MASK FORCING: IMPROVING AUTOREGRESSIVE VIDEO DIFFUSION DISTILLATION VIA DUAL-NOISE MASKING ROLLOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive (AR) video diffusion models have shown great potential in real-time video generation. Recent methods distill pretrained bidirectional video diffusion models into causal students through Distribution Matching Distillation (DMD), but the generated videos often suffer from over-saturation and over-smoothing issues, resulting in limited visual quality and realism. The key contributing factor is the mode-seeking behavior of the reverse KL objective in DMD, which can cause the student distribution to collapse onto only a few modes of the teacher distribution. To address this, we propose Mask Forcing, a Dual-Noise Masking Rollout strategy that perturbs the causal student self-rollout to mitigate mode collapse induced by reverse-KL mode seeking. The core idea is to inject cleaner signals into noisy rollout inputs via random masks along spatial and temporal axes during the self-rollout process of AR diffusion distillation. Such perturbations encourage the student rollouts to explore broader regions of the teacher distribution, allowing DMD to provide learning signals beyond the modes already covered by the student. Moreover, the cleaner tokens act as denoising guidance for other noisier tokens, improving the intermediate rollout predictions and reducing error accumulation. Extensive experiments demonstrate that our method improves multiple AR video diffusion distillation methods with higher visual quality efficiently, without incorporating real video data or additional post-training stages.

1 INTRODUCTION

Video diffusion models have advanced rapidly in recent years, enabling the generation of long-duration and high-fidelity videos (Wan et al., 2025; Kong et al., 2024; HaCohen et al., 2026). However, these models typically rely on bidirectional attention and multiple denoising timesteps, limiting their applicability to real-time streaming generation scenarios. Therefore, autoregressive (AR) diffusion models have emerged as a promising paradigm for real-time video generation (Huang et al., 2026; Zhu et al., 2026; Yang et al., 2025; Hu et al., 2026; Gao et al., 2026). By leveraging causal attention mechanisms, AR diffusion models can generate future chunks sequentially, conditioning each chunk on previously generated ones via causal dependencies. This formulation supports real-time inference and unbounded video generation without recomputing previously generated frames.

Teacher Forcing (Hu et al., 2024; Gao et al., 2025) and Diffusion Forcing (Chen et al., 2024) are two paradigms for training AR video diffusion models. Teacher Forcing trains the model to predict the next frames conditioned on the ground-truth frames, leading to exposure bias since the model can only condition on its own predictions during inference. Diffusion Forcing instead trains the model by assigning each frame with independently sampled noise. While this paradigm alleviates the distribution shift, it still fails to align training with inference. Self Forcing (Huang et al., 2026) bridges the train-test gap by training the model with self-rollout, generating the next frame based on previously self-generated frames and distilling the teacher's knowledge via a DMD loss (Yin et al., 2024b). However, a critical limitation of combining self-rollout with DMD is that it often produces over-saturated and over-smoothed videos, exhibiting low visual quality and limited realism.

This phenomenon can be attributed to two key factors. First, the reverse-KL objective in DMD exhibits mode-seeking behavior that tends to cover the high-probability regions of the teacher's

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.

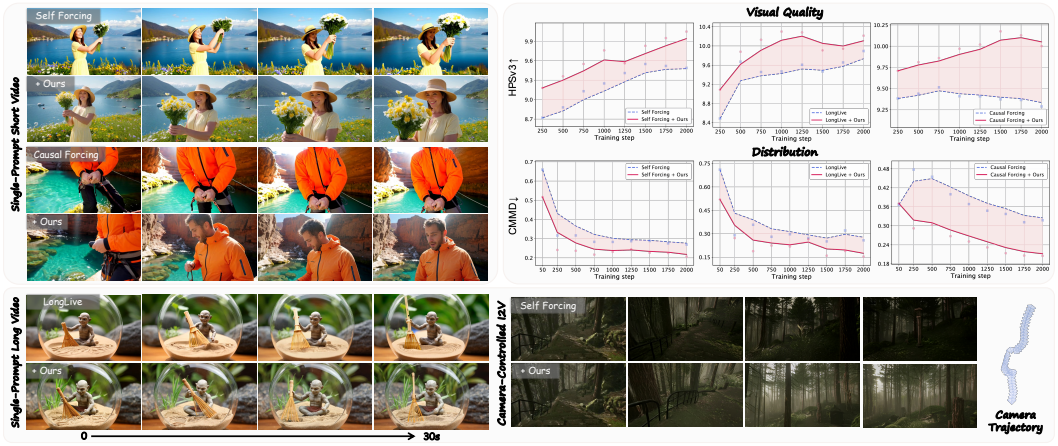

Figure 1: Mask Forcing improves autoregressive video generation through a Dual-Noise Masking Rollout strategy that alleviates the mode-seeking behavior of the reverse KL objective in DMD, consistently improving visual realism across short, long, and camera-controlled autoregressive video generation without additional real-video supervision or post-training.

distribution (Chen et al., 2025; Cai et al., 2026; Zheng et al., 2026b). This behavior induces mode collapse, leading to reduced diversity in the generated videos. Second, the intermediate rollout predictions receive no explicit training signal at each step, leading to error accumulation. Recent methods mitigate these issues by explicitly incorporating real data into the training objective (Yin et al., 2024a; Chen et al., 2026; Liu et al., 2026), but they rely on complex data-curation processes or additional post-training, which further complicates the multi-stage training pipeline. Another line of research incorporates reinforcement learning post-training to distilled AR models (Zhang et al., 2026), but its performance still relies on reward models and is sensitive to multiple hyperparameters. A complementary line of work balances mode-seeking and mode-covering objectives (Cai et al., 2026; Zheng et al., 2026b; Li et al., 2026a), but the generated videos can still appear over-saturated and lack fine-grained visual details. This begs the question: can we improve AR video diffusion distillation without real data curation or additional post-training stages?

To address this question, we focus on the self-rollout trajectory, which determines both the student samples exposed to DMD and the intermediate predictions reused in subsequent denoising and autoregressive conditioning. Perturbing the rollout trajectory can expose DMD to a broader range of student samples, allowing it to provide learning signals beyond the modes the student already covers. Moreover, cleaner tokens can serve as context for denoising noisier tokens (Chefer et al., 2026), helping improve intermediate rollout predictions and reduce error accumulation. Based on these motivations, we propose Mask Forcing, which injects randomly masked cleaner tokens into the rollout input to mitigate both mode-seeking behavior and error accumulation. At each rollout transition, Mask Forcing randomly samples a mask and an additional timestep corresponding to a lower noise level than the original one and injects the cleaner tokens into the rollout input. Specifically, the latent positions selected by the mask are re-noised to this lower noise level, while the remaining positions are kept at the original level, producing a dual-noise rollout input. The conditioning timestep fed to the model remains the original one, requiring the model to denoise inputs whose local noise levels are partially inconsistent with the global timestep. Randomizing the mask and the lower-noise timestep diversifies student rollout trajectories and encourages exploration of more modes in the teacher distribution during DMD training. Beyond perturbing rollout trajectories, Mask Forcing uses lower-noise tokens as context for denoising noisier tokens. This aligns with the observation in Self-Flow (Chefer et al., 2026) that cleaner context in mixed-noise inputs can facilitate denoising, thereby improving intermediate predictions and reducing error accumulation during self-rollout.

Extensive experiments across multiple AR video distillation methods validate the effectiveness of our method in both chunk-wise and frame-wise settings. Representative results in Fig. 1 show that our method significantly improves the visual quality against various baselines, with enhanced real-

ism and richer fine-grained details. Comprehensive evaluations demonstrate improvements across multiple benchmarks. In summary, our contributions are as follows:

- We propose Mask Forcing, a simple and effective approach that alleviates the mode-seeking behavior of the reverse-KL objective in self-rollout DMD training for AR video diffusion distillation, without incorporating real video data or additional post-training stages.
- We introduce a Dual-Noise Masking Rollout strategy that injects lower-noise signals into noisy rollout inputs through random masks along spatial and temporal axes. Such perturbations diversify student rollout trajectories to cover more teacher modes, while providing cleaner context for denoising to reduce error accumulation.
- Extensive experiments on multiple AR video distillation methods demonstrate the effectiveness of Mask Forcing in both chunk-wise and frame-wise settings, with significantly improved visual quality and faster convergence. Comprehensive ablations further validate the effects of different masking mechanisms.

2 Related Work

Autoregressive Video Generation. Autoregressive video diffusion models enable real-time video generation by sequentially generating videos conditioned on historical context. Teacher Forcing (Hu et al., 2024; Gao et al., 2025) denoises the current chunk conditioned on clean ground-truth context, which suffers from a train-test gap and exposure bias. Diffusion Forcing (Chen et al., 2024) assigns each frame with independently sampled noise to improve robustness to imperfect histories during rollout, but still fails to align training with inference. Self Forcing (Huang et al., 2026) bridges this gap by performing AR self-rollout on self-generated histories and distills a bidirectional teacher into the causal student via a DMD loss. Causal Forcing (Zhu et al., 2026) further uses an AR teacher for ODE initialization to reduce the architecture gap. LongLive (Yang et al., 2025) extends AR generation to long videos via short window attention with frame sink and streaming long tuning. Despite these advances, distilled AR models still exhibit limited visual quality and realism. Mask Forcing builds on AR distillation with self-rollout and improves visual quality via dual-noise masking.

Mode Seeking and Mode Covering in Video Diffusion Distillation. DMD and DMD2 (Yin et al., 2024a) use score-based reverse KL to match student and teacher distributions. This mode-seeking objective improves sample fidelity but restricts the student to limited teacher modes. DMD2 further introduces an adversarial loss on real data. In contrast, trajectory-based consistency objectives are associated with the mode-covering behavior of forward divergence (Song et al., 2023; Kim et al., 2024; Lu & Song, 2025). Such objectives cover more teacher modes, but may average across modes and reduce sample quality. Recent methods seek to balance these behaviors through complementary training objectives. Mode Seeking meets Mean Seeking (Cai et al., 2026) uses separate heads for supervised flow matching and reverse-KL distribution matching. rCM (Zheng et al., 2026b) augments continuous-time consistency with score distillation regularization. DistillAlign (Li et al., 2026a) jointly optimizes DMD and a consistency distillation loss. Mask Forcing instead retains the original DMD and perturbs student rollouts via dual-noise masking, broadening student distribution exposed to DMD and enabling learning signals from teacher modes that standard self-rollout rarely explores.

Masked Modeling. Masked modeling is a powerful paradigm for representation and generative learning. MAE (He et al., 2022) and BEiT (Bao et al., 2021) learn visual representations by reconstructing masked pixels or tokens. MaskGIT (Chang et al., 2022) generates images through iterative masked-token prediction. Masking can also be realized through heterogeneous noise levels that control the information retained by tokens. For multi-modal generation, Self-Flow introduces a self-supervised framework for flow matching that combines mix-timestep scheduling with masking for representation alignment. In AR generation, Diffusion Forcing associates each frame with a random and independent noise level, which can be regarded as partial masking along the time axis. Inspired by these works, we introduce dual-noise masking into AR video distillation to diversify self-rollout trajectories and provide cleaner context for denoising noisier tokens.

3 METHOD

3.1 PRELIMINARIES

Autoregressive (AR) Video Generation. An AR video model represents a video as a sequence of F chunks $x_{1:F} = (x_1, \ldots, x_F)$, where each chunk includes one or more latent frames. It factorizes the text-conditioned joint distribution as $p(x_{1:F} \mid c) = \prod_{i=1}^{F} p(x_i \mid x_{<i}, c)$, where c denotes the text prompt and $x_{<i} = (x_1, \ldots, x_{i-1})$ denotes the preceding video chunks. Each conditional distribution $p(x_i \mid x_{<i}, c)$ is modeled using a diffusion process following the flow-matching formulation, where the noisy sample is defined as $x_i^t = (1-t)x_i + t\epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, I)$ and $t \in [0, 1]$. Each video chunk is generated through this denoising process conditioned on c and the historical context $x_{<i}$, which is stored in the key-value (KV) cache. Self Forcing mitigates the train-test gap by unrolling the student on its own generated history during training, with $x_{1:F}^{\theta} \sim \prod_{i=1}^{F} p_\theta(x_i \mid x_{<i}, c)$.

Distribution Matching Distillation (DMD). DMD (Yin et al., 2024b;a) distills a pretrained multi-step teacher model into a few-step student model G_θ by minimizing the reverse KL divergence from the student generator induced distribution p_{fake} to the teacher distribution p_{real}. Specifically, DMD adopts a reverse KL objective, whose gradient is used to update the student model:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = \mathbb{E}[\nabla_\theta D_{KL}(p_{\text{fake},t} || p_{\text{real},t})], \tag{1}$$

$$D_{KL}(p_{\text{fake},t} || p_{\text{real},t}) = \mathbb{E}\left[\log \frac{p_{\text{fake},t}(x_t)}{p_{\text{real},t}(x_t)}\right] = -\mathbb{E}[\log p_{\text{real},t}(x_t) - \log p_{\text{fake},t}(x_t)], \tag{2}$$

where x_t is a noisy sample corresponding to timestep t: $x_t = (1-t)G_\theta(z) + t\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, and z is latent drawn from random noise. The gradient to update G is formulated as:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = \mathbb{E}\left[-(s_{\text{real}}(x_t, t) - s_{\text{fake}}(x_t, t))\frac{dG}{d\theta}\right], \tag{3}$$

where $s_{\text{real}}(x_t, t) = \nabla_{x_t} \log p_{\text{real},t}(x_t)$, $s_{\text{fake}}(x_t, t) = \nabla_{x_t} \log p_{\text{fake},t}(x_t)$ denote the scores of the teacher and student distributions, respectively. However, reverse-KL is mode-seeking, which can limit teacher-mode coverage and degrade visual realism in AR distillation (Chen et al., 2026).

3.2 DUAL-NOISE MASKING ROLLOUT

Overview. We propose a dual-noise masking rollout strategy to mitigate mode collapse induced by the mode-seeking behavior of reverse KL in self-rollout DMD training for AR diffusion distillation. The core idea is to inject low-noise signals into noisy rollout inputs during the rollout process via random masks applied both within and across chunks. These cleaner signals serve two purposes: perturbing student rollouts to encourage broader teacher-mode coverage and providing context for denoising noisier tokens. An overview of our method is shown in Fig. 2.

3.2.1 SELF-ROLLOUT WITH DUAL-NOISE MASKING

An AR video diffusion model represents a video as a sequence of F chunks $x = (x_1, x_2, ..., x_F)$, where each chunk contains one or more latent frames. Using the flow-matching formulation, a noisy chunk is defined as:

$$x_i^t = (1-t)\, x_i^0 + t\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{4}$$

where $t \in [0, 1]$ denotes the timestep, which interpolates between the clean chunk at $t = 0$ and pure Gaussian noise at $t = 1$. During the self-rollout training, the student model G_θ generates the video chunk by chunk and each chunk is produced by iterative denoising over a fixed schedule of T timesteps $\{t_1, \ldots, t_T\}$ selected from the $N_t = 1000$ training timesteps, where $t_T = 1$ denotes pure noise, $t_0 = 0$ denotes the clean output. We use $T = 4$ denoising steps throughout training. For each chunk index i, at denoising timestep t_j, the student model predicts a clean chunk estimate $\hat{x}_{i,j}^0$ from the noisy chunk $x_i^{t_j}$, conditioned on the previously generated clean chunks $x_{<i}$, the text prompt c, and the timestep t_j:

$$\hat{x}_{i,j}^0 = G_\theta\left(x_i^{t_j} \mid x_{<i}, c, t_j\right). \tag{5}$$

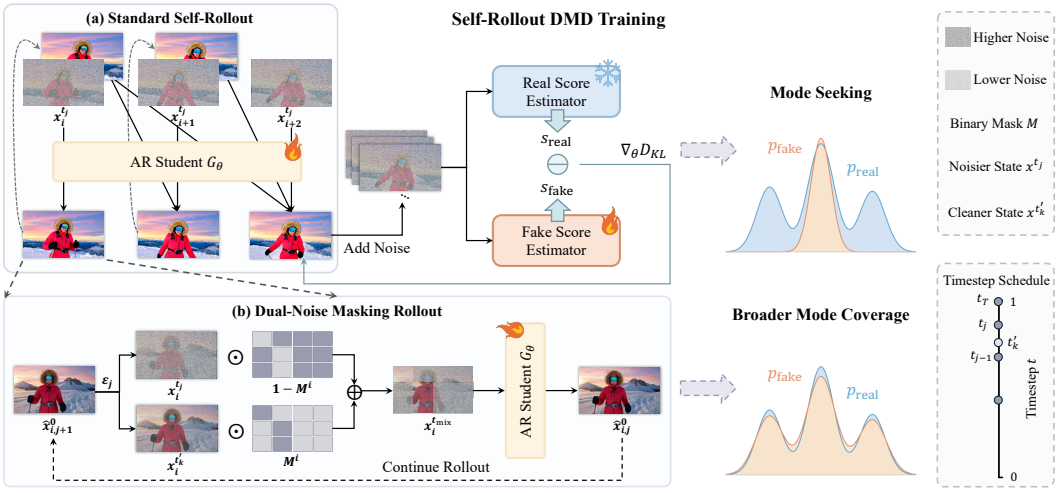

Figure 2: (a) Standard self-rollout DMD training suffers from mode collapse due to the mode-seeking reverse-KL objective, resulting in over-saturation and limited realism. (b) Mask Forcing perturbs the student rollout trajectory through dual-noise masking rollout, encouraging broader teacher-mode coverage while providing cleaner context for denoising noisy tokens.

The DMD objective is evaluated only on the completed self-rollout, providing no explicit training signal for intermediate predictions. Since intermediate predictions are reused in subsequent denoising steps and as context for later chunks, their errors can accumulate throughout the rollout. Meanwhile, the reverse-KL objective is mode-seeking and concentrates the student rollout distribution on a limited subset of teacher modes. Together, these issues degrade visual quality and realism.

We address this issue by injecting cleaner signals into the student model input to perturb the student rollout trajectory, encouraging the student rollouts to reach more high-density regions of the teacher distribution. Specifically, we adopt a dual-timestep scheduling strategy inspired by Self-Flow. We retain the denoising schedule of the base model and define Δ as the size of the timestep window measured in training timestep units, corresponding to a normalized width of Δ/N_t. At each denoising timestep t_j, we uniformly sample an additional timestep t'_k within the corresponding window in timestep space:

$$\ell_j = \max\left(t_{\min},\ t_j - \frac{\Delta}{N_t}\right), \tag{6}$$

$$t'_k \sim \mathrm{Uniform}\big([\ell_j, t_j]\big). \tag{7}$$

Here, t'_k corresponds to a lower noise level than t_j, and the floor $t_{\min}$ prevents the injected signal from being nearly clean.

For each chunk i, we construct a binary mask M^i with a masking ratio α. The mask is sampled independently across frames within a chunk (spatial axis) and across chunks (temporal axis), so that different frames of the same chunk and different chunks receive different dual-noise patterns. At the j-th denoising step, corresponding to timestep t_j, we sample $\epsilon_j \sim \mathcal{N}(0, I)$ and use it to re-noise the clean prediction from the previous step $\hat{x}^0_{i,j+1}$ at t'_k and t_j, yielding the cleaner sample $x_i^{t'_k}$ and the noisier sample $x_i^{t_j}$, respectively:

$$x_i^{t'_k} = (1 - t'_k)\,\hat{x}^0_{i,j+1} + t'_k\,\epsilon_j, \tag{8}$$

$$x_i^{t_j} = (1 - t_j)\,\hat{x}^0_{i,j+1} + t_j\,\epsilon_j. \tag{9}$$

We use the random mask M^i to select tokens from the cleaner sample at masked positions ($M^i = 1$), while retaining tokens from the original noisier sample at the remaining positions. This yields the dual-noise input:

$$x_i^{t_{\mathrm{mix}}} = M^i \odot x_i^{t'_k} + (1 - M^i) \odot x_i^{t_j}. \tag{10}$$

Consequently, the student generator G_θ receives the dual-noise input at denoising step j and produces the updated clean prediction:

$$\hat{x}^0_{i,j} = G_\theta\left(x_i^{t_{\text{mix}}} \mid x_{<i}, c, t_j\right). \tag{11}$$

During training, the student remains conditioned on the scheduled timestep t_j, treating the lower-noise tokens as input perturbations, while inference follows the original denoising schedule. At the next denoising step, $\hat{x}^0_{i,j}$ is re-noised at t_{j-1} and a newly sampled lower-noise timestep within its corresponding timestep window, then mixed by mask M^i to form the next dual-noise input. At the sampled exit step s, the clean prediction $\hat{x}^0_{i,s}$ updates the KV cache to condition subsequent chunks.

This construction broadens the rollout samples evaluated by DMD, enabling learning signals in teacher-supported regions underexplored by standard self-rollout. Meanwhile, lower-noise tokens provide context for denoising noisier tokens, improving intermediate predictions and reducing error accumulation. The detailed algorithm is provided in Appendix Alg. 1.

3.2.2 Distributional Analysis

We analyze the student rollout distribution induced by dual-noise masking. We characterize it as a mixture over masking trajectories and decompose its reverse-KL objective.

Fix a text condition c and a DMD score-noising timestep τ. Let X_τ be the completed student rollout noised at τ and let p_τ be the corresponding teacher noisy marginal. Let $V \sim \pi$ collect all masks and lower-noise timesteps sampled during rollout, with π determined by masking hyperparameters.

Conditioned on a masking trajectory $V = v$, $q^v_{\theta,\tau}(x)$ denotes the corresponding student distribution. Averaging over V gives the student rollout mixture $\bar{q}_{\theta,\tau}(x) := \mathbb{E}_{V\sim\pi}[q^V_{\theta,\tau}(x)]$. Since the fake-score estimator is not conditioned on V, DMD matches this mixture rather than its trajectory-conditioned components $q^v_{\theta,\tau}$. The reverse-KL objective of the rollout mixture can be decomposed as

$$\underbrace{D_{\text{KL}}\left(\bar{q}_{\theta,\tau}\|p_\tau\right)}_{\text{marginal KL}} = \underbrace{\mathbb{E}_{V\sim\pi}\left[D_{\text{KL}}\left(q^V_{\theta,\tau}\|p_\tau\right)\right]}_{\text{teacher alignment}} - \underbrace{\mathbb{E}_{V\sim\pi}\left[D_{\text{KL}}\left(q^V_{\theta,\tau}\|\bar{q}_{\theta,\tau}\right)\right]}_{\text{rollout-distribution diversity}}. \tag{12}$$

The second term in this decomposition measures how much the rollout distributions induced by different masking trajectories deviate from their marginal mixture. When different perturbations induce similar rollout distributions, this term is small. When different perturbations induce distinct rollout distributions, this term can be larger. Therefore, diversifying the modes reached across rollout trajectories can reduce the marginal reverse KL, while the first term measures their average deviation from the teacher distribution. The masking hyperparameters control the perturbation strength. Weak perturbations yield similar rollout distributions and limited additional coverage, whereas strong perturbations may shift the rollout distributions away from teacher-supported regions, weakening teacher alignment. Appropriate perturbation strength encourages broader mode coverage while maintaining teacher alignment. The derivation is provided in Appendix A and ablations are presented in Sec. 4.2.

4 Experiments

Implementation Details. We adopt Self Forcing (Huang et al., 2026), LongLive (Yang et al., 2025), and Causal Forcing (Zhu et al., 2026) as baselines, using Wan2.1-T2V-1.3B (Wan et al., 2025) as the base model and Wan2.1-T2V-14B as the teacher. Our method is applied during self-rollout DMD training. We evaluate chunk-wise (3 latent frames per chunk) and frame-wise settings, generating 81-frame videos at 832 $\times$ 480 with training prompts sampled from the VidProM dataset (Wang & Yang, 2024). Since frame-wise ODE initialization models are unavailable for Self Forcing and LongLive, we train them on ODE pairs distilled from the bidirectional teacher. For Causal Forcing, we use its causal model initialized via ODE distillation from the AR teacher. We set the mask ratio α to 0.2, timestep window Δ to 250, and floor $t_{\min}$ to schedule index 20. Training takes approximately 1.5k steps and 14 hours on 8 GPUs. All other training settings remain unchanged for each baseline.

Evaluation. We evaluate on the motion-rich 100-prompt set from Causal Forcing and the VBench benchmark. For the 100-prompt set, we report HPSv3 (Ma et al., 2025) for visual quality and use VisionReward (Xu et al., 2026) to report vision reward score (Vision.), instruction following (Instruct.),



Figure 3: Qualitative comparison of our method (+Ours) against different baselines. Visual results show that our method generates videos with higher visual quality and greater realism.

Table 1: Quantitative results on the 100-prompt set and VBench benchmarks under chunk-wise and frame-wise settings.

	Method	HPSv3 ↑	Vision. ↑	Instruct. ↑	MQ ↑	Dynamic. ↑	Total ↑	Quality ↑	Semantic ↑
Chunk-wise	Self Forcing	9.55	10.10	38.50	15.88	70	81.89	82.99	77.49
	+ Ours	9.84$_{+.29}$	11.37$_{+1.27}$	45.03$_{+6.53}$	20.49$_{+4.61}$	82$_{+12}$	82.61$_{+.72}$	83.68$_{+.69}$	78.74$_{+1.25}$
	Causal Forcing	9.37	10.36	40.41	17.73	76	82.67	83.58	78.98
	+ Ours	10.17$_{+.80}$	11.58$_{+1.22}$	46.30$_{+5.89}$	21.54$_{+3.81}$	82$_{+6}$	82.76$_{+.09}$	83.69$_{+.11}$	79.01$_{+.03}$
	LongLive	9.11	10.77	42.48	21.20	76	82.02	82.87	78.66
	+ Ours	10.14$_{+1.03}$	11.00$_{+.23}$	42.65$_{+.17}$	22.34$_{+1.14}$	69$_{-7}$	82.75$_{+.73}$	83.71$_{+.84}$	78.91$_{+.25}$
Frame-wise	Self Forcing	9.34	9.45	35.62	18.27	53	80.73	81.72	76.78
	+ Ours	9.79$_{+.45}$	10.49$_{+1.04}$	39.60$_{+3.98}$	19.07$_{+.80}$	61$_{+8}$	81.49$_{+.76}$	82.55$_{+.83}$	77.23$_{+.45}$
	Causal Forcing	9.67	10.58	37.56	20.25	28	80.64	81.55	77.00
	+ Ours	9.96$_{+.29}$	10.71$_{+.13}$	39.60$_{+2.04}$	23.69$_{+3.44}$	52$_{+24}$	82.28$_{+1.64}$	83.19$_{+1.64}$	78.62$_{+1.62}$
	LongLive	9.19	9.35	38.50	12.14	25	80.97	81.86	77.41
	+ Ours	9.46$_{+.27}$	10.55$_{+1.20}$	42.78$_{+4.28}$	19.32$_{+7.18}$	76$_{+51}$	81.47$_{+.50}$	82.33$_{+.47}$	78.05$_{+.64}$

and motion quality (MQ) scores following Causal Forcing. Dynamic Degree (Dynamic.) is used to evaluate motion dynamics. All scores are scaled by 100 for readability except HPSv3. For VBench, we report Total, Quality, and Semantic Scores with 946 standard prompts. For long-video generation, we further compare with LongLive on the first 100 prompts from the MovieGen (Polyak et al., 2024) extended version and VBench-Long on 30-second video generation following LongLive.

4.1 Comparisons with Baselines

Quantitative Comparisons. As shown in Tab. 1, Mask Forcing consistently improves visual quality, semantic alignment, and motion quality across all baselines in both chunk-wise and frame-wise settings, while generally increasing Dynamic Degree. These gains suggest that the dual-noise masking rollout strategy helps mitigate mode-collapse artifacts. Though the Dynamic Degree for LongLive is lower, this can be attributed to generic VBench rewarding drift-induced optical flow (Minar et al., 2026). For VBench, including Mask Forcing surpasses all baselines across the VBench metrics in both settings. Moreover, we evaluate Mask Forcing on single-prompt long video generation with LongLive, integrating it into both initialization and streaming long-video tuning stages. As shown in Tab. 2, our method effectively improves visual quality in the long video generation setting.

Qualitative Comparisons. Fig. 3 presents qualitative comparisons of our method and baselines on the 100-prompt set and MovieGen. The baselines exhibit limited visual quality, with over-saturation

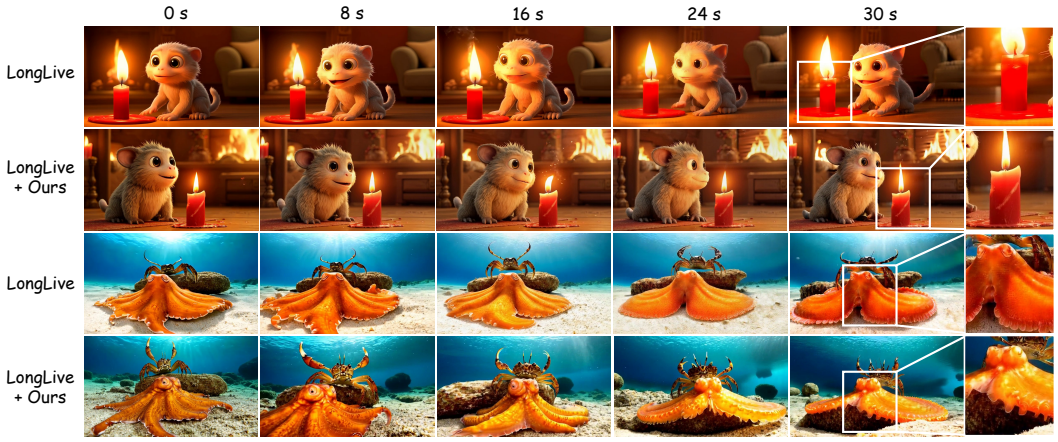

Figure 4: Qualitative results of single-prompt long-video setting on LongLive. Our method improves video quality with fine-grained visual details, such as candle and octopus textures.

Table 2: Single-prompt 30s long video generation on MovieGen and VBench-Long benchmarks.

Method	HPSv3 ↑	Vision. ↑	Instruct. ↑	MQ ↑	Dynamic. ↑	Total ↑	Quality ↑	Semantic ↑
LongLive	8.44	14.31	62.04	18.84	**70**	83.91	84.71	80.70
+ Ours	**9.11**$_{+.67}$	**14.93**$_{+.62}$	**66.02**$_{+3.98}$	**19.37**$_{+.53}$	64$_{-6}$	**84.51**$_{+.60}$	**85.28**$_{+.57}$	**81.44**$_{+.74}$

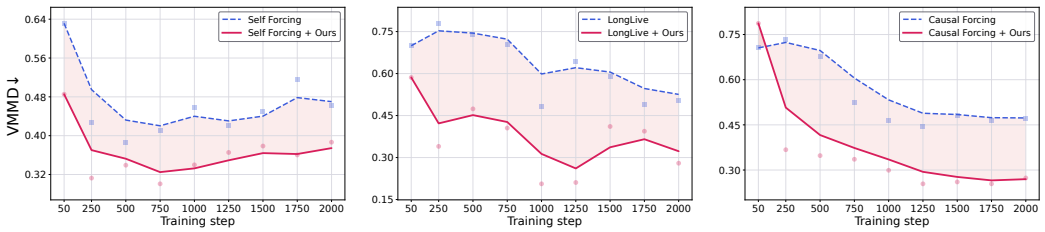

Figure 5: V-JEPA2 Maximum Mean Discrepancy (VMMD) over training steps on the 100-prompt set. Incorporating Mask Forcing leads to faster convergence across various baselines.

and over-smoothing artifacts. Specifically, Self Forcing produces the woman's face with unnaturally high-contrast coloration and the creature with a visually flat appearance, while LongLive loses fine details. The rock-man and kitchen examples from Causal Forcing show severe over-saturation. Mask Forcing alleviates these issues, enhancing visual quality and realism. Qualitative results for long video generation are shown in Fig. 4, where Mask Forcing improves visual quality with richer high-frequency details. More qualitative results are provided in Appendix G and supplementary file.

Convergence Speed. We compare convergence of different baselines with and without Mask Forcing on the 100-prompt set in Figs. 1 and 5 using HPSv3 for visual quality and maximum mean discrepancy (MMD) (Jayasumana et al., 2024) in the CLIP (Radford et al., 2021) and V-JEPA2 (Assran et al., 2025) feature spaces (CMMD and VMMD) for student–teacher distributional alignment. CMMD and VMMD are computed between a fixed teacher reference set and student outputs at each training checkpoint using the same 100-prompt set. Figs. 1 and 5 show that Mask Forcing improves convergence in both visual quality and distributional alignment across all baselines.

Human Evaluation. We conduct pairwise human evaluations between anonymized outputs from Mask Forcing (+Ours) against different baselines for short and long video generation with 24 users. Each user is required to select the video with higher visual quality and realism from each video pair. Fig. 6 shows that incorporating Mask Forcing receives 80%, 79%, and 83% of preference votes over Self Forcing, Causal Forcing, and LongLive baselines. For long video generation, Mask Forcing receives 72% of preference votes over LongLive, demonstrating the effectiveness of our method.

Table 3: Mask ratio α ablation.

α	HPSv3 ↑	Dynamic. ↑
0.1	10.00	47
0.2	9.84	82
0.3	9.82	65
0.4	10.15	57
0.5	10.17	44

Table 4: Timestep window Δ ablation.

Δ	HPSv3 ↑	Dynamic. ↑
50	9.16	92
150	9.82	70
250	9.84	82
450	9.98	80
600	9.90	67

Table 5: Mask scheme ablation.

Mask scheme	HPSv3 ↑	Dynamic. ↑
shared, per-rollout	10.12	56
shared, per-chunk	9.84	81
shared, per-step	9.64	72
per-frame, per-rollout	10.03	53
per-frame, per-chunk	9.84	82
per-frame, per-step	9.93	74

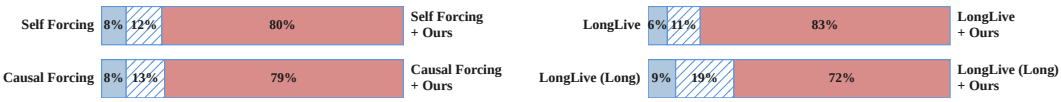

Figure 6: Pairwise human preferences for baselines without (blue) and with (red) Mask Forcing, while hatched regions denote ties.

4.2 Ablation Studies

We ablate Mask Forcing on the 100-prompt set with Self Forcing. As motion may decrease in DMD training, we report both HPSv3 and Dynamic Degree to assess visual quality and motion dynamics.

Mask Ratio α. We compare different mask ratios in Tab. 3. With a small ratio of 0.1, the masking trajectories only induce similar rollouts and provide limited exploration beyond the modes already covered by the student, achieving an HPSv3 score of 10.00 but a low Dynamic Degree of 47. Large ratios of 0.4-0.5 introduce strong guidance from cleaner tokens, yielding the highest HPSv3 scores of 10.15 and 10.17 but reducing Dynamic Degree to 57 and 44, respectively. This suggests that excessive reliance on cleaner tokens restricts the model's ability to revise or diversify motion patterns in subsequent denoising. We therefore adopt $\alpha = 0.2$ to balance visual quality and motion dynamics.

Timestep Window Δ. Tab. 4 ablates the timestep window, which controls the maximum difference between the original and lower noise levels. A small window of $\Delta = 50$ provides limited perturbation and denoising cues, yielding a high Dynamic Degree of 92 but a low HPSv3 score of 9.16. Increasing Δ strengthens rollout perturbations and provides more informative guidance from cleaner tokens, improving HPSv3 to 9.82–9.98 for $\Delta = 150$–450. However, at $\Delta = 600$, excessive cleaner-token guidance may dominate subsequent denoising and limit motion diversity. We adopt $\Delta = 250$, achieving a competitive HPSv3 score of 9.84 and a high Dynamic Degree of 82.

Mask Scheme. We ablate the mask schemes for chunk-wise generation in Tab. 5. Spatially, "shared" uses one mask across frames within a chunk and "per-frame" samples masks independently. Temporally, masks are fixed per rollout ("per-rollout"), resampled per chunk ("per-chunk") or denoising step ("per-step"). These mask choices control how mask-induced perturbations vary throughout the rollout. An effective scheme should diversify the student rollout trajectories to encourage broader coverage of the teacher modes while avoiding excessive variation that deviates from the teacher distribution. Tab. 5 shows that the per-frame, per-chunk scheme achieves both high visual quality and motion dynamics, with a HPSv3 score of 9.84 and a Dynamic Degree of 82.

5 Conclusion

In this paper, we introduce Mask Forcing, a Dual-Noise Masking Rollout strategy for AR video diffusion distillation that alleviates mode collapse induced by the mode-seeking reverse KL objective in DMD. By injecting randomly masked lower-noise signals into the rollout inputs, Mask Forcing diversifies student trajectories and exposes DMD to a broader range of student samples, enabling it to guide the student toward teacher modes rarely explored under standard self-rollout. Meanwhile, cleaner tokens provide context to denoise noisier tokens, improving intermediate rollout predictions and mitigating error accumulation during the self-rollout process. Extensive experiments across multiple baselines demonstrate improved visual quality and faster convergence without incorporating real video data or post-training.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and improve the readability of the manuscript, to prepare preliminary versions of figures and visualizations, and to provide limited assistance with code development and debugging. All AI-assisted figures and code were subsequently reviewed, modified, and verified by the authors. Generative AI tools were not used to formulate the main research ideas, design the methodology or experiments, interpret experimental results, or draw scientific conclusions. The authors take full responsibility for the final content of this work.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Shengqu Cai, Weili Nie, Chao Liu, Julius Berner, Lvmin Zhang, Nanye Ma, Hansheng Chen, Maneesh Agrawala, Leonidas Guibas, Gordon Wetzstein, et al. Mode seeking meets mean seeking for fast long video generation. *arXiv preprint arXiv:2602.24289*, 2026.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*, 2025.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation. *arXiv preprint arXiv:2606.18478*, 2026.

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. In *ICML*. PMLR, 2025.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

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.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Anthony Hu, Václav Volhejn, Adrien Ramanana Rahary, Chris Mulder, Aditya Makkar, Amélie Royer, Manu Orsini, Alyx Liao, Adam Jelley, Eloi Alonso, et al. Multiplayer interactive world models with representation autoencoders. *arXiv preprint arXiv:2607.05352*, 2026.

Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. Acdit: Interpolating autoregressive conditional modeling and diffusion transformer. *arXiv preprint arXiv:2412.07720*, 2024.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315. IEEE, 2024.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

Dongjun Kim, Chieh-Hsin Lai, WeiHsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *International Conference on Learning Representations*, volume 2024, pp. 44493–44525, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Jiaxing Li, Kai Zou, Cindy Zhou, Kaichen Huang, Junyao Gao, Zile Wang, Yang Liu, Bin Liu, Bo An, and Yangguang Li. Distillalign: Coordinating mode covering and mode seeking in autoregressive video distillation. *arXiv preprint arXiv:2607.26811*, 2026a.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *Advances in Neural Information Processing Systems*, 38, 2026b.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Hongyu Liu, Chun Wang, Feng Gao, Xuanhua He, Yue Ma, Ziyu Wan, Yong Zhang, Xiaoming Wei, and Qifeng Chen. Opsd-v: On-policy self-distillation for post-training few-step autoregressive video generators. *arXiv preprint arXiv:2607.08766*, 2026.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

Matiur Rahman Minar, Seunghun Oh, Ganghyeon Jeong, and Unsang Park. Steady-forcing: Balancing spatial persistence and motion continuity in long-horizon nature video diffusion. *arXiv preprint arXiv:2606.14732*, 2026.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026a.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026b.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026a.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026b.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=R3JzLU2qZj.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Songchun Zhang, Zeyue Xue, Siming Fu, Jie Huang, Xianghao Kong, Y Ma, Haoyang Huang, Nan Duan, and Anyi Rao. Astrolabe: Steering forward-process reinforcement learning for distilled autoregressive video models. *arXiv preprint arXiv:2603.17051*, 2026.

Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Wenqiang Sun, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, et al. minwm: A full-stack open-source framework for real-time interactive video world models. *arXiv preprint arXiv:2605.30263*, 2026.

Guangcong Zheng, Teng Li, Xianpan Zhou, and Xi Li. Realcam-vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. *arXiv preprint arXiv:2504.08212*, 2025. URL https://arxiv.org/abs/2504.08212.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026a.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=2uNlM353RI.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, Haoran Li, Haoyang Huang, Zeyue Xue, and Nan Duan. Self gradient forcing: Native long video extrapolation, 2026. URL https://arxiv.org/abs/2607.20368.

Appendix Overview

The appendix provides the theoretical justification, comparisons with joint distillation methods, additional long video and diversity evaluations, an application to camera-controlled video generation, algorithmic details, and additional qualitative results. It is organized as follows:

- Sec. A provides the theoretical justification for Mask Forcing.
- Sec. B compares Mask Forcing with joint distillation methods Causal-rCM (Zheng et al., 2026a) and DistillAlign (Li et al., 2026a).
- Sec. C provides additional evaluation of LongLive over different intervals of long video generation.
- Sec. D evaluates video diversity across different autoregressive video distillation baselines.
- Sec. E presents an application of Mask Forcing to camera-controlled image-to-video generation.
- Sec. F provides the detailed training algorithm for Mask Forcing.
- Sec. G presents additional qualitative comparisons across Self Forcing, LongLive, and Causal Forcing.

A THEORETICAL JUSTIFICATION

In this section, we characterize the distribution induced by Mask Forcing, derive an exact decomposition of its reverse-KL objective, and discuss its implications for mode coverage. Fix a text condition c and a DMD score-noising timestep $\tau > 0$. Let X_τ denote the completed student rollout noised at τ, and let x denote a generic realization of X_τ. Define

$$p_\tau(x) := p_{\mathrm{real},\tau}(x \mid c) \tag{13}$$

as the corresponding teacher noisy marginal. Let

$$V := \{(M_r, t'_{k,r})\}_{r=1}^{R}, \qquad V \sim \pi, \tag{14}$$

collect the masks and cleaner timesteps sampled during one Mask Forcing rollout, where π is the distribution induced by the corresponding samplers.

Given a masking trajectory $V = v$, let $q^v_{\theta,\tau}(x)$ denote the corresponding trajectory-conditioned student distribution of X_τ. Marginalizing over V gives the student rollout mixture

$$\bar{q}_{\theta,\tau}(x) := \mathbb{E}_{V\sim\pi}\left[q^V_{\theta,\tau}(x)\right]. \tag{15}$$

Under exact score estimation, the real and fake scores at noise level τ are

$$s_{\mathrm{real}}(x,\tau,c) = \nabla_x \log p_\tau(x), \qquad s_{\mathrm{fake}}(x,\tau,c) = \nabla_x \log \bar{q}_{\theta,\tau}(x). \tag{16}$$

Since the fake-score model is not conditioned on V, it estimates the score of the marginal distribution $\bar{q}_{\theta,\tau}$ rather than a component score $\nabla_x \log q^v_{\theta,\tau}$.

The DMD score-difference update corresponds to the gradient

$$\nabla_\theta D_{\mathrm{KL}}\left(\bar{q}_{\theta,\tau}\|p_\tau\right) = \mathbb{E}\left[\left(s_{\mathrm{fake}}(X_\tau,\tau,c) - s_{\mathrm{real}}(X_\tau,\tau,c)\right)\nabla_\theta X_\tau\right]. \tag{17}$$

The exact-score assumption is used only for this identification. In practice, both score networks approximate the corresponding marginal scores.

Proposition (Reverse-KL decomposition for the student rollout mixture). Suppose that the KL divergences below are finite. Then

$$D_{\mathrm{KL}}\left(\bar{q}_{\theta,\tau}\|p_\tau\right) = \mathbb{E}_V\left[D_{\mathrm{KL}}\left(q^V_{\theta,\tau}\|p_\tau\right)\right] - \mathbb{E}_V\left[D_{\mathrm{KL}}\left(q^V_{\theta,\tau}\|\bar{q}_{\theta,\tau}\right)\right]. \tag{18}$$

Proof. By the definition of KL divergence, the average conditional KL can be expanded by multiplying and dividing the density ratio by $\bar{q}_{\theta,\tau}$:

$$\begin{aligned}
&\mathbb{E}_V D_{\mathrm{KL}}\left(q^V_{\theta,\tau}\|p_\tau\right) \\
&= \int \pi(v) \int q^v_{\theta,\tau}(x) \log \frac{q^v_{\theta,\tau}(x)}{p_\tau(x)}\,\mathrm{d}x\,\mathrm{d}v \\
&= \mathbb{E}_{V\sim\pi}\left[\mathbb{E}_{X_\tau|V,c}\left[\log \frac{q^V_{\theta,\tau}(X_\tau)}{p_\tau(X_\tau)}\right]\right] \\
&= \mathbb{E}_{V,X_\tau|c}\left[\log \frac{q^V_{\theta,\tau}(X_\tau)}{p_\tau(X_\tau)}\right] \\
&= \mathbb{E}_{V,X_\tau|c}\left[\log \frac{q^V_{\theta,\tau}(X_\tau)}{\bar{q}_{\theta,\tau}(X_\tau)} + \log \frac{\bar{q}_{\theta,\tau}(X_\tau)}{p_\tau(X_\tau)}\right] \\
&= \mathbb{E}_{V,X_\tau|c}\left[\log \frac{q^V_{\theta,\tau}(X_\tau)}{\bar{q}_{\theta,\tau}(X_\tau)}\right] + \mathbb{E}_{V,X_\tau|c}\left[\log \frac{\bar{q}_{\theta,\tau}(X_\tau)}{p_\tau(X_\tau)}\right] \\
&= \mathbb{E}_V D_{\mathrm{KL}}\left(q^V_{\theta,\tau}\|\bar{q}_{\theta,\tau}\right) + D_{\mathrm{KL}}\left(\bar{q}_{\theta,\tau}\|p_\tau\right).
\end{aligned} \tag{19}$$

For the first term, expanding the joint expectation and applying the definition of KL divergence gives

$$\begin{aligned}
&\mathbb{E}_{V,X_\tau|c}\left[\log \frac{q^V_{\theta,\tau}(X_\tau)}{\bar{q}_{\theta,\tau}(X_\tau)}\right] \\
&= \int \pi(v)\left[\int q^v_{\theta,\tau}(x)\log\frac{q^v_{\theta,\tau}(x)}{\bar{q}_{\theta,\tau}(x)}\,\mathrm{d}x\right]\mathrm{d}v \\
&= \int \pi(v) D_{\mathrm{KL}}\left(q^v_{\theta,\tau}\|\bar{q}_{\theta,\tau}\right)\,\mathrm{d}v \\
&= \mathbb{E}_V D_{\mathrm{KL}}\left(q^V_{\theta,\tau}\|\bar{q}_{\theta,\tau}\right). \qquad (20)
\end{aligned}$$

For the second term, we first expand the joint expectation and then marginalize over V:

$$\begin{aligned}
&\mathbb{E}_{V,X_\tau|c}\left[\log \frac{\bar{q}_{\theta,\tau}(X_\tau)}{p_\tau(X_\tau)}\right] \\
&= \int \pi(v)\left[\int q^v_{\theta,\tau}(x)\log\frac{\bar{q}_{\theta,\tau}(x)}{p_\tau(x)}\,\mathrm{d}x\right]\mathrm{d}v \\
&= \int \left[\int \pi(v) q^v_{\theta,\tau}(x)\,\mathrm{d}v\right]\log\frac{\bar{q}_{\theta,\tau}(x)}{p_\tau(x)}\,\mathrm{d}x \\
&= \int \bar{q}_{\theta,\tau}(x)\log\frac{\bar{q}_{\theta,\tau}(x)}{p_\tau(x)}\,\mathrm{d}x \\
&= D_{\mathrm{KL}}\left(\bar{q}_{\theta,\tau}\|p_\tau\right). \qquad (21)
\end{aligned}$$

Combining the two terms and rearranging proves equation 18.

The first term in equation 18 is the average reverse KL of the conditional rollout distributions. The second term measures rollout-distribution diversity, reflecting how much the trajectory-conditioned rollout distributions deviate from their marginal mixture. It also follows from equation 18 that

$$D_{\mathrm{KL}}\left(\bar{q}_{\theta,\tau}\|p_\tau\right) \le \mathbb{E}_V D_{\mathrm{KL}}\left(q^V_{\theta,\tau}\|p_\tau\right). \qquad (22)$$

When the masking trajectory does not affect the conditional rollout distribution, each conditional distribution is the same as the mixture. The corresponding KL divergences are zero, so the rollout-distribution diversity term vanishes and equality holds. Otherwise, the diversity term is positive, and the inequality is strict.

Implication for mode coverage. Mask Forcing can broaden mode coverage by allowing different rollouts to seek different teacher modes without eliminating mode seeking. Different masking trajectories may induce conditional distributions concentrated on different teacher modes. Marginalizing over these trajectories combines their coverage, allowing the marginal mixture to cover multiple teacher modes even when individual conditional distributions remain mode-seeking. equation 18 identifies the average KL from the conditional rollout distributions to their mixture as the reduction from the average conditional reverse KL to the marginal KL. When the conditional rollout distributions concentrate on distinct teacher modes, this reduction can be interpreted as the benefit of combining different mode-seeking outcomes.

B Comparison with Joint Distillation

Recent methods seek to improve distillation performance by combining complementary objectives that balance mode-seeking and mode-covering behaviors, such as rCM and DistillAlign. We compare with Causal-rCM (Zheng et al., 2026a) and DistillAlign in terms of visual quality and motion dynamics on the 100-prompt set benchmark. While rCM jointly optimizes continuous-time consistency distillation and score-based distillation in bidirectional models, Causal-rCM decouples these objectives into two stages. It first performs consistency distillation under teacher forcing to obtain causal initialization with broad mode coverage, and then applies DMD under self-forcing. Since DistillAlign is built upon Causal Forcing and Causal-rCM pairs its two distillation objectives with two causal training paradigms, we compare Mask Forcing on the Causal Forcing baseline with

Table 6: Quantitative comparison of Mask Forcing (Ours) with Causal-rCM and DistillAlign on the 100-prompt set benchmark.

Method	HPSv3 ↑	Vision. ↑	Instruct. ↑	MQ ↑	Dynamic. ↑
DistillAlign (Li et al., 2026a)	9.29	9.46	37.69	14.36	76
Causal-rCM (Zheng et al., 2026a)	9.61	9.43	38.50	14.34	64
Ours	**10.17**	**11.58**	**46.30**	**21.54**	**82**

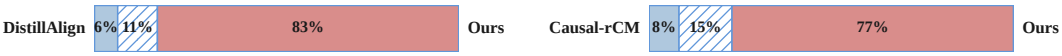

Figure 7: Pairwise human preference between joint distillation methods (DistillAlign and Causal-rCM) and Mask Forcing.

them. For DistillAlign, we use its final joint-distilled generator with the Wan2.1-T2V-14B teacher. For Causal-rCM, we follow its 4-step inference setting.

As shown in Tab. 6, our method outperforms both methods in visual quality, motion dynamics, as well as semantic alignment. Qualitative comparison in Fig. 8 is consistent with the quantitative results, showing that our method generates videos with richer details and greater realism. We further conduct pairwise human preference against joint distillation methods DistillAlign and Causal-rCM. As shown in Fig. 7, Mask Forcing receives more preference votes, with 83% and 77% over DistillAlign and Causal-rCM, respectively. These results demonstrate that Mask Forcing yields videos with higher visual quality and realism from the users' perspective.

C Additional Evaluation on LongLive

We evaluate LongLive and Mask Forcing on MovieGen extended version over consecutive 6-second intervals from 0 to 30 seconds. For each interval, we average the CLIP score over all frames and the HPSv3 score over 12 uniformly sampled frames. As reported in Tab. 8, both methods exhibit a certain decline in CLIP and HPSv3 scores as the video generation progresses, likely due to error accumulation during autoregressive self-rollout. However, incorporating Mask Forcing achieves higher CLIP and HPSv3 scores than LongLive across all intervals, demonstrating consistent improvements in semantic alignment and visual quality throughout long video generation.

D Diversity Evaluation

We further evaluate the video diversity across different baselines under the chunk-wise setting using CLIP-ViT-Large (Radford et al., 2021) and DINOv3-ViT-Large (Siméoni et al., 2025) following the diversity metric of DP-DMD (Wu et al., 2026). We use the 100-prompt set as the evaluation set and generate 8 videos for each prompt with random seeds 0-7, generating 800 videos for each baseline for evaluation in total. To adapt the image-level metric to videos, we uniformly sample 8 frames from each video clip and mean-pool their ℓ_2-normalized per-frame embeddings into a single video embedding, following the video adaptation of DFD (Chen et al., 2026).

As shown in Tab. 7, Mask Forcing improves CLIP diversity across all three baselines, indicating consistently greater semantic diversity. It also substantially improves DINOv3 diversity for Self Forcing, LongLive, and Causal Forcing, suggesting increased structural diversity. Together with the consistent gains in visual quality, semantic alignment, distributional alignment, and human preference, these results suggest that Mask Forcing jointly improves generation quality and diversity.

Table 7: Video diversity across baselines.

Method	CLIP Div.↑	DINO Div.↑
Self Forcing	0.0773	0.1704
+ Ours	**0.0859**	**0.1871**
LongLive	0.0784	0.1537
+ Ours	**0.0819**	**0.1750**
Causal Forcing	0.0675	0.1371
+ Ours	**0.0720**	**0.1577**



Figure 8: Qualitative comparison of our method with Causal-rCM and DistillAlign. Our method produces videos with better visual quality and motion dynamics.

Table 8: Quantitative evaluation on long video generation on MovieGen. We evaluate the CLIP Score and HPSv3 across 6-second intervals (0–30s).

Method	CLIP↑					HPSv3↑				
	0-6	6-12	12-18	18-24	24-30	0-6	6-12	12-18	18-24	24-30
LongLive	33.70	33.41	33.29	33.23	33.15	8.81	8.37	8.23	7.89	7.79
+ Ours	**34.12**	**33.81**	**33.59**	**33.63**	**33.38**	**9.65**	**9.44**	**9.05**	**8.87**	**8.68**

E Interactive Video World Model Application

We further apply Mask Forcing on autoregressive video generation with precise camera control. Current interactive world models extend video generation by conditioning future observations on user inputs (Team et al., 2026a;b; HunyuanWorld, 2025; Wang et al., 2026b; Zhao et al., 2026), such as actions, camera motions, or instructions. We evaluate Mask Forcing in a camera-controlled image-to-video setting, where an input image serves as the initial frame, and the model generates subsequent frames following a specified camera trajectory. We utilize Wan2.2-5B-TI2V as the base model and train the bidirectional model with open-source data, such as SpatialVID (Wang et al., 2026a), OmniWorld (Zhou et al., 2025), RealCam-Vid (Zheng et al., 2025), DL3DV (Ling et al., 2024), Sekai (Li et al., 2026b), and MiraData (Ju et al., 2024). Then we train an AR model using teacher forcing and adopt consistency distillation to obtain a few-step generator. Both bidirectional and autoregressive models are trained on 5s videos. We then compare Self Forcing and Mask Forcing for further causal student distillation to align with the teacher's distribution.

Fig. 9 presents qualitative comparisons in the Sekai-Walking and Sekai-Game test sets. We observe that in scenes with relatively dark conditioning first frames, videos generated by Self Forcing become progressively darker and lose fine-grained details as the rollout proceeds. This degradation may reflect error accumulation during self-rollout, further affected by the mode-seeking behavior of the reverse-KL objective. Mask Forcing mitigates this degradation, producing scenes with richer details throughout the rollout. We view this experiment as an initial evaluation of our method in the camera-controlled image-to-video setting. To enable long and precise interactive video generation, future work includes training with long sequences and rollouts (Yang et al., 2025) and the context update strategy for temporal extrapolation (Zhuang et al., 2026).

F Algorithm

We provide detailed pseudocode for training with Mask Forcing in Algorithm 1. Mask Forcing requires no additional forward passes, external data, or post-training stages and can be seamlessly integrated into existing self-rollout DMD training pipelines. Mask Forcing constructs a dual-noise input at each denoising step by mixing the original- and lower-noise states with a random mask. The perturbed input is then denoised at the original timestep, and the final clean prediction is cached as context for subsequent blocks.

G More Qualitative Results

We present additional qualitative comparisons on single-prompt short video generation on the 100-prompt set and MovieGen in Figs. 10-12. Qualitative comparisons across different methods show that Mask Forcing consistently improves the visual quality of the generated videos, enhancing them with greater realism and richer detail while preventing over-saturated and over-smoothed artifacts. See the supplementary file for more video results.

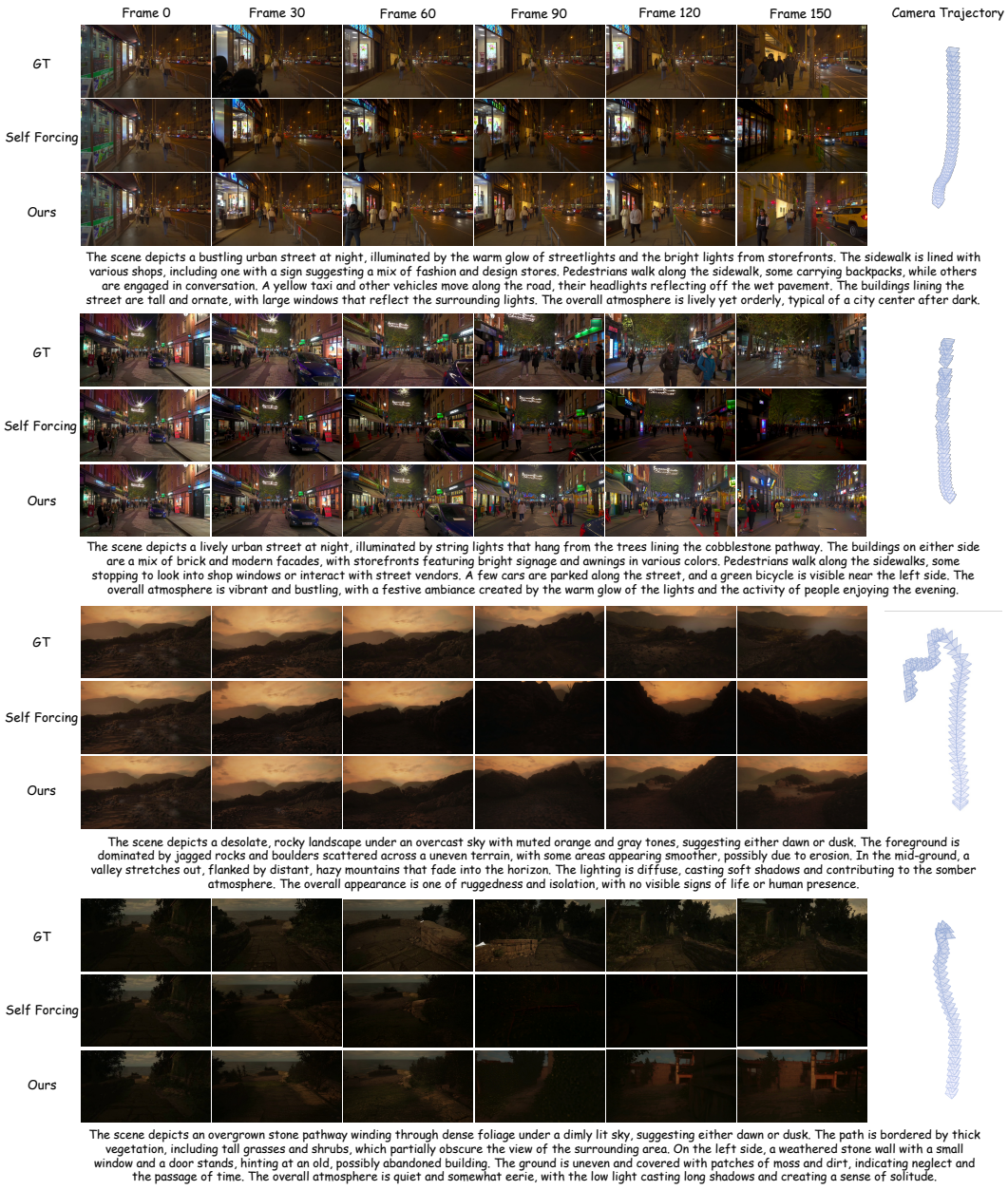

Figure 9: Qualitative comparison of Self Forcing and Mask Forcing (Ours) on Sekai-Walking and Sekai-Game for camera-controlled image-to-video generation. Compared with Self Forcing, Mask Forcing substantially alleviates progressive darkening during rollout and better preserves details.

Algorithm 1 Mask Forcing Training

Require: Denoising timesteps $\mathcal{T} = \{t_1, \ldots, t_T\}$ with $t_T = 1$ and clean endpoint $t_0 = 0$
Require: Text prompt c; number of autoregressive chunks F, latent frames per chunk L, spatial latent resolution $H \times W$
Require: AR diffusion model G_θ (returns KV embeddings via G_θ^{KV})
Require: Mask ratio α, timestep window size Δ in training timestep units, number of training timesteps N_t, normalized timestep floor $t_{\min}$
Notation: $\Psi(x, \varepsilon, t) = (1-t)x + t\varepsilon$

loop
  Initialize KV cache $\mathrm{KV} \leftarrow [\,]$ and model output $X_\theta \leftarrow [\,]$
  Sample exit index $s \sim \mathrm{Uniform}\{1, \ldots, T\}$
for chunk $i = 1, \ldots, F$ **do**
  Sample initial noisy latent $x_i^{t_T} \sim \mathcal{N}(0, I)$
  Sample a per-frame binary mask $M^i \in \{0,1\}^{L \times H \times W}$ with ratio α
  if $s = T$ **then**
    Enable gradient computation
  else
    Disable gradient computation
  end if
  $\hat{x}_{i,T}^0 \leftarrow G_\theta\big(x_i^{t_T};\, c,\, t_T,\, \mathrm{KV}\big)$ ▷ initial unperturbed prediction
  if $s < T$ **then**
    for denoising step $j = T-1, T-2, \ldots, s$ **do**
      $t'_k \sim \mathrm{Uniform}([\max(t_{\min}, t_j - \Delta/N_t),\, t_j])$
      Sample $\varepsilon_j \sim \mathcal{N}(0, I)$
      $x_i^{t'_k} \leftarrow \Psi(\hat{x}_{i,j+1}^0, \varepsilon_j, t'_k), \quad x_i^{t_j} \leftarrow \Psi(\hat{x}_{i,j+1}^0, \varepsilon_j, t_j)$
      $x_i^{t_{\mathrm{mix}}} \leftarrow M^i \odot x_i^{t'_k} + (1 - M^i) \odot x_i^{t_j}$
      if $j = s$ **then**
        Enable gradient computation
      else
        Disable gradient computation
      end if
      $\hat{x}_{i,j}^0 \leftarrow G_\theta\big(x_i^{t_{\mathrm{mix}}};\, c,\, t_j,\, \mathrm{KV}\big)$ ▷ perturbed prediction at step j
    end for
  end if
  Append $\hat{x}_{i,s}^0$ to X_θ
  Disable gradient computation
  $\mathrm{kv}^i \leftarrow G_\theta^{\mathrm{KV}}\big(\hat{x}_{i,s}^0;\, c,\, t_0,\, \mathrm{KV}\big)$ ▷ cache the exit prediction
  $\mathrm{KV}.\mathrm{append}(\mathrm{kv}^i)$
end for
  Update θ via the distribution matching loss $\mathcal{L}_{\mathrm{DMD}}(X_\theta)$
end loop

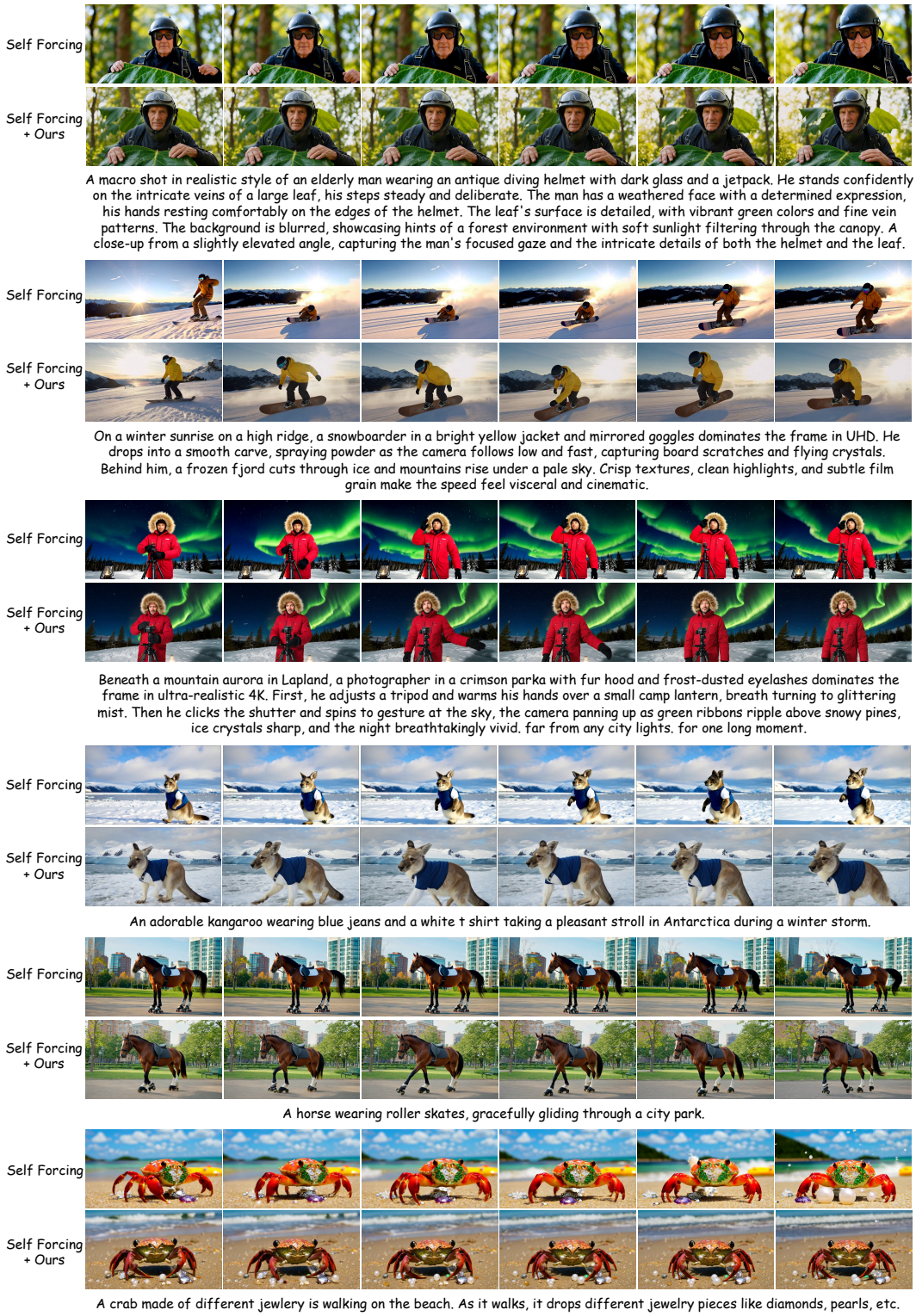

Figure 10: Additional qualitative comparisons between Self Forcing and Self Forcing (+Ours).



Figure 11: Additional qualitative comparisons between LongLive and LongLive (+Ours).

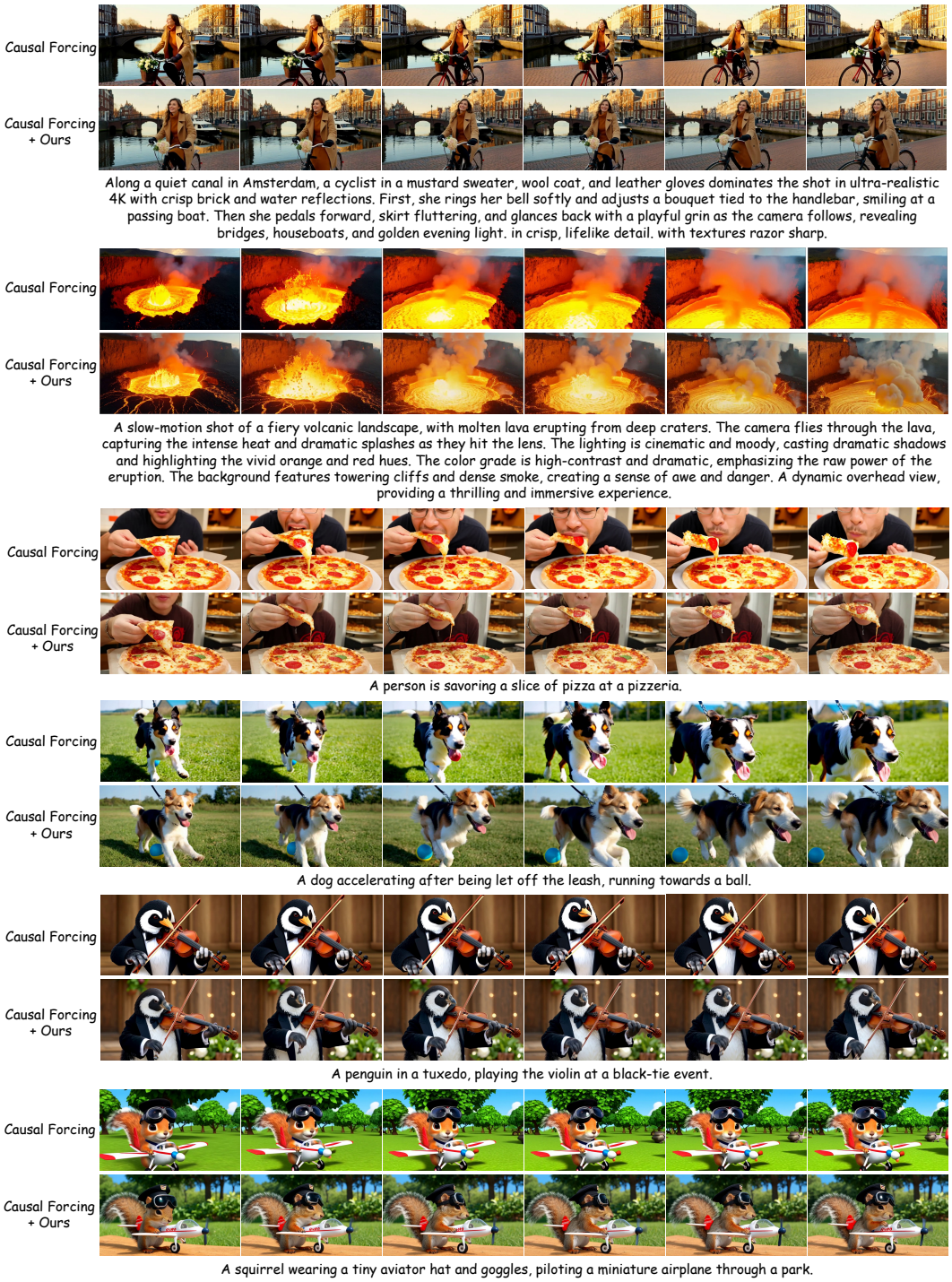

Figure 12: Additional qualitative comparisons between Causal Forcing and Causal Forcing (+Ours).