COEFFICIENT-CALIBRATED MEMORY RETRIEVAL VIA SPARSE CODING FOR LONG VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video generation models use fixed context windows to make long rollouts tractable, but progressively lose early information as earlier frames are evicted from the context. Existing retrieval methods retrieve the top-K historical latent entries from a memory bank by similarity, improving subject and background consistency. However, independent top-K selection may retrieve near-duplicate historical latents whose accumulated evidence causes *memory over-anchoring*, suppressing motion evolution and ultimately leading to motion collapse. We propose a training-free memory interface to address this failure, namely the Coefficient-calibrated Atom REtrieval of MEMory (CARE-MEM). We map historical latent entries in the memory bank to memory atoms that form an online dictionary. Our approach adopts constrained sparse coding to jointly select a fixed-budget set of memory atoms and applies the resulting coefficients to reweight the attention evidence of the corresponding memory entries before normalization. Across three autoregressive backbones and 30-, 60-, and 120-second generation under VBench-Long, the proposed method achieves the best or second-best performance in 47 of 54 metric–setting comparisons. Compared with the top-K retrieval baseline, our method improves Dynamic Degree by 3.48 points on average while maintaining comparable Motion Smoothness, with additional gains in quality and consistency.

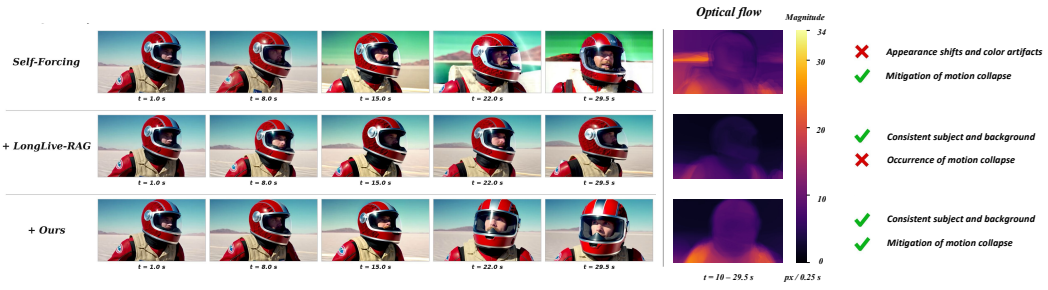


Figure 1: Self-Forcing (Huang et al., 2025) suffers from appearance shifts and color artifacts. To mitigate this, LongLive-RAG (Hu et al., 2026) adaptively retrieves historical states based on cosine similarity. While LongLive-RAG (Hu et al., 2026) improves subject and background consistency, its top-K retrieval may repeatedly select near-duplicate memories, causing *memory over-anchoring* that restricts continued motion evolution which leads to motion collapse. Our method alleviates this failure mode, preserving consistent subject and background appearance while maintaining motion evolution.

1 INTRODUCTION

Autoregressive video diffusion models generate long videos with fixed context windows, making long rollouts computationally practical but progressively discarding early information about subject identity and scene structure as earlier frames are evicted from the context (Huang et al., 2025; Yang et al., 2026; Zhu et al., 2026). Recent methods preserve distant history through persistent tokens, compression, or retrieval (Yesiltepe et al., 2025; Yi et al., 2025; Kim et al., 2026; Hu et al., 2026). In particular, LongLive-RAG (Hu et al., 2026) retrieves a fixed number of previously generated latent blocks from a memory bank without modifying the pretrained generator or causal context.

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.

Existing retrieval interfaces (Hu et al., 2026) typically rank historical latents independently and apply no memory-specific reweighting. Consequently, top-K retrieval may spend limited memory slots on near-duplicate states, whose repeated evidence pulls generation toward previously observed configurations. As illustrated in Figure 1, retrieval can improve long-term consistency while restricting continued motion evolution, potentially leading to motion collapse. We call this failure *memory over-anchoring*: redundant similar selection wastes the fixed memory budget, while uniform attention access strength amplifies repeated historical evidence.

To alleviate *memory over-anchoring* of the retrieval for long video generation, we introduce Coefficient-calibrated Atom REtrieval of MEMory (CARE-MEM), a training-free memory interface that jointly determines what to retrieve and how strongly each retrieved state should participate in attention. We view historical latent entries as a memory bank (Hu et al., 2026). A frozen retrieval encoder (Hu et al., 2026) maps each entry to a memory atom, yielding an online dictionary, while the latest latent is encoded as the query descriptor. We propose *constrained sparse coding optimization for memory selection*, which reconstructs the query from a fixed-budget set of memory atoms, while the atom-correlation constraint filter prevents near-duplicate atoms from occupying memory slots. We further introduce *coefficient-calibrated priors for attention control*, using the sparse coding coefficients as memory-level priors on attention evidence. Native query-key scores continue to determine token compatibility, while the coefficients control memory-specific access strength.

We evaluate the proposed method on Self-Forcing, LongLive, and Causal-Forcing for 30-, 60-, and 120-second generation under VBench-Long (Huang et al., 2024; Yang et al., 2026; Huang et al., 2025; Zhu et al., 2026). Across nine backbone–duration settings, our approach is best or second-best in 47 of 54 metric–setting comparisons. Relative to LongLive-RAG (Hu et al., 2026), dynamic degree improves by 3.48 points on average, while motion smoothness remains nearly unchanged. The Moving-MNIST example, selected-memory diversity analysis, and component ablation further support the complementary roles of memory selection and memory access strength.

In summary, our contributions are:

- We identify *memory over-anchoring* as a limitation of existing retrieval: top-K retrieval can waste a fixed budget on near-duplicate histories without memory-specific attention control.
- Based on a dictionary view of memory, we introduce *constrained sparse coding optimization for memory selection* to jointly retrieve complementary memory atoms under a fixed budget.
- We introduce *coefficient-calibrated priors for attention control*, using the terminal sparse coefficients to determine how strongly each selected memory participates in attention.

2 METHODOLOGY

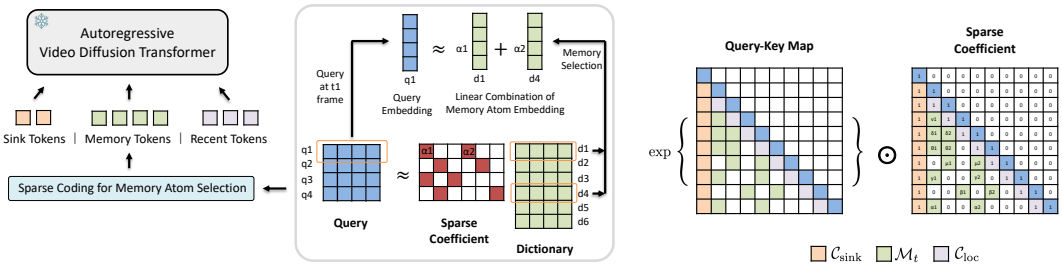
2.1 A DICTIONARY VIEW OF MEMORY BANK

At autoregressive block t, the latest completed clean latent $\mathbf{x}_t$ is mapped by the frozen LongLive-RAG retrieval encoder E_ψ to an ℓ_2-normalized query descriptor $\mathbf{q}_t = E_\psi(\mathbf{x}_t)/\|E_\psi(\mathbf{x}_t)\|_2 \in \mathbb{R}^d$ (Hu et al., 2026). Each eligible historical clean latent $\mathbf{x}_i$ is similarly mapped to $\mathbf{d}_i = E_\psi(\mathbf{x}_i)/\|E_\psi(\mathbf{x}_i)\|_2$, which we define as a memory atom. The memory atoms therefore define an online dictionary

$$\mathbf{D}_t = [\mathbf{d}_1, \ldots, \mathbf{d}_{N_t}] \in \mathbb{R}^{d \times N_t}. \tag{1}$$

This view separates the search space from the generator context. Columns of $\mathbf{D}_t$ are compact memory atoms used to explain the current state; after an atom is selected, CARE-MEM fetches its native latent block and places it in the unchanged attention context. The dictionary grows with the rollout, but the active explanation remains bounded by the limited historical slots. Thus the memory problem becomes a fixed-budget sparse approximation problem over generated history rather than a collection of independent similarity decisions. As illustrated in Figure 2a, the query descriptor is not matched against past frames independently; instead, it is reconstructed by a small set of complementary atoms whose native latent blocks are then fetched into the fixed memory interface.

The dictionary is not trained or updated by gradient descent. It is instantiated online from the same frozen descriptor space used by LongLive-RAG (Hu et al., 2026). Retrieval runs once per generated block and its selected latent blocks are reused throughout that block's denoising steps.



(a) Sparse coding optimization for memory selection. **(b)** Coefficient priors for attention control.

Figure 2: CARE-MEM replaces independent top-K retrieval with a joint sparse explanation. **(a)** The query descriptor is reconstructed from an online dictionary of historical memory atoms through the constrained sparse coding, which reduces the chance that several near-duplicate states occupy separate slots and repeatedly anchor the rollout. **(b)** The corresponding coefficients reweight only the retrieved-memory region of the original attention evidence. The symbol $\odot$ denotes the Hadamard product, i.e., element-wise multiplication.

2.2 Constrained Sparse Coding Optimization for Memory Selection

Inspired by the *minimum-redundancy maximum-relevance* (mRMR) principle (Peng et al., 2005), which maximizes relevance to the target while minimizing redundancy among selected features, we formulate memory retrieval as an analogous reconstruction–diversity problem. Specifically, we aim to *minimize the reconstruction error of the current query while maximizing the diversity among the selected memory atoms*. We denote by K the memory budget and by τ the cosine-similarity threshold that prevents overly similar memory atoms from coexisting. We then seek memory atoms which can jointly reconstruct the query descriptor $\mathbf{q}_t$ through a constrained reconstruction objective:

$$\boldsymbol{\alpha}_t^\star \in \underset{\boldsymbol{\alpha}\in\mathcal{A}_{K,\tau}(\mathbf{D}_t)}{\arg\min}\ f_t(\boldsymbol{\alpha}), \qquad f_t(\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{q}_t - \mathbf{D}_t\boldsymbol{\alpha}\|_2^2, \tag{2}$$

where $\mathcal{A}_{K,\tau}(\mathbf{D}_t)$ is a set:

$$\mathcal{A}_{K,\tau}(\mathbf{D}_t) = \left\{\boldsymbol{\alpha} \geq \mathbf{0} : \|\boldsymbol{\alpha}\|_0 \leq K,\ \mathbf{d}_i^\top \mathbf{d}_j < \tau,\ \forall\, i,j \in \mathrm{supp}(\boldsymbol{\alpha}),\ i \neq j\right\}. \tag{3}$$

The reconstruction objective enforces query relevance, whereas the correlation threshold τ controls inter-atom redundancy. As shown in Proposition 1, an appropriate τ can preserve reconstruction quality relative to top-K similarity retrieval while maximizing the guaranteed diversity lower bound within the admissible range. We approximately optimize Equation (2) with an IHT-style (Blumensath & Davies, 2009) gradient step followed by the implemented greedy filter as Equation (3). Let h denote the iteration index and η the step size. Starting from $\boldsymbol{\alpha}^{(0)} = \mathbf{0}$, each iteration forms

$$\mathbf{z}^{(h+1)} = \left[\boldsymbol{\alpha}^{(h)} + \eta\mathbf{D}_t^\top(\mathbf{q}_t - \mathbf{D}_t\boldsymbol{\alpha}^{(h)})\right]_+, \qquad \boldsymbol{\alpha}^{(h+1)} = \mathcal{P}_{K,\tau}^{\mathrm{greedy}}(\mathbf{z}^{(h+1)}; \mathbf{D}_t), \tag{4}$$

$$\mathcal{P}_{K,\tau}^{\mathrm{greedy}}(\mathbf{z}; \mathbf{D}_t) = \mathbf{z} \odot \mathbf{1}_{S(\mathbf{z})}, \qquad S(\mathbf{z}) = \mathrm{Greedy}_{\mathbf{z}\downarrow}\left\{S :\ \mathbf{z} \odot \mathbf{1}_S \in \mathcal{A}_{K,\tau}(\mathbf{D}_t)\right\}, \tag{5}$$

where $\mathrm{Greedy}_{\mathbf{z}\downarrow}$ scans positive coordinates in descending z_i and retains an index only if feasibility in $\mathcal{A}_{K,\tau}(\mathbf{D}_t)$ is preserved. Thus Equation (4) is an IHT-style update (Blumensath & Davies, 2009) with a greedy feasibility filter. After Ω iterations, we set $\widehat{\boldsymbol{\alpha}}_t = \boldsymbol{\alpha}^{(\Omega)}$, prune zero coordinates, and use the retained coefficients directly. The support $S_t = \mathrm{supp}(\widehat{\boldsymbol{\alpha}}_t)$ determines which historical entries are retrieved, while the coefficients quantify their contributions to the joint explanation.

2.3 Coefficient-Calibrated Priors for Attention Control

After constrained sparse coding jointly selects a fixed-budget set of complementary memory atoms, we directly use the resulting coefficients to reweight the attention evidence of their corresponding memory entries before normalization, yielding coefficient-calibrated priors for attention control. The support S_t from Equation (4) determines which historical blocks enter the context, while $\widehat{\alpha}_{t,j}$

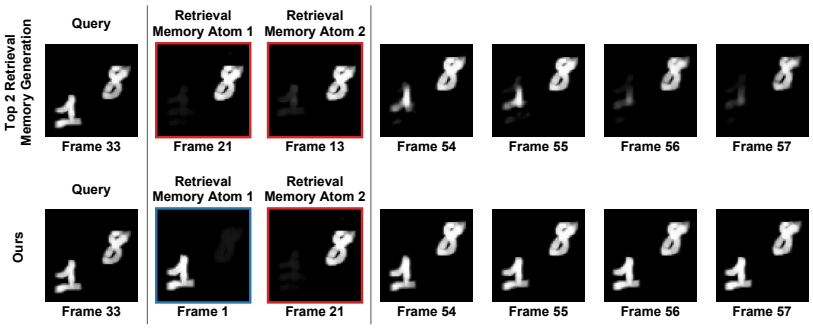


Figure 3: Toy example visualization. Top-2 retrieves a redundant pair (frames 21 and 13; cosine similarity 0.92), while CARE-MEM retrieves complementary memories (frames 1 and 21; cosine similarity 0.25). With the same generator, noise, and prefix, CARE-MEM avoids the over-anchoring caused by duplicated memory evidence; see Section B for details.

controls the access strength of memory $j \in S_t$. For one attention head at block t, let n index a current query token, $\mathcal{M}_{t,j}$ the tokens of memory entry j, and $\mathcal{F}_t$ other context tokens. The visible context is $\mathcal{C}_t = \mathcal{F}_t \cup_{j \in S_t} \mathcal{M}_{t,j}$.

To distinguish attention variables from the retrieval descriptor $\mathbf{q}_t$, we use $\mathbf{q}_{t,n}$, $\mathbf{k}_{t,m}$, and $\mathbf{v}_{t,m}$. For each $m \in \mathcal{C}_t$, define

$$\ell_{t,n,m} = \frac{(\mathbf{q}_{t,n})^\top \mathbf{k}_{t,m}}{\sqrt{d_h}}, \qquad \pi_{t,m} = \begin{cases} \widehat{\alpha}_{t,j}, & m \in \mathcal{M}_{t,j}, \\ 1, & m \in \mathcal{F}_t, \end{cases} \tag{6}$$

where d_h is the head dimension. Thus, all tokens of memory j share its sparse coefficient, while non-memory tokens retain unit prior. $\widehat{\alpha}_{t,j}$ determines how strongly memory j participates in attention.

CARE-MEM reweights only the unnormalized attention evidence:

$$\widetilde{e}_{t,n,m} = \pi_{t,m} \exp(\ell_{t,n,m}), \qquad p_{t,n,m} = \frac{\widetilde{e}_{t,n,m}}{\sum_{m' \in \mathcal{C}_t} \widetilde{e}_{t,n,m'}}, \qquad \mathbf{o}_{t,n} = \sum_{m \in \mathcal{C}_t} p_{t,n,m} \mathbf{v}_{t,m}. \tag{7}$$

This is equivalent to a Hadamard product between the native unnormalized attention map and the broadcast coefficient priors. The coefficients from Section 2.2 naturally guide memory contributions rather than merely select the support. Proposition 2 bounds how far native attention evidence can shift the resulting memory contribution distribution away from the coefficient-derived prior.

3 Analysis

We first illustrate complementary memory, then analyze the roles of residual fitting, nonredundancy, and fitted coefficients in attention. Complete proofs are in Section A.

3.1 Toy Example

Setting. We use a Moving-MNIST toy dataset to visualize retrieval redundancy based on a very small latent-video diffusion model. We first pretrain the toy VAE and a bidirectional full-video diffusion teacher on a 64-frame distribution with a deterministic visibility schedule: frames 0–7 emphasize digit 1, frames 8–27 emphasize digit 8, and frames 28–63 are for both digits. We then causally adapt a student from the pretrained teacher, followed by Self-Forcing DMD fine-tuning over self-generated rollouts. The two-slot retrieval setting is introduced only at inference to test whether this learned joint state remains available during the subsequent causal rollout; more details are deferred to Section B. For the Retrieval comparison, Top-2 Retrieval uses cosine similarity, while CARE-MEM uses 20 optimization steps with $\tau = 0.7$, matching the VBench-Long experiments.

Observations. From Figure 3, at query frame 33, top-2 retrieval selects frames 21 and 13 with cosine similarity 0.92. So both selected memories are "8" digits and highly redundant. CARE-MEM

with $\tau = 0.7$ selects frames 1 and 21, with cosine similarity 0.25, yielding complementary "1" digits and "8" digits memories. In Figure 3, the redundant pair over-anchors the rollout to digit "8", whereas CARE-MEM preserves both digits by using the released slot to recover the missing query component.

3.2 Reconstruction and Memory Atom Diversity

Section 2.2 minimizes query reconstruction error while excluding overly correlated atoms. It therefore seeks memories that explain the query together without repeating the same information. Figure 3 illustrates this: a "1" complements the retained "8", unlike another "8". We compare $\mathcal{E}_B^\star = \min_{\mathbf{a}\geq\mathbf{0}} \frac{1}{2}\|\mathbf{q} - \mathbf{D}_B\mathbf{a}\|_2^2$, the best nonnegative pair fits.

Two-memory setting. For $K = 2$ with unit-normalized descriptors, we compare the top-K similarity-based retrieval pair $T = \{\mathbf{u}, \mathbf{v}\}$ with the pair $S = \{\mathbf{u}, \mathbf{w}\}$ selected by CARE-MEM. Write $s = \mathbf{u}^\top\mathbf{q} \geq b = \mathbf{v}^\top\mathbf{q} \geq \xi = \mathbf{w}^\top\mathbf{q}$, $c = \mathbf{u}^\top\mathbf{w} \geq 0$, and $\rho = \mathbf{u}^\top\mathbf{v}$. For $0 < \tau < 1$, feasibility of S gives $c < \tau$; we consider the redundant case $\tau \leq \rho$, where T violates the correlation constraint. Assume $\xi > cs$, so $\mathbf{w}$ explains part of $\mathbf{q} - s\mathbf{u}$. Define $g = (\xi - cs)/\sqrt{1-c^2}$ and $\tau_0 = (sb - g^2)/(s^2 + g^2)$; $g^2/2$ is the added reconstruction gain. We measure memory diversity by the von Neumann entropy $\mathsf{H}(B)$ of $\mathbf{D}_B^\top\mathbf{D}_B/2$ (Watrous, 2018); higher entropy means less concentration in one descriptor direction. Let h_2 denote binary entropy in nats.

Proposition 1 (Reconstruction with a diversity constraint)**.** *Under this setting, if $\tau \leq b/s$, then*

$$\mathcal{E}_T^\star - \mathcal{E}_S^\star \geq \frac{(s^2+g^2)(\tau-\tau_0)}{2(1+\tau)}, \qquad \mathsf{H}(S) > h_2\left(\frac{1+\tau}{2}\right) \geq \mathsf{H}(T). \tag{8}$$

Remark 1. Within the stated range, $\tau \geq \tau_0$ guarantees $\mathcal{E}_S^\star \leq \mathcal{E}_T^\star$. Because the entropy lower bound of $\mathsf{H}(S)$ decreases with τ, $\tau = \tau_0$ maximizes the diversity of memory atoms without increasing reconstruction error of query compared to Top-K retrieval when $c < \tau_0 \leq \rho$.

3.3 How Sparse Coefficients Guide Attention

Section 2.3 reuses the sparse coefficients to guide attention, rather than discard their unequal contributions after selection. We examine how closely the memory weights follow these coefficients.

Attention setting. Fix a block, head, and query token. For a nonempty support S, let $\widehat{\alpha}_j > 0$ be the actual coefficients and $\mathcal{M}_j$ disjoint visible memory-token groups. Write $Z_j = \sum_{m\in\mathcal{M}_j} e^{\ell_m} \in (0, \infty)$ for native evidence. The normalized coefficients are $\omega_j = \widehat{\alpha}_j / \sum_i \widehat{\alpha}_i$. Equation 7 gives each memory's conditional share $\beta_j = \omega_j Z_j / \sum_i \omega_i Z_i$. Set $\rho_{\mathrm{att}} = Z_{\min}/Z_{\max}$.

Proposition 2 (Contribution-prior fidelity)**.** *For any nonempty support and positive terminal coefficients, using natural logarithms,*

$$0 \leq D_{\mathrm{KL}}(\boldsymbol{\omega}\|\boldsymbol{\beta}) \leq \frac{\log^2 \rho_{\mathrm{att}}}{8}. \tag{9}$$

Remark 2. When native evidence is similar across memories, attention follows the coefficient distribution closely. Equal evidence gives $\boldsymbol{\beta} = \boldsymbol{\omega}$; otherwise, the bound measures how far native compatibility can move attention away from this prior.

4 Experiments

4.1 Experimental Setup

We follow the 128-prompt VBench-Long protocol of LongLive-RAG (Hu et al., 2026) and evaluate Self-Forcing, LongLive, and Causal-Forcing (Huang et al., 2025; Yang et al., 2026; Zhu et al., 2026) at 30, 60, and 120 seconds with six VBench-Long metrics (Huang et al., 2024). All methods in full VBench-Long experiments share the frozen generator and encoder, candidate history, one sink, six retrieved latents, five recent latents, and the retrieval schedule. CARE-MEM uses $\Omega = 20$ optimization iterations and $\tau = 0.7$. We adaptively set $\eta = 1/\lambda_{\max}(\mathbf{D}_t^\top\mathbf{D}_t)$, where each column of $\mathbf{D}_t$ is an ℓ_2-normalized candidate atom. Every run fits on NVIDIA RTX 3090 (24GB) GPUs.

Table 1: VBench-Long results across 30/60/120-second rollouts. Backbones are Self-Forcing (Huang et al., 2025), LongLive (Yang et al., 2026), and Causal-Forcing (Zhu et al., 2026); compared memory methods are ∞-RoPE (Yesiltepe et al., 2025), Deep Forcing (Yi et al., 2025), and LongLive-RAG (Hu et al., 2026). Published values follow LongLive-RAG Table 2; CARE-MEM values are from our 128-prompt runs with $\tau = 0.7$ under the same protocol. Bold/underline mark best/second-best.

Method	Retrieval Based	Subject Consistency ↑	Background Consistency ↑	Motion Smoothness ↑	Dynamic Degree ↑	Aesthetic Quality ↑	Imaging Quality ↑	AVG ↑
30s generation								
Self-Forcing		96.21	95.39	98.39	52.03	56.69	63.31	77.00
+ ∞-RoPE	×	97.32	96.38	98.59	46.82	56.78	63.93	76.64
+ Deep Forcing	×	97.04	96.02	98.57	38.85	56.44	61.91	74.81
+ LongLive-RAG	✓	97.57	96.56	**98.76**	42.24	57.17	65.43	76.29
+ CARE-MEM	✓	**97.67**	**96.70**	98.53	**53.96**	**57.96**	**66.77**	**78.60**
LongLive		97.35	96.15	98.70	44.74	59.38	68.15	77.41
+ ∞-RoPE	×	97.27	96.19	98.68	48.18	58.69	67.99	77.83
+ Deep Forcing	×	97.52	96.43	**98.82**	41.46	59.00	67.61	76.81
+ LongLive-RAG	✓	97.53	96.39	98.77	44.84	59.24	68.42	77.53
+ CARE-MEM	✓	**97.55**	**96.51**	98.62	**48.49**	**59.53**	**69.60**	**78.38**
Causal-Forcing		94.60	94.68	96.56	73.96	54.58	65.53	79.99
+ ∞-RoPE	×	93.93	94.11	96.21	**90.83**	55.42	68.26	83.13
+ Deep Forcing	×	93.52	93.86	95.84	84.79	55.03	66.07	81.52
+ LongLive-RAG	✓	95.43	94.79	**97.16**	82.29	**57.31**	**70.07**	82.84
+ CARE-MEM	✓	**95.46**	**95.26**	96.72	87.97	56.78	68.81	**83.50**
60s generation								
Self-Forcing		95.84	95.27	98.20	**51.72**	56.05	62.22	76.55
+ ∞-RoPE	×	97.24	96.24	98.58	46.64	56.09	63.28	76.35
+ Deep Forcing	×	96.08	95.38	98.24	41.44	56.68	60.81	74.77
+ LongLive-RAG	✓	97.60	96.51	**98.70**	44.69	57.19	64.97	76.61
+ CARE-MEM	✓	**97.76**	**96.66**	98.60	49.53	**57.88**	**67.01**	**77.91**
LongLive		97.13	95.89	98.61	44.56	58.17	67.56	76.99
+ ∞-RoPE	×	97.00	95.85	98.53	**53.36**	57.48	66.94	78.19
+ Deep Forcing	×	97.17	96.04	**98.73**	45.13	57.48	67.27	76.97
+ LongLive-RAG	✓	97.32	96.08	98.62	49.90	58.30	67.79	78.00
+ CARE-MEM	✓	**97.50**	**96.39**	98.58	50.60	**58.95**	**69.87**	**78.65**
Causal-Forcing		93.52	94.12	95.74	72.32	51.24	62.30	78.21
+ ∞-RoPE	×	93.81	93.78	96.09	**92.47**	54.42	67.50	83.01
+ Deep Forcing	×	94.27	94.18	**96.62**	78.59	52.12	64.25	80.01
+ LongLive-RAG	✓	94.29	94.24	96.48	88.20	54.95	68.16	82.72
+ CARE-MEM	✓	**95.11**	**95.04**	96.48	88.70	**56.53**	**68.60**	**83.41**
120s generation								
Self-Forcing		96.12	95.32	98.27	43.39	55.64	61.57	75.05
+ ∞-RoPE	×	97.15	96.09	98.55	46.29	55.11	61.81	75.83
+ Deep Forcing	×	96.92	**96.83**	**98.97**	15.23	52.84	57.93	69.79
+ LongLive-RAG	✓	**97.64**	96.40	98.75	44.10	56.30	64.16	76.23
+ CARE-MEM	✓	97.56	96.59	98.60	**47.36**	**57.46**	**66.87**	**77.41**
LongLive		96.93	95.64	98.58	47.12	**57.90**	66.95	77.19
+ ∞-RoPE	×	96.81	95.65	98.48	**53.59**	56.73	66.19	77.91
+ Deep Forcing	×	97.17	95.72	98.57	46.03	56.98	66.03	76.75
+ LongLive-RAG	✓	**97.22**	95.88	**98.62**	50.25	57.57	66.95	77.75
+ CARE-MEM	✓	97.15	**96.21**	98.49	53.58	57.80	**69.09**	**78.72**
Causal-Forcing		92.98	94.66	95.41	63.79	47.31	58.23	75.40
+ ∞-RoPE	×	93.45	93.56	95.95	**93.44**	53.47	66.81	82.78
+ Deep Forcing	×	93.19	94.04	95.60	76.45	46.86	61.58	77.95
+ LongLive-RAG	✓	94.38	94.08	**96.56**	90.21	54.82	68.23	83.05
+ CARE-MEM	✓	**95.01**	**95.02**	96.46	87.84	**56.16**	**68.36**	**83.14**

4.2 VBench-Long Results

Consistent improvement with reduced memory over-anchoring. Table 1 shows that CARE-MEM improves the balance between historical consistency and continued motion, ranking best or second-best in 47 of 54 metric–setting comparisons. Self-Forcing at 30 seconds provides the clearest contrast: top-K retrieval improves appearance but suppresses dynamics relative to the retrieval-free backbone; CARE-MEM retains the appearance gains while recovering motion beyond that backbone.

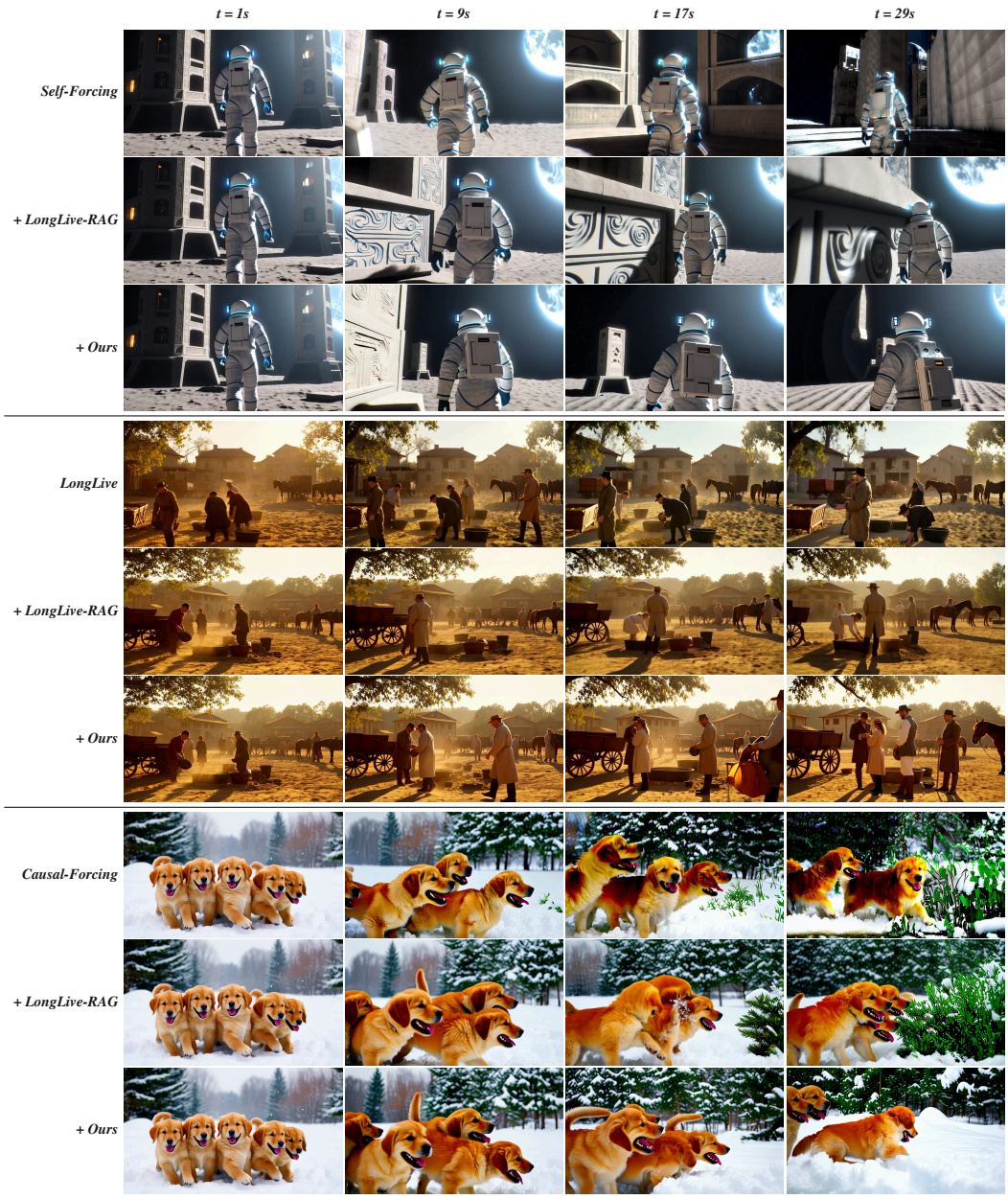


Figure 4: Qualitative 30-second rollouts. We compare three backbones, their +LongLive-RAG variants, and +Ours at four rollout times (Huang et al., 2025; Yang et al., 2026; Zhu et al., 2026; Hu et al., 2026). CARE-MEM preserves subject and scene appearance identity while allowing continued pose and viewpoint changes.

Across the nine settings, Dynamic Degree improves over LongLive-RAG in eight, averaging $+3.48$ points with only a 0.15-point average decrease in Motion Smoothness. The joint pattern supports continued evolution without a comparable loss of temporal smoothness. Background Consistency improves on every backbone at every duration, and Imaging Quality improves in all 60- and 120-second comparisons, showing that the gains over top-K persist beyond short rollouts. Overall, these results show that CARE-MEM consistently improves the balance between long-range consistency and motion evolution across backbones and rollout lengths.

Qualitative evidence. Figure 4 provides qualitative evidence consistent with the benchmark results, showing how CARE-MEM preserves long-range subject and scene consistency while allowing

Table 2: Component ablation on Self-Forcing at 30s (128 prompts). All variants share the frozen backbone and context budget; sparse variants use $K = 6$ and $\tau = 0.7$. SC-Mem denotes constrained sparse coding optimization for memory selection, and CP-Attn denotes coefficient-calibrated priors for attention control. Checkmarks indicate active components; bold/underline mark best/second-best.

Method	Components			VBench-Long					
	Long-term Retrieval	Mitigating Memory Over-Anchoring	Attention Control	Subject ↑	Background ↑	Motion ↑	Dynamic ↑	Aesthetic ↑	Imaging ↑
Self-Forcing				96.21	95.39	98.39	<u>52.03</u>	56.69	63.31
+ LongLive-RAG	✓			97.57	96.56	**98.76**	42.24	57.17	65.43
+ SC-Mem	✓	✓		**97.73**	<u>96.68</u>	<u>98.62</u>	47.97	<u>57.77</u>	<u>66.54</u>
+ SC-Mem + CP-Attn (Ours)	✓	✓	✓	<u>97.67</u>	**96.70**	98.53	**53.96**	**57.96**	**66.77**

continued motion evolution. Across the astronaut, outdoor human-scene, and dog examples, CARE-MEM retains the subject or landmark while allowing continued pose, framing, and viewpoint changes. The LongLive-RAG rows often preserve history strongly, but the CARE-MEM rows more frequently continue the local trajectory rather than returning to a similar retrieved composition. This observation shows that top-K memory retrieval can become over-anchoring when several independently similar states repeatedly enter the limited retrieval slots, restricting continued motion evolution.

Diversity of selected memory atoms. We measure the diversity of retrieved memories using the Vendi Score (Friedman & Dieng, 2023), where a higher score indicates a more diverse set of selected atoms (see Section C). As shown in Figure 5, CARE-MEM achieves consistently higher Vendi Scores across all three backbones, improving the 30-second averages by approximately 17–20%. Since all methods use the same number of historical slots, this indicates that CARE-MEM selects less redundant and more complementary memory atoms instead of repeatedly retrieving similar states. The simultaneous improvements in consistency and dynamics suggest that these more diverse memories remain relevant to the current query through sparse coding reconstruction optimization. Further analysis of memory diversity and sparse coding method comparison is provided in Section D.

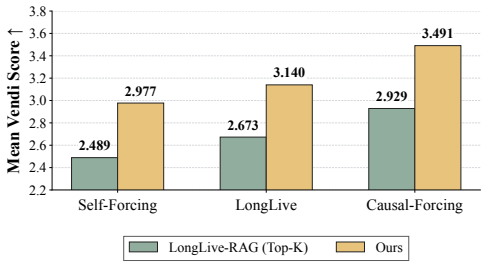


Figure 5: Selected-memory diversity comparison. Mean Vendi Score over retrieval events in 30-second rollouts. CARE-MEM selects more diverse memory atoms across three backbones.

Table 3: Sensitivity to the atom-correlation limit τ on Self-Forcing at 30s. Only τ is varied; all other SC-Mem + CP-Attn settings are fixed. Bold denotes the best value.

τ	Subject ↑	Background ↑	Motion ↑	Dynamic ↑	Aesthetic ↑	Imaging ↑
0.6	**97.71**	96.70	98.52	51.46	**58.00**	66.62
0.7	97.67	**96.70**	**98.53**	**53.96**	57.96	**66.77**
0.8	97.67	96.67	98.51	53.33	57.78	66.45
0.9	97.65	96.68	98.53	52.66	57.81	66.56

Table 4: Retrieved/recent allocation under a fixed 12-latent context on Self-Forcing at 30s. One sink latent is fixed and all other settings are unchanged. Bold denotes the best value.

Retrieved + Recent	Subject ↑	Background ↑	Motion ↑	Dynamic ↑	Aesthetic ↑	Imaging ↑
2 + 9	97.54	96.64	98.52	51.56	56.72	64.24
4 + 7	**97.70**	**96.74**	**98.53**	52.34	57.87	65.75
6 + 5	97.67	96.70	**98.53**	**53.96**	57.96	**66.77**
8 + 3	95.51	95.33	97.34	46.20	**58.23**	64.70

4.3 Ablation Studies

Component Decomposition. We first examine the individual contributions of memory selection and attention control. As shown in Table 2, replacing top-K retrieval with SC-Mem substantially restores the motion dynamics suppressed by redundant historical retrieval, while maintaining strong subject and background consistency. Introducing CP-Attn on top of SC-Mem further improves Dynamic Degree without sacrificing appearance-related quality. These results support the roles of the two components: Constrained sparse coding determines *what* memories should be retrieved, while coefficient-calibrated priors determine *how strongly* the selected memories should influence attention.

Sensitivity Analysis. We further evaluate the atom-correlation threshold τ and the retrieved/recent allocation. As τ decreases, Table 3 shows that Dynamic Degree first increases and then decreases, peaking at $\tau = 0.7$. This suggests that τ is close to the desired balance of minimizing query reconstruction error while maximizing diversity among selected memory atoms: a larger τ admits redundant memories, whereas a smaller τ can exclude useful complementary atoms. Table 4 further shows that a balanced retrieved/recent split best preserves both distant and recent context.

5 Related Work

Autoregressive long-video generation. Diffusion Forcing combines next-token prediction with sequence-level diffusion, providing a general route to causal generation beyond the training horizon (Chen et al., 2024). Self-Forcing trains on self-generated context to reduce exposure bias (Huang et al., 2025); LongLive introduces frame-level autoregression, KV recaching, long-video tuning, and frame sinks for interactive generation (Yang et al., 2026); and Causal Forcing aligns the causal teacher–student structure used in few-step diffusion distillation (Zhu et al., 2026). Reward Forcing further updates persistent sink information and emphasizes dynamic samples during distillation (Lu et al., 2025). These works improve the generator, its training distribution, or its temporal prior (Yan et al., 2021; Ge et al., 2022; Villegas et al., 2023; Yu et al., 2023; Kondratyuk et al., 2024). CARE-MEM is complementary: it keeps the trained generator fixed and changes only how generated history is retrieved and used.

Memory for long rollouts. A bounded causal context can preserve history through persistent anchors, positional remapping, compression, consolidation, or explicit retrieval (Dai et al., 2019; Rae et al., 2020; Wu et al., 2022; Zhang et al., 2023; Li et al., 2024; Munkhdalai et al., 2024). Attention sinks provide a stable prefix for streaming transformers (Xiao et al., 2024). In long-video generation, ∞-RoPE combines relative temporal coordinates with cache refresh (Yesiltepe et al., 2025); Deep Forcing uses deep sinks and participative compression (Yi et al., 2025); MemRoPE maintains evolving long- and short-term memory tokens with online positional assignment (Kim et al., 2026); and FadeMem progressively merges older KV blocks into a dense-near, sparse-far hierarchy (Lu et al., 2026). LongLive-RAG restores explicit non-local latent blocks through content-addressable retrieval (Hu et al., 2026), while OmniMem performs sparse full-range KV access and controls local bias and selection growth (Zhao et al., 2026). CARE-MEM addresses a different decision within a matched historical budget: it jointly chooses a complementary set of memories and assigns a memory-specific prior after selection.

Sparse representation and dictionary learning. Sparse coding represents a signal with a small set of atoms and has long been used to expose reusable structure in natural data (Aharon et al., 2006; Mairal et al., 2010; Tibshirani, 1996; Chen et al., 2001; Foucart, 2011; Hoyer, 2002). Classical recovery analyses relate support identifiability to dictionary conditioning and atom correlation (Donoho & Elad, 2003; Tropp, 2004), and iterative hard thresholding provides a practical fixed-cardinality optimization template (Blumensath & Davies, 2009). CARE-MEM does not learn a new visual dictionary: frozen retrieval descriptors of the generated history form the online dictionary. Its contribution is to use one constrained sparse explanation both as a retrieval decision and as a calibrated interface to attention.

6 Conclusion

We introduce CARE-MEM, a training-free memory retrieval that separates memory selection from memory access in long-video generation. Based on a dictionary view of memory, *constrained sparse coding optimization for memory selection* retrieves a jointly explanatory set of historical atoms, while the atom-correlation constraint reduces redundant memory selections. The *coefficient-calibrated priors for attention control* then use the sparse coding coefficients to determine how strongly each selected memory participates in attention. Across multiple autoregressive backbones and rollout lengths, CARE-MEM preserves long-range subject and scene information while allowing continued motion, pose, and viewpoint evolution. These results suggest that effective long-video memory should not rely on independent similarity ranking followed by uniform access; instead, historical states should explain the current query together and participate according to their respective contributions.

AI USE STATEMENT

In this work, we did not use generative AI tools for any task with required disclosure: we did not use them to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean or reformat datasets, or interpret results; supporting qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for two tasks with recommended disclosure: to aid and polish the writing and to support the retrieval and discovery of potentially relevant literature. We did not use generative AI tools to draft any section of the paper. We have reviewed all AI-assisted work: all AI-assisted writing was reviewed and revised by the authors, and all retrieved references and related-work information were manually verified against the original sources. We take responsibility for the final content of this work, including all text and references produced with the aid of generative AI.

REFERENCES

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11): 4311–4322, 2006. doi: 10.1109/TSP.2006.881199.

Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009. doi: 10.1016/j.acha.2009.04.002.

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

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001. doi: 10.1137/S003614450037906X.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019. doi: 10.18653/v1/P19-1285.

David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, 100(5): 2197–2202, 2003. doi: 10.1073/pnas.0437847100.

Simon Foucart. Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011. doi: 10.1137/100806278.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic VQGAN and time-sensitive transformer. *arXiv preprint arXiv:2204.03638*, 2022.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Patrik O. Hoyer. Non-negative sparse coding. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 557–565, 2002. doi: 10.1109/NNSP.2002.1030067.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Youngrae Kim, Qixin Hu, C.-C. Jay Kuo, and Peter A. Beerel. MemRoPE: Training-free infinite video generation via evolving memory tokens. *arXiv preprint arXiv:2603.12513*, 2026.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. VideoPoet: A large language model for zero-shot video generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25105–25124, 2024.

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

Yu Lu, Junjie Yang, Piotr Koniusz, YuXin Song, and Yi Yang. FadeMem: Distance-aware memory consolidation for autoregressive video diffusion. *arXiv preprint arXiv:2606.10671*, 2026.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 2024.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005. doi: 10.1109/TPAMI.2005.159.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004. doi: 10.1109/TIT.2004.834793.

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 descriptions. In *International Conference on Learning Representations*, 2023.

John Watrous. *The Theory of Quantum Information*. Cambridge University Press, 2018. doi: 10.1017/9781316848142.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022.

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

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-RoPE: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 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.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. MAGVIT: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lin Zhao, Yushu Wu, Yifan Gong, Yanzhi Wang, and Pu Zhao. OmniMem: Scalable and adaptive memory retrieval for long video generation. *arXiv preprint arXiv:2605.30519*, 2026.

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.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Proof. We compare the best nonnegative fits on the two supports. The setting gives $c < \tau \le \rho$; the additional condition $\tau \le b/s$ is used in the top-2 comparison.

The complementary fit. Let $\mathbf{z} = (\mathbf{w} - c\mathbf{u})/\sqrt{1-c^2}$. Then $\mathbf{u}, \mathbf{z}$ are orthonormal, $\mathbf{q}^\top \mathbf{z} = g$, and $\mathbf{q}_\perp = \mathbf{q} - s\mathbf{u} - g\mathbf{z}$ is orthogonal to both. Pythagoras gives, for any $a, d \ge 0$,

$$\frac{1}{2}\|\mathbf{q} - a\mathbf{u} - d\mathbf{w}\|_2^2 = \frac{1-s^2-g^2}{2} + \frac{(s-a-cd)^2}{2} + \frac{(g-\sqrt{1-c^2}\,d)^2}{2}. \tag{10}$$

Both squares vanish at $d^\star = (\xi - cs)/(1-c^2) > 0$ and $a^\star = (s - c\xi)/(1-c^2) > 0$, where positivity follows from $cs < \xi \le s$ and $0 \le c < 1$. Hence

$$\mathcal{E}_S^\star = \frac{1-s^2-g^2}{2}. \tag{11}$$

The top-2 comparison. Put $\mathbf{v}_q = (s, b)^\top$, $\mathbf{G}_\tau = \left(\begin{smallmatrix} 1 & \tau \\ \tau & 1 \end{smallmatrix}\right)$, and $C_\tau = \mathbf{v}_q^\top \mathbf{G}_\tau^{-1} \mathbf{v}_q = (s^2 + b^2 - 2\tau sb)/(1-\tau^2)$. For arbitrary $\mathbf{a} \ge \mathbf{0}$, $\rho \ge \tau$ gives $\|\mathbf{D}_T\mathbf{a}\|_2^2 \ge \mathbf{a}^\top \mathbf{G}_\tau \mathbf{a}$. Since $\mathbf{G}_\tau$ is positive definite, weighted Cauchy–Schwarz gives $\mathbf{v}_q^\top \mathbf{a} \le \sqrt{C_\tau}\sqrt{\mathbf{a}^\top \mathbf{G}_\tau \mathbf{a}}$. With $y = \sqrt{\mathbf{a}^\top \mathbf{G}_\tau \mathbf{a}}$, expansion and completing the square yield

$$\begin{aligned} \frac{1}{2}\|\mathbf{q} - \mathbf{D}_T\mathbf{a}\|_2^2 &\ge \frac{1}{2} - \sqrt{C_\tau}\, y + \frac{y^2}{2} \\ &= \frac{1-C_\tau}{2} + \frac{(y - \sqrt{C_\tau})^2}{2} \ge \frac{1-C_\tau}{2}. \end{aligned} \tag{12}$$

The score ordering and $\tau \le b/s$ imply

$$\frac{s(s+b)}{1+\tau} - C_\tau = \frac{(s-b)(b-\tau s)}{1-\tau^2} \ge 0.$$

Taking the minimum over $\mathbf{a} \ge \mathbf{0}$ and subtracting Equation (11),

$$\mathcal{E}_T^\star \ge \frac{1}{2} - \frac{s(s+b)}{2(1+\tau)}, \tag{13}$$

$$\mathcal{E}_T^\star - \mathcal{E}_S^\star \ge \frac{s^2+g^2}{2} - \frac{s(s+b)}{2(1+\tau)} = \frac{(s^2+g^2)(\tau - \tau_0)}{2(1+\tau)}. \tag{14}$$

The last expression is nonnegative for $\tau \ge \tau_0$ and positive for $\tau > \tau_0$.

The threshold and diversity. Write $d = \xi - cs \in (0, s(1-c)]$. Since $b \ge \xi$ and $g^2 = d^2/(1-c^2)$,

$$(s^2+g^2)(\tau_0 - c) = s(b - cs) - (1+c)g^2 \ge d\left(s - \frac{d}{1-c}\right) \ge 0,$$

$$\frac{b}{s} - \tau_0 = \frac{g^2(s+b)}{s(s^2+g^2)} > 0. \tag{15}$$

Thus $c \le \tau_0 < b/s \le 1$. A pair with correlation $z \in [0,1]$ has normalized Gram eigenvalues $(1 \pm z)/2$. Its von Neumann entropy (Watrous, 2018) is therefore

$$\mathsf{H}(z) = h_2\left(\frac{1+z}{2}\right), \qquad \mathsf{H}'(z) = \frac{1}{2}\log\frac{1-z}{1+z} < 0 \quad (0 < z < 1). \tag{16}$$

Using $0 \log 0 = 0$ at the endpoints, $c < \tau \le \rho$ implies

$$\mathsf{H}(S) > \underbrace{h_2\left(\frac{1+\tau}{2}\right)}_{\underline{H}(\tau)} \ge \mathsf{H}(T).$$

When $c < \tau_0 \le \rho$, every admissible $\tau \ge \tau_0$ satisfies $\underline{H}(\tau) \le \underline{H}(\tau_0)$. Hence $\tau = \tau_0$ gives the largest entropy lower bound among the thresholds certified to preserve reconstruction. □

A.2 Proof of Proposition 2

Proof. Fix one generation block, attention head, and query token. Let S be the nonempty returned support, with positive coefficients $\widehat{\alpha}_j$ and disjoint visible memory-token groups $\mathcal{M}_j$. The logits are finite, so each $Z_j = \sum_{m\in\mathcal{M}_j} e^{\ell_m}$ is finite and positive. We first express the attention received by each memory in terms of its coefficient. We then bound how much the native evidence changes their relative contributions.

How the coefficients enter the attention weights. In Section 2.3, every token in memory j receives the same multiplier $\widehat{\alpha}_j$, while non-memory tokens retain multiplier one. Let $Z_0 = \sum_{m\in\mathcal{F}} e^{\ell_m}$ denote the total non-memory evidence. Summing Equation 7 over memory j gives its attention mass

$$P_j = \sum_{m\in\mathcal{M}_j} \frac{\widehat{\alpha}_j e^{\ell_m}}{Z_0 + \sum_i \widehat{\alpha}_i Z_i} = \frac{\widehat{\alpha}_j Z_j}{Z_0 + \sum_i \widehat{\alpha}_i Z_i},$$

$$\beta_j = \frac{P_j}{\sum_i P_i} = \frac{\widehat{\alpha}_j Z_j}{\sum_i \widehat{\alpha}_i Z_i} = \frac{\omega_j Z_j}{\sum_i \omega_i Z_i}, \qquad \omega_j = \frac{\widehat{\alpha}_j}{\sum_i \widehat{\alpha}_i}. \tag{17}$$

Here β_j is memory j's share of the attention assigned to all memories, not its share of the full context. The common denominator cancels when taking the ratio between memories:

$$\frac{\beta_i}{\beta_j} = \frac{\widehat{\alpha}_i}{\widehat{\alpha}_j}\frac{Z_i}{Z_j}.$$

Thus the coefficient ratio directly influences the relative attention, while the evidence ratio allows native compatibility to adjust it. Within one memory, the multiplier cancels instead: for $m \in \mathcal{M}_j$, $p_m/P_j = e^{\ell_m}/Z_j$. The prior therefore changes how attention is distributed across memories without changing the relative token weights inside each memory. Normalization into $\boldsymbol{\omega}$ is only for this analysis; the implemented attention still uses the raw coefficients in P_j.

How far attention can move from the coefficient distribution. The KL divergence compares the normalized coefficients $\boldsymbol{\omega}$ with the resulting memory attention shares $\boldsymbol{\beta}$. All sums below are over S. Substituting Equation (17) gives

$$\begin{aligned} D_{\mathrm{KL}}(\boldsymbol{\omega}\|\boldsymbol{\beta}) &= \sum_j \omega_j \log \frac{\omega_j}{\omega_j Z_j / \sum_i \omega_i Z_i} \\ &= \sum_j \omega_j \left(\log \sum_i \omega_i Z_i - \log Z_j \right) \\ &= \log \sum_i \omega_i Z_i - \sum_j \omega_j \log Z_j. \end{aligned} \tag{18}$$

To bound this expression, let X take the value $\log Z_j$ with probability ω_j. Then

$$D_{\mathrm{KL}}(\boldsymbol{\omega}\|\boldsymbol{\beta}) = \log \mathbb{E}[e^X] - \mathbb{E}[X] = \log \mathbb{E}[e^{X-\mathbb{E}[X]}].$$

Jensen's inequality gives

$$\mathbb{E}[e^{X-\mathbb{E}[X]}] \geq e^{\mathbb{E}[X-\mathbb{E}[X]]} = 1,$$

so the divergence is nonnegative. For the upper bound, Hoeffding's lemma (Hoeffding, 1963) states that a random variable supported on $[a, b]$ satisfies

$$\log \mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \frac{\lambda^2 (b-a)^2}{8}.$$

Here $a = \log Z_{\min}$, $b = \log Z_{\max}$, and $\lambda = 1$. Consequently,

$$\begin{aligned} 0 \leq D_{\mathrm{KL}}(\boldsymbol{\omega}\|\boldsymbol{\beta}) &\leq \frac{(\log Z_{\max} - \log Z_{\min})^2}{8} \\ &= \frac{\log^2(Z_{\max}/Z_{\min})}{8} = \frac{\log^2 \rho_{\mathrm{att}}}{8}. \end{aligned} \tag{19}$$

This proves Equation 9. If every memory has the same native evidence, the evidence cancels in Equation (17) and $\boldsymbol{\beta} = \boldsymbol{\omega}$: attention exactly follows the relative coefficients. As $\rho_{\mathrm{att}} \to 1$, the upper bound tends to zero. Thus similar native evidence preserves the coefficient guidance, while differences in native evidence can change the allocation by the bounded amount. No optimality assumption on the sparse coefficients is used. □

B TOY-EXAMPLE DETAILS

Dataset. We procedurally generate 60,000 grayscale Moving-MNIST clips at 64×64 resolution. Each 64-frame clip contains a canonical digit 1 moving within the left half of the canvas and a canonical digit 8 moving independently within the right half under two-dimensional bouncing motion. Initial positions and velocities are randomized. Frames 0–7 render digit 1 at unit intensity and digit 8 at 0.08 intensity; frames 8–27 reverse these intensities; and frames 28–63 render both digits at unit intensity. The resulting training distribution therefore contains the transition $1 \to 8 \to (1 + 8)$. This transition is learned by the generator and is not created by retrieval.

Model architecture. The backbone is a compact latent-video diffusion model. A spatiotemporal convolutional VAE maps each one-channel video to eight latent channels, with temporal downsampling by 2 and spatial downsampling by 8; a 64-frame clip is represented by 32 latent frames of size 8×8. The VAE uses 64 hidden channels and deterministic latents ($\beta = 0$), and contains 7.56M parameters. The latent denoiser uses a convolutional spatial stem followed by a six-block Transformer with width 256, eight attention heads, head dimension 32, and feed-forward dimension 1,024. Separate left- and right-digit embeddings provide conditioning for the 100 ordered digit pairs. The base denoiser contains 5.306M parameters; the causal student contains 5.775M parameters after adding rank-16 LoRA adapters. The complete VAE and student contain 13.335M parameters.

Training. We train the VAE for 10,000 updates and a bidirectional full-video diffusion teacher for 12,000 updates. A causal student is initialized from the teacher and adapted for 8,000 updates using at most the two most recent ground-truth VAE latents as context. We then fine-tune the same rank-16 LoRA adapters with Self-Forcing distribution-matching distillation (Huang et al., 2025), replacing ground-truth context with the student's own autoregressive rollout. The VAE and copied base denoiser remain frozen during this stage. Generation uses four denoising steps per latent frame and EMA weights; the visualization in Figure 3 uses the checkpoint at distillation step 1,500.

C VENDI SCORE FOR SELECTED-MEMORY DIVERSITY

The Vendi Score (Friedman & Dieng, 2023) measures diversity through the spectrum of a similarity matrix. For a retrieval event e with $n_e \geq 1$ selected atoms, let $\mathbf{G}_e$ be their Gram matrix under a positive semidefinite kernel with unit diagonal. Given the eigenvalues $\{\lambda_{e,i}\}$ of $\mathbf{G}_e/n_e$,

$$\mathrm{VS}(S_e) = \exp\left(-\sum_{i=1}^{n_e} \lambda_{e,i} \log \lambda_{e,i} \right). \tag{20}$$

The logarithm is natural, and atoms are weighted equally, independently of sparse coefficients or attention weights.

For unit-normalized descriptor columns $\mathbf{X}_e$, the cosine kernel gives $\mathbf{G}_e = \mathbf{X}_e^\top \mathbf{X}_e$, which remains positive semidefinite even with negative pairwise similarities. The score captures spectral diversity rather than mean pairwise cosine and ranges from 1 for identical atoms to n_e for mutually orthogonal atoms. It thus reflects an effective number of descriptor directions, not semantic identities, and depends on both support size and geometry. The one-sided τ constraint does not guarantee a higher score.

Figure 5 reports the equally weighted mean over evaluated nonempty events $\mathcal{I}$:

$$\overline{\mathrm{VS}} = \frac{1}{|\mathcal{I}|} \sum_{e \in \mathcal{I}} \mathrm{VS}(S_e). \tag{21}$$

This differs from pooling atoms across events. Comparisons require consistent descriptor preprocessing, kernels, and event inclusion rules. Numerically, negligible negative eigenvalues may be clipped before renormalization; substantial negatives indicate an invalid matrix. Vendi Score is an offline diversity diagnostic, not a guarantee of reconstruction accuracy or video quality.

Table 5: Retrieved atom diversity and video quality on Self-Forcing at 30s (128 prompts). Sparse coding variants use the optimization coefficients as coefficient-calibrated priors for attention. AVG averages the six VBench-Long metrics. Bold and underline indicate best and second-best values, determined before rounding.

Method	Sparse Coding	Atom Diversity		VBench-Long	
		Mean Cosine	Vendi Score	Dynamic Degree ↑	AVG ↑
LongLive-RAG	✗	0.69	2.49	42.24	76.29
Lasso	✓	0.67	2.24	<u>51.72</u>	<u>78.26</u>
Elastic Net	✓	0.67	2.45	51.25	78.21
Ours (w/o τ)	✓	0.70	2.42	50.47	78.01
Ours (w/ τ)	✓	0.58	2.98	**53.96**	**78.60**

D Analysis of Different Sparse Coding Methods

Table 5 compares nonnegative Lasso (Tibshirani, 1996), Elastic Net (Zou & Hastie, 2005), and our method. We solve Lasso and Elastic Net using scikit-learn's coordinate descent, retain at most $K = 6$ positive coefficients, and apply the same coefficient-to-attention rule across all sparse variants.

The results show that *sparse coding alone does not ensure complementary memories*. Lasso promotes sparse reconstruction without explicitly excluding redundant atoms, while Elastic Net encourages coefficient sharing among correlated predictors. Both reduce mean cosine from 0.69 to 0.67, but their Vendi Scores remain below LongLive-RAG. This is not contradictory: mean cosine measures average pairwise alignment, whereas Vendi captures spectral diversity and also depends on support size. But in general, mean cosine and Vendi are negatively correlated with each other. Likewise, our variant without τ has the highest mean cosine, demonstrating that a cardinality budget alone does not prevent redundant retrieval.

Adding the correlation constraint in our method reduces mean cosine to 0.58 and raises Vendi Score from 2.42 to 2.98, while improving Dynamic Degree from 50.47 to 53.96 and AVG from 78.01 to 78.60. Because AVG includes Dynamic, this improvement primarily reflects stronger dynamics with otherwise stable aggregate performance. Importantly, Lasso and Elastic Net also improve dynamics despite lower Vendi Scores, so the benefits cannot be attributed solely to diversity, but also to the attention prior based on the coefficients of sparse coding. Together, these results support complementary roles for reconstruction, redundancy control, and coefficient-calibrated memory access, consistent with reduced memory over-anchoring.

E Generation Examples



Figure 6: For this example from the Self-Forcing backbone, CARE-MEM increases the Vendi Score (Friedman & Dieng, 2023) from 2.51 to 3.23, indicating that it selects a more diverse and complementary set of memory atoms. Three of the six retrieved atoms are shown and the Vendi Score is computed over all six.

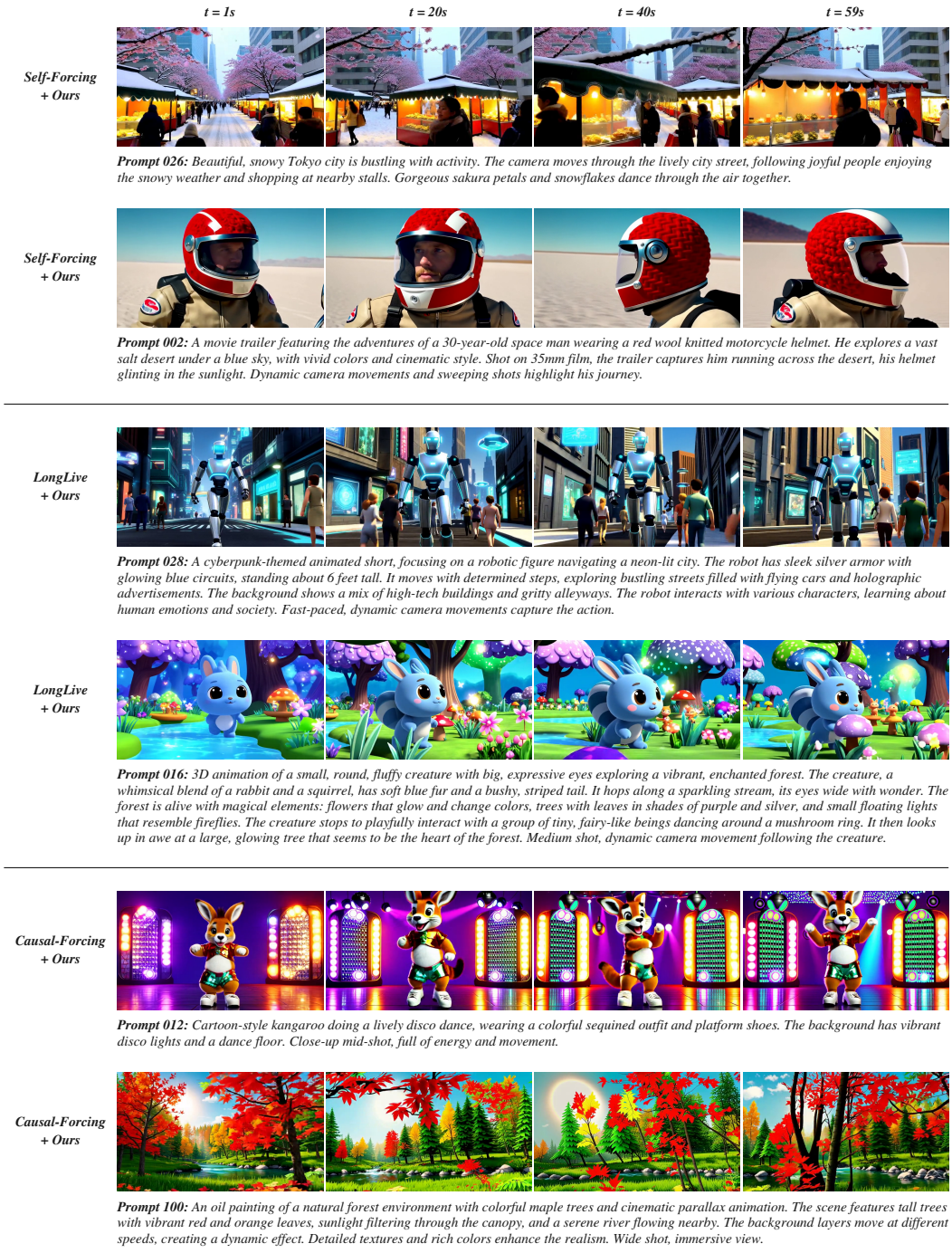


Figure 7: Qualitative 60-second rollouts with CARE-MEM. Two examples are shown for each backbone: Self-Forcing, LongLive, and Causal-Forcing. Frames sampled at 1, 20, 40, and 59 s show recognizable subjects and environments alongside changes in pose, viewpoint, and scene content. The full text prompt is provided below each example.

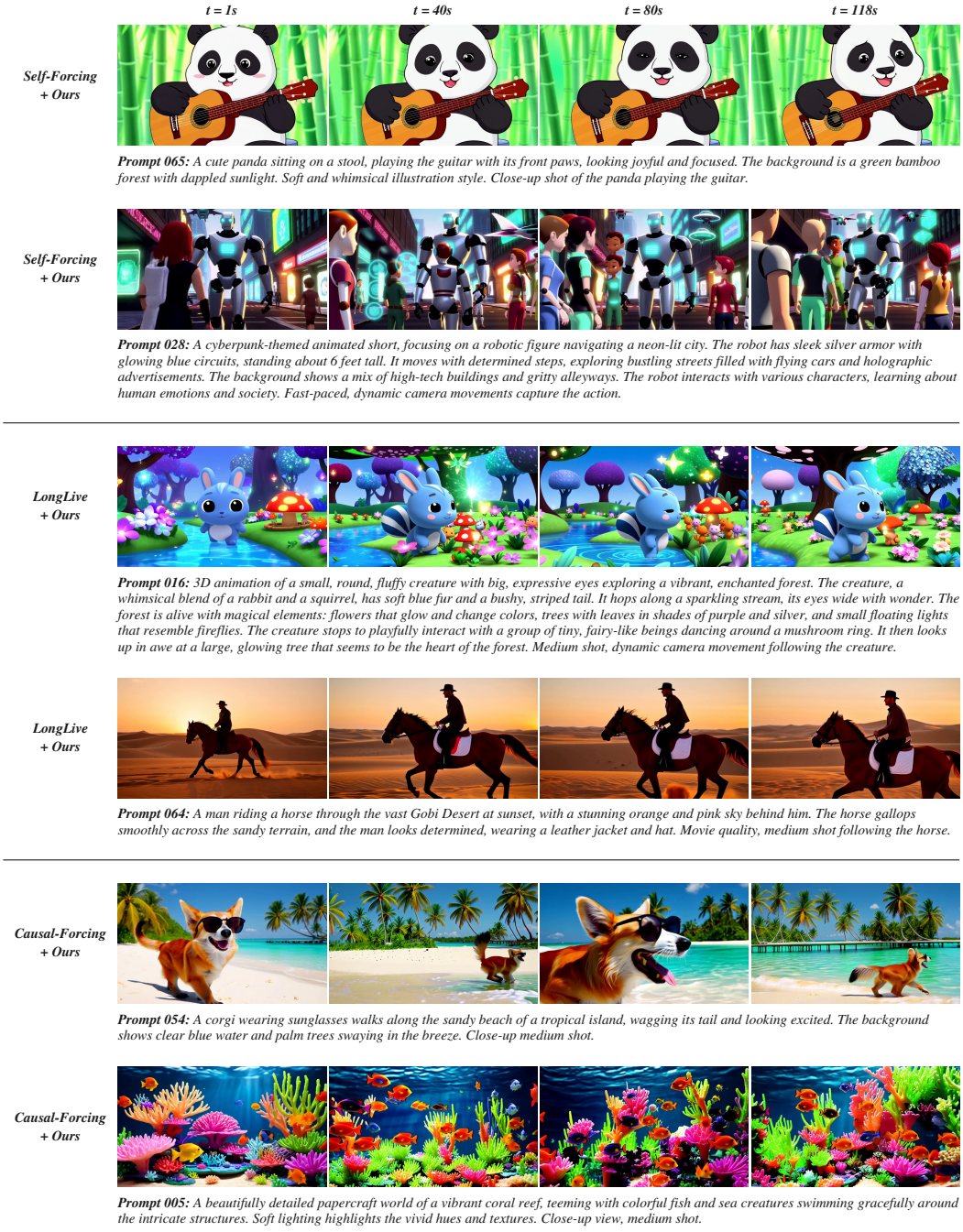


Figure 8: Qualitative 120-second rollouts with CARE-MEM. Two examples are shown for each backbone: Self-Forcing, LongLive, and Causal-Forcing, with frames sampled at 1, 40, 80, and 118 s. The panda-playing-guitar example illustrates long-horizon appearance consistency; the other examples show changes in interaction, pose, viewpoint, and scene content. The full text prompt is provided below each example.