FORGETTING TO KEEP GOING: HISTORY DROPOUT FOR LONG-HORIZON VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video generators use previously generated content to maintain continuity, but this history can also carry errors into later predictions. We introduce **Hi**story **Drop**out (HiDrop), a history regularization method for long-horizon video generation. During training, HiDrop varies historical value masks across denoising steps while retaining all keys for attention normalization. Inverse-retention scaling preserves the expected attention output for fixed, mask-independent inputs. At inference, each attention head independently samples historical key-value tokens, with sink and current-chunk tokens always retained. The two operations serve different purposes: training perturbs historical contributions without changing the softmax denominator, whereas inference uses shorter attention inputs. Applied to Self Forcing, Causal Forcing, and LongLive, HiDrop improves visual quality and dynamics at both 60 and 240 seconds, while inference latency decreases from 0.72 to 0.64 seconds on a single NVIDIA H20 GPU. Please refer to this link for our supplementary video.

1 INTRODUCTION

Autoregressive video diffusion generates a long video one chunk at a time. Few-step causal models make this process efficient (Yin et al., 2025), while self-rollout training and streaming long-video tuning extend its practical reach (Huang et al., 2025; Yang et al., 2025a). A key-value (KV) cache carries subject appearance, scene structure, and motion into each continuation, requiring the model to use its own generated history repeatedly as generation proceeds.

The same history can also carry generation errors into later predictions. Cached representations are computed from model outputs, not clean observations; repeatedly attending to them can preserve unwanted features along with useful context. Self Forcing addresses the train–test mismatch by training on self-generated histories (Huang et al., 2025). Causal Forcing improves autoregressive initialization (Zhu et al., 2026), and LongLive extends training to streaming long-video generation (Yang et al., 2025a). Yet subject distortion and texture degradation remain visible in minute-scale rollouts (Fig. 1). These failures motivate a complementary question: *how should a generator use an imperfect history, even after it has been trained on its own outputs?*

We introduce HiDrop, which regularizes historical conditioning without replacing the rollout objective or altering the stored cache. Training masks vary the available historical values across denoising steps, so a token excluded at one step can contribute at another. The masks do not identify errors; they vary history contributions rather than exclude a fixed subset throughout denoising.

The training rule separates value availability from attention normalization. Removing historical keys changes the softmax denominator and reweights surviving tokens, including the current chunk. HiDrop keeps all keys, zeros masked historical values, and scales surviving values by the inverse retention probability. For fixed, mask-independent inputs, this preserves the expected attention output. This local identity guarantees neither rollout-level unbiasedness nor error correction.

At inference, each head independently samples historical KV tokens at every attention call, retaining the sink and current chunk. Attention normalizes over this smaller set without value scaling. This shortens attention inputs without shrinking the stored cache or preserving dense attention in expectation. Across Self Forcing, Causal Forcing, and LongLive, our generated videos show that visual quality and dynamics improve at 60s and 240s.

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.

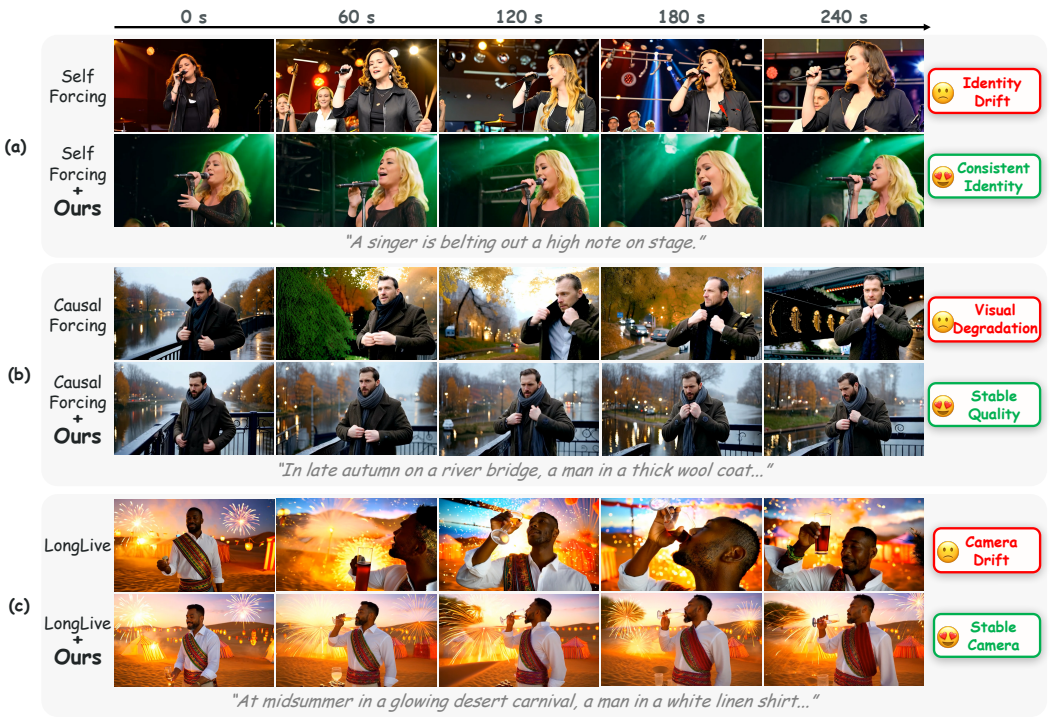

Figure 1: Long-horizon video generation with and without HiDrop.

Our contributions span history regularization, analysis, and long-video evaluation:

- **History regularization.** We combine step-dependent historical value dropout during training with independent per-head KV selection at inference, adding no parameters or losses.
- **Training and inference effects.** We separate their empirical contributions on Self Forcing and characterize the training operator's local expectation property and its scope.
- **Long-video evaluation.** On three autoregressive baselines, HiDrop improves visual quality and dynamics at 60s and 240s. Steady-state chunk latency falls from 0.72s to 0.64s on an H20, an 11.1% reduction at the reported precision.

2 Related Work

Long-video generation and training. Video diffusion architectures (Ho et al., 2022; Blattmann et al., 2023; Yang et al., 2025b) extend to longer sequences through noise rescheduling and temporal attention (Qiu et al., 2024; Lu et al., 2024), memory (Henschel et al., 2025), sliding denoising windows (Ruhe et al., 2024; Kim et al., 2024), or frame-dependent noise schedules (Sun et al., 2025; Chen et al., 2025; Sand.ai et al., 2025). Building on progressive and distribution-matching distillation (Salimans & Ho, 2022; Yin et al., 2024b;a), CausVid enables few-step causal generation (Yin et al., 2025). Following earlier approaches to train-test mismatch (Bengio et al., 2015; Lamb et al., 2016), Self Forcing trains on self-rollouts (Huang et al., 2025), Causal Forcing revises teacher-based ODE initialization (Zhu et al., 2026), and LongLive uses streaming long tuning (Yang et al., 2025a). Context Forcing uses a long-context teacher (Chen et al., 2026b), while Self Gradient Forcing propagates future supervision through recomputed history (Zhuang et al., 2026). HiDrop complements these strategies by regularizing historical attention without changing the objective.

History conditioning and regularization. Context perturbation is established in video diffusion: Diffusion Forcing varies token noise levels (Chen et al., 2024), History-Guided Video Diffusion studies history conditioning and guidance (Song et al., 2025), In-Context Forcing progressively noises contexts (Yang et al., 2026), and Stable Video Infinity recycles generated errors (Li et al., 2026). At the feature level, dropout regularizes neural networks and Transformers (Srivastava et al.,

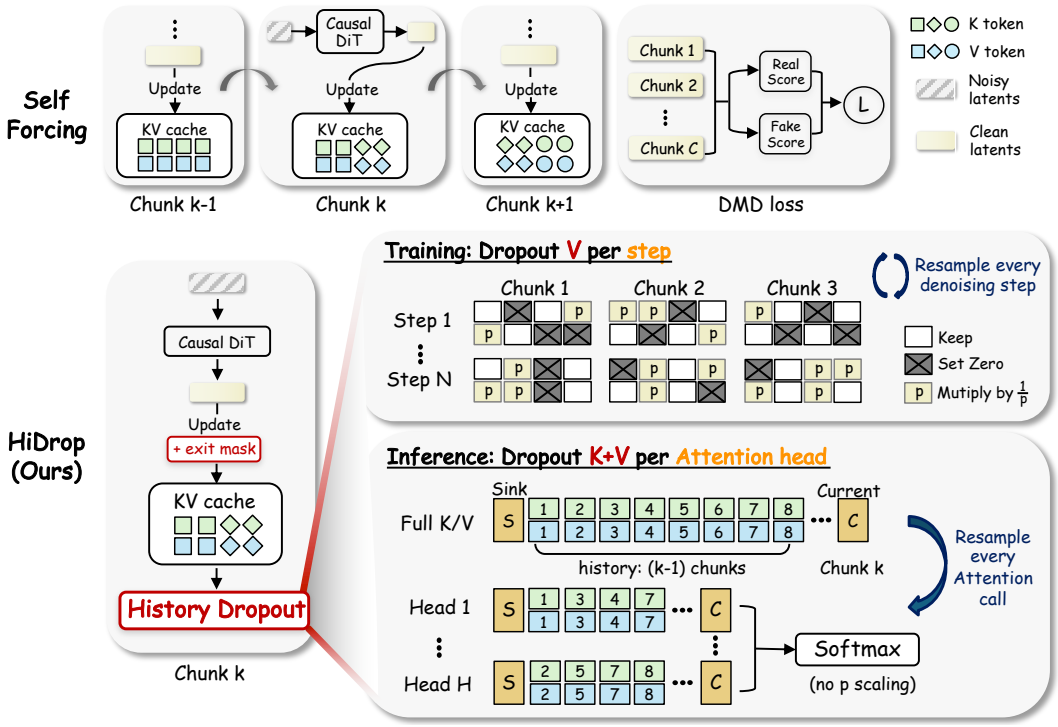

Figure 2: **Overview of HiDrop.** During training, masks vary across denoising steps: all keys are retained, masked historical values are zeroed, and surviving values are scaled by $1/p$, where p is the retention probability. This preserves the full-history attention contribution in expectation for fixed, mask-independent inputs. Cache refresh reuses the exit-step mask. During inference, each head independently samples historical KV tokens at every attention call, retaining sink (S) and current-chunk (C) tokens. Attention is normalized over the selected tokens without value rescaling.

2014; Vaswani et al., 2017), with shared temporal masks in recurrent models (Gal & Ghahramani, 2016). DropAttention masks and renormalizes attention weights (Lin et al., 2019); DropKey removes keys before softmax (Li et al., 2023). Structured masking also underlies DropBlock (Ghiasi et al., 2018) and reconstruction-based MAE/VideoMAE (He et al., 2022; Tong et al., 2022). HiDrop uses the standard inverse-retention scaling of dropout, but applies it only to historical values under full-key normalization. Its design concerns where masks act, how they vary over denoising steps, and how training is paired with per-head KV selection at inference. It does not re-noise context latents or require an auxiliary reconstruction objective.

History compression and KV selection. Pyramid Flow and FramePack compress video history by temporal scale or frame importance (Jin et al., 2025; Zhang et al., 2025). Language-model caches retain attention sinks, recent tokens, or heavy hitters (Xiao et al., 2024; Zhang et al., 2023), while SnapKV selects prompt positions per head (Li et al., 2024). Pyramid Forcing assigns video-cache budgets by head category (Chen et al., 2026a). HiDrop randomly selects historical KV reads per head without deleting the stored cache; this inference rule is distinct from its training-time estimator.

3 Forgetting to Keep Going

3.1 Preliminaries

Let $x_{1:B}$ denote a video represented by B latent chunks and c its text condition. At denoising-step index $s \in \{1, \ldots, S\}$, the student predicts chunk b from its noisy latents and a KV cache of preceding chunks. Let H_b denote the historical token indices under the cache policy, and let C_b contain all remaining self-attention token indices outside H_b. Thus, C_b includes the current chunk

and, at inference, any protected sink tokens. For query q_i, key k_j, and value v_j, attention is

$$\ell_{ij} = \frac{q_i^\top k_j}{\sqrt{d}}, \qquad a_i = \frac{\sum_{j \in H_b \cup C_b} e^{\ell_{ij}} v_j}{\sum_{j \in H_b \cup C_b} e^{\ell_{ij}}}. \tag{1}$$

Here, d is the query/key dimension of each attention head. The indices for layers, heads, and denoising steps are omitted from the attention notation. Tokens in C_b are always retained; the separate text cross-attention pathway is unchanged by history dropout.

3.2 Step-Dependent History Masks

We vary historical value masks across denoising steps so that the generator trains with different partial contributions from the same cache. Holding a mask fixed would exclude one subset throughout denoising and leave another continuously available. Step-dependent masks instead allow excluded features to return at later steps, without guaranteeing continuous access to any token.

We sample a history mask for each video-token position j and denoising-step index s:

$$m_j^{(s)} \sim \text{Bernoulli}(p), \qquad p = 1 - r, \qquad 0 \le r < 1. \tag{2}$$

Here, r is the training dropout rate and p is the corresponding retention probability for each eligible historical token. Masks are sampled independently across token positions and step indices. At a given step s, entries for the same historical KV chunk are shared across layers and heads and reused when generating later chunks. When encoding a generated chunk into the KV cache, the context-refresh pass reuses that chunk's exit-step history mask, from its last executed denoising step.

For a token eligible at each of S denoising steps, both a fixed mask and independent step masks retain it for Sp steps in expectation. The probability of retaining it at every step is p under a fixed mask and p^S under step-dependent masks. Thus the schedule changes persistent access, not the expected retention budget. This does not guarantee removal of errors already in the current latent.

3.3 Historical Value Dropout with Full-Key Normalization

Training-time dropout perturbs historical contributions while keeping the full-key attention weights. Using Equation (1), define the weights and history contribution as

$$\alpha_{ij} = \frac{e^{\ell_{ij}}}{Z_i}, \qquad Z_i = \sum_{k \in H_b \cup C_b} e^{\ell_{ik}}, \qquad h_i = \sum_{j \in H_b} \alpha_{ij} v_j. \tag{3}$$

Direct KV removal changes both the available values and their attention weights. Removing a historical token excludes its value from the output and its key from the softmax denominator, so attention is renormalized over the retained history and the tokens in C_b. With $m_j = m_j^{(s)}$, we obtain

$$a_i^{\text{drop}} = \frac{\sum_{j \in H_b} m_j e^{\ell_{ij}} v_j + \sum_{j \in C_b} e^{\ell_{ij}} v_j}{\sum_{j \in H_b} m_j e^{\ell_{ij}} + \sum_{j \in C_b} e^{\ell_{ij}}}. \tag{4}$$

Let $Z_i(M)$ denote the denominator above. Whenever a historical key is removed, $Z_i(M) < Z_i$, increasing each surviving token's weight from α_{ij} to $\alpha_{ij} Z_i / Z_i(M)$. Even tokens in C_b are reweighted although they are never dropped. This reweighting does not, in general, compensate for the missing historical contributions. Thus, direct KV removal generally produces biased estimates of the full-history contribution h_i and the full-attention output a_i, even when Q, K, V are fixed and randomness is solely due to sampling the history mask (Appendix A.1).

We therefore use *historical value dropout*: compute the softmax with all keys, then apply dropout only to historical values in the weighted sum. This preserves the denominator Z_i and softmax weights α_{ij}. We set dropped historical values to zero and scale retained values by $1/p$ to compensate for their retention probability p. The resulting historical contribution and full attention output are

$$\widehat{h}_i = \sum_{j \in H_b} \frac{m_j}{p} \alpha_{ij} v_j, \tag{5}$$

$$\widehat{a}_i = \widehat{h}_i + \sum_{j \in C_b} \alpha_{ij} v_j, \tag{6}$$

Table 1: Main quantitative comparison at 60s and 240s for Self Forcing, Causal Forcing, and LongLive, with and without HiDrop. Lat. is steady-state latency per chunk on a single NVIDIA H20 GPU with an established history cache, not first-frame latency (Appendix B.3).

Horizon	Method	AQ↑	BC↑	DD↑	IQ↑	MS↑	SC↑	TF↑	Lat. (s)↓
60s	Self Forcing	0.582	**0.953**	0.693	0.703	**0.979**	0.965	0.981	0.72
	Self Forcing + Ours	**0.595**	**0.953**	**0.785**	**0.708**	**0.979**	**0.966**	**0.985**	**0.64**
	Causal Forcing	0.562	0.958	0.451	0.702	**0.986**	0.964	**0.982**	0.72
	Causal Forcing + Ours	**0.604**	**0.960**	**0.565**	**0.710**	**0.986**	**0.972**	**0.982**	**0.64**
	LongLive	0.573	0.951	0.709	0.694	0.975	0.960	**0.982**	0.72
	LongLive + Ours	**0.612**	**0.966**	**0.765**	**0.706**	**0.980**	**0.977**	**0.982**	**0.64**
240s	Self Forcing	0.548	0.956	0.713	0.694	**0.979**	0.959	0.971	0.72
	Self Forcing + Ours	**0.560**	**0.959**	**0.790**	**0.714**	**0.979**	**0.965**	**0.979**	**0.64**
	Causal Forcing	0.503	0.956	0.536	0.688	**0.983**	0.957	0.976	0.72
	Causal Forcing + Ours	**0.565**	**0.962**	**0.645**	**0.704**	**0.983**	**0.969**	**0.981**	**0.64**
	LongLive	0.540	0.955	0.783	0.682	0.972	0.954	0.976	0.72
	LongLive + Ours	**0.574**	**0.965**	**0.873**	**0.697**	**0.979**	**0.972**	**0.982**	**0.64**

where the contribution from C_b is neither masked nor rescaled. Equivalently, historical values receive effective weights $\widehat{\alpha}_{ij} = (m_j/p)\alpha_{ij}$, with no renormalization after masking.

The $1/p$ scaling makes this estimator unbiased for fixed inputs Q, K, V that are independent of the mask M. Under this assumption, α_{ij} and v_j remain constant as the mask varies, while $\mathbb{E}_M[m_j/p \mid Q, K, V] = 1$. Taking the expectation term by term recovers the historical contribution; adding the unchanged C_b contribution recovers the full attention output:

$$\mathbb{E}_M[\widehat{h}_i \mid Q, K, V] = h_i, \qquad \mathbb{E}_M[\widehat{a}_i \mid Q, K, V] = a_i. \tag{7}$$

Equation (7) is a local operator identity. In a rollout, reused masks can affect the hidden states and cached representations that produce later Q, K, V. Those inputs need not be independent of the mask, so the identity does not establish unbiased network outputs, gradients, or future cache entries. It also preserves the contribution of the available history, including any errors in that history, rather than recovering clean context. Masking changes only the temporary values used by attention; stored KV entries, text conditioning, and positional treatment are unchanged.

3.4 Self-Rollout Training

We integrate this perturbation into Self Forcing training (Huang et al., 2025) without changing its distribution-matching objective. Let $G_\theta(z, c; M)$ denote the generated sequence under input noise z, condition c, and step-indexed masks $M = \{M^{(s)}\}_{s=1}^{S}$. The generator is trained with the existing loss, evaluated on sequences from these masked self-rollouts,

$$\mathcal{L}(\theta) = \mathbb{E}_{c,z,M}\left[\mathcal{L}_{\mathrm{DMD}}\big(G_\theta(z, c; M); c\big)\right]. \tag{8}$$

The fake-score model is trained on samples from the same masked rollout process. No auxiliary loss terms or additional learned parameters are introduced by this procedure.

3.5 Sink-Preserving Per-Head Inference

At inference, we reduce historical KV reads by sampling a subset $R_{b,h}$ of H_b for each attention head h. Within each non-sink historical frame containing N tokens, we select $k = \max(1, \lceil \rho N \rceil)$ tokens uniformly, where $\rho \in (0, 1]$ is the inference keep ratio. The retained set for that head is

$$I_{b,h} = R_{b,h} \cup C_b. \tag{9}$$

The same selected indices serve all current queries within a head, but different heads draw independently. We resample at every attention call, including across layers, denoising steps, chunks, and context-refresh forwards, rather than reusing the training masks.

We gather both K and V at the retained indices and apply ordinary attention:

$$a_{i,h}^{\text{inf}} = \frac{\sum_{j \in I_{b,h}} \exp(\ell_{ij,h}) v_{j,h}}{\sum_{j \in I_{b,h}} \exp(\ell_{ij,h})}. \tag{10}$$

Unlike training, this operation normalizes over selected keys and does not rescale values. It therefore has no dense-attention unbiasedness guarantee. The distinction is deliberate: training perturbs value contributions under the original normalization, whereas inference reduces the number of historical KV reads. The full cache remains stored, so other heads or calls can select a skipped token.

Empirically, on dropout-trained Self Forcing, this rule improves aesthetic quality, imaging quality, and dynamic degree over full-history inference at both horizons (Table 2).

4 Experiments

We evaluate 60-second and 240-second generation using Self Forcing, Causal Forcing, and LongLive, with paired comparisons within each baseline family using VBench (Huang et al., 2023). Following Self Gradient Forcing (Zhuang et al., 2026), we use VBench-Long prompts at 60s and MovieGen prompts at 240s. Comparisons are made within each horizon; the two prompt sets do not form a controlled measure of degradation from 60s to 240s. Scores are point estimates without confidence intervals. Settings, metrics, and implementation details are in Appendix B.

4.1 Long-Horizon Generation Quality

Across all three baseline families and both horizons, HiDrop improves aesthetic quality (AQ), imaging quality (IQ), and dynamic degree (DD) (Table 1). Background consistency (BC), subject consistency (SC), motion smoothness (MS), and temporal flickering (TF) are maintained or improved at the reported precision. The quality gains are therefore accompanied by more motion, not achieved by making videos nearly static. At 240s, Causal Forcing's AQ rises from 0.503 to 0.565 and its DD from 0.536 to 0.645; LongLive's DD rises from 0.783 to 0.873 alongside higher AQ and IQ. Self Forcing exhibits the same joint trend: at 60s, DD increases from 0.693 to 0.785 while IQ rises from 0.703 to 0.708; at 240s, the corresponding changes are 0.713 to 0.790 and 0.694 to 0.714. Reading these dimensions together matters because consistency alone does not distinguish a coherent moving scene from one with little motion.

Appendix C.4 reports complementary 5-second results. Fig. 3 compares scene composition and subject appearance at sampled points over four minutes. These frames illustrate visual differences; motion is assessed through the video metrics and human study. Additional examples appear in Appendix D.

The quality comparison is accompanied by lower steady-state chunk latency: 0.64s versus 0.72s on an H20. Inference selection reduces the history read by attention while leaving the stored cache intact, so this timing result should not be interpreted as a reduction in cache capacity or a measurement of end-to-end video generation time.

4.2 Training and Inference Contributions

Table 2 compares a cumulative sequence of changes: "+ Naive Dropout" directly removes a fixed subset of historical KV tokens; "+ Per-step" varies the subset across denoising steps; "+ Unbiased" uses the value-dropout estimator from Section 3.3; and "+ Infer Dropout" adds per-call, per-head KV selection at inference. All variants use Self Forcing with matched initialization and training budget. The baseline and the first three variants use full-attention inference. Each row changes the preceding configuration; the table does not test every combination of masking and normalization. Fig. 4 provides qualitative examples of these cumulative changes.

Fixed direct KV removal. Fixed KV removal lowers AQ, IQ, and DD at both horizons; at 240s, IQ falls from 0.694 to 0.610 and DD from 0.713 to 0.505. This comparison changes both token availability and normalization, so it does not establish that fixed-mask value dropout is ineffective.

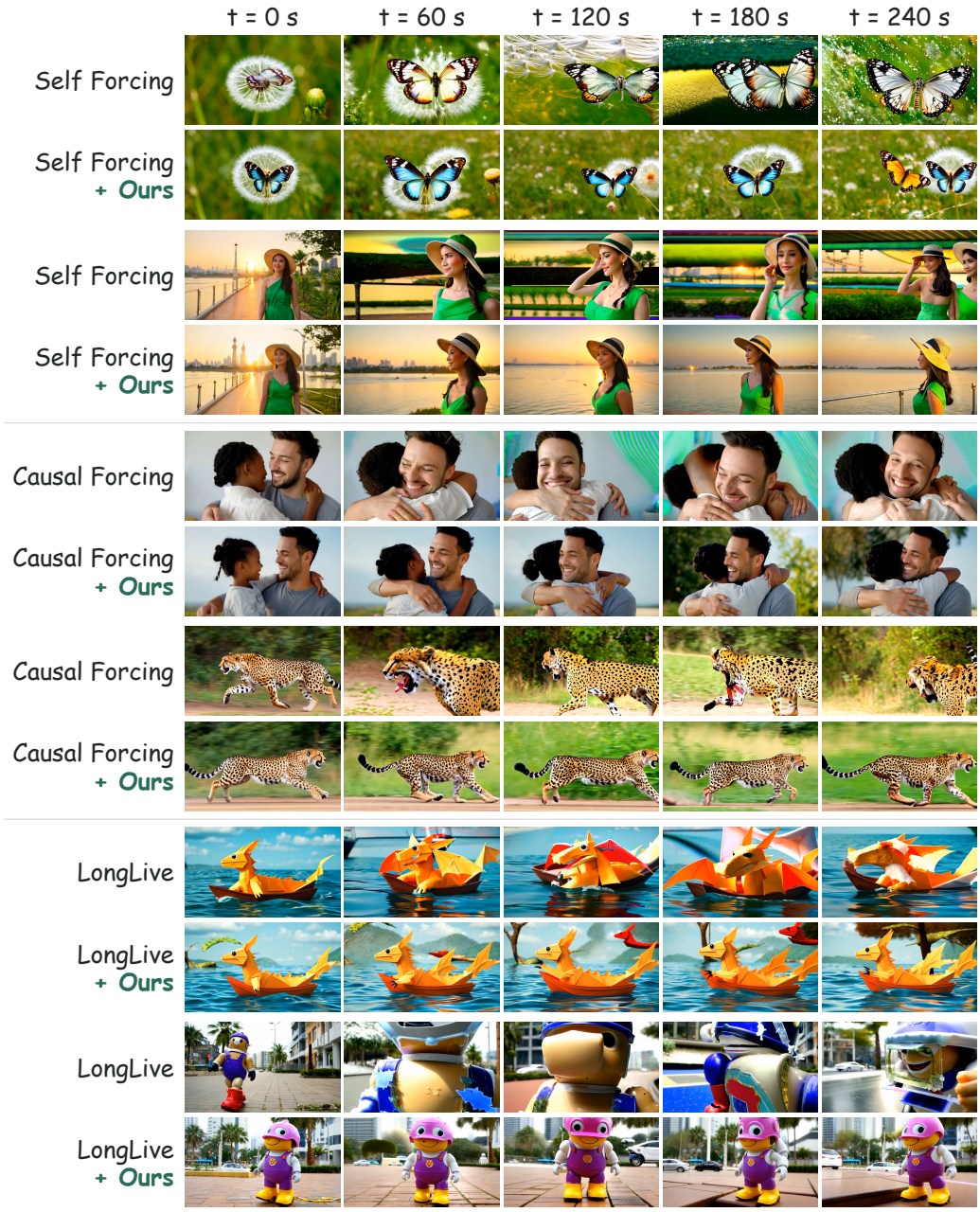

Figure 3: Main qualitative comparison over four minutes. *See the supplementary videos.*

Step-dependent selection. Varying the KV-removal mask improves DD at both horizons and recovers some 240s IQ, but visual quality remains below the baseline. This tests mask timing under KV removal, not under full-key normalization with value dropout.

Full-key normalization and value scaling. Replacing per-step KV removal with historical value dropout raises AQ, IQ, and DD at both horizons. At 240s, IQ recovers to 0.695 and DD reaches 0.776, compared with 0.694 and 0.713 for the baseline. Subject consistency remains lower, at 0.954 versus 0.959. This jointly tests full-key normalization and scaling, not their separate effects.

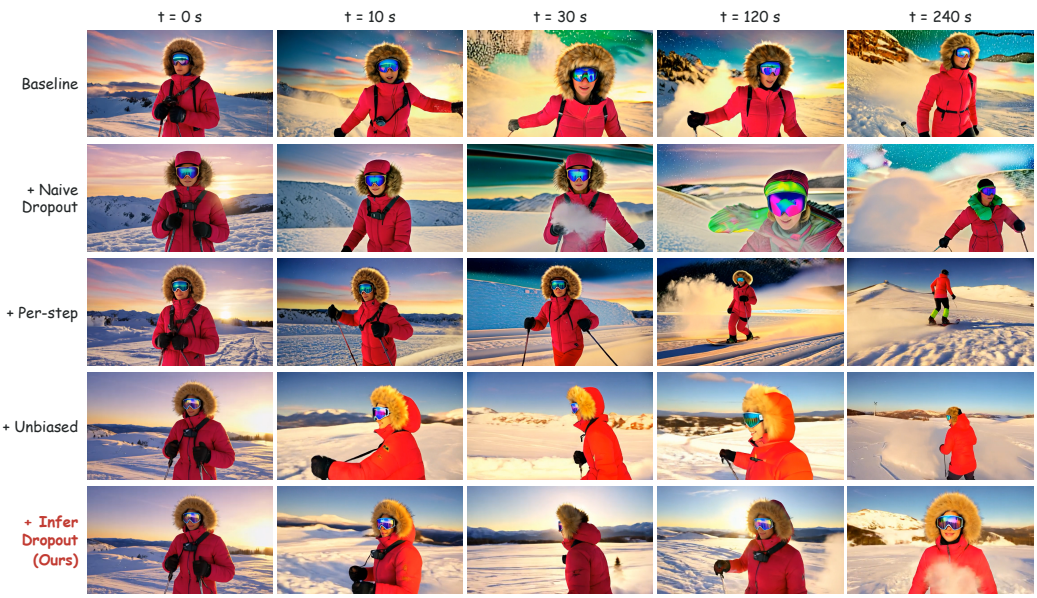

Figure 4: Qualitative examples from the cumulative component ablation. Rows successively add Naive Dropout, Per-step, Unbiased, and Infer Dropout (Ours).

Table 2: Cumulative component ablation. The first four rows at each horizon use the same full-attention inference settings; only "+ Infer Dropout (Ours)" adds per-call, per-head KV selection.

Horizon	Method	AQ↑	BC↑	DD↑	IQ↑	MS↑	SC↑	TF↑
60s	Baseline	0.582	0.953	0.693	0.703	0.979	0.965	0.981
	+ Naive Dropout	0.547	0.954	0.441	0.675	0.975	0.962	0.981
	+ Per-step	0.557	**0.958**	0.649	0.670	**0.983**	**0.968**	0.979
	+ Unbiased	0.586	0.954	0.757	0.699	0.976	0.957	**0.986**
	+ Infer Dropout (Ours)	**0.595**	0.953	**0.785**	**0.708**	0.979	0.966	0.985
240s	Baseline	0.548	0.956	0.713	0.694	0.979	0.959	0.971
	+ Naive Dropout	0.487	0.953	0.505	0.610	0.978	0.958	0.971
	+ Per-step	0.507	0.958	0.689	0.648	0.979	0.960	0.977
	+ Unbiased	0.555	0.952	0.776	0.695	**0.980**	0.954	**0.980**
	+ Infer Dropout (Ours)	**0.560**	**0.959**	**0.790**	**0.714**	0.979	**0.965**	0.979

Inference-time selection. Adding inference-time selection gives the highest AQ, IQ, and DD among these variants at both horizons and restores SC above the baseline. At 240s, IQ increases from 0.695 to 0.714 and SC from 0.954 to 0.965, while DD improves from 0.776 to 0.790. The improvement is not universal across metrics: MS and TF each decrease by 0.001 relative to training with full-attention inference. A shared-head control at the same token budget is needed to isolate the benefit of independent head sampling.

Retraining versus inference alone. Table 4 places all four training/inference configurations together. At 240s, inference alone raises IQ from 0.694 to 0.712; adding dropout training reaches 0.714, while DD rises from 0.734 to 0.790. The additional training benefit is more pronounced in dynamics. These controls cover Self Forcing, not all three baselines.

4.3 Human Preference Study

We randomly sampled 10 videos from the 240s test dataset for the human study. For each test case, we paired a baseline output with its HiDrop counterpart, giving 10 pairs per baseline and 30 pairs across the three baseline families. Twenty evaluators chose the video they preferred based on visual

quality and dynamics. The mean preference for HiDrop is 0.95 against Self Forcing, 0.96 against Causal Forcing, and 0.92 against LongLive (Table 3). These are descriptive preferences for 10 video pairs per baseline, not 20 independent sets of videos.

Table 3: Mean human preference scores on 240s video pairs.

Baseline family	Baseline ↑	Baseline + HiDrop ↑
Self Forcing	0.05	**0.95**
Causal Forcing	0.04	**0.96**
LongLive	0.08	**0.92**

4.4 Additional Experiments

Additional experiments examine the training/inference combination (Appendix C.1), jointly varied training and inference dropout rates (Appendix C.2), and history masking granularity (Appendix C.3). We also provide 5-second generation results (Appendix C.4) and additional four-minute qualitative comparisons (Appendix D). Appendix D.1 compares uniform inference sampling with top-k token selection in a separate qualitative comparison. These experiments also expose trade-offs hidden by a single score. At 240s, increasing the joint dropout rate from 0.5 to 0.7 raises IQ from 0.714 to 0.734, but lowers DD from 0.790 to 0.362. Likewise, coarser masking can score well on selected consistency metrics without matching token-level dropout on dynamics. The choice of rate and masking unit therefore needs to account for motion as well as frame quality.

5 Limitations

On short, 5-second videos, HiDrop performs comparably overall to the underlying baselines (Appendix C.4). Across Self Forcing, Causal Forcing, and LongLive, aesthetic and imaging quality remain close to baseline levels, with small changes in either direction rather than consistent improvements. Thus, the quality gains observed at 60s and 240s do not translate directly to short clips. These results position HiDrop primarily as a method for improving long-horizon generation, rather than a general quality enhancement across all video lengths.

6 Conclusion

We presented HiDrop, a history regularization method for long-horizon autoregressive video generation. During training, step-dependent masks vary access to historical values across denoising steps. Full-key normalization keeps the attention denominator unchanged, and inverse-retention scaling preserves the expected attention output for fixed, mask-independent inputs. At inference, each head independently selects historical keys and values while always retaining sink and current-chunk tokens. Neither operation adds parameters or losses.

Experiments on Self Forcing, Causal Forcing, and LongLive show improvements in aesthetic quality, imaging quality, and dynamic degree at both 60s and 240s. Qualitative comparisons illustrate improved subject appearance and texture preservation in minute-scale continuations, and evaluators prefer HiDrop in the sampled four-minute video pairs. Shorter attention inputs also reduce steady-state chunk latency from 0.72s to 0.64s on a single NVIDIA H20 GPU, an 11.1% reduction at the reported precision.

Self Forcing controls show that inference selection supplies much of the imaging-quality improvement at 240s, while adding dropout training gives a more pronounced gain in dynamics. Combining the two yields higher aesthetic quality, imaging quality, and dynamic degree than either operation alone at both evaluated long horizons. The cumulative ablation also highlights the importance of mask timing and attention normalization over fixed KV removal.

Selective history access thus complements learning from self-generated rollouts. Separating what is stored in the cache from what contributes to each attention call offers a simple way to improve long-video quality and dynamics while reducing inference cost.

AI USE STATEMENT

In this work, we used generative AI tools for drafting, code inspection, and literature lookup. We have not used generative AI tools for generating the reported experimental measurements or model-output frames in the qualitative comparisons. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, June 2023.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Jiayu Chen, Junbei Tang, Wenbiao Zhao, Maoliang Li, Jiayi Luo, Zihao Zheng, Jiawei Yang, Guojie Luo, and Xiang Chen. Pyramid Forcing: Head-aware pyramid KV cache policy for high-quality long video generation. *arXiv preprint arXiv:2605.13111*, 2026a.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context Forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026b.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. DropBlock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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 (CVPR)*, pp. 16000–16009, June 2022.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2568–2577, June 2025.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations (ICLR)*, 2025.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. FIFO-Diffusion: Generating infinite videos from text without training. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor Forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Bonan Li, Yinhan Hu, Xuecheng Nie, Congying Han, Xiangjian Jiang, Tiande Guo, and Luoqi Liu. DropKey for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22700–22709, June 2023.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable Video Infinity: Infinite-length video generation with error recycling. In *International Conference on Learning Representations (ICLR)*, 2026.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zehui Lin, Pengfei Liu, Luyao Huang, Junkun Chen, Xipeng Qiu, and Xuanjing Huang. DropAttention: A regularization method for fully-connected self-attention networks. *arXiv preprint arXiv:1907.11065*, 2019.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. FreeLong: Training-free long video generation with SpectralBlend temporal attention. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. FreeNoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations (ICLR)*, 2024.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42818–42835, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, Siyu Zhou, Qian He, and Jing Liu. AR-Diffusion: Asynchronous video generation with auto-regressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7364–7373, June 2025.

Team Wan et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Lingxiao Yang, Liu Liu, Moran Li, Han Feng, Wenjian Cao, Jiangning Zhang, and Ye Shi. In-Context Forcing: Uncovering context effects in autoregressive video diffusion. *arXiv preprint arXiv:2608.05237*, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025b.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo 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 (CVPR)*, pp. 6613–6623, June 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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. *arXiv preprint arXiv:2607.20368*, 2026.

A Analysis of History Dropout

A.1 Bias from Direct KV Removal

For fixed Q, K, V, let $Z_{i,H} = \sum_{j \in H_b} e^{\ell_{ij}}$, $Z_{i,C} = \sum_{j \in C_b} e^{\ell_{ij}} > 0$, and $Z_{i,H}(M) = \sum_{j \in H_b} m_j e^{\ell_{ij}}$. With $Z_i(M) = Z_{i,H}(M) + Z_{i,C}$, direct removal increases each surviving token's weight by $Z_i / Z_i(M) \geq 1$, yet reduces the total historical attention mass:

$$\sum_{j \in H_b} \frac{m_j e^{\ell_{ij}}}{Z_i(M)} = \frac{Z_{i,H}(M)}{Z_{i,H}(M) + Z_{i,C}} \leq \frac{Z_{i,H}}{Z_{i,H} + Z_{i,C}} = \sum_{j \in H_b} \alpha_{ij}. \tag{11}$$

The remaining mass shifts to C_b. This inequality does not order the outputs, since attention values can have arbitrary magnitudes and directions, including opposite signs.

For example, take one historical and one current token with equal logits, scalar values $(1, 0)$, and $p = 1/2$. Dense attention gives $a_i = 1/2$. Direct removal returns $1/2$ or zero with equal probability, so $\mathbb{E}_M[a_i^{\text{drop}}] = 1/4$. Historical value dropout instead returns 1 or zero, recovering a_i in expectation.

B Experimental and Implementation Details

B.1 Experimental Settings

We compare each of Self Forcing, Causal Forcing, and LongLive with its HiDrop counterpart at 5s, 60s, and 240s. All ablations use Self Forcing; the component study matches initialization and training budget. Full-method inference uses $\rho = 0.5$ unless varied in Appendix C.2.

Evaluation prompts. Following Self Gradient Forcing (Zhuang et al., 2026), we use VBench prompts for 5-second evaluation, VBench-Long prompts for 60-second evaluation, and MovieGen prompts for 240-second evaluation. Within each horizon, each baseline and its HiDrop counterpart are evaluated using the same prompts for a paired comparison.

Metrics. We report seven VBench dimensions (Huang et al., 2023): aesthetic quality (AQ), background consistency (BC), dynamic degree (DD), imaging quality (IQ), motion smoothness (MS), subject consistency (SC), and temporal flickering (TF). Higher is better for all scores; higher TF means less flickering. We assess motion as well as stability, which can favor static videos.

B.2 Training and Inference Algorithms

Algorithms 1 and 2 summarize the training and inference procedures, respectively.

B.3 Training Configuration and Inference Protocol

Baseline-aligned training. For each Wan-based baseline (Team Wan et al., 2025), we retain its default training hyperparameters, including the optimizer, learning rates, batch size, noise schedule, and generator and score-model update schedules. Self Forcing, Causal Forcing, and LongLive therefore retain their respective configurations rather than sharing a single set of hyperparameters. We introduce history dropout without changing the underlying training objectives.

Chunkwise inference. We use the standard chunkwise autoregressive inference scheme, as in Self Gradient Forcing (Zhuang et al., 2026): latent frames within each chunk are denoised jointly, and successive chunks condition on the KV cache of previously generated content. Each chunk contains three latent frames. For long-video generation, the cache retains the sink prefix and a sliding window of recent history, following the corresponding baseline's context policy. Each attention call selects non-sink history per head; sink and current tokens and the stored cache remain intact.

Steady-state chunk latency. The latency in Table 1 measures generation of a *continuation chunk in steady state*, with the historical KV cache already populated under the baseline's sliding-window

Algorithm 1 HiDrop training: step-dependent historical value dropout

Require: Text c, chunk noise $z_{1:B}$, timesteps $t_{1:S}$, baseline exit steps $s_{1:B}$, dropout rate $r \in [0, 1)$
Ensure: Generated chunks $(\widehat{x}_1, \ldots, \widehat{x}_B)$ for the baseline training objectives

$p \leftarrow 1 - r$; initialize history cache $\mathcal{C} \leftarrow \varnothing$
Sample $m_j^{(s)} \sim \text{Bernoulli}(p)$ for every token j and step s
▷ Independent draws; reuse $M^{(s)}$ across chunks, layers, and heads
for $b = 1, \ldots, B$ **do**
  $u \leftarrow z_b$
  for $s = 1, \ldots, s_b$ **do**
    $\widehat{x}_b \leftarrow \mathsf{D}_\theta(u, t_s, c, \mathcal{C}; M^{(s)})$
    if $s < s_b$ **then**
      $u \leftarrow \mathsf{ReNoise}(\widehat{x}_b, t_{s+1})$
    end if
  end for
  $\mathcal{C} \leftarrow \mathsf{Refresh}_\theta(\widehat{x}_b, c, \mathcal{C}; M^{(s_b)})$ ▷ Reuse the exit-step mask
end for
return $(\widehat{x}_1, \ldots, \widehat{x}_B)$

function TRAINATTENTION($Q, K, V, H_b; M^{(s)}, p$)
  $\widetilde{V} \leftarrow \textsc{Copy}(V)$ ▷ Temporary values; stored KV is unchanged
  $\widetilde{v}_j \leftarrow (m_j^{(s)}/p)v_j$ for all $j \in H_b$ ▷ Mask and rescale history only
  $\alpha \leftarrow \text{softmax}(QK^\top/\sqrt{d})$ ▷ Keep all keys
  return $\alpha\widetilde{V}$
end function

Algorithm 2 HiDrop inference: sink-preserving per-head KV selection at each attention call

Require: Q, K, V, non-sink history H_b, sink and current tokens C_b, keep ratio $\rho \in (0, 1]$
Ensure: Attention output O_h for each head h

for each head h **do**
  $I_{b,h} \leftarrow C_b$ ▷ Always keep sink and current tokens
  for each nonempty historical frame $F \subseteq H_b$ **do**
    $k \leftarrow \max(1, \lceil \rho|F| \rceil)$
    Uniformly sample k indices $R_{F,h}$ from F without replacement
    $I_{b,h} \leftarrow I_{b,h} \cup R_{F,h}$
  end for
  $(K'_h, V'_h) \leftarrow (K_h[I_{b,h}], V_h[I_{b,h}])$ ▷ Select both keys and values
  $O_h \leftarrow \text{softmax}(Q_h {K'_h}^\top/\sqrt{d})\, V'_h$ ▷ No value rescaling
end for
return $(O_h)_h$ ▷ Apply the baseline head concatenation and projection

policy. The timed unit is one chunk of three latent frames, reported in seconds per chunk. This is neither the time to the first frame from an empty cache nor the total time to generate a 60s or 240s video. Initial cache construction is outside this steady-state measurement. In this regime, non-sink historical tokens are available for selection, so HiDrop can shorten attention inputs. This mechanism does not apply to the first chunk when there is no eligible history.

Each baseline and its HiDrop counterpart use the same spatial resolution, chunk size, denoising schedule, and timing procedure on a single NVIDIA H20. The measured values, 0.72s and 0.64s, give an 11.1% reduction at the displayed precision for steady-state chunk latency. They do not measure first-frame latency, end-to-end throughput, or full-video speedup.

Scope of the reported evaluation. The tables report aggregate scores, not time-resolved measurements of a single long rollout. Establishing reduced error propagation would require within-video temporal evaluation or controlled history perturbations. The reported human preferences likewise describe the sampled video pairs; repeated ratings do not enlarge that video sample.

C Additional Quantitative Experiments

C.1 Separating Dropout Training and Inference Selection

Table 4 reorganizes existing Self Forcing results into a 2×2 comparison: historical value dropout during training on/off, and per-call, per-head KV selection at inference on/off. "Train on" is the "+ Unbiased" checkpoint from Table 2; "Infer on" uses $\rho = 0.5$. The other training and inference settings follow the corresponding controls in the component and inference-only studies. These are the same reported measurements, not additional runs.

Table 4: Training dropout $\times$ inference selection on Self Forcing. "Train" enables step-dependent historical value dropout; "Infer" enables per-head KV selection with $\rho = 0.5$. On/on combines both operations in full HiDrop; off/off uses the unmodified baseline.

Horizon	Train	Infer	AQ↑	BC↑	DD↑	IQ↑	MS↑	SC↑	TF↑
60s	Off	Off	0.582	0.953	0.693	0.703	**0.979**	0.965	0.981
	On	Off	0.586	**0.954**	0.757	0.699	0.976	0.957	**0.986**
	Off	On	0.587	0.952	0.726	0.704	**0.979**	**0.966**	0.982
	On	On	**0.595**	0.953	**0.785**	**0.708**	**0.979**	**0.966**	0.985
240s	Off	Off	0.548	0.956	0.713	0.694	0.979	0.959	0.971
	On	Off	0.555	0.952	0.776	0.695	**0.980**	0.954	**0.980**
	Off	On	0.556	0.958	0.734	0.712	0.979	0.960	0.976
	On	On	**0.560**	**0.959**	**0.790**	**0.714**	0.979	**0.965**	0.979

Inference selection alone improves AQ, IQ, and DD at both horizons. At 240s, IQ increases from 0.694 to 0.712 without dropout training, compared with 0.714 for the full method. With inference selection held fixed, enabling dropout training raises DD from 0.734 to 0.790, but IQ only from 0.712 to 0.714. Training with full-attention inference instead yields 0.695 IQ and 0.954 SC, versus 0.694 and 0.959 for the baseline. The evidence therefore supports a larger additional training benefit in dynamics, not a uniform training-only improvement in every quality or consistency metric.

The combined method has higher AQ, IQ, and DD than either operation alone at both horizons, but this does not establish a positive statistical interaction or generalize the decomposition to Causal Forcing and LongLive. Isolating the remaining choices requires fixed versus step-dependent masks under identical value dropout, full-key value masking with versus without $1/p$ scaling, and shared versus independent head selection at the same token budget.

C.2 How Does the Dropout Rate Affect Generation?

We jointly vary training dropout and inference dropout, $r = 1 - \rho \in \{0.3, 0.5, 0.7\}$, as shown in Table 5. A rate of 0.5 improves aesthetic and visual quality over 0.3 with little change in dynamic degree. Raising it to 0.7 further improves visual consistency but sharply reduces motion. Among the tested settings, 0.5 offers a quality–motion compromise. Training and inference rates are coupled in this comparison, so it cannot attribute the motion loss to either stage separately.

Table 5: Joint dropout rates on Self Forcing: training rate r and inference keep ratio $\rho = 1 - r$.

Horizon	$r = 1 - \rho$	AQ↑	BC↑	DD↑	IQ↑	MS↑	SC↑	TF↑
60s	0.3	0.581	0.953	**0.787**	0.702	0.978	0.959	0.984
	0.5	**0.595**	0.953	0.785	0.708	0.979	0.966	**0.985**
	0.7	0.585	**0.970**	0.390	**0.724**	**0.990**	**0.988**	0.979
240s	0.3	0.548	0.957	**0.793**	0.707	0.976	0.960	0.979
	0.5	**0.560**	0.959	0.790	0.714	0.979	0.965	0.979
	0.7	0.558	**0.971**	0.362	**0.734**	**0.991**	**0.985**	**0.980**

C.3 How Does History Masking Granularity Affect Generation?

We compare history dropout strategies with different masking units. "Chunk" and "Frame" apply chunk-level and frame-level dropout, respectively, during both training and inference. "VMAE" applies the same VideoMAE-style masking strategy (Tong et al., 2022) during both training and inference, with a spatial mask shared across time. This variant adopts only the masking pattern, without adding or optimizing a reconstruction objective.

At 240s, HiDrop has the highest aesthetic quality, imaging quality, and dynamic degree (Table 6). At 60s, Chunk and Frame score higher on aesthetic quality, while HiDrop leads on imaging quality and dynamic degree. Fig. 5 illustrates these comparisons.

Table 6: Masking granularity on Self Forcing. Chunk, Frame, and VMAE use chunk-level dropout, frame-level dropout, and VideoMAE-style masking, respectively, during both training and inference.

Horizon	Method	AQ↑	BC↑	DD↑	IQ↑	MS↑	SC↑	TF↑
60s	Chunk	**0.599**	0.940	0.568	0.700	0.975	0.961	**0.993**
	Frame	0.596	0.946	0.441	0.700	0.975	0.963	0.985
	VMAE	0.590	0.949	0.637	0.696	**0.981**	0.963	0.979
	Ours	0.595	**0.953**	**0.785**	**0.708**	0.979	**0.966**	0.985
240s	Chunk	0.547	0.955	0.537	0.680	0.974	0.962	**0.986**
	Frame	0.544	0.955	0.408	0.699	0.975	0.961	0.982
	VMAE	0.548	**0.961**	0.620	0.703	**0.980**	**0.969**	0.976
	Ours	**0.560**	0.959	**0.790**	**0.714**	0.979	0.965	0.979



Figure 5: Masking-granularity examples at 0, 60, 120, 180, and 240 seconds (Table 6).

C.4 Does History Dropout Also Help Short Videos?

On 5-second videos, HiDrop increases dynamic degree in all three baseline families but lowers motion smoothness and flickering scores; other metrics show mixed changes (Table 7).

Table 7: 5-second comparison with inference keep ratio $\rho = 0.5$. Better scores within each baseline/ours pair are **bold**; tied scores are bold for both methods.

Horizon	Method	AQ↑	BC↑	DD↑	IQ↑	MS↑	SC↑	TF↑
5s	Self Forcing	**0.636**	**0.944**	0.767	0.704	**0.982**	0.960	**0.978**
	Self Forcing + Ours	0.634	0.943	**0.858**	**0.706**	0.980	**0.961**	0.973
	Causal Forcing	**0.643**	**0.957**	0.550	**0.715**	**0.986**	**0.974**	**0.988**
	Causal Forcing + Ours	**0.643**	0.950	**0.675**	0.713	0.985	0.972	0.980
	LongLive	0.638	**0.945**	0.756	0.709	**0.977**	**0.956**	**0.973**
	LongLive + Ours	**0.643**	0.943	**0.822**	**0.711**	0.974	0.953	0.968

D Additional Qualitative Comparisons

Fig. 6 shows two additional four-minute examples per baseline family. These sampled frames illustrate visual appearance but cannot establish motion smoothness.

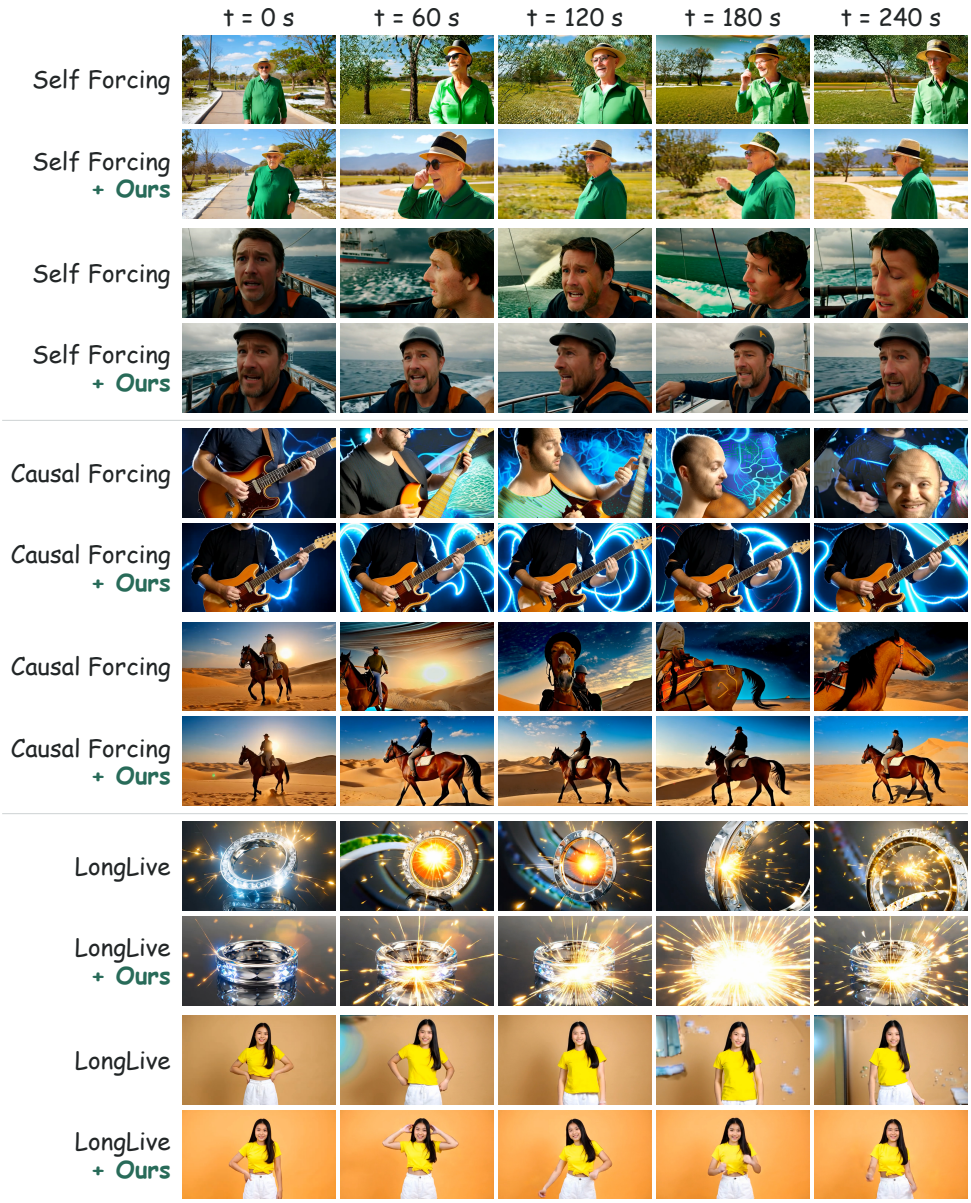

Figure 6: Additional baseline/HiDrop comparisons at timestamps 0, 60, 120, 180, and 240 seconds.

D.1 Does Top-k Token Selection Preserve Visual Quality?

We replace the uniform token sampling used by HiDrop at inference with top-k token selection, leaving training unchanged. Fig. 7 shows sampled frames from this variant, including visible artifacts. This example does not establish that uniform sampling outperforms top-k selection in general.

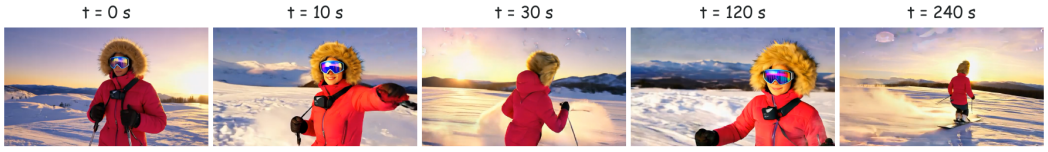

Figure 7: Inference-time top-k token selection in HiDrop.