OmniForcing: Unleashing Real-time Joint Audio-Visual Generation

Yaofeng Su[1,2*], Yuming Li[3*], Zeyue Xue[1,4], Jie Huang[1], Siming Fu[1], Haoran Li[1], Haoyang Huang[1], and Nan Duan[1]

[1] Joy Future Academy, JD, Beijing, China
[2] Fudan University, Shanghai, China
[3] Peking University, Beijing, China
[4] The University of Hong Kong, Hong Kong, China

Abstract. Recent joint audio-visual diffusion models achieve remarkable quality but suffer from high latency due to their bidirectional attention. We present OmniForcing, the first framework to distill an offline, dual-stream bidirectional audio-visual diffusion model (LTX-2) into a high-fidelity streaming autoregressive generator. Naive causal distillation here is severely unstable, owing to the extreme temporal asymmetry between modalities (3 vs. 25 latent frames/s) and the resulting audio-token sparsity. We bridge the information-density gap with an Asymmetric Block-Causal Alignment and a zero-truncation Global Prefix, and eliminate the gradient explosion from Softmax collapse on sparse audio blocks via Audio Sink Tokens with an Identity RoPE constraint. A three-stage pipeline, culminating in Joint Self-Forcing, then corrects cross-modal exposure bias over long rollouts. With a modality-independent rolling KV-cache, OmniForcing streams at ∼25 FPS on a single GPU, on par with the synchronization and visual quality of its bidirectional teacher. **Project Page:** https://omniforcing.com.

Keywords: Streaming Audio-Visual Generation · Diffusion Distillation · Autoregressive Video Synthesis

1 Introduction

Built on Diffusion Transformers (DiTs) [15,18], joint audio-visual models such as LTX-2 [3] and Veo 3 [19] recently achieved notable progress, leveraging modality-specific VAEs [3, 10, 12] to jointly model video and audio in continuous latent spaces. However, they rely on bidirectional full-sequence attention: the entire physical timeline is processed at once, incurring a high Time-To-First-Chunk (TTFC) latency [23] that precludes real-time or streaming use (Fig. 1).

Prior remedies follow two lines. Cascaded pipelines synthesize one modality after the other [1, 6, 13, 14, 17], but severing the joint distribution obstructs streaming. Causal autoregressive frameworks that adapt video-only diffusion

[*] Equal contribution. This paper is an extended abstract; the full version of this work has been accepted at the ECCV 2026 main conference.

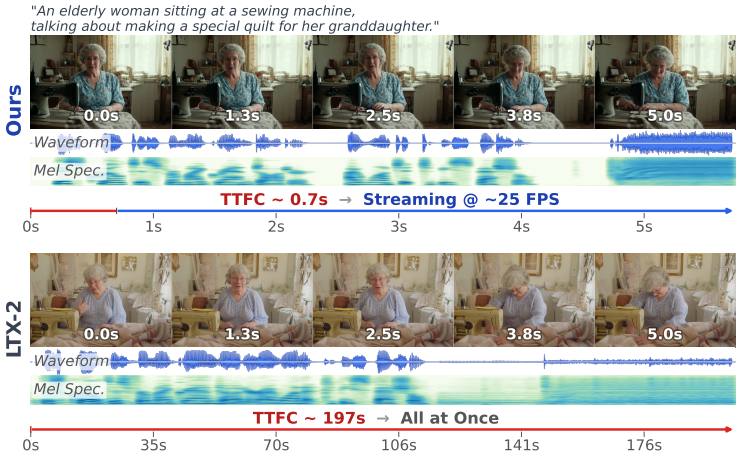

Fig. 1: OmniForcing streams joint audio-visual content at ∼25 FPS with a ∼0.7 s TTFC, versus ∼197 s for the offline bidirectional teacher (LTX-2), at comparable fidelity.

(*e.g.*, CausVid [23], Self-Forcing [4]) remain single-modality; extending them to dual-stream models is non-trivial, since the temporal asymmetry starves the sparser modality [7] and destabilizes training.

We present **OmniForcing**, the first framework to distill a heavy bidirectional audio-visual foundation model (LTX-2 [3]) into a high-fidelity streaming autoregressive generator that, by dynamically interleaving audio and video chunks, streams at ultra-low latency without sacrificing the teacher's holistic joint distribution. Our contributions are threefold: a unified streaming autoregressive framework for joint audio-visual generation; an **Asymmetric Block-Causal Alignment** and **Audio Sink Tokens** with Identity RoPE that resolve the Softmax collapse from multi-modal token-density mismatch; and a **Joint Self-Forcing** paradigm with a **Modality-Independent Rolling KV-Cache** that mitigates exposure bias and reaches ∼25 FPS on a single GPU.

2 OmniForcing

2.1 Problem Formulation

Given a text prompt c, we stream a temporally aligned video $\mathbf{V}$ and audio $\mathbf{A}$ in independent latent spaces via modality-specific VAEs [3, 10, 12]. Following CausVid [23] and Self-Forcing [4], we restructure the pretrained bidirectional dual-stream transformer (LTX-2 [3]) into a **block-causal autoregressive** model over $K{+}1$ synchronized blocks $\{\mathcal{B}_0, \ldots, \mathcal{B}_K\}$ ($K =$ physical seconds):

$$p(\mathbf{V}, \mathbf{A} \mid c) = p(\mathcal{B}_0 \mid c) \prod_{k=1}^{K} p(\mathcal{B}_k \mid \mathcal{B}_{<k}, c). \tag{1}$$

This retrofit must overcome the 3-vs-25 latent-frame-rate asymmetry, Softmax collapse from sparse causal audio history [7], and cross-modal exposure bias [4], which we tackle with block-causal masking and a three-stage pipeline.

2.2 Asymmetric Block-Causal Alignment and Mask Design

Block alignment. To factorize at the block level, we bridge the modalities' information-density gap: their VAEs use highly asymmetric temporal downsampling [3], giving f_v=3 video vs. f_a=25 audio latent frames per second in LTX-2. A frame-by-frame causal mask over this 25:3 non-integer ratio would truncate destructively, so we adopt a physical-time **Macro-block Alignment**: a one-second window holds exactly ΔN_v=3 video and ΔN_a=25 audio latents with no remainder. This matches the VAEs' causal-convolution stride (stride 1 for the first frame, full strides thereafter [10]), so the latent lengths are N_v=1+Kf_v and N_a=1+Kf_a. The constant-1 initial latents $\mathbf{V}_0, \mathbf{A}_0$ cannot fit into subsequent blocks, so we merge them into a **Global Prefix** $\mathcal{B}_0$ whose attention is unconditionally bidirectional and, like a system prompt, globally visible to all future tokens—a zero-truncation cross-modal anchor.

Causal mask. Let $\tau(q)$ be the physical block index of token q, from a floor division over the per-modality block sizes ($|\mathcal{B}^v|$=$f_v H_v W_v$ video tokens, as each frame is patchified into $H_v W_v$ tokens; $|\mathcal{B}^a|$=f_a audio tokens, one per frame). Enforcing strict causality with intra-block bidirectional flow, we define the four-pathway masks:

$$\mathbf{M}^{V\to V}_{q,k} = \mathbb{I}(\tau_v(k) \le \tau_v(q)), \qquad \mathbf{M}^{A\to A}_{q,k} = \mathbb{I}(\tau_a(k) \le \tau_a(q)), \tag{2}$$

$$\mathbf{M}^{V\to A}_{q,k} = \mathbb{I}(\tau_a(k) \le \tau_v(q)), \qquad \mathbf{M}^{A\to V}_{q,k} = \mathbb{I}(\tau_v(k) \le \tau_a(q)), \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator. The Global Prefix tokens (τ=0) satisfy $\tau(k) \le \tau(q)$ for all queries, staying globally visible; thus both modalities' receptive fields expand synchronously at physical block boundaries despite the token-density mismatch.

2.3 Three-Stage Distillation and Streaming Inference

We inject few-step denoising and causal generation through a three-stage pipeline, then deploy with asymmetric parallel inference.

Stage I: Bidirectional DMD. We first distill the pretrained model into a few-step bidirectional student via DMD [21, 22] ($\mathcal{L}_{\text{Bi-DMD}}$=$\lambda_v \mathcal{L}^v_{\text{DMD}}$+$\lambda_a \mathcal{L}^a_{\text{DMD}}$), preserving the global receptive field to provide an easily regressible teacher trajectory for the causal migration.

Stage II: Causal ODE Regression. We equip the model with the masks of Sec. 2.2 and regress the Stage-I ODE trajectories to adapt the weights. The resulting *conditional distribution shift* (globally informed → truncated causal) is asymmetric: a video block holds $f_v H_v W_v$=3×384=1,152 tokens, an audio block only 25. For early audio tokens the history is minuscule, so the Softmax collapses to a near-one-hot vector and minor logit perturbations explode through the exponential, surging the gradient into NaN under bf16. We stabilize this

by prepending S learnable **Audio Sink Tokens** [2, 20] to the audio stream, anchored in $\mathcal{B}_0$ with an *Identity RoPE* constraint ($\cos\theta_{\text{sink}}=\mathbf{1}, \sin\theta_{\text{sink}}=\mathbf{0}$, degenerating RoPE [16] to identity so they stay position-agnostic); acting as a soft memory buffer, they widen the attention denominator from i to $i+S$, breaking the collapse and restoring entropy.

Stage III: Causal DMD with Joint Self-Forcing. To combat exposure bias, we adopt a joint Self-Forcing [4] paradigm: the model unrolls autoregressively during training, appending its noise-free KV embeddings to a rolling cache, and the generated trajectory is scored against the frozen teacher via the DMD objective [21]:

$$\mathcal{L}_{\text{SF}} = \sum_{k=1}^{K} \mathbb{E}_{\hat{\mathcal{B}}_{<k},\, t,\, \boldsymbol{\epsilon}} \left[\lambda_v \left\| \nabla_{\text{KL}}^{v,k} \right\|_2^2 + \lambda_a \left\| \nabla_{\text{KL}}^{a,k} \right\|_2^2 \right], \tag{4}$$

where $\nabla_{\text{KL}}^{m,k}=s_{\text{gen}}^{m}-s_{\text{CFG}}^{m}$ is the per-modality ($m \in \{v, a\}$) score difference between critic and CFG-weighted teacher. This coupled unrolling forces the streams to adapt to each other's drift, ensuring cross-modal synchrony. At inference, since the two self-attention streams share no data dependency, a Modality-Independent Rolling KV-Cache ($\mathcal{O}(L)$ context) runs them concurrently at ∼25 FPS on a single GPU.

3 Experiments

Implementation. We build OmniForcing on LTX-2 [3] (14B video + 5B audio) on 32 GPUs (bf16, batch 32, lr 2×10^{-5}): Stages I/III for 2,000 steps each and Stage II for 3,000, with S=16 Audio Sink Tokens and CFG w_v=3, w_a=5. Data combine Mixkit clips with Open-Sora-Plan [11] and AudioCaps [9] captions rewritten by Gemma 3 12B [8]; distilling from a pretrained teacher, this compact set suffices. At inference, streaming decoding is strictly causal: a sliding-window video decoder (W=2 past blocks, 47.4 dB PSNR) and a lossless cumulative audio decoder ($\leq$ 14 ms) run concurrently for real-time output.

Distillation fidelity. We validate visual fidelity with VBench [5] against the teacher on identical prompts (Tab. 1). OmniForcing slightly *surpasses* the teacher on per-frame quality (aesthetic +0.026, imaging +0.020, subject consistency +0.010) while preserving motion smoothness and temporal flickering, echoing prior distillation findings [4, 21]. Critically, it streams a 5 s clip in ∼5.7 s (sustained ∼25 FPS, ∼0.7 s TTFC)—a ∼35× speedup over the offline teacher's 197 s and the only true-streaming method here.

Table 1: Distillation fidelity vs. the bidirectional LTX-2 teacher (same prompts; VBench [5]). Best in **bold** (↑/↓: higher/lower better).

	VBench Visual Quality					Streaming	
Model	Aesthetic Quality ↑	Imaging Quality ↑	Motion Smoothness ↑	Subject Consistency ↑	Temporal Flickering ↑	TTFC ↓	FPS ↑
LTX-2 [3]	0.569	0.574	0.993	0.945	0.988	197.0s	–
OmniForcing	**0.595**	**0.594**	**0.995**	**0.955**	**0.989**	**0.7s**	**25**

References

1. Cheng, H.K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., Mitsufuji, Y.: Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28901–28911 (2025)
2. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024)
3. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026)
4. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
5. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
6. Jeong, Y., Ryoo, W., Lee, S., Seo, D., Byeon, W., Kim, S., Kim, J.: The power of sound (tpos): Audio reactive video generation with stable diffusion. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7822–7832 (2023)
7. Jia, D., Chen, Z., Chen, J., Du, C., Wu, J., Cong, J., Zhuang, X., Li, C., Wei, Z., Wang, Y., et al.: Ditar: Diffusion transformer autoregressive modeling for speech generation. In: International Conference on Machine Learning. pp. 27255–27270. PMLR (2025)
8. Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
9. Kim, C.D., Kim, B., Lee, H., Kim, G.: 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)
10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
11. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024)
12. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: Audioldm: Text-to-audio generation with latent diffusion models. In: International Conference on Machine Learning. pp. 21450–21474. PMLR (2023)
13. Luo, S., Yan, C., Hu, C., Zhao, H.: Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. Advances in Neural Information Processing Systems **36**, 48855–48876 (2023)
14. Mei, X., Nagaraja, V., Le Lan, G., Ni, Z., Chang, E., Shi, Y., Chandra, V.: Foleygen: Visually-guided audio generation. In: 2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2024)
15. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (October 2023)

16. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
17. Sung-Bin, K., Senocak, A., Ha, H., Owens, A., Oh, T.H.: Sound to visual scene generation by audio-to-visual latent alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2023)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
19. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
20. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: The Twelfth International Conference on Learning Representations (2024)
21. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)
22. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
23. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22963–22974 (2025)