Unison: Harmonizing Motion, Speech, and Sound for Human-Centric Audio-Video Generation

Shihao Cheng[1,4†], Jiaxu Zhang[2†], Quanyue Song[3,4], Shansong Liu[4‡], Zhizhi Guo[4], Xiao-Lei Zhang[5], Chi Zhang[5], Xuelong Li[5], and Zhigang Tu[1✉]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
[2] ByteDance, China
[3] State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
[4] China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd., China
[5] Institute of Artificial Intelligence (TeleAI), China Telecom, China

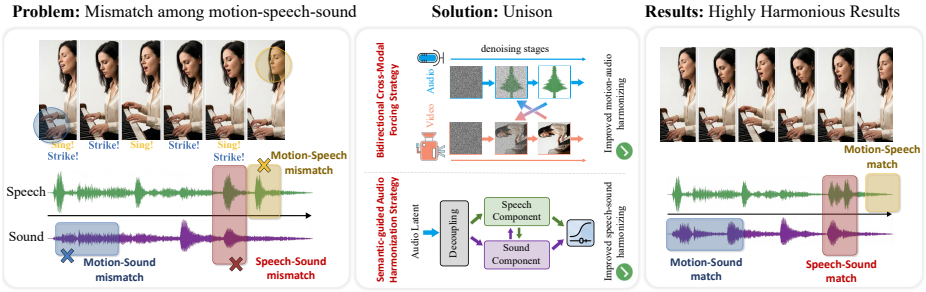

Fig. 1: Overview of the key challenges and our approach. Left: Two major misalignments in human-centric audio-video generation—(a) speech-sound interference within the audio stream and (b) motion-audio desynchronization, including motion-speech and motion-sound mismatches. **Middle:** Our proposed Unison framework addresses these issues through a semantic-guided harmonization strategy that decouples and adaptively balances speech-sound components via interactive refinement, and a progressive audio-video forcing strategy that enforces rigorous temporal alignment across modalities. **Right:** These designs jointly reduce speech dominance, preserve sound effects, and achieve coherent synchronization between motion and audio.

Abstract. Motion, speech, and sound effects are fundamental elements of human-centric videos, yet their heterogeneous temporal characteristics make joint generation highly challenging. Existing audio-video generation models often fail to maintain consistent alignment across these modalities, leading to noticeable mismatches between motion, speech, and environmental sounds. We present Unison, a unified framework that explicitly promotes coherence across the motion, speech, and sound modalities. Within the audio stream, Unison employs a semantic-guided harmonization strategy that decouples the generation of speech and sound-effect components. Leveraging bidirectional audio cross-attention and

† Equal contribution.
‡ Project leader.
✉ Corresponding author.

semantic-conditioned gating for semantic-driven adaptive recomposition, this approach effectively mitigates speech dominance and enhances acoustic clarity. For audio–motion synchronization, we propose a bidirectional cross-modal forcing strategy where the cleaner modality guides the noisier one through decoupled denoising schedules, reinforced by a progressive stabilization strategy. Extensive experiments demonstrate that Unison achieves state-of-the-art performance in both audio perceptual quality and cross-modal synchronization, highlighting the importance of explicit multimodal harmonization in human-centric video generation.

Keywords: Human-Centric Video Generation · Joint Audio-Video Synthesis · Multimodal Harmonization

1 Introduction

The rapid evolution of generative AI [2, 5, 12, 18–20, 24, 31, 39, 41, 45, 51–59] has catalyzed significant breakthroughs in synchronized audio-video generation. Proprietary models such as Veo 3 [11], Sora 2 [26], and Wan 2.5 [39], alongside open-source models like Ovi [24], UniAVGen [50], and LTX-2 [12], have demonstrated remarkable creative potential, inspiring users across online communities.

Although most open-source approaches have continuously improved visual fidelity [16,40], they still exhibit poor alignment between audio and video streams. Recent methods, such as Harmony [13], employ fine-grained cross-attention mechanisms to enable frame-level synchronization. However, these approaches primarily focus on architectural modifications and lack explicit consistency constraints. In contrast, closed-source commercial systems leverage large-scale datasets to achieve fully data-driven synchronization [26], facilitating visually and temporally coherent outputs. Nevertheless, the generated audio often lacks acoustic richness and balance: speech components tend to dominate the entire mix, suppressing ambient and contextual sounds that are essential for perceptual realism. This imbalance leads to flattened auditory scenes and weakens the overall sense of immersion, with these issues becoming especially evident in human-centric video generation.

As shown in the left part of Figure 1, our evaluation of representative audio–visual generation models reveals two fundamental challenges. (1) Intra-audio interference: speech and non-speech sound effects often overlap within the same audio stream, creating perceptual masking where structured speech components dominate and obscure transient environmental sounds—especially in human-centric scenarios such as “singing while playing an instrument." This indicates that current models lack mechanisms to disentangle and balance heterogeneous acoustic elements. (2) Cross-modal misalignment: motion and audio generation are typically linked only through architectural cross-attention, without explicit synchronization objectives during training. Consequently, audio and video exchange information implicitly but fail to form consistent temporal correspondence, leading to desynchronized gestures, lip motions, and acoustic onsets. These issues collectively limit the realism and coherence of human-centered video generation.

To overcome the aforementioned challenges, we propose Unison, a unified framework designed to generate motion, speech, and sound in a coherent and synchronized manner. To resolve acoustic interference, Unison implements a semantic-guided harmonization strategy that decouples speech and sound-effect generation. This approach integrates bidirectional audio cross-attention (Bi-ACA) for interactive refinement and semantic-conditioned gating (SCG) for semantic-driven adaptive recomposition. By modulating component interplay, Unison ensures cross-modal consistency while enabling dedicated modeling for each acoustic entity, effectively preventing speech from overshadowing subtle environmental sounds. Second, to tackle motion–audio desynchronization, Unison employs a bidirectional cross-modal forcing paradigm that explicitly aligns denoising progress between modalities. By allowing the temporally advanced modality to guide the lagging one during training, the model learns stable and causally consistent motion–audio coordination without relying solely on cross-attention. Together, these designs establish a principled way of generating human-centric videos where motion, speech, and acoustic context evolve in Unison–yielding perceptually balanced and temporally synchronized results.

Extensive experiments verify that Unison not only enhances overall generation fidelity but also achieves markedly better harmony among motion, speech, and sound. The resulting videos exhibit more balanced auditory layers and tighter temporal correspondence between modalities, confirming that both spectral disentanglement and cross-modal alignment substantially improve perceptual realism. Moreover, the framework's symmetric design naturally enables bidirectional generation—it can synthesize video conditioned on audio or, conversely, generate audio from visual cues—demonstrating its flexibility as a unified model for multimodal content creation.

The main contributions are summarized as follows:

- We propose Unison, a unified audio–video generation framework that produces motion, speech, and sound in a coherent and harmonized manner. Unlike prior models that loosely couple audio and visual streams, Unison explicitly enforces cross-modal consistency, achieving balanced and temporally aligned outputs.
- We propose a semantic-guided harmonization strategy that decouples speech and sound-effect generation to resolve acoustic interference. By integrating Bidirectional Audio Cross-Attention (Bi-ACA) for interactive refinement and Semantic-Conditioned Gating (SCG) for semantic-driven adaptive recomposition, resulting in controllable and acoustically layered audio generation.
- We introduce a bidirectional cross-modal forcing strategy that strengthens temporal dependencies between audio and motion. By allowing the modality in a clearer denoising state to guide the other, Unison achieves robust synchronization and stable training, effectively bridging the gap between motion and acoustics in human-centric video generation.

2 Related Work

2.1 Joint Audio-Video Generation

Early joint audio-video generation, exemplified by MM-Diffusion [29], utilized U-Net architectures for ambient sound synthesis, a scope later expanded by large-scale dataset training [21,61]. While UniVerse-1 [40] and Ovi [24] recently introduced human speech generation, their reliance on implicit global alignment often compromises temporal precision. Advanced mechanisms [12, 13, 50] have since achieved frame-level synchronization via fine-grained cross-attention. Nonetheless, these methods typically employ unidirectional conditioning, leaving bidirectional cross-modal reinforcement and the acoustic interference between speech and effects largely unaddressed.

Recent commercial systems, including Veo 3 [11], Sora 2 [26], and Wan 2.5 [39], have achieved cinematic audio-video synchronization. Nevertheless, when synthesizing scenes with concurrent dense speech and sound effects, these models often prioritize human voices, leading to the suppression or overriding of essential environmental acoustics. To address this, Unison implements a semantic-guided harmonization strategy that decouples speech and sound-effect generation. By isolating these acoustic components at the generative source, this approach resolves speech dominance and ensures intrinsic acoustic clarity.

2.2 Diffusion Forcing

Originally pioneered by Chen [3] to enable sequential modeling with independent per-step noise levels, the diffusion forcing paradigm has been extended to transformer-based architectures [34], self-corrective long-term synthesis [14], and sliding-window video generation [23, 35]. While these advancements effectively mitigate exposure bias and background forgetting in unimodal sequences, they lack a systematic mechanism to capture bidirectional audio–visual dependencies across heterogeneous denoising processes. Unison generalizes this principle to a multimodal setting by introducing a cross-modal forcing strategy. This approach enables audio and video signals to mutually guide their respective denoising trajectories, thereby reinforcing cross-modal interdependencies for superior audiovisual synchronization.

3 Method

We introduce Unison, a unified audio-video generation framework designed to transform text prompts into synchronized audiovisual content that harmonizes motion, speech, and sound effects. Figure 2 illustrates our dual-branch architecture. We formulate Unison's processes as follows.

$$\text{Unison}(\kappa, \tau, c_s, c_a) \mapsto (\nu, \alpha). \tag{1}$$

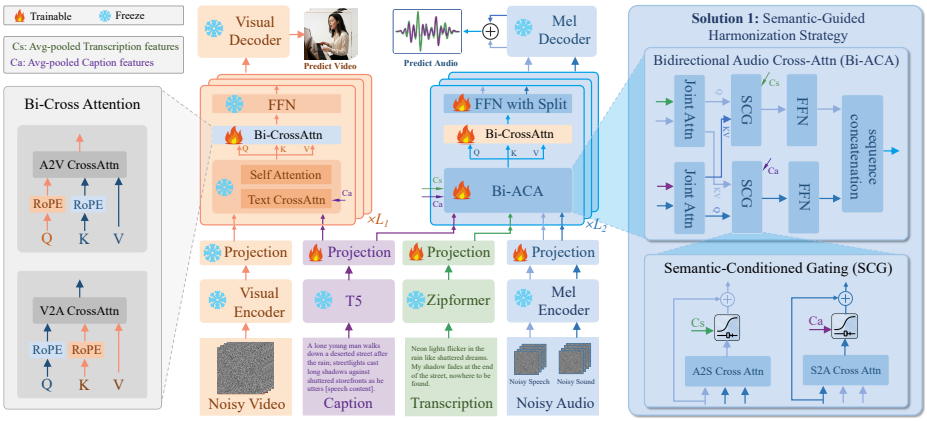

Fig. 2: Overview of Unison. Unison couples a video branch and an audio branch via bidirectional cross-attention. The audio branch employs a Semantic-Guided Harmonization Strategy for independent speech and sound-effect generation, utilizing a Bidirectional Audio Cross-Attention (Bi-ACA) module to mutually refine speech and sound-effect features, effectively enhancing their respective clarity. At each interaction node, a Semantic-Conditioned Gating (SCG) mechanism is employed that balances speech and sound-effect contributions based on c_s and c_a.

where ν and α denote the output video and audio, respectively. The inputs consist of the caption text κ, the transcription text τ, and their corresponding average-pooled feature representations, c_a and c_s.

Unison employs a dual-branch architecture comprising a video branch based on Wan2.2-5B [39] and an audio branch built upon MMAudio [6], with the latter augmented by Zipformer [46, 62] to enable speech generation. The video branch comprises 29 layers (L_1=29) and the audio branch 23 layers (L_2=23). Integration between these branches is achieved through frame-level bidirectional cross-attention, utilizing video and audio latents as mutual queries to facilitate seamless cross-modal information exchange.

To achieve word-level lip-sync, we enforce synchronization across three dimensions. At the architecture level, cross-attention aligns features within a three-frame window using a stride of one, and only the middle frame's representation is retained during restoration. At the data level, we introduce a lip-filtering operator that first detects the number and location of faces. SyncNet [7] is then applied exclusively within these bounding boxes to verify alignment, which naturally filtering out clips with unsynchronized speech and lip movements, as well as off-screen voice-overs. Finally, we design a Bidirectional Cross-Modal Forcing Strategy to enhance audio-visual synchronization by reinforcing the mutual dependence between the two modalities during training, as detailed in Section 3.3.

3.1 Preliminaries

Flow Matching (FM) [20] provides a simulation-free framework for training continuous normalizing flows by regressing a velocity field $v_\theta(x_t, t)$ onto a target

vector field $u_t(x)$ that transports samples from a prior p_0 to the data distribution p_1. For a conditional probability path $p_t(x|x_1)$ with optimal transport, $x_t = (1-t)x_0 + tx_1$ and $u_t(x_t|x_1) = x_1 - x_0$. The Conditional Flow Matching (CFM) objective is:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,x_1,x_t}\big[\|v_\theta(x_t,t) - (x_1 - x_0)\|^2\big]. \tag{2}$$

By defining a deterministic mapping via an associated ODE, FM offers straighter sampling trajectories and improved inference efficiency with fewer function evaluations compared to score-based diffusion.

Diffusion Forcing (DF) [3] extends diffusion to sequences by assigning an *independent* noise level to each element. For a sequence $\mathbf{x}_{1:L}$, it samples per-element timesteps $\mathbf{t} = (t_1, \ldots, t_L)$, training the model with a heterogeneous-noise objective:

$$\mathcal{L}_{\text{DF}} = \mathbb{E}_{\mathbf{x},\mathbf{t}}\Big[\sum_{i=1}^{L} \big\|\mathcal{T}(x_i,t_i) - f_\theta(\mathbf{x}_{1:L,\mathbf{t}},\mathbf{t})_i\big\|^2\Big], \tag{3}$$

where $\mathcal{T}(x_i,t_i)$ is the target and $f_\theta(\cdot)_i$ denotes the prediction for element i conditioned on the full heterogeneous sequence. As a diffusion analogue of next-token prediction, DF allows for flexible causal or bidirectional context windows, effectively reducing train–test mismatch and enhancing long-horizon stability by mitigating error accumulation.

3.2 Speech-Sound Coordination

We extend MMAudio [6] by integrating Zipformer [46,62] to empower the model with robust speech generation capabilities. Reflecting the inherent structural divergence between speech and sound effects, we propose a Semantic-Guided Harmonization Strategy that decouples the generation process into separate speech and sound-effect streams to ensure high-fidelity synthesis and eliminate acoustic ambiguity. Under this strategy, a Bidirectional Audio Cross-Attention (Bi-ACA) module facilitates hierarchical feature interaction between these disentangled components, while a Semantic-Conditioned Gating (SCG) mechanism adaptively modulates their synthesis based on captions and transcriptions.

Bidirectional Audio Cross-Attention (Bi-ACA). To enforce acoustic coherence while maintaining structural segregation, the audio latent is represented as a dual-stream tensor $h \in \mathbb{R}^{B\times 2\times N\times D}$, comprising speech and sound-effect components. The source audio is decoupled and encoded into two temporally aligned sequences, h^{sp} and h^{sfx}. Within each Transformer block, Bi-ACA facilitates mutual context exchange via bidirectional cross-attention before merging them for synchronized global modeling. Specifically, the latents are concatenated along the sequence dimension:

$$h_{\text{joint}} = \text{Concat}([\tilde{h}^{sp}, \tilde{h}^{sfx}], \text{dim} = N) \in \mathbb{R}^{B\times 2N\times D}. \tag{4}$$

To ensure precise temporal synchronization, both streams reuse identical temporal indices for Rotary Positional Embeddings (RoPE). To distinguish between

the two modalities within the shared self-attention layers, we incorporate a modality-specific learnable bias, preventing semantic confusion while allowing for joint acoustic modeling. After processing h_{joint} through the shared Transformer backbone, the representation is bifurcated at the block's exit to recover the independent generation trajectories:

$$h^{sp}, h^{sfx} = \text{Split}(h_{\text{joint}}, \text{dim} = N). \tag{5}$$

This interact-merge-split cycle enables the two streams to benefit from shared global context while maintaining their unique structural characteristics.

Bi-ACA executes *bidirectional* cross-attention between the two streams across multiple decoder layers (Fig. 2, right), facilitating intra-audio context exchange while preserving their structural independence:

$$\tilde{h}^{[sp|sfx]} = h^{[sp|sfx]} + \text{Attn}(\text{LN}(h^{[sp|sfx]}), \text{LN}(h^{[sfx|sp]})) \tag{6}$$

where $\text{Attn}(Q, K, V)$ follows the standard multi-head cross-attention mechanism and LN denotes layer normalization. While individual speech or sound-effect streams may lack sufficient acoustic grounding when performing cross-attention with video tokens, Bi-ACA provides mutual audio-to-audio priors. This ensures that the synthesized components are not only visually aligned but also acoustically consistent, preventing the high-fidelity synthesis from degrading due to isolated feature mapping.

Semantic-Conditioned Gating (SCG). To preclude acoustic interference and phonetic degradation from unconstrained context injection, SCG adaptively regulates cross-stream updates via modality-specific semantic priors. Global semantic vectors c_s and c_a are derived through average pooling of transcription and caption sequences to predict dual gating coefficients $[g^{sp}, g^{sfx}] = \sigma(\text{MLP}([c_s; c_a]))$. These coefficients function as semantic filters to prioritize salient content. In narration-dominant scenes, SCG attenuates the influx of sound-effect features to safeguard phonetic purity. The mechanism further modulates the gating to accommodate intricate environmental layers in complex soundscapes like musical performances. The gated updates are formulated as:

$$h_{\text{out}}^{[sp|sfx]} = h^{[sp|sfx]} + g^{[sp|sfx]} \cdot \text{Attn}(\text{LN}(h^{[sp|sfx]}), \text{LN}(h^{[sfx|sp]})) \tag{7}$$

The coefficients $g \in [0, 1]$, constrained by the sigmoid function, act as semantic valves to regulate information flow. For instances with prominent c_s (e.g., dense narration), SCG adaptively suppresses g^{sp} to shield the speech stream from SFX interference, preserving phonetic purity. Conversely, in scenarios dominated by c_a (e.g., complex soundscapes), it amplifies cross-stream influence to enrich non-speech acoustics, thereby achieving a balanced and controllable audio recomposition.

Audio Stream Supervision. To enforce explicit supervision within the decoupled architecture, we leverage Mel-Roformer to disentangle mixed audio into

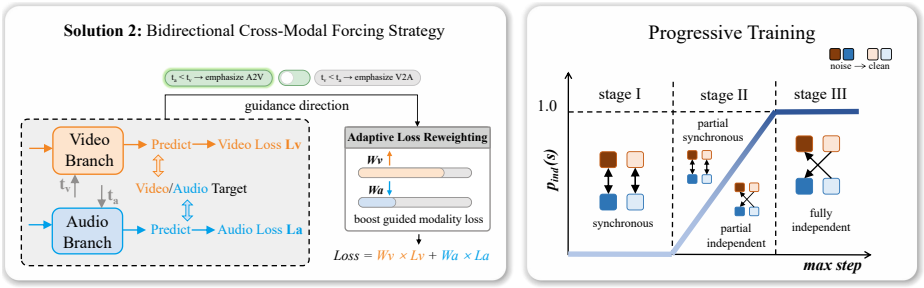

Fig. 3: Bidirectional Cross-Modal Forcing strategy for audio-visual alignment. By decoupling diffusion timesteps during training, our approach enables mutual guidance where the modality at a lower noise level provides enhanced conditioning to steer the denoising process of the noisier counterpart. To ensure stable optimization, a three-stage curriculum is employed, progressing from synchronous warmup to partial and eventually full temporal independence.

high-fidelity speech and sound-effect (SFX) components, which are encoded as ground-truth latents z_1^{sp} and z_1^{sfx}. During training, the dual-stream decoder is optimized via independent flow-matching losses: $\mathcal{L}_{\text{dual}} = \mathcal{L}_{\text{CFM}}^{\text{sp}} + \mathcal{L}_{\text{CFM}}^{\text{sfx}}$. This explicit per-stream supervision eliminates acoustic ambiguity and source interference, ensuring that the decoupled representations remain independent and precisely aligned with the visual modality.

3.3 Motion-Audio Coordination

Regarding the structural design of audio-visual alignment, we follow established methods [13, 50] and employ frame-level cross-attention as illustrated in the left part of Figure 2. Regarding the training strategy for audio-visual alignment, unlike prior works [13, 50] using identical timesteps, we independently sample timesteps for each modality. This allows the "cleaner" modality to guide the "noisier" one, forcing the model to rely more heavily on cross-modal information. Consequently, this bidirectional guidance strengthens the mutual dependency between audio and video.

Bidirectional Cross-Modal Forcing. Inspired by *Diffusion Forcing* [3], we decouple the denoising timesteps for video and audio, allowing each modality to evolve through independent noise levels rather than being strictly synchronized. Specifically, we separately sample t_v and t_a per iteration to generate their respective noisy latents, with the audio branch mapped to the $[0, 1]$ interval. This asynchronous schedule enables the model to learn cross-modal dependencies across varying corruption levels.

Our core premise is that smaller timesteps yield *cleaner* states with superior semantic reliability. Accordingly, we define a per-sample direction indicator $d = \mathbb{I}[t_a < t_v]$, where $d = 1$ (or $d = 0$) identifies audio (or video) as the leading modality. Rather than altering the fusion mechanism itself, d dynamically designates the *student* modality under noise-mismatched conditions. By upweighting

the loss for the noisier branch, we compel it to extract stronger supervisory signals from the cleaner counterpart via cross-modal conditioning.

To prevent optimization instability from extreme noise gaps, we constrain $|t_v - t_a| \leq \Delta_{\max}$ (with $\Delta_{\max} = 0.25$) and compute direction-aware loss weights governed by a guidance strength $\lambda = 0.5$:

$$w_v = 1 + \lambda \cdot d, \quad w_a = 1 + \lambda \cdot (1 - d). \tag{8}$$

The full TI2AV (Text-and-Image-to-Audio-Video) training objective is then defined as the direction-weighted sum of per-branch flow-matching losses:

$$\mathcal{L}_{\text{TI2AV}} = w_v \cdot \mathcal{L}_{\text{CFM}}^{v} + w_a \cdot (\mathcal{L}_{\text{CFM}}^{\text{sp}} + \mathcal{L}_{\text{CFM}}^{\text{sfx}}), \tag{9}$$

where $\mathcal{L}_{\text{CFM}}^{v}$ is the video flow-matching loss, and $\mathcal{L}_{\text{CFM}}^{\text{sp}}$, $\mathcal{L}_{\text{CFM}}^{\text{sfx}}$ are the per-stream audio losses from Sec. 3.2. This strategy prioritizes the learning signal for the more heavily corrupted modality, significantly enhancing motion-audio correspondence across disparate convergence rates.

Progressive Training Strategy. Directly optimizing with fully independent timesteps precipitates training instability due to pronounced cross-modal noise disparities and stochastic oscillations in the leading direction d. We mitigate this via a three-stage curriculum: (i) Synchronous Warmup, enforcing $t_v = t_a$ to establish foundational alignment; (ii) Incremental Decoupling, activating independent sampling with probability $p_{\text{ind}}(s)$ while constraining $|t_v - t_a| \leq 0.25$; and (iii) Full Independence, employing unconstrained decoupled updates. To stabilize this manifold, we apply direction-aware loss reweighting starting from Phase II, upscaling the video loss for $d = 1$ (cleaner audio) and prioritizing the audio loss for $d = 0$ (cleaner video). This progression ensures a robust optimization trajectory by gradually introducing increasingly complex cross-modal noise configurations.

4 Experiments

4.1 Experimental Setup

Training Corpora. Our training framework leverages a diverse collection of audio-visual and audio-only corpora, incorporating Mel-Roformer [42] to decouple both audio-visual tracks and mixed audio streams into independent speech and sound-effect components. For joint audio-visual training, we aggregate several large-scale open-source datasets including OpenHumanVid [17], HDTF [60], VFHQ [43], CelebV-Text [47], and VGGSound [4]. Regarding the audio branch training, we curate a comprehensive repository spanning speech, sound effects, music, and singing performances. Sound effects are collected from YouTube-8M [1], AudioSet [9], and WavCaps [25]. Music data is derived from VidMuse [37], and the singing portion is primarily extracted from the Yue collection [48, 49]. We also include internal speech data to further enrich the diversity and coverage of the training corpus. After refinement through our automated processing pipeline, the final dataset encompasses approximately 2 million synchronized audio-visual clips totaling over 3,000 hours, alongside 50 million high-quality audio segments exceeding 130,000 hours in total duration.

Implementation Details. Unison training involves two stages. Stage 1 (Audio Branch) utilizes 4 NVIDIA H100 GPUs with a 96-sample batch size. We adopt a 1×10^{-4} learning rate with 1k-step linear warmup and step decay ($\gamma = 0.1$) at 240k and 270k steps. Stage 2 (Joint Training) fine-tunes the coupled system via the TI2AV objective (Eq. 9) on 16 H100 GPUs using bf16 precision and ZeRO-2 optimization. At a 2×10^{-5} learning rate and 32-sample batch size, we optimize only the audio branch and fusion modules (bi-directional cross-attention and layer normalization), keeping the video backbone frozen. A progressive training strategy with phase ratios of 0.3, 0.4, and 0.3 ensures multimodal alignment stability. Inference employs a 50-step flow-matching sampler with classifier-free guidance (scale=6.0), producing 25 FPS videos. The code and models will be made publicly available upon acceptance.

Evaluation Metrics. We report a comprehensive set of objective metrics. (1) For video assessment, we employ LAION-Aesthetic Predictor V2.5 [30] to compute the Video Aesthetic Score (VA) as a proxy for artistic coherence, alongside DINOv3 [33] to measure inter-frame Identity consistency (ID). (2) For audio assessment, following the Audiobox [38] protocol, we employ Audiobox-Aesthetics to determine Perceptual Quality (PQ) and Content Usefulness (CU). To evaluate speech-text alignment, we isolate vocal components via Mel-RoFormer [42] and compute the Word Error Rate (WER) using Whisper-large-v3 [28]. (3) For cross-modal consistency, we utilize CLAP [8] for audio-text semantic consistency (TA), VideoCLIP-XL-V2 [44] for video-text semantic alignment (TV), and ImageBind [10] for audio-video semantic similarity (AV). Lip-audio synchronization is measured via SyncNet [27] (LSE-C and LSE-D) while audio-visual temporal correspondence is captured by the DeSync (DS) score from Synchformer [15] to evaluate absolute time offsets between modal onsets.

4.2 Qualitative Results

Qualitative Comparisons. We evaluate Unison against representative baselines under identical settings. As shown in Fig. 4, Universe-1 [40] and UniAVGen [50] fail to synthesize musical accompaniments and struggle to distinguish singing from standard speech. In locomotive scenes, the aforementioned baselines exhibit severe modal imbalance, with speech volume disproportionately overwhelming engine sounds. MOVA [32] generates plausible vocals in musical contexts yet suffers from substantial phonetic artifacts in complex environments. Conversely, Unison achieves precise synchronization between motion dynamics and diverse acoustic components, including lip movements and impact transients. Our model maintains superior acoustic layering, ensuring intelligible speech without suppressing salient environmental audio.

Audio-to-Video and Video-to-Audio Generation. Unison leverages decoupled denoising schedules and bidirectional guidance to achieve precise modal translation across both A2V and V2A tasks. In A2V scenarios, audio conditioning aligns motion with salient acoustic features, such as speech and impact transients, while

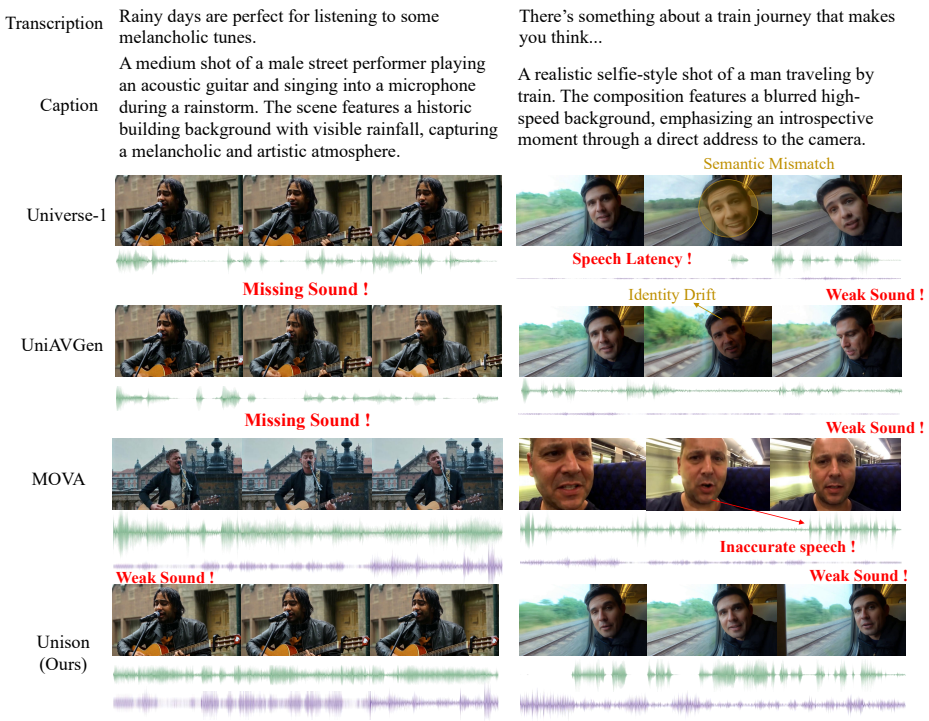

Fig. 4: Qualitative comparison between Unison and the state-of-the-art methods, including Universe-1 [40], UniAVGen [50] and MOVA [32].

for V2A, the model synthesizes coherent audio streams from visual cues. This capability arises from our progressive cross-modal forcing, which facilitates stable conditioning under heterogeneous denoising dynamics. As shown in Fig. 5, Unison delivers high-fidelity synthesis across all translation directions, maintaining robust semantic consistency regardless of the source modality.

Effectiveness of Semantic-Guided Audio Harmonization Strategy. We investigate the impact of decoupled audio generation, Bidirectional Audio Cross-Attention (Bi-ACA), and Semantic-Conditioned Gating (SCG) within the audio branch. As shown in the beach scene (Fig. 6), models lacking these components fail to synthesize harmonious speech and sound-effect representations, leading to speech-dominant audio where ambient waves are significantly attenuated. Conversely, our Semantic-guided Audio Harmonization Strategy enables the adaptive recomposition of structurally distinct speech and sound effects. This mechanism yields sharper acoustic transients and more balanced acoustic mixtures.

Effectiveness of Bidirectional Cross-modal Forcing Strategy. As shown in the piano sequence (Fig. 7), omitting cross-modal forcing leads to temporally drifted accompaniments where audio onsets no longer correspond to specific finger movements. With forcing enabled, the modality at a lower noise level provides an explicit denoising reference for its noisier counterpart, establishing a bidirectional

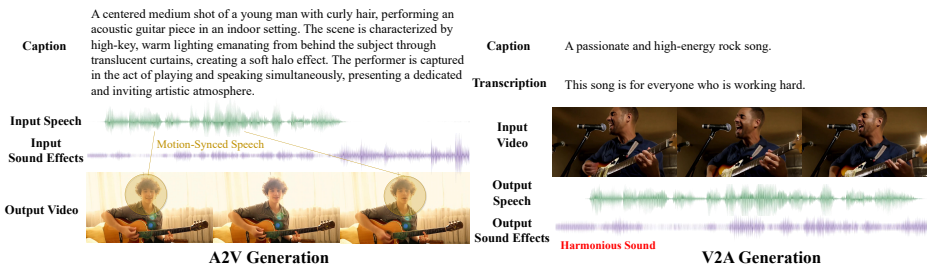

Fig. 5: Bidirectional Synthesis of Audio-to-Video and Video-to-Audio.

Table 1: Quantitative comparison with state-of-the-art methods. 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.

Type	Model	Video Quality		Audio Quality			Cross-Modal Consistency					
		VA ↑	ID ↑	PQ ↑	CU ↑	WER ↓	TA ↑	TV ↑	AV ↑	LSE-C ↑	LSE-D ↓	DS ↓
TI2AV	Universe-1 [40]	3.77	4.42	5.95	5.21	0.52	3.37	25.57	0.62	2.32	–	0.50
	Ovi [24]	3.94	4.42	6.25	5.51	0.43	3.48	25.86	0.87	2.81	9.12	0.12
	UniAVGen [50]	4.02	4.46	6.18	5.48	0.33	3.42	25.99	0.81	2.89	9.49	0.15
	MOVA [36]	4.01	4.52	6.28	5.52	0.29	3.58	25.97	0.88	3.24	7.92	0.13
	LTX-2 [12]	**4.15**	**4.61**	6.30	5.58	0.25	**3.65**	**26.24**	0.89	**3.45**	**7.62**	0.10
	Unison (Ours)	4.02	4.53	**6.34**	**5.61**	**0.22**	3.61	26.17	**0.91**	3.30	7.88	**0.08**
T2AV	JavisDiT [22]	3.29	4.52	4.83	3.73	1.81	3.53	24.31	0.49	1.81	–	0.53
	Ovi [24]	4.22	4.51	6.08	5.65	0.18	3.55	25.99	0.83	3.47	8.05	0.08
	LTX-2 [12]	**4.63**	**4.68**	6.12	5.72	0.11	**3.74**	**26.35**	0.81	**3.62**	**7.75**	0.07
	Unison (Ours)	4.51	4.59	**6.17**	**5.78**	**0.09**	3.62	26.21	**0.86**	3.55	7.95	**0.06**

guidance loop that corrects misalignment during generation. The resulting audio exhibits tightly synchronized note onsets and release transients that faithfully follow the observed hand dynamics.

4.3 Quantitative Results

Comparison with Baselines. We evaluate Unison on a curated test set of 1,000 held-out samples, with ground-truth annotations provided by Gemini to ensure rigorous T2AV and TI2AV assessment. As shown in Table 1, Unison achieves the highest audio fidelity across all methods, notably leading in PQ, CU, and WER. Despite utilizing a 5B video backbone—nearly 4× smaller than LTX-2's 19B—Unison exhibits superior cross-modal synchronization. While LTX-2 shows marginal gains in visual texture (VA), Unison remains highly competitive in overall perceptual quality. These results validate that our architectural innovations effectively capture high-fidelity audiovisual correlations without the need for massive parameter scaling.

Analysis of SCG Gating Behavior. To verify that SCG learns dynamic, content-aware modulation, we visualize the gate values g^{sp} and g^{sfx} across different dimensions (Fig. 8). *Layer-wise*: Shallow layers maintain balanced gates ($g \approx 0.5$)

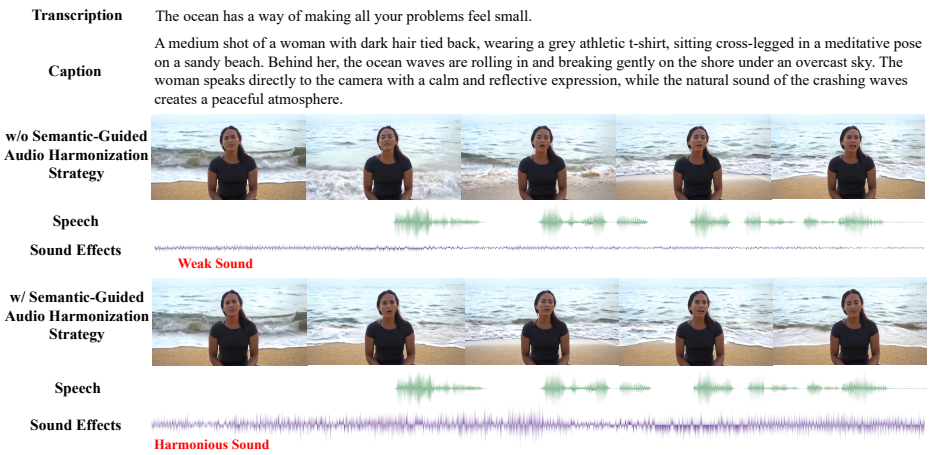

Fig. 6: Ablation experiments on the Semantic-Guided Audio Harmonization Strategy.

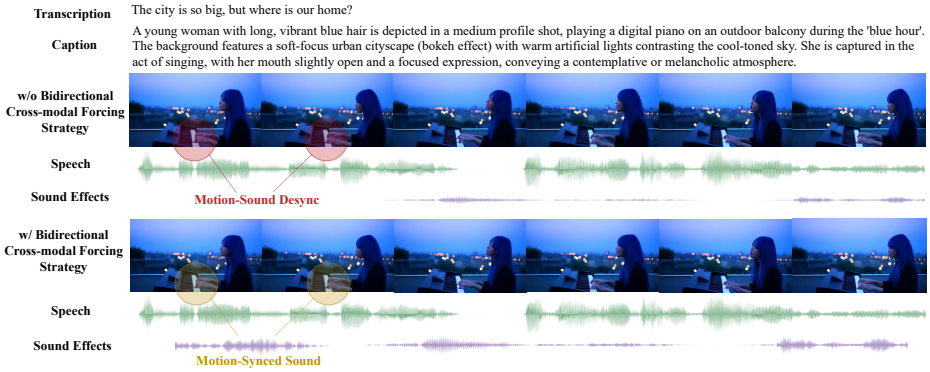

Fig. 7: Ablation experiments on the Bidirectional Cross-modal Forcing Strategy.

for coarse structure formation, while deeper layers exhibit increasing polarization for semantic refinement. *Timestep-wise*: Gate divergence intensifies as denoising progresses; g^{sp} and g^{sfx} stay moderate at high noise levels but diverge significantly as content crystallizes, ensuring interference suppression. *Instance-wise*: SCG mitigates the dominance of speech over subtle environmental textures via dynamic rebalancing. In sports broadcasting, the mechanism constrains the narration stream ($g^{sp} = 0.62$) to prevent it from masking critical acoustic cues. This safeguards high-frequency transients of the stadium atmosphere ($g^{sfx} = 0.38$), ensuring that impact sounds and crowd cheers remain perceptible despite high-volume vocal inputs.

Ablation Study on Components. Table 2 evaluates four core modules: SGHS (Semantic-Guided Harmonization Strategy), Bi-ACA (Bidirectional Audio Cross-Attention), SCG (Semantic-Conditioned Gating), and CMFS (Cross-Modal Forcing Strategy). We report video quality (VA), audio fidelity (PQ), and synchronization (LSE-C, DS). Among audio-side components, removing SGHS causes

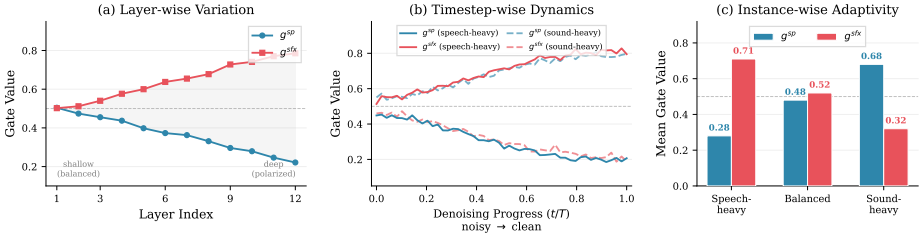

Fig. 8: Analysis of SCG gate behavior. (a) **Layer-wise**: gate polarization increases with model depth. (b) **Timestep-wise**: gate divergence intensifies as denoising progresses. (c) **Instance-wise**: mean gate values across semantic categories, demonstrating content-adaptive modulation.

Table 2: Ablation study on the core components of Unison.

Settings	VA ↑	PQ ↑	LSE-C ↑	DS ↓
w/o SGHS	3.99	6.12	3.08	0.15
w/o Bi-ACA	4.00	6.20	3.18	0.11
w/o SCG	4.01	6.21	3.22	0.10
w/o CMFS	3.91	6.24	3.02	0.19
Ours	**4.02**	**6.34**	**3.30**	**0.08**

Table 3: Ablation study on the training strategies of Unison.

Settings	VA ↑	PQ ↑	LSE-C ↑	DS ↓
SyncOnly	3.90	6.10	3.12	0.17
IndepOnly	3.95	6.18	3.28	0.14
PF(Ours)	**4.02**	**6.34**	**3.30**	**0.08**

the most significant PQ degradation (6.12 vs. 6.34) due to spectral interference, and eliminating Bi-ACA and SCG further reduces PQ and lip-sync precision by weakening bidirectional context and text-driven suppression. Notably, these audio-specific ablations leave VA largely unaffected (3.99–4.01), confirming their modularity. In contrast, removing CMFS reveals a distinct degradation pattern: DS rises sharply to 0.19 and LSE-C drops to 3.02, the poorest alignment scores across all settings. Furthermore, VA decreases to 3.91, indicating the video branch benefits from audio-stream guidance to reinforce visual coherence. These results confirm a functional complementarity: audio-side modules ensure acoustic fidelity, and CMFS establishes robust audio-visual synchronization.

Ablation of Training Strategy. Table 3 evaluates three training schedules for cross-modal forcing. The SyncOnly baseline, enforcing $t_v{=}t_a$ throughout, yields the lowest performance (VA 3.90, DS 0.17), confirming that uniform noise prevents the model from exploiting denoising asymmetries. IndepOnly, which applies fully decoupled timesteps without a warmup, improves alignment (DS 0.14) but destabilizes optimization, capping LSE-C at 3.28 due to abrupt noise gaps. In contrast, our ProgForcing (PF) schedule implements a three-phase curriculum—synchronous warmup, incremental decoupling, and full independence—to progressively introduce challenging noise configurations. This strategy achieves superior results across all metrics (VA 4.02, PQ 6.34, LSE-C 3.30, DS 0.08), demonstrating that a gradual transition is essential for both stable optimization and precise temporal alignment.

Table 4: Results of the user study. 40 samples × 25 participants (1,000 mean-rank votes; lower is better).

Method	Lip-sync↓	Speech-Sound Harmony↓	Motion-Audio Align.↓	Gemini Score↓
UniAVGen	3.51	3.45	3.42	3.48
MOVA	2.58	2.77	2.89	2.48
LTX-2	**1.74**	<u>1.95</u>	**1.89**	<u>2.05</u>
Unison	<u>1.86</u>	**1.55**	<u>1.92</u>	**1.68**

User Study. We conducted a user study with 40 samples × 25 participants from diverse backgrounds (1,000 mean-rank votes; lower is better), evaluating lip-speech synchrony, speech-sound harmony, and motion-audio alignment (considering both speech and environmental sounds). We further incorporate an additional evaluation column by scoring Motion-Speech-SFX coherence via Gemini. Participants were required to rank shuffled videos across different methods, including UniAVGen [50], MOVA [32], and LTX-2 [12]. As shown in Table 4, our approach achieved the highest scores in both speech-sound harmony and motion-audio alignment, as well as the best overall Gemini coherence score. While our lip-speech synchrony performance was second only to LTX-2, our method received the dominant overall preference across the integrated metrics, demonstrating superior holistic audio-visual quality.

5 Conclusion.

In this work, we present Unison, a framework that achieves superior motion-speech-sound synchronization. By implementing decoupled acoustic modeling alongside a progressive cross-modal forcing strategy, Unison effectively resolves the temporal misalignments and modal interference typical of joint synthesis. Extensive evaluations demonstrate that Unison achieves state-of-the-art performance in cross-modal consistency and audio perceptual fidelity.

Acknowledgements

This work was supported by the NSFC Regional Innovation and Development Joint Fund under Grant U25A20537, and the National Key Research and Development Program of China under Grant 2024YFC3015600.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, A.P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. In: arXiv:1609.08675 (2016), https://arxiv.org/pdf/1609.08675v1.pdf
2. An, H., Hu, W., Huang, S., Huang, S., Li, R., Liang, Y., Shao, J., Song, Y., Wang, Z., Yuan, C., Zhang, C., Zhang, H., Zhuang, W., Li, X.: Ai flow: Perspectives, scenarios, and approaches (2025), https://arxiv.org/abs/2506.12479
3. Chen, B., Monso, D.M., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion (2024), https://arxiv.org/abs/2407.01392
4. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2020)
5. Chen, K., Zhang, J., Li, M., Zheng, Z., Fan, H.: Clusterstyle: Modeling intra-style diversity with prototypical clustering for stylized motion generation. arXiv preprint arXiv:2512.02453 (2025)
6. Cheng, H.K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., Mitsufuji, Y.: MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In: CVPR (2025)
7. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Workshop on Multi-view Lip-reading, ACCV (2016)
8. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: Clap learning audio concepts from natural language supervision. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
9. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: Proc. IEEE ICASSP 2017. New Orleans, LA (2017)
10. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180–15190 (2023)
11. Google DeepMind: Veo: A text-to-video generation system (2025), https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf
12. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., Richardson, E., Shiran, G., Chachy, I., Chetboun, J., Finkelson, M., Kupchick, M., Zabari, N., Guetta, N., Kotler, N., Bibi, O., Gordon, O., Panet, P., Benita, R., Armon, S., Kulikov, V., Inger, Y., Shiftan, Y., Melumian, Z., Farbman, Z.: Ltx-2: Efficient joint audio-visual foundation model (2026), https://arxiv.org/abs/2601.03233
13. Hu, T., Yu, Z., Zhang, G., Su, Z., Zhou, Z., Zhang, Y., Zhou, Y., Lu, Q., Yi, R.: Harmony: Harmonizing audio and video generation through cross-task synergy. arXiv preprint arXiv:2511.21579 (2025)
14. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion (2025), https://arxiv.org/abs/2506.08009
15. Iashin, V., Xie, W., Rahtu, E., Zisserman, A.: Synchformer: Efficient synchronization from sparse cues. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5325–5329. IEEE (2024)

16. Jiang, W., Zhang, Y., Zheng, S., Liu, S., Yan, S.: Data augmentation in human-centric vision. Vicinagearth **1**(1), 8 (2024)
17. Li, H., Xu, M., Zhan, Y., Mu, S., Li, J., Cheng, K., Chen, Y., Chen, T., Ye, M., Wang, J., Zhu, S.: Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation (2025), https://arxiv.org/abs/2412.00115
18. Li, X., Wang, S., Zeng, S., et al.: A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth **1**(9) (2024). https://doi.org/10.1007/s44336-024-00009-2
19. Li, X.: Positive-incentive noise. IEEE Transactions on Neural Networks and Learning Systems **35**(6), 8708–8714 (2024). https://doi.org/10.1109/TNNLS.2022.3224577
20. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling (2023), https://arxiv.org/abs/2210.02747
21. Liu, H., Lan, G.L., Mei, X., Ni, Z., Kumar, A., Nagaraja, V., Wang, W., Plumbley, M.D., Shi, Y., Chandra, V.: Syncflow: Toward temporally aligned joint audio-video generation from text (2024), https://arxiv.org/abs/2412.15220
22. Liu, K., Li, W., Chen, L., Wu, S., Zheng, Y., Ji, J., Zhou, F., Jiang, R., Luo, J., Fei, H., et al.: Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. arXiv preprint arXiv:2503.23377 (2025)
23. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time (2025), https://arxiv.org/abs/2509.25161
24. Low, C., Wang, W., Katyal, C.: Ovi: Twin backbone cross-modal fusion for audio-video generation. arXiv preprint arXiv:2510.01284 (2025)
25. Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M.D., Zou, Y., Wang, W.: WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. IEEE/ACM Transactions on Audio, Speech, and Language Processing pp. 1–15 (2024)
26. OpenAI: Sora 2 system card (2025), https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf
27. Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. p. 484–492. MM '20, ACM (Oct 2020). https://doi.org/10.1145/3394171.3413532, http://dx.doi.org/10.1145/3394171.3413532
28. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518. PMLR (2023)
29. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q., Guo, B.: Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: CVPR (2023)
30. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
31. Shen, Y., Zhang, D.: A survey of language-guided video object segmentation: from referring to reasoning. Vicinagearth **2**(9) (2025). https://doi.org/10.1007/s44336-025-00018-9
32. SII-OpenMOSS Team, Yu, D., Chen, M., Chen, Q., Luo, Q., Wu, Q., Cheng, Q., Li, R., Liang, T., Zhang, W., Tu, W., Peng, X., Gao, Y., Huo, Y., Zhu, Y., Luo, Y., Zhang, Y., Song, Y., Xu, Z., Zhang, Z., Yang, C., Chang, C., Zhou, C.,

Chen, H., Ma, H., Li, J., Tong, J., Liu, J., Chen, K., Li, S., Wang, S., Jiang, W., Fei, Z., Ning, Z., Li, C., Li, C., He, Z., Huang, Z., Chen, X., Qiu, X.: Mova: Towards scalable and synchronized video-audio generation (Feb 2026). https://doi.org/10.48550/arXiv.2602.08794, https://arxiv.org/abs/2602.08794, technical report. Corresponding authors: Xie Chen and Xipeng Qiu. Project leaders: Qinyuan Cheng and Tianyi Liang.
33. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
34. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion (2025), https://arxiv.org/abs/2502.06764
35. Song, Q., He, Y., Zhang, Y., Cheng, S., He, Z., Guo, Z., Zhang, C., Li, X., Jiang, C.: Interactiveavatar: Real-time streaming video generation for consistent and intent-aware avatars (2026), https://arxiv.org/abs/2606.22905
36. Team, O., Yu, D., Chen, M., Chen, Q., Luo, Q., Wu, Q., Cheng, Q., Li, R., Liang, T., Zhang, W., Tu, W., Peng, X., Gao, Y., Huo, Y., Zhu, Y., Luo, Y., Zhang, Y., Song, Y., Xu, Z., Zhang, Z., Yang, C., Chang, C., Zhou, C., Chen, H., Ma, H., Li, J., Tong, J., Liu, J., Chen, K., Li, S., Jiang, S., Wang, S., Jiang, W., Fei, Z., Ning, Z., Li, C., Li, C., He, Z., Huang, Z., Chen, X., Qiu, X.: Mova: Towards scalable and synchronized video-audio generation (2026), https://arxiv.org/abs/2602.08794
37. Tian, Z., Liu, Z., Yuan, R., Pan, J., Liu, Q., Tan, X., Chen, Q., Xue, W., Guo, Y.: Vidmuse: A simple video-to-music generation framework with long-short-term modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18782–18793 (2025)
38. Vyas, A., Shi, B., Le, M., Tjandra, A., Wu, Y.C., Guo, B., Zhang, J., Zhang, X., Adkins, R., Ngan, W., et al.: Audiobox: Unified audio generation with natural language prompts. arXiv preprint arXiv:2312.15821 (2023)
39. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
40. Wang, D., Zuo, W., Li, A., Chen, L.H., Liao, X., Zhou, D., Yin, Z., Dai, X., Jiang, D., Yu, G.: Universe-1: Unified audio-video generation via stitching of experts. arXiv preprint arXiv:2509.06155 (2025)
41. Wang, J., Deng, H., Pan, T., Liu, Y., Wang, C., Zhang, F., Qi, Y., Wang, X.: Udm-grpo: Stable and efficient group relative policy optimization for uniform discrete diffusion models (2026), https://arxiv.org/abs/2604.18518
42. Wang, J.C., Lu, W.T., Chen, J.: Mel-roformer for vocal separation and vocal melody transcription (2024), https://arxiv.org/abs/2409.04702
43. Wang, L.X.X., Zhang, H., Dong, C., Shan, Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution (2022), https://arxiv.org/abs/2205.03409
44. Wang, W., Yang, Y.: Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. Advances in Neural Information Processing Systems **37**, 65618–65642 (2024)
45. Xiao, A., Cheng, S., Xu, Y., Ren, Y., Chen, H., Yokoya, N.: Geommbench and geommagent: Toward expert-level multimodal intelligence in geoscience and remote sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 34843–34853 (Jun 2026)
46. Yao, Z., Guo, L., Yang, X., Kang, W., Kuang, F., Yang, Y., Jin, Z., Lin, L., Povey, D.: Zipformer: A faster and better encoder for automatic speech recognition (2024), https://arxiv.org/abs/2310.11230

47. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: CelebV-Text: A large-scale facial text-video dataset. In: CVPR (2023)
48. Yuan, R., Lin, H., Guo, S., Zhang, G., Pan, J., Zang, Y., Liu, H., Du, X., Du, X., Ye, Z., Zheng, T., Jiang, Z., Ma, Y., Liu, M., Yu, L., Tian, Z., Zhou, Z., Xue, L., Qu, X., Li, Y., Shen, T., Ma, Z., Wu, S., Zhan, J., Wang, C., Wang, Y., Zhou, X., Chi, X., Zhang, X., Yang, Z., Liang, Y., Wang, X., Liu, S., Mei, L., Li, P., Chen, Y., Lin, C., Chen, X., Xia, G., Zhang, Z., Zhang, C., Chen, W., Zhou, X., Qiu, X., Dannenberg, R., Liu, J., Yang, J., Huang, S., Xue, W., Tan, X., Guo, Y.: Yue: Open music foundation models for full-song generation. https://github.com/multimodal-art-projection/YuE (2025), gitHub repository
49. Yuan, R., Lin, H., Guo, S., Zhang, G., Pan, J., Zang, Y., Liu, H., Liang, Y., Ma, W., Du, X., Du, X., Ye, Z., Zheng, T., Jiang, Z., Ma, Y., Liu, M., Tian, Z., Zhou, Z., Xue, L., Qu, X., Li, Y., Wu, S., Shen, T., Ma, Z., Zhan, J., Wang, C., Wang, Y., Chi, X., Zhang, X., Yang, Z., Wang, X., Liu, S., Mei, L., Li, P., Wang, J., Yu, J., Pang, G., Li, X., Wang, Z., Zhou, X., Yu, L., Benetos, E., Chen, Y., Lin, C., Chen, X., Xia, G., Zhang, Z., Zhang, C., Chen, W., Zhou, X., Qiu, X., Dannenberg, R., Liu, J., Yang, J., Huang, W., Xue, W., Tan, X., Guo, Y.: Yue: Scaling open foundation models for long-form music generation (2025), https://arxiv.org/abs/2503.08638
50. Zhang, G., Zhou, Z., Hu, T., Peng, Z., Zhang, Y., Chen, Y., Zhou, Y., Lu, Q., Wang, L.: Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. arXiv preprint arXiv:2511.03334 (2025)
51. Zhang, H., Huang, S., Guo, Y., Li, X.: Variational positive-incentive noise: How noise benefits models. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(9), 8313–8320 (2025). https://doi.org/10.1109/TPAMI.2025.3575295
52. Zhang, X., Cai, Y., Li, K., Yang, K., Zhou, Y., Li, Z., Chu, X., Zhang, J., Liu, H.: Personagesture: Single-reference co-speech gesture personalization for unseen speakers. arXiv preprint arXiv:2605.06064 (2026)
53. Zhang, X., Li, J., Ren, J., Zhang, J.: Mitigating error accumulation in co-speech motion generation via global rotation diffusion and multi-level constraints. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12834–12842 (2026)
54. Zhang, X., Li, J., Zhang, J., Dang, Z., Ren, J., Bo, L., Tu, Z.: Semtalk: Holistic co-speech motion generation with frame-level semantic emphasis. arXiv preprint arXiv:2412.16563 (2024)
55. Zhang, X., Li, J., Zhang, J., Ren, J., Bo, L., Tu, Z.: Echomask: Speech-queried attention-based mask modeling for holistic co-speech motion generation (2025), https://arxiv.org/abs/2504.09209
56. Zhang, Z., Foo, L.G., Rahmani, H., Liu, J., Soh, D.W.: Performing defocus deblurring by modeling its formation process. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5791–5801 (2025)
57. Zhang, Z., Huang, M.H., Tu, Z., Yang, M.H.: Dart: Difficulty-adaptive routing for zero-shot video temporal grounding. In: European Conference on Computer Vision (2026)
58. Zhang, Z., Tu, Z., Yuan, J., Soh, D.W., Du, B.: Leveraging text-to-image diffusion models for unsupervised visual object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
59. Zhang, Z., Zhou, Y., Peng, D., Lim, J.H., Tu, Z., Soh, D.W., Foo, L.G.: Visual prompting for one-shot controllable video editing without inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7784–7794 (2025)

60. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3661–3670 (2021)
61. Zhao, L., Feng, L., Ge, D., Chen, R., Yi, F., Zhang, C., Zhang, X.L., Li, X.: Uniform: A unified multi-task diffusion transformer for audio-video generation (2025), https://arxiv.org/abs/2502.03897
62. Zhu, H., Kang, W., Yao, Z., Guo, L., Kuang, F., Li, Z., Zhuang, W., Lin, L., Povey, D.: Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. arXiv preprint arXiv:2506.13053 (2025)