FROM SEEING TO HEARING: UNIFIED AUDIO-VIDEO GENERATION WITH PROGRESSIVE ADAPTATION AND MODALITY FORCING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating temporally synchronized and semantically coherent audio and video within a unified model remains challenging, particularly when supporting both joint and directional generation. In this work, we propose a multi-stage adaptation framework that extends a pretrained video generator (Wan2.1) into a unified audio–video model. Our framework employs modality-specific branches with bidirectional cross-attention and progressively acquires multimodal generation capabilities through three stages. First, audio-only adaptation equips the pretrained backbone with audio generation capability. Second, joint audio–video fine-tuning establishes cross-modal interaction for synchronized audiovisual generation. Finally, we introduce *Modality Forcing*, a task-dependent adaptation strategy that combines clean modality conditioning, directional information routing, and asymmetric optimization to enable video-to-audio and audio-to-video generation while retaining joint-generation capability. Extensive experiments demonstrate competitive joint audio–video generation, strong bidirectional conditional generation, and preservation of the pretrained video generation capability, highlighting the effectiveness of progressive adaptation and Modality Forcing for unified audiovisual generation.

1 INTRODUCTION

Joint audio–video generation aims to synthesize temporally synchronized visual and auditory signals for applications such as automated filmmaking, Foley sound design, and immersive multimedia. Despite advances in transformer-based diffusion models, generating high-quality audiovisual content with accurate cross-modal alignment remains challenging. Existing approaches either train multimodal models from scratch, such as MM-Diffusion Ruan et al. (2023) and MoNL Kim et al. (2024), or leverage pretrained components through feature fusion Haji-Ali et al. (2025); Wang et al. (2025a) and unified multimodal backbones Zhao et al. (2025); Low et al. (2025). An alternative is to build upon pretrained video generators, which have acquired rich visual and temporal priors from large-scale video data.

However, extending a video generator with audio capability is only part of the challenge. A unified audio–video model should support both joint generation from text and directional generation, where one modality conditions the other. Joint generation requires coordinated modeling of both modalities, whereas directional generation requires a fixed conditioning modality and a generated target. This raises a central question: *How can pretrained video generators be adapted to support both joint and directional audio–video generation while preserving their existing generative capabilities?*

We address this challenge through a progressive adaptation framework built upon Wan2.1, using modality-specific branches and bidirectional cross-attention for audiovisual interaction. The framework consists of three stages. Audio-only training equips a copy of the pretrained video backbone with audio generation capability using audio–text pairs. Joint audio–video fine-tuning then learns semantic correspondence and temporal synchronization from audio–video–text triplets. Finally, we introduce *Modality Forcing* to enable directional generation while retaining joint-generation capability. It explicitly separates conditioning and target modalities through task-dependent information routing and asymmetric optimization. For video-to-audio generation, video latents remain clean and the video

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

branch is frozen, while the reverse cross-modal pathway is disabled and only the audio branch is optimized. Audio-to-video generation follows the symmetric configuration. Joint-generation samples are retained during this stage to preserve existing capabilities. Together, these training modes enable a single model to support joint and bidirectional conditional generation.

Experiments demonstrate competitive joint audio–video generation, strong V2A and A2V performance, and preservation of the pretrained video generation capability. Applying Modality Forcing to JavisDiT further improves joint and conditional generation, demonstrating its effectiveness beyond our Wan2.1-based framework. Our contributions are summarized as follows:

- We introduce *Modality Forcing*, a task-dependent adaptation strategy that enables bidirectional conditional generation through asymmetric optimization and directional information routing while retaining joint-generation capability.
- We develop a progressive framework that adapts a pretrained video generator into a unified audio–video model through audio capability acquisition, joint audiovisual modeling, and directional specialization.
- We demonstrate competitive joint and conditional generation, validate the generalization of Modality Forcing to another audiovisual architecture, and analyze capability preservation and computational efficiency.

2 Related Works

2.1 Video Generation

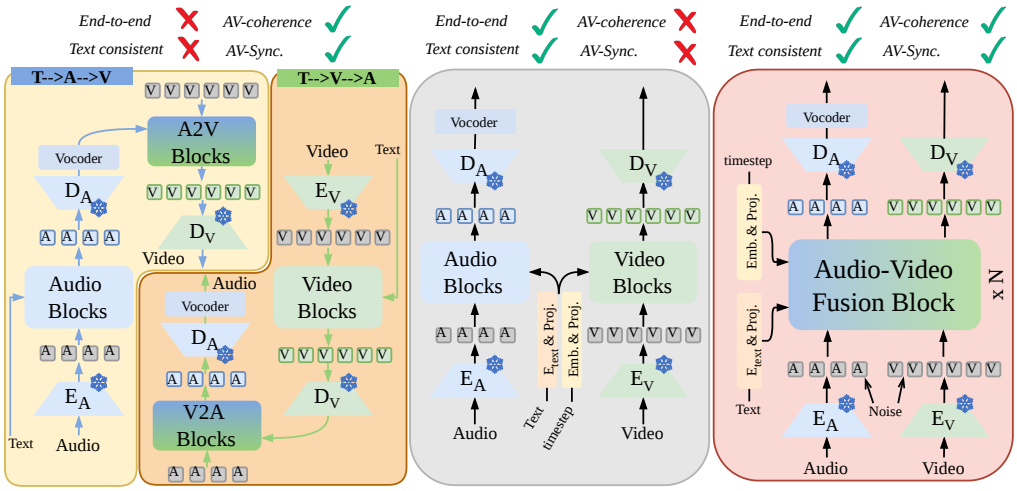
Video generation has advanced from diffusion models Ho et al. (2020) with UNet-based architectures Ronneberger et al. (2015) and pretrained image backbones Rombach et al. (2022) to Diffusion Transformers (DiTs) Peebles & Xie (2023) with spatio-temporal latent representations. Large-scale models such as Sora OpenAI (2025), Wan Wan et al. (2025), CogVideoX Yang et al. (2024), and HunyuanVideo Kong et al. (2024) demonstrate high-quality, temporally coherent video generation. These models provide rich visual and temporal priors for multimodal generation. We build upon Wan2.1 and progressively adapt it for joint and directional audio–video generation.

2.2 Video-Audio Cross-modal Generation

Cross-modal audio–video generation includes audio-to-video (A2V) and video-to-audio (V2A) synthesis. A2V methods commonly adapt pretrained video diffusion transformers with audio conditioning. HunyuanVideo Kong et al. (2024) and its variants Chen et al. (2025b); Yi et al. (2025) leverage representations from speech models such as Whisper Radford et al. (2023) and Wav2Vec Schneider et al. (2019) for synchronized video generation. Other approaches Gao et al. (2025); Chen et al. (2025a) introduce lightweight audio-conditioning modules while freezing most backbone parameters. V2A methods either train dedicated models on paired audiovisual data, such as Diff-Foley Luo et al. (2023), Frieren Wang et al. (2024b), MMAudio Cheng et al. (2025), and VATT Liu et al. (2024b), or adapt pretrained audio generators through modality mapping and control, including Seeing-and-Hearing Xing et al. (2024), V2A-Mapper Wang et al. (2024a), and FoleyCrafter Zhang et al. (2026). Recent methods such as AudioX Tian et al. (2025) and HunyuanVideo-Foley Shan et al. (2025) further improve conditional audio generation. Unlike these task-specific approaches, our framework supports joint generation and both conditional directions within a single model.

2.3 Joint Audio-Video Generation

Joint audio–video generation (JAVG) aims to synthesize synchronized audio and video within a unified model. Closed-source systems such as Veo3 Google DeepMind (2025) and Sora2 OpenAI (2025) demonstrate large-scale audiovisual generation capabilities. Existing open-source approaches employ either modality-specific branches with cross-modal fusion, including UniVerse-1 Wang et al. (2025b), JOVA Huang et al. (2025), MOVA Team et al. (2026), LTX-2 HaCohen et al. (2026), and OVI Low et al. (2025), or shared backbones, such as UniForm Zhao et al. (2025) and AV-DiT Wang et al. (2025c). Closely related to our work, JavisDiT Liu et al. (2025) introduces hierarchical spatio-temporal prior synchronization, while JavisDiT++ Liu et al. (2026) extends a pretrained



(1) Cascaded Framework (2) Dual-Branch w/o Fusion **(3) Ours**

Figure 1: Overview of existing JAVG pipelines (left and center) and our proposed method (right). (1) Cascaded framework. The model first generates one modality (audio or video) from the text prompt and then conditions on it to generate the other modality. (2) Dual-branch w/o fusion. Audio and video are generated in parallel using independent branches without cross-modal interaction. (3) Our architecture. Audio and video are first encoded into modality-specific latent representations and then fused through bidirectional cross-attention. Compared with (1) and (2), our method enables end-to-end joint generation with improved audio–video coherence and synchronization.

video backbone through unified token modeling, temporal alignment, and preference optimization. MoNL Kim et al. (2024) and JavisDiT also support conditional generation through variable noise levels and dynamic temporal masking, respectively. In contrast, our framework introduces *Modality Forcing* for directional specialization through task-dependent information routing and asymmetric optimization while retaining joint-generation capability.

3 Proposed Method

3.1 Preliminaries

Rectified Flow Matching. We adopt Rectified Flow Matching (RFM) Liu et al. (2022) for generative modeling. Given a clean data sample $x_1 \sim p_1(x)$ and a noise sample $x_0 \sim p_0(x) = \mathcal{N}(0, I)$, RFM defines a linear interpolation between noise and data at time $t \in [0, 1]$:$x_t = tx_1 + (1-t)x_0$. The corresponding target velocity is $u_t = \frac{dx_t}{dt} = x_1 - x_0$. A neural network v_θ is trained to predict this velocity conditioned on the intermediate state, time, and conditioning information C:

$$\mathcal{L}_{\text{RFM}} = \mathbb{E}_{x_0 \sim p_0,\, x_1 \sim p_1,\, t \sim \mathcal{U}(0,1)} \left[\| v_\theta(t, x_t, C) - (x_1 - x_0) \|_2^2 \right]. \tag{1}$$

During inference, generation starts from $\tilde{x}_0 \sim \mathcal{N}(0, I)$ and integrates the learned velocity field $\frac{d\tilde{x}_t}{dt} = v_\theta(t, \tilde{x}_t, C)$ from $t = 0$ to $t = 1$ using an ODE solver, yielding the generated sample $\tilde{x}_1$.

Joint Audio-Video Generation. We extend rectified flow matching to joint audio–video generation in modality-specific latent spaces. Let (x_1^a, x_1^v) denote paired audio and video latents, and let (x_0^a, x_0^v) be independently sampled Gaussian noise. For each modality $m \in \{a, v\}$, the intermediate state and target velocity are defined as $x_t^m = tx_1^m + (1-t)x_0^m$ and $u_t^m = x_1^m - x_0^m$, respectively.

We learn a joint audio–video network parameterized by θ_{av}, consisting of modality-specific branches coupled through cross-modal attention. The two branches predict separate velocity fields while incorporating information from both modalities and the shared text condition C. The joint flow-matching objective is:

$$\mathcal{L}_{\text{joint}} = \mathbb{E}\Big[\beta_v \left\| v^v_{\theta_{av}}(t, x_t^v, x_t^a, C) - u_t^v \right\|_2^2 + \quad \beta_a \left\| v^a_{\theta_{av}}(t, x_t^a, x_t^v, C) - u_t^a \right\|_2^2 \Big], \tag{2}$$

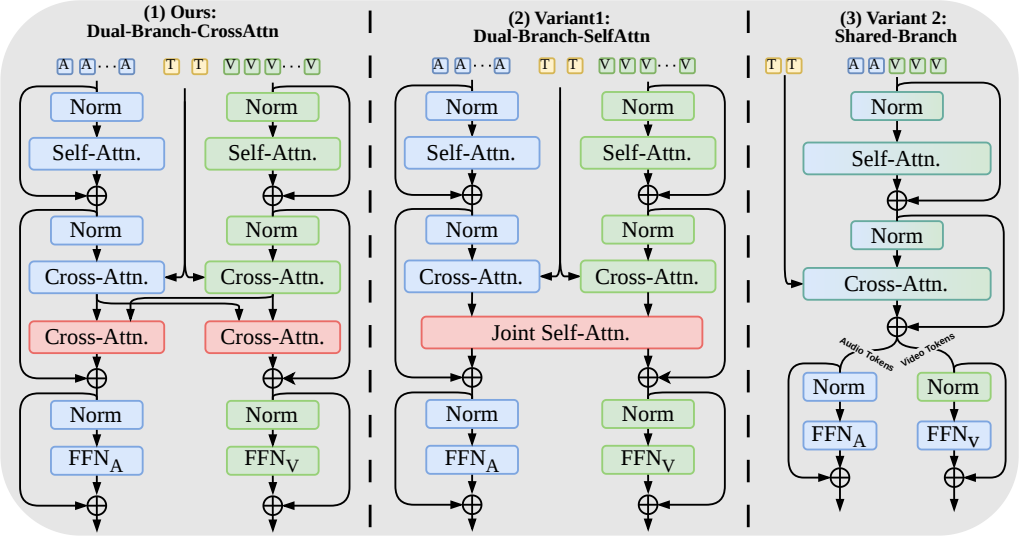


Figure 2: Proposed audio–video fusion block (left) with two additional variants (center and right). (1) `Dual-Branch-CrossAttn`: two parallel branches with bidirectional cross-attention for cross-modal interaction. (2) Variant 1: replaces cross-attention with joint self-attention. (3) Variant 2: uses a shared branch to jointly model audio and video modalities.

where the expectation is taken over paired audio–video samples, independent Gaussian noise for each modality, and $t \sim \mathcal{U}(0, 1)$. The weights β_v and β_a balance the video and audio flow-matching losses, respectively. During inference, both modalities are generated by jointly integrating their predicted velocity fields from Gaussian noise.

3.2 Proposed Architecture

Model Overview. We extend the pretrained Wan2.1 video generator into a unified audio–video model with modality-specific branches and cross-modal interaction. Unlike cascaded frameworks (Figure 1(1)) and independent dual-branch pipelines (Figure 1(2)), our model jointly generates both modalities through latent flow matching. As shown in Figure 1(3), frozen encoders E_A and E_V map audio and video into their respective latent spaces, which are processed by N Audio-Video Fusion Blocks conditioned on timestep and text embeddings. During inference, the predicted velocity fields are jointly integrated, and frozen decoders D_A and D_V reconstruct the audiovisual outputs.

Audio, Video and Text Encoding. We use the frozen video VAE from Wan2.1 Wan et al. (2025), which temporally downsamples video by a factor of 4 (except the first frame) and spatially by 8×8, producing 16-channel latents. A 2×2 spatial patchification further reduces the token length. For audio, we adopt the pretrained 1D audio VAE from MMAudio Cheng et al. (2025) to encode mel-spectrograms obtained through Short-Time Fourier Transform (STFT). Generated audio latents are decoded into spectrograms and converted into waveforms using BigVGAN Lee et al. (2022). We use the 16 kHz audio encoder variant and the frozen umT5-XXL text encoder from Wan2.1 Wan et al. (2025), which supports up to 512 tokens.

Aligned Audio-Video Positional Embedding. Audio latents typically contain more temporal frames (f_a) than video latents (f_v) Low et al. (2025). We apply Rotary Positional Embeddings (RoPE) to both branches and scale the audio RoPE frequency by f_v/f_a to align their temporal resolutions.

Audio-Video Fusion Block. Our `Dual-Branch-CrossAttn` block (Figure 2(1)) processes audio and video through two Wan2.1-initialized branches, each containing self-attention, text-conditioned cross-attention, and a modality-specific Feed-Forward Network (FFN). Bidirectional cross-attention enables explicit information exchange, with features from one modality serving as keys and values for the other. We also consider two variants: `Dual-Branch-SelfAttn`, which replaces bidirectional cross-attention with joint self-attention (Figure 2(2)), and `Shared-Branch`, which models both

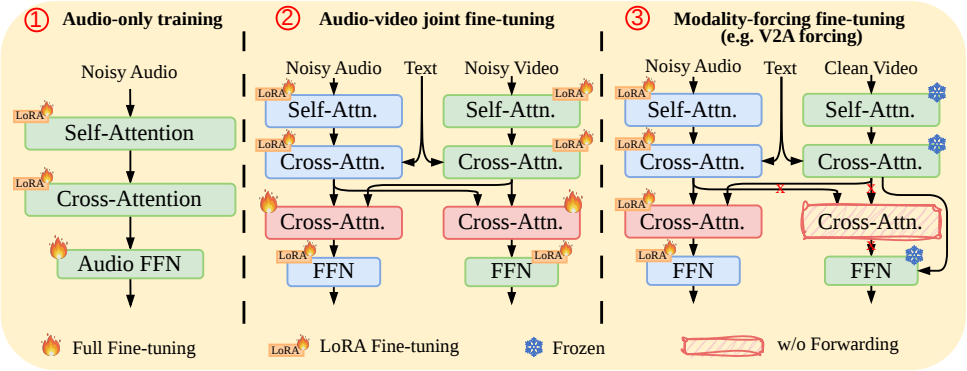


Figure 3: Proposed multi-stage training pipeline. (1) **Audio-only training.** The audio branch initialized from Wan2.1 is trained using audio–text data. (2) **Audio–video joint fine-tuning.** Both audio and video branches, including the proposed cross-attention modules (red), are jointly optimized. (3) **Modality-forcing fine-tuning.** One modality remains clean to condition the generation of the other modality. Video-to-audio (V2A) forcing is shown as an example, where only the audio branch is updated while the video branch is frozen and the A2V cross-attention is removed.

modalities within a shared backbone (Figure 2(3)). Our design preserves modality-specific modeling while providing directional cross-modal pathways for the subsequent Modality Forcing stage.

3.3 Multi-stage Training

As illustrated in Figure 3, we progressively adapt Wan2.1 through three stages: audio-only training, joint audio–video fine-tuning, and Modality Forcing, enabling audio generation, joint generation, and bidirectional conditional generation, respectively.

Audio-only Training. We equip a copy of the pretrained Wan2.1 backbone with audio generation capability using audio–text pairs. Audio-specific modules, including the embedding layer and FFN, are fully fine-tuned, while Low-Rank Adaptation (LoRA) is applied to the pretrained self-attention and cross-attention layers. This hybrid strategy reuses pretrained weights while adapting to audio representations, offering a favorable quality–cost trade-off compared with full or FFN-only fine-tuning Low et al. (2025); Wang et al. (2025b).

Audio-Video Joint Fine-tuning. Using audio–video–text triplets, we jointly optimize the original Wan2.1 video branch and the audio branch adapted in Stage 1 with the joint flow-matching objective. Newly introduced bidirectional cross-attention modules are fully trained to establish cross-modal interaction, while the remaining adapted layers are fine-tuned with LoRA.

Modality Forcing. To extend joint generation to directional generation, we introduce *Modality Forcing*, which separates conditioning and target modalities through task-dependent information routing and asymmetric optimization. As shown in Figure 3(3), training alternates among text-to-audio-video (T2AV), video-to-audio (V2A), and audio-to-video (A2V) modes. In T2AV mode, both modalities follow their flow-matching paths, and both branches are optimized. In V2A mode, clean video latents condition audio generation; the video branch is frozen, the reverse audio-to-video cross-attention pathway is disabled, and only the audio branch is optimized to predict the target velocity field. A2V mode follows the symmetric configuration, using clean audio to condition the video branch while disabling the reverse video-to-audio pathway. T2AV samples are retained throughout this stage to preserve joint-generation capability. All three modes share the rectified flow-matching formulation, differing only in conditioning, information flow, and trainable parameters, enabling a single model to support T2AV, V2A, and A2V generation.

4 Experiments

4.1 Datasets

For audio-only training, we use 780K audio–text pairs collected by JavisDiT Liu et al. (2025) from AudioSet Gemmeke et al. (2017), AudioCaps Kim et al. (2019), VGGSound Chen et al. (2020),

WavCaps Mei et al. (2024), Clotho Drossos et al. (2020), ESC50 Piczak (2015), GTZAN Sturm (2013), MACS Martín-Morató & Mesaros (2021), and UrbanSound8K Salamon et al. (2014). For joint audio–video fine-tuning, we use 611K audio–video–text samples collected by JavisDiT, comprising 136K from MMTrail Chi et al. (2024) and 475K from TAVGBench Mao et al. (2024) after cleaning and filtering. For Modality Forcing, we follow the filtering strategy of Liu et al. (2026; 2025) to select TAVGBench triplets. We remove videos containing human speech using FunASR Gao et al. (2023) and filter low-quality samples using aesthetic Schuhmann (2022), motion Xu et al. (2023), and OCR Liao et al. (2020) scores, following OpenSora Zheng et al. (2024). The resulting dataset contains 355K samples.

4.2 Evaluation Metrics

We evaluate joint audio–video generation on JavisBench Liu et al. (2025), comprising 10,140 text prompts. Following its protocol, we assess **generation quality** using Fréchet Video Distance (FVD) Carreira & Zisserman (2017) and Fréchet Audio Distance (FAD) Guzhov et al. (2022); **semantic consistency** using ImageBind Girdhar et al. (2023), CLIP Radford et al. (2021), CLAP Wu et al. (2023), and AVHScore Mao et al. (2024); and **audiovisual synchronization** using JavisScore Liu et al. (2025) and DeSync Iashin et al. (2024). ImageBind measures text–video, text–audio, and audio–video similarity, while CLIP and CLAP assess text–video and text–audio alignment, respectively. Lower FVD, FAD, and DeSync scores indicate better performance. All joint-generation models produce 4-second samples at 240p and 16 kHz. We additionally evaluate V2A and A2V on the VGGSound test set. V2A metrics include FAD, KL divergence, Inception Score (IS), ImageBind similarity (IB-score), and DeSync; A2V metrics include FVD, audio–video ImageBind similarity (AV-IB), and DeSync.

4.3 Experimental Setup

Following Liu et al. (2025), we train using OpenSora Zheng et al. (2024) with AdamW, dynamic batch sizing, and a LoRA compression ratio of 64. Stage 1 trains the audio branch for 50 epochs at a learning rate of 1×10^{-4} with 1,000 warm-up steps. Stage 2 performs joint fine-tuning for 10 epochs with the same learning rate and warm-up schedule, using loss weights $\beta_v = 0.3$ and $\beta_a = 0.7$. Stage 3 applies Modality Forcing for 10 epochs at 5×10^{-5} with 500 warm-up steps, retaining T2AV training alongside V2A and A2V forcing. Additional training configurations and computational costs are provided in the appendix.

4.4 Experimental Results

Table 1 compares our method with cascaded pipelines (Figure 1(1)) and end-to-end text-to-audio-video (T2AV) models. Cascaded baselines include T2A+A2V, using AudioLDM2 Liu et al. (2024a) followed by TempoToken Yariv et al. (2024) or TPoS Jeong et al. (2023), and T2V+V2A, using Wan2.1 followed by V2A models. Beyond earlier methods such as ReWaS Jeong et al. (2025) and Seeing&Hearing Xing et al. (2024), we include FoleyCrafter Zhang et al. (2026), MMAudio Cheng et al. (2025), AudioX Tian et al. (2025), and HunyuanVideo-Foley Shan et al. (2025). End-to-end baselines further include JavisDiT++, MOVA, and LTX-2. Our method achieves competitive joint-generation performance, particularly in text alignment and audiovisual coherence. Compared with the same-data OVI* baseline, it improves FVD (167.2 to 148.1), FAD (6.2 to 5.9), and AV-IB (0.207 to 0.211), while reducing DeSync (0.850 to 0.828). It also outperforms JavisDiT++ in TV-IB and AV-IB, although JavisDiT++ achieves better distributional quality. MOVA and LTX-2 perform better on several quality and synchronization metrics. Beyond T2AV, our framework supports bidirectional conditional generation through Modality Forcing, as evaluated next.

4.5 Bidirectional Conditional Generation

Video-to-Audio Generation. We evaluate V2A generation on the VGGSound test set. As shown in Table 2, our model achieves an FAD of 5.5, compared with 7.0 for AudioX and 8.2 for HunyuanVideo-Foley. It also obtains an IB-score of 0.338 and a DeSync score of 0.52. AudioX performs better in KL, IS, and DeSync, while HunyuanVideo-Foley achieves higher IB-score, indicating that our unified model remains competitive with dedicated V2A systems.

Table 1: Comparison results to existing JAVG models on JavisBench, each one generating 4s 240p and 16kHz sounding videos. Best results are in **bold**, second best are underlined. OVI⋆ is the same architecture as OVI Low et al. (2025) but trained from scratch using our datasets.

Model	AV-Quality		A&T—V&T				AV-Seman.		AV-Temp.	
	FVD↓	FAD↓	TV-IB↑	TA-IB↑	CLIP↑	CLAP↑	AV-IB↑	AVH↑	Javis↑	DeSync↓
Cascaded: T2A(AudioLDM2)+A2V										
TempoToken	539.8	6.5	0.084	0.132	0.205	0.387	0.139	0.122	0.103	1.532
TPoS	839.7	6.5	0.201	0.132	0.229	0.387	0.124	0.129	0.095	1.493
Cascaded: T2V(Wan2.1)+V2A										
ReWaS	197.5	9.4	0.255	0.123	0.281	0.280	0.110	0.104	0.079	1.071
See&Hear	197.5	7.6	0.255	0.129	0.281	0.263	0.160	0.143	0.112	1.099
FoleyCrafter	197.5	9.1	0.255	0.149	0.281	0.383	0.193	0.186	0.151	0.952
MMAudio	197.5	6.1	0.255	0.160	0.281	0.407	0.198	0.182	0.150	0.849
AudioX	197.5	6.2	0.255	0.170	0.281	0.401	0.198	0.187	0.159	0.891
HunyuanVideo-Foley	197.5	5.9	0.255	0.160	0.281	0.413	0.207	**0.191**	**0.165**	0.833
End-to-End: T2AV										
JavisDiT	204.1	7.2	0.263	0.143	0.302	0.391	0.197	0.179	0.154	1.039
OVI⋆	167.2	6.2	0.28	0.158	0.308	0.399	0.207	0.181	0.159	0.850
JavisDiT++	**141.5**	**5.5**	0.28	0.160	**0.316**	**0.424**	0.198	0.184	0.159	0.832
Ours	148.1	5.9	**0.29**	**0.171**	0.315	0.421	**0.211**	**0.191**	0.162	**0.828**
Uses Proprietary Data:										
UniVerse-1	194.2	8.7	0.272	0.111	0.309	0.245	0.104	0.098	0.077	0.929
OVI	121.7	5.2	**0.34**	**0.20**	**0.352**	**0.455**	**0.260**	0.209	0.188	0.792
MOVA	122.5	5.2	0.24	0.18	0.307	0.402	0.228	0.209	0.181	0.695
LTX-2	**116.7**	**4.9**	0.27	0.17	0.311	0.412	0.232	**0.223**	**0.192**	**0.569**

Table 2: Bidirectional conditional generation results on the VGGSound test set. Best results are in **bold**.

(a) Video-to-Audio Generation

Model	FAD↓	KL↓	IS↑	IB-score↑	DeSync↓
HunyuanVideo-Foley	8.2	2.14	16.14	**0.360**	0.53
AudioX	7.0	**1.48**	**17.91**	0.329	**0.46**
Ours	**5.5**	1.57	17.62	0.338	0.52

(b) Audio-to-Video Generation

Model	FVD↓	AV-IB↑	DeSync↓
TempoToken	627.5	0.166	0.97
TPoS	916.8	0.153	1.05
Ours	**278.1**	**0.278**	**0.75**

Audio-to-Video Generation. On the same test set, our model achieves an FVD of 278.1, compared with 627.5 for TempoToken and 916.8 for TPoS (Table 2). It also achieves an AV-IB of 0.278 and a DeSync of 0.75. Together, these results demonstrate that Modality Forcing enables both directions of conditional generation within a unified model.

4.6 Ablation Studies

We evaluate Modality Forcing and the supporting architectural and training components. Unless otherwise specified, joint-generation ablations are conducted on JavisBench under identical settings.

Effects of Modality Forcing. Removing the final adaptation stage reduces AV-IB from 0.211 to 0.192 and JavisScore from 0.162 to 0.146, while increasing DeSync from 0.828 to 0.922 (Table 3). TV-IB also decreases from 0.29 to 0.25. Thus, Modality Forcing improves joint-generation performance while enabling the bidirectional conditional generation evaluated above.

Loss Weight Ratios. We compare our default loss weights $\beta_v : \beta_a = 0.3 : 0.7$ with $0.7 : 0.3$ and $0.5 : 0.5$. The default setting achieves the lowest FAD and DeSync and the highest AV-IB and JavisScore among the evaluated ratios (Table 3). Although a higher video loss weight slightly improves CLIP, our setting provides a better balance of audiovisual quality, coherence, and synchronization.

Effects of Audio-only Training. We compare our hybrid adaptation strategy with full fine-tuning, FFN-only fine-tuning, and training the audio branch from scratch. As shown in Table 3, training

Table 3: Comparison results with different training settings.

Model	AV-Quality		A&T—V&T				AV-Seman.		AV-Temp.	
	FVD↓	FAD↓	TV-IB↑	TA-IB↑	CLIP↑	CLAP↑	AV-IB↑	AVH↑	Javis↑	DeSync↓
Effects of modality-forcing fine-tuning:										
No Modality-Forcing	150.7	6.0	0.25	0.159	0.305	0.408	0.192	0.186	0.146	0.922
Loss weight ratios:										
β_v:β_a=0.7:0.3	148.8	6.2	0.28	0.171	0.318	0.410	0.208	0.181	0.157	0.840
β_v:β_a=0.5:0.5	149.6	6.0	0.27	0.169	0.310	0.419	0.210	0.188	0.160	0.832
Effects of various audio-only training schemes:										
Audio-Fully-FT	161.5	6.0	0.32	0.165	0.311	0.411	0.205	0.180	0.155	0.908
Audio-FFN-only	167.2	6.3	0.35	0.155	0.306	0.403	0.201	0.180	0.151	0.951
Audio-from-scratch	183.7	6.8	0.46	0.156	0.297	0.389	0.193	0.172	0.125	1.031
Ours	**148.1**	**5.9**	**0.29**	**0.171**	**0.315**	**0.421**	**0.211**	**0.191**	**0.162**	**0.828**

Table 4: Generalization of Modality Forcing to JavisDiT. T2AV generation is evaluated on JavisBench, while V2A and A2V generation are evaluated on the VGGSound test set. Best results are in **bold**.

Model	T2AV				V2A			A2V		
	FVD↓	FAD↓	AV-IB↑	DeSync↓	FAD↓	IB-score↑	DeSync↓	FVD↓	AV-IB↑	DeSync↓
JavisDiT	204.1	7.2	0.197	1.039	8.5	0.316	0.81	415.8	0.192	0.89
+ Modality Forcing	**188.2**	**6.5**	**0.206**	**1.010**	**6.8**	**0.325**	**0.69**	**352.5**	**0.235**	**0.81**

from scratch increases FVD from 148.1 to 183.7 and DeSync from 0.828 to 1.031, while reducing JavisScore from 0.162 to 0.125. Our hybrid strategy also outperforms full and FFN-only fine-tuning in audiovisual coherence and synchronization, supporting the reuse of pretrained video weights for audio adaptation.

Generalization of Modality Forcing. To assess transferability beyond our Wan2.1-based framework, we apply Modality Forcing to JavisDiT without changing its architecture. Table 4 shows consistent improvements across T2AV on JavisBench and V2A/A2V on the VGGSound test set. T2AV FVD decreases from 204.1 to 188.2, V2A FAD from 8.5 to 6.8, and A2V FVD from 415.8 to 352.5, with improved cross-modal alignment and synchronization across all three tasks. These results demonstrate that Modality Forcing also benefits an audiovisual generator with a different synchronization mechanism.

Architectural Variants. We compare our design with `Dual-Branch-SelfAttn`, which replaces bidirectional cross-attention with joint self-attention, and `Shared-Branch`, which uses a shared backbone (Figure 2). Our model outperforms both variants in generation quality, semantic correspondence, and synchronization (Table 5). Compared with our design, the shared-branch variant reduces AV-IB from 0.211 to 0.157 and increases DeSync from 0.828 to 1.051, supporting modality-specific modeling with explicit cross-modal interaction.

Fusion Block Variants. We compare bidirectional fusion with no fusion and V2A-only or A2V-only fusion. Removing fusion reduces AV-IB from 0.211 to 0.121 and increases DeSync from 0.828 to 1.81 (Table 5). Both unidirectional variants improve upon no fusion but remain below the full model, highlighting the importance of bidirectional interaction for joint generation.

User Study. We evaluate 80 samples generated by our method and three baselines from 20 randomly selected JavisBench prompts. Five participants rate each sample on a five-point Likert scale across video quality, audio quality, text–audio–video consistency, audio–video semantic alignment, and temporal synchronization, with samples shuffled within each prompt. As shown in Table 6, our method achieves the highest average scores in four of the five dimensions and overall quality, while tying with UniVerse-1 in semantic alignment.

Table 5: Comparison results with model settings including architectural variants, and fusion variants.

Model	AV-Quality		A&T—V&T				AV-Seman.		AV-Temp.	
	FVD↓	FAD↓	TV-IB↑	TA-IB↑	CLIP↑	CLAP↑	AV-IB↑	AVH↑	Javis↑	DeSync↓
v.s. Architectural Variants:										
Variant-1	150.8	6.1	0.27	0.156	0.307	0.409	0.198	0.176	0.160	0.895
Variant-2	161.5	6.7	0.25	0.122	0.291	0.363	0.157	0.138	0.129	1.051
v.s. Fusion Block Variants:										
V+A, w/o fusion	172.7	6.7	0.22	0.109	0.267	0.349	0.121	0.115	0.085	1.81
V+A, v2a attn.	164.6	6.5	0.24	0.137	0.281	0.392	0.165	0.134	0.113	1.39
V+A, a2v attn.	165.7	6.5	0.25	0.140	0.297	0.389	0.162	0.135	0.117	1.41
Ours	**148.1**	**5.9**	**0.29**	**0.171**	**0.315**	**0.421**	**0.211**	**0.191**	**0.162**	**0.828**

Table 6: Comparison results from the user study.

Model	Video Qual.	Audio Qual.	Text-AV Align.	Sem. Align.	Temp. Sync.	Overall
MM-Diff	2.4	2.6	3.4	2.6	2.4	2.7
JavisDiT	3.4	3.6	3.6	3.6	3.6	3.6
Universe-1	3.8	4.0	4.0	**4.6**	4.0	4.1
Ours	**4.0**	**4.6**	**4.2**	**4.6**	**4.4**	**4.4**

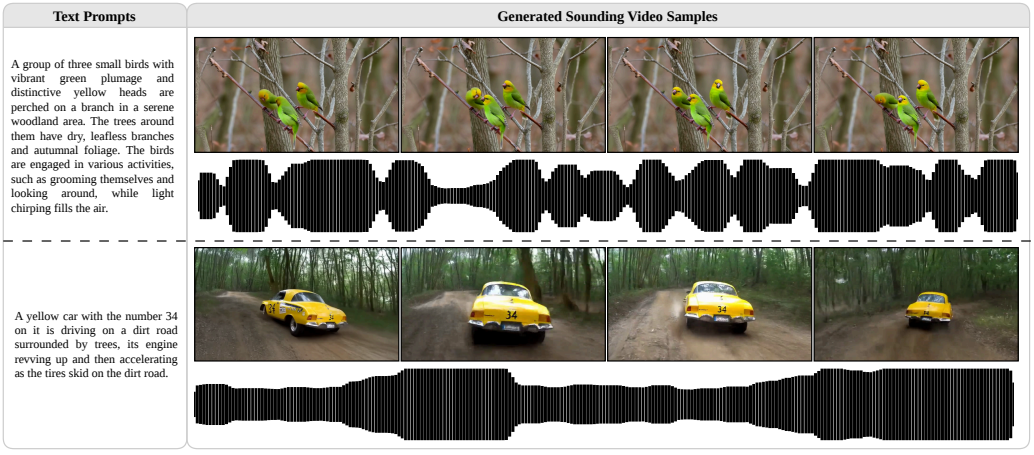


Figure 4: Visualization of generated video-audio samples from our proposed model.

Qualitative Results. Figure 4 visualizes generated sounding video samples using video frames and audio waveforms conditioned on text prompts. The results show that our model produces high-quality and text-consistent audio and video while maintaining reasonable audio–video synchronization.

5 Conclusion

In this work, we present a progressive adaptation framework that extends a pretrained video generator (Wan2.1) into a unified audio–video model. The framework first acquires audio generation capability through audio-only training and then learns synchronized joint generation through bidirectional cross-modal interaction. Our key contribution, *Modality Forcing*, further enables bidirectional conditional generation through task-dependent information routing and asymmetric optimization while retaining joint-generation capability. Extensive experiments demonstrate competitive joint audio–video generation, strong V2A and A2V performance, and preservation of the pretrained video generation capability. These results highlight the effectiveness of progressive adaptation and Modality Forcing for unified audiovisual generation.

AI USE STATEMENT

We used AI tools for language polishing to improve the clarity, grammar, and conciseness of the manuscript. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Humo: Human-centric video generation via collaborative multi-modal conditioning. *arXiv preprint arXiv:2509.08519*, 2025a.

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*, 2025b.

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*, pp. 28901–28911, 2025.

Xiaowei Chi, Yatian Wang, Aosong Cheng, Pengjun Fang, Zeyue Tian, Yingqing He, Zhaoyang Liu, Xingqun Qi, Jiahao Pan, Rongyu Zhang, et al. Mmtrail: A multimodal trailer video dataset with language and music descriptions. *arXiv preprint arXiv:2407.20962*, 2024.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Google DeepMind. Veo 3. `https://deepmind.google/models/veo/`, 2025.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 976–980. IEEE, 2022.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Xiaohu Huang, Hao Zhou, Qiangpeng Yang, Shilei Wen, and Kai Han. Jova: Unified multimodal learning for joint video-audio generation. *arXiv preprint arXiv:2512.13677*, 2025.

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

Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7822–7832, 2023.

Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17590–17598, 2025.

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)*, pp. 119–132, 2019.

Gwanghyun Kim, Alonso Martinez, Yu-Chuan Su, Brendan Jou, José Lezama, Agrim Gupta, Lijun Yu, Lu Jiang, Aren Jansen, Jacob Walker, et al. A versatile diffusion transformer with mixture of noise levels for audiovisual generation. *Advances in Neural Information Processing Systems*, 37: 11837–11865, 2024.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

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.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11474–11481, 2020.

Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

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

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.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. *arXiv preprint arXiv:2602.19163*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Xiulong Liu, Kun Su, and Eli Shlizerman. Tell what you hear from what you see-video to audio generation through text. *Advances in Neural Information Processing Systems*, 37:101337–101366, 2024b.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36: 48855–48876, 2023.

Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Tavgbench: Benchmarking text to audible-video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6607–6616, 2024.

Irene Martín-Morató and Annamaria Mesaros. What is the ground truth? reliability of multi-annotator data for audio tagging. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 76–80. IEEE, 2021.

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.

OpenAI. Sora 2. `https://openai.com/zh-Hans-CN/index/sora-2/`, 2025.

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

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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*, pp. 10219–10228, 2023.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Christoph Schuhmann. Improved aesthetic predictor, March 2022. URL https://github.com/christophschuhmann/improved-aesthetic-predictor.

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.

Bob L Sturm. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: Diffusion transformer for anything-to-audio generation. *arXiv e-prints*, pp. arXiv–2503, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, and Daxin Jiang. 1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 1(2):3, 2025a.

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

Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15492–15501, 2024a.

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*, pp. 10486–10495, 2025c.

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

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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*, pp. 7151–7161, 2024.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13941–13958, 2023.

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.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6639–6647, 2024.

Hongwei Yi, Tian Ye, Shitong Shao, Xuancheng Yang, Jiantong Zhao, Hanzhong Guo, Terrance Wang, Qingyu Yin, Zeke Xie, Lei Zhu, et al. Magicinfinite: Generating infinite talking videos with your words and voice. *arXiv preprint arXiv:2503.05978*, 2025.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, Bin Liu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *International Journal of Computer Vision*, 134(1):46, 2026.

Lei Zhao, Linfeng Feng, Dongxu Ge, Rujin Chen, Fangqiu Yi, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. Uniform: A unified multi-task diffusion transformer for audio-video generation. *arXiv preprint arXiv:2502.03897*, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

A Appendix

A.1 Model Implementation

Model Backbone. Our model builds upon the Wan2.1-1.3B text-to-video diffusion backbone Wan et al. (2025) and extends it to text-conditioned joint audio–video generation. The underlying diffusion transformer comprises 30 Transformer layers with a hidden dimension of 1536. To accommodate multimodal generation, we introduce a dual-branch architecture that jointly models visual and acoustic representations, while largely preserving the architectural design and generative capacity of the original video model.

Audio Representation. We employ a pretrained 1D variational autoencoder from MMAudio Cheng et al. (2025) to obtain compact acoustic latent representations. Raw audio waveforms are first converted into mel-spectrograms using Short-Time Fourier Transform and then encoded into latent tokens by the audio VAE encoder. During inference, the generated audio latents are decoded into spectrograms and subsequently synthesized into waveforms using a pretrained HiFi-GAN vocoder Kong et al. (2020). Operating in this latent acoustic space facilitates efficient integration of audio generation within the diffusion transformer framework while retaining fine-grained temporal structure.

Video Representation. We reuse the pretrained video VAE from Wan2.1 and keep it frozen to encode raw video frames into a compact latent space. Specifically, all frames except the first are temporally downsampled by a factor of four, while each frame is spatially compressed by 8×8, producing latent embeddings with 16 channels. For additional computational efficiency, the resulting latent feature maps are further partitioned into 2×2 spatial patches before being fed into the diffusion transformer. This hierarchical compression substantially reduces the token sequence length while preserving sufficient spatiotemporal structure for high-fidelity video synthesis.

Text Condition. We reuse the pretrained umT5-XXL text encoder from Wan2.1 and keep it frozen during training. The encoder supports a maximum context length of 512 tokens. The resulting text embeddings serve as shared semantic conditioning for both the audio and video branches, enabling the model to perform unified multimodal generation guided by natural language prompts.

A.2 Training Details

We adopt a three-stage training scheme to progressively extend the pretrained video generation backbone toward unified audio–video synthesis while preserving training stability and cross-modal coherence. The detailed configurations are summarized in Table 7.

In the first stage, we introduce the audio branch on top of the Wan2.1 backbone and focus on audio-only optimization to establish reliable acoustic modeling. We employ a hybrid fine-tuning strategy that combines parameter-efficient LoRA adaptation for attention modules (self-attention and cross-attention) with full updates of audio-specific components, including the audio embedder, output head, and feed-forward layers. This setting facilitates efficient transfer of the pretrained video prior to the acoustic domain. Training is performed on approximately 780K audio samples for 50 epochs using a learning rate of 1×10^{-4} with 1000 warm-up steps. To improve duration generalization, we adopt a variable-length training scheme in which audio segments of different temporal spans are randomly sampled, and the batch size is dynamically adjusted to maintain high hardware utilization.

In the second stage, we enable joint optimization of the audio and video branches to learn cross-modal representations and audiovisual interaction dynamics. Newly introduced audio–video cross-attention layers are fully updated, while other parameters continue to be adapted via LoRA. This stage is trained on approximately 611K paired audio–video samples for 10 epochs with a learning rate of 1×10^{-4} and 1000 warm-up steps. The spatial resolution, audio duration, and batch size are dynamically scheduled during training to maximize computational efficiency.

In the final stage, we apply a modality-forcing fine-tuning strategy to further enhance conditional controllability and cross-modal alignment. At each iteration, the audio branch, the video branch, or both are randomly selected as optimization targets. When one modality is designated for generation, the other remains unnoised and serves as a clean conditioning signal. To retain the model's native joint generation capability, a fraction of standard audio–video joint training is preserved. LoRA adaptation is applied to all relevant layers under each modality-forcing configuration. We adopt a reduced

learning rate of 5×10^{-5} with 500 warm-up steps and train on approximately 355K samples for 10 epochs. As in previous stages, spatial resolution, temporal duration, and batch size are dynamically adjusted throughout training. For all training stages, experiments are conducted using PyTorch on 4 NVIDIA H200 GPUs.

Table 7: Details for multi-stage training pipeline. 'A' and 'V' mean Audio branch and Video branch, respectively. 'Full FT' denotes the full fine-tuning.

Setting	**Stage 1**	**Stage 2**	**Stage 3**
Training Stage	Audio-only Training	Audio-Video Training	Modality-Forcing
Trainable branches	A	A+V	A, or V or A+V
Trainable type	LoRA + Full FT	LoRA + Full FT	LoRA
Trainable params	925M	512M	172M
Learning rate	1e-4	1e-4	5e-5
Warm-up steps	1000	1000	500
Weight decay	0.0	0.0	0.0
Training samples	780K	611K	355K
Resolution	-	Dynamic	Dynamic
Duration	Variable	Variable	Variable
Batch size	Dynamic	Dynamic	Dynamic
Epoch	50	10	10
GPU Hours	390	680	360

A.3 Subjective Evaluation

We additionally conduct a subjective user study to assess the perceptual performance of our model in comparison with three representative baseline approaches, namely MM-Diffusion Ruan et al. (2023), JavisDiT Liu et al. (2025), and Universe-1 Wang et al. (2025b). To construct the evaluation set, we randomly select 20 text prompts from the JavisBench benchmark Liu et al. (2025). Each of the four models (three baselines and our method) is used to generate corresponding sounding video samples, resulting in a total of 80 generated videos. For each prompt, the generated samples from different methods are grouped together and randomly shuffled before being presented to participants in order to reduce potential ordering bias and ensure a fair comparison.

We recruit five human participants to evaluate all generated samples across five perceptual dimensions, including video quality, audio quality, text–audio–video consistency, audio–video semantic alignment, and audio–video temporal synchronization. For each dimension, participants are provided with separate evaluation instructions describing the specific aspect to be considered. Ratings are collected using a five-point Likert scale Likert (1932), where scores from 1 to 5 correspond to strongly disagree, disagree, neutral, agree, and strongly agree, respectively. During evaluation, participants are asked to focus on the designated criterion for each dimension while disregarding potential influences from other perceptual factors. This protocol is intended to encourage more disentangled and reliable subjective judgments. The detailed definitions of each evaluation dimension are provided as follows:

(a) **Video quality.** This criterion measures the perceptual fidelity of the generated videos, focusing on visual sharpness, motion coherence, and the extent of observable artifacts. The evaluation is conducted independently of the accompanying audio signal or any form of audiovisual correspondence.

(b) **Audio quality.** This dimension evaluates the perceptual realism of the generated audio, including clarity, naturalness, and the presence of noise or distortion. Judgments under this criterion are made without considering the visual stream or multimodal semantic consistency.

(c) **Text–audio–video consistency.** This criterion assesses whether the generated audiovisual content jointly conveys the semantic intent of the input textual prompt. The evaluation emphasizes overall semantic coherence across modalities, rather than perceptual fidelity or temporal synchronization accuracy.

(d) **Audio–video semantic alignment.** This dimension measures the semantic plausibility of the generated audio with respect to the visual scene, i.e., whether the produced sound events are consistent

with the depicted visual context. Audio quality and fine-grained temporal correspondence are not considered in this assessment.

(e) **Audio–video temporal synchronization.** This criterion evaluates the temporal alignment between audio events and visual dynamics, focusing on whether sound occurrences coincide with visually meaningful moments without noticeable delay or anticipation. The assessment is restricted to timing accuracy, independent of semantic relevance or perceptual quality.

A.4 Limitation and Future Work

Despite achieving encouraging performance in synchronized sounding video generation, the proposed framework still has several limitations. First, the current design primarily targets ambient sound synthesis and does not explicitly address other important audio categories such as speech and music, which play a central role in many real-world audiovisual scenarios. As a result, the model remains less suitable for complex multimedia settings involving structured acoustic events or human communication.

Second, the generation process follows a diffusion-based paradigm, leading to relatively high inference latency, especially when synthesizing long audiovisual sequences. This computational overhead restricts the applicability of the model in time-critical scenarios, such as interactive systems and real-time immersive media experiences.

Future work will focus on extending the diversity and scale of training data by incorporating richer sound categories, including speech, music, and structured environmental audio. Such data could facilitate more robust and generalizable joint audio–video generation. Building on this direction, developing scalable audiovisual world models that capture complex scene dynamics and long-range temporal dependencies is an important next step. In addition, exploring more efficient generation mechanisms and model acceleration strategies is necessary to improve inference efficiency and support real-time deployment in practical multimedia applications.

A.5 More Visualizations

We provide additional visualizations of generated video frames and corresponding audio spectrograms produced by our model using randomly sampled prompts from the JavisBench benchmark. Representative examples are shown in Figure 5 and Figure 6.

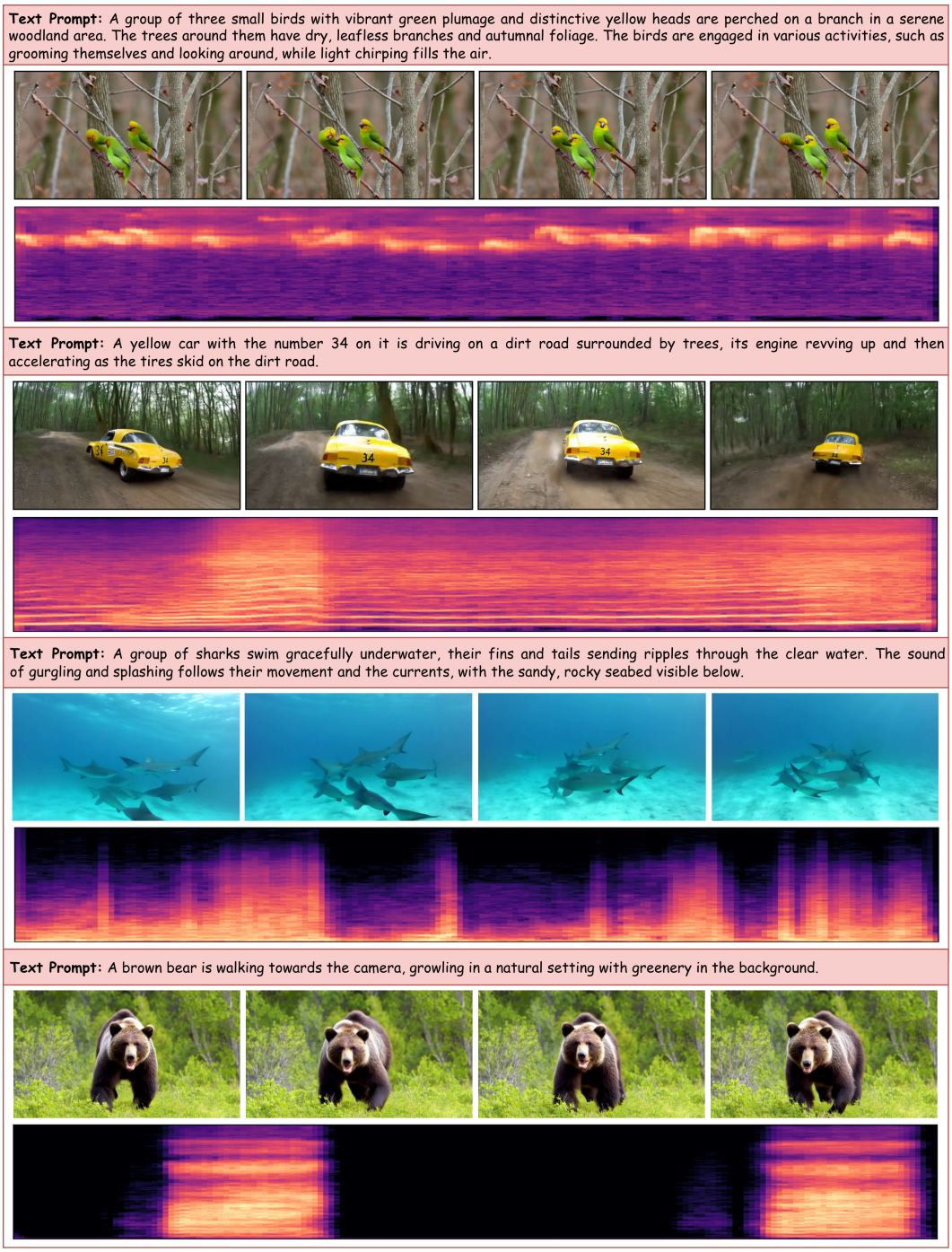


Figure 5: More visualizations of generated video-audio samples from our proposed model.

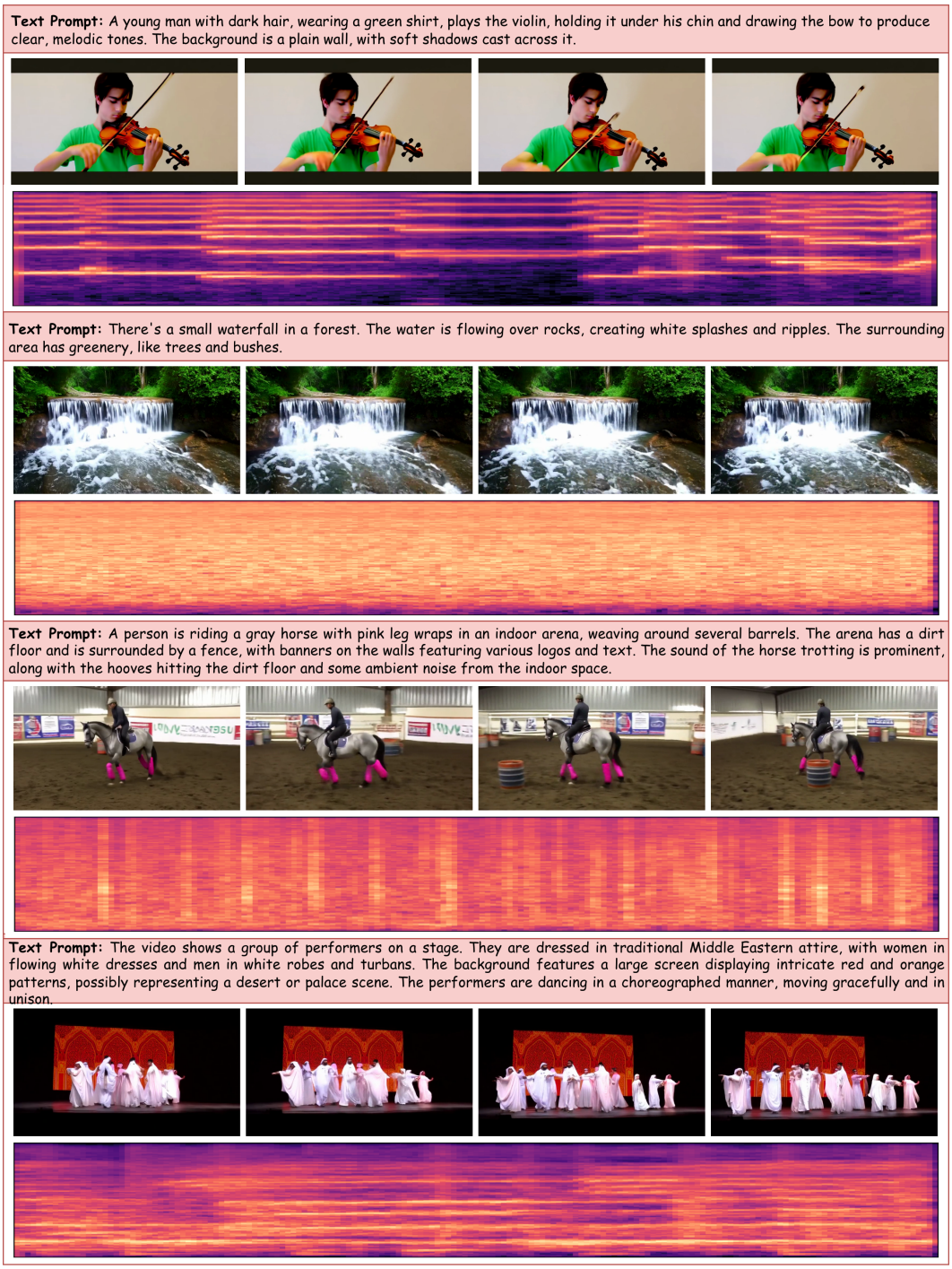


Figure 6: More visualizations of generated video-audio samples from our proposed model.