IMPROVED QUALITY, SYNCHRONY, AND PREFERENCE ALIGNMENT FOR JOINT AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

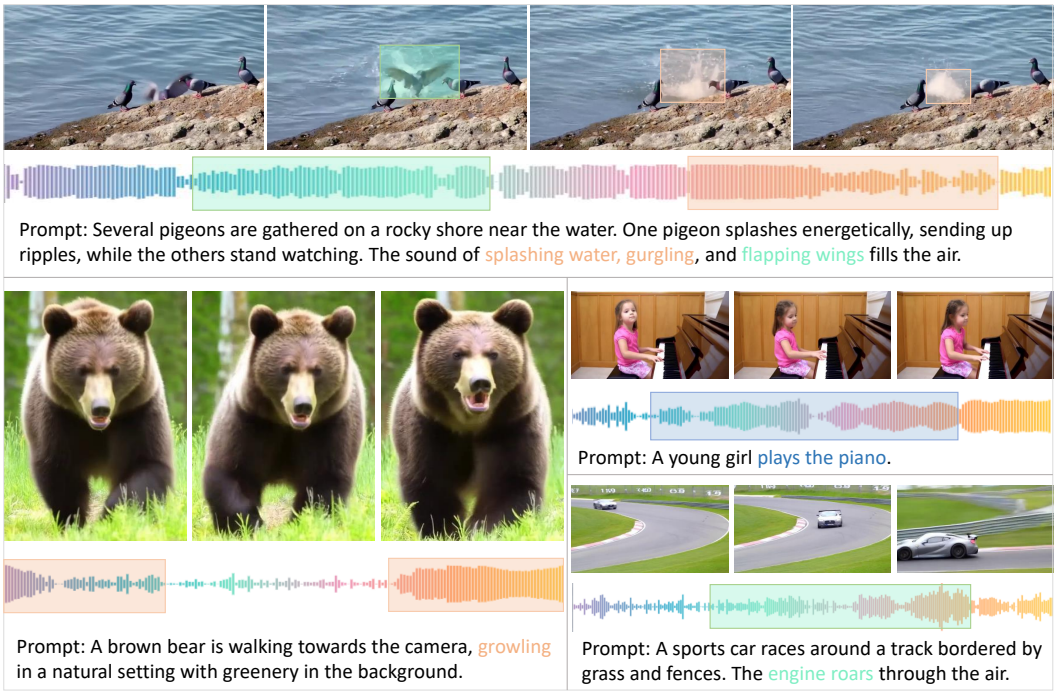


Figure 1: Examples of joint audio-video generation results by our proposed method.

ABSTRACT

Recent AIGC advances have rapidly expanded from text-to-image generation toward high-quality multimodal synthesis across video and audio. Within this context, joint audio-video generation (JAVG) has emerged as a fundamental task, enabling synchronized and semantically aligned sound and vision from textual descriptions. However, compared with advanced proprietary systems such as Veo3, existing open-source methods still suffer from limitations in generation quality, temporal synchrony, and alignment with human preferences. This paper presents a concise yet powerful framework for efficient and effective JAVG. First, we introduce a modality-specific mixture-of-experts (MS-MoE) design that enables effective cross-modal communication while enhancing single-modality generation quality. Then, we propose a temporal-aligned RoPE (TA-RoPE) strategy to achieve explicit, frame-level synchronization between audio and video tokens. Besides, we develop an audio-video direct preference optimization (AV-DPO) method to align model outputs with human preference across quality, consistency, and synchrony dimensions. Built upon Wan2.1-1.3B-T2V, our model achieves state-of-the-art performance merely with around 1M training entries, significantly outperforming prior approaches in both qualitative and quantitative evaluations. Comprehensive ablation studies have been conducted to validate the effectiveness of our proposed modules. We hope this work can set a milestone for the field of native JAVG and bring new inspiration to the community. The code implementation can be found in `https://anonymous.4open.science/r/iclr26-19605-code/`.

1 INTRODUCTION

AIGC (Artificial Intelligence Generated Content) technologies have evolved from text-to-image generation towards more complex domains such as video and audio (Wan et al., 2025; Zheng et al., 2024; Liu et al., 2024b; Jiang et al., 2025). The intrinsic correlation between audio and video modalities has also been increasingly explored, giving rise to related research on video-to-audio (Xing et al., 2024; Cheng et al., 2025; Tian et al., 2025) and audio-to-video (Jeong et al., 2023a; Yariv et al., 2024b; Gao et al., 2025b) generation. With the rapidly growing demands for AI-generated content in domains like short videos, film, gaming, and VR, native joint audio-video generation (JAVG) from textual inputs has become increasingly crucial in this era (Wang et al., 2025a; Gao et al., 2025b).

MM-Diffusion (Ruan et al., 2023) represents a series of early works that study unconditional audio-video generation, with a scope limited to natural landscapes (Landscape (Lee et al., 2022)) and human dancing (AIST++ (Li et al., 2021)). Then, Uniform (Zhao et al., 2025) extended the task to label-name-conditioned JAVG and evaluated the model on the relatively broader VGGSound dataset (Chen et al., 2020). More recently, JavisDiT (Liu et al., 2025c) and Universe-1 (Wang et al., 2025a) have begun to employ free-form text prompts as inputs, systematically exploring sounding video generation in real-world scenarios. Despite these efforts, current approaches still struggle to produce high-quality, temporally synchronized sounding videos compared with advanced proprietary models such as Veo3 (DeepMind, 2025). Fig. 2 quantitatively show the cap between previous methods and Veo3 (details are provided in Sec. B.3). Current JAVG methods fail to capture human preferences regarding the aesthetics and harmony of audio-visual content.

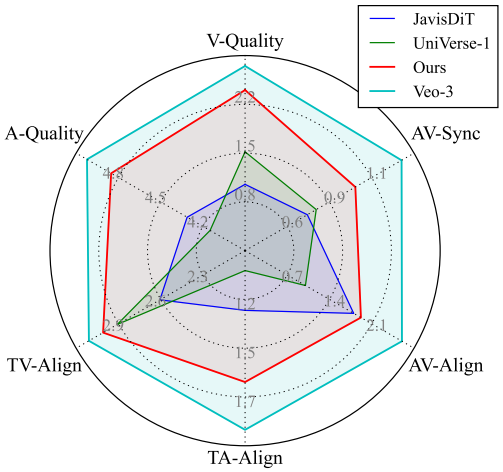


Figure 2: Comparison with recent JAVG models.

To bridge the gap, we propose a concise and elegant framework to efficiently train a DiT model to generate *high-quality*, *better-synchronized*, and *human-preference-aligned* sounding videos.

First, to improve generation quality, we introduce a **modality-specific mixture-of-experts (MS-MoE)** module, where audio and video tokens exchange information through shared multi-head self-attention layers, and then each modality aggregates information via two separate FFN layers (Deng et al., 2025). This architecture enhances single-modality generation quality compared with Uniform (Zhao et al., 2025), which processes aggregated audio-video tokens using a single FFN module. On the other hand, MS-MoE is simpler, more efficient, and more unified than dual-DiT plus audio-visual interaction block designs such as JavisDiT (Liu et al., 2025c) and Universe-1 (Wang et al., 2025a). Sec. C provides detailed architectural comparison with related works.

Second, to enhance audio-video synchrony, we propose an **temporal-aligned multimodal RoPE (TA-RoPE)** strategy, which aligns the position IDs of audio and video tokens on a unified temporal axis Xu et al. (2025), enabling direct and frame-level fine-grained temporal synchronization. On one hand, this strategy modulates audio-video temporal synchrony more explicitly and effectively against the ST-Prior in JavisDiT (Liu et al., 2025c) and the Stitching strategy in Universe-1 (Wang et al., 2025a). On the other hand, TA-RoPE is actually compatible with these methods — it can be combined with ST-Prior and frame-level cross-attention for slightly better performance. However, we ultimately abandon these combinations to maintain overall simplicity and efficiency.

Finally, to capture human preference, we design an **audio-video direct preference optimization (AV-DPO)** method to align the baseline model with curated preference data (Liu et al., 2025b), further enhancing the video-audio quality and synchronization. To curate robust preference data, we leverage diverse reward models to comprehensively evaluate the generated audio and video across multiple dimensions (*e.g.*, quality, consistency, and synchrony) and then adopt normalized modality-aware ranking to select winning-losing pairs. To the best of our knowledge, we are the first to in-

troduce preference alignment into JAVG, enabling the model to generate high-quality, synchronized sounding videos, more faithfully aligned with the input text and human preferences.

Our model is built upon Wan2.1-1.3B-T2V (Wan et al., 2025), and is efficiently trained using only 780K diversified audio-text pairs (Liu et al., 2025c) and 360K high-quality sounding videos (Mao et al., 2024), supporting arbitrary duration and resolution ranging from 2–5 seconds and 240p–480p across different aspect ratios. It achieves state-of-the-art performance, surpassing JavisDiT (Liu et al., 2025c) and Universe-1 (Wang et al., 2025a) across both qualitative and quantitative evaluations on various dimensions. Extensive ablation studies demonstrate the effectiveness and rationality of the proposed modules. We hope this work will set a milestone for the field of native joint audio-video generation and provide further inspiration in the field.

In summary, our main contributions are threefold:

- We propose a concise JAVG model architecture, which employs a modality-specific MoE strategy for efficient and high-quality audio-video generation, and introduces temporally aligned RoPE to achieve precise temporal synchronization.
- We are the first to introduce human preference alignment into JAVG, with the AV-DPO algorithm to consistently improve quality, consistency, and synchrony of sounding video generation.
- We train a state-of-the-art JAVG model using only 1M data entries, providing an important reference milestone and new inspirations for the field of native joint audio-video generation.

2 Related Work

Joint Audio-Video Generation. Recent JAVG approaches have taken several forms. Some studies use a unified representation, projecting both modalities into a shared latent space, such as CoDi (Tang et al., 2023; 2024), MM-LDM (Sun et al., 2024), and UniForm (Zhao et al., 2025). However, this strong constraint can cause modality-specific information loss and provide insufficient temporal control. Another line of studies performs intermediate fusion. MM-Diffusion (Ruan et al., 2023), SyncFlow (Liu et al., 2024a), and AV-DiT (Wang et al., 2024) exchange information between modalities within the model's layers using cross-modal attention or adapters. Other methods like Seeing (Xing et al., 2024) and MMDisCo (Hayakawa et al., 2024) use online discriminators to adjust the outputs of separately pre-trained models. More recently, JavisDiT (Liu et al., 2025c) develops a two-stream DiT with a spatio-temporal prior estimator to guide alignment, and UniVerse-1 (Wang et al., 2025a) leverages two pretrained DiTs with a stitching strategy for crossmodal information exchange. However, these methods all rely on complicated model architectures and ad-hoc implicit synchrony modulation, hindering the scalability of unified sounding video generation.

RL in Generative Models. Reinforcement learning (RL) has been widely applied to align generative models with human preferences. Early approaches use policy-based algorithms such as Proximal Policy Optimization (PPO) to improve diffusion models (Black et al., 2023; Fan et al., 2023), and Direct Preference Optimization (DPO) is subsequently introduced to align text-to-image generation without explicit reward models (Wallace et al., 2024; Yang et al., 2024). This area continues to develop, with recent work improving DPO (Dang et al., 2025; Huang et al., 2025) and adopting newer strategies like Group-wise Ranking Preference Optimization (GRPO) (Wang et al., 2025b; Xue et al., 2025; Liu et al., 2025a; Yuan et al., 2025). Although these alignment techniques are now widely used for other modalities like video and audio (Zhang et al., 2023; Liu et al., 2025b;d; Furuta et al., 2024; Wu et al., 2025; Gao et al., 2025a; Chen et al., 2025), their application to complex, cross-modal tasks remains limited. To the best of our knowledge, we are the first to successfully apply a preference alignment algorithm to the field of joint audio-video generation.

3 Methodology

3.1 Preliminary

We use flow-matching as the noise scheduler to optimize our DiT model, where RoPE plays a vital role in ensuring temporally synchronized audio-video generation.

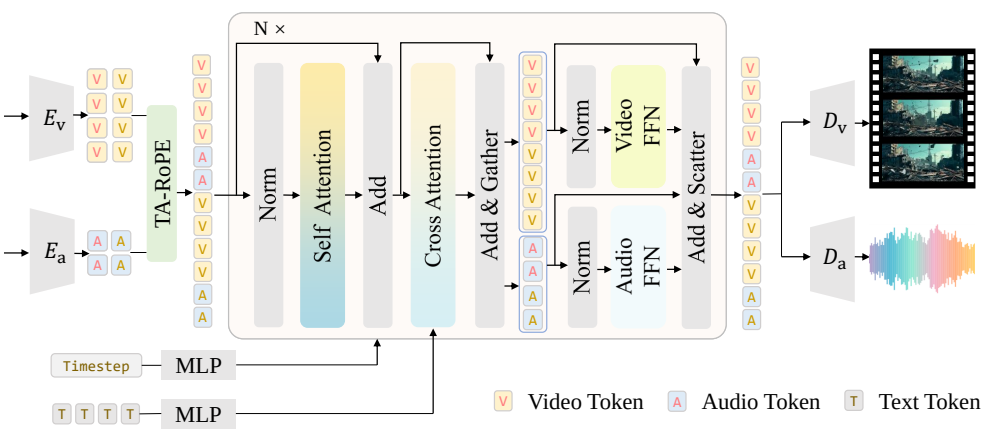


Figure 3: Illustration of the DiT architecture. We use shared attention layers to encourage audio-visual mutual information modeling, with modality-specific FFN layers to enhance intra-modal aggregation. The Temporal-Aligned RoPE strategy is applied to ensure audio-video synchrony.

Flow-Matching. Let p_0 denote the target data distribution and p_1 be a simple prior distribution (*e.g.*, a standard Guassian $\mathcal{N}(\mathbf{0}, \mathbf{I})$). Rectified Flow (Liu et al., 2023) constructs straight-line paths from noise samples $\mathbf{x}_1 \sim p_1$ to data samples $\mathbf{x}_0 \sim p_0$, where the trajectory $\mathbf{x}_t$ at time $t \in [0, 1]$ is defined as $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$. The corresponding target velocity field becomes $\mathbf{v} = \mathbf{x}_1 - \mathbf{x}_0$. A neural network $\mathbf{v}_\theta(\mathbf{x}_t, t)$ is then trained to regress this target field by minimizing the following objective:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim[0,1],\mathbf{x}_0\sim p_0,\mathbf{x}_1\sim p_1}\left[||\mathbf{v}_\theta(\mathbf{x}_t, t) - \mathbf{v}||^2\right] \tag{1}$$

Once trained, new samples can be generated by solving the ordinary differential equation $\mathrm{d}\mathbf{x}/\mathrm{d}t = v_\theta(\mathbf{x}, t)$ from $t = 0$ to $t = 1$, starting from an initial noise sample $\mathbf{x}_0 \sim p_0$.

Joint Audio-Video Generation. The goal of JAVG is to model the conditional distribution $p(A, V \mid c)$, where $V \in \mathbb{R}^{T_v \times H \times W \times 3}$ denotes a video with T_v frames of resolution $H \times W$, and $A \in \mathbb{R}^{T_a \times M}$ denotes the audio in a mel-spectrogram with T_a temporal steps and M frequency bins. Given a condition c (*e.g.*, a text prompt), a model p_θ is trained to generate synchronized audio-video pairs:

$$(\hat{\mathbf{x}}^a, \hat{\mathbf{x}}^v) = \mathbf{v}_\theta(\mathbf{x}_t^a, \mathbf{x}_t^v, t, c); \quad \mathcal{L}_{\text{FM}}^{av}(\theta) = \mathbb{E}_{t\sim[0,1],\mathbf{x}_0\sim p_0,\mathbf{x}_1\sim p_1}\left[||\hat{\mathbf{x}}^a - \mathbf{v}^a||^2 + ||\hat{\mathbf{x}}^v - \mathbf{v}^v||^2\right] \tag{2}$$

Rotary Position Embedding (RoPE). In video generation (Wan et al., 2025), position IDs are assigned along temporal (T), height (H), and width (W) dimensions, and their rotational position embeddings are applied to queries and keys in attention layers:

$$R(t, h, w) = [R_T(t); R_H(h); R_W(w)]\,; \quad (q_i', k_j') = (R(t_i, h_i, w_i)q_i,\ R(t_j, h_j, w_j)k_j). \tag{3}$$

This design captures the relative positional relationships of tokens across all three dimensions.

3.2 The DiT Model Architecture

From Dual-Stream DiT to a Unified Backbone. Unlike dual-stream DiT frameworks such as JavisDiT (Liu et al., 2025c) and UniVerse-1 (Wang et al., 2025a), this paper aims to design a *concise*, *efficient*, and *unified* DiT architecture to jointly process audio and video tokens, ensuring better scalability. As shown in Fig. 3, we first flatten and concatenate video and audio tokens, feeding them into subsequent modules for full self-attention to enable dense and rich cross-modal interaction. The tokens are then separated and passed through modality-specific FFNs, which ensure sufficient intra-modal aggregation. The video VAE from Wan2.1 (Wan et al., 2025) and audio VAE from AudioLDM2 (Liu et al., 2024b) are retained and frozen during the whole training process.

Modality-Specific Feed-Forward Network. In contrast to conventional MoE architectures with dynamic routing (Cai et al., 2025), our MS-FFN (or MS-MoE) deterministically assigns audio and

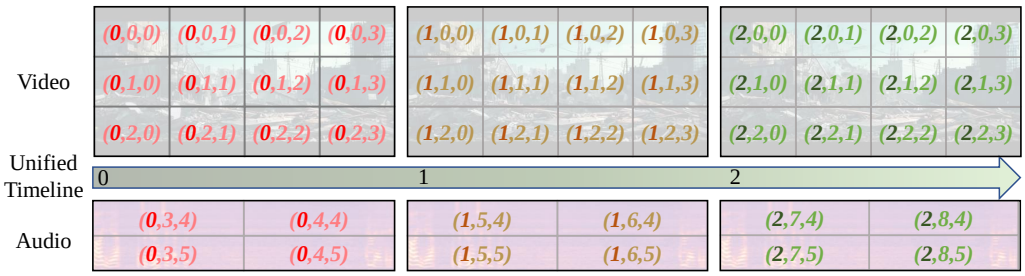


Figure 4: Illustration of temporal-aligned rotary position encoding for video and audio tokens.

video tokens to their modality-specific FFNs/Experts. This design is similar to BAGEL (Deng et al., 2025), which allocates understanding and generation tokens to separate FFNs, while we instead assign tokens based on modality. The advantage is that, after sufficient cross-modal interaction through attention, modality interference in FFN is isolated, allowing each branch to focus on intra-modal feature modeling. Similar to traditional MoE benefits, although the total parameter size increases from 1.3B to 2.1B, the number of activated parameters per token remains 1.3B. Thus, we expand model capacity to improve performance without adding inference overhead.

3.3 Temporal-Aligned Rotary Position Encoding

Establishing a shared temporal reference for audio and video tokens is crucial for achieving generation synchrony (Liu et al., 2025c; Wang et al., 2025c). In contrast to the spatial-temporal prior (ST-Prior) employed in JavisDiT (Liu et al., 2025c) and the frame-level cross-attention mechanism in UniVerse-1 (Wang et al., 2025a), this paper proposes a more direct and precise control strategy: an audio-visual Temporally Aligned Rotary Position Encoding (TA-RoPE). Specifically, absolute temporal alignment is enforced along the first dimension (dimension 0) of the 3D position IDs for both audio and video tokens, as illustrated in Fig. 4.

We first retain the 3D RoPE formulation for video tokens as introduced in Wan2.1 (Wan et al., 2025). Given video tokens of shape $T_v \times H \times W$, their 3D position IDs range from $(0, 0, 0)$ to $(T_v - 1, H - 1, W - 1)$. For audio tokens of shape $T_a \times M$ (extracted from mel-spectrograms as in AudioLDM2 (Liu et al., 2024b)), we first augment them with an additional leading dimension (dimension 0) to align with the video tokens along the absolute time axis. This means, audio tokens corresponding to the same time window of video tokens with temporal ID i are assigned temporal ID i as well. Subsequently, to ensure that audio and video position IDs remain strictly non-overlapping, we offset the original mel-spectrogram dimensions (T_a and M) by adding H and W, respectively. Consequently, the audio position IDs span from (H, W) to $(H + T_a - 1, W + M - 1)$. Further discussion and comparisons are provided in the Sec. C.

Formally, for an audio token at timestamp t and frequency bin m, its positional ID is defined as:

$$R_a(t, m) = \left(\left[t \cdot \frac{T_v}{T_a} \right], t + H, m + W \right) \tag{4}$$

where $[\cdot]$ denotes the round operation. This formulation explicitly enforces temporal alignment and synchrony between audio and video tokens.

Note that, thanks to the full attention design in Wan2.1, we can logically emulate the interleaved audio-video temporal arrangement shown in Fig. 3 purely through position ID manipulation, without physically reordering tokens. In contrast, within a causal (autoregressive) DiT framework, achieving temporal alignment would require physically interleaving audio and video tokens such that those with smaller temporal position IDs are placed earlier in the sequence. However, such physical reordering entails non-contiguous memory accesses and incurs additional computational overhead.

3.4 Direct Performance Optimization for Joint Audio-Video Generation

To further improve the video and audio quality and synchronization, we propose an audio-video direct preference optimization (AV-DPO) algorithm to align JAVG models with human preferences.

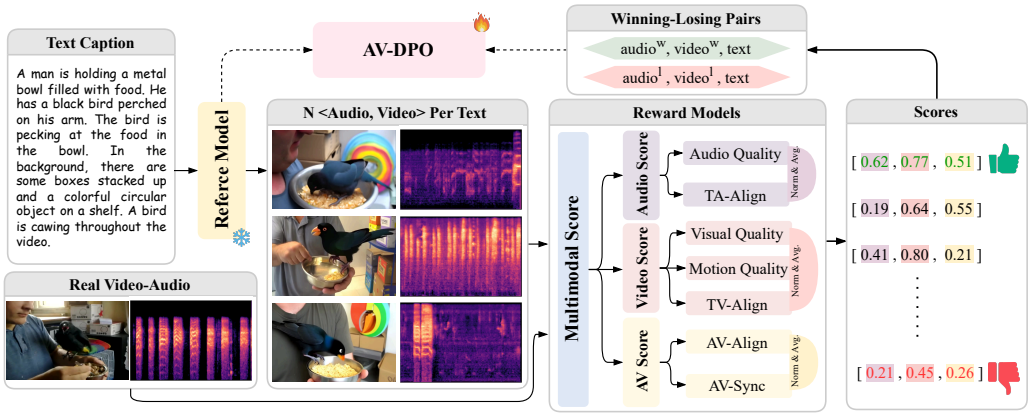


Figure 5: Illustration of preference data collection and training pipeline of audio-video DPO.

The overall pipeline is shown in Fig. 5, and our designed AV-DPO features some core contributions: (1) AV-DPO is the first one to align rectified flows with preference data for joint audio-video generation; (2) AV-DPO adopts various reward models to automatically and comprehensively evaluate the generated audio-video samples, and perform ranking based on modality-aware dimensions to ensure modality consistency in selected chosen-rejection pairs. The details are introduced as follows.

Reward Models. We employ various reward models to evaluate audio-video data from three modality-aware dimensions: audio reward, video reward, and audio-video alignment. Specifically, for audio reward, we leverage AudioBox (Tjandra et al., 2025) to comprehensively evaluate the audio quality, with ImageBind (Girdhar et al., 2023) to measure text-audio semantic similarity (alignment). As for video reward, we take VideoAlign (Liu et al., 2025b) to assess visual and motion quality, with ImageBind (Girdhar et al., 2023) to measure text-video semantic alignment. For audio-video alignment, we also utilize ImageBind to calculate audio-video semantic similarity, with Syncformer (Iashin et al., 2024) to estimate temporal synchrony. To solve the problem that different reward indicators yield various magnitudes, we first normalize the scores of each metric and average them to obtain the reward for the corresponding modality-aware dimension.

Preference Data Acquisition. To construct the preference data, we first curate a prompt pool P_t with $30k$ text captions apart from the SFT training data, and then prompt the reference model to generate $N = 3$ audio-video pairs for each prompt. To stabilize the preference optimization, we also add the ground truth audio-video pairs into the generated ones to form the candidate preference data, which are evaluated by previously mentioned reward models from three modality-specific dimensions. Afterwards, we select the winner sample a_i^w, v_i^w yielding better ranking scores than the loser sample a_i^l, v_i^l across all dimensions, forming around $25k$ audio-video preference pairs $\mathcal{D}_\sqcup = \{(a_i^w, v_i^w, a_i^l, v_i^l, y_i) \mid y_i \in P_t\}$. It is worth mentioning that winning pairs from our generated samples occupy around 30% of the total preference data, indicating that the baseline model itself already possesses fairly strong generative capabilities.

Direct Preference Optimization. Unlike prior single-modality DPO approaches (Liu et al., 2025b), our AV-DPO enhances joint audio-video generation by considering modality-aware preference:

$$\begin{cases} \text{Diff}^v_{\text{policy}} &= \| \mathbf{v}_\theta(\mathbf{x}_t^{v,w}, t) - \mathbf{v}^{v,w} \|_2^2 - \| \mathbf{v}_\theta(\mathbf{x}_t^{v,l}, t) - \mathbf{v}^{v,l} \|_2^2, \\ \text{Diff}^v_{\text{ref}} &= \| \mathbf{v}_{\text{ref}}(\mathbf{x}_t^{v,w}, t) - \mathbf{v}^{v,w} \|_2^2 - \| \mathbf{v}_{\text{ref}}(\mathbf{x}_t^{v,l}, t) - \mathbf{v}^{v,l} \|_2^2, \\ \text{Diff}^a_{\text{policy}} &= \| \mathbf{v}_\theta(\mathbf{x}_t^{a,w}, t) - \mathbf{v}^{a,w} \|_2^2 - \| \mathbf{v}_\theta(\mathbf{x}_t^{a,l}, t) - \mathbf{v}^{a,l} \|_2^2, \\ \text{Diff}^a_{\text{ref}} &= \| \mathbf{v}_{\text{ref}}(\mathbf{x}_t^{a,w}, t) - \mathbf{v}^{a,w} \|_2^2 - \| \mathbf{v}_{\text{ref}}(\mathbf{x}_t^{a,l}, t) - \mathbf{v}^{a,l} \|_2^2, \\ \mathcal{L}^{av}_{\text{DPO}} &= -\mathbb{E}_{t\sim[0,1],(\mathbf{x}_0^{a,w},\mathbf{x}_0^{v,w},\mathbf{x}_0^{a,l},\mathbf{x}_0^{v,l})\sim\mathcal{D}}\big[\log\sigma\big(-\beta_v\big(\text{Diff}^v_{\text{policy}} - \text{Diff}^v_{\text{ref}}\big) - \beta_a\big(\text{Diff}^a_{\text{policy}} - \text{Diff}^a_{\text{ref}}\big)\big)\big] \end{cases} \tag{5}$$

Using curated preference pairs, AV-DPO increases the likelihood of well-aligned (winner) samples and decreases that of misaligned (loser) ones. To prevent over-optimization, we follow (Hung et al., 2024) to incorporate the flow matching loss as a regularization term.

Table 1: Main results on JavisBench for generating 240p4s sounding videos. The best results are marked with **bold**, and the second ones are marked with underline. "†" means reproduction.

Method	AV-Quality			Text-Consistency				AV-Consistency		AV-Synchrony	
	FVD ↓	KVD ↓	FAD ↓	TV-IB ↑	TA-IB ↑	CLIP ↑	CLAP ↑	AV-IB ↑	AVHScore ↑	JavisScore ↑	DeSync ↓
- T2A+A2V											
TempoTkn	539.8	7.2	-	0.085	-	0.198	-	0.137	0.122	0.103	-
TPoS	839.7	4.7	-	0.103	-	0.238	-	0.142	0.129	0.095	-
- T2V+V2A											
ReWaS	-	-	9.4	-	0.109	-	0.236	0.111	0.104	0.079	-
See&Hear	-	-	7.6	-	0.072	-	0.248	0.164	0.143	0.112	-
FoleyC	-	-	9.1	-	**0.173**	-	0.299	0.204	**0.186**	0.151	-
MMAudio	-	-	6.1	-	0.160	-	0.407	**0.198**	0.182	0.150	-
- T2AV											
MM-Diff	2311.9	12.2	27.5	0.080	0.032	0.173	0.048	0.119	0.109	0.070	-
JavisDiT †	204.1	0.6	7.2	0.195	0.151	0.308	0.328	0.197	0.179	0.154	1.039
UniVerse-1	194.2	0.5	8.7	0.272	0.111	0.309	0.245	0.104	0.098	0.077	0.929
Ours	**141.5**	**0.4**	**5.5**	**0.282**	0.164	**0.316**	**0.424**	**0.198**	0.184	**0.159**	**0.832**

4 EXPERIMENTS

Sec. 4.1 and Sec. 4.2 present the setup and results of the main experiments, while Sec. 4.3 provides an in-depth analysis of the key modules. More ablation studies are included in Sec. D.

4.1 EXPERIMENTAL SETUP

Implementation Details. Our model is built upon Wan2.1-1.3B-T2V (Wan et al., 2025) and progressively adapted for joint audio-video generation: audio pre-training, audio-video SFT, and audio-video DPO. Rectified flow (Liu et al., 2023) is adopted as the noise scheduler for diffusion optimization. All the hyper-parameter details are provided in Sec. B.1.

Training Datasets. For audio data, we directly adopt the 780K audio-text pairs collected by JavisDiT (Liu et al., 2025c) for audio pre-training. For video data, we filtered 330K audio-video-text triplets from TAVGBench for audio-video SFT, along with an additional 25K samples for audio-video DPO. Further details and investigations are provided in Sec. B.2 and Sec. D.2.

Evaluation Benchmarks. We mainly follow JavisDiT (Liu et al., 2025c) to conduct experiments on JavisBench (10,140 samples for the main experiments) and JavisBench-mini (1,000 samples for ablation studies). The evaluation covers 11 metrics across various audio-video dimensions, including quality, consistency, and synchrony. Further details and explanations are provided in Sec. B.3.

4.2 MAIN RESULTS

As shown in Tab. 1, our model significantly outperforms previous methods in all three dimensions of audio-video generation (*i.e.*, quality, consistency, and synchrony), achieving new state-of-the-art performance. In particular, we surpass UniVerse-1 (Wang et al., 2025a) by a large margin on quality and consistency, which also adopts Wan2.1-1.3B (Wan et al., 2025) as the backbone. This improvement is attributed to the unified and efficient design of MS-MoE, rather than simply stitching two pretrained models together as in UniVerse-1. On the other hand, our model also substantially outperforms both JavisDiT and UniVerse-1 on synchrony metrics, benefiting from the explicitly designed TA-RoPE strategy that provides accurate audio-video temporal alignment. Fig. 6 further presents qualitative comparisons, where our generated results consistently surpass JavisDiT and UniVerse-1, narrowing the gap with powerful proprietary models such as Veo3 (DeepMind, 2025). In addition, Fig. 6 compares inference latency across different JAVG models, showing that our model introduces around only 18% additional inference cost over the Wan2.1 baseline (1m5s), while being significantly more efficient than two-stream methods like UniVerse-1 and JavisDiT.

4.3 IN-DEPTH ANALYSIS

In this section, we JavisBench-mini (Liu et al., 2025c) with 1,000 prompts as the test set for evaluation. To avoid redundancy, we report 7 out of the 11 full metrics of audio-video generation.

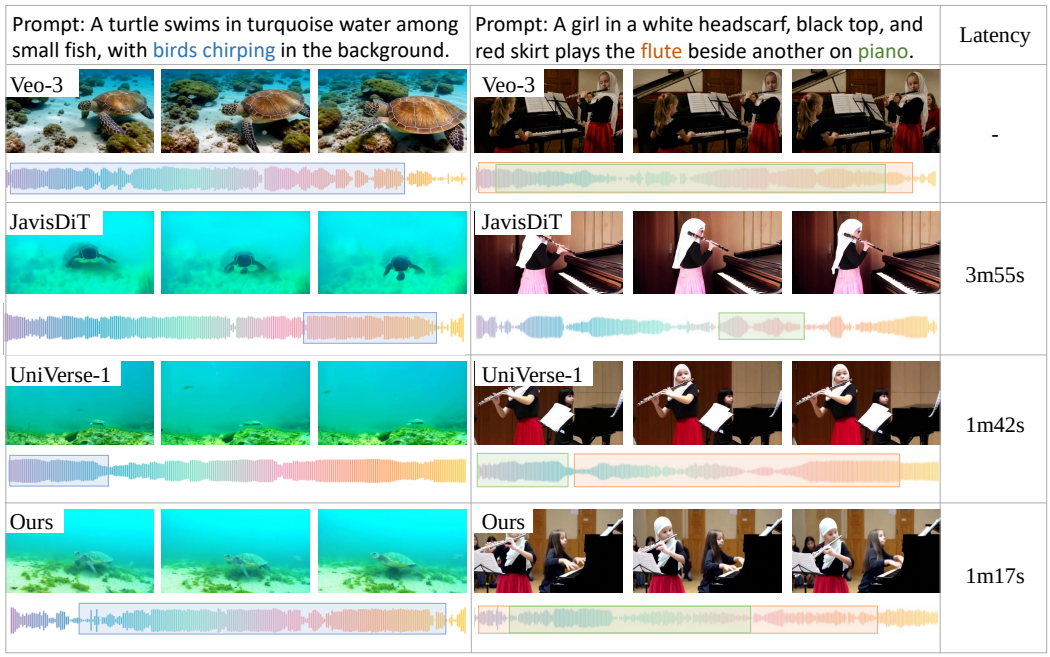


Figure 6: Generations from recent JAVG models. Best viewed in zoom or supplementary materials.

Table 2: Investigation on architectural designs to adapt Wan2.1-T2V to joint audio-video generation.

Arch Design	Quality		Consistency			Synchrony	
	FVD ↓	FAD ↓	TV-IB ↑	TA-IB ↑	AV-IB ↑	JavisScore ↑	DeSync ↓
Shared-DiT + LoRA	227.6	6.51	**0.283**	0.138	0.127	0.098	1.018
Shared-DiT + Full-FT	269.3	5.66	0.276	0.159	0.164	0.137	1.026
MS-MoE (Ours)	**221.3**	**5.51**	**0.283**	**0.163**	**0.194**	**0.153**	**0.901**

The proposed modality-specific MoE is capable for JAVG. Tab. 2 compares three different strategies for performing audio pre-training and audio-video SFT within a unified model. First, we reuse Wan2.1-T2V as the shared DiT (Zhao et al., 2025) and apply either LoRA or full-parameter finetuning for text-to-audio-video (T2AV) adaptation, which serve as two important baselines. According to Tab. 2, the LoRA scheme suffers from poor audio quality and consistency due to its limited trainable capacity; the full-finetuning scheme, on the other hand, shifts too many parameters during the audio pre-training stage, which severely degrades video quality and consistency. In contrast, our MS-MoE design preserves strong video generation ability while equipping the model with high-quality audio generation, and simultaneously maintains excellent audio-video synchrony.

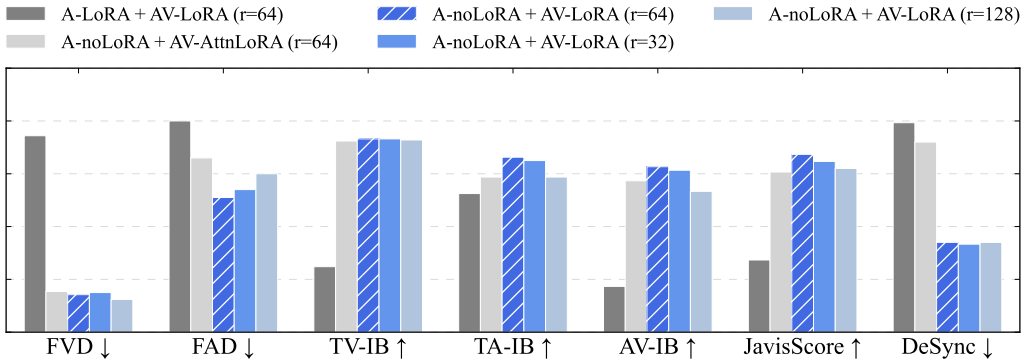


Figure 7: Comprehensive ablation studies on LoRA configurations.

Our LoRA configuration is a suitable choice. Fig. 7 illustrates the effects of different settings for T2AV adaptation, from which we can draw several conclusions: (1) Adding LoRA to the attention layers during the audio pre-training stage cannot improve audio generation quality but significantly reduces video generation performance, as it alters parameters in the video branch. (2) Compared with adding LoRA only to the attention layers, also applying LoRA to FFN leads to a notable improvement in audio-video generation performance, since the T2AV adaptation task remains relatively challenging. (3) The model performance is not highly sensitive to the choice of LoRA rank and alpha parameters; empirically, setting rank = 64 brings slightly better performance.

The proposed TA-RoPE synchronization mechanism is effective and efficient. Tab. 3 compares several designs for audio-video generation synchrony. First, although both the ST-Prior proposed in JavisDiT (Liu et al., 2025c) and the frame-level cross-attention used in UniVerse-1 (Wang et al., 2025a) improve synchrony, they bring an unaffordable increase in inference latency. In contrast, our TA-RoPE strategy achieves better performance with zero additional inference cost. On the other hand, TA-RoPE can also be combined with ST-Prior or FrameAttn to yield slightly better synchrony, but we ultimately discard these combinations to maintain inference efficiency and overall simplicity.

Table 3: Comparison of synchronization mechanisms.

Mechanism	JavisScore ↑	DeSync ↓	Latency ↓
None	0.142	1.044	**1m17s**
ST-Prior	0.145	0.973	1m33s
FrameAttn	0.124	0.951	1m45s
TA-RoPE (Ours)	**0.153**	**0.901**	**1m17s**
TA-RoPE + ST-Prior	0.155	0.953	1m33s
TA-RoPE + FrameAttn	0.151	0.890	1m45s

The AV-DPO reward strategy is reasonable and effective. Tab. 4 compares different reward strategies for selecting win–lose pairs in DPO training. First, applying modality-agnostic micro averaging (averaging across all metrics before ranking (Liu et al., 2025b)) or macro averaging (ranking within each metric and then averaging (Xue et al., 2025)) fails to achieve consistent improvements in audio-video generation. This is because such strategies may form a win sample by combining better video but worse audio, which conflicts with eq. (5). In contrast, calculating rewards separately for audio and video and ensuring modality-consistent chosen samples effectively improves generation quality, consistency, and synchrony. Meanwhile, removing normalization for single-dimension rewards (*i.e.*, w/o norm) reduces the accuracy of pair ranking due to scale and range differences across rewards, which in turn degrades DPO performance. Likewise, discarding ground-truth samples and forming pairs only from generated candidates (*i.e.*, w/o gt) even gets worse results, as differences among generated samples are often too small to guide preference shifts.

Table 4: Investigation on the effectiveness of different AV-DPO reward strategies.

Reward Design	**Quality**		**Consistency**			**Synchrony**	
	FVD ↓	FAD ↓	TV-IB ↑	TA-IB ↑	AV-IB ↑	JavisScore ↑	DeSync ↓
None (baseline)	221.3	5.51	0.283	0.163	0.194	0.153	0.901
Average-Micro	199.7	5.28	0.281	0.166	0.199	0.154	0.902
Average-Macro	203.3	5.31	0.281	0.166	0.196	0.152	0.903
Modality-Micro	**198.5**	**5.32**	**0.284**	**0.168**	**0.201**	**0.156**	0.888
Modality-Macro	191.1	5.41	0.282	0.166	0.197	**0.156**	**0.882**
Modality-Micro (w/o norm)	210.0	5.34	0.281	0.167	0.197	0.153	0.925
Modality-Micro (w/o gt)	234.7	5.43	0.281	0.164	0.197	0.154	0.937

5 Conclusion

In this work, we presented a concise and efficient framework for native joint audio-video generation. By introducing the MS-MoE design for modality-specific quality enhancement, the TA-RoPE strategy for explicit temporal alignment, and the AV-DPO algorithm for preference alignment, our model achieves state-of-the-art performance in quality, consistency, and synchrony. Built upon Wan2.1-1.3B-T2V and trained with only 1M data entries, our model significantly outperforms existing open-source approaches while maintaining efficiency. We believe this work establishes an important milestone for JAVG and opens new directions for future research in the field.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

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.

Huakang Chen, Yuepeng Jiang, Guobin Ma, Chunbo Hao, Shuai Wang, Jixun Yao, Ziqian Ning, Meng Meng, Jian Luan, and Lei Xie. Diffrhythm+: Controllable and flexible full-length song generation with preference optimization. *arXiv preprint arXiv:2507.12890*, 2025.

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.

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

Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, and Jiaming Song. Personalized preference fine-tuning of diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8020–8030, 2025.

Google DeepMind. Veo 3: Tech report, 2025. URL `https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf`.

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

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.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and Sherry Yang. Improving dynamic object interactions in text-to-video generation with ai feedback. *arXiv preprint arXiv:2412.02617*, 2024.

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a.

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*, 2025b.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*, 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.

Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. Ace-step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*, 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.

Akio Hayakawa, Masato Ishii, Takashi Shibuya, and Yuki Mitsufuji. Mmdisco: Multi-modal discriminator-guided cooperative diffusion for joint audio and video generation. *arXiv preprint arXiv:2405.17842*, 2024.

Qihan Huang, Long Chan, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, and Jie Song. Patchdpo: Patch-level dpo for finetuning-free personalized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18369–18378, 2025.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024.

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, 2023a.

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, 2023b.

Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. *arXiv preprint arXiv:2407.05551*, 2024.

Yuxuan Jiang, Zehua Chen, Zeqian Ju, Chang Li, Weibei Dou, and Jun Zhu. Freeaudio: Training-free timing planning for controllable long-form text-to-audio generation. *arXiv preprint arXiv:2507.08557*, 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.

Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *European Conference on Computer Vision*, pp. 34–50. Springer, 2022.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13401–13412, 2021.

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.

Haohe Liu, Gael Le Lan, Xinhao Mei, Zhaoheng Ni, Anurag Kumar, Varun Nagaraja, Wenwu Wang, Mark D Plumbley, Yangyang Shi, and Vikas Chandra. Syncflow: Toward temporally aligned joint audio-video generation from text. *arXiv preprint arXiv:2412.15220*, 2024a.

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*, 2024b.

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

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

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*, 2025c.

Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8009–8019, 2025d.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

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*, 2024.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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.

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

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

Mingzhen Sun, Weining Wang, Yanyuan Qiao, Jiahui Sun, Zihan Qin, Longteng Guo, Xinxin Zhu, and Jing Liu. Mm-ldm: Multi-modal latent diffusion model for sounding video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10853–10861, 2024.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27425–27434, 2024.

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 preprint arXiv:2503.10522*, 2025.

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

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

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, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025a.

Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025b.

Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation. *arXiv preprint arXiv:2406.07686*, 2024.

Le Wang, Jun Wang, Chunyu Qiang, Feng Deng, Chen Zhang, Di Zhang, and Kun Gai. Audiogen-omni: A unified multimodal diffusion transformer for video-synchronized audio, speech, and song generation. *arXiv preprint arXiv:2508.00733*, 2025c.

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

Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025.

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 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, 2024a.

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, 2024b.

Shihao Yuan, Yahui Liu, Yang Yue, Jingyuan Zhang, Wangmeng Zuo, Qi Wang, Fuzheng Zhang, and Guorui Zhou. Ar-grpo: Training autoregressive image generation models via reinforcement learning. *arXiv preprint arXiv:2508.06924*, 2025.

Hui Zhang, Zuxuan Wu, Zhen Xing, Jie Shao, and Yu-Gang Jiang. Adadiff: Adaptive step selection for fast diffusion. *arXiv preprint arXiv:2311.14768*, 2023.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024.

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, March 2024. URL `https://github.com/hpcaitech/Open-Sora`.

A Discussion

A.1 Potential Limitations

While our framework demonstrates state-of-the-art performance in joint audio-video generation, several limitations remain and open promising directions for future work:

- **Training Data Scale**. Our model is trained on roughly 1M entries, which, although efficient, may constrain scalability compared with larger proprietary systems. Expanding the dataset with more diverse and high-quality audio-video pairs could further improve generalization and robustness.
- **Model Size**. We adopt a 1.3B-parameter backbone with parameter-efficient adaptations. Scaling up to larger backbones may unlock stronger representational capacity, especially for capturing subtle temporal and semantic correlations across modalities.
- **Full-Parameter Training**. Our approach relies on parameter-efficient tuning (e.g., LoRA). Exploring full-parameter finetuning could provide additional performance gains, albeit with higher computational cost.
- **Controllable Generation**. Current experiments focus on general text-to-audio-video generation. Extending controllability to domains such as music or speech, with fine-grained control over rhythm, pitch, timbre, or lexical content, is an important next step.
- **Unified Cross-Modal Generation**. Beyond text-conditioned JAVG, broader tasks such as audio-to-video (A2V), video-to-audio (V2A), and audio-image-to-video (AI2V) offer opportunities for a unified multimodal generative framework. Developing models that seamlessly perform across these modalities would mark a significant milestone toward general-purpose audio-visual content generation.

We hope these directions inspire future research toward building more scalable, controllable, and unified multimodal generative systems.

A.2 Ethics Statement

All datasets and models used in this work are publicly available on the internet and do not involve any private or sensitive information. In addition, part of the DPO data we plan to release is generated by models themselves, ensuring that no personal privacy is infringed.

A.3 Reproducibility Statement

We provide detailed descriptions of model design, training, and evaluation in both the main paper and the appendix. Furthermore, all code, pretrained checkpoints, and processed datasets will be publicly released to ensure full reproducibility of our results.

A.4 LLM Usage Statement

Large Language Models (LLMs) were used solely as writing assistants, including tasks such as language polishing and presentation refinement. They were not involved in the conception of core ideas or designs.

B Detailed Implementations

B.1 Model Details

Audio VAE. We reuse and freeze the audio encoder and decoder from AudioLDM2 (Liu et al., 2024b). All 1D audio signals are resampled to 16 kHz and converted into 64-bin mel-spectrograms using a window size of 64 ms and a hop size of 10 ms. The spectrograms are then 8×8 compressed via VAE into 8-channel audio embeddings. To further reduce the token count, we apply a 2×2 patchify operation before feeding the tokens into the DiT for diffusion and denoising.

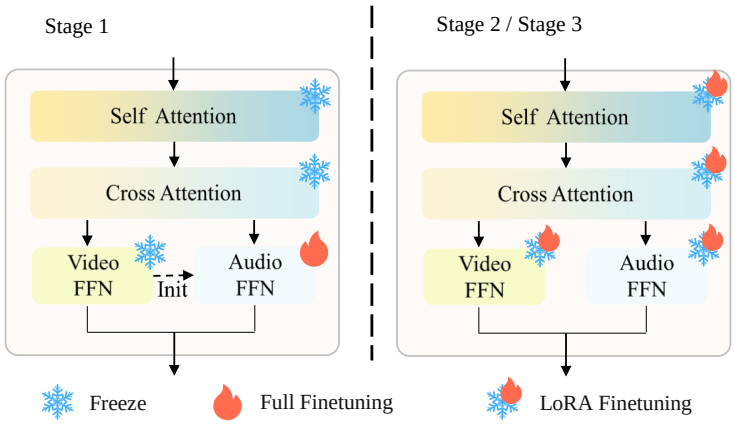


Figure A1: Illustration of trainable parameters at different stages.

Table A1: Detailed settings for the three-stage training pipeline.

Setting	Stage 1	Stage 2	Stage 3
training purpose	Audio PreTrain	Audio-Video SFT	Audio-Video DPO
trainable modules	Audio FFN/Embedder/Head	LoRA	LoRA
trainable params	794M	121M	121M
learning rate	1e-4	1e-4	1e-5
warm-up steps	1000	1000	100
weight decay	0.0	0.0	0.0
training samples	780K	330K	25K
resolution	-	dynamic	dynamic
duration	dynamic	dynamic	dynamic
batch size	dynamic	dynamic	dynamic
epoch	50	2	1
GPU days (H100)	16	16	3

Video VAE. We reuse and freeze the video VAE from Wan2.1 (Wan et al., 2025). Except for the first frame, all subsequent video frames are temporally compressed by a factor of 4, while every frame is spatially compressed by 8×8, resulting in 16-channel video embeddings. To ensure further compactness, these video embeddings are also compressed with a 2×2 spatial patchify operation before being fed into the DiT.

Text Encoder. We also reuse and freeze Wan2.1's umT5-xxl (Chung et al., 2023) as the text-encoder, whose context length is 512.

Backbone DiT. Our model is built on the powerful Wan2.1-1.3B-T2V (Wan et al., 2025) model, and progressively extends from text-to-video generation to text-conditioned joint audio-video generation. The Wan2.1-1.3B base model has 30 layers with a hidden dimension of 1536. We keep the original parameters frozen throughout the entire training process, updating only the newly introduced audio FFN (along with audio embedder, audio head, etc) and the LoRA components at different stages. The final model has only 2.1B parameters after merging LoRA components. Detailed training settings are presented as follows.

B.2 Training Details

Details of the three-stage training pipeline. Our model progressively extends from Wan2.1-1.3B-T2V (Wan et al., 2025) to JAVG through three stages: (1) Audio Pre-Training, where the Audio FFN is trained on 780K audio-text pairs for 50 epochs with a learning rate of 1e-4; (2) Audio-Video SFT, where LoRA is applied to train on 330K audio-video-text triplets for 1 epochs with a learning rate of 1e-4; and (3) Audio-Video DPO, where the LoRA parameters are retained and further trained on

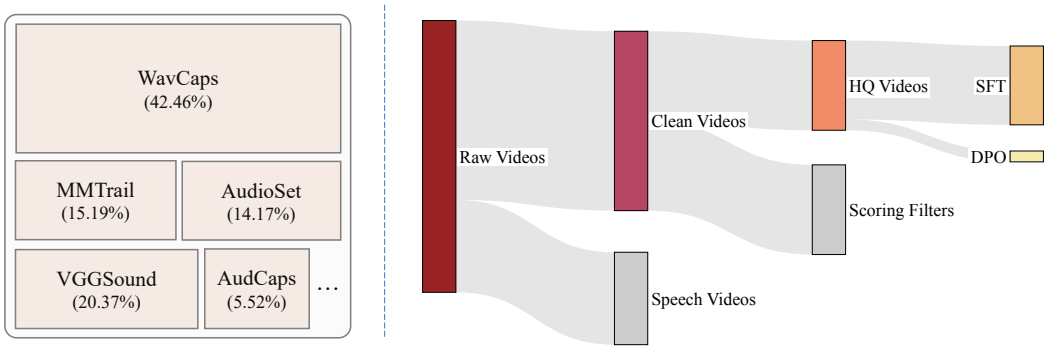


Figure A2: **(Left)**: Diversified audio-text sources. **(Right)**: Data filtering process from TAVGBench.

25K entries for 1 epoch with a learning rate of 1e-5 and β of 500. The specific trainable modules and other training settings can be found in Fig. A1 and Tab. A1.

Details of the audio-video training data. Fig. A2 demonstrates the detailed data composition and filtering procedure. For audio data, we directly adopt 780K audio-text pairs from JavisDiT (Liu et al., 2025c) for audio pretraining, which covers various public datasets, including 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). We apply no data filtering strategy to ensure maximal text-to-audio generation capability spanning general sound, music, and speech.

For video data, we adopt a subset of 1.1 million text-video-audio triplets from TAVGBench (Mao et al., 2024) and conduct a series of filtering strategies. We first eliminate a large part of videos containing human speech by using the FunASR (Gao et al., 2023) detection tool, and then follow OpenSora (Zheng et al., 2024) to filter out the videos with relatively lower quality through aesthetic scoring (Schuhmann, 2022), flow (motion) scoring (Xu et al., 2023), and OCR scoring (Liao et al., 2020). In the final filtered pool, 330K data entries are divided for audio-video SFT, and the other 25K samples are used for audio-video DPO to avoid overlapping. We also provide empirical evaluations on the data diversity and quality in Sec. D.2.

B.3 Evaluation Details

Evaluation Setup. We mainly follow JavisDiT (Liu et al., 2025c) to evaluate the models on JavisBench, which consists of 10,140 prompts for joint audio-video generation in various real-world scenarios. In ablation studies, we take the JavisBench-mini for fast evaluation, where the 1,000 prompts are randomly selected from the total 10,140 samples from JavisBench. All models are required to generate 240P, 4-second sounding videos for quantitative evaluation.

JavisBench provides a comprehensive evaluation of quality, consistency, and synchrony for audio-video generation results. Here, we briefly introduce the mechanisms of each evaluation dimension:

- **Audio / Video Quality**: measuring the perceptual quality of the generated audio and video, including (1) *Fréchet Video Distance (FVD)*: $\mathrm{FVD} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$, where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of ground-truth and generated video features extracted by a pretrained I3D encoder (Carreira & Zisserman, 2017). Lower is better, indicating the generated video distribution is closer to the real one; (2) *Kernel Video Distance (KVD)*: similar to FVD, but estimates distribution differences via a kernel-based method (Kernel Inception Distance style), which is more stable on smaller datasets; lower is better; and (3) *Fréchet Audio Distance (FAD)*: same concept as FVD, but computed on audio features extracted by a pretrained AudioClip model (Guzhov et al., 2022), measuring distribution distance between generated and real audio; lower is better.
- **Text Consistency**: evaluating how well the generated audio and video semantically match the input text description, including (1) *ImageBind (Girdhar et al., 2023) text-video cosine similarity*: $\mathrm{sim}(t, v) = \frac{f_{\text{text}}(t) \cdot f_{\text{video}}(v)}{\|f_{\text{text}}(t)\| \cdot \|f_{\text{video}}(v)\|}$; (2) *ImageBind text-audio cosine similarity:* same pro-

cess but with the audio encoder f_{audio}; (3) *CLIP-Score*: using CLIP (Radford et al., 2021) to compute semantic similarity between text and video (video frames are sampled, encoded, and averaged); and (4) *CLAP-Score*: using CLAP (Wu* et al., 2023) to compute semantic similarity between text and audio.

- **Audio–Video Semantic Consistency**: measuring the semantic alignment between generated audio and generated video, including (1) *ImageBind audio-video cosine similarity*, encoding both modalities into the same space and computing cosine similarity between video and audio features; and (2) *Audio-Visual Harmony Score (AVHScore)*: introduced in TAVGBench (Mao et al., 2024) as a way to quantify how well the generated audio and video align semantically in a shared embedding space. It is defined by computing the cosine similarity between each video frame and the entire audio, then averaging across all frames: $\text{AVHScore} = \frac{1}{N}\sum_{i=1}^{N}\cos\big(f_{\text{frame}}(v_i),\ f_{\text{audio}}(a)\big)$. A higher AVHScore indicates stronger audio–video semantic consistency. Note that we remove the CAVP-Score (Luo et al., 2023) used in JavisDiT (Liu et al., 2025c) because this metric keeps a range from 0.798 to 0.801 and cannot capture the difference when evaluating semantic consistency.
- **Audio–Video Spatio-Temporal Synchrony**: evaluating spatiotemporal alignment in generated audio-video pairs, including (1) *JavisScore*: a new metric proposed in JavisDiT (Liu et al., 2025c). The core idea is to use a sliding window along the temporal axis to split the audio-video pair into short segments. For each segment, compute cross-modal similarity with ImageBind and take the mean score: $\text{JavisScore} = \frac{1}{N}\sum_{i=1}^{N}\sigma(a_i, v_i), \quad \sigma(v_i, a_i) = \frac{1}{k}\sum_{j=1}^{k}\min\text{-}k\{\cos\left(E_v(v_{i,j}), E_a(a_i)\right)\}$; and (2) *DeSync*: a metric adapted from Synchformer(Iashin et al., 2024), which measures fine-grained temporal misalignment between audio and video streams. Specifically, it estimates the temporal offset of audio–visual events by performing a 21-category classification task (predicting the offset/asynchrony degree ranging from -10 to 10 and taking the absolute values). A lower DeSync score indicates better synchronization.

Compared Methods. Since open-source models that support text-conditioned joint audio-video (JAVG) generation are still very limited, we also follow JavisDiT (Liu et al., 2025c) and include cascaded generation pipelines (*i.e.*, T2A+A2V (Yariv et al., 2024a; Jeong et al., 2023b) and T2V+V2A (Jeong et al., 2024; Xing et al., 2024; Zhang et al., 2024; Cheng et al., 2025)) for comparison. For JAVG-capable models, we treat the unconditional MMDiffusion (Ruan et al., 2023) as a simple baseline, while focusing on comparisons against text-conditional models like JavisDiT (Liu et al., 2025c) and UniVerse-1 (Wang et al., 2025a). For JavisDiT, we directly downloaded its released checkpoints and performed JavisBench generation locally, enabling us to compute the newly introduced DeSync (Iashin et al., 2024) metric and update other metrics consistently. For UniVerse-1, since it essentially supports audio-video generation from Text + Reference Image, we use the first frame of our model's generated video as the reference image, allowing UniVerse-1 to perform audio-synchronized image animation.

Details of Fig. 2. In Sec. 1, we present a radar chart (Fig. 2) to illustrate the performance differences among various JAVG models. Specifically, we randomly sample 100 prompts from JavisBench (Liu et al., 2025c) for audio-video generation. For open-source models, we run inference locally, while for Veo3 (DeepMind, 2025), we obtain results via its API. After collecting outputs from all models, we follow the DPO reward setup in Sec. 3.4 to comprehensively evaluate the performance, including (1) VideoAlign (Liu et al., 2025b) and AudioBox (Tjandra et al., 2025) to assess video and audio quality; (2) ImageBind (Girdhar et al., 2023) to compute semantic similarity across TV-Align, TA-Align, and AV-Align; and (3) SynchFormer (Iashin et al., 2024) to compute DeSync as the synchrony metric. Since DeSync is defined as a "lower-is-better" score, whereas all other metrics are "higher-is-better", we invert DeSync values in Fig. 2 to ensure visual consistency.

C Detailed Comparison with Related Works

Architectural Difference with Recent JAVG Models. Fig. A3 presents a comparison between our model design and recent JAVG approaches. First, UniForm (Zhao et al., 2025) attempts to use a single set of attention and FFN parameters to process both audio and video tokens. This poses a significant challenge when extending a pretrained T2V model with audio generation capability

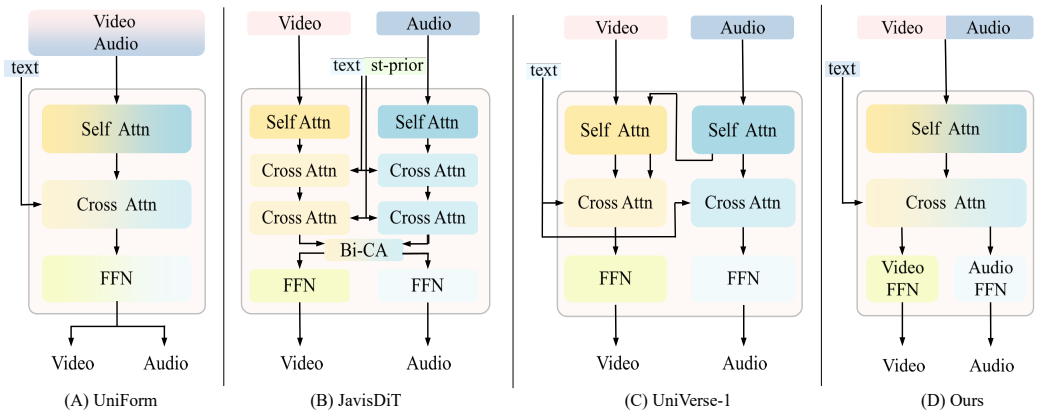


Figure A3: Architectural comparison with Uniform, JavisDiT, UniVerse-1.

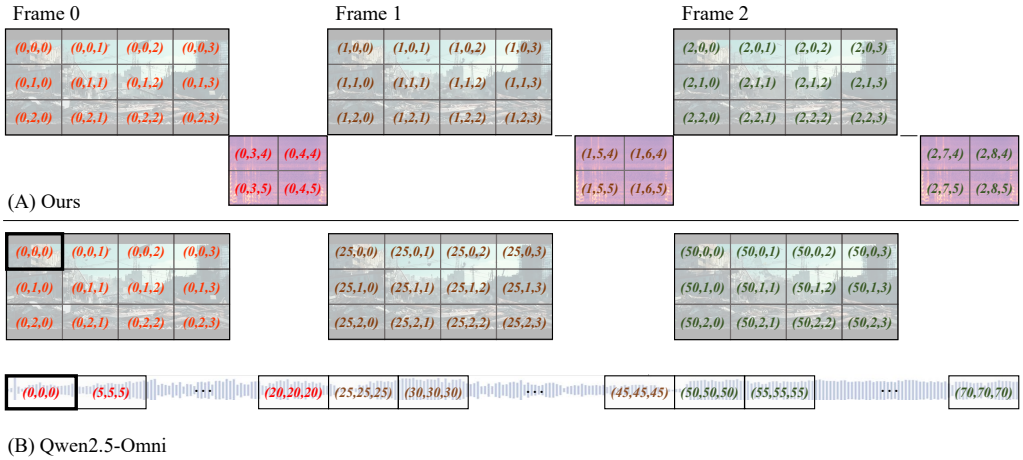


Figure A4: Comparison of our audio-video frame interleaving with Qwen2.5-Omni's strategy.

while preserving its original video generation performance, as validated in Tab. 2. To address this, JavisDiT (Liu et al., 2025c) introduces a dual-stream architecture with separate parameter sets for video and audio generation, along with ST-Prior and frame-level audio-video bidirectional cross-attention to enhance synchrony. However, this design introduces a large number of additional parameters, making training more difficult and inference more expensive. UniVerse-1 (Wang et al., 2025a), on the other hand, employs a pretrained T2V model (Wan et al., 2025) and a pretrained T2A model (Gong et al., 2025), with a complex cross-attention mechanism to enable information exchange between audio and video, which still suffers from inefficiencies in both training and inference stages.

In contrast to these prior works, our proposed framework is simpler and more effective: it uses shared attention to enable cross-modal interaction between audio and video tokens, while adopting modality-specific FFNs to enhance intra-modal modeling quality. This design achieves a better trade-off between efficiency and performance, and provides higher scalability.

Design Difference with Qwen2.5-Omni's RoPE Strategy. Considering audio-video alignment, our TA-RoPE design shares some similarity with Qwen2.5-Omni Xu et al. (2025)'s RoPE solution: both approaches align audio and video through the position IDs along a specific dimension to ensure temporal alignment. However, as shown in Fig. A4, there exists a key difference between us: whether the position IDs of audio and video tokens overlap. Qwen2.5-Omni treats audio, like language, as a 1D token sequence and assigns the same position IDs across the three RoPE dimensions. This inevitably introduces overlaps between audio and video, *e.g.*, an audio token with position ID $(0, 0, 0)$ coinciding with a video token $(0, 0, 0)$, or potentially $(25, 25, 25)$, $(50, 50, 50)$, and so on. For multimodal understanding models such as Qwen2.5-Omni, such overlaps have negligible influ-

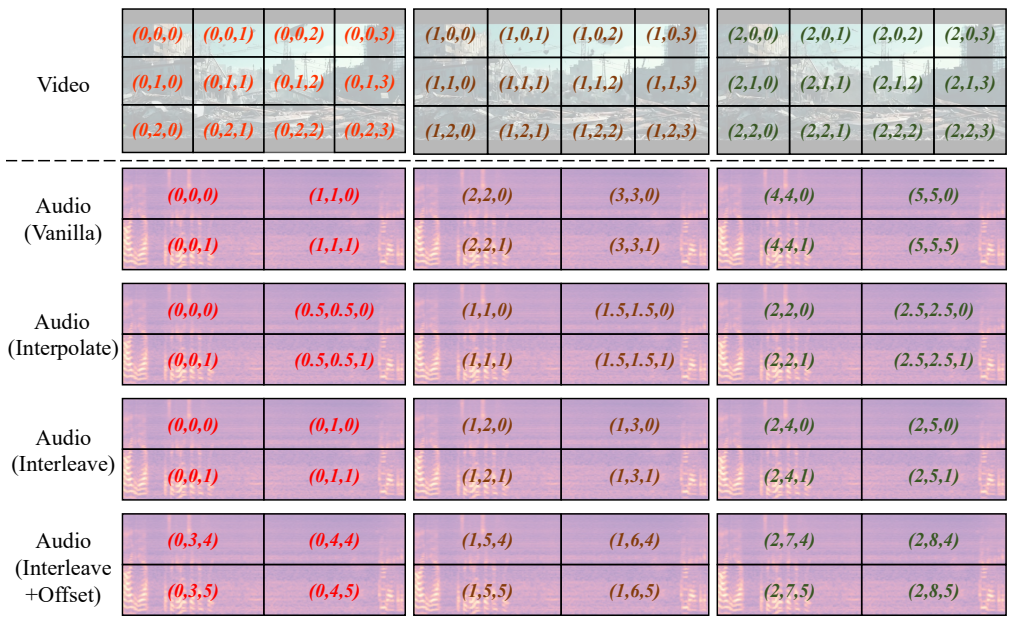


Figure A5: Illustration of different audio positional encoding strategies.

Table A2: Ablation study on positional encoding strategy for both audio and video generation.

Strategy	AudioCaps			JavisBench-mini (audio)			JavisBench-mini (video)					
	FAD ↓	TA-IB ↑	CLAP ↑	FAD ↓	TA-IB ↑	CLAP ↑	FVD ↓	TV-IB ↑	CLIP ↑	AV-IB ↑	JavisScore ↑	DeSync ↓
AudioLDM2	5.06	**0.200**	**0.460**	8.81	0.153	0.300	-	-	-	-	-	-
JavisDiT-audio	6.58	0.162	0.306	8.11	**0.182**	0.381	-	-	-	-	-	-
Vanilla	6.03	0.193	0.411	6.41	0.157	0.417	238.2	0.281	0.316	0.184	0.142	1.044
Interpolate	12.87	0.155	0.325	10.73	0.149	0.349	239.9	0.282	**0.320**	0.183	0.144	1.011
Interleave	5.49	0.195	0.420	6.20	0.153	**0.418**	225.8	**0.284**	0.318	0.187	0.144	0.926
Interleave+Offset	**4.65**	0.198	0.420	**6.81**	0.154	0.417	**221.3**	0.283	0.317	**0.200**	**0.153**	**0.901**

ence on inference performance. However, for generative tasks like JAVG, overlaps cause non-trivial position confusion and lead to performance degradation, as quantitatively demonstrated in Sec. D.1.

By comparison, our TA-RoPE design treats audio as a 2D image (represented by the form of mel-spectrogram (Liu et al., 2024b)). When ensuring temporal alignment by matching the 0-th dimension of position IDs between audio and video along the time axis, we introduce adaptive offsets in the other two dimensions of the mel-spectrogram corresponding to video width and height (as formulated in eq. (4)). This strategy completely avoids overlap between the two modalities and enables more accurate audio-video synchrony.

About AudioGen-Omni's RoPE Strategy. Focusing on the domain of video-to-audio generation, AudioGen-Omni Wang et al. (2025c) also identifies RoPE as a key component for achieving temporal alignment between audio and video, and proposes a specific design. However, due to the lack of detailed descriptions in their paper and the unavailability of their code, we are currently unable to conduct a thorough comparative analysis.

D Additional Experiments

D.1 Ablation on Positional Encoding Strategy

This section systematically investigates the impact of different audio position encoding strategies (illustrated in Fig. A5) on final audio-video generation quality, including:

Table A3: Ablation study on training data composition in the AV-SFT stage.

Data	Quality	Epoch	Quality		Consistency			Synchrony	
			FVD ↓	FAD ↓	TV-IB ↑	TA-IB ↑	AV-IB ↑	JavisScore ↑	DeSync ↓
120K	High	1.0	230.6	6.24	0.282	0.154	0.183	0.144	0.936
120K	High	2.0	239.9	5.81	**0.283**	0.157	0.189	0.151	0.905
120K	High	3.0	233.7	5.64	0.281	0.159	0.188	0.148	0.911
330K	Medium	1.0	225.5	5.62	**0.283**	0.161	0.190	0.145	0.912
330K	Medium	2.0	221.3	5.51	**0.283**	**0.163**	**0.194**	**0.153**	**0.901**
720K	Low	0.5	229.8	5.53	0.281	0.154	0.182	0.142	0.929
720K	Low	1.0	217.8	5.64	0.280	0.159	0.185	0.145	0.919
+120K	Low+High	2.0	223.8	5.46	0.282	0.161	0.191	0.150	0.918
+330K	Low+Med	1.0	**212.0**	**5.45**	**0.283**	0.160	0.187	0.146	0.918

- **Vanilla**, ignores video positions entirely, encoding audio purely along the time and frequency axes of its mel-spectrogram: $R_a(t, m) = (t, t, m)$
- **Interpolate**, aligns the audio temporal axis with video frames by interpolating intermediate IDs: $R_a(t, m) = \left(t \cdot \frac{T_v}{T_a}, t \cdot \frac{T_v}{T_a}, m\right)$
- **Interleave**, aligns the audio temporal axis (0-th dimension) with the video temporal axis, while unfolding the other two dimensions along the mel-spectrogram: $R_a(t, m) = \left(\left[t \cdot \frac{T_v}{T_a}\right], t, m\right)$
- **Interleave+Offset**, similarly aligns the audio 0-th dimension with the video temporal axis, but shifts the remaining two dimensions by the video width and height to fully avoid overlapping position IDs between modalities: $R_a(t, m) = \left(\left[t \cdot \frac{T_v}{T_a}\right], t + H, m + W\right)$ (eq. (4)).

We conduct full audio-pretraining and AV-SFT training on Wan2.1-1.3B-T2V (Wan et al., 2025) to examine how these strategies affect both audio and video generation. Results in Tab. A2 lead to three main conclusions:

1. Preserving integer audio position IDs is essential for audio quality — e.g., Interpolate performs significantly worse, since frozen attention layers during audio pretraining cannot learn relative offsets represented by fractional IDs;
2. The more overlap exists between audio and video position IDs, the poorer the video quality becomes (Vanilla → Interleave → Interleave+Offset), consistent with our analysis in Sec. C;
3. Generation models require non-overlapping position IDs to disentangle modalities, as well as temporal alignment along one axis for audio-video synchrony (*e.g.*, Vanilla yields much worse synchrony than Interleave-based designs).

Based on these findings, we adopt Interleave+Offset as our final position encoding scheme.

D.2 Investigation on Training Data Quality and Diversity

In this section, we conduct an in-depth investigation into the impact of diversity and quality of training data during the Audio-Video SFT stage. Specifically, we first construct a low-quality dataset of 720K text–audio–video triplets by extracting a large portion of speech videos from the full data pool (see Sec. B.2). Next, we filter out low-quality data using multiple scoring metrics such as aesthetics, motion, and OCR, resulting in a medium-quality dataset of 330K samples, which is the set adopted for AV-SFT in the main paper. Finally, we further raise the thresholds of the scoring filters, obtaining a high-quality dataset of 120K samples for comparative analysis. After audio pretraining, we conduct experiments with different data compositions in the AV-SFT stage, and the results are presented in Tab. A3. Accordingly, several conclusions can be drawn:

First, data with high quality but low diversity cannot bring optimal performance. Models trained solely on the 120K high-quality dataset underperform those trained on the 330K medium-quality dataset. This is because transitioning from unimodal audio/video generation to joint generation is

a relatively challenging task, requiring sufficient data quantity or diversity to enable the model to acquire new capabilities.

Second, data quality is also indispensable. Models trained on the 720K low-quality dataset show clearly inferior generation quality compared to those trained on the 330K medium-quality dataset. This degradation occurs because low-quality data undermine the priors learned by the Wan2.1 (Wan et al., 2025) backbone during pretraining on high-quality data. Even introducing a second round of SFT with 120K high-quality or 330K medium-quality data cannot fully recover the lost performance.

In summary, we adopt the 330K medium-quality dataset for AV-SFT training as a reasonable trade-off. We believe that further enhancing both the diversity and quality of training data will yield better scaling properties.

D.3 Ablation on Hyper-Parameters of AV-DPO

In this section, we further investigate the stability of hyperparameters in the AV-DPO algorithm proposed in eq. (5), focusing primarily on two factors: the choice of β and the learning rate. Since different parameters influence the magnitude of the loss unevenly, we adopt *implicit accuracy* as a unified proxy metric to evaluate the impact of various hyperparameter settings:

$$Acc^a = \frac{1}{N}\sum_{i=1}^{n}\mathbb{I}(\text{Diff}^a_{\text{policy}} < \text{Diff}^a_{\text{ref}}); \quad Acc^v = \frac{1}{N}\sum_{i=1}^{n}\mathbb{I}(\text{Diff}^v_{\text{policy}} < \text{Diff}^v_{\text{ref}}) \tag{A1}$$

Recalling eq. (5), the implicit accuracy can measure whether the model successfully shifts toward the distribution of the chosen data while moving away from that of the rejected data. The experimental results, shown in Fig. A6 and Fig. A7, lead to the following conclusions:

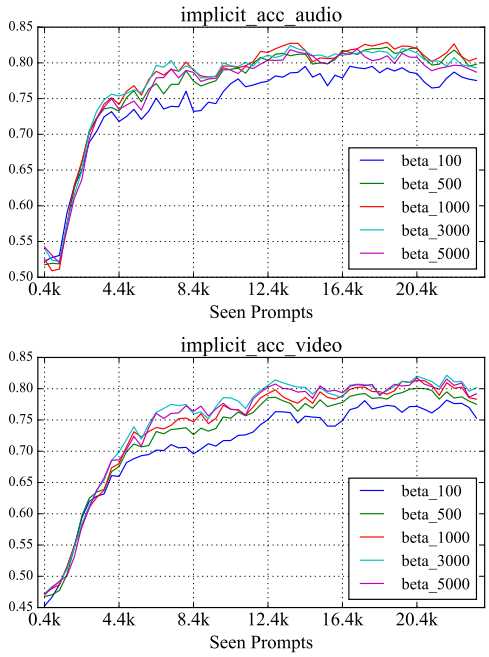


Figure A6: Implicit accuracy on β selections.

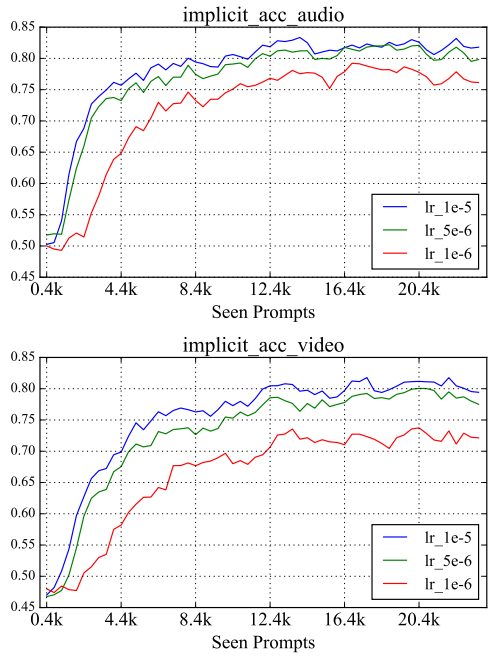


Figure A7: Implicit accuracy on learning rates.

Analysis on β selection. In the original DPO algorithm, β primarily controls the divergence between the policy model and the reference model (Rafailov et al., 2023; Liu et al., 2025b). As shown in Fig. A6, audio achieves its best performance at $\beta = 1000$, whereas video converges faster and reaches higher final accuracy at $\beta = 3000$. This observation is closely related to our proposed framework. Our model is built upon the pretrained Wan2.1-T2V, which already aligns well with human preferences; thus, a larger β (*e.g.*, 3000 or 5000) is needed to keep the video policy model closer to the reference model. In contrast, the audio branch is newly trained and initially less aligned

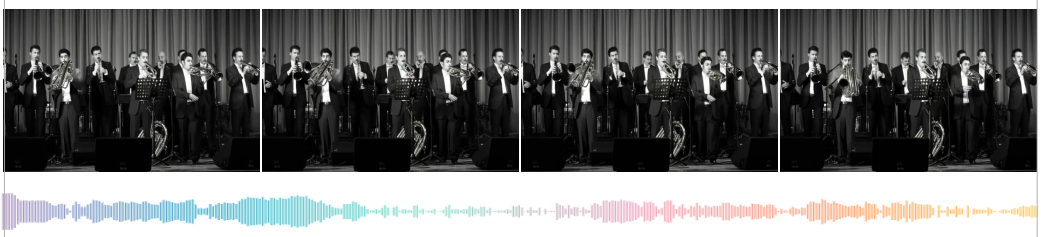
with human preferences, so a smaller β (*e.g.*, 1000) is required to shift the model closer to the preferred data distribution. However, setting β too small (*e.g.*, 100) leads to excessive divergence from the reference model and overfitting to imperfect preference data. Based on these findings, we set $\beta = 3000$ for the audio DPO loss and $\beta = 1000$ for the video DPO loss to achieve relatively better performance.

Analysis on learning rate. Fig. A7 presents three experimental groups with learning rates set to 1×10^{-5}, 5×10^{-6}, and 1×10^{-6}, respectively. The results show that 1×10^{-5} achieves both the fastest convergence and the highest final accuracy. This observation is consistent with prior works (Liu et al., 2025b), which suggests that setting the learning rate in the DPO stage to approximately one-tenth of that in the SFT stage is a suitable choice. Therefore, we adopt a learning rate of 1×10^{-5} for training.

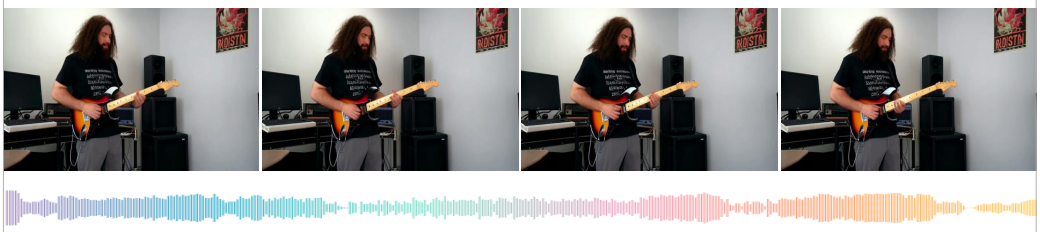
D.4 More Visualizations

Fig. A8 and Fig. A9 showcase additional cases of joint audio-video generation, illustrating the strong generative capability of our proposed model across multiple dimensions.

Prompt: A small wooden cabin stands in a rainy forest, with misty trees around it and the sound of heavy rain and distant thunder.



Prompt: A black-and-white image shows suited musicians on stage playing saxophones, trumpets, and a tuba, music filling the air before a curtain backdrop.



Prompt: A man with long curly hair and a beard plays an electric guitar in a studio, wearing a black t-shirt and gray pants. Behind him are a monitor, speakers, and a poster, as the sound of the guitar fills the room.

Prompt: A group of sharks swim gracefully underwater, their fins and tails sending ripples through the clear water. The sound of gurgling and splashing follows their movement and the currents, with the sandy, rocky seabed visible below.

Prompt: A large cartoonish alien head with big eyes and a small mouth appears on screen, looking concerned. Behind it, a dusk cityscape with tall buildings and a river sets the scene, as a deep boom echoes, followed by a softer one.

Figure A8: More examples for high-quality audio-video generation results.

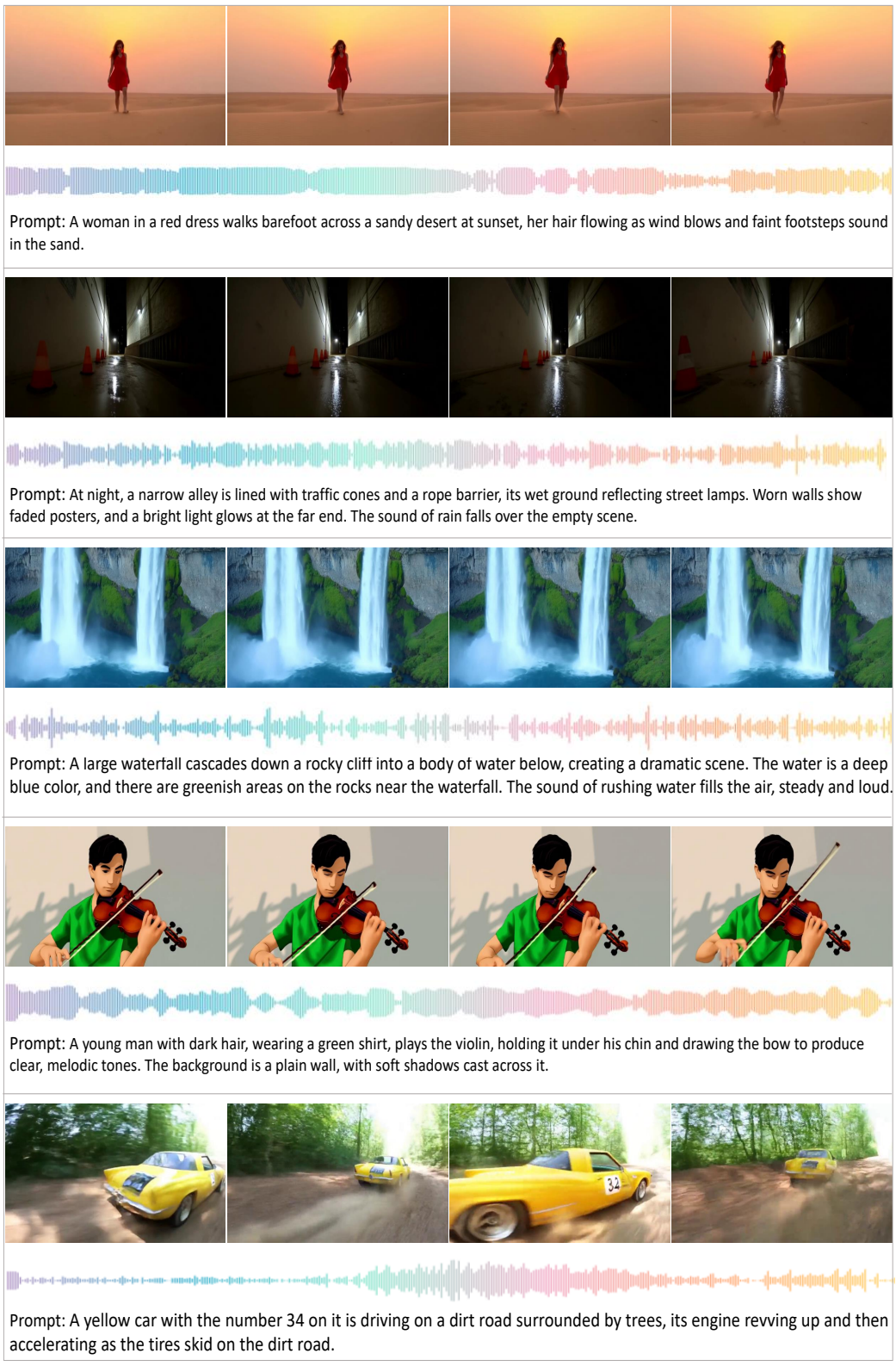


Figure A9: More examples for high-quality audio-video generation results.