics, and anticipate events before they become visible Somayazulu et al. (2024). Joint MLLM-WM architectures are emerging where MLLMs provide semantic reasoning and task decomposition while WMs offer physics-aware simulation including audio-visual dynamics Feng et al. (2025c). Such multimodal world models support planning in environments where audio cues are essential—for example, navigating toward sound sources, predicting acoustic consequences of actions, or generating appropriate audio feedback in interactive simulations Wang et al. (2025f).

Challenges and Future Directions. Extending world models to multimodal audio-visual generation presents several challenges: (1) ensuring precise temporal synchronization between visual dynamics and audio events at millisecond granularity; (2) modeling the physical acoustics of environments including reverberation, occlusion, and spatial audio; (3) generating semantically appropriate sounds for novel visual concepts not seen during training; and (4) scaling multimodal world models to support long-horizon generation while maintaining audio-visual coherence. Addressing these challenges will be essential for building world simulators that capture the full multisensory nature of physical reality.

6 Limitation and Challenges

The field of video-audio generation is experiencing a rise in popularity and rapid advancement. State-of-the-art models, architectures, and applications are advancing at an unprecedented pace. While this review highlights pioneering works and outlines potential research directions for multimodal integration, it cannot claim to be exhaustive given the field's hype. Instead, we aim to provide a foundational overview, concluding with a critical discussion of the current limitations and remaining challenges in multimodal video generation.

Evaluation. Critical challenges still remain in multimodal video generation evaluations. There is no universally adopted metric for audio-visual synchrony or semantic alignment, leading models to report diverse and sometimes incomparable metrics. Many audio metrics (e.g., FAD Kilgour et al. (2019a)) operate on downsampled mono audio, making them insensitive to high-frequency content and stereo characteristics Liu et al. (2023a). Furthermore, the correlation between automatic metrics and human perception varies significantly

across content types and generation quality levels, necessitating continued reliance on human evaluation for comprehensive assessment Hua et al. (2025).

Model Efficient Deployment and Latency. As multimodal video generations are going to be deployed and come into our daily life, massive users will use the service for various purposes Chakraborty and Biswal (2025). Especially for interactive and real-time interaction. In such cases, challenges remain for multimodal generation, where high inference latency disrupts the experience of human-AI interaction, and the computational burden of processing synchronized audio-visual streams necessitates aggressive model compression (e.g., quantization, distillation) that often degrades temporal coherence Chern et al. (2025); Chen et al. (2025e).

Modality Fusion and Unified Generation. Although a line of works already studied the fusion and interaction between different modalities (text, video, audio) and try to unify those modalities together for synchronous generation and understanding, many challenges remain for studying the relationship between different modalities and how modalities affect each other Yan et al. (2025b). Open research questions are whether unifying these modalities could affect the final result, the training data recipe composition for different modalities, and how to *effectively mitigate modality hallucination, where the model over-relies on dominant signals (like text semantics) while ignoring subtle cues from others (like audio texture), to ensure balanced, high-fidelity generation across all streams* Wang et al. (2025g). Furthermore, designing unified tokenizers that can jointly compress video, text and audio into a shared latent space without suffering from interference or resolution loss remains a critical architectural hurdle Ma et al. (2025b).

7 Conclusion

This paper reviews the rapid shift from visual-only video synthesis toward native audio–visual generation, and provides the landscape around shared architectural building blocks, post-training strategies, evaluation protocols, and emerging application domains. We discuss how modern systems increasingly rely on scalable diffusion backbones and explicit cross-modal fusion/alignment to synchronize video dynamics with semantically and temporally consistent audio, while streaming, human-centric generation, and long-horizon storytelling/world-modeling are becoming the most active frontiers for real-world applications.

References

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025a.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation, 2025. URL `https://arxiv.org/abs/2505.04512`.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL `https://arxiv.org/abs/2503.20314`.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, Dec 1976. doi: 10.1038/264746a0. PMID: 1012311.

H. Jahncke, K. Eriksson, and S. Naula. The effects of auditive and visual settings on perceived restoration likelihood. *Noise & Health*, 17(74):1–10, Jan-Feb 2015. doi: 10.4103/1463-1741.149559.

Gianluca Sergi. Knocking at the door of cinematic artifice: Dolby Atmos, challenges and opportunities. *The New Soundtrack*, 3(2):107–121, 2013. ISSN 2042-8855. doi: 10.3366/sound.2013.0041. URL `https://doi.org/10.3366/sound.2013.0041`.

Ishita Babbar. Evolution of cinema. *International Journal for Multidisciplinary Research (IJFMR)*, 6(2):1–4, March-April 2024. ISSN 2582-2160. URL `https://www.ijfmr.com/papers/2024/2/17578.pdf`. Article ID: IJFMR240217578.

Lucas Goncalves, Prashant Mathur, Chandrashekhar Lavania, Metehan Cekic, Marcello Federico, and Kyu J. Han. Perceptual evaluation of & nbsp;audio-visual synchrony grounded in viewers' opinion scores. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXIX*, page 288–305, Berlin, Heidelberg, 2024a. Springer-Verlag. ISBN 978-3-031-72985-0. doi: 10.1007/978-3-031-72986-7_17. URL `https://doi.org/10.1007/978-3-031-72986-7_17`.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL `https://arxiv.org/abs/2412.20404`.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025.

Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Alper Canberk, Kwot Sin Lee, Vicente Ordonez, and Sergey Tulyakov. Av-link: Temporally-aligned diffusion features for cross-modal audio-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19373–19385, 2025.

Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in Neural Information Processing Systems*, 37:128118–128138, 2024.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023a.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 48855–48876. Curran Associates, Inc., 2023a. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/98c50f47a37f63477c01558600dd225a-Paper-Conference.pdf`.

Junwon Lee, Jaekwon Im, Dabin Kim, and Juhan Nam. Video-foley: Two-stage video-to-sound generation via temporal event condition for foley sound. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3813–3825, 2025. doi: 10.1109/TASLPRO.2025.3597477.

Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18770–18781, June 2025a.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024a. URL `https://arxiv.org/abs/2402.17177`.

OpenAI. Sora 2. `https://openai.com/index/sora-2/`, September 2025a. State-of-the-art video and audio generation model with synchronized audio, improved physics, and enhanced controllability.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328.

xAI. Grok 4. https://x.ai/news/grok-4, July 2025. Accessed: 2026-02-02.

Kuaishou Technology. Kling video 2.6 model. https://klingai.com/global/, December 2025. AI video generation model with simultaneous audio-visual generation capability.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation, 2025. URL https://arxiv.org/abs/2510.01284.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

Nantheera Anantrasirichai, Fan Zhang, and David Bull. Advances in artificial intelligence: A review for the creative industries, 2026. URL https://arxiv.org/abs/2501.02725.

Ruihan Zhang, Borou Yu, Jiajian Min, Yetong Xin, Zheng Wei, Juncheng Nemo Shi, Mingzhen Huang, Xianghao Kong, Nix Liu Xin, Shanshan Jiang, Praagya Bahuguna, Mark Chan, Khushi Hora, Lijian Yang, Yongqi Liang, Runhe Bian, Yunlei Liu, Isabela Campillo Valencia, Patricia Morales Tredinick, Ilia Kozlov, Sijia Jiang, Peiwen Huang, Na Chen, Xuanxuan Liu, and Anyi Rao. Generative ai for film creation: A survey of recent advances, 2025a. URL https://arxiv.org/abs/2504.08296.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. Filmaster: Bridging cinematic principles and generative ai for automated film generation, 2025a. URL https://arxiv.org/abs/2506.18899.

Pauline Leininger, Christoph Weber, and Sylvia Rothe. Understanding creative potential and use cases of ai-generated environments for virtual film productions: Insights from industry professionals. pages 60–78, 05 2025. doi: 10.1145/3706370.3727853.

Vidi Team, Celong Liu, Chia-Wen Kuo, Dawei Du, Fan Chen, Guang Chen, Jiamin Yuan, Lingxi Zhang, Lu Guo, Lusha Li, Longyin Wen, Qingyu Chen, Rachel Deng, Sijie Zhu, Stuart Siew, Tong Jin, Wei Lu, Wen Zhong, Xiaohui Shen, Xin Gu, Xing Mei, Xueqiong Qu, and Zhenfang Chen. Vidi: Large multimodal models for video understanding and editing, 2025. URL https://arxiv.org/abs/2504.15681.

Xinyue Guo, Xiaoran Yang, Lipan Zhang, Jianxuan Yang, Zhao Wang, and Jian Luan. Av-edit: Multimodal generative sound effect editing via audio-visual semantic joint control, 2025. URL https://arxiv.org/abs/2511.21146.

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing, 2025a. URL https://arxiv.org/abs/2506.21448.

Masato Ishii, Akio Hayakawa, Takashi Shibuya, and Yuki Mitsufuji. Coherent audio-visual editing via conditional audio generation following video edits, 2025. URL https://arxiv.org/abs/2512.07209.

Torin Anderson and Shuo Niu. Making ai-enhanced videos: Analyzing generative ai use cases in youtube content creation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, page 1–7. ACM, April 2025. doi: 10.1145/3706599.3719991. URL http://dx.doi.org/10.1145/3706599.3719991.

Deheng Ye, Fangyun Zhou, Jiacheng Lv, Jianqi Ma, Jun Zhang, Junyan Lv, Junyou Li, Minwen Deng, Mingyu Yang, Qiang Fu, Wei Yang, Wenkai Lv, Yangbin Yu, Yewen Wang, Yonghang Guan, Zhihao Hu, Zhongbin Fang, and Zhongqian Sun. Yan: Foundational interactive video generation, 2025. URL https://arxiv.org/abs/2508.08601.

Wenping Ma, Xiaoting Yang, Licheng Jiao, Lingling Li, Xu Liu, Fang Liu, Puhua Chen, Yuting Yang, Mengru Ma, Long Sun, Ruohan Zhang, Xueli Geng, Yuwei Guo, Shuyuan Yang, and Zhixi Feng. Video diffusion generation: comprehensive review and open problems. *Artificial Intelligence Review*, 58, 2025a. URL https://api.semanticscholar.org/CorpusID:280853499.

Yimu Wang, Xuye Liu, Wei Pang, Li Ma, Shuai Yuan, Paul Debevec, and Ning Yu. Survey of video diffusion models: Foundations, implementations, and applications, 2025a. URL https://arxiv.org/abs/2504.16081.

Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Bingyuan Wang, Qinghe Wang, Xuanhua He, Hongfa Wang, Chenyang Zhu, Hongyu Liu, Yingqing He, Zeyu Wang, Zhifeng Li, Xiu Li, Sirui Han, Yike Guo, Wei Liu, Dan Xu, Linfeng Zhang, and Qifeng Chen. Controllable video generation: A survey, 2026. URL https://arxiv.org/abs/2507.16869.

Kadhim Hayawi and Sakib Shahriar. Generative ai for text-to-video generation: Recent advances and future directions, 12 2025.

Mohamed Elmoghany, Ryan Rossi, Seunghyun Yoon, Subhojyoti Mukherjee, Eslam Bakr, Puneet Mathur, Gang Wu, Viet Dac Lai, Nedim Lipka, Ruiyi Zhang, Varun Manjunatha, Chien Nguyen, Daksh Dangi, Abel Salinas, Mohammad Taesiri, Hongjie Chen, Xiaolei Huang, Joe Barrow, Nesreen Ahmed, Hoda Eldardiry, Namyong Park, Yu Wang, Jaemin Cho, Anh Totti Nguyen, Zhengzhong Tu, Thien Nguyen, Dinesh Manocha, Mohamed Elhoseiny, and Franck Dernoncourt. A survey on long-video storytelling generation: Architectures, consistency, and cinematic quality, 2025. URL https://arxiv.org/abs/2507.07202.

Rishika Bhagwatkar, Saketh Bachu, Khurshed Fitter, Akshay Kulkarni, and Shital Chiddarwar. A review of video generation approaches. In *2020 International Conference on Power, Instrumentation, Control and Computing (PICC)*, pages 1–5, 2020. doi: 10.1109/PICC51425.2020.9362485.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis, 2025a. URL https://arxiv.org/abs/2412.15322.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10219–10228, June 2023b.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL https://api.semanticscholar.org/CorpusID:216078090.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8633–8646. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022b. URL https://arxiv.org/abs/2210.02303.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4), November 2023. ISSN 0360-0300. doi: 10.1145/3626235. URL https://doi.org/10.1145/3626235.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Taming multimodal joint training for high-quality video-to-audio synthesis. *arXiv e-prints*, pages arXiv–2412, 2024.

Mingzhen Sun, Weining Wang, Yanyuan Qiao, Jiahui Sun, Zihan Qin, Longteng Guo, Xinxin Zhu, and Jing Liu. Mm-ldm: Multi-modal latent diffusion model for sounding video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 10853–10861, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3680889. URL https://doi.org/10.1145/3664647.3680889.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/liu23f.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, October 2023a.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Trans. Mach. Learn. Res.*, 2025, 2024. URL https://api.semanticscholar.org/CorpusID:266844878.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, June 2023.

Haohang Xu, Longyu Chen, Yichen Zhang, Shuangrui Ding, and Zhipeng Zhang. Msf: Efficient diffusion model via multi-scale latent factorize, 2025a. URL https://arxiv.org/abs/2501.13349.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020b. URL https://arxiv.org/abs/2006.11239.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022c. URL https://arxiv.org/abs/2204.03458.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021. URL https://arxiv.org/abs/2102.05095.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. URL https://arxiv.org/abs/2203.12602.

P. Kingma Diederik and Welling Max. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, November 2019. ISSN 1935-8245. doi: 10.1561/2200000056. URL http://dx.doi.org/10.1561/2200000056.

Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation, 2024. URL https://arxiv.org/abs/2310.05737.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024b. doi: 10.1109/TASLP.2024.3399607.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

NVIDIA. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts, 2021. URL https://arxiv.org/abs/2106.05974.

Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran Xu, Yicheng Ma, Meishan Zhang, Baotian Hu, and Min Zhang. Uni-moe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data, 2025b. URL https://arxiv.org/abs/2511.12609.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding, 2020. URL https://arxiv.org/abs/2006.16668.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3424–3439, 2025c. doi: 10.1109/TPAMI.2025.3532688.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *CoRR*, abs/2401.06066, 2024. URL https://arxiv.org/abs/2401.06066.

Evandro S. Ortigossa and Eran Segal. Seg-moe: Multi-resolution segment-wise mixture-of-experts for time series forecasting transformers, 2026. URL https://arxiv.org/abs/2601.21641.

Yike Yuan, Ziyu Wang, Zihao Huang, Defa Zhu, Xun Zhou, Jingyi Yu, and Qiyang Min. Expert race: A flexible routing strategy for scaling diffusion transformer with mixture of experts, 2025a. URL https://arxiv.org/abs/2503.16057.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2023. URL https://arxiv.org/abs/2211.01324.

Shaobin Zhuang, Yiwei Guo, Yanbo Ding, Kunchang Li, Xinyuan Chen, Yaohui Wang, Fangyikang Wang, Ying Zhang, Chen Li, and Yali Wang. Timestep master: Asymmetrical mixture of timestep lora experts for versatile and efficient diffusion models in vision, 2025. URL https://arxiv.org/abs/2503.07416.

Kun Cheng, Xiao He, Lei Yu, Zhijun Tu, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Diff-MoE: Diffusion transformer with time-aware and space-adaptive experts. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10010–10024. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/cheng25d.html.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A Ross, Bryan Seybold, and Lu Jiang. VideoPoet: A large language model for zero-shot video generation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25105–25124. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kondratyuk24a.html.

Sand. ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. Magi-1: Autoregressive video generation at scale, 2025. URL https://arxiv.org/abs/2505.13211.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1587–1606, June 2025d.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025b. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025c. URL https://arxiv.org/abs/2509.17765.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26429–26445, 2023. URL https://api.semanticscholar.org/CorpusID:266573555.

Xin He, Longhui Wei, Jianbo Ouyang, Minghui Liao, Lingxi Xie, and Qi Tian. Emma: Efficient multimodal understanding, generation, and editing with a unified architecture, 2025a. URL https://arxiv.org/abs/2512.04810.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *ArXiv*, abs/2402.08268, 2024c. URL `https://api.semanticscholar.org/CorpusID:267637090`.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025b.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025c.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

Shawn Hershey, Daniel P W Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. The benefit of temporally-strong labels in audio event classification, 2021. URL `https://arxiv.org/abs/2105.07031`.

Lei Zhao, Linfeng Feng, Dongxu Ge, Rujin Chen, Fangqiu Yi, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. Uniform: A unified multi-task diffusion transformer for audio-video generation, 2025. URL `https://arxiv.org/abs/2502.03897`.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds, 2024.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025b.

Yuechen Zhang, Jinbo Xing, Bin Xia, Shaoteng Liu, Bohao Peng, Xin Tao, Pengfei Wan, Eric Lo, and Jiaya Jia. Training-free efficient video generation via dynamic token carving. *arXiv preprint arXiv:2505.16864*, 2025b.

Assaf Singer, Noam Rotstein, Amir Mann, Ron Kimmel, and Or Litany. Time-to-move: Training-free motion controlled video generation via dual-clock denoising. *arXiv preprint arXiv:2511.08633*, 2025.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024a.

Jibin Song, Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Syncphony: Synchronized audio-to-video generation with diffusion transformers, 2025. URL `https://arxiv.org/abs/2509.21893`.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arxiv. *arXiv preprint arXiv:2208.12242*, 2022.

Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Taming image diffusion transformers for efficient joint audio and video generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 10486–10495, New York, NY, USA, 2025c. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755713. URL https://doi.org/10.1145/3746027.3755713.

Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7827–7839, 2024b.

Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Taming image diffusion transformers for efficient joint audio and video generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 10486–10495, New York, NY, USA, 2025d. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755713. URL https://doi.org/10.1145/3746027.3755713.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329, 2024. doi: 10.1109/ICASSP48485.2024.10448489.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions, 2025c. URL https://arxiv.org/abs/2511.03334.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation, 2025. URL https://arxiv.org/abs/2508.16930.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: synchronized video-to-audio synthesis with latent diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023b. Curran Associates Inc.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12978–12988, 2025b.

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7763–7772, 2025a.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*, 2025c.

Jingxi Chen, Brandon Y Feng, Haoming Cai, Tianfu Wang, Levi Burner, Dehao Yuan, Cornelia Fermuller, Christopher A Metzler, and Yiannis Aloimonos. Repurposing pre-trained video diffusion models for event-based video interpolation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12456–12466, 2025c.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025.

Shuchen Weng, Haojie Zheng, Zheng Chang, Si Li, Boxin Shi, and Xinlong Wang. Audio-sync video generation with multi-stream temporal control, 2025. URL https://arxiv.org/abs/2506.08003.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019a. URL https://api.semanticscholar.org/CorpusID:202725406.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. URL https://arxiv.org/abs/1812.01717.

Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b. doi: 10.1109/TPAMI.2025.3633890.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. 2023b. URL https://api.semanticscholar.org/CorpusID:264172222.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. URL https://arxiv.org/abs/1705.06950.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Fuxiao Liu, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *arXiv preprint arXiv:2505.01481*, 2025e.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, and Ben Sangbae Chon. Kad: No more fad! an effective and efficient evaluation metric for audio generation. *arXiv:2502.15602*, 2025a. URL https://arxiv.org/abs/2502.15602.

Soham Deshmukh et al. Pam: Prompting audio-language models for audio quality assessment. *arXiv preprint arXiv:2402.00282*, 2024.

Apoorv Vyas, Bowen Shi, Matt Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, W.K.F. Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. Audiobox: Unified audio generation with natural language prompts. *ArXiv*, abs/2312.15821, 2023. URL https://api.semanticscholar.org/CorpusID:266551778.

Huan Zhang, Jinhua Liang, Huy Phan, Wenwu Wang, and Emmanouil Benetos. From aesthetics to human preferences: Comparative perspectives of evaluating text-to-music systems. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2025d. doi: 10.1109/MLSP62443.2025.11204254.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, June 2023a.

Lucas Goncalves, Prashant Mathur, Chandrashekhar Lavania, Metehan Cekic, Marcello Federico, and Kyu J. Han. Peavs: Perceptual evaluation of audio-visual synchrony grounded in viewers' opinion scores. In *European Conference on Computer Vision*, 2024b. URL `https://api.semanticscholar.org/CorpusID:269043272`.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation, 2023a.

Xingbo Feng, Zhuyun Qi, Yi Wang, Ziyao Huang, Yan Liu, Jiashuo Lin, Chenxi Ling, Weichao Li, Jin Zhang, and Jianping Wang. Desync: Proactive congestion control via random delay offsets for large-scale ml training. In *2025 IEEE/ACM 33rd International Symposium on Quality of Service (IWQoS)*, pages 1–2, 2025a. doi: 10.1109/IWQoS65803.2025.11143372.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023b.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019b. URL `https://arxiv.org/abs/1812.08466`.

Methods for subjective determination of transmission quality summary. URL `https://api.semanticscholar.org/CorpusID:143430507`.

Thomas Unterthiner et al. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Ziqi Huang et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024b.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019c.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, and Ben Sangbae Chon. Kad: No more fad! an effective and efficient evaluation metric for audio generation. *arXiv preprint arXiv:2502.15602*, 2025b.

Benjamin Elizalde et al. Clap: Learning audio concepts from natural language supervision. *arXiv preprint arXiv:2206.04769*, 2022.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Guy Yariv et al. Diverse and aligned audio-to-video generation via text-to-video model adaptation. *arXiv preprint arXiv:2309.16429*, 2023b.

Huai Zhou et al. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. *arXiv preprint (see: MMAudio paper citing DeSync)*, 2025. Uses DeSync predicted by Synchformer for AV synchrony evaluation.

Shentong Mo et al. Text-to-audio generation synchronized with videos. *arXiv preprint arXiv:2403.07938*, 2024.

Kazuki Shimada et al. Benchmarking spatially aligned audio-video generation. *arXiv preprint arXiv:2412.13462*, 2024a.

Itu-t recommendation p.800.1: Mean opinion score (mos) terminology. Technical report, International Telecommunication Union (ITU-T), 7 2016. Rec. ITU-T P.800.1 (07/2016).

Rethinking human evaluation protocol for text-to-video models: Enhancing reliability, reproducibility, and practicality. In *NeurIPS*, 2024. Poster / protocol paper (T2VHE).

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, et al. Video-bench: Human-aligned video generation benchmark. *CVPR*, 2025.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Defossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. In *ICLR*, 2023. Human eval commonly reports overall quality and text relevance.

A. Vinay et al. Evaluating generative audio systems and their metrics. In *ISMIR*, 2022.

L. Goncalves et al. Perceptual evaluation of audio-visual synchrony. In *ECCV*, 2024c.

Kazuki Shimada et al. Benchmarking spatially aligned audio-video generation. *arXiv preprint arXiv:2412.13462*, 2024b.

Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, page 52–69, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58609-6. doi: 10.1007/978-3-030-58610-2_4. URL https://doi.org/10.1007/978-3-030-58610-2_4.

OpenAI. Sora 2 is here. https://openai.com/index/sora-2/, September 2025b. Accessed 2026-02-04.

Google AI Studio. Veo 3. https://aistudio.google.com/models/veo-3, 2025. Accessed 2026-02-04.

Google DeepMind. Veo. https://deepmind.google/models/veo/, 2025. Accessed 2026-02-04.

AtlasCloud. Wan2.6 video models api collection. https://www.atlascloud.ai/collections/wan2.6, 2025. Accessed 2026-02-04.

Higgsfield. Wan 2.6 ai video generator. https://higgsfield.ai/wan-2.6, 2025. Accessed 2026-02-04.

xAI. Grok imagine api. https://x.ai/news/grok-imagine-api, 2026. Accessed 2026-02-04.

Kling AI. Kling video 2.6 user guide (audio). https://app.klingai.com/global/quickstart/klingai-video-26-audio-user-guide, 2025. Accessed 2026-02-04.

Runway. Introducing runway gen-4. https://runwayml.com/research/introducing-runway-gen-4, 2025a. Accessed 2026-02-04.

Runway. Runway models (developer docs). https://dev.runwayml.com/models, 2026. Accessed 2026-02-04.

Pika. Sound effects on pika. https://pikalabs.org/sound-effects-on-pika/, March 2024. Accessed 2026-02-04.

Pika. Pika. https://pika.art/, 2026. Accessed 2026-02-04.

Yahoo Finance. Bytedance adds video generator to china's most popular ai chatbot doubao. https://finance.yahoo.com/news/bytedance-adds-video-generator-chinas-093000339.html, November 2024. Accessed 2026-02-04.

WIRED. How bytedance made china's most popular ai chatbot. https://www.wired.com/story/bytedance-doubao-chatbot-popularity, 2025. Accessed 2026-02-04.

ComfyUI Blog. Stable audio 2.5 is now in comfyui! https://blog.comfy.org/p/stable-audio-25-is-now-in-comfyui, September 2025. Accessed 2026-02-04.

ComfyUI Blog. Grok imagine now available in comfyui. https://blog.comfy.org/p/grok-imagine-now-available-in-comfyui, January 2026. Accessed 2026-02-04.

gitmylo. Comfyui-audio-nodes. https://github.com/gitmylo/ComfyUI-audio-nodes, 2026. Accessed 2026-02-04.

Kuaishou. Kling ai. *Kling AI*, 2024. URL https://www.klingai.com/global/.

Jake Rosenberg. Forgiveness before permission: The legal and ethical implications of openai's sora. *University of Miami Law Review*, nov 2025. Insights Section.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *ArXiv*, abs/2406.08801, 2024a. URL https://api.semanticscholar.org/CorpusID:270440539.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditioning, 2024.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation, 2024.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXIII*, page 244–260, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73009-2. doi: 10.1007/978-3-031-73010-8_15. URL https://doi.org/10.1007/978-3-031-73010-8_15.

Linrui Tian, Siqi Hu, Qi Wang, Bang Zhang, and Liefeng Bo. Emo2: End-effector guided audio-driven avatar video generation. *ArXiv*, abs/2501.10687, 2025. URL https://api.semanticscholar.org/CorpusID:275757784.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models, 2025a. URL https://arxiv.org/abs/2502.01061.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. Artificial intelligence index report 2024, 2024. URL https://arxiv.org/abs/2405.19522.

Laurie Sullivan. Google solves holiday miracles in series of storytelling ads. *MediaPost*, nov 2025. URL https://www.mediapost.com/publications/article/411028/google-solves-holiday-miracles-in-series-of-storyt.html. Details the release of Veo 3 generated commercials, including "Mr. Fuzzy's Big Adventure" and "Big Night Out".

Adobe Newsroom. Adobe firefly delivers groundbreaking ai audio, video and imaging innovations and new models in all-in-one creative ai studio, oct 2025. URL https://news.adobe.com/news/2025/10/adobe-max-2025-firefly. Press Release: Announced at Adobe MAX 2025.

Katelyn Chedraoui. Adobe's new ai is all about audio: How to create music for your videos with firefly. *CNET*, oct 2025. Reviews the Firefly Audio Model's commercial safety and new Generate Speech features.

ElevenLabs. Power game worlds with natural ai voices. ElevenLabs Official Website, 2025. URL https://elevenlabs.io/use-cases/gaming. Documentation on Turbo v2.5 low-latency models and NPC voice integration.

Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, Masayoshi Tomizuka, and Kurt Keutzer. Streamdiffusion: A pipeline-level solution for real-time interactive generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12371–12380, October 2025a.

Yubo Huang, Hailong Guo, Fangtai Wu, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live avatar: Streaming real-time audio-driven avatar generation with infinite length, 2025c. URL https://arxiv.org/abs/2512.04677.

Wilson Yan, Yunzhi Zhang, P. Abbeel, and A. Srinivas. Videogpt: Video generation using vq-vae and transformers. *ArXiv*, abs/2104.10157, 2021. URL https://api.semanticscholar.org/CorpusID:233307257.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls, 2025. URL https://arxiv.org/abs/2511.01266.

Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, Masayoshi Tomizuka, and Kurt Keutzer. Streamdiffusion: A pipeline-level solution for real-time interactive generation, 2025b. URL https://arxiv.org/abs/2312.12491.

Tianrui Feng, Zhi Li, Shuo Yang, Haocheng Xi, Muyang Li, Xiuyu Li, Lvmin Zhang, Keting Yang, Kelly Peng, Song Han, et al. Streamdiffusionv2: A streaming system for dynamic and interactive video generation. *arXiv preprint arXiv:2511.07399*, 2025b.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025d.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025b.

Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. Videogpa: Distilling geometry priors for 3d-consistent video generation, 2026. URL https://arxiv.org/abs/2601.23286.

Zhiyuan Li, Chi-Man Pun, Chen Fang, Jue Wang, and Xiaodong Cun. Personalive! expressive portrait image animation for live streaming. *arXiv preprint arXiv:2512.11253*, 2025f.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation, 2024a. URL https://arxiv.org/abs/2311.17117.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *arXiv*, 2023.

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time, 2024b. URL https://arxiv.org/abs/2404.10667.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11040–11049, 2022. doi: 10.1109/CVPR52688.2022.01077.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *arXiv preprint arXiv:2505.22647*, 2025.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, and Xiaoming Wei. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025c.

Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. *arXiv preprint arXiv:2504.04842*, 2025e.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025.

Meituan LongCat Team. Longcat-video-avatar technical report. Technical report (PDF), 2025. Available at: `https://github.com/meituan-longcat/LongCat-Video/blob/main/assets/LongCat-Video-Avatar-Tech-Report.pdf`.

Kling Team, Jialu Chen, Yikang Ding, Zhixue Fang, Kun Gai, Yuan Gao, Kang He, Jingyun Hua, Boyuan Jiang, Mingming Lao, Xiaohan Li, Hui Liu, Jiwen Liu, Xiaoqiang Liu, Yuan Liu, Shun Lu, Yongsen Mao, Yingchao Shao, Huafeng Shi, Xiaoyu Shi, Peiqin Sun, Songlin Tang, Pengfei Wan, Chao Wang, Xuebo Wang, Haoxian Zhang, Yuanxing Zhang, and Yan Zhou. Klingavatar 2.0 technical report. *arXiv preprint arXiv:2512.13313*, 2025.

Zhizhou Zhong, Yicheng Ji, Zhe Kong, Yiying Liu, Jiarui Wang, Jiasun Feng, Lupeng Liu, Xiangyi Wang, Yanjia Li, Yuqing She, Ying Qin, Huan Li, Shuiyang Mao, Wei Liu, and Wenhan Luo. Anytalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475*, 2025.

Yubo Huang, Hailong Guo, Fangtai Wu, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025e.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*, 2025b.

Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, Qinglin Lu, and Chengjie Wang. Sonic: Shifting focus to global audio perception in portrait animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Also available as arXiv:2411.16331.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Feng Wang, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-animate: Unified character animation and replacement with holistic replication, 2025d. URL `https://arxiv.org/abs/2509.14055`.

Wenhao Yan, Sheng Ye, Zhuoyi Yang, Jiayan Teng, ZhenHui Dong, Kairui Wen, Xiaotao Gu, Yong-Jin Liu, and Jie Tang. Scail: Towards studio-grade character animation via in-context learning of 3d-consistent pose representations. *arXiv preprint arXiv:2512.05905*, 2025a.

Jiaming Zhang, Shengming Cao, Rui Li, Xiaotong Zhao, Yutao Cui, Xinglin Hou, Gangshan Wu, Haolan Chen, Yu Xu, Limin Wang, and Kai Ma. Steadydancer: Harmonized and coherent human image animation with first-frame preservation, 2025e. URL https://arxiv.org/abs/2511.19320.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black. Generating holistic 3d human motion from speech. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, 2023. doi: 10.1109/CVPR52729.2023.00053.

Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. Edge: Editable dance generation from music. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 448–458, 2023. doi: 10.1109/CVPR52729.2023.00051.

Liyang Chen et al. Humo: Human motion-aware video generation. *arXiv preprint arXiv:2504.17204*, 2025d.

Xiyang Wu et al. First frame guidance for human video generation. *arXiv preprint arXiv:2511.15700*, 2025. FFGo.

Chang Xue et al. Stand-in: Fine-grained subject-driven video generation with spatial-temporal consistency. *arXiv preprint arXiv:2508.07901*, 2025.

Kaiming Hu et al. Hunyuancustom: A multimodal customization framework for video generation. *arXiv preprint arXiv:2408.13753*, 2024b.

Shichao Xu et al. Hypermotion: High-fidelity motion control for video generation. *arXiv preprint arXiv:2505.22977*, 2025d.

Wenbo Liu et al. Phantom: Subject-driven video generation via phantom tokens. *arXiv preprint arXiv:2502.11079*, 2025d.

Ziqian Qiang et al. Mm-sonate: Training-free high-quality audio-video generation with zero-shot voice cloning. *arXiv preprint arXiv:2601.01568*, 2026.

Yinan He et al. Audiogen-omni: A unified multi-modal diffusion transformer for video-synchronized audio generation. *arXiv preprint arXiv:2508.00733*, 2025b.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025f.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025e.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025g.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

Jianxiong Gao, Zhaoxi Chen, Xian Liu, Jianfeng Feng, Chenyang Si, Yanwei Fu, Yu Qiao, and Ziwei Liu. Longvie: Multimodal-guided controllable ultra-long video generation. *arXiv preprint arXiv:2508.03694*, 2025.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *arXiv preprint arXiv:2407.19918*, 2024.

Haydar Yesiltepe et al. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 2025.

Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, Chenyang Zhang, Tao Xiang, Fanny Yang, Serge Belongie, and Tian Xie. Onestory: Coherent multi-shot video generation with adaptive memory. *arXiv preprint arXiv:2512.07802*, 2025.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, Yujun Shen, and Huamin Qu. Holocine: Holistic generation of cinematic multi-shot long video narratives. *arXiv preprint arXiv:2510.20822*, 2025.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025b.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. Storymem: Multi-shot long video storytelling with memory. *arXiv preprint arXiv:2512.19539*, 2025g.

Yehang Zhang, Xinli Xu, Xiaojie Xu, Doudou Zhang, Li Liu, and Yingcong Chen. Orchestrating audio: Multi-agent framework for long-video audio synthesis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22267–22282. Association for Computational Linguistics, November 2025h.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö Arık. Vista: A test-time self-improving video generation agent. *arXiv preprint arXiv:2510.15831*, 2025.

Achint Soni, Sreyas Venkataraman, Abhranil Chandra, Sebastian Fischmeister, Percy Liang, Bo Dai, and Sherry Yang. Videoagent: Self-improving video generation. *arXiv preprint arXiv:2410.10076*, 2024.

Xiaoxuan Tang, Xinping Lei, Chaoran Zhu, Shiyun Chen, Ruibin Yuan, Yizhi Li, Changjae Oh, Ge Zhang, Wenhao Huang, Emmanouil Benetos, et al. Automv: An automatic multi-agent system for music video generation. *arXiv preprint arXiv:2512.12196*, 2025.

Chenyu Mu, Xin He, Qu Yang, Wanshun Chen, Jiadi Yao, Huang Liu, Zihao Yi, Bo Zhao, Xingyu Chen, Ruotian Ma, et al. The script is all you need: An agentic framework for long-horizon dialogue-to-cinematic video generation. *arXiv preprint arXiv:2601.17737*, 2026.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Xun Huang. Towards video world models, 2025. URL `https://www.xunhuang.me/blogs/world_model.html`.

Kai Zhang, Xiangchao Chen, Bo Liu, Tianci Xue, Zeyi Liao, Zhihan Liu, Xiyao Wang, Yuting Ning, Zhaorun Chen, Xiaohan Fu, et al. Agent learning via early experience. *arXiv preprint arXiv:2510.08558*, 2025i.

Adrian Bolton, Alexander Lerchner, Alexandra Cordell, Alexandre Moufarek, Andrew Bolt, Andrew Lampinen, Anna Mitenkova, Arne Olav Hallingstad, Bojan Vujatovic, Bonnie Li, et al. Sima 2: A generalist embodied agent for virtual worlds. *arXiv preprint arXiv:2512.04797*, 2025.

Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. Learning spatially-aware language and audio embeddings. *Advances in Neural Information Processing Systems*, 37:33505–33537, 2024.

Yuelei Li, Hyunjin Kim, Fangneng Zhan, Ri-Zhao Qiu, Mazeyu Ji, Xiaojun Shan, Xueyan Zou, Paul Liang, Hanspeter Pfister, and Xiaolong Wang. Visual acoustic fields. *arXiv preprint arXiv:2503.24270*, 2025h.

Xiulong Liu, Anurag Kumar, Paul Calamia, Sebastia V Amengual, Calvin Murdock, Ishwarya Ananthabhotla, Philip Robinson, Eli Shlizerman, Vamsi Krishna Ithapu, and Ruohan Gao. Hearing anywhere in any environment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5732–5741, 2025f.

Derong Jin and Ruohan Gao. Differentiable room acoustic rendering with multi-view vision priors. *arXiv preprint arXiv:2504.21847*, 2025.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. Visage: Video-to-spatial audio generation. *arXiv preprint arXiv:2506.12199*, 2025.

Jiahua Wang, Shannan Yan, Leqi Zheng, Jialong Wu, and Yaoxin Mao. Audio-visual world models: Towards multisensory imagination in sight and sound. *arXiv preprint arXiv:2512.00883*, 2025f.

Fan Zhang and Michael Gienger. Learning robot manipulation from audio world models. *arXiv preprint arXiv:2512.08405*, 2025.

Runway. Introducing runway gwm-1, December 2025b. URL https://runwayml.com/research/introducing-runway-gwm-1. Accessed: 2026-02-02.

Veo-Team. Veo 2. *DeepMind Blog*, 2024. URL https://deepmind.google/technologies/veo/veo-2/.

Arjun Somayazulu, Sagnik Majumder, Changan Chen, and Kristen Grauman. Activerir: Active audio-visual exploration for acoustic environment modeling. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13830–13836, 2024. URL https://api.semanticscholar.org/CorpusID:269362234.

Tongtong Feng, Xin Wang, Yu-Gang Jiang, and Wenwu Zhu. Embodied ai: From llms to world models [feature]. *IEEE Circuits and Systems Magazine*, 25:14–37, 2025c. URL https://api.semanticscholar.org/CorpusID:281505303.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. Vabench: A comprehensive benchmark for audio-video generation, 2025. URL https://arxiv.org/abs/2512.09299.

Uttam Chakraborty and Santosh Kumar Biswal. Scaling generative ai: key factors driving user growth and community. *J. Decis. Syst.*, 35, 2025. URL https://api.semanticscholar.org/CorpusID:284257938.

Ethan Chern, Zhulin Hu, Bohao Tang, Jiadi Su, Steffi Chern, Zhijie Deng, and Pengfei Liu. Livetalk: Real-time multimodal interactive video diffusion via improved on-policy distillation, 2025. URL https://arxiv.org/abs/2512.23576.

Ming Chen, Liyuan Cui, Wenyuan Zhang, Haoxian Zhang, Yan Zhou, Xiaohan Li, Songlin Tang, Jiwen Liu, Borui Liao, Hejia Chen, Xiaoqiang Liu, and Pengfei Wan. Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation, 2025e. URL https://arxiv.org/abs/2508.19320.

Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Zhendong Wang, Hao Liu, Bin Lin, Hao Li, Xue Xu, Xinyan Xiao, Jingdong Wang, Haifeng Wang, and Li Yuan. Unified multimodal model as auto-encoder, 2025b. URL https://arxiv.org/abs/2509.09666.

Shijie Wang, Li Zhang, Xinyan Liang, Yuhua Qian, and Shen Hu. Balanced multimodal learning: An unidirectional dynamic interaction perspective, 2025g. URL https://arxiv.org/abs/2509.02281.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. UniTok: A unified tokenizer for visual generation and understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b. URL https://neurips.cc/virtual/2025/poster/116864.