DELTA VIDEO:LEARNING RESIDUAL UPDATES FOR LONG-TERM VIDEO MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

World modeling requires maintaining persistent information about scenes and objects as the environment evolves. Causal streaming video generation offers a promising route toward this goal, yet limited attention windows restrict access to distant history, making long-term consistency difficult to maintain. Existing context-based memory methods extend this accessible history, but often retain redundant visual information across neighboring frames, leading to inefficient use of memory capacity. Efficiently retaining this history therefore requires exploiting the substantial visual redundancy across frames. We introduce *Delta Video*, a learnable online memory mechanism that learns residual memory updates, suppressing information already captured by the current state while preserving newly observed content. For memory updating, its frame-wise Delta rule predicts incoming visual representations from the current memory state and writes only the prediction residual, suppressing redundant updates for already well-represented information. Retrieved historical context is injected into attention layers to guide ongoing generation. To further teach the memory to handle redundant observations, we introduce redundancy-aware post-training that enforces consistency in both the recurrent memory state and its readout under constructed redundant history, encouraging stable and redundancy-aware memory preservation. On 30-second autoregressive generation, *Delta Video* improves Continuity of Memory by 14.1% over the strongest baseline. Controlled analyses further show that redundant observations induce 53.0% smaller changes in subsequent memory reads, providing direct evidence that *Delta Video* selectively suppresses already-represented information. Project page: `https://delta-video.github.io/`.

1 INTRODUCTION

World models seek to predict how the state of the world changes over time (Mathieu et al., 2016), making generative video a natural substrate for modeling visual dynamics and future states (Ho et al., 2022). A central challenge persists: the underlying world does not reset when the observation window shifts. Long-horizon video generation must go beyond locally realistic appearance and motion to retain, update, and retrieve information across time. Such persistent memory is important for robotics and autonomous driving, where relevant states may temporarily disappear from view.

As video generation advances toward longer-term modeling of evolving environments (Villegas et al., 2022), the research frontier has shifted from polished short clips to longer, temporally coherent generation.However, directly extending the spatiotemporal context of bidirectional video models is computationally expensive (Yin et al., 2025). Causal streaming approaches instead generate videos autoregressively within a bounded local attention window, with advances in causal modeling and distillation improving rollout efficiency and stability (Huang et al., 2025; Zhu et al., 2026; Yang et al., 2025a). However, a longer generation horizon does not imply a longer memory horizon (Henschel et al., 2025). A model may continue producing new frames for an arbitrarily long time, while information from earlier observations ceases to influence generation once it falls outside the local context window. In other words, existing models may learn to generate for longer without necessarily learning to remember for longer.

Prior work addresses limited memory horizons through sliding windows, retrieval, or history compression. Sliding-window methods retain recent context, sometimes augmented with the first frame

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.

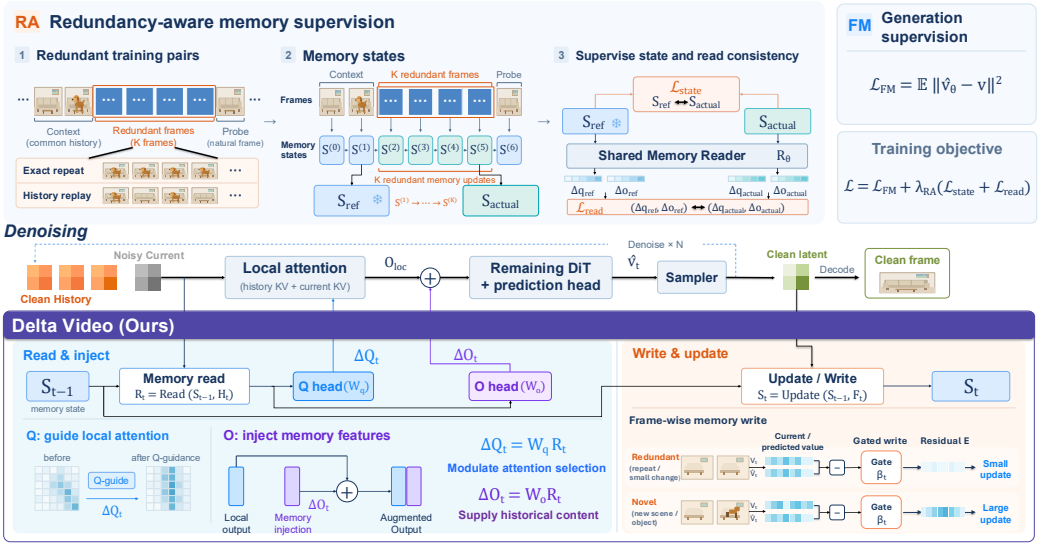


Figure 1: Overview of *Delta Video*. *Delta Video* retrieves persistent history through dual $\Delta Q/\Delta O$ readout branches and performs frame-wise gated residual updates after each generated frame. Redundancy-aware supervision encourages stable memory states and readouts under redundant histories, enabling selective long-term memory updates.

or selected keyframes (Huang et al., 2025; Yang et al., 2025a), but discard most distant history and cannot capture states that emerge later. Retrieval-based methods select relevant historical frames or tokens (Li et al., 2025; Cai et al., 2025), reducing active context but not necessarily consolidating the underlying historical store. Other approaches compress out-of-window history into lower-resolution representations, hierarchical caches, or contextual summaries (Zhang et al., 2025; Yu et al., 2025). However, these methods largely organize history by temporal position rather than by how visual content persists and changes over time.

Temporal distance is a poor proxy for visual novelty. Distant observations may preserve information important for future generation, whereas adjacent frames are often highly redundant. Memory organized primarily by recency may therefore repeatedly store already represented content while discarding older but persistent information. Effective long-term memory should instead preserve stable information and selectively incorporate meaningful changes.

To this end, we introduce *Delta Video*, a learnable online memory mechanism for causal streaming video generation. It maintains a fixed-size recurrent state and adapts Delta-style residual updates (Schlag et al., 2021; Yang et al., 2025c;b) to frame-wise video memory. The current memory predicts each incoming frame representation, and only the gated residual is written back, yielding small updates for already represented content and larger corrections for novel information. Historical context is retrieved through query- and output-side readout heads, while redundancy-aware supervision stabilizes memory under redundant observations. This enables long-range memory with storage independent of the generated video length.

Our main contributions are summarized as follows:

- **Frame-wise Residual Memory Updates.** We introduce a fixed-size recurrent video memory that adapts Delta-style prediction-residual updates to frame-wise generation, enabling already represented visual information to induce small updates while preserving sensitivity to novel content.
- **Redundancy-Aware Memory Supervision.** We introduce state- and readout-consistency objectives that explicitly stabilize memory under redundant histories, improving the recovery of previously observed content by 33.1% in our ablation setting.
- **Dual-Path Memory Readout.** We introduce lightweight query- and output-side readout heads that respectively modulate local attention and inject retrieved historical features.

Combining both pathways improves long-term memory consistency by 19.1–21.3% over either single-path variant.

Overall, *Delta Video* improves long-term memory consistency by 14.1% over the strongest baseline on 30-second autoregressive generation, while maintaining competitive generation quality.

2 PRELIMINARIES

Rectified Flow for Video Generation. Rectified Flow Liu et al. (2022) learns a velocity field along the linear interpolation between data and noise. Given $\mathbf{x}_0 \sim p_{\text{data}}$ and Gaussian noise ϵ,

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\epsilon, \qquad t \in [0,1], \tag{1}$$

and the velocity model is trained by

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\|v_\theta(\mathbf{x}_t,t) - (\epsilon - \mathbf{x}_0)\|_2^2\right]. \tag{2}$$

Linear Attention and Delta Rule. Linear attention summarizes past key-value pairs into fixed-size recurrent states:

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{v}_t\phi(\mathbf{k}_t)^\top, \qquad \mathbf{z}_t = \mathbf{z}_{t-1} + \phi(\mathbf{k}_t), \tag{3}$$

with normalized readout

$$\mathbf{o}_t = \frac{\mathbf{S}_t\phi(\mathbf{q}_t)}{\mathbf{z}_t^\top\phi(\mathbf{q}_t)}. \tag{4}$$

Because additive updates accumulate every observation, repeated content can be overrepresented in memory. The Delta Rule Yang et al. (2025b) instead writes the prediction residual:

$$\hat{\mathbf{v}}_t = \mathbf{S}_{t-1}\phi(\mathbf{k}_t), \qquad \mathbf{S}_t = \mathbf{S}_{t-1} + \beta_t(\mathbf{v}_t - \hat{\mathbf{v}}_t)\phi(\mathbf{k}_t)^\top, \tag{5}$$

where $\beta_t \in [0,1]$ controls the update strength and $\phi(\cdot)$ includes key normalization. Thus, already well-represented content induces smaller memory updates.

3 METHOD

3.1 ONLINE UPDATED MEMORY FOR CAUSAL VIDEO GENERATION

For causal video generation with autoregressive rollout under a bounded local attention window, we augment self-attention with a recurrent memory state $S_{t-1} \in \mathbb{R}^{d_v \times d_k}$, to preserve information fall outside the attention window. Here, d_k and d_v denote the memory key and value dimensions, respectively. Each attention layer and attention head maintains its own memory state shared across time. *Delta Video* follows a simple read–generate–write cycle. Before generating frame F_t, the current hidden representation queries S_{t-1} to retrieve historical information. The resulting readout is injected into local self-attention by q and o readout head. When the frame has been fully generated, we update the state using its clean-frame features:

$$S_{t-1} \xrightarrow{\text{read}} R_t \xrightarrow{\text{generation}} F_t \xrightarrow{\text{write}} S_t. \tag{6}$$

3.2 READING PERSISTENT HISTORY INTO LOCAL ATTENTION

During frame generation, *Delta Video* reads the fixed state S_{t-1} and injects the retrieved history through two residual paths (Alg. 1, Lines 1–5). Specifically,

$$Q_t^m = \phi(H_t W_Q^m), R_t = \frac{Q_t^m S_{t-1}^\top}{Q_t^m z_{t-1} + \epsilon}, \widetilde{Q}_t = Q_t + W_q R_t, \widetilde{O}_t = \text{LocalAttn}(\widetilde{Q}_t, K_t, V_t; C_t) + W_o R_t. \tag{7}$$

The two paths serve different roles. ΔQ_t provides history-guided selection by changing how current tokens attend to the local context, whereas ΔO_t provides history-conditioned supplementation by directly injecting information that may no longer exist in the local window. During denoising, the memory state S_{t-1} is read-only and remains unchanged. Memory updating is triggered only after the clean frame x_t has been generated.

Algorithm 1 Framewise read–write with *Delta Video*.

Input: Hidden tokens H_t; previous memory states (S_{t-1}, z_{t-1}); local context $\mathcal{C}_t$; clean-frame features $\{H_{t,f}\}_{f=1}^{F}$; memory parameters $\Theta_m = \{W_Q^m, W_K^m, W_V^m, W_\beta^m, W_\lambda^m, b_\lambda, W_q, W_o\}$; stabilization constant ϵ

Output: Memory-augmented output $\widetilde{O}_t$; updated memory states (S_t, z_t)

Memory read during frame generation

$(Q_t, K_t, V_t) \leftarrow \textsc{Proj}(H_t)$

Query memory: $Q_t^m \leftarrow \phi(H_t W_Q^m), \quad R_t \leftarrow \dfrac{Q_t^m S_{t-1}^\top}{Q_t^m z_{t-1} + \epsilon}$

Inject the memory readout into the attention queries: $\Delta Q_t \leftarrow W_q R_t, \quad \widetilde{Q}_t \leftarrow Q_t + \Delta Q_t$

$O_t^{\mathrm{loc}} \leftarrow \textsc{LocalAttn}(\widetilde{Q}_t, K_t, V_t; \mathcal{C}_t)$

Fuse the memory readout with the attention output: $\Delta O_t \leftarrow W_o R_t, \quad \widetilde{O}_t \leftarrow O_t^{\mathrm{loc}} + \Delta O_t$

Frame-wise delta write after clean generation

Initialize the working states: $S \leftarrow S_{t-1}, \quad z \leftarrow z_{t-1}$

for $f = 1, \ldots, F$ **do**

  Encode the current frame as memory keys and values: $K_f^m \leftarrow \phi(H_{t,f} W_K^m), \quad V_f^m \leftarrow H_{t,f} W_V^m$

  Compute token-wise update gates: $\beta_f^m \leftarrow \sigma(H_{t,f} W_\beta^m)$

  Compute the frame-level scalar retention gate: $\bar{H}_{t,f} \leftarrow \textsc{AvgPool}(H_{t,f}), \quad \lambda_{t,f} \leftarrow \sigma(\bar{H}_{t,f} W_\lambda^m + b_\lambda)$

  Write the delta residual into memory:
  $S \leftarrow \lambda_{t,f} S + \left[\beta_f^m \odot \left(V_f^m - \lambda_{t,f} S K_f^m\right)\right] (K_f^m)^\top$

  Update the denominator state:
  $z \leftarrow \lambda_{t,f} z + K_f^m \left[\beta_f^m \odot \left(\mathbf{1} - \lambda_{t,f} (K_f^m)^\top z\right)\right]$

end for

Commit the updated states: $S_t \leftarrow S, \quad z_t \leftarrow z$

return $(\widetilde{O}_t, S_t, z_t)$

3.3 Frame-wise Delta Rule Update

After frame t is generated, we re-encode the clean frame through the DiT to obtain $H_t^{\mathrm{clean}} \in \mathbb{R}^{N_t \times d}$. We compute

$$K_t = \frac{1}{\sqrt{rN_t}} \phi(H_t^{\mathrm{clean}} W_K^m),\ V_t \quad = H_t^{\mathrm{clean}} W_V^m,\ \beta_t = \sigma(H_t^{\mathrm{clean}} W_\beta^m), \tag{8}$$

where $\phi(\cdot)$ denotes RMS normalization followed by ReLU. A frame-level retention gate λ_t is predicted from the spatially averaged frame feature.

The memory is updated by

$$S_t = \lambda_t S_{t-1} + \left[\beta_t \odot \left(V_t - \lambda_t K_t S_{t-1}^\top\right)\right]^\top K_t. \tag{9}$$

The delta residual $V_t - \lambda_t K_t S_{t-1}^\top$ captures information not already represented by the retained memory, while β_t controls token-wise writing and λ_t controls memory retention.

All spatial tokens within a frame are evaluated against the same previous state and aggregated into a single update, so recurrence occurs once per frame rather than once per token.

For each attention head, memory requires $\mathcal{O}(d_v d_k)$ storage for S_t, with an additional $\mathcal{O}(d_k)$ denominator state. Both reading and writing scale linearly with the number of current-frame tokens and remain independent of the generated history length.

3.4 Redundancy-Aware Memory Supervision

The Delta update in Eq. (9) enables selective writing by suppressing predictable content, but prediction error alone does not reliably distinguish semantic novelty from trivial appearance variations. More importantly, causal-forcing training provides no explicit supervision for distinguishing redundant observations from novel information. As a result, the redundancy-aware behavior enabled by the Delta update may not be fully learned. We therefore introduce redundancy-aware post-training to explicitly encourage stable memory states and readouts under redundant observations while preserving sensitivity to meaningful changes.

Repeated-history construction. Starting from a memory state $S^{(0)}$, which we denote as the reference state S_{ref}, we replay K frames whose content is already represented by the current history. Applying the regular Delta write recurrently gives

$$S^{(k)} = \text{Write}\left(S^{(k-1)}, H_k^{\text{clean}}\right), \qquad k = 1, \ldots, K, \tag{10}$$

with $S_{\text{actual}} = S^{(K)}$. The reference and replay trajectories therefore contain the same historical information, but the latter exposes the memory to redundant writes.

State consistency. If a replayed observation carries no new historical information, its write should not substantially perturb the existing memory. We directly regularize every redundant transition using a normalized state reconstruction loss:

$$\mathcal{L}_{\text{state}} = \frac{1}{K}\sum_{k=1}^{K} \text{RSE}\left(S^{(k)}, \text{sg}\left[S^{(k-1)}\right]\right), \tag{11}$$

where $\text{sg}[\cdot]$ denotes stop-gradient and

$$\text{RSE}(A, B) = \frac{\|A - B\|_F^2}{\|B\|_F^2 + \epsilon}. \tag{12}$$

This objective directly encourages repeated writes to preserve the established persistent state.

Readout consistency. State similarity alone does not guarantee that two memories behave equivalently when queried by the generator. We therefore additionally constrain their functional effect on the next natural frame. Let H_p denote a probe frame immediately following the replayed history. Using the same memory reader for both states, we compute

$$\begin{aligned}(\Delta Q_p^{\text{ref}}, \Delta O_p^{\text{ref}}) &= \text{Reader}\left(H_p, \text{sg}[S_{\text{ref}}]\right), \\ (\Delta Q_p^{\text{actual}}, \Delta O_p^{\text{actual}}) &= \text{Reader}\left(H_p, S_{\text{actual}}\right).\end{aligned} \tag{13}$$

We then minimize the discrepancy between the query and output corrections produced by the two histories:

$$\mathcal{L}_{\text{read}} = \text{RSE}\left(\Delta Q_p^{\text{actual}}, \Delta Q_p^{\text{ref}}\right) + \text{RSE}\left(\Delta O_p^{\text{actual}}, \Delta O_p^{\text{ref}}\right). \tag{14}$$

While $\mathcal{L}_{\text{state}}$ regularizes what is written, $\mathcal{L}_{\text{read}}$ constrains what remains functionally retrievable. Together, they prevent redundant observations from progressively altering the memory while preserving its effect on future generation.

The full post-training objective is

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda_{\text{RA}}(\mathcal{L}_{\text{state}} + \mathcal{L}_{\text{read}}), \tag{15}$$

where $\mathcal{L}_{\text{FM}}$ is the standard flow-matching loss of the base DiT, corresponding to the original training objective used in each stage of CF++. In our implementation, the causal-distillation objective is retained as $\mathcal{L}_{\text{FM}}$, and we use $\lambda_{\text{RA}} = 0.1$.

Importantly, redundancy-aware supervision is used only during post-training. It introduces no additional modules or computation at inference time; the resulting model performs the same online read–generate–write recurrence described above.

4 Experiments

4.1 Experimental Details

Implementation We build on the frame-wise causal Wan2.1-T2V-1.3B checkpoint released with Causal Forcing++ (Zhao et al., 2026; Wan et al., 2025). Starting from this checkpoint, we introduce *Delta Video* at Stage-2 and train the augmented model using the original Causal Forcing++ causal-distillation objective. The memory-specific Q/K/V projections, update gates, and Q/O readout heads are newly initialized at this stage. We then insert an additional Stage-2.5 redundancy-aware post-training stage, where the memory is further optimized with the proposed state- and readout-consistency objectives. Afterward, we continue with Stage-3 DMD following Causal Forcing++, and finally perform Stage-3.5 long-video adaptation using a LongLive-style DMD objective.

Table 1: Quantitative comparison with existing video generation methods. Long-term memory is evaluated on MemoBench-Auto and MemoBench-VQA, and general video quality on VBench-Long, using 30 s autoregressive generation. All metrics are reported on a 0–100 scale.

	Memory			General			
	MemoBench-Auto		MemoBench-VQA	VBench-Long			
Method	ORS ↑	Camera Controllability ↑	Continuity of Memory ↑	Quality ↑	Subject Consistency ↑	Dynamic Degree ↑	Imaging Quality ↑
Self Forcing	35.86	16.75	25.37	78.43	97.67	30.13	68.46
Causal Forcing++ (framewise-2step)	35.18	17.34	41.81	78.36	98.21	23.26	70.24
LongLive 1.0	36.93	15.36	32.55	79.39	**98.61**	19.51	70.89
LongLive 2.0	31.69	**17.61**	38.18	79.71	98.39	33.43	69.87
Rolling Forcing	42.27	14.60	36.37	78.86	98.35	17.65	71.06
Deep Forcing	31.62	11.79	37.62	80.47	97.88	39.90	**71.40**
Delta Video (Ours)	**45.02**	16.49	**47.70**	**80.99**	96.74	**58.71**	71.12

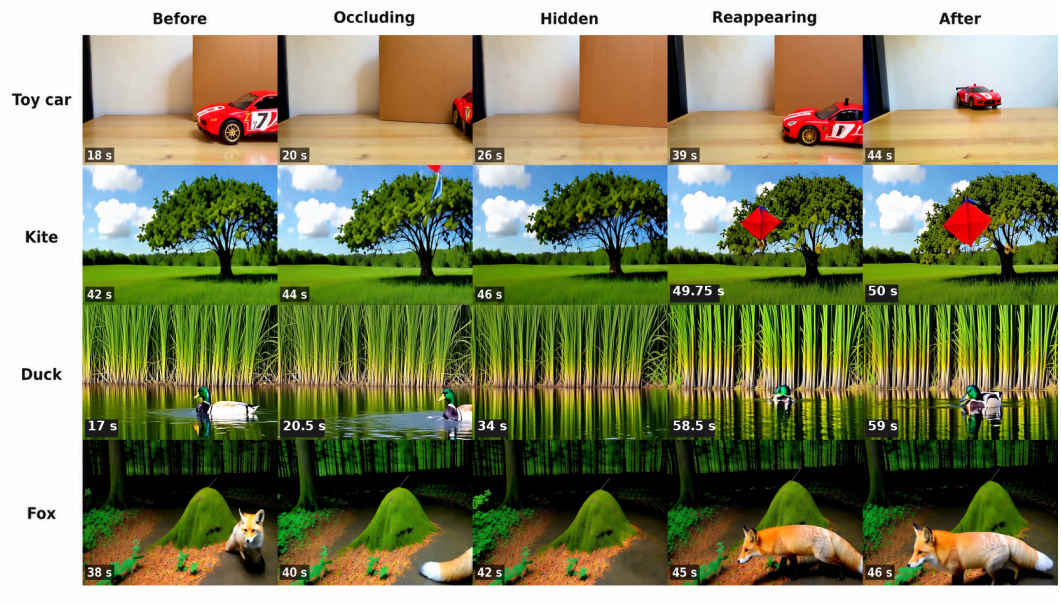


Figure 2: Qualitative examples are shown from longer rollouts of up to 60s. Each row shows a target before disappearance, while hidden, upon reappearance, and afterward.

Evaluation metric. We use MemoBench to evaluate long-term memory (Chen et al., 2026a). following its existing evaluation protocol. We report Object Revisit Score (ORS), Camera Controllability, and Continuity of Memory (CoM). ORS measures target reappearance based on SAM-3 detection, Camera Controllability measures consistency with the specified camera trajectory, and CoM evaluates whether the identity and state of revisited targets remain consistent over long temporal gaps using Gemini-3.1-Pro as a VQA evaluator. ORS and Camera Controllability are evaluated on 360 videos (196 Synthetic and 164 Real).

We use the VBench-I2V benchmark suite to evaluate general video quality (Huang et al., 2023; 2024; Zheng et al., 2025). We adopt the extended prompts from Self Forcing, obtained by expanding the official VBench prompts. For the 30s generation results, we evaluate on the VBench-I2V data using the long-video evaluation protocol of VBench-Long. All metrics are reported on a 0–100 scale.

4.2 Experimental Results

We compare our method with relevant open-source video generation baselines, we evaluate 30-second generations. (Huang et al., 2025; Yin et al., 2025; Zhao et al., 2026; Yang et al., 2025a; Chen et al., 2026b; Liu et al., 2026; Yi et al., 2025). As shown in Table 1, *Delta Video* achieves the highest ORS of 45.02 on MemoBench-Auto and a CoM score of 47.70 on MemoBench-VQA, demonstrating substantial improvements in long-term memory over existing video generation meth-

Table 2: Ablation results with a 20s inference horizon. We report Object Revisit Score (ORS), Continuity of Memory (CoM), and Quality, all on a 0–100 scale (↑: higher is better).

(a) RA and Update Rule

Update rule	RA	ORS↑	CoM↑	Quality↑
Additive	×	29.46	27.30	**80.87**
Additive	✓	29.02	27.96	78.46
Delta Rule	×	24.60	30.64	80.02
Delta Rule	✓	**32.74**	**32.04**	80.16

(b) Readout branches

ΔQ	ΔO	ORS↑	CoM↑	Quality↑
×	✓	31.51	26.41	78.84
✓	×	28.48	26.89	79.31
✓	✓	**32.74**	**32.04**	**80.16**

(c) Update granularity

Setting	ORS↑	CoM↑	Quality↑
Global pool	31.07	29.63	79.34
2×2 pool	23.21	26.07	79.24
Full tokens	**32.74**	**32.04**	**80.16**

ods. These results support the effectiveness of the proposed memory mechanism in preserving information throughout long autoregressive rollouts.

Beyond memory-specific metrics, *Delta Video* also maintains comparable general video quality to existing long-video baselines. On VBench-I2V, it achieves 80.99 in Quality, 58.71 in Dynamic Degree, and 71.12 in Imaging Quality, indicating that the improved memory capability does not come with an evident degradation in general generation quality. This is further supported by the qualitative results in Fig. 2, where *Delta Video* preserves previously observed content over long horizons while maintaining coherent and visually plausible generation. Overall, these results suggest that the proposed memory mechanism improves long-horizon memory while largely preserving the generation quality of the underlying video model.

4.3 Ablation

We conduct ablation studies to assess the contribution of each key component of our method. The ablation variants are illustrated in Figure 3, with their quantitative results summarized in Table 2. For computational efficiency, all variants follow the same multi-stage training pipeline as the main model, but use fewer optimization steps at each stage. All ablation experiments are evaluated with a 20-second inference horizon. Detailed training configurations are provided in Appendix A.6.

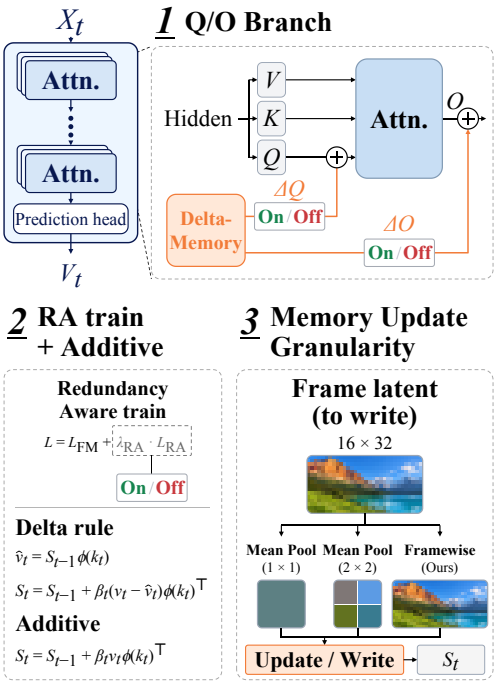


Figure 3: Overview of the ablated components, including memory readout branches, redundancy-aware supervision with different update rules, and memory-update granularity.

Readout branches We examine the two memory readout branches, ΔQ and ΔO. Removing ΔQ decreases ORS and CoM by 1.23 and 5.63 points, respectively, while removing ΔO leads to drops of 4.26 and 5.15 points. These results indicate that both branches contribute to effective memory readout.

Redundancy-aware supervision and update rule. We jointly ablate the memory update rule and redundancy-aware supervision (RA). The Delta Rule alone does not outperform the additive update, but combining it with RA yields the strongest memory performance, reaching 32.74 ORS and 32.04 CoM. In contrast, applying RA to the additive update degrades generation quality. These results reveal a clear interaction between the two designs, suggesting that RA is effective specifically with the Delta-style residual update rather than as a generic auxiliary supervision.

Update granularity. We compare different update granularity. Following prior work (Yu et al., 2025), global pooling compresses each frame into a single token for one memory update. The 2×2

variant divides the frame into four regions and sequentially updates memory with their averaged tokens, while the full-token variant writes all spatial tokens directly. Full-token updates achieve the best memory performance, highlighting the importance of preserving fine-grained spatial information. Interestingly, the 2×2 variant performs substantially worse, suggesting that repeated regional updates may require more effective redundancy-aware control to avoid unnecessary interference with the memory state.

4.4 Interpretability Experiments

Redundancy-aware temporal updates We first investigate whether *Delta Video* selectively attenuates redundant visual information. We construct 128 held-out 30s videos, each with two teacher-forced histories sharing the same prefix: a natural sequence and a controlled sequence with four repeated consecutive frames. We quantify this effect using two metrics: the next-read change, which measures how much the current write alters the subsequent memory read, and the memory state change, which measures the relative magnitude of the corresponding state update. As shown in Figure4, across the 128 samples, repeated inputs reduce the mean next-read change from 0.0105 to 0.0049, corresponding to a 53.0% reduction. The mean memory state change similarly decreases from 0.0370 to 0.0179, corresponding to a 51.7% reduction. These results indicate that repeated inputs induce weaker memory updates and exert less influence on subsequent memory reads.

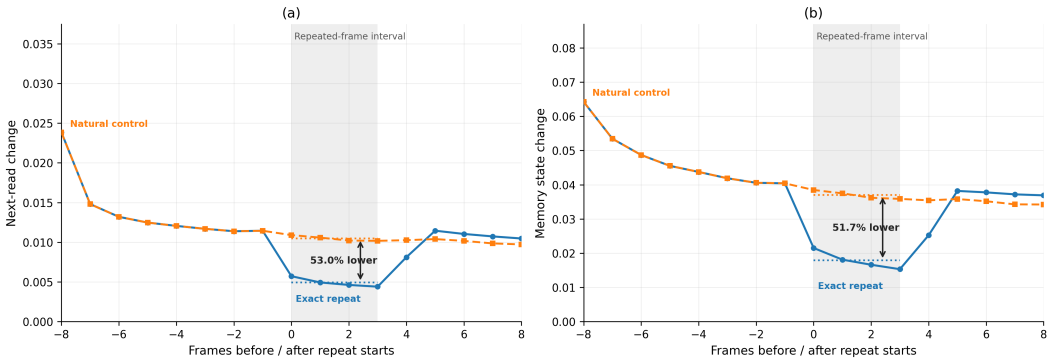


Figure 4: **Redundancy-aware temporal updates.** **(a)** Next-read change and **(b)** memory-state change under natural observations and exact frame repeats. During the repeated-frame interval, exact repeats reduce the two quantities by 53.0% and 51.7%, respectively. Both recover toward the natural trajectory once repetition ends, indicating that *Delta Video* selectively suppresses redundant writes rather than globally reducing memory updates.

Spatial alignment between visual change and memory correction We next examine whether *Delta Video* allocates its updates to spatial regions containing new visual information. For each spatial token, we measure the RMS latent change between adjacent natural frames and the normalized magnitude of the corresponding memory correction,

$$\widetilde{R}_i = \frac{\mathrm{RMS}(\beta_i\,(v_i - \lambda_t S_{t-1} k_i))}{\mathrm{RMS}(V_i) + \epsilon},$$

where $S_{t-1}k_i$ denotes the content predicted from the current memory state.We then compute the Spearman correlation between the two token-grid maps for each sample. Regions with strong latent changes consistently coincide with regions receiving large memory corrections. As shown in Figure5, across 128 held-out samples, the mean spatial Spearman correlation is $\rho = 0.605$, with a variance of 0.013, and all 128 samples exhibit positive correlations. This spatial correspondence shows that *Delta Video* concentrates its updates on regions containing stronger temporal changes.

Structured and causally effective memory states. We examine whether *Delta Video* stores content-dependent structure, whether such information persists beyond the local context, and whether the retained state is actually used by the model. As shown in Fig. 6, we first construct a controlled set of six object types across four scenes, placing segmented objects on a shared

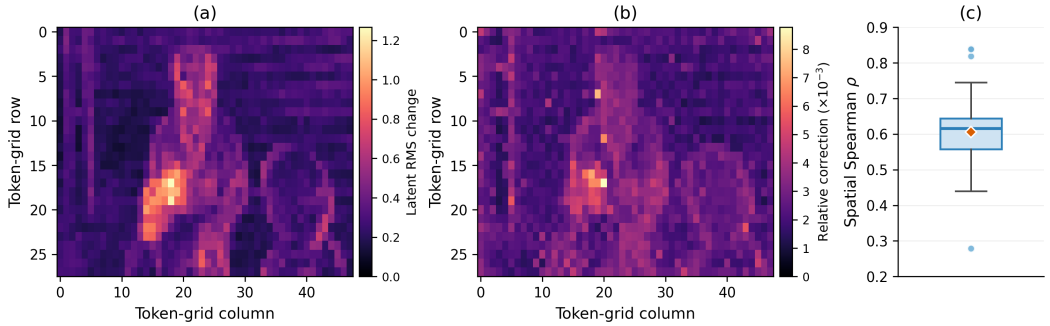


Figure 5: **Spatial alignment between visual change and memory correction. (a)** Temporal latent change and **(b)** relative memory-correction magnitude exhibit similar spatial patterns. **(c)** Their spatial Spearman correlation remains consistently positive across samples ($\rho \approx 0.61$). Regions that change more receive stronger memory corrections, showing that *Delta Video* spatially focuses its updates on newly changing visual content.

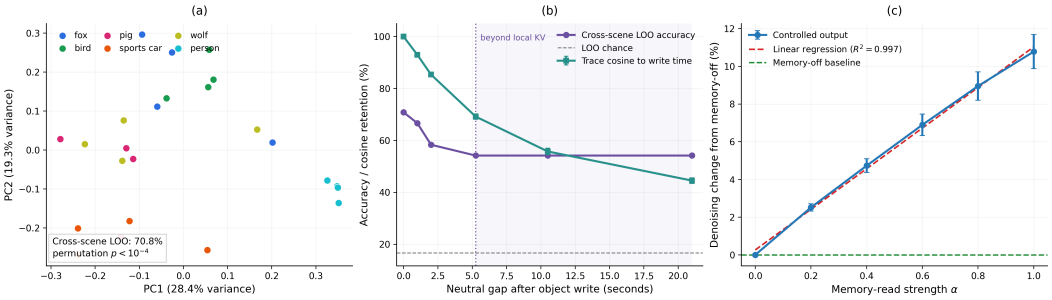


Figure 6: **Structured, persistent, and causally effective recurrent memory. (a)** Visual contents induce distinct memory-update directions, revealing content-structured representations. **(b)** These traces remain distinguishable beyond the local KV window, demonstrating persistent long-range retention. **(c)** Memory-only interventions produce a graded, nearly linear change in the predicted flow ($R^2 = 0.997$), showing that the recurrent state actively conditions denoising. Together, *Delta Video* stores structured information, preserves it beyond local context, and causally influences generation.

gray background while fixing the text conditioning. In panel (a), memory updates exhibit clear content-dependent structure, achieving 70.8% cross-scene classification accuracy, far above the 16.6% chance level. while PCA visualizes this separation. We test whether these traces persist through subsequent updates by feeding identical neutral frames after the object write and subtracting a matched all-neutral baseline. As shown in panel (b), the traces gradually decay but remain content-dependent well beyond the local KV window: Even after a 21s neutral gap, cross-scene classification remains well above chance, retaining an accuracy of around 50%. We further scaling the memory-read strength α while holding all other inputs fixed produces an almost linear increase in the predicted-flow change ($R^2 = 0.997$), reaching 10.8% at full readout. This demonstrates that the recurrent memory acts as a causally effective conditioning signal for denoising.

5 Limitations and Conclusion

We present *Delta Video*, a training framework that augments autoregressive video generation with a fixed-size recurrent memory, dual-path memory readout, and redundancy-aware supervision, improving long-horizon memory consistency while largely preserving the visual quality and temporal dynamics of the underlying video model. Despite these improvements, the fixed-capacity recurrent state may still lose fine-grained information over increasingly long horizons. Moreover, the proposed memory module incurs additional computation during inference, which may affect latency-sensitive streaming generation scenarios.

ETHICS STATEMENT

This work is primarily methodological and does not involve human-subject data collection. As with other video generation systems, it may inherit risks related to synthetic-media misuse, dataset bias, privacy, and copyright. Downstream deployment, particularly in world-modeling or action-selection systems, should therefore include appropriate safety evaluation and safeguards.

REPRODUCIBILITY STATEMENT

We provide detailed implementation and evaluation settings to facilitate reproduction of our results. A describes the training data, multi-stage training pipeline, optimization hyperparameters, Delta-Video architecture, inference configuration, and ablation settings. We additionally provide the construction details for redundancy-aware post-training and the protocols used in our interpretability analyses. The memory read–write procedure is fully specified in Algorithm 1, and the redundancy-aware objectives are defined in Sec. 3.4.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation. We did not use generative AI tools to develop theoretical models or conceptual frameworks; implement methods; clean or reformat datasets; propose or refine hypotheses; design or provide feedback on research methodology or experiments; or interpret results, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to edit the paper for readability. We reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation, 2025. URL https://arxiv.org/abs/2508.21058.

Haoyu Chen, Kaichen Zhou, Hang Hua, Kaile Zhang, Jingwen Qian, Wufei Ma, Haonan Chen, Chunjiang Liu, Yizhou Zhao, Xiaoyuan Wang, Weiyue Li, Alan Yuille, Paul Pu Liang, and Yilun Du. Memobench: Benchmarking world modeling in dynamically changing environments, 2026a. URL https://arxiv.org/abs/2606.27537.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, Haozhe Liu, Hongwei Yi, Hao Zhang, Muyang Li, Yukang Chen, Han Cai, Sanja Fidler, Ping Luo, Song Han, and Enze Xie. Sana-video: Efficient video generation with block linear diffusion transformer, 2025. URL https://arxiv.org/abs/2509.24695.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, Yuyang Zhao, Huizi Mao, Ying-Cong Chen, Enze Xie, Xiaojuan Qi, and Song Han. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation, 2026b. URL https://arxiv.org/abs/2605.18739.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

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, 2022. URL https://arxiv.org/abs/2204.03458.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion, 2025. URL https://arxiv.org/abs/2506.08009.

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, 2023. URL https://arxiv.org/abs/2311.17982.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models, 2024. URL https://arxiv.org/abs/2411.13503.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention, 2020. URL https://arxiv.org/abs/2006.16236.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory, 2025. URL https://arxiv.org/abs/2506.18903.

Shanchuan Lin, Ceyuan Yang, Hao He, Jianwen Jiang, Yuxi Ren, Xin Xia, Yang Zhao, Xuefeng Xiao, and Lu Jiang. Autoregressive adversarial post-training for real-time interactive video generation, 2025. URL https://arxiv.org/abs/2506.09350.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep Multi-Scale Video Prediction Beyond Mean Square Error. In *International Conference on Learning Representations*, 2016. URL https://mlanthology.org/iclr/2016/mathieu2016iclr-deep/.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling, 2024. URL https://arxiv.org/abs/2310.15169.

Nedko Savov, Naser Kazemi, Deheng Zhang, Danda Pani Paudel, Xi Wang, and Luc V Gool. Statespacediffuser: Bringing long context to diffusion world models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 68865–68898. Curran Associates, Inc., 2025. doi: 10.52202/085713-2316. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/63943ee9fe347f3d95892cf87d9a42e6-Paper-Conference.pdf.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers, 2021. URL https://arxiv.org/abs/2102.11174.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description, 2022. URL https://arxiv.org/abs/2210.02399.

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, 2025. URL https://arxiv.org/abs/2503.20314.

Zeqi Xiao, Yushi LAN, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 49632–49652. Curran Associates, Inc., 2025. doi: 10.52202/085713-1659. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/470629a47e2d65ce0606c40055df5d26-Paper-Conference.pdf.

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*, 2025a.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training, 2024. URL https://arxiv.org/abs/2312.06635.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule, 2025b. URL https://arxiv.org/abs/2412.06464.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length, 2025c. URL https://arxiv.org/abs/2406.06484.

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, 2025d. URL https://arxiv.org/abs/2408.06072.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models, 2025. URL https://arxiv.org/abs/2412.07772.

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, and Xiaojuan Qi. Videossm: Autoregressive long video generation with hybrid state-space memory, 2025. URL https://arxiv.org/abs/2512.04519.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models, 2025. URL https://arxiv.org/abs/2504.12626.

Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers, 2025. URL https://arxiv.org/abs/2502.15894.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. URL https://arxiv.org/abs/2503.21755.

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.

A Training and Inference Details

A.1 Training Data

Stage 2 and Stage 3 use OpenVid-HD latent data recaptioned by Tarsier2. Stage 2 uses a fixed subset of 80K high-quality and high-dynamic samples, while the subsequent redundancy-aware post-training stage uses a 15K dataset constructed by sampling from OpenVid-80K, and the DMD stages use OpenVid prompts and their corresponding first frames as image-to-video conditions. OpenVid samples are pre-encoded into Wan latents of shape $21 \times 16 \times 56 \times 96$ and are paired with UMT5 text embeddings.

A.2 Training Pipeline

We initialize our model from the frame-wise causal checkpoint released with Causal Forcing++ and train the proposed memory path through a multi-stage pipeline.

Stage 2: causal consistency distillation. Stage 2 follows the causal distillation pipeline of Causal Forcing++. We train on 80K OpenVid subset using their natural sampling ratio. Each training sample contains 21 latent frames and uses pure T2V parallel teacher forcing. Both the causal generator and the Delta-Memory modules are optimized.We train Stage 2 for 5k steps.

Redundancy-aware post-training. After Stage 2, we perform an additional memory-only post-training stage on an OpenVid redundancy-recipe subset. The generator backbone is frozen and only the Delta-Memory modules are updated using the redundancy-aware objectives.We train redundancy-aware post training for 1k steps.

Stage 3: asymmetric DMD. We initialize Stage 3 from the redundancy-aware Stage-2.5 checkpoint and follow the asymmetric DMD training pipeline. Training uses 15K quality-filtered OpenVid samples with a 1:1 mixture of T2V and I2V examples. The generator and fake-score network are jointly optimized, while the real-score model is kept fixed.We train this stage for 1.2k steps.

Stage 3.5: long-rollout training. Finally, we extend the model to long autoregressive rollouts following the LongLive-style training setup.We train this stage for 1k steps.

A.3 Training Hyperparameters

All stages use BF16 mixed precision and AdamW. Unless otherwise specified, we use $\beta_1 = 0$, $\beta_2 = 0.999$, and a weight decay of 0.01. The base generator is Wan2.1-T2V-1.3B, while the real-score network used during Stage 3 is Wan2.1-T2V-14B. The causal generator operates autoregressively at the granularity of one latent frame, with a local attention/KV window of 21 latent frames.

The training hyperparameters are summarized in Table 3.

Table 3: Training hyperparameters for each stage.

Stage	Global batch	Steps	Generator LR	Critic LR	Additional settings
Stage 2	48	5,000	2×10^{-6}	4×10^{-7}	
RA post-training	48	1,000	5×10^{-6}	–	memory only; state/read loss weights 0.1
Stage 3	48	1,200	2×10^{-6}	4×10^{-7}	T2V:I2V = 1:1
Stage 3.5	48	1,000	LoRA 1×10^{-5}; memory 2×10^{-6}	2×10^{-6}	LoRA rank/alpha = 256/256; T2V:I2V = 1:1

EMA is enabled with a decay of 0.99 from step 200 for Stage 2 and Stage 3, and from step 0 for the redundancy-aware stages.

A.4 Delta-Memory Architecture

The Delta-Memory module is inserted into the attention blocks of the causal DiT and maintains an independent recurrent state for each layer and attention head. Each state uses a memory rank of 128 and a scaling factor of 256. The local softmax-attention window remains fixed at 21 latent frames.

Table 4: Shared Delta-Memory architecture hyperparameters.

Hyperparameter	Value
Memory rank	128
Scaling factor	256
Local attention window	21 latent frames
Write granularity	Frame tokens
Readout branches	Query + output
Q/K/V projections	Independent
Update rule	Gated delta update
Q/K kernel	Sana-ReLU
β bias initialization	4.0
λ bias initialization	4.0
Denominator ϵ	10^{-6}
Readout gate initialization	5×10^{-3}
AdaLN gate initialization	10^{-2}

Table 5: Default inference configuration.

Setting	Value
Resolution	768×448
Frame rate	16 fps
Autoregressive unit	1 latent frame
Regular-frame denoising	$[1000, 500]$
First-chunk denoising	$[1000, 750, 500, 250]$
CFG scale	3.0
Timestep shift	5.0
Local KV window	21 latent frames
Precision	BF16

The memory uses independent projections for its queries, keys, and values. Memory readout is injected through two lightweight residual branches: a query-side branch that modifies the attention query and an output-side branch that directly augments the attention output. The default model uses both branches, denoted as *q+o*.

Memory writing operates on all spatial tokens of a completed frame and follows the gated delta update introduced in the main paper. The query/key feature map uses the Sana-ReLU kernel. The biases of the token-level update gate β and the frame-level retention gate λ are initialized to 4.0. Keys are normalized according to the spatial dimension, and memory readout uses denominator normalization with $\epsilon = 10^{-6}$. The memory readout residual and AdaLN gates are initialized to 5×10^{-3} and 10^{-2}, respectively.

A.5 Inference Details

General settings. We use EMA weights and BF16 precision for inference. Unless otherwise specified, all videos are generated at a resolution of 768×448 and 16 fps. The generator uses the two-step denoising schedule $[1000, 500]$ for regular frames, while the first chunk uses the four-step schedule $[1000, 750, 500, 250]$. Timesteps are transformed using the Wan shift parameter of 5.0, and the classifier-free guidance scale is set to 3.0.

Generation proceeds autoregressively one latent frame at a time. The local KV window is fixed to 21 latent frames, while the Delta-Memory state is propagated throughout the complete rollout. For I2V generation, the input image is encoded as one conditional latent frame, such that an output containing F latent frames requires sampling only $F-1$ noisy latent frames. The Wan VAE decodes F latent frames into $4F - 3$ RGB frames.

A.6 Ablation Settings

All structural ablations use the same initialization, training budget, and inference schedule, while changing only the memory component under study.For all ablation experiments, we use the same datasets as in the main experiments, while adopting a reduced training schedule consisting of 2K steps for Stage-2, 500 steps for Stage-2.5 redundancy-aware post-training, 750 steps for Stage-3 DMD, and 300 steps for Stage-3.5 LongLive-style DMD. All ablation variants are evaluated with a fixed generation horizon of 20 seconds.

B Related Work

Autoregressive/Diffusion Models for Video Generation. Diffusion-based models have achieved remarkable success in video generation (Yang et al., 2025d; Wan et al., 2025), yet scaling generation to substantially longer sequences remains a persistent challenge. Training-free approaches address this issue by employing techniques such as noise scheduling (Qiu et al., 2024) and positional-encoding remapping (Zhao et al., 2025), with the goal of keeping long-sequence inference close to the short-sequence distribution observed during training. A more widely adopted line of work instead equips diffusion models with causal predictive capabilities during training (Zhu et al., 2026; Huang et al., 2025; Zhao et al., 2026), and has been successfully applied to large-scale production systems. To alleviate the error accumulation inherent to autoregressive video generation, self-rollout training further exposes the model to its own generated trajectories during training (Huang et al., 2025; Liu et al., 2026; Lin et al., 2025), substantially improving inference-time rollout quality and generalization to sequences beyond the training horizon (Yang et al., 2025a; Chen et al., 2026b).

Context Learning and Compression Maintaining long-context consistency is crucial for video generation. Retrieval-based memory methods maintain a memory bank (Xiao et al., 2025) and employ a learned router to dynamically select relevant historical context for each query or generation block (Li et al., 2025; Ji et al., 2025), effectively sparsifying long-range attention (Cai et al., 2025). Orthogonal to retrieval, memory-compression approaches construct compact context representations either through distance-aware compression (Zhang et al., 2025; Gu et al., 2025) or by recurrently summarizing historical information into a fixed-size state (Savov et al., 2025).

Linear Attention and Recurrent Memory. Linear attention provides an efficient alternative to softmax attention by reformulating sequence processing with a fixed-size recurrent state, enabling efficient long-context modeling and memory-efficient autoregressive inference (Katharopoulos et al., 2020). DeltaNet introduces prediction-residual updates (Schlag et al., 2021; Yang et al., 2025c) to recurrent linear attention, while Gated DeltaNet further incorporates data-dependent state decay for adaptive memory retention (Yang et al., 2025b). More broadly, recent works improve recurrent linear attention through gating, decay, and Delta-style updates, allowing the state to be maintained more flexibly over long sequences (Yang et al., 2024). Beyond language modeling, these recurrent architectures have also been explored in visual generation (Chen et al., 2025; Yu et al., 2025), offering an efficient way to model long spatiotemporal contexts without explicitly attending to the entire history.