FlowLong: Training-free Long Audio-Video Generation via Manifold-constrained Tweedie Matching

Jangho Park[1,⋆], Geon Yeong Park[1,⋆], Gihyun Kwon[2,†], and Jong Chul Ye[1,†]

[1]KAIST [2]Amazon

Abstract. Extending audio-video generation beyond a model's native window is an open problem: bidirectional extensions are backbone-specific and degrade over long horizons, while autoregressive models drift from exposure bias and require a bidirectional teacher for distillation—which recent joint audio-video generators lack. We present *FlowLong*, a training-free, architecture-agnostic, inference-time framework that harmonizes overlapping windows sampled in parallel via *Tweedie matching*, a manifold-constrained blend of predicted clean samples on overlaps, and *stochastic early-phase sampling*, which injects fresh noise in the high-noise phase to synchronize trajectories before deterministic ODE sampling. On the joint audio-video transformer LTX-2, without fine-tuning, FlowLong yields arbitrarily long, phase-locked audio-video, among the first training-free approaches to long audio-video generation.

Keywords: Long audio-video generation · Diffusion models · Flow matching · Training-free inference

1 Introduction

Long-form generation is increasingly in demand, yet video and joint audio-video diffusion models are trained on short fixed-length clips and collapse when rolled out past that length [4, 11, 13]. Training-free remedies come in two limited families. Bidirectional extensions [8, 17, 18] rely on backbone-specific interventions and still degrade with length. Autoregressive methods [2, 5, 9, 14–16] reuse a KV-cache—accumulating exposure bias, drift, and repetitive motion—and must be distilled from a bidirectional teacher. This is decisive for *audio-video*: recent joint audio-video transformers such as LTX-2 [4] have no distilled autoregressive counterpart, so no training-free method extends them to long horizons.

We close this gap with **FlowLong**, which leaves the backbone untouched and reshapes only the sampling process. FlowLong generates a long sequence as K overlapping windows sampled in parallel from independent noise, harmonized by (i) *Tweedie matching*, a manifold-constrained closed-form per-frame blend of adjacent windows' clean estimates derived from a diffusion inverse-problem

⋆Equal contribution. †Co-corresponding authors.

Project page: `https://jhq1234.github.io/flowlong-audio-video/`

view [1], and (ii) *stochastic early-phase sampling*, which injects fresh noise in the high-noise phase to break trajectory inertia before switching to deterministic ODE sampling. Never reusing a KV-cache, FlowLong is exposure-bias-free by construction. We (1) give a training-free, architecture-agnostic recipe for long generation from pretrained flow models, and (2) instantiate it on LTX-2 for—to our knowledge—among the first training-free long *audio-video* generation (Sec. 2.4).

2 Method

2.1 Preliminaries: Flow Models

Flow matching [10] transports a source $\boldsymbol{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to data $\boldsymbol{x}_0 \sim p_0$ along $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$, learning a velocity field $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$ under text condition $\boldsymbol{c}$. Tweedie's formula [3] gives the denoised and noisy estimates $\hat{\boldsymbol{x}}_{0|t} = \boldsymbol{x}_t - t\boldsymbol{v}_\theta$ and $\hat{\boldsymbol{x}}_{1|t} = \boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta$, so an Euler step $t \to s$ becomes the interpolation $\boldsymbol{x}_s = (1-s)\hat{\boldsymbol{x}}_{0|t} + s\hat{\boldsymbol{x}}_{1|t}$ [7]. We use $\boldsymbol{x}$ for encoded latents throughout.

2.2 Tweedie Matching

We generate a long sequence $\boldsymbol{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^K)$ of $N > F$ frames as K overlapping chunks of native length F, each sampled from independent noise with $\boldsymbol{v}_\theta$ invoked *per chunk*. Adjacent chunks must agree on an overlap of O frames, which we relax into a clean-manifold guidance loss

$$\ell_k(\boldsymbol{x}; t) = \tfrac{1}{2}\left\| M_k\, \boldsymbol{x} - M'_{k+1}\, \hat{\boldsymbol{x}}_{0|t}^{(k+1)}(\boldsymbol{c}_{k+1}) \right\|^2, \tag{1}$$

where M_k, M'_{k+1} select the last / first O frames of a chunk—exactly the inverse-problem template $\frac{1}{2}\|\boldsymbol{y} - \mathcal{A}\boldsymbol{x}\|^2$. Following decomposed diffusion sampling (DDS) [1], optimizing ℓ_k over the denoised estimate reduces to one closed-form gradient step on the overlap, the per-frame update

$$\bar{\boldsymbol{x}}_{0|t}^{(k)}[j] = \begin{cases} \hat{\boldsymbol{x}}_{0|t}^{(k)}(\boldsymbol{c}_k)[j], & j \notin \Omega_k, \\ (1-\lambda_j)\, \hat{\boldsymbol{x}}_{0|t}^{(k)}(\boldsymbol{c}_k)[j] + \lambda_j\, \hat{\boldsymbol{x}}_{0|t}^{(k+1)}(\boldsymbol{c}_{k+1})[j'], & j \in \Omega_k, \end{cases} \tag{2}$$

with $j' = j - (F - O)$ and a symmetric linear schedule $\lambda_j = \frac{j-(F-O)}{O-1}$. Blending *clean* estimates keeps the correction on-manifold, and the symmetric schedule makes overlaps boundary-consistent and collapses all pairwise updates into a single buffer aggregation. We call this *Tweedie matching*.

2.3 Stochastic Early-Phase Sampling

Under a deterministic ODE, the renoising step drags $\boldsymbol{x}_s^k$ back toward each chunk's original—and divergent—trajectory, so Tweedie matching alone cannot

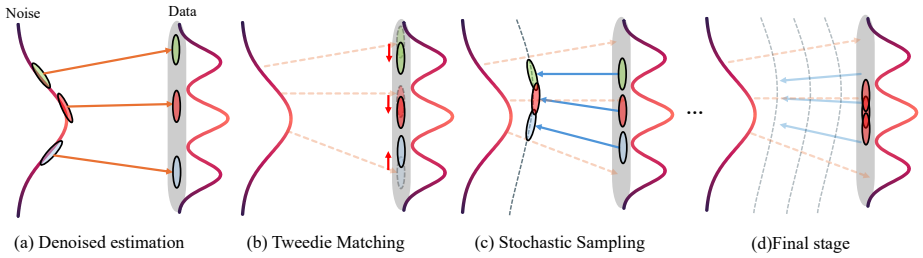

Fig. 1: FlowLong pipeline. Chunks follow independent ODEs (a); Tweedie matching interpolates their clean estimates on overlaps (b); early stochastic noise prevents reversion to the original paths (c); iterating (b)–(c) synchronizes all chunks into one coherent long sequence (d).

fully synchronize windows. We break this inertia by casting the early renoising as an SDE: following FlowDPS [7], we replace the noisy estimate with

$$\tilde{\boldsymbol{x}}^k_{1|t} = \sqrt{1-\eta_t}\,\bar{\boldsymbol{x}}^k_{1|t} + \sqrt{\eta_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

and adopt a binary schedule $\eta_t = \mathbb{1}(t \geq t^*)$: the high-noise phase remixes trajectories after every matching step, while the low-noise phase reverts to deterministic ODE sampling—combining the temporal consistency of SDE with the visual quality of ODE.

2.4 FlowLong on LTX-2 for Long Audio-Video Generation

Backbone and chunked joint sampling. LTX-2 [4] is a flow-matching video DiT with an audio branch and cross-modal attention that *jointly* denoises a video latent $\boldsymbol{x}^v$ and audio latent $\boldsymbol{x}^a$ under a shared text condition $\boldsymbol{c}$; the audio-video correspondence (motion-sound phase alignment) is produced by this pretrained cross-modal attention. FlowLong adds no module or fine-tuning: it decomposes each stream into K overlapping chunks and invokes the joint denoiser per chunk, $(\hat{\boldsymbol{x}}^{v,(k)}_{0|t}, \hat{\boldsymbol{x}}^{a,(k)}_{0|t}) = \mathrm{LTX2}(\boldsymbol{x}^{v,(k)}_t, \boldsymbol{x}^{a,(k)}_t, \boldsymbol{c})$, so cross-modal attention keeps the two streams synchronized *within* each window while FlowLong harmonizes consecutive windows.

Per-modality Tweedie matching. We apply the update (2) to *both* clean estimates with the same schedule λ_j, blending each modality on its own overlap region Ω^m_k for $m \in \{v, a\}$, enforcing cross-window consistency separately on the video and audio manifolds.

Cross-modal temporal alignment. The two streams carry very different numbers of latents per second: after the VAE's temporal stride r=8, video latents arrive at rate $\rho_v = \mathrm{fps}_v/r$ (e.g. 24/8=3 latents/s), while audio latents arrive at ρ_a=25 latents/s. Slicing both into chunks of the *same latent count* would make each chunk span a different real duration, desynchronizing the two modalities. We

Table 1: Long generation on the audio-video backbone LTX-2 (30s, VBench). Best per metric in **bold**; FlowLong (Ltx2+Ours) is applied training-free. AQ/BC/DD/IQ/MS/SC/TF: Aesthetic, Background, Dynamic Degree, Imaging, Motion, Subject, Temporal Flicker. Audio-video phase-locking is shown qualitatively on the project page.

Model	AQ	BC	DD	IQ	MS	SC	TF
Ltx2 (sliding window)	**0.541**	0.885	**0.625**	0.612	0.981	0.815	0.948
Ltx2 + Ours	0.534	**0.902**	0.616	**0.639**	**0.985**	**0.820**	**0.977**

instead match the chunk geometry *in seconds*: given the video pixel window W and overlap-start index w,

$$F_a = \mathrm{round}\Big(\tfrac{W}{\mathrm{fps}_v}\rho_a\Big), \quad S_a = \mathrm{round}\Big(\tfrac{W-w}{\mathrm{fps}_v}\rho_a\Big), \quad O_a = F_a - S_a. \tag{4}$$

For LTX-2 with $(W, w, \mathrm{fps}_v, \rho_a) = (121, 64, 24, 25)$ this yields video $(F, O, S) = (16, 8, 7)$ and audio $(F_a, O_a, S_a) = (126, 67, 59)$, both satisfying $O \geq S$ so every blended frame is co-predicted by the next chunk; rounding leaves ≤ 1 audio latent (≈ 40 ms) of slack, clamped to zero.

Per-modality stochastic sampling. Finally we advance both buffers with the hybrid schedule (Sec. 2.3), drawing *independent* noise $\boldsymbol{\epsilon}^v, \boldsymbol{\epsilon}^a \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ per modality: each stream breaks its own inertia across windows, while the shared per-second geometry keeps video and audio aligned—yielding an arbitrarily long, phase-locked sequence at inference time.

3 Experiments

We evaluate FlowLong on the joint audio-video transformer LTX-2 [4] using VBench [6] (seven dimensions) over 30-second generations from 100 MovieGen Bench [12] prompts, on a single NVIDIA H100. To our knowledge FlowLong is among the first training-free methods to extend a joint audio-video transformer beyond its native window. As no training-free baseline supports LTX-2 past its native 5-second window, we compare against a sliding-window baseline and obtain higher overall VBench scores across most metrics (Table 1, "Ltx2+Ours"), while the stochastic early-phase sampling keeps the two modalities phase-locked over long horizons (see the project page for audio-video results).

4 Conclusion

We presented FlowLong, a training-free, architecture-agnostic framework that extends pretrained flow models beyond their native horizon via Tweedie matching and stochastic early-phase sampling. Instantiated on LTX-2, it is among the first training-free long *audio-video* generators.

References

1. Chung, H., Lee, S., Ye, J.C.: Decomposed diffusion sampler for accelerating large-scale inverse problems. arXiv preprint arXiv:2303.05754 (2023)
2. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. arXiv preprint arXiv:2510.02283 (2025)
3. Efron, B.: Tweedie's formula and selection bias. Journal of the American Statistical Association **106**(496), 1602–1614 (2011)
4. 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)
5. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
6. 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)
7. Kim, J., Kim, B.S., Ye, J.C.: Flowdps: Flow-driven posterior sampling for inverse problems. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12328–12337 (2025)
8. Kim, J., Kang, J., Choi, J., Han, B.: Fifo-diffusion: Generating infinite videos from text without training. In: NeurIPS (2024)
9. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025)
10. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
11. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
12. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
13. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
14. Yesiltepe, H., Meral, T.H.S., Akan, A.K., Oktay, K., Yanardag, P.: Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. arXiv preprint arXiv:2511.20649 (2025)
15. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. arXiv preprint arXiv:2512.05081 (2025)
16. 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: CVPR (2025)

17. Zhao, M., He, G., Chen, Y., Zhu, H., Li, C., Zhu, J.: Riflex: A free lunch for length extrapolation in video diffusion transformers. arXiv preprint arXiv:2502.15894 (2025)
18. Zhao, M., Zhu, H., Wang, Y., Yan, B., Zhang, J., He, G., Yang, L., Li, C., Zhu, J.: Ultravico: Breaking extrapolation limits in video diffusion transformers. arXiv preprint arXiv:2511.20123 (2025)