This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Harmony: Harmonizing Audio and Video Generation through Cross-Task Synergy

Teng Hu[1*] Zhentao Yu[2*] Guozhen Zhang[2] Zihan Su[1] Zhengguang Zhou[2]
Youliang Zhang[2] Yuan Zhou[2] Qinglin Lu[2] Ran Yi[1†]
[1]Shanghai Jiao Tong University [2]Tencent Hunyuan

Project page: https://sjtuplayer.github.io/projects/Harmony

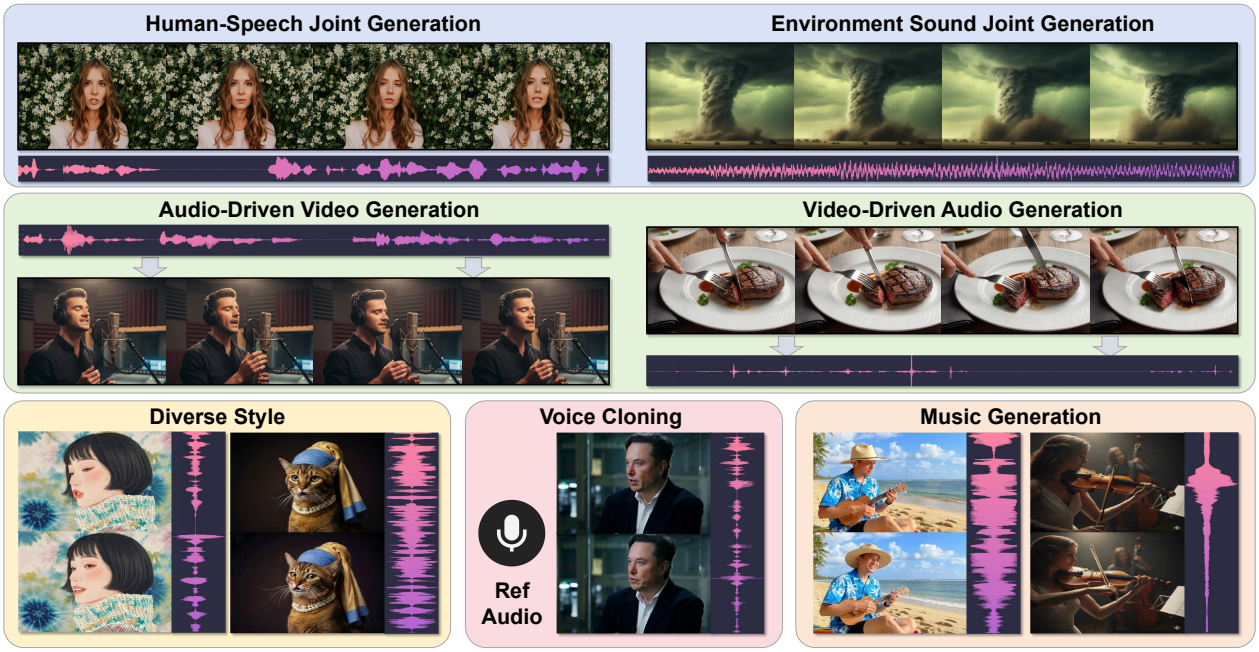


Figure 1. **Harmony** employs a cross-task synergy training strategy to achieve robust audio-visual synchronization. This versatile framework supports multiple generation paradigms, including joint audio-video synthesis as well as audio-driven and video-driven generation, while also demonstrating strong generalizability to diverse audio types (e.g., music) and visual styles.

Abstract

The synthesis of synchronized audio-visual content is a key challenge in generative AI, with open-source models facing challenges in robust audio-video alignment. Our analysis reveals that this issue is rooted in three fundamental challenges of the joint diffusion process: (1) ***Correspondence Drift****, where concurrently evolving noisy latents impede stable learning of alignment; (2) inefficient global attention mechanisms that fail to capture fine-grained temporal cues; and (3) the intra-modal bias of conventional Classifier-Free Guidance (CFG), which enhances conditionality but not cross-modal synchronization. To overcome these challenges, we introduce* ***Harmony****, a novel framework that mechanistically enforces audio-visual synchronization. We first propose a* ***Cross-Task Synergy*** *training paradigm to mitigate drift by leveraging strong supervisory signals from audio-driven video and video-driven audio generation tasks. Then, we design a* ***Global-Local Decoupled Interaction Module*** *for efficient and precise temporal-style alignment. Finally, we present a novel* ***Synchronization-Enhanced CFG (SyncCFG)*** *that explicitly isolates and amplifies the alignment signal during inference. Extensive experiments demonstrate that Harmony establishes a new state-of-the-art, significantly outperforming existing methods in both generation fidelity and, critically, in achieving fine-grained audio-visual synchronization.*

[*] Equal Contribution.
[†] Corresponding author.

1. Introduction

The unified synthesis of audio and video represents a pivotal frontier in contemporary generative AI, with profound implications for content creation, digital avatars, and immersive virtual worlds. Industry-leading proprietary models, such as Veo 3 [10] and Sora 2 [40], have set a high benchmark, delivering outputs with remarkable fidelity and demonstrating substantial practical utility. However, a significant gap persists between these closed-source systems and the capabilities of existing open-source methods [8, 41, 45]. A fundamental challenge, in particular, remains largely unsolved in the open-source community: achieving precise and harmonious audio-visual alignment.

While recent open-source models have made strides in generation quality, they often struggle with robust audio-visual synchronization. Recent explorations into end-to-end **joint audio-video generation** [14, 23, 32, 34, 44, 48] underscore this limitation. Specifically, these methods often exhibit specialized limitations: many are confined to generating ambient sounds and fail to synthesize natural human speech [14, 23, 32, 44]; while others, such as JAM-Flow [27], focus solely on speech generation but lack the capability to generate environmental sounds. Even among more general models, Ovi [34] exhibits deficiencies in *robust alignment*, while UniVerse-1 [48] suffers from poor *audio-video synchronization*. Table 1 presents a capability comparison. These shortcomings reveal a critical gap in current research: few methods investigate the root causes of audio-video misalignment from a methodological standpoint. Consequently, there remains a lack of highly generalizable and well-aligned audio-video joint generation methods. *This leaves a significant void in the open-source landscape for a unified framework capable of generating a comprehensive audio spectrum—from ambient sounds to human speech—while maintaining precise audio-visual harmony.*

In this work, we posit that the difficulty in achieving robust synchronization stems from three fundamental challenges inherent to the joint diffusion process. *(1)* During joint generation, both modalities are progressively denoised from pure noise. In the early, highly stochastic stages, attempting to align two concurrently evolving, highly noisy latents causes a phenomenon we term **Correspondence Drift**, where the optimal mapping continuously shifts, impeding stable learning. *(2)* Audio-visual synchronization presents a fundamental architectural tension between two competing objectives: precise, frame-level temporal alignment (*e.g.*, lip movements) and holistic, global style consistency (*e.g.*, emotional tone). Existing designs, often relying on a single, monolithic mechanism like global cross-attention, conflate these distinct goals, forcing the model into a suboptimal trade-off where neither objective is fully achieved. *(3)* Conventional Classifier-Free Guidance (CFG) [15] operates by amplifying conditioning signals for each modality in isolation. Consequently, it does not inherently promote or enhance the crucial cross-modal correspondence between the generated audio and video.

Table 1. Comparison of capabilities among existing joint audio-video generation models. We evaluate their ability to generate different sound types and the quality of their temporal alignment with the video. (✓: Good, ❍: Fair/Limited, ✗: Poor/Unsupported)

Model	Human Speech	Environmental Sound	Speech-Video Alignment	Sound-Video Alignment
MM-Diffusion [44]	✗	❍	✗	❍
JavisDiT [32]	✗	❍	✗	❍
AnimateSI [51]	✗	❍	✗	❍
JAM-Flow [27]	✓	✗	✓	✗
UniVerse-1 [48]	✓	✓	✗	❍
Ovi [34]	✓	✓	❍	❍
Harmony (Ours)	✓	✓	✓	✓

To overcome these challenges, we propose **Harmony**, a novel joint audio-video generation framework designed to generate highly synchronized audio-video content with cross-task synergy. Harmony's design is centered on three core innovations, each targeting one of the aforementioned challenges. To mitigate Correspondence Drift, we employ a **Cross-Task Synergy** training paradigm, co-training the joint generation task with auxiliary audio-driven video and video-driven audio generation tasks to leverage the latter's strong supervisory signal for instilling robust alignment priors. To resolve the conflation of local and global synchronization objectives, we further propose a **Global-Local Decoupled Interaction Module** that ensures both holistic style consistency and precise temporal synchronization through the decoupled global style attention and localized, RoPE-aligned frame-wise attention. Finally, to address the fact that conventional CFG lacks a mechanism for enhancing audio-visual alignment, we propose **Synchronization-Enhanced CFG (SyncCFG)**. This novel technique redefines the negative condition learned from the cross-task training stage to explicitly isolate and amplify the guidance vector corresponding to audio-visual alignment.

Extensive experiments on our newly proposed **Harmony-Bench** validate our framework on the challenging task of jointly generating human speech and ambient sounds. Harmony achieves the best audio-video alignment, maintaining fine-grained temporal synchronization in complex acoustic scenes, which confirms the efficacy of our approach. The main contributions of our work are summarized as follows:

- We propose **Harmony**, a novel joint audio-video generation framework built upon the principle of **Cross-Task Synergy** to resolve the fundamental *Correspondence Drift* problem in joint diffusion models.
- We design a **Global-Local Decoupled Interaction Module** that achieves comprehensive alignment in both overall style and fine-grained temporal details.
- We propose a novel **Synchronization-Enhanced CFG**

(SyncCFG) that guides the model towards better audio-visual correspondence during inference by using mute-audio and static-video condition as negative guidance.
- We establish a new state-of-the-art in audio-visual generation, with extensive experiments validating Harmony's superior performance in cross-modal synchronization.

2. Related Work

2.1. Video Generation

The field of video generation has rapidly advanced [5, 13, 20, 26, 36, 47, 54, 61], transitioning from early Generative Adversarial Networks (GANs) [11] to the now-dominant diffusion models [16]. Building on their success in image synthesis [22, 42, 43], models like AnimateDiff [13] and SVD [1] extended diffusion to the temporal domain. Architectures have also evolved from UNets to more powerful Diffusion Transformers (DiT) [31, 58], with recent open-source models like HunyuanVideo [26] and Wan [47] achieving state-of-the-art visual quality, which inspired a lot of downstream video generation methods, like video customization [2, 18, 19], motion control [17, 21, 28, 37, 38, 50, 52, 55–57, 62], and video editing [4, 7, 30, 33]. However, a critical limitation persists across this body of work: a singular focus on the visual modality. By generating silent videos, these models produce content that feels incomplete and lacks the immersive quality of real-world experiences, underscoring the need for cohesive audio-visual synthesis.

2.2. Joint Audio-Video Generation

To incorporate audio into video synthesis, early methods typically relied on audio-driven approaches [6, 24, 35, 59]; however, these techniques inherently require the audio track to be generated in advance. Recently, a growing body of research has begun to explore the simultaneous generation of audio and video within a single, unified framework [32, 34, 44, 48, 49, 53, 60]. However, most of the early open-source approaches in this domain were restricted to synthesizing coarse environmental sounds and were unable to generate meaningful human speech [23, 32, 44]. A significant advancement came with models like UniVerse-1 [48] and Ovi [34], which integrated more powerful audio synthesis components to enable the joint generation of both ambient sounds and human vocals. Despite this progress, a key challenge remains in the detailed alignment of the overall soundscape. These models often struggle to cohesively blend human speech with its surrounding environmental audio in a manner that is acoustically and semantically consistent with the visual context, highlighting a remaining gap in creating truly immersive audio-visual experiences.

3. Method

In this section, we introduce **Harmony**, a novel framework for joint audio-video synthesis designed to overcome the fundamental challenge of cross-modal alignment in diffusion models, capable of joint audio-video generation on speech, sound effects, and ambient audio. We introduce three core innovative designs: (1) a **Cross-Task Synergy** training strategy that combines the standard joint audio-video generation task with auxiliary audio-driven video and video-driven audio generation tasks, leveraging the strong, uni-directional supervisory signals from both to accelerate and stabilize the learning of audio-video alignment (2) a **Global-Local Decoupled Interaction Module** that efficiently ensures both fine-grained temporal correspondence and holistic stylistic consistency; and (3) a **Cross-Task Alignment-Enhanced CFG** mechanism that repurposes guidance by designing more meaningful negative anchors, to explicitly amplify audio-video synchronization during inference.

3.1. Preliminary: Joint Audio-Video Diffusion

Joint audio-video synthesis typically employs a dual-stream Latent Diffusion Model. After encoding a video V and audio A into latents (z_v, z_a), a denoising network ϵ_θ is trained to reverse a standard Gaussian noising process. The network consists of parallel video and audio backbones that process their respective noisy latents, $z_{v,t}$ and $z_{a,t}$. Synchronization is learned through an interaction module (e.g., cross-attention) that couples the two streams. The model is optimized by minimizing the noise prediction error for both modalities:

$$\mathcal{L} = ||\epsilon_v - \hat{\epsilon}_v(z_{v,t}, z_{a,t}, t)||^2 + ||\epsilon_a - \hat{\epsilon}_a(z_{a,t}, z_{v,t}, t)||^2. \quad (1)$$

However, this standard approach struggles to learn robust alignment from two concurrently noisy signals—a core challenge our work addresses.

3.2. Cross-Task Synergy for Enhanced Alignment

3.2.1. The Challenge of Correspondence Drift

Problem Formulation. While recent advancements in joint audio-video generation have focused on novel architectures, a fundamental challenge persists: achieving robust cross-modal alignment. We identify the root cause not as an architectural limitation, but as an inherent instability in the training paradigm, a phenomenon we term **Correspondence Drift**. During the initial stages of joint training, both audio and video signals are heavily diffused with noise. Attempting to learn a correspondence between two concurrently evolving, highly stochastic latent variables results in an unstable and inefficient learning target, causing the alignment process to drift and converge slowly.

Empirical Motivation. To empirically validate this hypothesis, we present a comparative analysis in Fig. 3. We

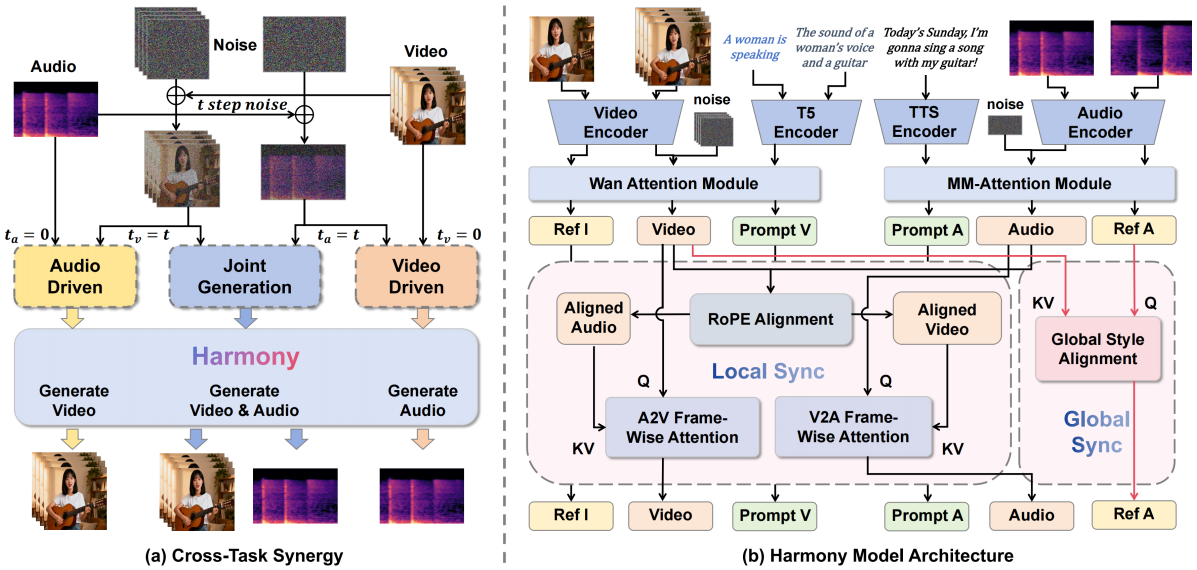


Figure 2. (a) **Mitigating Correspondence Drift with Cross-Task Synergy.** Our training paradigm leverages a supervised audio- and video-driven task to provide a strong alignment signal. This instills robust synchronization features in the model, stabilizing the otherwise stochastic joint generation process. (b) **Overview of the Harmony Model.** The architecture features parallel branches for multimodal inputs. The video stream is conditioned on a reference image and a descriptive prompt. The audio stream is conditioned on a reference audio, an ambient sound description, and a speech transcript. The model then generates a single, synchronized audio-visual result.

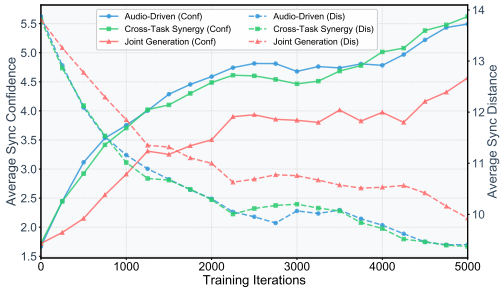


Figure 3. Comparison of the audio-video alignment score among different training strategies.

compare the lip-sync alignment scores of an audio-driven video generation task versus a joint audio-video generation task, utilizing an identical network architecture (detailed in Sec. 3.3). The results reveal that the audio-driven model, conditioned on a clean audio signal, rapidly converges to a high alignment score; in contrast, the joint generation model exhibits markedly slower convergence. This disparity strongly indicates that anchoring one modality with a deterministic, noise-free signal, as implemented in the audio-driven model, provides a stable learning gradient, enabling the cross-modal interaction module to efficiently capture alignment cues.

3.2.2. Cross-Task Synergy

Overview of Cross-Task Synergy. Based on this insight, we propose a novel training framework, *Cross-Task Synergy*, which synergistically combines the standard joint audio-video generation task (the primary task) with audio-driven video and video-driven audio generation tasks. By leveraging the high-quality, noise-free learning signal from the uni-directional supervisory task, our model efficiently learns intricate audio-video correspondences. This pre-learned alignment knowledge then acts as a powerful catalyst, accelerating the convergence and enhancing the final alignment quality of the primary joint generation task.

Dual-Branch Model Architecture. Our model features a dual-branch architecture for video and audio generation. The video branch adapts the pre-trained Wan2.2-5B model [47]. To ensure structural parity, we design a symmetric audio generation branch that synthesizes an audio clip A conditioned on a speech transcript T_s (phonetic content), a descriptive caption T_a (describing the acoustic scene, e.g., vocal emotion or ambient sounds), and a reference audio A_r (timbre). We process these inputs with a multi-encoder setup: an audio VAE [8] encodes A and A_r into latents z_a and z_r. Crucially, departing from prior works [32, 34], we use separate text encoders to preserve phonetic precision: a dedicated speech-encoder [3] for the transcript T_s ($\rightarrow \mathbf{e}_{\text{speech}}$) and a T5 encoder [9] for the descriptive prompt T_a ($\rightarrow \mathbf{e}_{\text{prompt}}$). During denoising, the reference latent z_r is prepended to the noisy target latent $z_{a,t}$, forming a composite input latent $z'_{a,t}$. This composite latent, along with the speech and prompt embeddings, is then processed by a Multi-Modal Diffusion Transformer (MM-DiT) to predict the noise:

$$\hat{\epsilon}_a = \text{MM-DiT}(\text{concat}(z'_{a,t}, \mathbf{e}_{\text{speech}}, \mathbf{e}_{\text{prompt}}), t_a). \quad (2)$$

To facilitate effective cross-modal interaction between the two branches, we instantiate a bidirectional global-local decoupled interaction module at each layer, with further details provided in Sec. 3.3.

Cross-Task Synergy Training. We design a hybrid training strategy that realizes the principle of Cross-Task Synergy. By concurrently training on both the joint generation task and and two deterministic, single-modality-driven tasks (audio-driven video gen and video-driven audio gen), we provide the model with a stable alignment signal to counteract Correspondence Drift.

The audio-driven task conditions video generation on the clean audio latent by setting the audio timestep t_a to 0. Symmetrically, the video-driven task conditions audio generation on the clean video latent by setting the video timestep t_v to 0. The total training objective is a weighted sum of the three corresponding losses:

$$\mathcal{L} = \mathcal{L}_{\text{joint}} + \lambda_v \mathcal{L}^{\text{audio}}_{\text{driven}} + \lambda_a \mathcal{L}^{\text{video}}_{\text{driven}}, \tag{3}$$

where λ_v and λ_a are balancing hyperparameters, and $\mathbf{c}$ represents the set of auxiliary conditions (e.g., text prompts and speech embeddings). The loss components are defined as:

$$\begin{aligned} \mathcal{L}_{\text{joint}} =& ||\epsilon_v - \hat{\epsilon}_v(z_{v,t}, z_{a,t}, \mathbf{c}, t)||^2 \\ &+ ||\epsilon_a - \hat{\epsilon}_a(z_{a,t}, z_{v,t}\mathbf{c}, t)||^2, \\ \mathcal{L}^{\text{audio}}_{\text{driven}} =& ||\epsilon_v - \hat{\epsilon}_v(z_{v,t}, z_{a,0}, \mathbf{c}, t)||^2, \\ \mathcal{L}^{\text{video}}_{\text{driven}} =& ||\epsilon_a - \hat{\epsilon}_a(z_{a,t}, z_{v,0}, \mathbf{c}, t)||^2. \end{aligned} \tag{4}$$

This bidirectional, synergistic training approach enables our model to achieve faster convergence and a superior degree of final audio-video alignment.

3.3. Global-Local Decoupled Interaction Module

Robust audio-video synchronization presents a fundamental tension between two objectives: ***(1)*** precise, fine-grained temporal alignment and ***(2)*** holistic, global style consistency (*e.g.*, emotional tone, ambient features). Prior works [32, 34] often attempt to address both with a single, monolithic mechanism like global cross-attention, which conflates these goals and leads to a suboptimal trade-off. To resolve this, we propose a novel **Global-Local Decoupled Interaction Module** with two specialized components: ***(1)*** a **RoPE-Aligned Frame-wise Attention** module for precise local synchronization, and ***(2)*** a **Global Style Alignment** module for holistic consistency. This decoupled design allows each component to excel at its specific task, resolving the conflict between fine-grained temporal alignment and global style propagation.

3.3.1. RoPE-Aligned Frame-wise Attention

To achieve precise temporal synchronization, we employ a local frame-wise attention strategy, which is more computationally efficient and better suited for fine-grained alignment than global cross-attention. However, a key challenge arises from the mismatched sampling rates of video and audio latents ($T_v \neq T_a$). This discrepancy means a specific event in one modality can occur at a timepoint that falls *between* two discrete frames in the other. A standard attention mechanism, forced to operate on a discrete set of keys, must attend to the nearest but temporally imperfect frame. This forced approximation introduces temporal jitter and fundamentally degrades fine-grained synchronization.

Temporal Alignment via RoPE Scaling. To resolve this mismatch, we introduce an alignment step prior to the attention operation [8]. Our key insight is to unify the temporal coordinate spaces of both modalities by dynamically scaling their Rotary Positional Embeddings (RoPE) [46]. Before attention operation, we rescale the positional indices of the source modality to match the timeline of the target. For instance, in Audio-to-Video (A2V) attention, an audio frame at index j is mapped to a virtual position $j' = j \cdot (T_v/T_a)$ for its RoPE calculation. This ensures their positional encodings are directly comparable, establishing a strong inductive bias for correct temporal correspondence.

Frame-wise Cross-Attention Mechanism. With the latents now temporally aligned in the RoPE space, we apply a symmetric, bidirectional cross-attention mechanism. Each frame's attention is confined to a small, relevant temporal window in the other modality. Taking A2V as an example, given a video latent z_v and an audio latent z_a, we first reshape z_v to expose its temporal dimension (z'_v). For each video frame i, we construct a local context window $C_{a,i}$ from adjacent audio frames. Cross-attention is then applied independently for each video frame, attending to its corresponding audio context window:

$$\begin{aligned} &\Delta z'_v[:, i, :, :] = \text{Cross-Attn}(Q_{v,i}, K_{a,i}, V_{a,i}), \forall i \in [0, T_v\text{-}1], \\ &Q_{v,i} = z'_v[:, i, :, :]W^Q_{v,i}, K_{a,i} = C_{a,i}W^K_{a,i}, V_{a,i} = C_{a,i}W^V_{a,i}. \end{aligned} \tag{5}$$

The Video-to-Audio (V2A) frame-wise alignment operates analogously. The updates are then integrated via residual connections:

$$z_v^{\text{updated}} = z_v + \Delta z'_v, \quad z_a^{\text{updated}} = z_a + \Delta z_a. \tag{6}$$

This RoPE-aligned frame-wise mechanism efficiently enforces mutual temporal synchronization, leveraging the benefits of local attention while correctly handling disparate timescales.

3.3.2. Global Style Alignment

While frame-wise attention excels at establishing fine-grained temporal correspondence, its localized nature inherently limits the propagation of holistic stylistic attributes, such as the overall emotional tone or ambient characteristics, which require a global context to be consistently maintained. Prior methods often rely on a single, monolithic global attention mechanism, which conflates the distinct

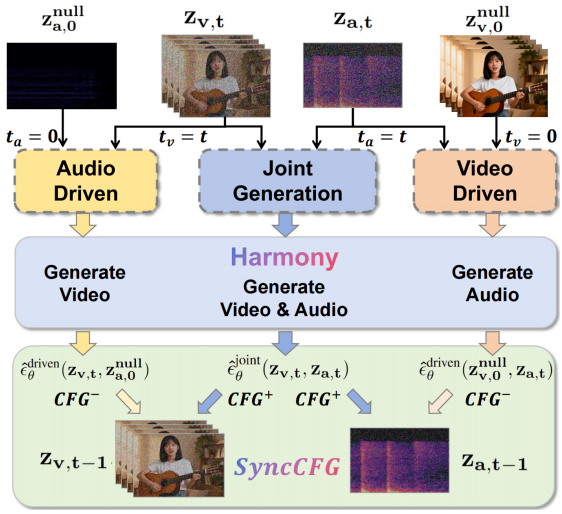


Figure 4. **SyncCFG** employs the mute audio and static video as the negative anchors to capture the synchronization feature, which can effectively enhance the audio-video alignment.

tasks of temporal alignment and global style consistency, overburdening the module. To address this, we propose a principled decoupling: our RoPE-Aligned Frame-wise Attention is exclusively responsible for precise temporal correspondence, while a dedicated **Global Style Alignment** module handles holistic consistency. This separation allows each component to specialize, preventing interference between the two objectives.

Our core insight for global alignment is to leverage the reference audio latent, z_r (which provides speaker identity and timbre), as a compact carrier for style information. Instead of directly modifying the target audio z_a and disrupting its fine-grained denoising, we modulate z_r with the global context from the entire video latent z_v. This is achieved by treating z_r as the query and z_v as the key and value within a residual cross-attention block:

$$\begin{aligned} z_r^{\text{updated}} &= z_r + \text{Cross-Attn}(Q_r, K_v, V_v), \\ Q_r &= z_r W_r^Q, K_v = z_v W_v^K, V_v = z_v W_v^V. \end{aligned} \tag{7}$$

The resulting visually-informed reference audio latent z_r^{updated} is then prepended to the noisy audio latent $z_{a,t}$ (as described in Sec. 3.2.2), allowing the audio generation to condition on a visually-grounded global style. This decoupled design offers a key advantage: by confining the global style injection to the reference latent, it prevents interference between holistic style consistency and precise frame-wise temporal alignment, preserving the stability and fidelity of the final audio generation.

3.4. Synchronization-Enhanced CFG

While Classifier-Free Guidance (CFG) is a powerful technique for conditional generation, its standard application in audio-video synthesis fails to explicitly amplify the crucial correspondence between modalities. To address this, we introduce **Synchronization-Enhanced (SyncCFG)**, a novel scheme that repurposes the guidance mechanism to specifically target and enforce audio-video synchronization. Our approach leverages the dual capabilities—joint generation and audio&video-driven synthesis—acquired during our cross-task training to enhance the alignment signal.

3.4.1. Analysis of Standard Guidance Limitations

The standard CFG formulation in previous work [34] is used to strengthen the conditioning on a text prompt c:

$$\tilde{\epsilon} = \hat{\epsilon}_\theta(z_{v,t}, z_{a,t}, \emptyset_c) + s\Big(\hat{\epsilon}_\theta(z_{v,t}, z_{a,t}, c) - \hat{\epsilon}_\theta(z_{v,t}, z_{a,t}, \emptyset_c)\Big). \tag{8}$$

Here, the guidance pushes the denoising process away from an unconditional (null-text $\emptyset_c$) prediction and towards one that aligns with the text prompt c.

The key limitation of standard CFG is that its guidance is oriented solely towards text-adherence. The guidance vector, computed by contrasting the text-conditioned output with a text-unconditioned one, exclusively strengthens how well the output matches the prompt. This process, however, is agnostic to the internal consistency between audio and video. It provides no mechanism to isolate or amplify the crucial **synchronization** signal between the two streams.

3.4.2. SyncCFG Formulation for Video Guidance

To explicitly compute an alignment-enhancing direction, we aim at isolating the visual dynamics caused by the audio. Our key insight is to design a *more meaningful Negative Anchor* that represents a static baseline—how the video should look in the absence of sound. For instance, for a person speaking, the correct video for a silent audio track would be a still face with a closed mouth.

We achieve this by creating a "silent audio" negative anchor. We leverage the audio-driven pathway of our model to predict the noise for the video latent $z_{v,t}$ conditioned on a "muted" audio input, $z^{null}_{a,0}$. The resulting prediction, $\hat{\epsilon}^{\text{driven}}_\theta(z_{v,t}, z^{null}_{a,0})$, represents the model's expectation for this visually static scene. The guided prediction for the video noise $\tilde{\epsilon}_v$ is then formulated as:

$$\begin{aligned} \tilde{\epsilon}_v = &\hat{\epsilon}^{\text{driven}}_\theta(z_{v,t}, z^{null}_{a,0}) + \\ &s_v\Big(\hat{\epsilon}^{\text{joint}}_\theta(z_{v,t}, z_{a,t}) - \hat{\epsilon}^{\text{driven}}_\theta(z_{v,t}, z^{null}_{a,0})\Big). \end{aligned} \tag{9}$$

The subtraction term isolates the precise visual modifications—such as mouth movements or object impacts—that are directly correlated with the audio. By amplifying this difference, SyncCFG specifically enhances the synchronization between sound and motion.

3.4.3. SyncCFG Formulation for Audio Guidance

Similarly, for audio guidance, we design a null video-based negative anchor to isolate motion-driven sounds. Using a



Figure 5. **Qualitative Comparison** between Harmony and the state-of-the-art methods, including Universe-1 [48] and Ovi [34].

"static video" latent $z_{v,0}^{null}$, we predict a baseline audio signal $\hat{\epsilon}_\theta^{\text{driven}}(z_{v,0}^{null}, z_{a,t})$ that represents the ambient sound of a motionless scene, as the video content is static. The guided prediction for audio noise $\tilde{\epsilon}_a$ is then formulated as:

$$\begin{aligned}\tilde{\epsilon}_a =& \hat{\epsilon}_\theta^{\text{driven}}(z_{v,0}^{null}, z_{a,t}) + \\ & s_a\left(\hat{\epsilon}_\theta^{\text{joint}}(z_{v,t}, z_{a,t}) - \hat{\epsilon}_\theta^{\text{driven}}(z_{v,0}^{null}, z_{a,t})\right).\end{aligned} \tag{10}$$

This approach transforms CFG from a generic conditional amplifier into a targeted mechanism, effectively enforcing fine-grained audio-video correspondence.

4. Experiments

4.1. Experimental Settings

Datasets and Training. Our model is trained on a diverse corpus of over 4 million audio-visual clips, covering both human speech and environmental sounds. The data is aggregated from public sources like OpenHumanVid [29], AudioCaps [25], and WavCaps [39], and supplemented with our own curated high-quality collections. All data is uniformly annotated using Gemini [12]. Our training follows a three-stage curriculum: (1) foundational audio pre-training on all audio data, (2) timbre disentanglement finetuning using multi-utterance speech data, and (3) final cross-task joint audio-visual training. The video branch is initialized from Wan2.2 [47]. The final joint stage is trained for 10,000 iterations with a batch size of 128 and a learning rate of 1e-5. A comprehensive list of datasets, data processing details, and full training hyperparameters are provided in the supplementary material.

Harmony-Bench and Metrics. To facilitate a rigorous evaluation, we introduce **Harmony-Bench**, a new benchmark of 150 test cases designed to assess core audio-visual generation capabilities. It is structured into three 50-item subsets with increasing complexity:

1. **Ambient Sound-Video:** Evaluates temporal alignment of non-speech sounds using AI-generated scenarios conditioned on audio and video captions.
2. **Speech-Video:** Assesses lip-sync and speech quality on a mix of real-world and synthetic multilingual data, conditioned primarily on a transcript.
3. **Complex Scene (Ambient + Speech):** Tests the model's ability to generate and synchronize co-occurring speech and ambient sounds in complex scenes, using a full set of multimodal prompts.

We evaluate performance using a comprehensive suite of automated metrics. **For Video**, we measure *Aesthetic Quality* (aesthetic-predictor-v2-5), *Imaging Quality* (MUSIQ), *Dynamic Degree* (RAFT), *Motion Smoothness* and *Identity Consistency* (DINOv3). **For Audio**, we report *AudioBox-Aesthetics* (PQ, PC, CE, CU), *WER* (Whisper-large-v3), and *IB-A Score*. **For AV-Sync**, we use lip-sync metrics (*Sync-C*, *Sync-D*), and overall consistency (*IB-score*). Further details on the benchmark construction and metric implementations are available in the appendix.

4.2. Comparison on audio-video generation

To evaluate the performance of our model, we compare it with state-of-the-art audio-video generation methods on the three types of datasets (Ambient Sound-Video, Speech-Video, and Complex Scene), including MM-Diffusion [44], JavisDiT [32], UniVerse-1 [48], and Ovi [34]. The quantitative results are shown in Tab. 2. Our model, Harmony, demonstrates a highly competitive performance, achieving state-of-the-art or comparable results in both video quality

Table 2. **Quantitative comparison** with state-of-the-art methods averaging on ambient sound-video, speech-video, and complex scene-video generation. We evaluate performance across three categories: video quality, audio fidelity, and audio-visual synchronization. Best results are in **bold**, second-best are underlined. For more comprehensive and detailed evaluations, please refer to the supplementary material. Some metrics could not be generated for MM-Diffusion, as it is an unconditional generation model.

Method	Video Quality & Coherence					Audio Fidelity & Quality						Audio-Visual Synchronization			
	AQ ↑	IQ ↑	DD ↑	MS ↑	ID ↑	PQ ↑	PC ↓	CE ↑	CU ↑	WER ↓	IB-A ↑	Sync-C ↑	Sync-D ↓	DeSync ↓	IB ↑
MM-Diffusion [44]	0.32	0.43	0.13	0.99	-	5.37	4.07	4.27	**5.89**	-	-	-	-	-	0.12
JavisDiT [32]	0.34	0.53	**0.38**	0.99	0.38	5.46	2.24	3.19	4.54	1.00	**0.14**	0.89	11.62	1.13	0.18
UniVerse-1 [48]	0.52	**0.67**	0.24	0.99	0.89	5.52	2.13	3.63	4.84	0.24	0.07	0.97	10.71	1.10	0.12
Ovi [34]	0.57	0.65	0.34	0.99	0.90	6.19	2.13	4.44	5.84	0.49	0.12	4.04	9.62	1.14	0.18
Harmony (Ours)	**0.59**	0.65	0.36	0.99	**0.91**	**6.39**	**2.05**	**4.73**	5.67	**0.15**	0.12	**5.61**	**7.53**	**0.92**	**0.19**

(e.g., AQ, DD, ID) and audio fidelity (e.g., PC, PQ). Most notably, its primary advantage lies in audio-visual synchronization. Harmony significantly outperforms all baselines on key synchronization metrics, achieving the highest Sync-C score of **5.61** and the lowest (best) Sync-D score of **7.53**. This substantial improvement in temporal alignment directly validates the effectiveness of our proposed cross-task synergy mechanism in enhancing cross-modal coherence.

We provide qualitative comparisons with UniVerse-1 and Ovi in Fig. 5. In the talking head example (left), both competing methods fail to produce synchronized lip movements. For the music-driven case (right), their limitations persist: UniVerse-1 generates irrelevant noise, while Ovi produces audio that, while musically correct, is less dynamic—a fact reflected in its simpler waveform. Visually, both methods yield videos with minimal motion. In contrast, our Harmony generates a fluid video of a person playing the mandolin with motions that are dynamically synchronized with the rich, corresponding music, as evidenced by the more complex audio waveform.

4.3. Ablation Studies

We conduct a comprehensive ablation study to validate our core components, with the results presented in Tab. 3. In this study, we train all ablated models on the human speech dataset and evaluate their audio-visual alignment. Our baseline model replaces the proposed Global-Local Decoupled Interaction (GLDI) module with a standard global cross-attention mechanism, similar to Ovi [34], and is trained without cross-task synergy (CTS). As shown in the table, progressively integrating our contributions yields consistent improvements. First, introducing the GLDI module demonstrates the benefit of decoupling local and global interactions. This is further enhanced by the RoPE Alignment, which effectively resolves timescale mismatches and boosts fine-grained synchronization (Sync-C from 4.29 to 4.80). Subsequently, the Cross-Task Synergy (CTS) training strategy further refines the model's alignment capabilities. Finally, applying the Synchronization-Enhanced CFG (SyncCFG) during inference provides the most substantial performance gain, catapulting Sync-C from 5.09 to **6.51**. This systematic improvement validates that each component of Harmony is crucial for achieving the state-of-the-art audio-video synchronization performance.

Table 3. Ablation study on the core components of Harmony. We start with a baseline and progressively add each module: Global-local decoupled interaction module (GLDI), RoPE Alignment (RoPE), Cross-Task Synergy (CTS) , and Synchronization-Enhanced CFG (SyncCFG). Note that this experiment is evaluated on the human-speech dataset; therefore, it has different synchronization results compared to Tab. 2.

Model Structure		Methodology		Synchronization Metrics		
GLDI	RoPE	CTS	SyncCFG	Sync-C ↑	Sync-D ↓	IB ↑
				4.20	10.93	0.13
✓				4.29	10.67	0.14
✓	✓			4.80	10.30	0.14
✓	✓	✓		5.09	10.16	0.15
✓	✓	✓	✓	**6.51**	**8.63**	**0.18**

5. Conclusion

In this work, we presented Harmony, a novel framework addressing the audio-visual synchronization gap in generative models. We find this gap stems from key methodological flaws: Correspondence Drift, an architectural conflict between global style and local timing, and the limitations of standard CFG for cross-modal alignment. To address these issues, Harmony introduces three core components. Cross-Task Synergy training instills robust alignment priors to counteract drift. A Global-Local Decoupled Module resolves the architectural conflict by handling style and timing separately. Finally, our novel Synchronization-Enhanced CFG (SyncCFG) provides an explicit mechanism to amplify the alignment signal during inference. Our experiments validate that Harmony establishes a new state-of-the-art in audio-video synchronization, proving more effective than simply scaling up models. We believe this work provides a strong foundation for a new generation of accessible and well-aligned audio-visual models.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No. 625B2115, 62302297, 62272447, 62472285), the Fundamental Research Funds for the Central Universities (YG2023QNB17, YG2024QNA44).

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[2] Junhao Chen, Mingjin Chen, Jianjin Xu, Xiang Li, Junting Dong, Mingze Sun, Puhua Jiang, Hongxiang Li, Yuhang Yang, Hao Zhao, et al. Dancetogether! identity-preserving multi-person interactive video generation. *arXiv preprint arXiv:2505.18078*, 2025. 3

[3] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024. 4

[4] Yiyang Chen, Xuanhua He, Xiujun Ma, and Yue Ma. Contextflow: Training-free video object editing via adaptive context enrichment. *arXiv preprint arXiv:2509.17818*, 2025. 3

[5] Yuheng Chen, Teng Hu, Jiangning Zhang, Zhucun Xue, Ran Yi, and Lizhuang Ma. Instancev: Instance-level video generation. *arXiv preprint arXiv:2511.23146*, 2025. 3

[6] Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025. 3

[7] Yinan Chen, Jiangning Zhang, Teng Hu, Yuxiang Zeng, Zhucun Xue, Qingdong He, Chengjie Wang, Yong Liu, Xiaobin Hu, and Shuicheng Yan. Ivebench: Modern benchmark suite for instruction-guided video editing assessment. *arXiv preprint arXiv:2510.11647*, 2025. 3

[8] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28901–28911, 2025. 2, 4, 5

[9] Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*, 2023. 4

[10] Google DeepMind. Veo3. https://deepmind.google/models/veo/, 2025. 2

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[12] Google. Gemini. https://gemini.google.com/, 2025. 7

[13] 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. In *ICLR*, 2024. 3

[14] 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. *arXiv preprint arXiv:2412.15191*, 2024. 2

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 3

[17] Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789*, 2024. 3

[18] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025. 3

[19] Teng Hu, Zhentao Yu, Zhengguang Zhou, Jiangning Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Polyvivid: Vivid multi-subject video generation with cross-modal interaction and enhancement. *arXiv preprint arXiv:2506.07848*, 2025. 3

[20] Teng Hu, Jiangning Zhang, Zihan Su, and Ran Yi. Ultragen: High-resolution video generation with hierarchical attention. *arXiv preprint arXiv:2510.18775*, 2025. 3

[21] Teng Hu, Jiangning Zhang, Ran Yi, Hongrui Huang, Yabiao Wang, and Lizhuang Ma. High-efficient diffusion model fine-tuning with progressive sparse low-rank adaptation. In *13th International Conference on Learning Representations, ICLR 2025*, pages 92066–92078. International Conference on Learning Representations, ICLR, 2025. 3

[22] Teng Hu, Jiangning Zhang, Ran Yi, Jieyu Weng, Yabiao Wang, Xianfang Zeng, Zhucun Xue, and Lizhuang Ma. Improving autoregressive visual generation with cluster-oriented token prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9351–9360, 2025. 3

[23] Masato Ishii, Akio Hayakawa, Takashi Shibuya, and Yuki Mitsufuji. A simple but strong baseline for sounding video generation: Effective adaptation of audio and video diffusion models for joint generation. *arXiv preprint arXiv:2409.17550*, 2024. 2, 3

[24] Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 193–203, 2025. 3

[25] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 119–132, 2019. 7

[26] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3

[27] Mingi Kwon, Joonghyuk Shin, Jaeseok Jung, Jaesik Park, and Youngjung Uh. Jam-flow: Joint audio-motion synthesis with flow matching. *arXiv preprint arXiv:2506.23552*, 2025. 2

[28] Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Long Chen. Dispose: Disentangling pose guidance for controllable human image animation. *arXiv preprint arXiv:2412.09349*, 2024. 3

[29] Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7752–7762, 2025. 7

[30] Sen Liang, Zhentao Yu, Zhengguang Zhou, Teng Hu, Hongmei Wang, Yi Chen, Qin Lin, Yuan Zhou, Xin Li, Qinglin Lu, et al. Omniv2v: Versatile video generation and editing via dynamic content manipulation. *arXiv preprint arXiv:2506.01801*, 2025. 3

[31] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 3

[32] Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025. 2, 3, 4, 5, 7, 8

[33] Zeqian Long, Mingzhe Zheng, Kunyu Feng, Xinhua Zhang, Hongyu Liu, Harry Yang, Linfeng Zhang, Qifeng Chen, and Yue Ma. Follow-your-shape: Shape-aware image editing via trajectory-guided region control. *arXiv preprint arXiv:2508.08134*, 2025. 3

[34] Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025. 2, 3, 4, 5, 6, 7, 8

[35] Boyao Ma, Yuanping Cao, and Lei Zhang. Decoupled two-stage talking head generation via gaussian-landmark-based neural radiance fields. *Computational Visual Media*, 2025. 3

[36] Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025. 3

[37] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6018–6026, 2025. 3

[38] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 3

[39] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. 7

[40] Openai. Sora2. https://openai.com/zh-Hans-CN/index/sora-2/, 2025. 2

[41] Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *arXiv preprint arXiv:2505.21448*, 2025. 2

[42] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 3

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[44] 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*, pages 10219–10228, 2023. 2, 3, 7, 8

[45] 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. *arXiv preprint arXiv:2508.16930*, 2025. 2

[46] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[47] Ang Wang, Baole Ai, and et al Bin Wen. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 4, 7

[48] 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*, 2025. 2, 3, 7, 8

[49] 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*, pages 10486–10495, 2025. 3

[50] Ruiyan Wang, Teng Hu, Kaihui Huang, Zihan Su, Ran Yi, and Lizhuang Ma. Poseanything: Universal pose-guided

video generation with part-aware temporal coherence. *arXiv preprint arXiv:2512.13465*, 2025. 3

[51] Xihua Wang, Ruihua Song, Chongxuan Li, Xin Cheng, Boyuan Li, Yihan Wu, Yuyue Wang, Hongteng Xu, and Yunfeng Wang. Animate and sound an image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23369–23378, 2025. 2

[52] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[53] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7161, 2024. 3

[54] Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025. 3

[55] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16284–16295, 2024. 3

[56] Yuhang Yang, Wei Zhai, Chengfeng Wang, Chengjun Yu, Yang Cao, and Zheng-Jun Zha. Egochoir: Capturing 3d human-object interaction regions from egocentric views. *Advances in Neural Information Processing Systems*, 37: 54529–54557, 2024.

[57] Yuhang Yang, Fengqi Liu, Yixing Lu, Qin Zhao, Pingyu Wu, Wei Zhai, Ran Yi, Yang Cao, Lizhuang Ma, Zheng-Jun Zha, et al. Sigman: Scaling 3d human gaussian generation with millions of assets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5122–5133, 2025. 3

[58] 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. 3

[59] Chenxu Zhang, Chao Wang, Jianfeng Zhang, Hongyi Xu, Guoxian Song, You Xie, Linjie Luo, Yapeng Tian, Jiashi Feng, and Xiaohu Guo. Magictalk: Implicit and explicit correlation learning for diffusion-based emotional talking face generation. *Computational Visual Media*, 2025. 3

[60] 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. *arXiv preprint arXiv:2511.03334*, 2025. 3

[61] Jiangning Zhang, Junwei Zhu, Teng Hu, Yabiao Wang, Donghao Luo, Weijian Cao, Zhenye Gan, Xiaobin Hu, Zhucun Xue, and Chengjie Wang. Transform trained transformer: Accelerating naive 4k video generation over 10 times. *arXiv preprint arXiv:2512.13492*, 2025. 3

[62] Youliang Zhang, Zhengguang Zhou, Zhentao Yu, Ziyao Huang, Teng Hu, Sen Liang, Guozhen Zhang, Ziqiao Peng, Shunkai Li, Yi Chen, et al. Making avatars interact: Towards text-driven human-object interaction for controllable talking avatars. *arXiv preprint arXiv:2602.01538*, 2026. 3