Inference-Time On-Manifold Steering for Autoregressive Long Video Generation

Taesung Kwon [*1] **TaeHoon Lee** [*1] **Jong Chul Ye** [1]

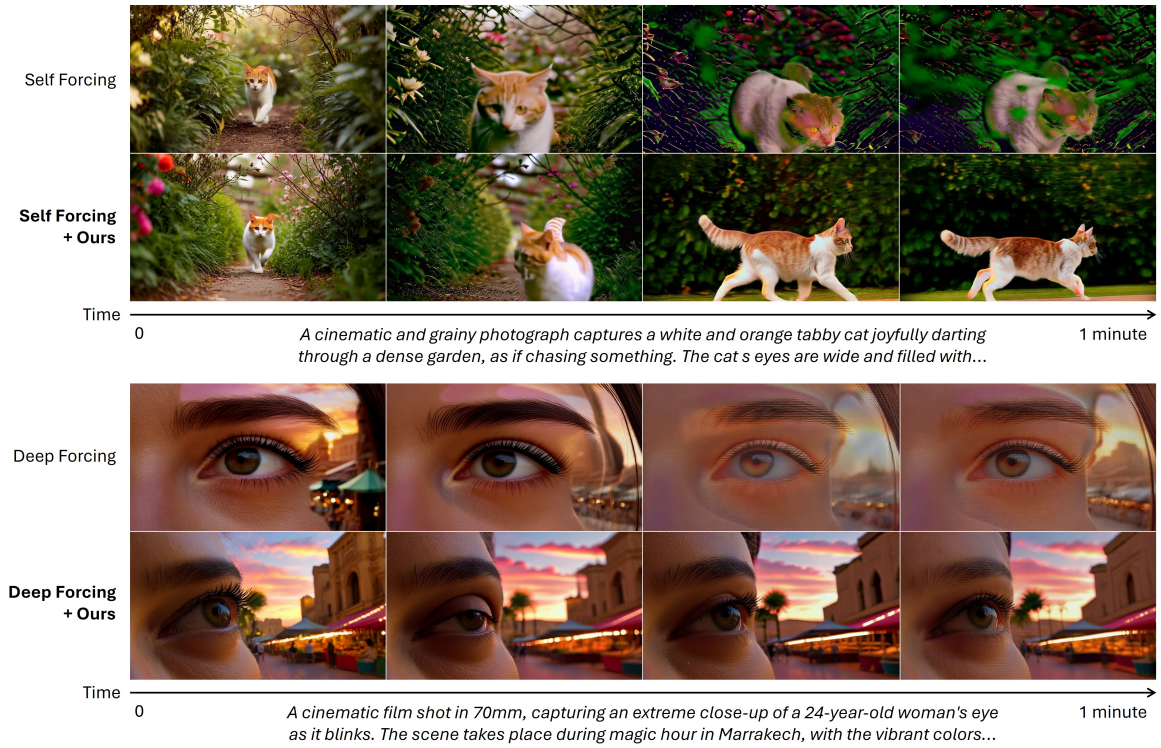


Figure 1. **On-Manifold Steering with Time Reward.** Autoregressive long video generation suffers from temporal drift as sequential errors push the latent states off the desired manifold. Our proposed method actively guides the sampling trajectory back onto the manifold using a lightweight *time predictor* to provide a *time reward*, ensuring temporal coherence in long video generation.

Abstract

Autoregressive video diffusion models have shown remarkable success in real-time video generation. However, generating long video sequences with these models remains challenging due to the progressive accumulation of errors and temporal drift. We attribute this degradation to the sequential generation process, which gradually pushes latent representations off the desired manifold. To address this, we introduce *On-Manifold Steering with Time Reward*, a novel framework designed for long video generation. By training a lightweight *time predictor* to classify the timesteps of noisy latents, we derive a *time reward* gradient that continuously steers the diffusion sampling trajectory. This guidance actively pulls the temporal latent states back onto the manifold at each step, effectively correcting inference misalignments on the fly. Consequently, our approach mitigates feature drift and preserves structural integrity across long video sequences. Extensive experiments demonstrate that our framework seamlessly integrates with existing baselines such as Self Forcing and Deep Forcing, consistently improving temporal coherence and visual quality in 30- and 60-second video generation tasks.

[1]Graduate School of AI, KAIST, South Korea. Correspondence to: Jong Chul Ye <jong.ye@kaist.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

The advancement of autoregressive (AR) video diffusion models (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025; Ruhe et al., 2024; Kim et al., 2024; Zhang et al., 2025; Jin et al., 2025) has revolutionized the generation of high-fidelity, short-form visual content in real-time. By generating future video segments conditioned on a cached context of previously generated frames, AR video models achieve significant memory efficiency and low-latency generation. However, extending these capabilities to generate long, temporally coherent videos remains technically challenging due to error accumulation over long horizons, as each prediction depends on previously generated and potentially imperfect frames (Wang et al., 2025).

To address this limitation, several recent works introduce long-horizon training strategies or explicit memory-management mechanisms (Cui et al., 2025; Liu et al., 2025; Chen et al., 2026; Lu et al., 2025; Yang et al., 2025). Despite their effectiveness, these approaches generally require additional training, distillation, or fine-tuning of the video model itself. Recent training-free approaches instead attempt to address long-horizon degradation purely at inference time by handling the long-context KV-cache effectively (Yi et al., 2025; Li et al., 2026).

In contrast to the existing inference-time methods that predominantly focus on optimizing or managing the KV-cache to retain long-term context, we provide a novel perspective in this paper: we view the error accumulation in long video generation inherently as an *off-manifold problem*, as studied extensively in the context of inverse problems and guided sampling (Chung et al., 2023; 2024; 2025; Park et al.; Jung et al., 2024). If the error accumulation occurs because the sampling points deviate from the desired manifold of the diffusion process at each timestep, we can correct it by explicitly guiding the points back to the manifold. To achieve this efficiently, we introduce *On-Manifold Steering with Time Reward*, utilizing a lightweight neural network—a time predictor (Park et al.; Jung et al., 2024)—to provide a time reward during inference.

Specifically, we train the time predictor to classify the discrete diffusion timesteps of noisy video segments. During autoregressive inference, the gradient of the time predictor serves as an on-manifold reward to evaluate whether the noisy latent is accurately located at the desired timestep. By actively correcting the noisy latent segment at each step, our method maximizes its alignment with the desired manifold. This process requires no retraining of the backbone video diffusion model and introduces negligible computational overhead.

We demonstrate that our method acts as a plug-and-play module for state-of-the-art AR video generative baselines, such as Self Forcing (Huang et al., 2025) and Deep Forcing (Yi et al., 2025). Extensive experiments show that our approach consistently improves these models in both 30- and 60-second video generation tasks. Ultimately, we hope our framework provides a fresh, complementary perspective for long video generation: demonstrating that actively steering the sampling trajectory on-manifold is crucial as conventional KV-cache management strategies.

Our main contributions are summarized as follows:

- We formulate error accumulation in autoregressive long video generation inherently as an *off-manifold problem*, providing a novel, complementary perspective that extends beyond conventional KV-cache management strategies.
- We propose *On-Manifold Steering with Time Reward*, a framework that utilizes a lightweight time predictor to provide on-the-fly gradient guidance. This actively corrects the sampling trajectory back onto the desired manifold without requiring retraining of the backbone model.
- We demonstrate the plug-and-play versatility of our method by seamlessly integrating it with state-of-the-art AR video generative baselines. Extensive evaluations show that our approach consistently improves temporal coherence and visual quality in both 30- and 60-second video generation tasks.

2. Preliminaries

Autoregressive Formulation. Consider a full video sequence divided into N distinct, non-overlapping segments, structured as $\mathbf{z} = [\mathbf{z}^1, \mathbf{z}^2, \dots, \mathbf{z}^N]$. In the autoregressive formulation, rather than synthesizing the entire latent sequence simultaneously, we decompose the generation process into a sequence of conditional steps. Here, each segment relies on the context provided by its predecessors, allowing the joint distribution to be factorized as follows:

$$p(\mathbf{z}) = \prod_{n=1}^{N} p(\mathbf{z}^n \mid \mathbf{z}^{<n}). \tag{1}$$

Each conditional distribution $p(\mathbf{z}^n|\mathbf{z}^{<n})$ is parameterized by a model $\mathbf{v}_\theta$, which generates the current segment $\mathbf{z}^n$ conditionally on the past context $\mathbf{z}^{<n}$ (where the initial context is defined as $\mathbf{z}^0 = \emptyset$).

Sampling process. For the sampling of the n-th latent $\mathbf{z}^n$ from the conditional distribution $p(\mathbf{z}^n|\mathbf{z}^{<n})$, we adopt the same sampling procedure as Self Forcing (Huang et al., 2025). Let $\mathbf{z}_0^n$ denote the clean data from a data distribution p_{data}, and $\mathbf{z}_1^n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represent the initial noise. Given a

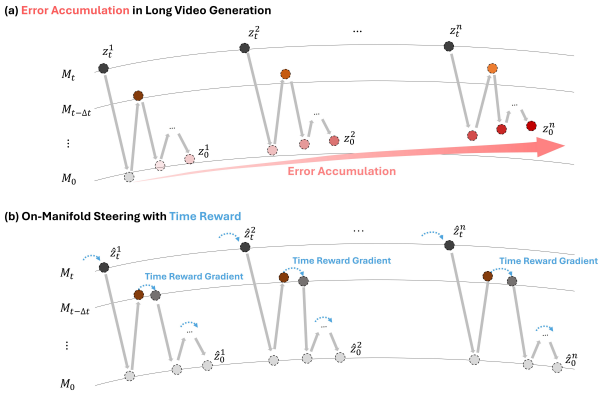


Figure 2. **Overview of On-manifold Steering with Time Reward.** (a) Rolling out self forcing progressively accumulates temporal drift, causing predictions to deviate from the target timestep manifold. (b) Our method steers the sampling trajectory with the gradient of the time reward, guiding samples back toward desired trajectories.

noisy latent $\mathbf{z}_t^n$ at an intermediate timestep t, the network directly outputs the denoised estimate $\mathbf{v}_\theta(\mathbf{z}_t^n, t; \mathbf{z}_0^{<n})$:

$$\hat{\mathbf{z}}_{0|t}^n = \mathbf{v}_\theta(\mathbf{z}_t^n, t; \mathbf{z}_0^{<n}). \tag{2}$$

In practice, to avoid the computational overhead of processing the past context, the conditioning on $\mathbf{z}_0^{<n}$ can be efficiently implemented using its KV cache (Yin et al., 2025; Huang et al., 2025), denoted as $\text{KV}^{<n}$. Subsequently, the process continues to the next timestep $t - \Delta t$ by re-noising the clean estimate $\hat{\mathbf{z}}_{0|t}^n$ from Eq. (2):

$$\mathbf{z}_{t-\Delta t}^n = (1 - (t - \Delta t))\hat{\mathbf{z}}_{0|t}^n + (t - \Delta t)\mathbf{z}_1^n. \tag{3}$$

This process of denoising and re-noising is repeated until $t = 0$, yielding the full sampled latent $\mathbf{z}_0^n$, after which the model proceeds to sample the next segment in the sequence.

3. On-manifold Steering with Time Reward

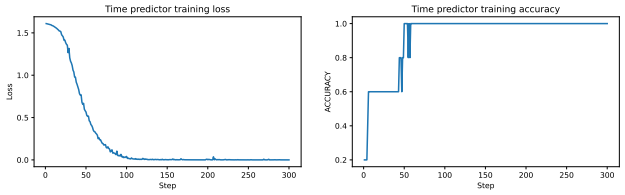
In this section, we detail our framework, designed to harness the underexplored potential of reward guidance in AR video diffusion models for long-video generation. We begin by introducing a training method for the time predictor in Section 3.1. Building upon this, Section 3.2 presents our core framework, On-manifold Steering with Time Reward, which steers the AR process by computing the gradient of the time reward and applying it to the sampling steps in the AR process.

3.1. Training Time Predictor

To provide a reward signal indicating whether a noisy segment resides on the correct manifold at a given timestep, we train a lightweight neural network classifier, the time predictor parameterized by ϕ by following prior works (Park et al.; Jung et al., 2024). Videos are encoded into the latent space using Wan-VAE (Wan et al., 2025). We extract random segments $\mathbf{z}_0$ and inject noise to simulate noisy data of specific inference timesteps $t \in \mathcal{T} = \{250, 500, 750, 1000\}$, matching the discrete timesteps used in the Self Forcing backbone (Huang et al., 2025). Specifically, to estimate the proper timestep of given noisy data, we parameterize the time predictor ϕ with a simple CNN architecture. Then, the cross-entropy loss between the one-hot embedding of the timestep vector and the logit vector of the model output is used for the following objective function:

$$\mathcal{L}_{\text{CE}}(\phi) = -\mathbb{E}_{t\sim\mathcal{T},\mathbf{z}_0}\left[\log(\hat{p}_\phi(\mathbf{z}_t)_t)\right], \tag{4}$$

where $\hat{p}_\phi(\mathbf{z}_t)_t$ is the t-th component of the model output for a given noisy segment $\mathbf{z}_t$. As shown in Figure 3, the predictor exhibits fast convergence, demonstrating the efficiency of the framework, which can be easily adapted for other diffusion priors.



(a) Training loss convergence. *(b)* Training accuracy progression.

Figure 3. **Training dynamics of the time predictor.** (a) The time predictor training loss steadily converges throughout optimization. (b) The timestep prediction accuracy exhibits fast convergence, demonstrating the efficiency of our framework.

3.2. On-Manifold Guidance with Time Reward Gradient

In standard diffusion models, fixed timestep scheduling suffices because the sampling trajectory strictly follows the expected reverse path. However, in AR video generation, the sequential conditioning on previously generated segments ($\text{KV}^{<n}$) potentially introduces external drift. As errors accumulate, the intermediate noisy latent $\mathbf{z}_t^n$ loses its temporal identity and falls off the desired manifold $\mathcal{M}_t$. Once off-manifold, the conventional vector field of the diffusion model struggles to denoise accurately, leading to temporal degradation.

To resolve this, we reinterpret the timestep not merely as a fixed scheduling parameter, but as an active conditioning variable, a *time reward*. If the structural drifting occurs because it deviates from the desired manifold at time t, we can project $\mathbf{z}_t^n$ back onto the correct manifold $\mathcal{M}_t$ by maximizing this time reward.

We introduce a gradient term derived from our lightweight

time predictor ϕ. Similar to prior work (Park et al.; Jung et al., 2024), we define the *Time Reward Gradient* (TRG) as the direction that maximizes the log-probability of the sampling point belonging to the target timestep t:

$$\textit{Time Reward Gradient}\ (\mathbf{z}_t^n, t) := \nabla_{\mathbf{z}_t^n} \log p_\phi(t \mid \mathbf{z}_t^n). \quad (5)$$

This gradient vector field actively directs off-manifold samples toward high-probability regions of the desired manifold at the target timestep. Without this correction, the base diffusion model would propagate the off-manifold error. By incorporating this TRG into the AR sampling process, we formulate the On-manifold Steering with Time Reward. Specifically, at each timestep t, before evaluating the diffusion vector field $\mathbf{v}_\theta$, we apply a guidance step to update $\mathbf{z}_t^n$:

$$\hat{\mathbf{z}}_t^n = \mathbf{z}_t^n + \lambda \nabla_{\mathbf{z}_t^n} \log p_\phi(t \mid \mathbf{z}_t^n), \quad (6)$$

where λ is a hyperparameter that controls the guidance strength. Applying this correction provides a mathematical shortcut for the drifted sample to return to the correct manifold. By ensuring that the sampling point of the diffusion model is always structurally aligned with the expected noise level at time t, our framework effectively prevents temporal drift by simply adding reward-guided steering to the original sampling process. Notably, our framework introduces reward-based steering as a novel and complementary perspective for long video generation. The complete pipeline is detailed in Algorithm 1.

Algorithm 1 On-manifold Steering with Time Reward

Require Diffusion model $\mathbf{v}_\theta$, Time Predictor ϕ, number of segments N, guidance scale λ.
$\text{KV}^0 \leftarrow \emptyset$
for $n = 1 : N$ **do**
  for $t : T \to 0$ **do**
    $\hat{\mathbf{z}}_t^n \leftarrow \mathbf{z}_t^n + \lambda \nabla_{\mathbf{z}_t^n} p_\phi(t \mid \mathbf{z}_t^n)$ ▷ Time Reward Gradient (Eq. (6))
    $\hat{\mathbf{z}}_{0|t}^n \leftarrow \mathbf{v}_\theta(\hat{\mathbf{z}}_t^n, t; \text{KV}^{<n})$ ▷ Obtain denoised estimate via diffusion model
    $\mathbf{z}_{t-\Delta t}^n \leftarrow (1 - (t - \Delta t))\hat{\mathbf{z}}_{0|t}^n + (t - \Delta t)\mathbf{z}_1, \quad \mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Re-noise step
  end for
end for
return $[\hat{\mathbf{z}}_0^1, \ldots, \hat{\mathbf{z}}_0^N]$

4. Experiments

4.1. Results

Quantitative results. As shown in Table 1, our method demonstrates consistently improved performance across major VBench-Long metrics when integrated with both Self Forcing (Huang et al., 2025) and Deep Forcing (Yi et al., 2025) baselines. Specifically, for 30-second video generation, our method improves the dynamic degree (*i.e.*, the capability to generate large motions) by a large margin. Furthermore, our approach consistently enhances both frame-wise quality (imaging and aesthetic quality) and temporal coherence (subject and background consistency).

Method	Dyn. Deg.↑	M. Smooth.↑	Imag.↑	Aesth.↑	Sub. Con.↑	Bg. Con.↑
30 seconds						
Self Forcing	42.19	**97.85**	68.00	59.58	97.43	96.56
Self Forcing + Ours	**57.59**	97.73	**69.19**	**62.46**	**97.45**	**96.65**
Deep Forcing	24.30	**98.77**	68.00	61.45	98.36	97.28
Deep Forcing + Ours	**33.33**	98.54	**69.76**	**64.69**	**98.40**	**97.29**
60 seconds						
Self Forcing	42.41	97.85	68.00	**59.58**	97.43	96.56
Self Forcing + Ours	**44.44**	**98.00**	**69.84**	59.11	**97.48**	**96.61**
Deep Forcing	26.39	98.76	67.58	61.77	98.38	**97.34**
Deep Forcing + Ours	**38.19**	**98.79**	**68.89**	**64.49**	**98.40**	97.26

Table 1. **Quantitative comparison on long video generation.** Integrating our on-manifold steering mitigates temporal drift, achieving superior coherence metrics on VBench-Long.

This trend generalizes to 60-second video generation. Notably, when applied to Deep Forcing, the dynamic degree improves significantly, and the majority of VBench-Long metrics exhibit consistent gains simply by guiding the sampling trajectory with our time reward. This demonstrates the strong versatility and effectiveness of our proposed method.

Qualitative results. Consistent with our quantitative findings, this positive trend is clearly observed in our qualitative results. As shown in Figure 4, integrating our method into the baselines consistently mitigates temporal drift and improves overall generation quality. Specifically, as observed in the first row, a 60-second rollout of Self Forcing introduces significant temporal drift. In contrast, our method successfully alleviates this degradation, maintaining visual coherence even in 1-minute video generation. Similarly, the second row demonstrates that our approach effectively corrects failure cases present in the Deep Forcing.

Furthermore, as highlighted in the third and fourth rows, both Self Forcing and Deep Forcing frequently suffer from severe artifacts, such as objects suddenly disappearing or collapsing (indicated by red arrows). When our method is applied, the structural integrity of these objects is well-preserved throughout the long sequence (indicated by blue arrows). These visual improvements demonstrate the versatility and robust performance of our proposed method.

4.2. Ablation Studies

λ	Dyn. Deg.↑	M. Smooth.↑	Imag.↑	Aesth.↑	Sub. Con.↑	Bg. Con.↑
0.01	50.89	97.67	67.88	**60.06**	97.03	96.31
0.1	**55.13**	97.84	68.02	59.58	97.29	96.44
1	51.34	97.78	68.47	58.98	97.23	96.31
10	44.42	**98.00**	**69.84**	59.12	**97.48**	**96.61**
20	53.57	97.84	68.37	58.55	96.98	96.22

Table 2. **Ablation study on the guidance scale (λ).**

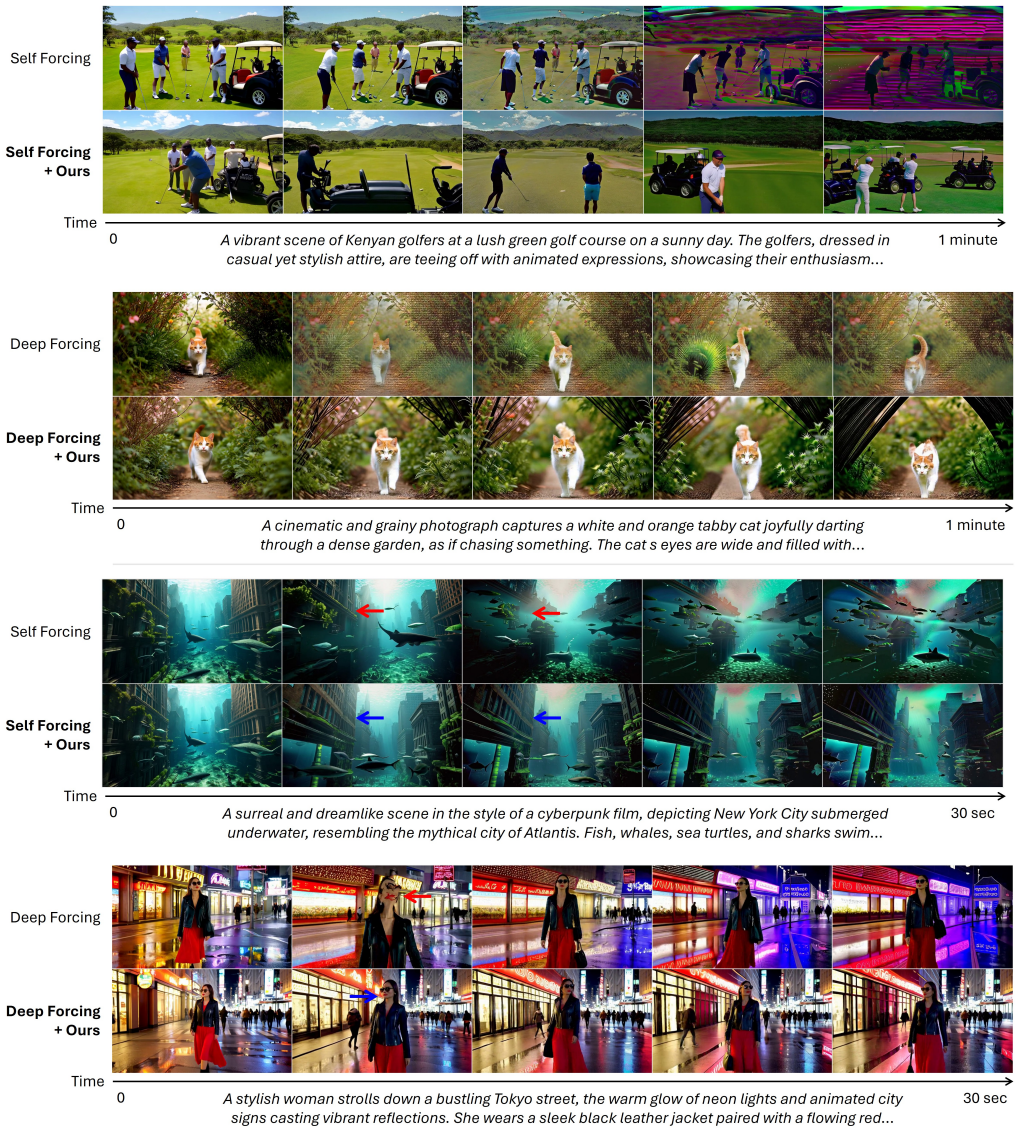


Figure 4. **Qualitative comparison on 30-second and 60-second video generation.** Compared to the baselines, our approach consistently reduces temporal artifact and maintains subject identity over long-horizon generations.

Effect of Guidance Scale (λ). In this section, we analyze the impact of the time reward guidance scale λ on 60-second video generation with the Self Forcing (Huang et al., 2025) backbone. As shown in Table 2, we observe that larger values of λ generally enhance the overall VBench-Long metrics by more strictly enforcing the on-manifold guidance. However, scaling λ excessively from 10 to 20 causes the performance to degrade. This indicates that excessive guidance may actually hinder, rather than improve, the generation quality. Consequently, we adopt $\lambda = 10$ as our default setting throughout the paper, as it achieves the highest overall performance.

5. Conclusion

In this paper, we introduced *On-Manifold Steering with Time Reward*, a novel framework designed to mitigate temporal drifting in autoregressive long video generation. By explicitly framing this drift as an *off-manifold problem*, we employed a lightweight time predictor to guide the sampling process back to the desired manifold using time reward steering. Our plug-and-play approach seamlessly integrates into existing AR video diffusion models, significantly improving their temporal stability and maintaining structural coherence for long video generations. Future work could explore more cost-efficient and hyperparameter-robust time-reward guidance that is applied only at selected informative sampling steps and generalizes better across diverse data distributions.

References

Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.-W., et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., and Sitzmann, V. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Chen, S., Wei, C., Sun, S., Nie, P., Zhou, K., Zhang, G., Yang, M.-H., and Chen, W. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Chung, H., Lee, S., and Ye, J. C. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DsEhqQtfAG.

Chung, H., Kim, J., Park, G. Y., Nam, H., and Ye, J. C. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E77uvbOTtp.

Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., and Hsieh, C.-J. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

DeepMind, G. Veo 3. https://deepmind.google/models/veo/, 2025.

HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Huang, X., Li, Z., He, G., Zhou, M., and Shechtman, E. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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.

Jin, Y., Sun, Z., Li, N., Xu, K., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., MU, Y., and Lin, Z. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

Jung, H., Park, Y., Schmid, L., Jo, J., Lee, D., Kim, B., Yun, S.-Y., and Shin, J. Conditional synthesis of 3d molecules with time correction sampler. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=gipFTlvfF1.

Kim, B. and Ye, J. C. Denoising mcmc for accelerating diffusion-based generative models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 16955–16977, 2023.

Kim, J., Kang, J., Choi, J., and Han, B. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37: 89834–89868, 2024.

Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Li, H., Liu, S., Lin, Z., and Chandraker, M. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026.

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.

Liu, K., Hu, W., Xu, J., Shan, Y., and Lu, S. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Lu, Y., Zeng, Y., Li, H., Ouyang, H., Wang, Q., Cheng, K. L., Zhu, J., Cao, H., Zhang, Z., Zhu, X., et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

OpenAI. Sora. https://openai.com/sora/, 2024.

Park, Y., Jung, H., Bae, S., and Yun, S.-Y. Temporal alignment guidance: On-manifold sampling in diffusion models. In *NeurIPS 2025 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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.

Ruhe, D., Heek, J., Salimans, T., and Hoogeboom, E. Rolling diffusion models. In *International Conference on Machine Learning*, pp. 42818–42835. PMLR, 2024.

San-Roman, R., Nachmani, E., and Wolf, L. Noise estimation for generative diffusion models. *arXiv preprint arXiv:2104.02600*, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wang, J., Zhang, F., Li, X., Tan, V. Y., Pang, T., Du, C., Sun, A., and Yang, Z. Error analyses of auto-regressive video diffusion models: A unified framework. *arXiv preprint arXiv:2503.10704*, 2025.

Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Yi, J., Jang, W., Cho, P. H., Nam, J., Yoon, H., and Kim, S. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Yin, T., Zhang, Q., Zhang, R., Freeman, W. T., Durand, F., Shechtman, E., and 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.

Zhang, L., Cai, S., Li, M., Wetzstein, G., and Agrawala, M. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

A. Experimental Setup

Implementation details. We encode the MixKit dataset (Lin et al., 2024) using Wan-VAE (Wan et al., 2025) to train the time predictor. The predictor architecture is designed to be highly lightweight, consisting of a shallow 1D CNN with four convolutional layers operating on the encoded features. Specifically, the hidden representations are first projected into a lower-dimensional embedding space, followed by stacked temporal convolution and average pooling layers to progressively aggregate temporal information. The final pooled representation is passed through a linear classification head to predict the diffusion timestep. Training converges efficiently using the AdamW optimizer with a learning rate of 1×10^{-4}, requiring only approximately 300 optimization steps. To demonstrate the versatility of our method, we apply our reward guidance to the official implementations of Self Forcing (Huang et al., 2025) and Deep Forcing (Yi et al., 2025). The guidance scale is set to $\lambda = 10$ based on our ablation study. All training and inference experiments are conducted on NVIDIA A100 GPUs with 80GB memory. Further details and the implementations are provided in the supplementary material.

Evaluation. We conduct extensive evaluations on long video generation using VBench-Long (Huang et al., 2024). Following the evaluation protocol of Deep Forcing (Yi et al., 2025), we utilize prompts from MovieGen (Polyak et al., 2024) to generate 30- and 60-second videos. We assess key metrics, including Dynamic Degree (Dyn. Deg.), Motion Smoothness (M. Smooth.), Imaging Quality (Imag.), Aesthetic Quality (Aesth.), Subject Consistency (Sub. Con.), and Background Consistency (Bg. Con.).

B. Related Work

Autoregressive Video Generation with Diffusion. Video generation has experienced rapid growth driven by the introduction of high-performance foundational models, including both proprietary systems and open-weight releases (OpenAI, 2024; DeepMind, 2025; Kong et al., 2024; HaCohen et al., 2024; Wan et al., 2025; Agarwal et al., 2025; Ali et al., 2025). Among recent open-source models (Kong et al., 2024; HaCohen et al., 2024; Wan et al., 2025; Ali et al., 2025), diffusion transformers (DiTs) (Peebles & Xie, 2023) have become the standard backbone, typically relying on bidirectional attention (Vaswani et al., 2017) to generate all frames simultaneously. To achieve faster, low-latency generation, recent research (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025; Ruhe et al., 2024; Kim et al., 2024; Zhang et al., 2025; Jin et al., 2025) has shifted towards autoregressive (AR) formulations. In this formulation, models generate new frames sequentially by conditioning on previously generated information. A notable methodology involves adapting bidirectional teacher models into causal student transformers through distillation, utilizing key-value (KV) caching to accelerate training and inference (Yin et al., 2025; Huang et al., 2025). In this line of work, Self Forcing (Huang et al., 2025) mitigates early-stage temporal drift by employing its previously generated outputs to condition current predictions during training, thereby reducing the train-test gap caused by relying solely on ground-truth data.

Autoregressive Long Video Generation with Diffusion. Despite the efficiency gains of the aforementioned AR formulations, they inherit a fundamental limitation when applied to extended temporal horizons. Because these models are often distilled from bidirectional teachers with finite context windows, their supervision is inherently bounded by the short-window generation capability of the teacher model. As a result, while the causal student can theoretically be rolled out for longer video generation, the model struggles to maintain temporal consistency and visual quality over long horizons. To address this limitation, several recent works introduce long-horizon training strategies or explicit memory-management mechanisms (Cui et al., 2025; Liu et al., 2025; Chen et al., 2026; Lu et al., 2025; Yang et al., 2025). Despite their effectiveness, these approaches generally require additional training, distillation, or fine-tuning of the video model itself. This increases computational cost and limits their applicability to already trained autoregressive video models. Recent training-free approaches instead attempt to address long-horizon degradation purely at inference time (Yi et al., 2025; Li et al., 2026). In this line of work, Deep Forcing (Yi et al., 2025) introduces Deep Sink and Participative Compression to handle the long-context KV-cache effectively.

Diverging from these paradigms that either demand extensive retraining or focus predominantly on KV-cache manipulation, we approach the long-horizon degradation problem from a fundamentally different and complementary perspective. We view this error accumulation inherently as an *off-manifold problem*, as studied extensively in the context of inverse problems and guided sampling (Chung et al., 2023; 2024; 2025; Park et al.; Jung et al., 2024). To the best of our knowledge, ours is the first work to investigate whether actively steering the sampling trajectory—ensuring it remains on-manifold—can effectively sustain temporal coherence and improve autoregressive long-video generation without heavy computational overhead.

Noise Predictors for Improved Diffusion Sampling. Noise prediction neural networks are often employed to accelerate

or improve the diffusion sampling process (Nichol & Dhariwal, 2021; San-Roman et al., 2021; Kim & Ye, 2023). For example, learning the covariance of the reverse distribution (Nichol & Dhariwal, 2021) or adjusting the noise schedule with a neural network (San-Roman et al., 2021) can speed up the integration of the reverse SDE/ODE. Moreover, Denoising MCMC (Kim & Ye, 2023) utilizes the noise predictor within an MCMC framework to search for optimal timesteps, further accelerating generation. More recently, predicting the *diffusion timestep* (noise level) itself has emerged as a powerful strategy to correct off-manifold trajectories. For instance, recent works have proposed denoising intermediate states based on predicted timesteps for 3D molecular generation (Jung et al., 2024), or utilizing the gradient of a learned diffusion time predictor to enforce on-manifold sampling in image generation (Park et al.).

In this work, we identify the temporal drift issue of autoregressive long video generation inherently as an *off-manifold problem*. To correct these deviations, we are the first to extend this time correction philosophy to the AR video domain. We employ a lightweight, pre-trained time predictor that provides a time reward signal to continuously guide the AR generation process on the fly. This mechanism effectively ensures that the sampling trajectory stays on the desired manifold across long video generation, seamlessly functioning as a plug-and-play module.