EVA: EFFICIENT TRAINING FOR JOINT VIDEO-AUDIO GENERATION VIA SELECTIVE CROSS-MODAL INTERACTION

Anonymous authors
Paper under double-blind review



Figure 1: **Diverse, high-quality audio–video generation with efficient training.** EVA generates visually rich videos with synchronized audio across diverse scenarios. The upper-right plot compares synchronization ($1/\mathrm{DeSync}$) against an estimated resource index $\log_{10}(N_{\mathrm{M}}C)$, where N_{M} is the training data scale in millions of samples and C is GPU-days; upper-left is better. Missing resource entries are median-imputed; methods with fully undisclosed resources are shown in gray. Full videos are provided in the supplementary material.

ABSTRACT

Joint video–audio generation has advanced rapidly, with most recent approaches building upon pretrained text-to-video and text-to-audio models to construct joint generators. However, dense cross-modal interaction throughout training can introduce redundant or noisy signals in randomly initialized interaction modules, impairing both training efficiency and generation quality. Thus, we present an *E*fficient training framework for joint *V*ideo–*A*udio generation, namely EVA, for fast adaptation from pretrained uni-modal generators through multi-aspect selective cross-modal interaction: (1) at the sample level, EVA selectively enables audio-to-video, video-to-audio, or bidirectional interaction according to the dominant cross-modal dependency of each sample; (2) at the model level, it restricts cross-modal interaction to deeper layers to reduce unnecessary computation and avoid premature interference with uni-modal representations; (3) at the token level, adaptive gating modulates cross-modal attention to emphasize informative visual–audio correspondences and ignore irrelative interactions, improving synchronization while preserving uni-modal generation quality. On Verse-Bench, our 7B model achieves competitive visual, audio, and speech quality with only 3% training cost. It also obtains the lowest DeSync of 0.1902 and the highest LSE-C of 7.8018, demonstrating strong audio–visual synchronization. By reducing the cost of cross-modal adaptation, EVA can free training budget for scaling model capacity or further optimization. Code and weights will be released soon.

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.

1 INTRODUCTION

Joint video–audio generation (JAVG) has advanced rapidly, with recent methods (Liu et al., 2026a;b; Wang et al., 2025a; Low et al., 2025; Zhang et al., 2026b; Wang et al., 2026a; Ji et al., 2026; Kuaishou Technology, 2026; ByteDance Seed, 2026; MiniMax-AI, 2026; HaCohen et al., 2026) achieving high-quality visual and audio synthesis with increasingly precise cross-modal synchronization. Most recent approaches build joint generators upon strong pretrained text-to-video and text-to-audio models, thereby inheriting their established uni-modal generative capabilities and reduce adaption cost for JAVG. However, establishing cross-modal coordination between these pretrained generators still often relies on dense interaction and large-scale joint training. For example, Ovi (Low et al., 2025) processed approximately 30 million samples during training, while MOVA (Team et al., 2026b) was trained on 1,024 GPUs for 42 days, and NAVA (Ji et al., 2026) on 160 GPUs for one month. Moreover, newly introduced cross-modal interaction modules are typically randomly initialized, and dense coupling can inject redundant or noisy cross-modal signals into pretrained representations, slowing adaptation and potentially degrading generation quality. This motivates a more efficient approach that selectively introduces cross-modal interaction while retaining the strong generative priors of pretrained uni-modal models.

To characterize the redundancy of dense cross-modal interaction, we systematically analyze cross-modal attention in the representative dual-tower video–audio generation model LTX-2.3 (HaCohen et al., 2026) from both sample and model perspectives.

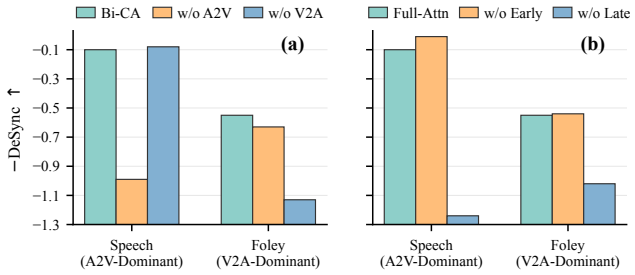


Figure 2: **Cross-modal interaction exhibits sample-dependent directionality and depth-wise redundancy.** **(a)** Different samples allow different non-dominant cross-attention directions to be removed. **(b)** Different samples exhibit consistent redundancy at the same network depths.

First, **the dominant direction of cross-modal interaction varies across samples**. Figure 2(a) shows for speech-dominant samples, removing audio-to-video (A2V) interaction substantially degrades video–audio synchronization, whereas removing video-to-audio (V2A) interaction has little effect. Foley-dominant samples exhibit the opposite pattern, with synchronization degrading more noticeably when V2A interaction is removed. This indicates that different samples rely on different directions of cross-modal information flow, allowing the non-dominant interaction direction to be selectively removed during training. This sample-dependent directionality is also consistent with existing generation paradigms: speech-related video generation is typically audio-driven (Gao et al., 2025; Wang et al., 2026b; Team et al., 2026a), whereas Foley and general sound generation are more commonly video-driven (Cheng et al., 2025; Shan et al., 2025; Yang et al., 2026a; Wang et al., 2025b).

Second, **cross-modal interaction exhibits consistent depth-wise redundancy across different sample types**. As shown in Figure 2(b), removing cross-modal interaction from the earlier half of the network has little effect on synchronization for both *Speech* and *Foley*, whereas removing interaction from the latter half causes a clear performance drop. Despite their different directional dependencies, these sample types exhibit a similar depth-wise interaction pattern. A related observation has also been reported in OmniNFT (Zhang et al., 2026a), where shallower layers primarily support intra-modal representation learning, while deeper layers contribute more strongly to cross-modal alignment. These results suggest that cross-modal interaction can be consistently restricted to deeper layers, avoiding redundant coupling in earlier layers.

In addition, even within the selected interaction paths, **not all cross-modal features contribute equally to synchronization**. Directly injecting all attended features may still introduce irrelevant information during cross-modal interaction. For example, in A2V scenarios, visual tokens around sound-producing regions such as the mouth are typically more relevant than background regions, while in V2A scenarios, audio tokens aligned with the occurrence of sound events are more informative than temporally unrelated ones. This suggests that cross-modal interaction should be further regulated at the token level to concentrate information exchange on synchronization-relevant correspondences.

Based on these findings, we propose EVA, an *E*fficient training framework for joint *V*ideo–*A*udio generation through multi-level selective cross-modal interaction. At the **sample level**, EVA selectively enables A2V, V2A, or bidirectional interaction according to the dominant cross-modal dependency of each sample. This avoids uniformly optimizing both interaction directions when only one direction provides the dominant cross-modal signal. At the **model level**, EVA introduces cross-modal interaction only in deeper layers, while preserving shallow layers for uni-modal representation learning, thereby reducing cross-modal computation and premature interference with pretrained representations. At the **token level**, adaptive gating modulates cross-modal attention to emphasize informative visual–audio correspondences while suppressing irrelevant interactions. Together, these designs shift joint training from dense cross-modal coupling toward selective adaptation of the interactions required for video–audio coordination.

Building on this framework, we construct a 7B joint video–audio generation model using Wan2.2-5B-TI2V (Wan Team, 2025) and the audio branch of MOVA (Team et al., 2026b) as pretrained generators. On Verse-Bench (Wang et al., 2025a), EVA achieves the lowest DeSync score of 0.1902 and the highest LSE-C of 7.8018, while maintaining competitive video, audio, and speech quality. Our model is trained using 64 GPUs for 20 days, which costs only approximately 3.0% of MOVA (Team et al., 2026b) and 26.7% of the NAVA (Ji et al., 2026). Further ablation studies validate the effectiveness of selective interactions at the sample, model, and token levels. By reducing the cost of cross-modal adaptation, EVA frees substantial training budget for further model scaling and optimization.

Our contributions are summarized as follows:

- We empirically analyze cross-modal interaction in dual-tower video–audio generators, revealing sample-dependent interaction directionality and consistent depth-wise redundancy across different scenarios.
- We propose EVA, an efficient training framework based on sample-level directional selection, model-level depth selection, and token-level adaptive regulation to reduce redundant cross-modal coupling and irrelevant feature injection.
- We build a 7B joint video–audio generation model that achieves the lowest DeSync and highest LSE-C on Verse-Bench with competitive generation quality, using 1,280 nominal GPU-days for approximately 3.0% of MOVA's and 26.7% of NAVA's training budgets.

2 Related Work

Joint Audio-Video Generation. Recent works jointly model audio and video to improve semantic consistency and temporal synchronization. JavisDiT and JavisDiT++ explore joint diffusion architectures and optimization strategies for synchronized generation (Liu et al., 2026a;b), while UniVerse-1 integrates pretrained audio and video experts (Wang et al., 2025a). Ovi adopts twin diffusion backbones with cross-modal fusion (Low et al., 2025), and LTX-2 employs an asymmetric dual-stream architecture (HaCohen et al., 2026). UniAVGen further introduces asymmetric cross-modal interactions with face-aware modulation (Zhang et al., 2026b). More recent systems, including MOVA and NAVA, scale joint audio-video generation while explicitly improving cross-modal synchronization and alignment (Team et al., 2026b; Ji et al., 2026), whereas MiniMax-H3 extends joint modeling to broader multimodal generation and editing settings (MiniMax-AI, 2026). In contrast to these works, we focus on reducing the joint-training cost of coupling pretrained audio and video generators by selectively allocating cross-modal interactions and optimization across content types, interaction directions, and network depths.

Directional Audio-Video Generation. Directional audio-video generation explicitly conditions one modality on the other. In audio-to-video generation, Wan-S2V and TalkVerse use audio signals to guide facial articulation and visual dynamics (Gao et al., 2025; Wang et al., 2026b). Audio conditioning has also been extended beyond speech to music-driven human motion, as explored by MACE-Dance and OmniDance (Yang et al., 2026c;b). Conversely, video-to-audio methods such as MMAudio and HunyuanVideo-Foley synthesize temporally aligned sounds from visual and textual cues (Cheng et al., 2025; Shan et al., 2025). These directional paradigms highlight that different content types naturally favor different directions of cross-modal conditioning. Building on this observation, EVA adopts content-dependent interaction directions during training while retaining a unified joint audio-video generator at inference.

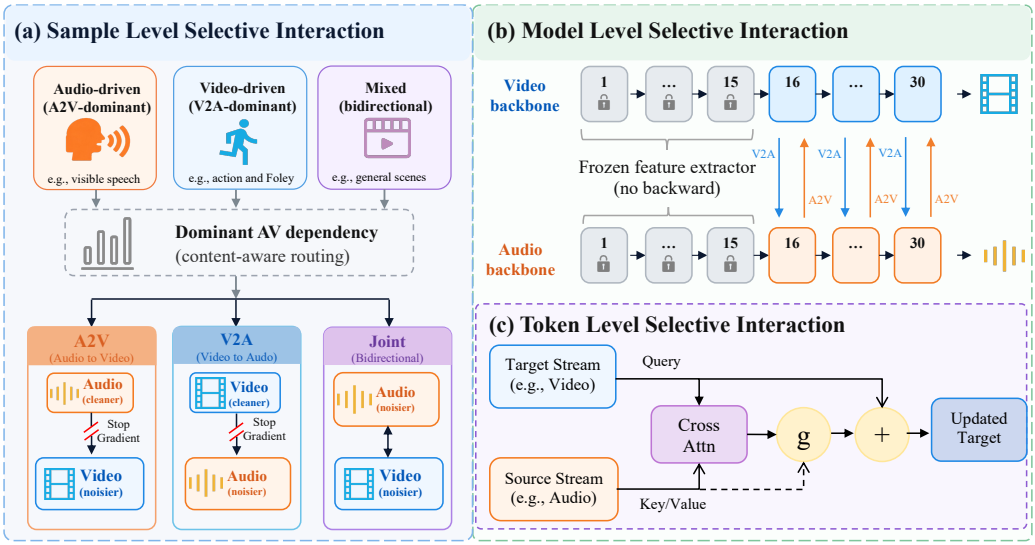


Figure 3: **Overview of EVA. (a)** Sample level selective interaction emphasizes A2V, V2A, or Joint interaction according to the dominant cross-modal dependency of each content type. **(b)** Model level selective interaction restricts the cross modal interaction to half later layers. **(c)** Token level selective interaction regulate the cross-modal contribution through a token-level gating.

3 METHOD

3.1 PRELIMINARY: DUAL-TOWER AUDIO-VIDEO GENERATION MODEL

Architecture. A typical dual-tower audio-video generator consists of a text-to-video DiT and a text-to-audio DiT, with cross-modal attention connecting the two streams (Liu et al., 2026a; Low et al., 2025; Team et al., 2026b; HaCohen et al., 2026). Let h^v and h^a denote the video and audio hidden states, respectively. For a target stream h^t and a source stream h^s, we define cross-modal attention as

$$\mathcal{C}(h^t, h^s) = \mathrm{Attn}\left(Q_t(h^t), K_s(h^s), V_s(h^s)\right), \tag{1}$$

where the target stream provides the queries, while the source stream provides the keys and values. Accordingly, audio-to-video (A2V) and video-to-audio (V2A) interactions are given by

$$\mathcal{C}_{a\to v} = \mathcal{C}(h^v, h^a), \qquad \mathcal{C}_{v\to a} = \mathcal{C}(h^a, h^v). \tag{2}$$

Thus, A2V retrieves audio information to update the video stream, whereas V2A retrieves video information to update the audio stream. The resulting cross-modal features are projected and added back to the corresponding target stream through residual connections.

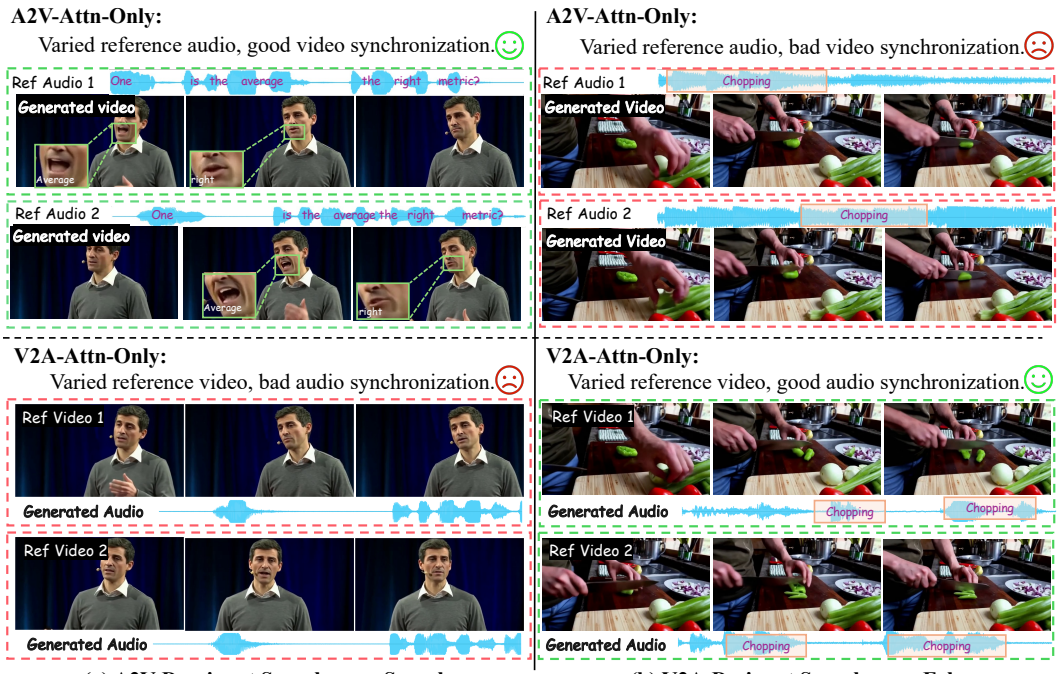
Training. The video and audio branches are jointly optimized with separate rectified flow objectives (Liu et al., 2023). Let x_0^v and x_0^a denote clean video and audio latents, and ϵ^v and ϵ^a independent Gaussian noise. At noise levels σ_v and σ_a, the corrupted latent of modality $m \in \{v, a\}$ is constructed as

$$x_{\sigma_m}^m = (1-\sigma_m)x_0^m + \sigma_m \epsilon^m, \tag{3}$$

with target velocity $u^m = \epsilon^m - x_0^m$. The two branches predict their respective velocity fields and are optimized jointly as

$$\mathcal{L}_{\mathrm{FM}} = \mathcal{L}_v + \lambda_a \mathcal{L}_a, \qquad \mathcal{L}_m = \mathbb{E}\left[\left\| v_\theta^m(x_{\sigma_v}^v, x_{\sigma_a}^a, \sigma_v, \sigma_a) - u^m \right\|_2^2\right]. \tag{4}$$

Because the two branches exchange information through A2V and V2A cross-attention, the modality-specific flow-matching losses jointly optimize the two generation streams together with the cross-modal attention modules.



(a) A2V-Dominant Sample, e.g., Speech (b) V2A-Dminant Sample, e.g., Foley

Figure 4: **Sample-dependent cross-modal directionality.** **(a)** For A2V-dominant samples (e.g., in speech scenarios), A2V interaction better preserves lip synchronization, as highlighted by the mouth close-ups variance. **(b)** For V2A-dominant samples (e.g., in foley scenarios), V2A interaction yields more consistent sound-event alignment, as indicated by waveform annotations.

3.2 Sample Level Selective Interaction

The preliminary analysis in Figure 2(a) reveals a clear directional asymmetry: speech synchronization depends more strongly on A2V interaction, whereas Foley synchronization relies more on V2A interaction. We further examine this behavior by retaining only one interaction direction at a time in Figure 4, where LTX-2.3 (Lightricks, 2026) serves as the base model. For A2V interaction, audio serves as the reference modality to guide video denoising; for V2A interaction, video guides audio denoising. We fix the initial noise of the denoised target modality while varying the reference modality, and examine whether the generated target follows the corresponding temporal changes.

As shown in Figure 4(a), speech samples exhibit an A2V-dominant dependency: changes in reference speech timing induce corresponding changes in generated mouth articulation, whereas varying reference mouth movements under V2A interaction have limited influence on the generated speech timing. Foley samples in Figure 4(b) exhibit the opposite pattern: changes in reference action timing lead to corresponding changes in generated sound events under V2A interaction, while varying reference sound timing under A2V interaction has limited influence on the generated action sequence.

These responses suggest that audio primarily provides temporal guidance for speech articulation, whereas video provides temporal guidance for Foley sound generation, explaining the directional asymmetry observed in Figure 2(a). We therefore adopt **sample-level selective interaction**, assigning A2V, V2A, or bidirectional interaction according to each sample's dominant cross-modal dependency.

Data organization. We use Qwen3.6-27B (Qwen Team, 2026) to infer each sample's dominant cross-modal dependency from structured audio–visual annotations and group samples into Audio-Driven, Video-Driven, and Mixed pools for A2V, V2A, and bidirectional training, respectively. Details are provided in Section B.

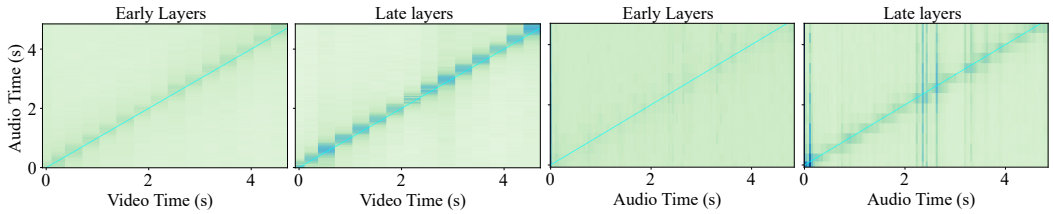
Directional interaction. We apply masks $(m_{a\to v}, m_{v\to a})$ to the cross-attention outputs, using $(1, 0)$, $(0, 1)$, and $(1, 1)$ for A2V, V2A, and bidirectional interaction. To provide a cleaner reference signal, we set $\sigma_a < \sigma_v$ for A2V and $\sigma_v < \sigma_a$ for V2A, while bidirectional training uses $\sigma_a = \sigma_v$.

Gradient routing. For A2V, we use $\mathcal{C}(h^v, \mathrm{sg}(h^a))$; for V2A, $\mathcal{C}(h^a, \mathrm{sg}(h^v))$. The stop-gradient operator detaches the reference modality before key/value projection, preventing the target loss from updating the reference backbone through cross-modal attention. Bidirectional training allows gradients to pass through both directions.

3.3 Model Level Selective Interaction

The layer-wise investigation in Figure 2(b) shows that the contribution of cross-modal interaction to synchronization varies with network depth. To understand this behavior, we compare temporal attention patterns across layers. As shown in Figure 5, early-layer attention is relatively divergent, whereas later layers exhibit more structured temporal correspondence, particularly the pronounced diagonal band in V2A attention. These patterns suggest that cross-modal temporal associations are more explicitly captured in later layers, helping interpret the depth-dependent investigation results.

Motivated by these observations, we introduce **model-level selective interaction**, placing cross-modal attention only in the latter half of each backbone. We also restrict lightweight LoRA adaptation to these layers. This placement allows the earlier backbone layers to remain frozen during adaptation, eliminating backpropagation through them and reducing training cost.



(a) V2A cross attention map **(b) A2V cross attention map**

Figure 5: **Cross-modal attention across model depth.** Early layers exhibit relatively diffuse attention patterns, while later layers show more structured temporal correspondence between audio and video.

3.4 Token Level Selective Interaction

Even within the selected interaction directions and layers, cross-modal features contribute unequally to synchronization, and uniformly injecting them may introduce irrelevant information into pretrained uni-modal representations. We therefore introduce **token-level selective interaction** through adaptive gates that regulate the cross-modal contribution to each target token. For interaction from source modality s to target modality t, let h_i^t denote the target hidden state and c_i the retrieved cross-modal feature at token i. A linear projection followed by a sigmoid produces the gating weights. Omitting normalization for brevity, the gated update is

$$g_i = \mathrm{sigmoid}(W_g[h_i^t; c_i] + b_g), \qquad h_i^{t\prime} = h_i^t + m_{s\to t} W_o(g_i \odot c_i). \tag{5}$$

Here, $[\cdot;\cdot]$ denotes concatenation, W_g and b_g parameterize the gating projection, and W_o is the output projection. The gating weights are broadcast over the corresponding feature dimensions for element-wise multiplication $\odot$, while $m_{s\to t}$ is the directional mask selected at the sample level. Conditioning on both the target state and retrieved cross-modal features allows the gate to adaptively regulate feature injection across spatial and temporal positions, emphasizing informative correspondences while suppressing irrelevant interactions.

4 Experiments

4.1 Experimental Setup

Implementation Details. We initialize our 7B-parameter model with Wan2.2-TI2V-5B (Wan et al., 2025) for video and MOVA's 1.3B audio branch (Team et al., 2026b) for audio. We adopt a three-stage training strategy using public and internally collected audio–video data filtered for quality and cross-modal alignment. Data preparation and training details are provided in Section B.

Table 1: Quantitative comparison on Verse-Bench. We report video quality, audio aesthetics, speech quality, and audio-video alignment metrics. **Bold** and underlined values indicate the best and second-best results, respectively. * MOVA is a MoE model, with 32B total parameters and 18B active.

		Video	Audio Aesthetics				Speech	AV-Align	
Model	**Params**	**VQ↑**	**CE↑**	**CU↑**	**PC↓**	**PQ↑**	**WER↓**	**LSE-C↑**	**DeSync↓**
NAVA	6.3B	0.6116	4.7695	6.0053	2.3027	6.3779	0.1574	7.7261	0.2342
UniAVGen	7.1B	**0.6644**	4.5938	5.8605	2.1534	**6.7643**	0.1661	4.9556	0.5371
Ovi	10B	0.6542	4.7756	5.9832	**2.1280**	6.1427	**0.1053**	7.3095	0.4730
MagiHuman-256p	15B	0.6067	4.9288	6.0619	2.4117	6.2781	0.1311	5.7972	0.4150
MagiHuman-512p	15B	0.6013	4.9361	6.0720	2.4156	6.2836	0.1228	6.8997	0.4010
LTX-2.3	22B	0.6285	**5.1876**	**6.4033**	2.6585	6.7169	0.1113	7.6768	0.2412
MOVA	32B*	0.6118	4.8042	6.0768	2.3508	6.4768	0.1344	7.4011	0.4304
Ours	7B	0.6568	4.7210	6.0818	2.1739	6.3519	0.1112	**7.8018**	**0.1902**

Evaluation Metrics. We evaluate EVA on Verse-Bench (Wang et al., 2025a), which contains 600 prompts and accompanying images across two general audio–visual subsets and one speech-centric subset. We assess video quality (VQ), audio aesthetics (CE, CU, PC, and PQ), speech accuracy (WER), and audio–video alignment (LSE-C, and DeSync). Detailed metric definitions and evaluation protocols are provided in the Section B.6.

4.2 Main Results

Quantitative Comparison. As shown in Table 1, EVA achieves the best DeSync (0.1902) and LSE-C (7.8018) scores among the compared methods, indicating strong temporal audio–video alignment and lip synchronization. Its DeSync score is 18.8% lower than that of the next-best model, NAVA (Ji et al., 2026). Meanwhile, EVA ranks second in VQ (0.6568), WER (0.1112), and CU (6.0818), achieving strong synchronization alongside competitive visual quality, speech accuracy, and audio content usefulness. Comparisons with the pretrained uni-modal backbones show that EVA largely preserves their generation quality, suggesting that stronger backbones could further improve its video and audio quality. Detailed results are provided in Section D.

With 7B parameters, EVA also achieves higher VQ and LSE-C scores and lower DeSync than the 22B LTX-2.3 (Lightricks, 2026), while attaining comparable WER (0.1112 vs. 0.1113). These results show that a relatively compact joint model can achieve competitive generation quality and strong cross-modal synchronization.

Training Efficiency. As shown in Table 2, EVA requires approximately 1,280 GPU-days on 64 GPUs for joint audio–video training, corresponding to 28.6% of the GPU-days reported for NAVA (Ji et al., 2026). It processes approximately 5M training samples, one-sixth of the approximately 30M reported for Ovi (Low et al., 2025). Here, *Samples Seen* denotes the cumulative number of samples processed throughout training. Together with the results in Table 1, these comparisons show that EVA achieves competitive generation quality and strong cross-modal synchronization with a substantially smaller training budget.

Table 2: Training-efficiency comparison with representative joint audio-video generation models.

Model	GPUs	GPU Days	Samples Seen
Ovi	–	–	∼30M
MOVA	1024	∼43008	–
NAVA	160	∼4480	–
Ours	64	∼1280	∼5M

Qualitative Comparison. Figure 6 presents qualitative comparisons on an A2V-dominant speech example and a V2A-dominant explosion example. EVA generates lip movements synchronized with speech in the former and sound events temporally aligned with the visible explosion in the latter. Additional comparisons and full videos are provided in the supplementary material.

4.3 In-depth Analysis

Sample level selective interaction enables more efficient cross-modal learning. We compare our strategy with naive joint training over 50K optimization steps, using identical model initialization and



Figure 6: **Qualitative comparison on speech and explosion scenarios.** Red text highlights key prompt content; blue traces show audio waveforms.

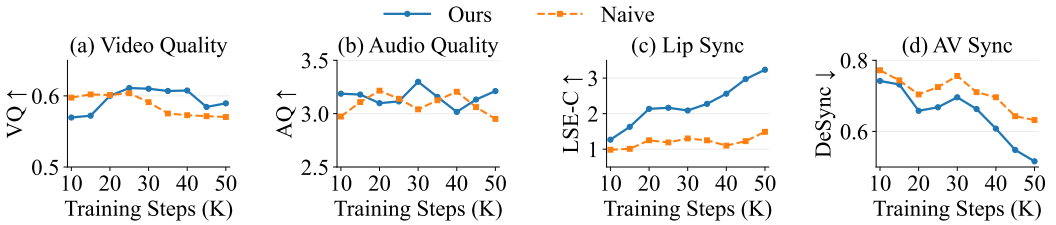


Figure 7: **Effect of sample-level selective interaction.** Our strategy achieves better synchronization while maintaining comparable video and audio quality. Audio quality is summarized as $\mathrm{AQ} = (\mathrm{CU} + \mathrm{CE} - \mathrm{PC} + \mathrm{PQ})/4$.

training dataset. As shown in Figure 7, both strategies achieve comparable video and audio quality but exhibit markedly different synchronization behavior. With directional training, LSE-C increases and DeSync decreases overall, with the synchronization gap relative to naive joint training widening as training progresses. These results indicate that configuring cross-modal interactions according to content dependencies enables the model to establish stronger audio–visual correspondence in fewer training steps while maintaining comparable generation quality.

Model level selective interaction reduces training cost. We train models with cross-modal interaction in all layers or only the later half for 10K steps each on the high-quality data used in Stage 3. As shown in Table 3, both settings achieve comparable video quality, audio quality, and audio–visual synchronization (captured by LSE_C), while late-layer interaction reduces training time per iteration from 41 to 27 seconds, a reduction of approximately 34.1%. These results suggest that concentrating cross-modal interaction in later layers enables more efficient joint audio–video training while maintaining generation quality and synchronization.

Table 3: **Ablation of cross-modal interaction.** Per-iteration training time is measured on NVIDIA H20 GPUs.

Setting	VQ↑	AQ↑	Sync↑	Time(s/iter)
Full-layer	0.6617	3.3270	1.2940	41
Late-layer	0.6631	3.4663	1.4177	27

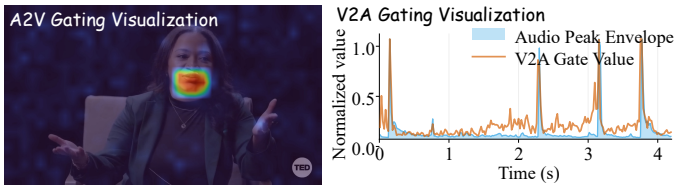


Figure 8: Visualization of learned cross-modal gates.

Table 4: Ablation of gated cross-modal interaction on uni-modal generation quality.

Setting	VQ↑	AQ↑	Sync↑
w/o Gate	0.63	3.28	1.27
w/ Gate	**0.66**	**3.47**	**1.42**

Token level selective interaction helps preserve uni-modal generation quality. We train the model with and without gating for 10K steps in each setting on the high-quality data used in Stage 3 to evaluate the effect of gating on video and audio generation quality. As shown in Table 4, enabling gating improves VQ from 0.6329 to 0.6631 and AQ from 3.2787 to 3.4663. Figure 8 further illustrates the selective responses of the learned gates: A2V gates exhibit higher values around the speaker's mouth, while V2A gate values increase near several audio energy peaks. These results indicate that gating selectively regulates cross-modal feature injection according to content, yielding better uni-modal generation quality after the same number of training steps.

4.4 User Study

We conduct a user study comparing EVA with six baselines in terms of text–video alignment, audio–visual alignment, video quality, and audio quality. As shown in Figure 9, EVA receives more wins than losses against every baseline in audio–visual alignment and audio quality. In particular, its audio–visual alignment win rates against Ovi (Low et al., 2025) and UniAVGen (Zhang et al., 2026b) reach 61.9% and 69.0%, respectively. Human evaluations also indicate competitive video quality and text–video alignment, with more wins than losses against five baselines and broadly comparable preferences to MOVA (Team et al., 2026b). These results complement the quantitative evaluation, demonstrating favorable human assessments of our model's generation quality and cross-modal alignment.

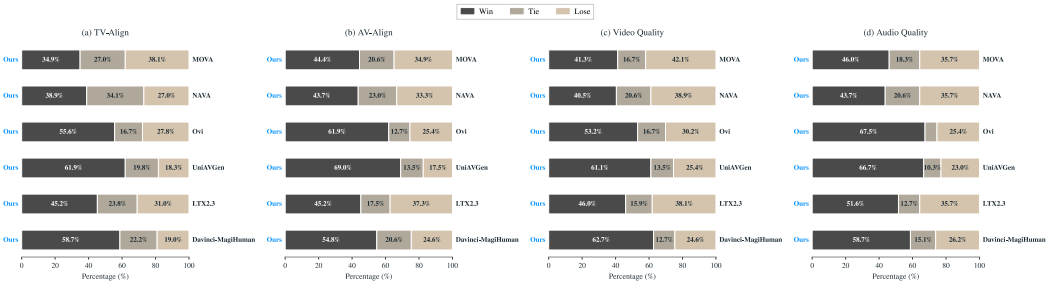


Figure 9: **Pairwise human preference against joint audio-video baselines.** Each row compares EVA (Ours) with the indicated baseline for (a) text-video alignment, (b) audio-video alignment, (c) video quality, and (d) audio quality. Dark gray, gray, and beige show the percentages of EVA wins, ties, and losses, respectively. Detailed study settings are collected in Appendix B.7.

5 Conclusion

This work presents EVA, a joint audio–video generation model that integrates pretrained video and audio generators through selective cross-modal training. Our approach selects interaction directions according to content dependencies, adaptively gates cross-modal feature injection, and restricts interaction to the latter half of the network, alongside early-layer freezing and progressive training to reduce optimization cost. Experiments show that EVA achieves competitive generation quality and strong audio–visual synchronization with a modest training budget. These results suggest that building on pretrained uni-modal capabilities and selectively learning cross-modal interactions provides a practical path toward efficient training for joint audio–video generation.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, including translation, language refinement. We also used AI models for audio–video data annotation and caption refinement, as described in the Section B.2, and AI coding assistants for implementation. Generative AI was not used to originate the research ideas or design the proposed method. The authors reviewed and revised the AI-assisted manuscript text, and AI-generated code was reviewed by at least three authors. We take full responsibility for the final content of this work, including its claims, data annotations, code, and other AI-assisted materials.

ETHICS STATEMENT

This work aims to improve the training efficiency of joint audio–video generation. The resulting technology can support creative applications but may also be misused to produce misleading content, impersonate individuals, or reproduce personal likenesses and voices without consent. We advocate clear disclosure of AI-generated content and respect for privacy and consent. Our model may inherit biases from its pretrained backbones and training data, which should be considered when interpreting its outputs and deploying it in practice.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of model design, training, and evaluation in both the main paper and the appendix. Furthermore, code and trained models will be publicly released to ensure reproducibility of our results.

REFERENCES

ByteDance Seed. Seedance 2.5. Official model announcement, 2026. URL https://seed.bytedance.com/en/seedance2_5.

Delong Chen, Tejaswi Kasarla, Yejin Bang, Mustafa Shukor, Willy Chung, Jade Yu, Allen Bolourchi, Theo Moutakanni, and Pascale Fung. Action100m: A large-scale video action dataset. *arXiv preprint arXiv:2601.10592*, 2026.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020. URL https://arxiv.org/abs/2004.14368.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911. IEEE, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/.

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.

Jort F. Gemmeke, Daniel P. W. 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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.

Longbin Ji, Guan Wang, Xuan Wei, Chenye Yang, Xiangrui Liu, Zhenyu Zhang, Shuohuan Wang, Yu Sun, and Jingzhou He. Native audio-visual alignment for generation. *arXiv preprint arXiv:2605.30073*, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5148–5157, October 2021.

Kuaishou Technology. Kling ai launches 3.0 model, ushering in an era where everyone can be a director. Official model announcement, 2026. URL https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-30-model-ushering-era-where-everyone-can-be.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, and Siyu Zhu. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7752–7762, June 2025.

Lightricks. LTX-2.3. Official model card, 2026. URL https://huggingface.co/Lightricks/LTX-2.3.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Scofield Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations*, volume 2026, pp. 139160–139194, 2026a.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Scofield Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. In *International Conference on Learning Representations*, volume 2026, pp. 150592–150618, 2026b.

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*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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

MiniMax-AI. Minimax-h3. Official model release, 2026. URL https://github.com/MiniMax-AI/MiniMax-H3.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

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.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

SII-GAIR, Sand.ai, Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026. URL https://arxiv.org/abs/2603.21986.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Meituan LongCat Team, Xunliang Cai, Meng Cheng, Feng Gao, Zhe Kong, Jiamu Li, Le Li, Weiheng Li, Hongyu Liu, Shuai Tan, et al. Longcat-video-avatar 1.5 technical report. *arXiv preprint arXiv:2605.26486*, 2026a.

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*, 2026b.

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

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.

Wan Team. Wan2.2: Open and advanced large-scale video generative models. Official model release, 2025. URL https://github.com/Wan-Video/Wan2.2.

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.

Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, et al. Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*, 2025b.

Jun Wang, Chunyu Qiang, Yuxin Guo, Yiran Wang, Xijuan Zeng, and Feng Deng. Apollo: Unified multi-task audio-video joint generation. *arXiv preprint arXiv:2601.04151*, 2026a.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8428–8437. IEEE, 2025c.

Zhenzhi Wang, Jian Wang, Ke Ma, Dahua Lin, and Bing Zhou. Talkverse: Democratizing minute-long audio-driven video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4516–4527, 2026b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

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*, 2023.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Jianxuan Yang, Xinyue Guo, Zhi Cheng, Kai Wang, Lipan Zhang, Jinjie Hu, Qiang Ji, Yihua Cao, Yihao Meng, Zhaoyue Cui, et al. Controlfoley: Unified and controllable video-to-audio generation with cross-modal conflict handling. *arXiv preprint arXiv:2604.15086*, 2026a.

Kaixing Yang, Jiashu Zhu, Xulong Tang, Ziqiao Peng, Xiangyue Zhang, Chubin Chen, Puwei Wang, Jiahong Wu, Xiangxiang Chu, Hongyan Liu, and Jun He. OmniDance: Multimodal driven dance video generation with large-scale internet data. In *Computer Vision – ECCV 2026*, pp. 571–588, 2026b. doi: 10.1007/978-3-032-37369-4_31. URL https://link.springer.com/chapter/10.1007/978-3-032-37369-4_31.

Kaixing Yang, Jiashu Zhu, Xulong Tang, Ziqiao Peng, Xiangyue Zhang, Puwei Wang, Jiahong Wu, Xiangxiang Chu, Hongyan Liu, and Jun He. MACE-Dance: Motion-appearance cascaded experts for music-driven dance video generation. In *Proceedings of the ACM SIGGRAPH Conference*, 2026c. URL https://arxiv.org/abs/2512.18181.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1190–1199. IEEE, 2022.

Zhen Ye, Xu Tan, Aoxiong Yin, Hongzhan Lin, Guangyan Zhang, Peiwen Sun, Yiming Li, Chi-Min Chan, Wei Ye, Shikun Zhang, et al. Talker-t2av: Joint talking audio-video generation with autoregressive diffusion modeling. *arXiv preprint arXiv:2604.23586*, 2026.

Guohui Zhang, XiaoXiao Ma, Jie Huang, Hang Xu, Hu Yu, Siming Fu, Yuming Li, Zeyue Xue, Lin Song, Haoyang Huang, et al. Omninft: Modality-wise omni diffusion reinforcement for joint audio-video generation. *arXiv preprint arXiv:2605.12480*, 2026a.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1950–1960, 2026b.

Youliang Zhang, Zhaoyang Li, Duomin Wang, jiahe zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 117896–117926, 2026c. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/bf7dbac50ed7f6e12ad529c5b9396bc4-Paper-Conference.pdf.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

A Appendix Overview

This appendix provides implementation details and complementary evaluations for EVA. It is organized into five parts, including this overview. Appendix B describes the training data, model adaptation, optimization, and evaluation settings, and user-study protocol. Appendix C compares EVA with MiniMax-H3 using automatic metrics and human preference. Appendix D examines how joint adaptation affects the generation quality of the pretrained video and audio backbones. Finally, Appendix E discusses qualitative results, diagnostic failure modes, and remaining limitations.

B Implementation Details

We first describe our three-stage training strategy, followed by the data construction pipeline, training details, and evaluation protocols.

B.1 Progressive Joint Training

To exploit pretrained generative capabilities while controlling training cost, we adopt a three-stage schedule that progresses from cross-modal adaptation to full-model refinement:

- **Stage 1: Cross-modal adaptation.** We freeze the pretrained backbone weights and train LoRA adapters (Hu et al., 2022) in the latter halves of both streams, together with the newly introduced cross-attention modules and gates. This concentrates initial optimization on cross-modal coordination with a limited set of trainable parameters.
- **Stage 2: Full-model optimization.** We merge the LoRA weights into the backbones and unfreeze both streams, jointly optimizing the backbones and cross-modal modules to enable broader adaptation.
- **Stage 3: High-quality finetuning.** We continue full-parameter finetuning on a filtered high-quality subset with progressively reduced learning rates to refine generation quality and synchronization.

The A2V, V2A, and Joint configurations are used throughout all stages, while cross-modal interaction remains restricted to the latter half of the network.

B.2 Training Data and Capability Assignment

Data Collection. We construct our audio-video corpus by integrating multiple public datasets, including Koala-36M (Wang et al., 2025c), VGGSound (Chen et al., 2020), AudioSet (Gemmeke et al., 2017), OpenHumanVid (Li et al., 2025), SpeakerVid-5M (Zhang et al., 2026c), Action100M (Chen et al., 2026), and Talker-T2AV (Ye et al., 2026), together with internally collected data. These heterogeneous sources provide complementary coverage across different content domains and audio-video interactions.

Data Filtering. We first segment the collected videos at scene boundaries using PySceneDetect[1], remove near-silent clips based on audio RMS energy, and discard invalid samples. We assess visual quality using Q-Align (Wu et al., 2023), estimate motion magnitude using UniMatch-based optical flow (Xu et al., 2023), and evaluate audio quality using Audiobox Aesthetics (Tjandra et al., 2025). We use Synchformer (Iashin et al., 2024) to assess general audio–visual synchronization and additionally apply SyncNet (Chung & Zisserman, 2016) to clips with visible speech for fine-grained lip–audio alignment assessment. The filtering criteria and cumulative training sample counts for each stage are detailed below.

- **Stage 1.** We discard videos shorter than 5 seconds or with a resolution below 360p. We randomly sample from the remaining pool and process approximately 4.5M examples during training.
- **Stage 2.** We filter samples based on audio–visual synchronization, retaining those satisfying $-0.2 < \text{DeSync} < 0.2$ and LSE-C ≥ 5. Approximately 0.6M examples are processed during this stage.

[1] https://github.com/breakthrough/pyscenedetect

- **Stage 3.** We select high-quality samples with a Q-Align score of at least 0.6, a motion score within $[0.01, 25.0]$, and an audio quality score of at least 6.0. Approximately 0.1M examples are processed during this stage.

Data annotation. We use Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025) to jointly analyze the audio and video of each filtered clip. The resulting structured annotations describe visual content, speech, music, sound events, and ambience, while recording event order and distinguishing visually grounded sounds from off-screen audio. We additionally use Qwen3-ASR-1.7B (Shi et al., 2026) to transcribe speech. Qwen3.6-27B (Qwen Team, 2026) then combines the multimodal annotations and ASR transcripts into coherent captions, preserving temporal structure and spoken content.

Capability Assignment We use Qwen3.6-27B (Qwen Team, 2026) to infer the dominant cross-modal dependency of each sample from its structured audio–video annotations and assign it to one of three subsets: Audio-Driven (A2V-dominant), Video-Driven (V2A-dominant), or Mixed (Joint). Audio-Driven samples emphasize audio cues for visual generation, such as speech timing that guides visible articulation. Video-Driven samples emphasize visual cues for sound generation, such as actions and object interactions associated with corresponding sound events. The Mixed subset includes samples with bidirectional dependencies, as well as those that do not fall into either the Audio-Driven or Video-Driven category. Samples from these subsets are used to optimize audio-to-video, video-to-audio, and bidirectional interactions, respectively.

B.3 Backbones and Cross-Modal Adaptation

EVA initializes its video branch from Wan2.2-TI2V-5B (Wan et al., 2025; Wan Team, 2025) and its audio branch from MOVA's 1.3B audio generator (Team et al., 2026b), yielding approximately 7B parameters in total. Cross-modal attention is introduced only in the latter half of each backbone. During Stage 1, we freeze all pretrained backbone weights and train rank-256 LoRA adapters (Hu et al., 2022) in the latter halves, together with the newly introduced cross-modal attention modules and gates.

We map audio and video positions to a shared temporal coordinate system before applying rotary position embeddings (Su et al., 2021) in cross-modal attention (Cheng et al., 2025; Team et al., 2026b). This provides a consistent temporal reference for streams with different temporal token rates without resampling their latent sequences. Directional masks control which cross-modal paths are active, while token- and head-wise gates modulate the injected cross-modal features.

B.4 Directional Training and Optimization

We use a single model for A2V, V2A, and Joint training. A2V activates only the audio-to-video path with $\sigma_a < \sigma_v$, whereas V2A activates only the video-to-audio path with $\sigma_v < \sigma_a$. Joint training enables both directions with $\sigma_a = \sigma_v$.

We sample the interaction direction according to each sample's capability assignment. For Audio-Driven samples, A2V, V2A, and Joint training are selected with probabilities 0.7, 0.1, and 0.2, respectively. For Video-Driven samples, the corresponding probabilities are 0.1, 0.7, and 0.2. Mixed samples use Joint training. This sampling strategy prioritizes the dominant dependency of each content type while retaining opportunities for reverse and bidirectional interaction, allowing each subset to contribute to multiple forms of cross-modal learning.

The active interaction paths, noise levels, and gradient routing follow the sampled direction. During A2V or V2A training, we apply stop-gradient to the conditioning hidden representation before the key and value projections. This prevents the target-stream loss from updating the conditioning backbone through cross-modal attention, while the conditioning stream continues to receive its own flow-matching loss. This control of cross-modal gradient flow is independent of backbone freezing.

Both streams are trained with rectified flow objectives (Liu et al., 2023). The combined objective is

$$\mathcal{L} = \mathbb{E}\big[\lambda_v \|v_\theta^v - u^v\|_2^2 + \lambda_a \|v_\theta^a - u^a\|_2^2\big], \tag{A1}$$

where v_θ^m and u^m denote the predicted and target velocities for modality m, respectively. We use $\lambda_v = \lambda_a = 1$ in Stage 1 and $\lambda_v = 1, \lambda_a = 0.2$ in Stages 2 and 3.

B.5 Progressive Training Schedule

Table A1 summarizes the three-stage training schedule. Cross-modal adaptation first establishes cross-modal correspondence through lightweight adaptation. Full-model optimization then jointly refines both backbones and the cross-modal modules, followed by high-quality finetuning on the curated audio–visual subset.

Table A1: Progressive training schedule. Pretrained backbone weights are frozen in Stage 1 and unfrozen in Stages 2 and 3. Sample counts include repeated presentations during training.

Stage	**Trainable components**	**Learning rate**	**Samples seen**
Cross-modal adaptation	Later-layer rank-256 LoRA, cross-modal attention, and gates.	LoRA: 10^{-4}; cross-modal modules: 2×10^{-5}.	$\sim$4.5M
Full-model optimization	Full model after merging LoRA into the backbones.	Backbones: 10^{-5}; cross-modal modules: 2×10^{-5}.	$\sim$600K
High-quality finetuning	Full model.	Cosine decay from 10^{-5} to 10^{-6}.	$\sim$100K

We train on 64 NVIDIA H100 GPUs with a global batch size of 64, using the AdamW optimizer (Loshchilov & Hutter, 2019) and bfloat16 (bf16) precision. We enable gradient checkpointing (Chen et al., 2016) and Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023), with gradient accumulation disabled. Stage 1 uses separate learning rates for the LoRA adapters and newly introduced cross-modal modules. Stage 2 lowers the backbone learning rate while maintaining the cross-modal learning rate. Stage 3 applies cosine learning-rate decay for full-model finetuning.

The total joint-training budget is approximately 1,280 GPU-days, with approximately 5M samples processed cumulatively across the three stages. This budget excludes the original pretraining of the uni-modal generators.

B.6 Evaluation

Metrics. We follow the Verse-Bench evaluation protocol (Wang et al., 2025a). The benchmark contains 600 prompts across three subsets: Sets 1 and 2 cover general audio–visual scenes, while Set 3 focuses on speech. We evaluate generated samples from four complementary aspects:

- **Video quality.** We compute VQ by combining perceptual quality and identity consistency. Perceptual quality is assessed using Aesthetic Predictor[2], MANIQA (Yang et al., 2022), and MUSIQ (Ke et al., 2021), while identity consistency (ID) is measured using DINOv3 (Siméoni et al., 2025). Specifically, $\mathrm{VQ} = (\mathrm{Aesthetic} + \mathrm{MANIQA} + \mathrm{MUSIQ})/6 + \mathrm{ID}/2$, assigning equal weight to the average perceptual quality score and identity consistency.
- **Audio quality.** We use Audiobox Aesthetics (Tjandra et al., 2025) and report four dimensions: Content Enjoyment (CE), Content Usefulness (CU), Production Complexity (PC), and Production Quality (PQ).
- **Speech generation.** We evaluate speech content accuracy on Set 3 using Word Error Rate (WER), comparing transcriptions of generated speech with the reference transcripts.
- **Audio–visual alignment.** We assess audio–visual alignment using Synchformer-based DeSync (Iashin et al., 2024). We additionally report SyncNet-based LSE-C (Chung & Zisserman, 2016) on Set 3 to assess lip–audio synchronization.
  For DeSync, we use MOVA's publicly released evaluation implementation (Team et al., 2026b).[3] The per-clip score is the average absolute temporal offset predicted over two temporal windows. DeSync is measured in seconds, with lower values indicating closer temporal alignment.

WER and DeSync are lower-is-better. We also retain the lower-is-better PC convention used in the main comparison, although greater production complexity does not necessarily indicate poorer audio quality. All remaining metrics are higher-is-better.

[2] https://github.com/discus0434/aesthetic-predictor-v2-5

[3] https://github.com/OpenMOSS/MOVA/tree/main/evaluation

Compared Methods. We compare EVA with NAVA(Ji et al., 2026), MOVA (Team et al., 2026b), Ovi (Low et al., 2025), LTX-2.3 (HaCohen et al., 2026), daVinci-MagiHuman (SII-GAIR et al., 2026), and UniAVGen (Zhang et al., 2026b). For daVinci-MagiHuman, we report results at both 256p and 512p. The quantitative results are presented in Table 1.

B.7 User-Study Settings

The pairwise studies complement the automatic metrics with perceptual judgments. The main comparison in Figure 9 evaluates EVA against MOVA, NAVA, Ovi, UniAVGen, LTX-2.3, and Davinci-MagiHuman. The MiniMax-H3 comparison in Appendix C uses the same four reported dimensions. The backbone comparison in Appendix D focuses on video and audio quality separately, since the standalone backbones do not produce jointly aligned audio-video pairs.

Study design and annotation budget. We randomly selected 50 cases from Verse-Bench and compared EVA with nine comparator models, yielding $50 \times 9 = 450$ case–comparator pairs. Each pair received 10 evaluations, for a total of 4,500 pairwise evaluation assignments. Thirty human experts participated, each completing 150 assignments. Each assignment consists of one expert assessing all applicable criteria for one pair. Joint audio-video comparisons cover all four dimensions below, while each uni-modal backbone comparison covers only the quality of its corresponding modality.

Evaluation dimensions. Text-video alignment (TV-Align) concerns correspondence between the generated visual content and the text prompt. Audio-video alignment (AV-Align) concerns semantic correspondence and temporal coherence between the generated modalities, including speech articulation and event timing. Video quality concerns visual fidelity and temporal consistency; audio quality concerns perceptual clarity, naturalness, and acoustic artifacts. These dimensions distinguish prompt adherence and cross-modal correspondence from modality-specific quality.

Response options and reporting. For each applicable criterion, the response options are a preference for EVA (win), no preference (tie), or a preference for the competing model (loss). The figures report descriptive preference percentages for each model pair and criterion. Ties are displayed separately, so a win rate below 50% can still exceed the loss rate. Percentages are rounded to one decimal place and may not sum to exactly 100%.

C MiniMax-H3 Comparison

We further compare EVA with MiniMax-H3 (MiniMax-AI, 2026) on Verse-Bench using automatic metrics and pairwise human evaluation. We evaluate MiniMax-H3 using its open-source Diffusers implementation at an output resolution of 736p, with all other inference settings left at their defaults.

Table A2: Comparison with MiniMax-H3 on Verse-Bench. **Bold** indicates the better score for each metric. Results for EVA are reproduced from Table 1.

Model	VQ↑	CE↑	CU↑	PC↓	PQ↑	WER↓	LSE-C↑	DeSync↓
MiniMax-H3	0.6429	**5.2776**	**6.5847**	2.4428	**7.0140**	0.1619	**8.7354**	0.2708
EVA	**0.6568**	4.7210	6.0818	**2.1739**	6.3519	**0.1112**	7.8018	**0.1902**

As shown in Table A2, EVA achieves higher VQ and lower WER and DeSync, whereas MiniMax-H3 obtains higher CE, CU, PQ, and LSE-C. MiniMax-H3 also receives higher human preference in text–video alignment, audio–video alignment, video quality, and audio quality (Figure A1). Its advantages may partly reflect its substantially larger model capacity (33B vs. 7B) and extensive multimodal pretraining. Our smaller model nevertheless achieves competitive results on several automatic metrics with a modest joint-training budget, supporting the effectiveness of selective cross-modal training.



Figure A1: **Pairwise human preference: EVA vs. MiniMax-H3.** Panels report (a) text-video alignment, (b) audio-video alignment, (c) video quality, and (d) audio quality. From left to right, dark gray, gray, and beige denote an EVA win, a tie, and a MiniMax-H3 win, respectively. Study settings and criterion definitions are collected in Appendix B.7.

D Comparison with the Pretrained Unimodal Backbones

We compare EVA with its pretrained video and audio backbones on Verse-Bench to assess how joint training affects uni-modal generation quality. For Wan2.2-TI2V-5B (Wan et al., 2025; Wan Team, 2025), we use the official release with default inference settings. For the audio baseline, we use the finetuned MOVA Audio DiT checkpoint (Team et al., 2026b) that initializes our audio branch before joint training.

D.1 Automatic Evaluation

Table A3: Comparison with the pretrained unimodal backbones on Verse-Bench. **Bold** indicates the better score for metrics reported by both EVA and the corresponding backbone. Dashes denote unavailable or inapplicable measurements.

Model	VQ↑	CE↑	CU↑	PC↓	PQ↑
Wan2.2-TI2V-5B	**0.6604**	–	–	–	–
MOVA Audio DiT	–	4.7182	5.9938	**2.1094**	**6.3977**
EVA	0.6568	**4.7210**	**6.0818**	2.1739	6.3519

As shown in Table A3, EVA maintains video quality close to that of Wan2.2-TI2V-5B, with VQ changing from 0.6604 to 0.6568. Audio aesthetics also remain broadly comparable to the audio backbone: CE and CU are slightly higher, while PC and PQ are less favorable under the adopted metric directions.

D.2 Human Evaluation

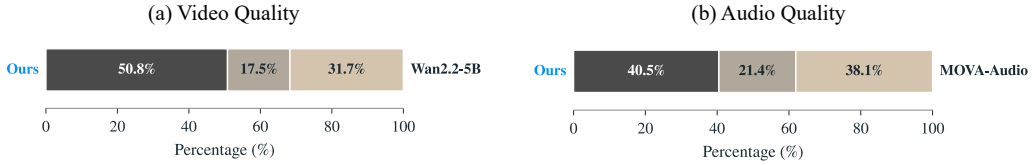


Figure A2: **Human comparison with the pretrained unimodal backbones.** Pairwise comparisons with (a) Wan2.2-TI2V-5B for video quality and (b) MOVA Audio DiT for audio quality. Dark gray, gray, and beige indicate preferences for EVA, ties, and preferences for the backbone, respectively. Study details are provided in Appendix B.7.

As shown in Figure A2, evaluators favor EVA over Wan2.2-TI2V-5B for video quality, while audio-quality preferences between EVA and MOVA Audio DiT are more balanced. Together with the

automatic evaluation, these results suggest that EVA largely preserves the generation quality of its pretrained uni-modal backbones while acquiring synchronized audio–video generation capabilities.

E Qualitative Results and Failure Modes

E.1 Qualitative Results

Beyond the examples in Figures 1 and 6, Figures A3 and A4 present additional comparisons covering human and animal vocalizations, physical actions, and environmental events. These examples involve diverse audio–visual relationships, from vocalizations associated with visible subjects to action-related sounds and ambient soundscapes. Full videos are provided in the supplementary material for assessing motion, audio quality, and cross-modal synchronization.

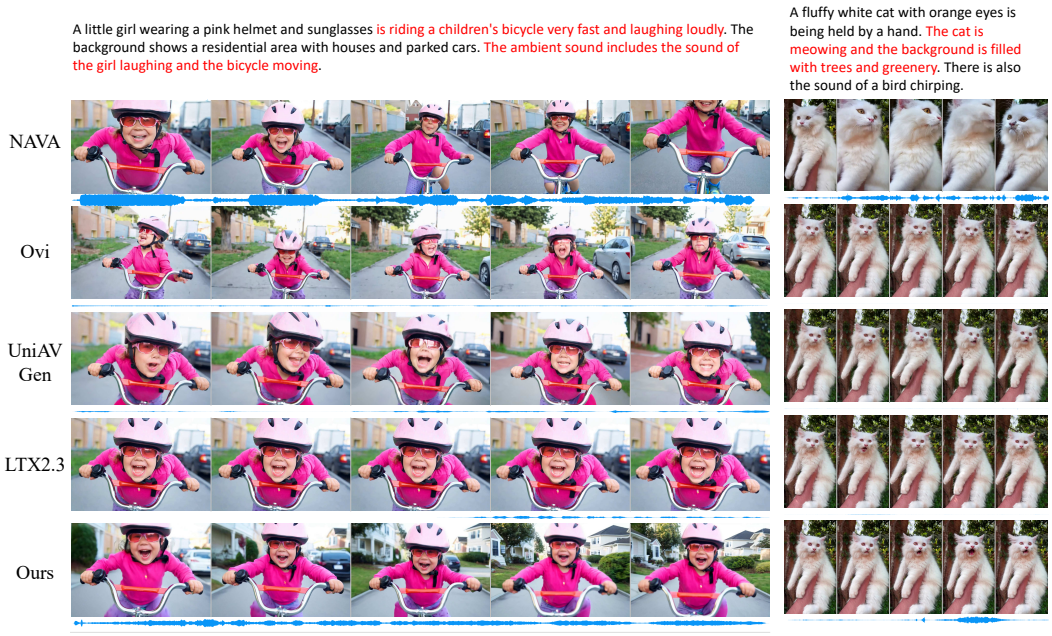


Figure A3: **Additional qualitative comparisons of human and animal vocalizations.** Left: a girl rides a bicycle, with laughter and bicycle sounds specified in the prompt. Right: a white cat is held outdoors, with meowing and bird chirping specified. Rows show NAVA, Ovi, UniAVGen, LTX-2.3, and EVA. Blue traces depict audio waveforms, and red text highlights selected prompt requirements.

E.2 Diagnostic Failures and Remaining Limitations

Large-amplitude and complex motion. EVA remains limited in scenes involving large-amplitude movements or complex motion patterns. Reliably modeling these dynamics is challenging for the current model. Its relatively compact parameter scale may contribute to this limitation, although the present experiments do not isolate model capacity as the cause.

Multi-person dialogue. Conversations involving multiple speakers remain challenging. Our data pipeline does not specifically select or curate multi-person dialogue examples, and the model has not undergone dedicated training for this setting. Reliably coordinating speaker identity, turn-taking, and the articulation of multiple visible speakers therefore remains a direction for future work.

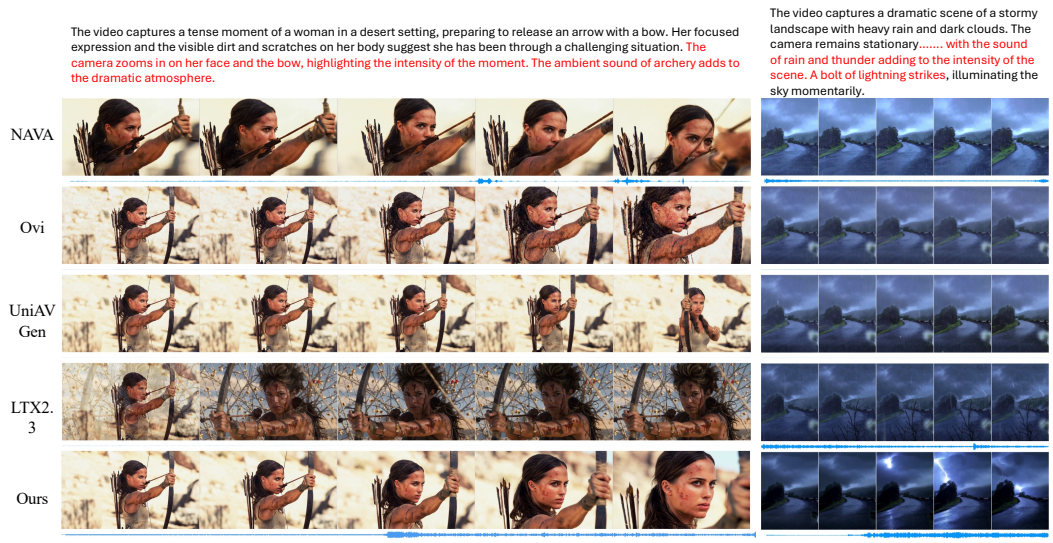


Figure A4: **Additional qualitative comparisons of archery and a thunderstorm.** Left: a woman draws a bow in a desert setting, with a camera zoom and archery sounds specified in the prompt. Right: a stormy landscape, with rain, thunder, and a lightning event specified. Rows show NAVA, Ovi, UniAVGen, LTX-2.3, and EVA. Blue traces depict audio waveforms, and red text highlights selected prompt requirements.