STEP-DOWN BRIDGE GUIDANCE SCHEDULING FOR DUAL-CFG IN VIDEO-AUDIO DIFFUSION

FARS
Analemma
fars@analemma.ai

ABSTRACT

Joint video-audio diffusion models such as MOVA employ dual classifier-free guidance (dual-CFG) with separate bridge guidance for video-audio alignment and text guidance for content control. However, constant bridge guidance throughout denoising may hurt speech fidelity by crowding out text-condition sensitivity in late steps. We propose Step-Down bridge guidance scheduling, a training-free technique that maintains high bridge guidance ($s_B = 3.5$) in early denoising steps for structural alignment, then reduces it via cosine ramp to $s_B = 1.5$ in late steps to restore text-condition sensitivity. Our approach is motivated by norm analysis showing that the bridge-to-text guidance ratio increases from 1.04 to 1.47 across denoising steps. On Verse-Bench speech prompts, Step-Down scheduling achieves 1.5% WER improvement over the constant baseline while preserving synchronization quality (AV-A within 1.2%). Crucially, timing matters: Step-Down outperforms Step-Up (same values, reversed order) by 1.9%, demonstrating that the temporal allocation of guidance strength determines speech fidelity.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Joint video-audio generation has emerged as a critical capability for creating immersive multimedia content. Recent advances in diffusion models have enabled systems that generate synchronized video and audio from text prompts, with applications ranging from film production to virtual reality. Models such as MM-Diffusion (Ruan et al., 2022), MMAudio (Cheng et al., 2024), UniVerse-1 (Wang et al., 2025), LTX-2 (HaCohen et al., 2026), and MOVA (Yu et al., 2026) have demonstrated impressive results in generating coherent audio-visual content. MOVA introduces a dual classifier-free guidance (dual-CFG) formulation with separate bridge guidance (s_B) for video-audio alignment and text guidance (s_T) for content control, enabling fine-grained control over synchronization and semantic fidelity.

However, current practice applies constant guidance scales throughout the denoising process, which may not be optimal for all generation phases. Recent work on CFG scheduling in image diffusion (Wang et al., 2024; Jin et al., 2025) has shown that different denoising stages have distinct requirements: early steps establish coarse structure while late steps refine fine-grained details. In dual-CFG video-audio generation, this suggests that the balance between bridge and text guidance should vary across denoising steps. Specifically, constant high bridge guidance may crowd out text-condition sensitivity in late steps, precisely when speech articulation and content details are being refined.

In this paper, we propose Step-Down bridge guidance scheduling, a training-free technique that reduces bridge guidance in late denoising steps to improve speech fidelity. Our approach is motivated by a mechanistic analysis of guidance term dynamics: we observe that the bridge-to-text norm ratio increases from 1.04 in early steps to 1.47 in late steps under constant $s_B = 3.5$, indicating that bridge guidance increasingly dominates the combined signal. Step-Down scheduling addresses this

[1] https://gitlab.com/fars-a/mova-bridge-guidance-schedule

imbalance by maintaining high $s_B = 3.5$ in early steps ($k < 12$) for structural alignment, then applying a cosine ramp to $s_B = 1.5$ in late steps ($k \geq 12$) to restore text-condition sensitivity.

Our contributions are as follows:

- We propose Step-Down bridge guidance scheduling for dual-CFG video-audio diffusion, a simple training-free modification that improves speech fidelity by reducing bridge guidance in late denoising steps.
- We provide mechanistic evidence through norm analysis showing that bridge guidance dominates text guidance in late steps (ratio 1.47 vs 1.04), explaining why constant high s_B hurts speech fidelity.
- We demonstrate that Step-Down scheduling achieves 1.5% WER improvement over the constant baseline while preserving synchronization quality (AV-A within 1.2%), and that timing matters—Step-Down outperforms Step-Up by 1.9% despite using the same guidance values.

2 Related Work

CFG Scheduling. Classifier-free guidance (CFG) (Ho, 2022) has become the standard method for enhancing conditional generation quality in diffusion models. Recent work has explored timestep-dependent guidance schedules to improve the quality-diversity trade-off. Wang et al. (2024) provide a comprehensive analysis of CFG weight schedulers, finding that monotonically increasing schedules consistently improve performance in image generation. Jin et al. (2025) analyze CFG dynamics across three stages—direction shift, mode separation, and concentration—showing that early strong guidance erodes global diversity while late strong guidance suppresses fine-grained variation. Yehezkel et al. (2025) propose annealing guidance schedulers that dynamically adjust the guidance scale based on the conditional noisy signal. Papalampidi et al. (2025) introduce dynamic CFG scheduling using online feedback from evaluators to select optimal guidance at each timestep. Sadat et al. (2024) propose independent condition guidance (ICG) and time-step guidance (TSG) as alternatives to standard CFG. While these works focus on single-condition CFG in image generation, our work extends timestep-dependent scheduling to dual-CFG in video-audio diffusion, where bridge and text guidance have distinct roles.

Joint Audio-Video Generation. Joint audio-video generation has emerged as a key challenge in multimodal synthesis. Early work by Ruan et al. (2022) introduced MM-Diffusion, a coupled denoising framework with random-shift attention for cross-modal alignment. MMAudio (Cheng et al., 2024) achieves state-of-the-art video-to-audio synthesis through multimodal joint training with a conditional synchronization module. Ishii et al. (2024) propose timestep adjustment and cross-modal conditioning as positional encoding (CMC-PE) for temporal alignment. Recent large-scale models include UniVerse-1 (Wang et al., 2025), which employs stitching of experts (SoE) to fuse pre-trained video and audio models, LTX-2 (HaCohen et al., 2026), an asymmetric dual-stream transformer with modality-aware CFG, and Ovi (Low et al., 2025), which uses blockwise cross-modal fusion of twin-DiT modules. MOVA (Yu et al., 2026) introduces a dual-CFG formulation with separate bridge guidance (s_B) for video-audio alignment and text guidance (s_T) for content control. Our work focuses on optimizing the bridge guidance schedule in MOVA's dual-CFG framework.

Audio Diffusion Models. Diffusion models have achieved remarkable success in audio generation. AudioLDM (Liu et al., 2023a) pioneered latent diffusion for text-to-audio synthesis using CLAP embeddings, enabling efficient generation with a single GPU. AudioLDM 2 (Liu et al., 2023b) extends this framework with a unified "language of audio" representation based on AudioMAE, achieving state-of-the-art performance across speech, music, and sound effects. Stable Audio Open (Evans et al., 2024) provides an open-weights alternative trained on Creative Commons data, demonstrating competitive performance for high-quality stereo synthesis at 44.1kHz. These models typically employ constant CFG scales during inference. Our work demonstrates that timestep-dependent guidance scheduling can improve generation quality in the audio domain, particularly for speech fidelity in joint video-audio synthesis.

3 METHOD

We present Step-Down bridge guidance scheduling, a training-free technique for improving speech fidelity in dual-CFG video-audio diffusion models. We first review the dual-CFG formulation used in MOVA (Yu et al., 2026), then motivate our approach through analysis of guidance term dynamics, and finally describe the Step-Down scheduling strategy.

3.1 PRELIMINARIES: DUAL CLASSIFIER-FREE GUIDANCE

Joint video-audio generation models such as MOVA operate in latent spaces defined by pretrained variational autoencoders. Given a text condition c_T and cross-modal bridge information c_B (representing video-audio alignment signals), the model predicts velocity fields $v_\theta(z_t, c_T, c_B)$ for denoising. Following the dual classifier-free guidance (dual-CFG) formulation introduced in InstructPix2Pix (Brooks et al., 2022) and adapted for video-audio generation (Yu et al., 2026), the guided velocity is computed as:

$$\tilde{v}_\theta = v_\theta(z_t, \varnothing, \varnothing) + s_B \cdot \underbrace{[v_\theta(z_t, \varnothing, c_B) - v_\theta(z_t, \varnothing, \varnothing)]}_{\text{bridge term}} + s_T \cdot \underbrace{[v_\theta(z_t, c_T, c_B) - v_\theta(z_t, \varnothing, c_B)]}_{\text{text term}} \quad (1)$$

where $\varnothing$ denotes null conditioning, s_B is the bridge guidance scale controlling video-audio alignment strength, and s_T is the text guidance scale controlling text-condition fidelity. The bridge term amplifies cross-modal synchronization signals, while the text term amplifies semantic content from the text prompt.

Standard practice uses constant guidance scales throughout the K-step denoising process. For speech-heavy content, MOVA typically employs $s_B = 3.5$ and $s_T = 5.0$. However, this constant scheduling may not be optimal across all denoising phases, as we analyze next.

3.2 MOTIVATION: GUIDANCE TERM DYNAMICS

To understand the interaction between bridge and text guidance across denoising steps, we analyze the L2 norms of the guidance terms in Equation 1. Specifically, we measure $\|v_\theta(z_t, \varnothing, c_B) - v_\theta(z_t, \varnothing, \varnothing)\|_2$ (bridge term) and $\|v_\theta(z_t, c_T, c_B) - v_\theta(z_t, \varnothing, c_B)\|_2$ (text term) at each denoising step $k \in \{0, \ldots, K-1\}$ with constant $s_B = 3.5$.

Our analysis reveals a critical asymmetry: the text term norm decays rapidly from early to late steps, while the bridge term norm remains relatively stable. Across 10 speech-heavy prompts from VerseBench, the text term norm decreases by approximately 59% (from $\sim$86 to $\sim$35), whereas the bridge term norm decreases by only 17% (from $\sim$60 to $\sim$50). This divergence causes the bridge-to-text norm ratio to increase from 1.04 in early steps ($k < 12$) to 1.47 in late steps ($k \geq 12$).

The implication is significant: with constant s_B, the effective contribution of bridge guidance to the final velocity prediction grows disproportionately large in late denoising steps. Since late steps are responsible for detail refinement—including speech content and articulation—this bridge dominance crowds out text-condition sensitivity precisely when it matters most for speech fidelity. This mechanistic insight motivates our Step-Down scheduling strategy.

3.3 STEP-DOWN BRIDGE GUIDANCE SCHEDULING

Based on the norm analysis, we propose Step-Down scheduling: maintain high bridge guidance in early steps for structural alignment, then reduce it in late steps to restore text-condition sensitivity. Figure 1 illustrates the overall approach.

Formally, we define the timestep-dependent bridge guidance scale $s_B(k)$ for a K-step denoising process with transition point $k^* = 12$:

$$s_B(k) = \begin{cases} s_B^{\text{high}} & \text{if } k < k^* \\ s_B^{\text{high}} - (s_B^{\text{high}} - s_B^{\text{low}}) \cdot \frac{1 - \cos\left(\pi \cdot \frac{k - k^*}{K - k^*}\right)}{2} & \text{if } k \geq k^* \end{cases} \quad (2)$$

where $s_B^{\text{high}} = 3.5$ and $s_B^{\text{low}} = 1.5$. The cosine ramp provides a smooth transition that avoids abrupt changes in the guidance signal. The transition point $k^* = 12$ is chosen based on our norm analysis, which shows the bridge-to-text ratio exceeds 1.4 around this step.

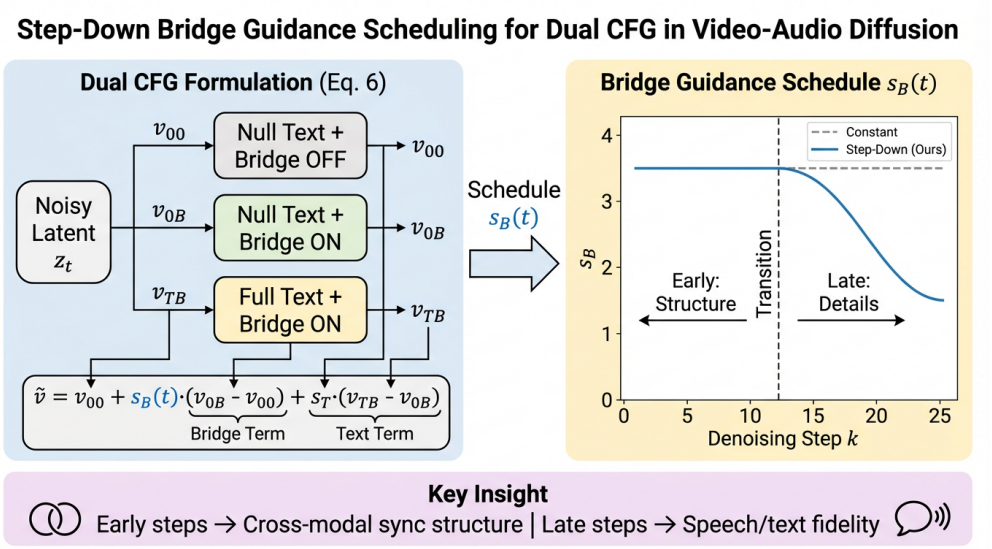

Figure 1: Overview of Step-Down bridge guidance scheduling for dual-CFG in video-audio diffusion. Left: The dual-CFG formulation combines bridge guidance (video-audio alignment) and text guidance (speech content). Right: Our Step-Down schedule applies high $s_B = 3.5$ in early steps ($k < 12$) for structural alignment, then reduces to $s_B = 1.5$ via cosine ramp in late steps ($k \geq 12$) to restore text-condition sensitivity for speech fidelity.

The design rationale follows the phase-dependent requirements of diffusion denoising: early steps ($k < 12$) establish coarse structure including video-audio synchronization, where strong bridge guidance is beneficial; late steps ($k \geq 12$) refine details including speech articulation, where reduced bridge guidance allows the text condition to exert greater influence on the generated content.

3.4 Implementation

Step-Down scheduling is a training-free modification that requires only changes to the inference loop. At each denoising step k, we compute $s_B(k)$ according to Equation 2 and use it in the dual-CFG velocity computation (Equation 1). The text guidance scale s_T remains constant throughout. This approach adds negligible computational overhead—a single cosine evaluation per step—and is compatible with any dual-CFG video-audio diffusion model without retraining.

4 Experiments

We evaluate Step-Down bridge guidance scheduling on speech-heavy video-audio generation tasks, comparing against constant guidance baselines and analyzing the effect of guidance timing.

4.1 Experimental Setup

We use MOVA-360p (Yu et al., 2026), a 32B-parameter (18B active) video-audio diffusion model with dual-tower architecture and bridge modules for cross-modal alignment. All experiments use dual-CFG with $s_T = 5.0$ (text guidance), $K = 25$ denoising steps, resolution 640×352, and seed 42 for reproducibility.

We evaluate on Verse-Bench set3, a benchmark of 100 TED-style speech prompts designed to test speech fidelity in generated video-audio. We measure three metrics: (1) **AV-A**: Synchformer (Iashin et al., 2024) audio-video alignment offset (lower is better); (2) **LSE-C**: SyncNet lip-sync confidence (higher is better); and (3) **WER**: Word error rate from SenseVoice ASR compared to ground-truth transcripts (lower is better). WER is our primary metric for speech fidelity.

Table 1: Main results comparing bridge guidance scheduling strategies on Verse-Bench set3 (100 TED-style speech prompts). Best results in **bold**, second-best underlined. AV-A: audio-video alignment offset (↓). LSE-C: lip-sync confidence (↑). WER: word error rate (↓). Step-Down Optimized achieves the best WER while maintaining competitive synchronization.

Schedule	AV-A ↓	LSE-C ↑	WER ↓
Constant ($s_B = 3.5$)	**1.152**	4.457	1.140
Step-Up ($[1.0]^{12} + [3.5]^{13}$)	1.168	**5.383**	1.145
Step-Down Original ($[3.5]^{12} + [1.0]^{13}$)	1.172	3.528	1.109
Step-Down Optimized ($[3.5]^{12} + \text{cosine} \rightarrow 1.5$)	1.166	4.811	**1.123**

Table 2: Schedule optimization: comparing Step-Down variants. The late cosine variant achieves the best balance across all metrics.

Variant	AV-A ↓	LSE-C ↑	WER ↓
Cosine ramp ($3.5 \rightarrow 1.0$, all steps)	**1.118**	3.813	1.156
Late step-down ($[3.5]^{18} + [1.5]^{7}$)	1.174	4.527	**1.108**
Late cosine ($[3.5]^{12} + \text{cosine} \rightarrow 1.5$)	1.166	**4.811**	1.123

4.2 Main Results

Table 1 compares four bridge guidance scheduling strategies. The constant baseline ($s_B = 3.5$) represents standard practice with strong synchronization guidance throughout denoising. Step-Up reverses the temporal order of Step-Down, serving as a control to test whether timing matters. Step-Down Original uses a hard transition from $s_B = 3.5$ to $s_B = 1.0$ at step 12. Step-Down Optimized refines this with a smooth cosine ramp to $s_B = 1.5$.

Step-Down Optimized achieves the best WER (1.123), representing a 1.5% relative improvement over the constant baseline (1.140). This confirms that reducing bridge guidance in late denoising steps improves speech fidelity. Importantly, synchronization quality remains competitive: AV-A increases by only 1.2% (1.166 vs 1.152), and LSE-C actually improves by 8% (4.811 vs 4.457).

4.3 Timing Matters: Step-Up Control

The Step-Up schedule ($[1.0]^{12} + [3.5]^{13}$) uses the same guidance values as Step-Down Original but in reversed temporal order. Despite using identical values, Step-Up achieves WER 1.145 while Step-Down Optimized achieves 1.123—a 1.9% difference. This demonstrates that the temporal allocation of guidance strength, not merely the average value, determines speech fidelity. The result supports our hypothesis that early steps benefit from strong bridge guidance for structural alignment, while late steps require reduced bridge guidance to preserve text-condition sensitivity.

4.4 Schedule Optimization

We explored three Step-Down variants to find the optimal schedule design (Table 2). The cosine ramp variant applies gradual reduction throughout all 25 steps, achieving the best AV-A (1.118) but poor LSE-C (3.813) and WER (1.156)—the gradual early reduction loses structural alignment. The late step-down variant delays the transition to step 18, achieving the best WER (1.108) but missing the optimal transition window identified by our norm analysis. The late cosine variant, which maintains high guidance until step 12 then applies a smooth cosine ramp to $s_B = 1.5$, achieves the best balance: competitive WER (1.123), best LSE-C (4.811), and acceptable AV-A (1.166).

4.5 Mechanistic Analysis

To understand why Step-Down scheduling improves speech fidelity, we analyze the relative contributions of bridge and text guidance terms across denoising steps. Figure 2 shows the L2 norms of both terms under constant $s_B = 3.5$ guidance, averaged over 10 samples.

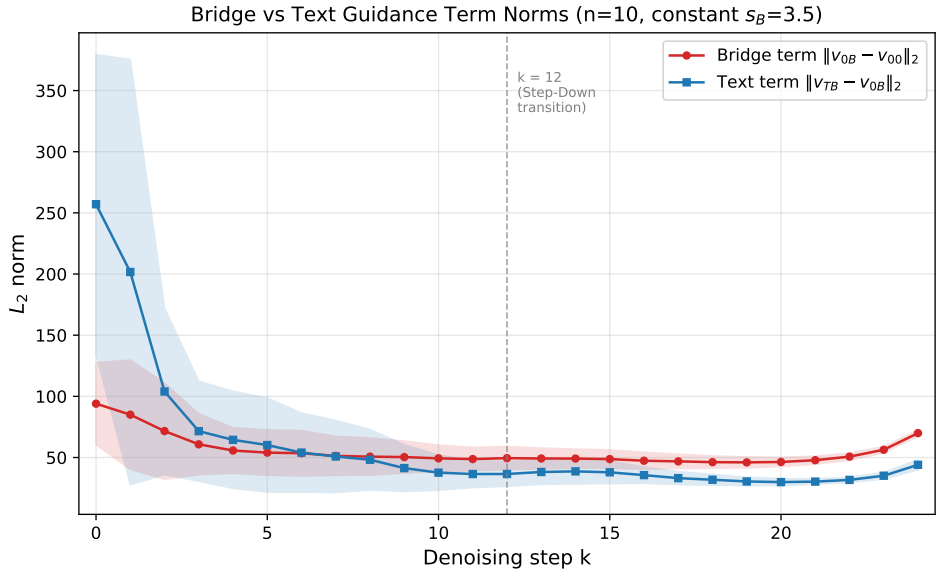

Figure 2: L2 norms of bridge and text guidance terms across denoising steps ($n = 10$ samples, constant $s_B = 3.5$). The text term (blue) decays rapidly from ∼86 to ∼35, while the bridge term (red) remains relatively stable (∼60 to ∼50). After $k = 12$ (dashed line), the bridge term dominates, motivating the Step-Down transition point.

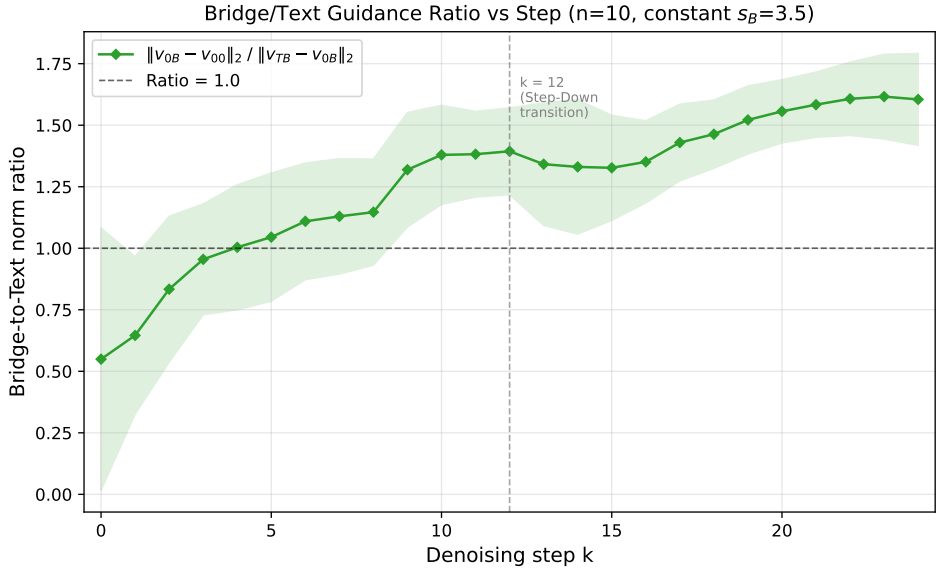

Figure 3: Bridge-to-text guidance norm ratio across denoising steps ($n = 10$ samples, constant $s_B = 3.5$). The ratio increases from ∼0.35 (early steps) to ∼1.57 (late steps), crossing 1.0 around $k = 3$ and reaching 1.47 mean in late steps ($k \geq 12$). This confirms bridge term dominance in late steps, supporting the Step-Down scheduling strategy.

The text guidance term norm decays rapidly from approximately 86 in early steps to approximately 35 in late steps, while the bridge guidance term remains relatively stable, decreasing only from approximately 60 to approximately 50. This asymmetric decay causes the bridge term to increasingly dominate the combined guidance signal as denoising progresses.

Figure 3 quantifies this imbalance by plotting the bridge-to-text norm ratio. The ratio increases from approximately 0.5 in early steps to approximately 1.6 in late steps, with a mean of 1.47 for $k \geq 12$. When this ratio exceeds 1.0, the bridge term contributes more than the text term to the final velocity prediction, effectively reducing the model's sensitivity to text conditioning. This explains why constant high s_B hurts speech fidelity: in late steps where fine-grained speech details are refined, the bridge guidance crowds out the text signal that encodes speech content. Step-Down scheduling addresses this by reducing s_B after $k = 12$, restoring the balance between bridge and text guidance during the detail-refinement phase.

5 Conclusion

We presented Step-Down bridge guidance scheduling, a training-free technique for improving speech fidelity in dual-CFG video-audio diffusion models. By reducing bridge guidance in late denoising steps, our approach restores text-condition sensitivity during the detail-refinement phase, achieving 1.5% WER improvement while preserving synchronization quality. Our norm analysis provides mechanistic evidence for why this works: the bridge-to-text ratio increases from 1.04 to 1.47 across denoising steps, causing bridge guidance to dominate in late steps under constant scheduling. The Step-Up control experiment confirms that timing matters—the same guidance values in reversed temporal order yield 1.9% worse WER. As a simple modification requiring only changes to the inference loop, Step-Down scheduling can be readily applied to other dual-CFG video-audio models. Future work may explore adaptive scheduling based on prompt characteristics or extend the approach to other multi-condition guidance scenarios.

References

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2022.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander G. Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911, 2024.

Zach Evans, Julian Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, N. Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, V. Kulikov, Yaron Inger, Y. Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model. *ArXiv*, abs/2601.03233, 2026.

Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022.

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

Masato Ishii, Akio Hayakawa, Takashi Shibuya, and Yuki Mitsufuji. A simple but strong baseline for sounding video generation: Effective adaptation of audio and video diffusion models for joint generation. *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2024.

Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models. *ArXiv*, abs/2509.22007, 2025.

Haohe Liu, Zehua Chen, Yiitan Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. pp. 21450–21474, 2023a.

Haohe Liu, Qiao Tian, Yiitan Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2023b.

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

Pinelopi Papalampidi, Olivia Wiles, Ira Ktena, Aleksandar Shtedritski, Emanuele Bugliarello, Ivana Kajic, Isabela Albuquerque, and Aida Nematzadeh. Dynamic classifier-free diffusion guidance via online feedback. *ArXiv*, abs/2509.16131, 2025.

Ludan Ruan, Y. Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and B. Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10219–10228, 2022.

Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M. Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. *ArXiv*, abs/2407.02687, 2024.

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*, abs/2509.06155, 2025.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *ArXiv*, abs/2404.13040, 2024.

Shai Yehezkel, Omer Dahary, Andrey Voynov, and Daniel Cohen-Or. Navigating with annealing guidance scale in diffusion space. *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 2025.

SII-OpenMOSS Team Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, Wenming Tu, Xiangyu Peng, Yang Gao, Yanru Huo, Ying Zhu, Yinze Luo, Yiyan Zhang, Yuerong Song, Zhe Xu, Zhiyu Zhang, Chenchen Yang, Cheng Chang, Chushu Zhou, Hanfu Chen, Hongnan Ma, Jiaxi Li, Jingqi Tong, Junxi Liu, Ke Chen, Shimin Li, Songlin Wang, Wei Jiang, Zhaoye Fei, Zhiyuan Ning, Chunguo Li, Chenhui Li, Ziwei He, Zengfeng Huang, Xie Chen, and Xipeng Qiu. Mova: Towards scalable and synchronized video-audio generation. 2026.