LEARNABLE COMPRESSED MEMORY FOR AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

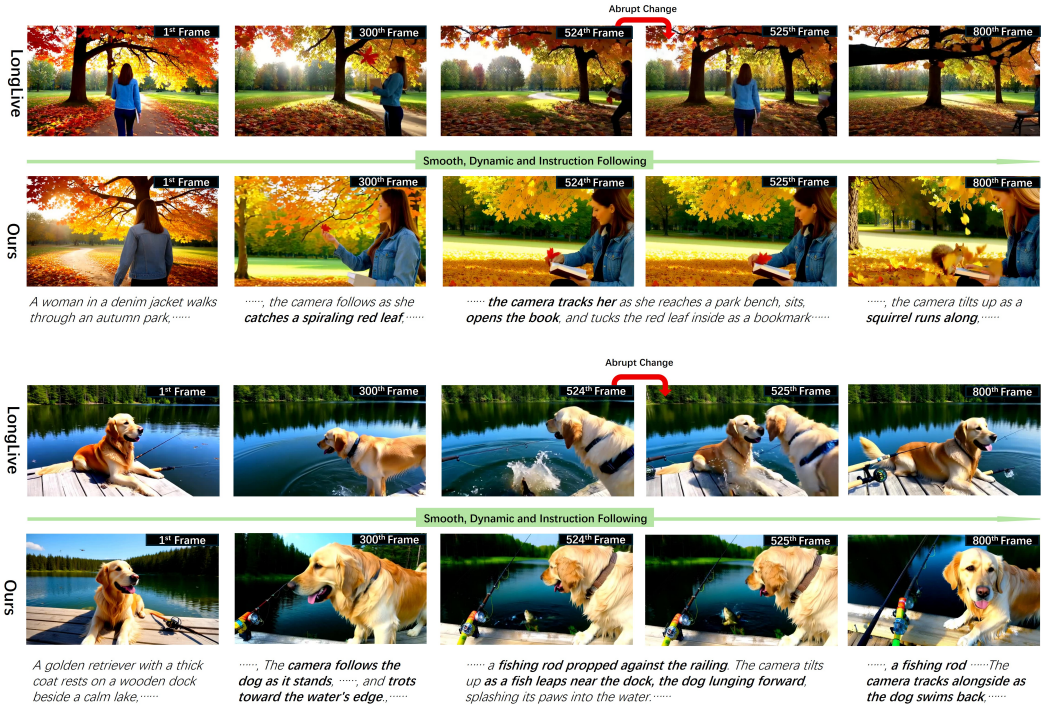

Figure 1: One-minute generations from the same prompts. LongLive (Yang et al., 2025) shows a return toward the opening appearance, while KVMEM maintains a more continuous progression. Our memory updates the long-range context as the video evolves.

ABSTRACT

Autoregressive video diffusion transformers generate videos chunk-by-chunk with causal attention over a sliding-window KV cache. While the window bounds computation and memory, it discards early context and can lead to long-horizon drift. Frame-sink variants retain a few anchor frames, but often introduce temporal discontinuities and abrupt regressions toward anchor-like appearances. The key challenge is therefore not simply to store more history, but to make more history usable within a limited context. In this paper, we present KVMEM, a learnable memory system that maintains access to a compact, evolving summary of long-range history under a fixed token budget rather than relying only on fixed frame sinks. KVMEM trains a KV Compress Adapter end-to-end with the base model and compresses history in space and time to extend the usable context. It combines local spatial summaries with learned global residuals, compressing KV tokens to represent 7× more history while maintaining attention fidelity. Compressed memories are organized in a fixed-capacity memory bank with exponential temporal binning and consolidated through content-aware eviction and merging, yielding a diverse and temporally balanced summary without hard boundaries. Experiments on long-form video generation show that KVMEM achieves better visual quality and long-range consistency, while reducing drift and sink-induced temporal artifacts.

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.

1 INTRODUCTION

Video generation Brooks et al. (2024); Wan et al. (2025); Kong et al. (2024) has witnessed remarkable progress, evolving from short-clip synthesis to longer, narrative-driven sequences. Early bidirectional diffusion-based methods Ho et al. (2020); Blattmann et al. (2023); Ma et al. (2024) generate entire videos holistically but are limited in length by GPU memory and quadratic attention costs. Autoregressive video diffusion (Yin et al., 2025; Huang et al., 2025a) enables streaming generation by producing successive chunks conditioned on a key-value (KV) cache. Maintaining coherence over a long video requires access to content far beyond the recent window. Keeping the full cache makes attention increasingly expensive, while a bounded sliding window discards earlier states and allows subject appearance and scene structure to drift.

Frame sinks retain a few opening frames as persistent anchors (Yang et al., 2025; Liu et al., 2025; Cui et al., 2025; Henschel et al., 2025; Ji et al., 2025). They stabilize early appearance but leave intervening events unrepresented. As the video evolves, the fixed references can pull generation back toward its opening state as shown in Fig. 1. Retrieval-based methods (Ji et al., 2025; Hu et al., 2026) refresh context using selected past frames. The challenge is therefore not only to preserve more history, but to make it usable within a bounded context. Our approach separates two decisions: how much spatial detail to store per chunk, and how finely to represent different parts of the timeline. Compact per-frame summaries allow more history to remain available, while progressive temporal merging keeps the total memory bounded. We learn to compress historical keys and values directly in each layer's KV space, producing compact memory summaries for the generator's attention.

In this paper, we present KVMEM, a learnable compressed KV memory system that maintains a continuously evolving summary of long-range video context under a fixed token budget. After each generation block, a lightweight KV compress adapter constructs compact spatiotemporal tokens directly in KV space, combining local pooling with a bounded, learned global correction. The compressed KV representations are inserted into a hierarchical memory bank with temporal bins at multiple granularities, preserving recent context at higher temporal resolution and older context at lower resolution. When a bin overflows, a pair is selected for merging based on a combination of content distance and normalized temporal distance. The selected pair is merged via key-norm-weighted interpolation and promoted to the next coarser bin, with in-place merging in the coarsest bin. This cascade maintains a multi-resolution summary of long-range video history within the fixed attention-token budget. Although our memory is learnable, it requires no additional manual annotations or compression-specific labels. We train the KV compress adapter with two complementary training signals: an attention-reconstruction term that encourages attention fidelity between compressed and original KV, and the standard DMD objective Yin et al. (2024b;a; 2025) used for AR diffusion training to provide task-level gradients through the compressed context. We jointly optimize the KV compression adapter and the generator's trainable adapters so that the learned compression co-adapts to the generator's attention patterns.

Our experiments show that KVMEM preserves long-range visual context while maintaining generation quality under a constrained token budget. Compared with PyramidKV Cai et al. (2024) or related KV-based method at the same long-range token budget, KVMEM retains context from the opening frame and achieves higher image quality, with better cross-clip subject and background consistency over minute-long videos. Evaluation on VBench Huang et al. (2024) further shows improvements in subject consistency and image quality over MeMFlow Ji et al. (2025) using the same backbone. Together, these findings support the value of broader historical coverage for preserving visual appearance and consistency in long-video generation.

Our main contributions are:

1. **A learnable compressed memory for autoregressive video generation.** We propose KVMEM, a structured memory system that combines learnable spatial KV compression with hierarchical temporal merging to maintain an evolving summary of long-range history within a fixed token budget. The compressed historical context is accessed through the generator's existing attention mechanism.
2. **Learnable KV compression with dual-signal training.** We design a KV compression adapter that combines spatially structured summaries with learned global refinements directly in each layer's KV space. The adapter is jointly optimized with the generator's trainable adapters through an attention-reconstruction loss for compression fidelity and a DMD loss for task-level co-adaptation.

3. **A hierarchical memory bank with content-aware merging.** We introduce a fixed-capacity memory bank organized into temporal bins at multiple granularities, coupled with a merge policy that selects redundant KV pairs based on content similarity and temporal proximity. Overflow-triggered merging progressively consolidates historical context, retaining finer temporal detail for recent history and coarser summaries for older history.

2 Related Work

2.1 Autoregressive Video Generation

Recent diffusion-based autoregressive generators adapt pretrained bidirectional models for causal, chunk-wise generation (Yin et al., 2025; Huang et al., 2025a; Zhu et al., 2026). Training on generated rollouts helps reduce exposure bias (Liu et al., 2025; Cui et al., 2025), and LongLive (Yang et al., 2025) extends this approach to interactive streaming. Sparse Forcing (Xu et al., 2026) improves efficiency through trainable sparse attention. FramePack (Zhang et al., 2026), MoC (Cai et al., 2025), and Context Forcing (Chen et al., 2026b) expand or reorganize the conditioning history. Building on these advances, our focus is the compressed KV representation that carries long-range history under a fixed attention budget.

2.2 Memory-Augmented Video Generation

Memory-augmented video generators preserve historical context through retrieval, compressed representations, or structured memory. Frame sinks retain early context, while retrieval-based methods selectively access historical frames (Yang et al., 2025; Ji et al., 2025; Hu et al., 2026). Other approaches encode historical inputs or maintain scene-level memory (Zhang et al., 2025b; Wu et al., 2025; 2026). WorldPack (Oshima et al., 2025) compresses frame context for long-context world modeling, while StoryMem (Zhang et al., 2025a) maintains memory across shots for narrative consistency. SlotMemory (Dou et al., 2026) uses object-centric slots to organize and retrieve native KV tokens, and MemRoPE (Kim et al., 2026) maintains evolving memory through training-free temporal aggregation. These methods offer different ways to organize and preserve video history. Our approach combines learned spatial KV summaries with hierarchical temporal merging, maintaining finer temporal detail for recent context and progressively coarser summaries of older history within a fixed attention budget.

2.3 KV-Cache Compression

In language models, KV-cache methods reduce memory costs through token selection (Zhang et al., 2023; Li et al., 2024), layer-wise budget allocation (Cai et al., 2024), state merging (Liu et al., 2024a), and quantization (Liu et al., 2024b; Hooper et al., 2024). StreamingLLM (Xiao et al., 2023) combines a sliding window with persistent attention sinks. For autoregressive video, PaFu-KV (Chen et al., 2026a) and PackCache (Li et al., 2026) guide token retention using salience or spatiotemporal relevance; Forcing-KV (Ji et al., 2026) and Deep Forcing (Yi et al., 2025) introduce head-dependent pruning and sink-aware retention, respectively. FAST-AR (Samuel et al., 2026) combines temporal KV-cache compression with sparse attention. Unlike token-selection policies, our adapter learns compact KV representations rather than retaining only a subset of existing tokens. Our KV compress adapter operates on video history as it is generated and is trained jointly with generator adapters using attention reconstruction and DMD supervision, aligning the compressed memory with its use in subsequent generation.

3 Method

KVmem compresses history in space and time, as shown in Fig. 2. The KV Compress Adapter introduced in Section 3.2 forms compact spatial summaries from incoming KV states, while the hierarchical memory bank described in Section 3.3 progressively coarsens the temporal granularity of older entries. The bank complements the recent sliding window with an evolving long-range memory, integrating new summaries and consolidating historical context within a fixed attention-token budget. Unlike fixed opening-frame references, this memory incorporates newly generated content throughout the video.

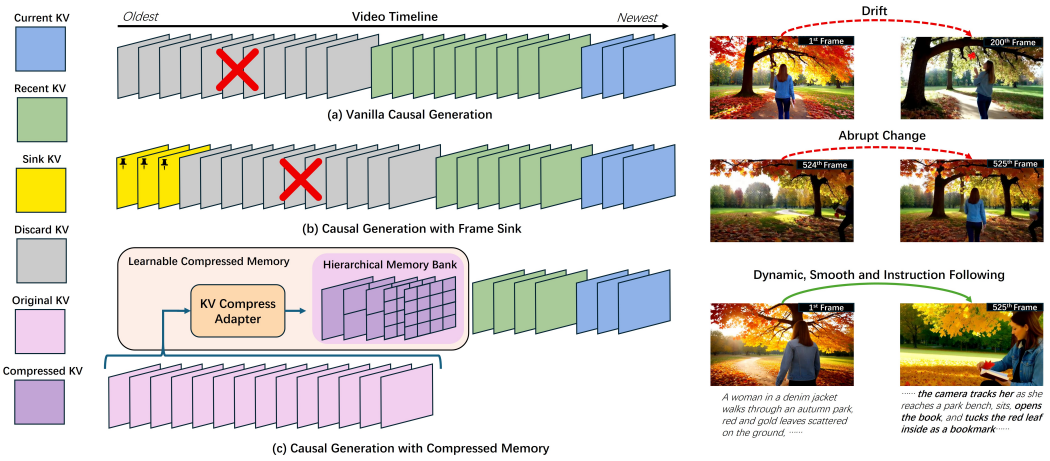

Figure 2: Long-range context in causal video generation. (a) A sliding window retains recent frames. (b) Frame sinks add fixed opening frames. (c) KVMEM maintains a compressed, evolving memory alongside the recent window.

3.1 PRELIMINARIES

A causal video diffusion transformer (Yin et al., 2025; Huang et al., 2025a) generates a chunk of C latent frames at each autoregressive step. Each frame contains $S = h \times w$ spatial tokens, processed by H attention heads of dimension d. Let t denote the first frame index of the current chunk. At layer ℓ, chunk-causal attention allows the current chunk to attend to preceding frames and all frames within the chunk. Omitting the head index and positional encoding operations for clarity, we write

$$\mathrm{Attn}_{\mathrm{causal}} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{(\ell)}_{t:t+C-1}\mathbf{K}^{(\ell)\top}_{1:t+C-1}}{\sqrt{d}}\right)\mathbf{V}^{(\ell)}_{1:t+C-1}. \tag{1}$$

To bound KV-cache growth, practical generators use a recent window, sometimes supplemented by historical context (Yang et al., 2025; Ji et al., 2025). We maintain a hierarchical bank of up to M compressed entries, each containing $S' < S$ tokens. The current chunk attends to this bank, up to R preceding frames, and its own KV states.

Let $r_t = \max(1, t - R)$. The current chunk attends to the compressed bank and recent context through

$$\mathrm{Attn}_{\mathrm{ours}} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{(\ell)}_{t:t+C-1}\left[\tilde{\mathbf{K}}^{(\ell)}_{\mathrm{bank}},\ \mathbf{K}^{(\ell)}_{r_t:t+C-1}\right]^{\top}}{\sqrt{d}}\right)\left[\tilde{\mathbf{V}}^{(\ell)}_{\mathrm{bank}},\ \mathbf{V}^{(\ell)}_{r_t:t+C-1}\right]. \tag{2}$$

Here, brackets denote token-wise concatenation, and tildes indicate compressed KV representations.

The attended-token count is bounded by $MS' + (R + C)S$. Given a fixed attention-token budget B, we allocate the compressed bank and recent context such that $MS' + (R + C)S \leq B$. Temporal merging accommodates new summaries without expanding the bank, keeping the attention-context capacity independent of video length.

3.2 LEARNABLE COMPRESSED MEMORY

KV Compress Adapter. Operating independently within each transformer layer ℓ, the KV Compress Adapter in Fig. 3 transforms the original keys and values into compact KV tokens. Its inputs are $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{S\times H\times d}$, where S is the input token count, H is the number of attention heads, and d is the per-head dimension. KVs are compressed before rotary positional encoding (RoPE), and the adapter produces $S' < S$ memory tokens while retaining the head and feature dimensions. To gather information from all input tokens, here we maintain S' learnable latent queries $\mathbf{Q}_{\mathrm{lat}} \in \mathbb{R}^{S'\times H\times d}$ and per-head projections $\mathbf{W}^{(h)}_q, \mathbf{W}^{(h)}_k \in \mathbb{R}^{d\times d}$. Following Perceiver-style cross-attention (Jaegle et al., 2021b;a), each head computes an attention map $\mathbf{A}^{(h)} \in \mathbb{R}^{S'\times S}$ and uses it to aggregate both keys

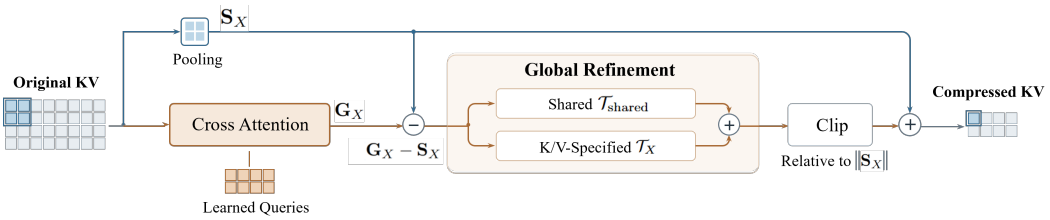

Figure 3: Illustration of our KV Compress Adapter. Local spatial summaries are refined by global context through shared and K/V-specific residual branches.

and values:

$$\mathbf{A}^{(h)} = \text{softmax}\left(\frac{(\mathbf{Q}_{\text{lat}}^{(h)}\mathbf{W}_q^{(h)})(\mathbf{K}^{(h)}\mathbf{W}_k^{(h)})^\top}{\sqrt{d}}\right), \quad \mathbf{G}_X^{(h)} = \mathbf{A}^{(h)}\mathbf{X}^{(h)}, \quad \mathbf{X} \in \{\mathbf{K}, \mathbf{V}\}. \tag{3}$$

In parallel, pooling each non-overlapping 2×2 token group yields a spatial compressed representation $\mathbf{S}_X \in \mathbb{R}^{S' \times H \times d}$. Treating $\mathbf{S}_X$ as a reference, the adapter learns a global refinement from the discrepancy $\mathbf{G}_X - \mathbf{S}_X$:

$$\widetilde{\mathbf{X}} = \mathbf{S}_X + \text{clip}_{\|\mathbf{S}_X\|}[\mathcal{T}_{\text{shared}}(\mathbf{G}_X - \mathbf{S}_X) + \mathcal{T}_X(\mathbf{G}_X - \mathbf{S}_X)]\,. \tag{4}$$

This formulation separates two roles: the spatial representation fixes where each memory token belongs, while the global refinement recovers information not captured by local pooling. $\mathcal{T}_{\text{shared}}$, shared by keys and values, sums a per-head linear projection, which corrects each head within its own subspace, and a cross-head MLP over the concatenated Hd features, which lets heads exchange information. $\mathcal{T}_K$ and $\mathcal{T}_V$ are separate low-rank MLPs that add key- and value-specific corrections, reflecting their distinct roles in attention, where keys determine relevance, while values carry the retrieved content. The operator $\text{clip}_{\|\mathbf{S}_X\|}$ bounds the correction relative to the reference, and zero-initialized branch outputs make the adapter start from $\widetilde{\mathbf{X}} = \mathbf{S}_X$. Global context therefore refines the reference rather than replacing it, keeping compressed tokens anchored to the input structure the generator expects.

Training Objective. The compressed memory should preserve past information and support future generation. We train it with two complementary signals. First, an attention-reconstruction loss encourages compressed and original KV states to produce similar attention outputs. Every k-th input key provides a parameter-free probe query, $\mathbf{Q}_{\text{probe}} = \mathbf{K}_{::k}$:

$$\mathcal{L}_{\text{recon}} = \left\|\text{Attn}\Big(\mathbf{Q}_{\text{probe}}, \tilde{\mathbf{K}}, \tilde{\mathbf{V}}\Big) - \text{Attn}(\mathbf{Q}_{\text{probe}}, \mathbf{K}, \mathbf{V})\right\|_2^2 \tag{5}$$

This loss regularizes local attention fidelity, while the DMD objective below provides task-level supervision through subsequent generation. The DMD objective (Huang et al., 2025a; Yin et al., 2024b;a) trains the compressed memory for subsequent generation. Gradients through later attention reads update newly written, unmerged entries within the training rollout, while a detached historical merge cascade bounds training memory. The combined objective is

$$\mathcal{L}_{\text{adapt}} = \lambda_{\text{recon}}\,\mathcal{L}_{\text{recon}} + \lambda_{\text{DMD}}\,\mathcal{L}_{\text{DMD}}. \tag{6}$$

3.3 Hierarchical Memory Bank

KV compression reduces the cost of each memory summary, but retaining every summary would still make storage grow with video length. The hierarchical bank in Fig. 4 integrates new summaries and progressively merges historical entries into coarser representations. This maintains an evolving long-range context within a fixed capacity. In that case, later chunks therefore read an evolving global context rather than only fixed opening-frame references.

Bank Structure and Eviction. The bank contains N bins with capacities C_i, for a total of $M = \sum_i C_i$ slots. New entries enter bin 0. Each promotion merges two summaries into one, giving bin i a temporal granularity of 2^i base chunks. The final bin has no coarser destination and therefore absorbs

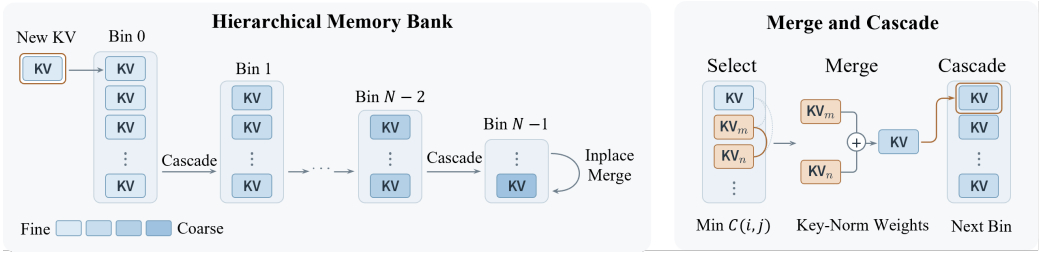

Figure 4: Hierarchical Memory Bank and its merge cascade. Left: N temporal bins with decreasing capacity and coarser granularity. Right: on overflow, the most redundant pair is merged and cascaded.

further history in place. When a bin overflows, we select the pair with the smallest content-and-time distance:

$$\mathcal{D}(i,j) = \|\mathbf{z}_i - \mathbf{z}_j\|_2 + \lambda_{\text{temp}} \cdot \frac{|t_i - t_j|}{t_{\max} - t_{\min}} \tag{7}$$

Here $\mathbf{z}_i$ is the mean compressed key over spatial tokens and heads, and t_i is the temporal index of entry i. The temporal term favors merging nearby entries, while the content term identifies redundant summaries. The pair is selected within the overflowing bin, combined using key-norm weights, and cascaded to the next bin, while overflow at the final bin is resolved by merging in place. These lightweight updates preserve a summary of the stored history within the fixed capacity. At read time, the stored summaries are concatenated with the recent window.

Positional Encoding. For the KV compress adapter, the bank stores keys before RoPE and values in their native form so that compression and merging operate independently of position. Merging entries with temporal indices t_1, t_2 and spans s_1, s_2 assigns the index $\bar{t} = (t_1 s_1 + t_2 s_2)/(s_1 + s_2)$. At read time, each token receives the spatial position of its local 2×2 cell and the temporal phase at $\bar{t}$. Unit-modulus rotations at fractional temporal positions keep these summaries compatible with the base model's positional encoding.

4 Experiments

4.1 Experimental Setup

Implementation Details We adopt Wan2.1-T2V-1.3B Wan et al. (2025) as our base generator and convert it from bidirectional to causal attention following Causal Forcing Zhu et al. (2026), with Wan2.1-T2V-14B Wan et al. (2025) as the DMD teacher. The causal generator is trained with rank-256 LoRA, and the KV compress adapter is trained with a separate optimizer via the dual-signal objective described in Section 3.2. Training proceeds in streaming autoregressive tuning following Self-Forcing Huang et al. (2025a) and LongLive Yang et al. (2025), generating up to 240 frames per iteration in chunks of 21 latent frames, which corresponds to 5s in pixel space. The memory bank uses N=4 bins with capacities $(16, 8, 8, 4)$, a $4\times$ downsampling factor via the KV compress adapter, and R=9 preceding frames, sized to exactly match the original model's attention capacity. For the prompt switching dataset, we follow LongLive Yang et al. (2025) and extend it to multi-switch dataset using GPT-4o to generate scripts of 2 to 6 sequential scene descriptions. The entire training takes approximately 2 days on 4×A100 GPUs for 3000 steps.

Evaluation We evaluate generated videos using VBench dimensions Huang et al. (2024; 2025b): *Subject Consistency*, *Background Consistency*, *Aesthetic Quality*, *Imaging Quality*, *Motion Smoothness*, and *Dynamic Degree*. For multi-switch videos, scores are averaged across prompt-aligned segments and videos, with cross-clip Subject and Background Consistency reported separately. We additionally report *Abrupt Change*, an automatic measure of large perceptual changes between consecutive frames. We compute frame-to-frame LPIPS Zhang et al. (2018) distances and count transitions whose distance exceeds a fixed threshold τ, using the same threshold for all methods. The metric is reported as the average count per video, where lower values indicate fewer abrupt perceptual changes. To complement single-prompt evaluation, we construct a multi-scene benchmark with six sequential scene descriptions per script. The scripts preserve a shared subject and setting while varying camera movements and actions, testing historical appearance retention under changing conditions. Examples are shown in Fig. 6.



Figure 5: Comparison of our KVMEM with LongLive Yang et al. (2025) and Memflow Ji et al. (2025), each tested with and without frame sink on a 1 minute multi-scene video.



Figure 6: Additional multi-scene generation results from KVMEM, demonstrating diverse environments with rich camera movements and smooth scene transitions.

4.2 Qualitative Analysis

Fig. 5 compares sampled frames from one-minute videos of a woman performing a sequence of actions in a flower shop under continuous camera motion. LongLive and MeMFlow produce plausible early frames but then abruptly snap back toward the opening appearance, as generation realigns to the fixed sink frames once the content diverges from them. Removing the sinks avoids these snap-backs but causes progressive drift in the character and environment. In contrast, KVMEM keeps the character and scene consistent while following the prompted camera movements, including pushing in, panning, and pulling back. Its evolving compressed memory adapts with the generation instead of anchoring to the opening frames. Fig. 6 shows additional multi-scene results, from dense forests to underwater reefs, where each segment follows its description with smooth transitions between scenes.

Table 1: Quantitative comparison on our 1-minute and 6-scenes multi-scene benchmark. H is the number of original latent frames represented by the attention context, and TB is the context token budget excluding the current chunk. Semantic is frame–prompt CLIP similarity and Abruptness is the LPIPS-based abrupt-change count. SF and CF denote Self-Forcing and Causal Forcing. Other metrics come from VBench. Bold marks the best value per column.

Model	H	TB	Subject(↑)	Background(↑)	Aesthetic(↑)	Imaging(↑)	Motion(↑)	Semantic(↑)	Abruptness(↓)
SnapKV Li et al. (2024)	132	23.4k	93.77	94.21	59.35	67.50	98.50	26.89	0.43
PyramidKV Cai et al. (2024)	128	23.4k	94.69	94.73	59.81	66.49	98.83	27.11	0.38
FramePack Zhang et al. (2026)	20	3.8k	93.76	95.08	60.55	63.24	98.83	26.41	0.32
Self-Forcing (SF) Huang et al. (2025a)	12	18.7k	91.23	93.94	59.18	65.74	98.77	26.45	1.92
LongLive (SF) Yang et al. (2025)	15	23.4k	91.25	**95.13**	58.17	68.12	49.72	26.04	1.75
w/o sink (SF)	12	18.7k	90.50	94.26	57.40	67.27	98.43	25.33	0.50
LongLive (CF) Yang et al. (2025)	15	23.4k	88.66	93.65	60.14	67.66	98.33	25.62	0.96
Deep Forcing Yi et al. (2025)	18	23.4k	92.66	93.61	56.51	67.04	97.86	25.63	1.42
Memflow (SF) Ji et al. (2025)	18	28.1k	92.80	95.10	60.47	68.80	98.84	26.59	1.34
w/o sink (SF)	15	23.4k	91.91	94.73	59.60	67.37	98.32	25.81	0.31
Memflow (CF) Ji et al. (2025)	18	28.1k	89.08	92.69	60.35	65.41	98.60	26.76	0.71
KVMEM	105	23.4k	**94.98**	94.79	**61.50**	**72.16**	**98.97**	**27.18**	**0.25**

Table 2: Quantitative comparison on single-prompt VBench with 1 minute per video by our own. SF and CF denote Self-Forcing and Causal Forcing. Bold marks the best value per column.

Model	Subject(↑)	Background(↑)	Aesthetic(↑)	Imaging(↑)	Motion(↑)	Semantic(↑)
SnapKV Li et al. (2024)	90.45	93.04	51.01	60.91	97.77	22.81
PyramidKV Cai et al. (2024)	93.26	94.42	53.84	62.83	98.62	23.00
LongLive (CF) Yang et al. (2025)	94.64	93.94	56.23	71.99	98.54	22.02
Memflow (SF) Ji et al. (2025)	95.61	96.13	56.42	**72.75**	98.76	**23.44**
Memflow (CF) Ji et al. (2025)	92.71	92.94	57.68	70.00	98.78	22.97
Causal Forcing (CF) Zhu et al. (2026)	93.17	94.11	57.69	66.43	97.97	21.63
Forcing-KV (CF) Ji et al. (2026)	95.63	95.77	57.23	70.71	97.37	23.20
KVMEM	**96.26**	**96.38**	**58.09**	71.92	**99.32**	23.01

4.3 Quantitative Analysis

VBench Metrics We report multi-scene and single-prompt results in Table 1 and Table 2, respectively. On the multi-scene benchmark, KVMEM achieves the best Subject Consistency, Aesthetic Quality, Imaging Quality, and Motion Smoothness, together with the fewest abrupt changes. Under the same token budget, it consistently outperforms selection-based and layer-allocation-based memory policies, indicating that learned compression retains more useful long-range context than keeping or selecting raw tokens. Baselines that persistently condition on opening frames obtain slightly higher Background Consistency, since fixed early context biases generation toward the initial scene, and this favors static backgrounds but also causes abrupt reversions when the content evolves. Removing the frame sinks reduces these reversions but leads to drift in appearance and image quality. KVMEM avoids both failure modes by maintaining an evolving summary of the full history, while preserving strong alignment with each scene prompt. Dynamic Degree measures motion amount rather than quality, and higher values can accompany drift, so a moderate value is consistent with stable long-horizon generation. On single-prompt VBench, KVMEM achieves the highest Subject Consistency, Background Consistency, Aesthetic Quality, and Motion Smoothness, showing that the compressed memory also strengthens appearance retention without prompt switching, while its Imaging Quality remains competitive with the strongest baselines.

Abruptness Abrupt changes are among the most visible failures in long-video generation, since viewers readily notice sudden reversions in appearance or layout. Table 1 reports the LPIPS-based abrupt-change count. Frame-sink baselines produce substantially more abrupt changes than KVMEM. Their fixed opening anchors compete with the evolving recent window: as the content moves away from the opening scene, attention can periodically realign to the anchors, pulling the generation back toward earlier appearance. Removing the sinks reduces these reversions but sacrifices long-range consistency and image quality, as discussed above. KVMEM yields fewer abrupt changes than both the frame-sink baselines and their sink-free variants while maintaining strong consistency. Rather than holding a frozen reference, the hierarchical memory bank continuously incorporates new content and consolidates older history, so the long-range context evolves smoothly with the video.

Throughput Throughput is governed mainly by the active attention context, since self-attention scales quadratically with the number of tokens. We evaluate all methods with Wan2.1-T2V-1.3B on a single H100. LongLive attends to 18 frames and runs at 23.7 fps. MeMFlow attends to 21 frames (12 recent, 3 sink, 3 memory-bank, and 3 current) and runs at 21.4 fps. KVMEM attends to 9 preceding frames, 3 current frames, and a compressed memory with the token count of 9 frames, for

Table 3: Ablations and additional evaluations on the multi-scene benchmark. Cons., Qual., Dyn., Sem., and Abr. denote Background Consistency, Imaging Quality, Dynamic Degree, semantic alignment, and abrupt-change count. Bins are listed from recent to older; *uses 40 slots instead of 36. $|q|$: number of latent queries; ΔQ: Overall Quality change from the reference. (c) reports the percentage of participants preferring each method. Bold marks the best value within each panel; Dynamic is not ranked.

(a) Components

Variant	H	Cons.	Qual.	Dyn.	Sem.	Abr.↓
w/o hierarchy	45	94.35	65.49	56.39	26.86	0.26
w/o adapter	105	94.46	66.74	57.50	25.42	0.33
KVMEM	105	**94.79**	**72.16**	55.73	**27.18**	**0.25**

(b) Video length

Frames	Cons.	Qual.	Dyn.	Sem.
240 (1×)	94.79	72.16	55.73	27.18
480 (2×)	94.21	69.18	53.50	27.11
720 (3×)	93.50	68.30	54.00	27.03

(c) User study

Question	KVMEM	LongLive	MeMFlow
Fewest jumps	**83%**	9%	8%
Consistency	**42%**	38%	20%
Smoothness	**58%**	28%	14%

(d) Bin allocation and sink controls

Variant	Cons.	Qual.	Dyn.	Sem.
Uniform [9, 9, 9, 9]	94.62	70.76	56.10	27.10
Reversed [4, 8, 8, 16]	94.55	70.41	54.37	27.05
Five bins* [16, 8, 8, 4, 4]	94.70	71.18	58.45	**27.22**
Sink token only	94.44	70.12	54.05	27.02
Frozen sink	94.36	70.09	53.23	26.67
KVMEM [16, 8, 8, 4]	**94.79**	**72.16**	55.73	27.18

(e) KV Compress Adapter

Variant	$\|q\|$	Heads	Layers	ΔQ
Fewer queries	8	Separate	Separate	−1.80
Fewer queries	32	Separate	Separate	−0.40
More queries	128	Separate	Separate	**+0.05**
Cross-head	64	Shared	Separate	−1.10
Shared-layer	64	Separate	Shared	−0.40
KVMEM	64	Separate	Separate	0.00

21 frame-equivalents in total. It runs at 20.6 fps, slightly slower than both baselines because of the KV Compress Adapter pass at each chunk boundary. All methods sustain near-real-time generation.

4.4 Ablation Study

Component Ablations Table 3(a) isolates the two core components. Replacing the hierarchical bank with a single flat bin of equal capacity sharply reduces the history the memory can represent, which substantially lowers image quality. Replacing the learned KV Compress Adapter with average pooling degrades every metric, most notably video quality and semantic alignment. Naive pooling blurs motion-relevant and semantic details, pushing generation toward static patterns. The full model performs best on all ranked metrics, indicating that learned compression and hierarchical organization are complementary. Within the adapter (Table 3(e)), fewer latent queries reduce quality while more queries bring negligible gains. Sharing queries across heads is more harmful than sharing them across layers, suggesting that head-specific queries matter more than query count.

Memory Organization and Scaling Allocating more slots to recent bins outperforms both uniform and reversed schedules (Table 3(d)). A fifth bin with additional capacity changes the results only marginally, so the gains come from temporal structure rather than raw capacity. Replacing the evolving memory with a sink token or a frozen sink consistently degrades performance, confirming that fixed anchors cannot substitute for continuously updated history. Performance also remains stable, with only mild degradation, when videos are extended to two and three times the default length, as shown in Table 3(b). As shown in Table 3(c), Participants most often preferred KVMEM for fewer jumps, consistency, and smoothness.

5 Conclusion

Generating temporally coherent long-form videos remains a central challenge for autoregressive video generation, as the finite KV cache forces a trade-off between memory efficiency and access to historical context. In this work, we have presented KVMEM, a learnable compressed memory system that addresses this tension through two complementary components. The KV compress adapter learns to distill each frame's KV representation into a compact fixed-size token set via perceiver-style cross-attention, trained with a dual-signal objective combining attention reconstruction and distilled generation quality. The Hierarchical Memory Bank organizes these compressed entries into an exponential temporal pyramid through content-aware eviction and cascade merging, maintaining multi-granularity coverage of the entire video history within a token budget matching the original model's attention capacity. Experiments demonstrate that KVMEM substantially outperforms frame-sink-based methods in subject and background consistency, image quality, semantic alignment, and temporal smoothness while maintaining comparable throughput. We believe learnable, structured KV compression offers a general path toward temporally coherent, extendable long-form video generation.

AI USE STATEMENT

Generative AI tools were employed to improve the clarity and readability of the manuscript and to support the search for relevant prior work. The authors carefully reviewed all AI-assisted text and verified literature suggestions against the original sources. The authors remain fully responsible for the paper's final content.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. *OpenAI Technical Report*, 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, et al. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Hanmo Chen, Chenghao Xu, Xu Yang, Xuan Chen, and Cheng Deng. Past- and future-informed kv cache policy with salience estimation in autoregressive video diffusion. *arXiv preprint arXiv:2601.21896*, 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.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

Weijia Dou, Hui Li, Jiahao Cui, Lei Zhou, Jingdong Wang, and Siyu Zhu. Slotmemory: Object-centric kv memory for streaming long-video generation. *arXiv preprint arXiv:2605.31033*, 2026.

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, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. In *Advances in Neural Information Processing Systems*, 2024.

Qixin Hu, Shuai Yang, Wei Huang, Song Han, and Yukang Chen. Longlive-rag: A general retrieval-augmented framework for long video generation. *arXiv preprint arXiv:2606.02553*, 2026.

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*, 2025a.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021a.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021b.

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.

Yicheng Ji, Zhizhou Zhong, Jun Zhang, Qin Yang, Xitai Jin, Ying Qin, Wenhan Luo, Shuiyang Mao, Wei Liu, and Huan Li. Forcing-kv: Hybrid kv cache compression for efficient autoregressive video diffusion models. *arXiv preprint arXiv:2605.09681*, 2026.

Youngrae Kim, Qixin Hu, C-C Jay Kuo, and Peter Beerel. Memrope: Training-free infinite video generation via evolving memory tokens. In *European Conference on Computer Vision*, pp. 244–261. Springer, 2026.

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

Kunyang Li, Mubarak Shah, and Yuzhang Shang. Packcache: A training-free acceleration method for unified autoregressive video generation via compact kv-cache. *arXiv preprint arXiv:2601.04359*, 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. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Akide Liu, Jing Liu, Zizheng Pan, Yefei He, Gholamreza Haffari, and Bohan Zhuang. Minicache: Kv cache compression in depth dimension for large language models. *Advances in Neural Information Processing Systems*, 37:139997–140031, 2024a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *International Conference on Machine Learning*, 2024b.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Worldpack: Compressed memory improves spatial consistency in video world modeling. *arXiv preprint arXiv:2512.02473*, 2025.

Dvir Samuel, Issar Tzachor, Matan Levy, Michael Green, Gal Chechik, and Rami Ben-Ari. Fast autoregressive video diffusion and world models with temporal cache compression and sparse attention. *arXiv preprint arXiv:2602.01801*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Mingqiang Wu, Weilun Feng, Zhefeng Zhang, Haotong Qin, Yuqi Li, Guoxin Fan, Xiaokun Liu, Zhulin An, Libo Huang, Yongjun Xu, et al. Echo-forcing: A scene memory framework for interactive long video generation. *arXiv preprint arXiv:2605.16003*, 2026.

Xiaofei Wu, Guozhen Zhang, Zhiyong Xu, Yuan Zhou, Qinglin Lu, and Xuming He. Pack and force your memory: Long-form and consistent video generation. *arXiv preprint arXiv:2510.01784*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Boxun Xu, Yuming Du, Zichang Liu, Siyu Yang, Ziyang Jiang, Siqi Yan, Rajasi Saha, Albert Pumarola, Wenchen Wang, and Peng Li. Sparse forcing: Native trainable sparse attention for real-time autoregressive diffusion video generation. *arXiv preprint arXiv:2604.21221*, 2026.

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

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, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

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

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. Storymem: Multi-shot long video storytelling with memory. *arXiv preprint arXiv:2512.19539*, 2025a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. Pretraining frame preservation in autoregressive video memory compression. *arXiv preprint arXiv:2512.23851*, 2025b.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *Advances in Neural Information Processing Systems*, 38:30546–30566, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 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.

A HYPERPARAMETERS SETUP

We provide a comprehensive listing of all hyperparameters introduced by KVMEM in Table 4. The base video generation backbone follows Self-Forcing and its DMD distillation setup; we only list the additional hyperparameters specific to our method. All experiments are conducted at 480×832 resolution on Wan2.1-T2V-1.3B. The adapter receives gradients from both $\mathcal{L}_{\text{recon}}$ (direct, via attention reconstruction on all compressed entries) and $\mathcal{L}_{\text{DMD}}$ (indirect, through the attention mechanism for intra-chunk compressed KVs).

B VISUALIZATION AND ANALYSIS

To understand the internal compression behavior of KVMEM, we instrument a full 240-frame generation and capture attention maps from the KV compress adapter and memory bank states from the hierarchical memory bankat every keyframe extraction step. All visualizations use the same prompt sequence on Wan2.1-T2V-1.3B.

B.1 TOKEN RECEPTIVE FIELDS ACROSS LAYERS

Fig. 7 compares the adapter's cross-attention patterns at three representative depths. At Layer 0, each compressed token spreads attention broadly across the 30×52 input grid with a mild horizontal banding bias, indicating a soft average pooling regime. At Layer 14, receptive fields tighten and begin to respect object boundaries—the person's silhouette and flower regions create distinguishable "shadows" in neighboring token maps. At Layer 27, a qualitative transition occurs: each token's attention becomes sharply localized with high contrast, closely resembling a semantic segmentation mask. Tokens over the foreground subject trace the human contour, tokens over the flower bouquet confine their attention to the floral region, and background tokens attend only to the wall. This depth-dependent transition—from global averaging to content-adaptive spatial selection—emerges end-to-end from $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{DMD}}$ without any explicit architectural bias, forming a hierarchical compression strategy where shallow layers preserve global context and deep layers retain spatially precise detail.

B.2 TEMPORAL CONSISTENCY OF RECEPTIVE FIELDS

Fig. 8 tracks Layer 27 receptive fields across early, mid, and late frames. The overall spatial tiling—which grid positions attend to which image regions—remains remarkably consistent across 120 frames, confirming that the learned queries $\mathbf{Q}_{\text{lat}}$ encode robust positional priors. Simultaneously, the fine-grained attention within each tile adapts to track scene content: as flowers transition from pink/white tulips to yellow chrysanthemums, tokens previously focused on pink regions smoothly redistribute toward the new floral content. This dual behavior—stable spatial structure with content-adaptive routing—explains how the adapter handles multi-prompt scene changes without requiring re-initialization.

B.3 PER-HEAD ATTENTION SPECIALIZATION

Fig. 9 reveals a striking division of labor among the 12 attention heads that progressively sharpens with depth. At Layer 0, all head types are already recognizable—horizontal-band heads (H4, H6), vertical-stripe heads (H7, H11), and content-driven heads (H0, H2, H5, H8)—but with low contrast, consistent with near-maximum entropy. At Layer 14, positional stripe patterns soften and more heads become content-influenced, with object silhouettes visibly warping the attention bands. At Layer 27, the factorization reaches its sharpest form:

- *Horizontal-band heads* (H0, H1, H6) encode the vertical coordinate through row-aligned stripes, consistent across all query positions.
- *Vertical-stripe heads* (H7, H10, H11) encode the horizontal coordinate through column-aligned stripes.
- *Content/semantic heads* (H2, H4, H5, H8) directly reflect scene structure—flower bouquets, human silhouettes, and background regions appear as coherent patterns.

Table 4: Hyperparameters of KVMEM.

Hyperparameter	Value
KV Compress Adapter	
Number of adapters	30 (one per transformer layer)
Input tokens per frame S	1,560 (30×52)
Output tokens per frame S'	390 (15×26)
Spatial pool size p	2 (2×2)
Local pooling	Key-norm weighted; shared weights for K/V
Number of global latent queries	390 per head
Global Q/K projections (per head)	$\mathbf{W}_q^{(h)}, \mathbf{W}_k^{(h)} \in \mathbb{R}^{128\times128}$
Global attention weights	Shared between K and V
Shared output projection (per head)	$\mathbf{W}_o^{(h)} \in \mathbb{R}^{128\times128}$
Shared FFN hidden dimension	768 (cross-head)
Separate K/V branches	Two low-rank MLPs; rank 128 each
FFN activation / normalization	SiLU / RMSNorm
Global refinement input	Global summary minus local summary
Residual norm bound	$0.5\times$ local-summary norm, per token/head
Output initialization	Zero for shared and K/V branch outputs
Initial adapter output	Local pooled summary
Query initialization	Truncated normal ($\sigma = 0.02$)
Total parameters	130.2M ($4.34\text{M} \times 30$)
Hierarchical Memory Bank	
Number of bins N	4
Bin capacities $[C_0, C_1, C_2, C_3]$	$[16, 8, 8, 4]$
Total memory slots M	36
Bin temporal spans	$[2^0, 2^1, 2^2, 2^3]$
Tokens per entry S'	390
Extraction interval	3 (one per chunk)
Extracted frame position	Middle frame
Merge cost λ_{temp}	0.1
Per-layer independent merge	Yes
Token Budget	
Memory bank	$36 \times 390 = 14{,}040$
Recent window R	$9 \times 1{,}560 = 14{,}040$
Current chunk C	$3 \times 1{,}560 = 4{,}680$
Total	**32,760** (= base 21-frame capacity)
Training	
Optimizer (all components)	AdamW (β_1=0, β_2=0.999, wd=0.01)
Generator / Adapter learning rate	1×10^{-5}
Critic learning rate	2×10^{-6}
Max gradient norm	10.0
Total iterations	3,000
Streaming chunk size	21 frames
Maximum sequence length	240 frames
λ_{recon} / λ_{DMD}	1.00 / 1.00
EMA decay / start step	0.99 / 200

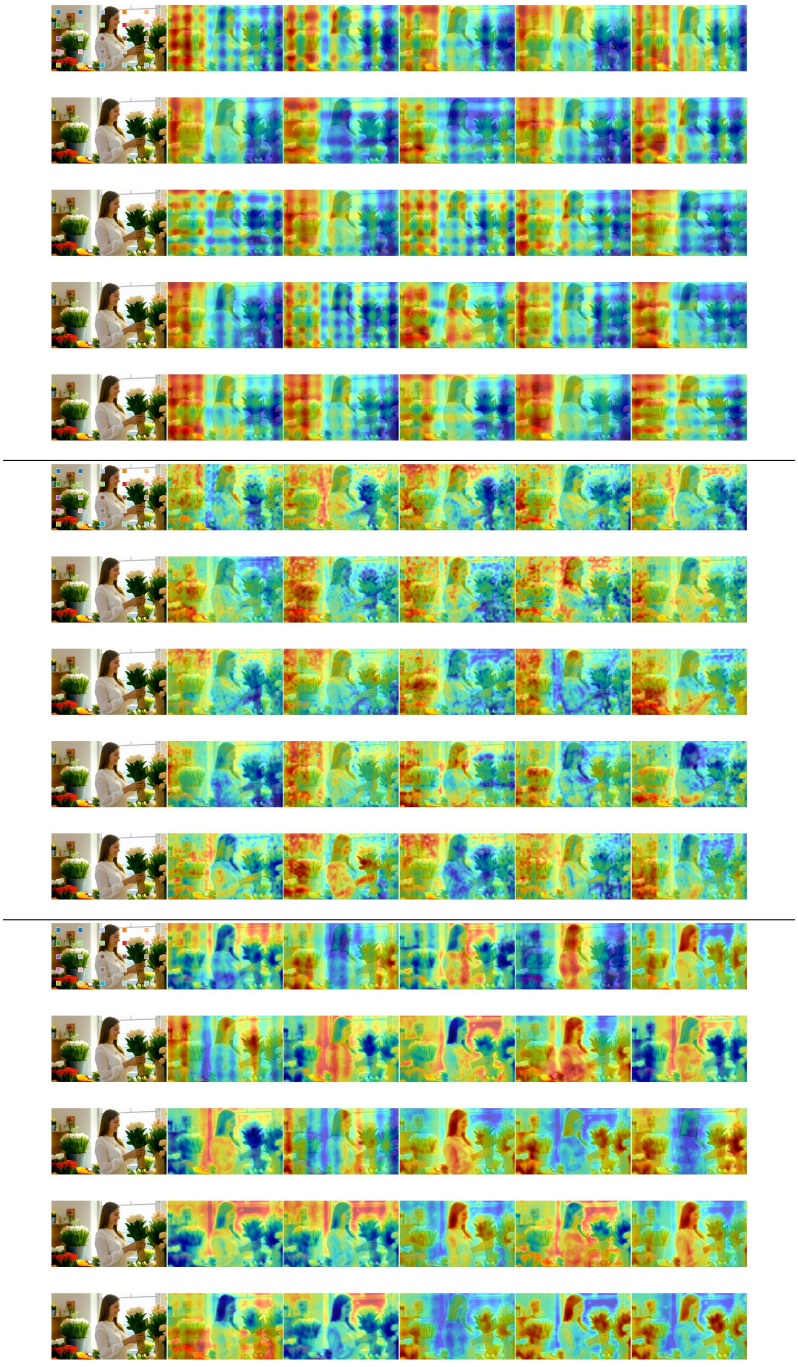

Figure 7: Token receptive fields at Frame 61 for Layers 0 (top), 14 (middle), and 27 (bottom). Each panel shows 25 sampled compressed tokens on the 15×26 output grid; the heatmaps are head-averaged cross-attention weights $\mathbf{A}^{(h)}$ overlaid on the decoded frame. Shallow layers produce diffuse, near-uniform attention, while deeper layers exhibit sharply localized, content-adaptive receptive fields.

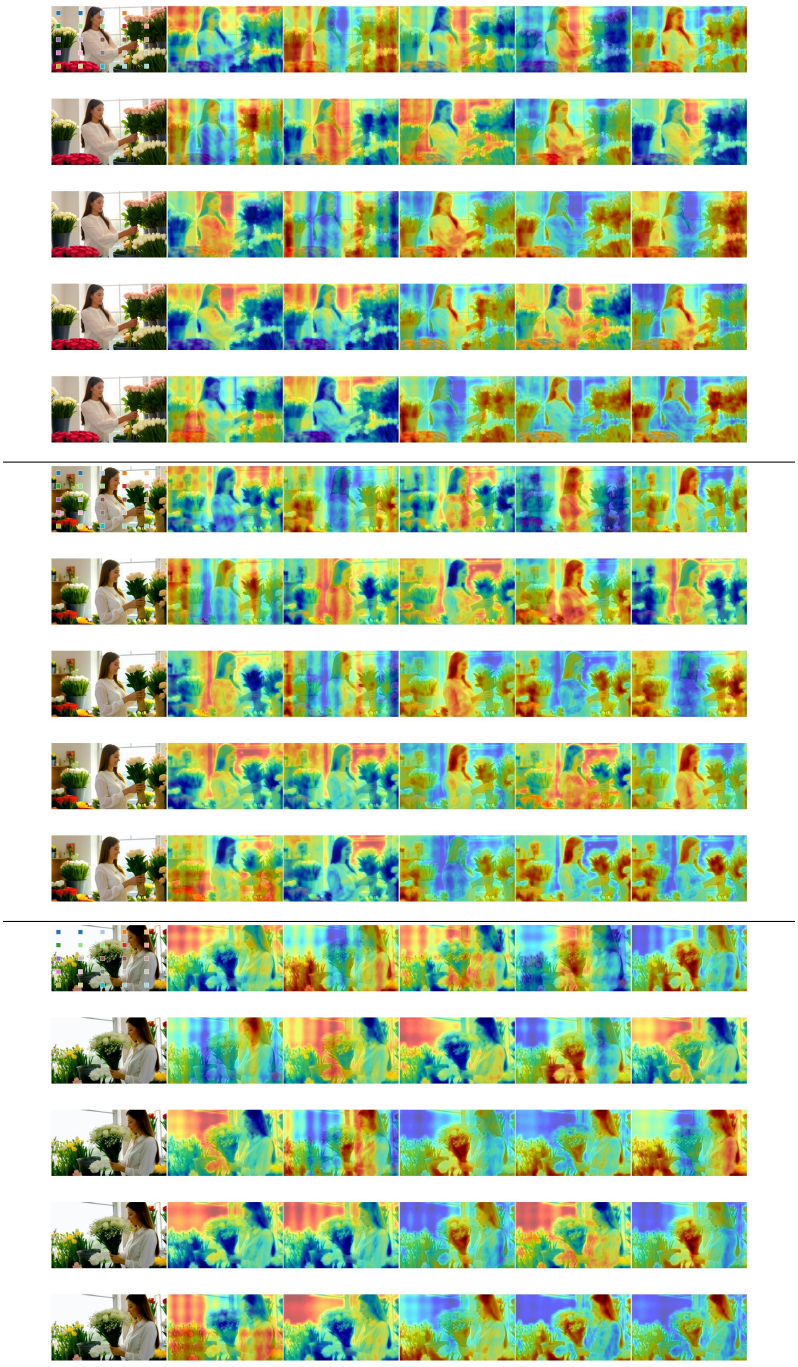

Figure 8: Token receptive fields at Layer 27 across three generation stages: Frame 4 (top, early), Frame 61 (middle, mid), and Frame 121 (bottom, late). Despite substantial scene changes (flower type, color, camera angle), the spatial structure of each token's receptive field remains stable while its fine-grained routing adapts to the current content.

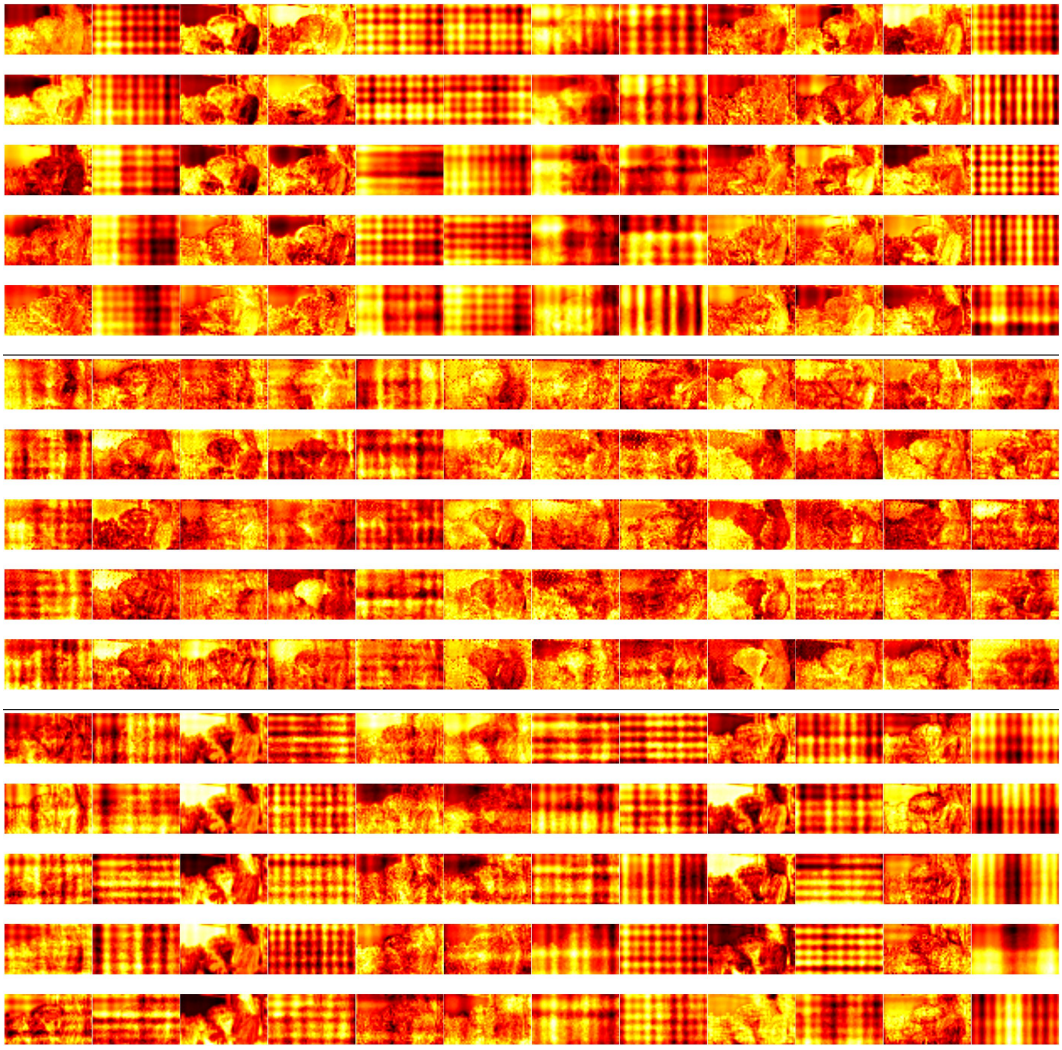

Figure 9: Per-head attention patterns at Frame 121 for Layers 0 (top), 14 (middle), and 27 (bottom). Rows correspond to 5 sampled token positions; columns to the 12 attention heads (H0–H11). Deeper layers develop sharply factored specialization: horizontal-band heads, vertical-stripe heads, and content/semantic heads emerge as distinct functional groups.



Figure 10: Hierarchical memory bank state at seven time points during a 240-frame generation. Each horizontal bar represents one compressed keyframe entry; its position indicates the source frame index and its width reflects the temporal span. Bin 0 (blue, span 1) stores recent keyframes at full resolution; deeper bins hold progressively coarser merged summaries.

Together, the positional heads provide a factorized 2D coordinate system while the content heads capture *what* is present at each location. The cross-head FFN in the adapter then fuses these complementary signals into a coherent compressed representation. This specialization is not architecturally enforced but emerges from the dual-signal training objective, where the pressure for complementary information capture intensifies at deeper layers whose key representations are more semantically structured.

B.4 Hierarchical Memory Bank Evolution

Fig. 10 traces the bank's evolution from a single entry at Frame 1 to a fully stratified pyramid at Frame 238. In the early phase ($\lesssim$Frame 40), Bin 0 fills linearly as a simple FIFO buffer. Once Bin 0 reaches its capacity of 16 entries, the merge-and-cascade mechanism activates: the most redundant pair is merged and promoted to Bin 1 (span 2). This cascade progressively populates deeper bins—Bin 2 (span 4) appears by Frame 118 and Bin 3 (span 8) by Frame 196—forming the designed exponential pyramid within bounded memory.