RECOLLAGE: SELF-SUPERVISED TOKEN COLLAGES AS MEMORY FOR CONTROLLABLE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long video generation requires compact memory that preserves past content despite prolonged absences, repetitive observations, and distractors. Existing approaches often rely on hand-designed compression rules that overlook uneven information density, favor recent observations, or learn new memory representations that require additional adaptation by the generator. An effective compressor should allocate memory according to content, preserve informative evidence across the full history, and remain compatible with the generator's native representation. Guided by these principles, we introduce ReCollage, a self-supervised reconstruction method that learns to select historical VAE tokens as collage memory for future generation. A context encoder uses learnable queries to form the collage memory, which is supplied as an inpainting condition to a video diffusion model for reconstruction. During training, content-dependent selections that retain informative regions while reducing redundant storage emerge, making collage an effective history representation. The frozen selector supports three interfaces: sparse memory conditioning, frame-level conditioning, and autoregressive KV-cache management, with the latter two requiring no downstream training. We evaluate these interfaces on three revisit benchmarks spanning hand-controlled, camera-controlled, and text-to-video generation. Under matched memory budgets, ReCollage improves recovery of historical objects, scenes, and subjects despite intervening distractors and long repetitive intervals, demonstrating the effectiveness of reconstruction-trained, content-adaptive token selection for long video history compression.

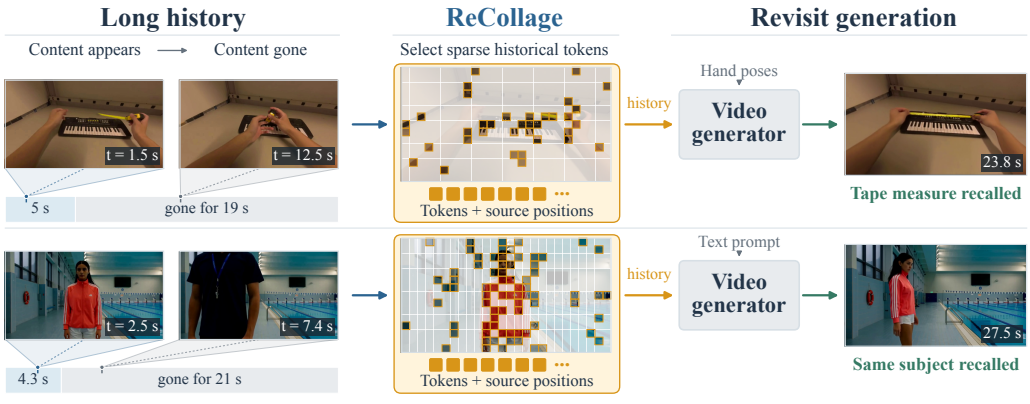
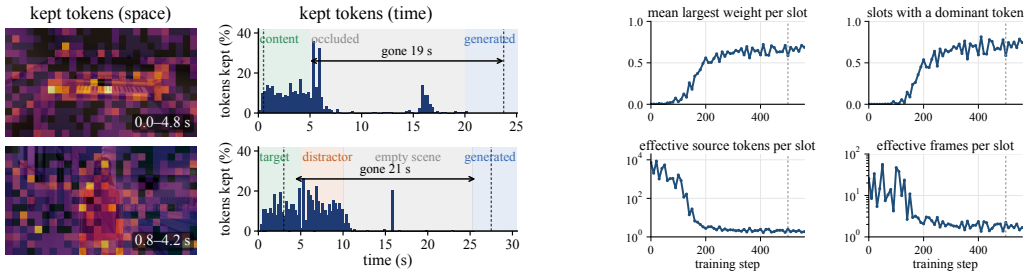


Figure 1: **ReCollage compresses history into collage memory for future generation.** Left: content appears and later leaves the view. Middle: ReCollage selects sparse video tokens with source positions. Right: conditioning on this memory recovers previously observed content during a revisit.

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.



(a) What the selector keeps.

(b) How the selection emerges.

Figure 2: **What ReCollage keeps, and how it learns to keep it.** (a) For the two clips of Figure 1, the spatial and temporal density of kept tokens. (b) Selection statistics of the encoder during reconstruction training: mean largest weight per slot; fraction of slots with largest weight > 0.5; median effective source tokens and frames per slot (log scale).

1 Introduction

Video generation is increasingly used to simulate environments that respond to text, camera motion, and human actions (Wang et al., 2026b; Sun et al., 2025; Gao et al., 2026). Maintaining consistency over long-term interactions requires remembering past content, even after it leaves the current view. However, retaining all the history quickly exceeds the generator's context budget. Existing methods achieve history management through hand-designed sampling, sliding windows, or importance-based packing (Huang et al., 2025; Zhang et al., 2025; Oshima et al., 2026; Yi et al., 2026), while other approaches learn compact history embeddings or evolving memory states (Zhang et al., 2026; Bian et al., 2026; Yu et al., 2025b).

These approaches expose three challenges in history compression. First, information is distributed unevenly across a video: a brief object interaction may contain evidence needed much later, while a long static interval adds little new information. Fixed sampling and compression schedules cannot adapt their allocation to this variation, and hand-designed importance rules provide only proxies for what should be remembered. Second, policies that prioritize recent observations can discard early evidence before it becomes relevant again. Third, compressing history into a newly learned representation introduces an additional interface that the generator must learn to interpret, often through dedicated memory modules or joint adaptation.

We argue that effective history compression should have three properties. It should be (a) *content-adaptive*, allocating capacity according to the information present in the video; (b) provide *coverage across the full history*, allowing informative early observations to remain available despite long interruptions; and (c) maintain *compatibility with the generator*, so that retained evidence can be conveniently consumed through its existing video representation.

Together, these properties motivate us to design **ReCollage**, a self-supervised approach that learns to compress video history into *collage memory* in the VAE domain. The main component of ReCollage is a context encoder that uses a fixed set of learnable queries to select informative tokens from historical VAE tokens. We train this encoder by asking a video diffusion model (VDM) to reconstruct the input video from the resulting compressed information. Reconstruction provides a training signal for historical information retention without requiring annotations of important objects, keyframes, or token locations. After training, the encoder is frozen, and its selected collage memory is used as history for future prediction.

However, a central challenge is making token selection, including its source-position information, differentiable. Discrete token indices interrupt gradient propagation, while simply aggregating tokens into query slots obscures where their content originated. ReCollage addresses this through a coupled collage-and-scatter construction. Each query first produces assignment weights over the history, forming a collage token as a weighted linear combination of the original VAE patches. We then reuse these weights to scatter the collage back onto the source video grid. The soft-masked history has the original history's spatiotemporal shape, with both its values and their placement determined differentiably by the assignments. The VDM takes it as an inpainting condition and reconstructs

the input video. Reconstruction gradients thus train the encoder's selection and placement together, without discrete sampling during training.

Our analysis shows that sharp, content-dependent selection emerges from the reconstruction objective. At initialization, each slot distributes its weight broadly across source tokens. As training progresses, the assignments sharpen and concentrate on a few spatiotemporally localized patches (Figure 2b). The resulting selections also adapt to video content. For the two examples in Figure 1, we visualize spatial selection frequencies as heatmaps, with brighter regions indicating more frequent retention. These regions highlight the manipulated object and hands, or the person's appearance. The temporal distributions further show sparse allocations over long stretches of repeated static observations during the subjects' absence (Figure 2a). These observations indicate that the encoder allocates its memory budget according to content, preserving informative evidence while reducing redundant storage.

We demonstrate three uses of the frozen context encoder for memory-conditioned generation. For *sparse memory conditioning*, selected native tokens retain their source spatial and temporal coordinates, and a predictor is fine-tuned to generate future segments from this sparse history. For *frame-level conditioning*, we aggregate query selections into frame scores and supply the highest-scoring full frames to an existing memory bank. For an *autoregressive KV-cache policy*, new and retained historical tokens compete for query slots, determining which corresponding KV entries remain available for subsequent generation. The latter two interfaces require no downstream parameter updates, allowing the learned selection policy to operate through existing generation interfaces.

We evaluate these applications in hand-controlled, camera-controlled, and text-to-video generation, respectively. Each benchmark requires a revisit: important content first appears in the history, disappears from view, and must be recovered later. Distracting observations and long repetitive intervals separate the initial appearance from the revisit, testing whether a compressor retains the evidence needed for recovery. Under matched memory budgets, ReCollage improves consistency across all three settings. For example, its hand-controlled predictions improve the VLM consistency score over TinyHistory (Zhang et al., 2026) by 7.5 points on both EgoDex (Hoque et al., 2026) and ARCTIC (Fan et al., 2023). In qualitative comparisons, competing methods often lose object details, change a subject's appearance, or fail to recover the target, whereas ReCollage preserves evidence that supports faithful revisits. Figure 1 illustrates recovery after roughly 20 seconds of target absence, despite the intervening distractors or repeated static observations.

Our contributions are threefold:

- We introduce self-supervised reconstruction training for content-adaptive video history compression in the VAE domain, using a differentiable collage-and-scatter interface to learn token selection with source-position information.
- We demonstrate three memory interfaces for the learned selector: sparse token conditioning, frame-level conditioning, and autoregressive KV-cache retention.
- We evaluate ReCollage on three revisit benchmarks spanning hand control, camera control, and text-to-video generation, demonstrating improved recovery of historical objects, scenes, and subjects under matched memory budgets.

2 Related work

2.1 Video Compression

Learned video compression spans rate-distortion optimization (Lu et al., 2019), holistic query-based tokenization (Wang et al., 2025), and content-adaptive token budgets (Yan et al., 2025). Efficient visual representation learning reduces tokens through learned aggregation (Ryoo et al., 2021), similarity-based merging (Bolya et al., 2023), or motion-guided pruning (Hwang et al., 2024). VideoMAE (Tong et al., 2022) and AdaMAE (Bandara et al., 2023) learn through masked reconstruction, with AdaMAE adapting visible-token sampling via policy gradients. Long-VMNet (Gurukar & Kadav, 2025) learns patch selection for fixed-size memory in video understanding. ReCollage learns differentiable, content-adaptive selection through reconstruction, retaining native VAE tokens within a fixed conditioning budget for future generation.

2.2 Memory Management for Long Video Generation

Retrieval-based memory. Historical frames are retrieved using camera geometry and FoV overlap (Xiao et al., 2025; Sun et al., 2025; Yu et al., 2025a) or learned content descriptors (Hu et al., 2026), limiting active context without necessarily bounding history storage. Learned access mechanisms use shallow-layer query tokens (Yu et al., 2026a) or chunk-level attention routing (Cai et al., 2026). ReCollage provides complementary, content-adaptive compression that can be combined with these retrieval mechanisms.

Rule-based retention and compression. Fixed temporal sampling and rolling windows (Huang et al., 2025) bound context but can discard distant evidence. Compression policies use temporal proximity or feature similarity (Zhang et al., 2025), viewpoint relevance (Oshima et al., 2026), attention participation (Yi et al., 2026), head-specific static and similarity-based pruning (Ji et al., 2026), or exponential moving averages (Kim et al., 2026). WorldTrace (Wu et al., 2026b) assigns virtual positions to compressed summaries or retained scene traces. Geometry-based methods use depth and camera poses to compose retrieved latent patches in the queried view (Yu et al., 2026b) or maintain a persistent 3D latent cache (Wang et al., 2026a).

Hierarchical generation. Global planning reduces context demands through keyframes with recursive diffusion-based interpolation (Yin et al., 2023) or video-wide semantic tokens guiding clip synthesis (Ouyang et al., 2025; Bai et al., 2025), maintaining consistency without a growing, fully detailed causal history.

Learned memory representations and selection. Learned compression uses pretrained history embeddings jointly adapted with the generator (Zhang et al., 2026), video–text representations (Jiang et al., 2026), or recursive fixed-budget latent memory (Wu et al., 2026a). Recurrent alternatives employ gated attention queries (Bian et al., 2026), hybrid local attention and state-space memory (Yu et al., 2025b), or test-time-trained neural hidden states (Dalal et al., 2025). SlotMemory (Dou et al., 2026) instead preserves original KV tokens, indexed by object-centric slots for prompt- and context-dependent retrieval. ReCollage preserves native values through self-supervised selection of input-space VAE patches.

3 ReCollage: Self-Supervised Token Collages as Memory

3.1 Problem Setup and Overview

We study how to represent a video history H as compact memory for generating a future segment Y conditioned on c, such as text, camera motion, or hand poses. We learn to select a sparse collection of historical video tokens, termed a *collage memory*. Specifically, a pretrained video VAE encodes H into latents, which are patchified into N tokens $V = [v_1, \ldots, v_N]^\top \in \mathbb{R}^{N \times d_v}$. Each token has a source position $p_j = (t_j, h_j, w_j)$ on the latent patch grid. Given a budget of $B \ll N$ memory slots, our goal is to retain informative collage memory $\mathcal{M}(H) = \{(v_j, p_j) : j = \{0, 1, ..., B\}\}$ for future generation, where $\mathcal{M}(H)$ is a list of value-position pairs of video tokens.

ReCollage learns this collage memory through self-supervised reconstruction. A context encoder uses learnable queries to form the collage memory, which is then supplied as an inpainting condition to a video diffusion model to reconstruct the input video. For future video generation applications, we freeze the context encoder and use the selected collage memory as history for future prediction, as shown in the right of Figure 3.

3.2 Self-Supervised Token Collages

Content-dependent assignments with native token values. The N history patches and B learnable query tokens are concatenated together and sent to the **context encoder**. The context encoder consists of a stack of Transformer blocks followed by a final selection head. The former employs self-attention to jointly process the tokens, resulting in history features f_h and query features f_q. The selection head then produces query projections $Q = W_q f_q, Q \in \mathbb{R}^{B \times d_k}$ and history projections $K = W_k f_h, K \in \mathbb{R}^{N \times d_k}$. The selection head then computes the cosine similarity matrix with

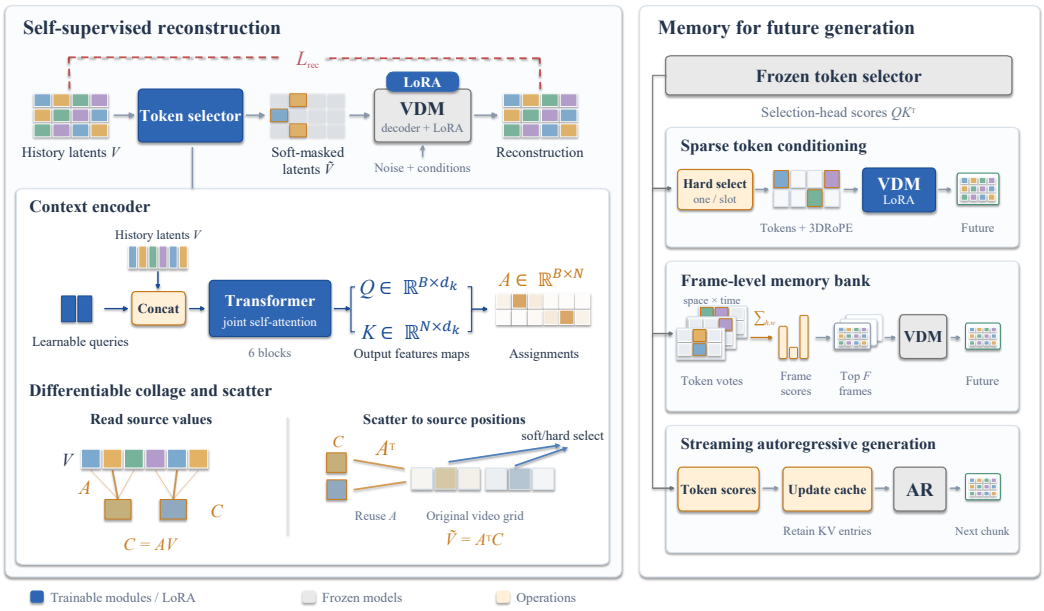


Figure 3: **ReCollage overview. Left:** Query–key assignments A read unprojected VAE patches V into a token collage $C = AV$, which is scattered back as $A^\top C$ to condition video diffusion model (VDM) reconstruction. **Right:** The frozen selector provides (1) sparse tokens with source spatiotemporal 3D-RoPE, requiring predictor fine-tuning; (2) full keyframes selected by aggregated query votes; or (3) streaming AR memory updated by per-query competition over new and retained token logits.

softmax:

$$A = \mathrm{softmax}_N\left(\frac{QK^\top}{\tau\sqrt{d_k}}\right), \qquad C = AV, \tag{1}$$

where $\tau > 0$ is a temperature and the softmax is taken over valid history positions. Each row of $A \in \mathbb{R}^{B\times N}$ specifies how one memory slot draws information from the history, and $C \in \mathbb{R}^{B\times d_v}$ is the **token collage**. We use the original VAE patches as values V, without a learned value or output projection in the selection head. During training, each collage slot is a convex combination of source patches. This construction makes the learned routing policy transferable to the prediction model's own token embedding interface.

Learning without discrete sampling through the selection interface. The token collage C produced by Eq. 1 lacks source patch position information. However, directly choosing token indices is non-differentiable. To make the position information differentiable, we map the collage C back to the original history grid using the transpose of the same assignment matrix:

$$\widetilde{V} = A^\top C = A^\top AV, \qquad \widetilde{Z} = \mathrm{unpatchify}(\widetilde{V}). \tag{2}$$

This operation accumulates each slot's content at its source positions without a second softmax or column normalization. Positions receiving little assignment weight are attenuated. When assignments are one-hot and select distinct patches, it places the retained patches at their original locations and leaves other positions at zero. Reusing A for both reading and placement couples the selection decision to the reconstruction performance. In this way, gradients can flow through the soft-masked history $\widetilde{V}$ into the query and key features. No token-selection labels or policy-gradient estimators are required. The conversion of the softmax parameterization to discrete memory is specified in Section 3.3.

Self-Supervised Learning by Reconstruction We provide $\widetilde{Z}$ through the video diffusion model's inpainting interface and train it to reconstruct the history. Let $Z = E(H)$ denote the clean history latents. Following the flow-matching parameterization, we sample Gaussian noise ϵ and a noise

level $s \in (0, 1)$, and construct

$$Z_s = (1 - s)Z + s\epsilon,$$
$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{H,c,s,\epsilon}\left[w(s)\left\|D_\theta(Z_s, s; \widetilde{Z}_\phi, c) - (\epsilon - Z)\right\|_2^2\right], \tag{3}$$

where ϕ denotes the context encoder parameters, D_θ is the reconstruction model, and $w(s)$ is the loss weighting. We keep the VAE fixed and optimize the context encoder together with LoRA parameters in the pretrained video model. The reconstruction loss propagates through both the collage and its placement weights. It encourages the selector to allocate its limited slots to evidence most useful for recovering the history. Its usefulness for future generation is evaluated through downstream prediction rather than assumed to follow monotonically from reconstruction quality.

3.3 Token Collages as Memory for Future Generation

Sparse token conditioning. For each query slot, we select the history patch with the highest assignment weight and remove duplicates:

$$j_i^\star = \arg\max_{1 \le j \le N} A_{ij}, \qquad \mathcal{I} = \text{unique}\big(\{j_i^\star\}_{i=1}^B\big),$$
$$\mathcal{M}(H) = \{(v_j, p_j) : j \in \mathcal{I}\}. \tag{4}$$

The collage memory $\mathcal{M}(H)$ contains at most B original patches with their source positions. We embed them through the generator's patchify layer, retain associated controls, and concatenate their embeddings with noisy target tokens. Each memory token receives 3D RoPE based on its source coordinates. To adapt the generator to collage memory, we fine-tune it via LoRA, using the conditional flow-matching objective. For more details, refer to Appendix A.1.

Training-free frame-level conditioning. For models with frame-level memory banks (Sun et al., 2025), we rank eligible history frames by their query-slot allocations before token deduplication and retain the top F:

$$n_t = \sum_{i=1}^{B} \mathbb{I}\big[t_{j_i^\star} = t\big], \qquad \mathcal{F} = \text{TopK}_t(n_t; F), \tag{5}$$

where F is the frame budget. The selected full frames and their associated controls enter the generator through its existing conditioning interface, requiring no additional training.

Training-free streaming autoregressive generation. For autoregressive generators with a KV cache, the selector acts as an online retention and eviction policy (Algorithm 1). After chunk r is generated, its clean latent patches form N_r candidates $\mathcal{N}_r$ with absolute positions p_j. The context encoder and selection head produce queries $Q^{(r)} \in \mathbb{R}^{B \times d_k}$ and candidate keys K_j, yielding

$$\ell_{ij} = \frac{Q_i^{(r)} K_j^\top}{\tau\sqrt{d_k}}, \qquad i = 1, \dots, B, \quad j \in \mathcal{N}_r. \tag{6}$$

These logits are computed once and stored. Across rounds, we compare raw logits from the shared head rather than softmax probabilities, whose normalization depends on the candidate pool.

Each query slot selects its highest-scoring token from the previously retained tokens $\mathcal{I}_{r-1}$ and new candidates:

$$j_i^{(r)} = \arg\max_{j \in \mathcal{I}_{r-1} \cup \mathcal{N}_r} \ell_{ij}, \qquad \mathcal{I}_r = \{j_i^{(r)}\}_{i=1}^B. \tag{7}$$

A token is retained if it wins at least one slot and evicted otherwise, ensuring $|\mathcal{I}_r| \le B$. Retained tokens preserve their original logits; tokens already covered by the generator's protected sink and recent context $\mathcal{W}_r$ are skipped. Each round encodes N_r candidates and compares at most $B + N_r$ scores per slot, so its cost is independent of total video length.

4 Experiments

We evaluate the three applications in Section 3.3: sparse token conditioning for hand-controlled task, frame-level conditioning for camera-controlled task (Section 4.3), and KV-cache management for autoregressive text-conditioned generation (Section 4.4). Sections 4.1 and 4.5 present the shared setup and ablations, respectively.

Table 1: **Hand-controlled generation.** EgoDex (200 clips) and ARCTIC (40 clips).

	EgoDex (200)					ARCTIC (40)					
Method	PSNR↑	SSIM↑	LPIPS↓	DINO↑	VLM↑	PSNR↑	SSIM↑	LPIPS↓	DINO↑	VLM↑	Halluc.↓
Recency	21.63	0.837	0.253	0.873	37.7	15.89	0.683	0.373	0.823	43.3	108
Uniform	22.30	0.852	0.232	0.886	51.8	16.05	0.679	**0.355**	0.840	55.8	93
Random	22.32	0.852	0.231	0.889	51.2	**16.06**	**0.683**	0.362	0.841	57.5	82
TinyHistory	21.97	0.843	0.238	0.893	62.3	14.71	0.655	0.399	0.855	67.5	85
FoV	22.21	0.846	0.225	0.899	65.0	15.73	0.674	0.369	**0.867**	63.3	68
ReCollage	**22.46**	**0.854**	**0.222**	**0.900**	**69.8**	16.01	0.679	0.360	0.856	**75.0**	**64**

4.1 Common Setup

Context encoder and reconstruction training. The context encoder comprises 6 Transformer blocks matching the backbone's hidden size, followed by the selection head. We fine-tune the backbone via rank-256 LoRA, training one encoder per setting, varying only the data, backbone, and slot count. See Appendix A.2 for details.

Evaluation protocol. Across all three benchmarks, relevant content appears in the history, leaves the view, and must be recovered during generation. Uninformative prefix frames prevent fixed-budget memory from retaining the entire history. Within each benchmark, all methods share the generator, controls, and history, differing only in memory selection.

4.2 Hand-Controlled Generation

Setting. We train a 1,620-slot encoder by reconstruction on EgoDex (Hoque et al., 2026) and ARCTIC (Fan et al., 2023) using Hand2World (Wang et al., 2026b), then fine-tune the generator for sparse history conditioning (Section 3.3). Test clips from both datasets contain an object visible in the history, absent from the first target frame, and reappearing later. Each comprises 81 history and 81 target frames at 16 fps. We extend the history to 321 frames, randomly positioning it between repeated copies of its first frame and the target's first frame. Informative content thus occupies roughly one quarter of the prefix at an unknown location. See Appendix A.2.1 for details.

Baselines. Under matched memory budgets, we compare Uniform, Random, FoV (Sun et al., 2025), TinyHistory (Zhang et al., 2026), and Recency.

Metrics. We report frame-averaged PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and DINO (Caron et al., 2021) CLS cosine similarity to ground truth. Background and hand alignment can obscure object-identity errors, motivating a VLM consistency score following TinyHistory (Zhang et al., 2026). Claude-Fable-5.1 (Anthropic, 2026) compares shuffled, method-anonymous outputs with references, scoring the annotated object's reappearance and appearance/state/layout agreement. The score averages the three attributes upon reappearance and is zero otherwise; hallucinations are counted separately.

Results. ReCollage leads all EgoDex metrics (Table 1), substantially outperforming FoV and TinyHistory on VLM; Recency ranks last as relevant content leaves its window. On ARCTIC, ReCollage leads VLM with the fewest hallucinations, while FoV leads DINO but ranks third on VLM. Uniform and random selection achieve the highest PSNR with blurred, aligned outputs yet trail by 15–20 VLM points, distinguishing pixel fidelity from memory retention. Figure 4 shows FoV struggling with object-state changes under limited camera motion and TinyHistory misremembering appearance and position; ReCollage preserves both through informative token retention.

4.3 Camera-Controlled Generation

Setting. We train a 1,620-slot encoder by reconstruction on DL3DV (Ling et al., 2024) using Wan2.2-Fun-5B-Control-Camera (aigc-apps, 2026). From initial frames and captions of 100 test scenes, HY-WorldPlay (Sun et al., 2025) generates 637-frame revisit trajectories with FoV, which serve as references. All selection methods receive the same prefix and regenerate the revisit tail. ReCollage selects $F = 4$ memory frames via Eq. 5.

Baselines. *Uniform* and *random* pin four frames from the same history.



Figure 4: **Hand-controlled revisits** on three EgoDex clips.

Table 2: **Camera-controlled generation** on HY-WorldPlay.

Method	DINO↑	Hit↑	VLM↑	≤2.5↓
Uniform	0.941	0%	3.10	34
Random	0.963	69%	3.90	11
ReCollage	**0.974**	**91%**	**4.31**	**1**

Table 3: **Streaming text-conditioned generation** on LingBot-World v2.

Method	$\cos_t$ ↑	Margin↑	Correct↑	VLM↑
Deep Forcing	0.670	0.019	0.615	0.0
ToMe	0.702	0.015	0.580	3.0
MemRoPE	0.677	0.016	0.613	0.0
LongLive-RAG	0.683	0.024	0.641	3.0
ReCollage	**0.756**	**0.048**	**0.698**	**72.0**

Metrics. We report final-turn DINO similarity to references, the fraction of cases retaining the turned-away view (Hit), and blind 1–5 VLM scene-consistency ratings with counts scoring ≤ 2.5.

Results. ReCollage retains the turned-away view in 91% of cases versus Random's 69%, improving VLM ratings by 0.41 over Random and 1.21 over Uniform (Table 2). For more visual results, please refer to Figure 6.

4.4 Auto-regressive Text-Conditioned Generation

Setting. We train by reconstruction using Wan2.2-Fun-5B-InP (aigc-apps, 2026) on MiraData (Ju et al., 2024), OpenVid (Nan et al., 2025), and MiniMax H3-generated videos (MiniMax, 2026). Test prefixes comprise 81 frames of a subject, 81 of a distractor, and a 243-frame hold, totaling 405 frames. Each method generates 84 frames prompted to recall the original subject using LingBot-World v2 (Gao et al., 2026). ReCollage is equipped with the autoregressive policy (Section 3.3).

Baselines. We compare Deep Forcing (Yi et al., 2026), ToMe (Bolya et al., 2023), MemRoPE (Kim et al., 2026), and LongLive-RAG (Hu et al., 2026) (sink=1, memory=6, window=14).

Metrics. Using DINOv2 ViT-B/14 (Oquab et al., 2024), we compare generated subject crops with target and distractor crops. We report subject appearance frequency (*appear*), mean target similarity ($\cos_t$), target–distractor similarity margin ($\cos_t - \cos_n$), and the fraction of subject frames closer to the target (*correct*). VLM consistency follows Section 4.2; see Appendix A.2.3 for details.

Results. ReCollage achieves 69.8% correct subject identification versus 58–64% for competing policies, with the highest target similarity and twice the strongest target–distractor margin (Table 3). ToMe ranks second in target similarity but last in margin (0.015), suggesting plausible appearance

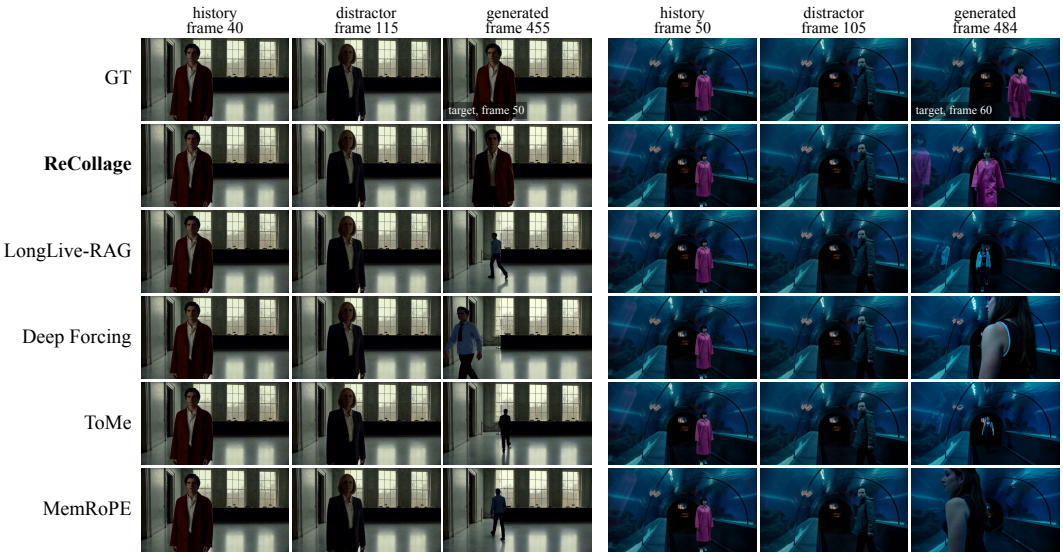


Figure 5: **Streaming recall** on LingBot-World v2 for two groups.

Table 4: **Ablations.** Left: EgoDex (100 clips) with the prediction model, prefixes, and metrics of Table 1; Right: the autoregressive benchmark of Table 3 (100 groups).

	Hand-controlled					Text-to-video			
Variant	PSNR↑	SSIM↑	LPIPS↓	DINO↑	VLM↑	$\cos_t$ ↑	Margin↑	Correct↑	VLM↑
ReCollage (one-shot, 1,620 slots)	22.63	0.855	0.213	0.901	82.0	0.751	0.046	0.684	51.0
chunk-wise selection	22.63	0.854	0.213	0.902	80.7	0.747	0.054	**0.734**	58.0
3,192 slots	–	–	–	–	–	**0.773**	**0.059**	**0.734**	**76.0**
798 slots	–	–	–	–	–	0.743	0.039	0.669	30.0

without faithful recall. Other policies often recover a person with incorrect clothing or facial details (Figure 5).

4.5 Ablation Study

Number of memory slots. Increasing chunk-wise memory slots from 798 to 1,620 to 3,192 raises VLM consistency from 30.0 to 58.0 to 76.0 and target–distractor margins from 0.039 to 0.054 to 0.059 (Table 4). Identification accuracy plateaus at 73.4% after rising from 66.9%, while target similarity keeps improving, suggesting finer appearance retention.

Chunk-wise versus one-shot selection. Chunk-wise selection encodes each chunk once without re-encoding memory (Eq. 7). Compared with one-shot encoding, EgoDex alignment metrics remain stable while VLM drops from 82.0 to 80.7. Autoregressive generation improves in margin (0.046→0.054), identification (68.4→73.4%), and VLM (51.0→58.0), despite lower target similarity (0.751→0.747). This may reflect avoiding length extrapolation: training clips span at most 321 frames, with few long examples, versus up to 470 frames at one-shot inference.

For more ablation studies, refer to the Appendix A.3.

5 Conclusion

ReCollage learns content-adaptive selection of native VAE tokens through self-supervised reconstruction and differentiable collage-and-scatter operations that preserve source positions. The frozen selector supports sparse token conditioning, frame-level memory, and streaming KV-cache management, improving recall after prolonged distractions or repetition across three generation tasks under matched budgets. Assignments sharpen during training, supporting reconstruction as effective supervision for memory selection. Limitations are discussed in Appendix A.5.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets, to provide critical ingredients for proving mathematical claims and to assist in the writing of proofs, to provide feedback on the research methodology, to design experiments, to implement methods, to assist with translation, to clean and reformat datasets, to support qualitative and thematic data analysis, and to interpret results. We have not used generative AI tools to design the research methodology, to help develop theoretical models or conceptual frameworks, to formulate mathematical claims, or to propose or refine hypotheses. We have reviewed all AI-assisted work. All AI-generated code was checked and tested by the authors. All AI-assisted writing, including translated and reformatted text and data, was checked against the originals. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

aigc-apps. VideoX-Fun: A video generation pipeline for diffusion transformer. https://github.com/aigc-apps/VideoX-Fun, 2026. Wan2.2-Fun-5B-Control and Wan2.2-Fun-5B-Control-Camera checkpoints.

Anthropic. System card: Claude Fable 5.1 & Claude Mythos 5.1. https://www.anthropic.com/claude-fable-5-1-mythos-5-1-system-card, September 2026. Model: Claude Fable 5.1, https://www.anthropic.com/claude/fable.

Jianhong Bai, Xiaoshi Wu, Xintao Wang, Xiao Fu, Yuanxing Zhang, Qinghe Wang, Xiaoyu Shi, Menghan Xia, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Kun Gai. SemanticGen: Video Generation in Semantic Space. *arXiv preprint arXiv:2512.20619*, 2025. URL https://arxiv.org/abs/2512.20619.

Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M. Patel. AdaMAE: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14507–14517, 2023. URL https://arxiv.org/abs/2211.09120.

Yuxuan Bian, Zeyue Xue, Songchun Zhang, Shiyi Zhang, Weiyang Jin, Yaowei Li, Junhao Zhuang, Haoran Li, Jie Huang, Haoyang Huang, Nan Duan, and Qiang Xu. Echo-Infinity: Learning Evolving Memory for Real-Time Infinite Video Generation. *arXiv preprint arXiv:2606.04527*, 2026. URL https://arxiv.org/abs/2606.04527.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.09461.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of Contexts for Long Video Generation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.21058.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2104.14294.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-Minute Video Generation with Test-Time Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17702–17711, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dalal_One-Minute_Video_Generation_with_Test-Time_Training_CVPR_2025_paper.html.

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. URL https://arxiv.org/abs/2605.31033.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2204.13662.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026. URL https://arxiv.org/abs/2607.07534. LingBot-World 2.0 technical report.

Saket Gurukar and Asim Kadav. Long-VMNet: Accelerating Long-Form Video Understanding via Fixed Memory. *arXiv preprint arXiv:2503.13707*, 2025. URL https://arxiv.org/abs/2503.13707.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.11709.

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. URL https://arxiv.org/abs/2606.02553.

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

Sunil Hwang, Jaehong Yoon, Youngwan Lee, and Sung Ju Hwang. EVEREST: Efficient masked video autoencoder by removing redundant spatiotemporal tokens. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2211.10636.

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. URL https://arxiv.org/abs/2605.09681.

Jiaxiu Jiang, Wenbo Li, Jingjing Ren, Yuping Qiu, Yong Guo, Xiaogang Xu, Han Wu, and Wangmeng Zuo. LoViC: Efficient Long Video Generation with Context Compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Findings Track*, 2026. URL https://arxiv.org/abs/2507.12952.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. MiraData: A large-scale video dataset with long durations and structured captions. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2407.06358.

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. URL https://arxiv.org/abs/2603.12513.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.16256.

Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. DVC: An End-to-end Deep Video Compression Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1812.00101.

MiniMax. MiniMax H3 (Hailuo 3.0) video generation model. https://www.minimax.io/blog/minimax-h3, 2026. First-and-last-frame conditioned video generation service (https://hailuoai.video) used to synthesize benchmark prefixes.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.02371.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. WorldPack: Dynamic Frame Compression for Long-context Video World Modeling. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2512.02473.

Wenqi Ouyang, Zeqi Xiao, Danni Yang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. TokensGen: Harnessing Condensed Tokens for Long Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2507.15728.

Michael S. Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: Adaptive Space-Time Tokenization for Videos. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/6a30e32e56fce5cf381895dfe6ca7b6f-Abstract.html.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards Long-Term Geometric Consistency for Real-Time Interactive World Modeling. *arXiv preprint arXiv:2512.14614*, 2025. URL https://arxiv.org/abs/2512.14614.

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

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.12602.

Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. LARP: Tokenizing Videos with a Learned Autoregressive Generative Prior. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.21264.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y. Chen, Yuqing Yang, and Bohan Zhuang. Latent Spatial Memory for Video World Models. *arXiv preprint arXiv:2606.09828*, 2026a. URL https://arxiv.org/abs/2606.09828.

Yuxi Wang, Wenqi Ouyang, Tianyi Wei, Yi Dong, Zhiqi Shen, and Xingang Pan. Hand2World: Autoregressive egocentric interaction generation via free-space hand gestures. *arXiv preprint arXiv:2602.09600*, 2026b. URL https://arxiv.org/abs/2602.09600.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-World: Scaling Interactive World Models to 1000-Frame Horizons via Pose-Free Hierarchical Memory. In *International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2602.02393.

Xindi Wu, Sven Elflein, James Lucas, Olga Russakovsky, Laura Leal-Taixé, Despoina Paschalidou, Jonathan Lorraine, and Aljoša Ošep. Addressable Memory for Video World Models. *arXiv preprint arXiv:2608.07408*, 2026b. URL https://arxiv.org/abs/2608.07408.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.12369.

Wilson Yan, Volodymyr Mnih, Aleksandra Faust, Matei Zaharia, Pieter Abbeel, and Hao Liu. ElasticTok: Adaptive Tokenization for Image and Video. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5e6cec2a9520708381fe520246018e8b-Abstract-Conference.html.

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. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2512.05081.

Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Ming Gong, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. URL https://aclanthology.org/2023.acl-long.73/.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as Memory: Scene-Consistent Interactive Long Video Generation with Memory Retrieval. In *SIGGRAPH Asia 2025 Conference Papers*, 2025a. URL https://arxiv.org/abs/2506.03141.

Jiwen Yu, Jianxiong Gao, Jianhong Bai, Yiran Qin, Kaiyi Huang, Quande Liu, Xintao Wang, Pengfei Wan, Kun Gai, and Xihui Liu. MemLearner: Learning to Query Context Memory for Video World Models. In *European Conference on Computer Vision*, 2026a. URL https://arxiv.org/abs/2606.31734.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Sri Siddarth Chakaravarthy P, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, and Animesh Garg. MosaicMem: Hybrid Spatial Memory for Controllable Video World Models. *arXiv preprint arXiv:2603.17117*, 2026b. URL https://arxiv.org/abs/2603.17117.

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, and Xiaojuan Qi. VideoSSM: Autoregressive Long Video Generation with Hybrid State-Space Memory. *arXiv preprint arXiv:2512.04519*, 2025b. URL https://arxiv.org/abs/2512.04519.

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

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. TinyHistory: Lightweight video history embeddings via two-stage context learning. *arXiv preprint arXiv:2512.23851*, 2026. URL https://arxiv.org/abs/2512.23851v6. Version 6.

Algorithm 1 ReCollage as a training-free streaming memory policy

Require: frozen context encoder with B query slots
$\mathcal{I}_0 \leftarrow \emptyset$
for chunk $r = 1, 2, \ldots$ **do**
  encode chunk r; obtain $\{Q_i^{(r)}\}_{i=1}^{B}$ and $\{K_j^{(r)}\}_{j=1}^{N_r}$
  $\ell_{ij} \leftarrow Q_i^{(r)} K_j^{(r)\top} / (\tau\sqrt{d_k})$ for all i and j; store ▷ new candidates are scored once
  $j_i^{(r)} \leftarrow \arg\max_{j \in \mathcal{I}_{r-1} \cup \mathcal{N}_r} \ell_{ij}$ for $i = 1, \ldots, B$ ▷ retained logits are reused
  $\mathcal{I}_r \leftarrow \{j_i^{(r)}\}_{i=1}^{B}$; discard the logits of $\mathcal{I}_{r-1} \setminus \mathcal{I}_r$ ▷ eviction
  $n_j \leftarrow |\{i : j_i^{(r)} = j\}|$ and $s_j \leftarrow \max_i \ell_{ij}$ for $j \in \mathcal{I}_r$
  $\mathcal{M}_r \leftarrow$ tokens of $\mathcal{I}_r \setminus \mathcal{W}_r$ in decreasing (n_j, s_j) order, up to M generator tokens
  generate chunk $r{+}1$ attending to the KV of sink $\|$ $\mathcal{M}_r$ $\|$ recent
end for

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*, 2018. URL https://arxiv.org/abs/1801.03924.

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. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02214.

A Appendix

A.1 Methods Details

To adapt the generator to sparse histories, we fine-tune it via LoRA, using the conditional flow-matching objective, as shown in the conditional flow-matching objective below:

$$\mathcal{L}_{\text{pred}}(\psi) = \mathbb{E}_{H,Y,c,s,\epsilon}\Big[w(s)\,\|G_\psi(Y_s, s; \mathcal{M}(H), c) - (\epsilon - Y_0)\|_2^2\Big], \tag{8}$$

where $Y_0 = E(Y)$, $Y_s = (1-s)Y_0 + s\epsilon$, and ψ denotes the trainable LoRA parameters. The loss is evaluated only on the future segment.

We detail the algorithms for streaming autoregressive video generation used in Section 3.3 in Algorithm 1.

A.2 Implementation details

Context encoder and reconstruction training. The context encoder consists of 6 Transformer blocks with the hidden size of the backbone, followed by the selection head. History latents come from the Wan2.2 VAE (Team Wan et al., 2025) (48 channels, $16\times$ spatial and $4\times$ temporal compression) and are patchified by the patch-embedding layer of the pretrained Wan2.2 backbone. Query slots are arranged on a 2-D grid with a rotary embedding over slot positions; the temperature is $\tau = 1$, and the temporal rotary embedding of the history is normalised to a fixed length so that clips of different durations share one positional range. The soft-masked history $\widetilde{Z}$ enters through the backbone's inpainting channels with an all-zero mask, and the backbone is adapted by LoRA (rank 256) as the reconstruction model D_θ. Optimisation uses AdamW with learning rate 10^{-4} and 48 clips per step on 4 GPUs. We train one encoder per setting, differing only in data, backbone, and slot count. *Hand-controlled*: $B = 30 \times 54 = 1{,}620$ slots, trained for 500 steps on EgoDex (Hoque et al., 2026) and ARCTIC (Fan et al., 2023) clips of 41–321 frames at 16 fps and 480p with the Wan2.2-Fun-5B-Control backbone of Hand2World (Wang et al., 2026b). *Camera-controlled*: $B = 30 \times 54 = 1{,}620$ slots, trained on DL3DV (Ling et al., 2024) with the camera-conditioned variant of the backbone (aigc-apps, 2026). *Auto-regressive text-conditioned*: $B = 2{,}430$

slots, trained for 300 steps on text-captioned clips of up to 321 frames from OpenVid (Nan et al., 2025), MiraData (Ju et al., 2024) and videos synthesised by MiniMax H3 (MiniMax, 2026) with the Wan2.2-Fun-5B-InP backbone. After training, each encoder is frozen.

A.2.1 HAND-CONTROLLED GENERATION

Setting. We select clips from the EgoDex (Hoque et al., 2026) and ARCTIC (Fan et al., 2023) datasets in which an object is visible in the history, absent from the first target frame, and returns later in the target. Each clip has an 81-frame history and an 81-frame target at 16 fps. We pad the history to 321 frames by placing the real history at a random position between copies of its first frame and copies of the target's first frame, so that the informative segment occupies roughly a quarter of the prefix at an unknown location.

Memory interface. We use the sparse token conditioning of Section 3.3 for this application. We fine-tune the prediction model G_ψ with LoRA initialised from the Hand2World control LoRA. Targets are 81 frames; histories contain 45–321 frames; each selected token receives its source rotary position, and its temporal positions are normalized within $[-81, -1]$.

Baselines. We compare against all the baseline methods at the same 1,620-token budget, including *Uniform*, *random*, *FoV*, and *Recency*. *FoV* and *Recency* use a frame-conditioned model trained in the same way. *TinyHistory* (Zhang et al., 2026) is re-implemented at the same budget with its two-stage training.

Metrics. We report PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018) and the DINO (Caron et al., 2021) CLS cosine similarity against the ground-truth target, averaged over frames. The background and the hand trajectory dominate these frame-level alignment metrics: a plausible but wrong object in the right place scores almost as well as the correct one. We therefore also report a VLM consistency score following TinyHistory (Zhang et al., 2026): a blind judge of Claude-Fable-5.1 (Anthropic, 2026) sees the reference frames and the outputs of all methods in a shuffled order and answers, *for the object named in the clip annotation, whether it re-appears (gating) and whether its appearance, state, and layout match*; the score is the mean of the three attribute answers when the gate passes, and zero otherwise, and hallucinated objects are counted separately.

Results. Table 1 reports the results. On EgoDex, ReCollage is best on every measure; on the VLM score, it is significantly ahead of the second tier, FoV and TinyHistory. Recency is last, confirming that the relevant content has indeed left the recent window. On ARCTIC, whose objects carry printed markings and articulated states, the picture sharpens: ReCollage obtains the best VLM score and the fewest hallucinations, while FoV drops to third despite the highest DINO. The alignment metrics disagree with the judges here: uniform and random tokens obtain the highest PSNR because they produce a blurred but well-aligned scene, while the VLM ranks them 15–20 points below ReCollage. We therefore treat PSNR as a fidelity measure and the VLM score as the memory measure, and note that TinyHistory's own evaluation also abandoned PSNR (Zhang et al., 2026). A further conclusion is that the geometric retrieval that dominates camera-controlled world models does not transfer to hand-object interaction, where the camera slightly moves, and the relevant change is the state of an object rather than a viewpoint.

Figure 4 shows representative revisits. TinyHistory's fixed temporal grid brings objects back but with the wrong appearance and position, whereas ReCollage, which copies native tokens, preserves them.

A.2.2 CAMERA-CONTROLLED GENERATION

Setting. We take the first frames and captions of 100 DL3DV test scenes, and drive HY-WorldPlay (Sun et al., 2025) along a yaw-only trajectory: turn by 108° (left or right, chosen per case), turn back, hold the original view for 17 s, then repeat the same turn (637 frames in total). The final turn therefore revisits exactly the views seen at the beginning, 23 s earlier. The stock FoV-retrieval model generates the complete video, which serves as the ground truth; every frame-level selection method then receives the history before the final turn as the prefix, and the tail is regenerated with the pinned memory replacing the retrieval slots.

Memory interface. The encoder is plugged into HY-WorldPlay through Eq. 5 with $F = 4$ pinned frames, keeping the model's 4 sink frames and 12-frame window.



Figure 6: **Camera-controlled revisits** on three DL3DV scenes with HY-WorldPlay.

Baselines. *Uniform* and *random* pin 4 frames from the same prefix through the same interface. FoV retrieval serves as the ground-truth reference.

Metrics. We report the DINO similarity between the regenerated and reference frames of the final turn, the fraction of cases in which at least one pinned frame shows the turned-away view (Hit), and a blind 1–5 VLM rating of whether the revisited view shows the same scene content as the reference, followed by the number of cases rated ≤ 2.5.

Results. Table 2 reports the results. Uniform spacing hardly pins a frame of the turned-away view, producing a second building or a different room (Figure 6). ReCollage pins such a frame in 91% of the cases against 69% for random draws, and improves the blind rating by $+0.41$ over random and $+1.21$ over uniform.

A.2.3 Auto-regressive Text-Conditioned Generation

Setting. We ask CLAUDE-Fable-5.1 to generate text prompts and synthesise prefixes with a first-and-last-frame video model (MiniMax, 2026): a content segment (81 frames) in which a person is present, a distractor segment (81 frames) with a different person, and a 243-frame hold on the last frame, for a 405-frame prefix; the model then generates 84 frames under a prompt that asks the original subject to return but without any detailed descriptions.

Memory interface. The context encoder of Section 4.1 is plugged into the four-step distilled LingBot-World v2 (Gao et al., 2026) with a 14-latent attention window (1 sink, 6 memory slots, 7 recent latents); each new chunk of 4 latents is encoded once, survivors fill the memory budget in rank order, and the retained tokens are uniformly grouped into 6 temporal buckets and assigned the corresponding temporal RoPE.

Baselines. We compare with the training-free KV-cache policies adaptable for this interface: Deep Forcing (Yi et al., 2026), ToMe (Bolya et al., 2023), MemRoPE (Kim et al., 2026) and LongLive-RAG (Hu et al., 2026), each run with the same memory budget and window.

Metrics. We embed the generated subject crops with DINOv2 ViT-B/14 (Oquab et al., 2024) and compare them with crops from the content segment (target) and the distractor segment. We report the fraction of frames in which a subject appears (*appear*), the mean similarity to the target ($\cos_t$), the margin $\cos_t - \cos_n$ over the distractor, and the fraction of subject frames closer to the target than to the distractor (*correct*). We also report a VLM consistency score with the protocol of Section 4.2. The blind judge sees reference frames and the outputs of every method in a shuffled order, and answers whether the target subject re-appears (gating) and whether its appearance (face, hair, clothing, and printed patterns) matches the target rather than the distractor.

Results. Table 3 reports the results. ReCollage recovers the correct subject in 69.8% of the frames against 58–64% for other KV-cache policies, with the largest target similarity and twice the margin over the distractor of the best policy. The margin is the informative quantity: ToMe reaches the second-highest target similarity, but its margin over the distractor is the smallest (0.015), i.e. it renders a plausible person rather than the remembered one. Qualitatively, other KV-cache policies bring back *a* person in the right place with the wrong clothing or face (Figure 5).

Table 5: **Ablations.** Left: EgoDex (100 clips) with the prediction model, prefixes, and metrics of Table 1; Right: the autoregressive benchmark of Table 3 (100 groups).

	Hand-controlled					Text-to-video			
Variant	PSNR↑	SSIM↑	LPIPS↓	DINO↑	VLM↑	$\cos_t$ ↑	Margin↑	Correct↑	VLM↑
ReCollage (one-shot, 1,620 slots)	22.63	0.855	0.213	0.901	82.0	0.751	0.046	0.684	51.0
chunk-wise selection	22.63	0.854	0.213	0.902	80.7	0.747	0.054	**0.734**	58.0
encoder trained w/o control	22.61	0.855	0.214	**0.905**	**83.7**	–	–	–	–
3,192 slots	–	–	–	–	–	**0.773**	**0.059**	**0.734**	**76.0**
798 slots	–	–	–	–	–	0.743	0.039	0.669	30.0
collage w/o source positions	22.50	0.853	0.226	0.889	65.7	–	–	–	–
learned tokens (5,000 slots, 1.3B backbone)	22.13	0.839	0.228	0.884	39.7	–	–	–	–

Table 6: **Soft versus hard selection** in reconstruction. The 1,620-slot hand-controlled encoder (step 500) reconstructs ten 81-frame EgoDex clips through the inpainting path, either from the soft-masked video scattered with the convex assignments ($A^\top AV$, as in training) or from the argmax tokens placed at their source positions (Eq. 4, as in every memory interface). Metrics are against the input clip; the last column counts the clips on which the row has the higher PSNR.

Selection	PSNR↑	SSIM↑	LPIPS↓	DINO↑	Higher PSNR (of 10)
Soft assignments ($A^\top AV$)	**26.56**	0.946	0.129	0.934	9
Hard selection (argmax, unique)	26.35	0.946	**0.118**	**0.939**	1

Why the baselines score near zero on the VLM measure. The VLM score is the product of the gating answer and the appearance answer. The gate passes for 98–100% of the groups with every policy, i.e. each of them brings a person back into the scene when the prompt asks for it; the appearance question, however, passes for at most 3 of 100 groups (Deep Forcing 0, ToMe 3, MemRoPE 0, LongLive-RAG 3, against 72 for ReCollage), so the product vanishes. The failure is not an absent subject but a substituted one. By the time of the recall prompt the target has been out of view for 324 frames, far beyond the 14-latent window, so its tokens have long left the recent context, and what each policy keeps instead does not carry the target's identity: Deep Forcing keeps the tokens that recent frames attend to, which during the empty hold are background; ToMe merges similar tokens, so the few target tokens are absorbed into the dominant scene tokens; MemRoPE averages evicted frames into memory tokens, which blurs a person seen for 81 frames into a 405-frame mean; and LongLive-RAG retrieves by similarity to the current, empty scene.

A.3 Ablations

Influence of control conditions. Training the encoder for 500 steps without reconstruction control inputs leaves EgoDex PSNR, SSIM, and LPIPS nearly unchanged, while DINO rises from 0.901 to 0.905 and VLM from 82.0 to 83.7 (Table 5), suggesting reconstruction alone suffices to learn informative memory selection.

Variation of context encoder. *Without source positions.* Feeding the collage directly to the generator without scattering discards source positions, producing nearly content-independent selection layouts and degrading all metrics at the same 1,620-slot budget. Its spatial selection profile is almost identical from clip to clip (mean pairwise correlation 0.997 against 0.19 for ReCollage; temporal profile 0.94 against 0.58; Table 7) and uncorrelated with motion ($r = -0.04$ against $+0.19$; Appendix A.4). On EgoDex, it is worse on LPIPS and DINO and loses 16.3 VLM points. *Learned tokens.* Replacing VAE tokens with 5,000 learned embeddings accessed through a dedicated context branch further degrades performance (Table 5), although it uses a smaller backbone and the comparison is indicative only.

Soft versus hard selection. Training uses the convex assignments of Eq. 1, whereas every memory interface uses its argmax (Eq. 4). Table 6 compares the two on ten 81-frame EgoDex clips: reconstructing from the argmax tokens instead of the soft-masked video lowers PSNR by 0.21 dB (soft is higher on 9 of 10 clips) but improves LPIPS from 0.129 to 0.118 and DINO from 0.934 to 0.939, with SSIM unchanged. The small PSNR loss is due to the residual softness of the assignments (Appendix A.4: a median of 2.2 effective source tokens per slot), which averages a few neighbouring

patches into a slightly smoother but less faithful token; the discrete memory keeps sharper, native tokens and is what all prediction results in the paper use.

A.4 Method Analysis

Nothing in Eq. 3 asks the assignments to be discrete, so we track during training what the encoder actually learns to do. Every 10 optimisation steps we read the assignment matrix A of the hand-controlled encoder on the current training clip and record, per slot, the largest assignment weight, the effective number of source tokens the slot averages over, $1/\sum_j A_{ij}^2$, and the effective number of frames it spreads over (Figure 2b).

Selection sharpens without supervision. At initialisation, every slot is a near-uniform average: the largest weight is 10^{-3}, a slot mixes about 7,000 source tokens and spreads over all frames of the clip. The slot statistics barely move during the first 100 steps and then sharpen within a narrow window: between steps 100 and 200 the effective number of tokens per slot drops from about 1,600 to under 3 and the effective number of frames from the whole clip to about 2, while the mean largest weight rises from 0.03 to 0.56. From step 240 on, the curves plateau at a largest weight of 0.64 (range 0.52–0.72 over the probed clips), 68% of slots holding a dominant token and a median of 2.2 effective source tokens per slot; the remaining fluctuations follow the length of the probed clip (6–78 latent frames), longer clips giving slightly softer rows. The assignments therefore converge to near-hard, spatially and temporally local selections, which is what the collage-and-scatter of Section 3.2 favours: a slot that spreads its weight scatters an attenuated mixture back to many positions, whereas a concentrated slot places one native token at full scale. This is also why the argmax conversion of Eq. 4 loses nothing (Appendix A.3).

What is selected. Once sharp, the retained tokens are not a fixed pattern: the per-slot layout differs from clip to clip (cross-clip correlation of the spatial profile 0.19 at the deployed checkpoint, see below), and the kept tokens cluster on the manipulated object and the hands rather than on the static background, and on the moments when the content is visible rather than on the padded or occluded part of the history (Figure 2a). The reconstruction objective alone thus produces a content-dependent, near-discrete selection of native tokens, the three properties Section 3.1 asks of a memory.

Selection layout across clips and its relation to motion. To quantify how much of the selection is content-dependent, we compute for every source token its coverage $c_j = \sum_i A_{ij}$, the expected number of slots that choose it, on six test clips per setting, and summarise it by the statistics of Table 7. *Spatial* and *temporal* cross-clip correlation: c is summed over time to give a spatial profile over the (h, w) grid, or over space to give a temporal profile over frames; each is normalised to sum to one, and we report the mean Pearson correlation between the profiles of any two clips. A value near 1 means the encoder places its slots in the same positions or at the same moments whatever the clip shows; a value near 0 means the layout is decided by the clip. *Motion*: the per-token Pearson correlation between c_j and $\|v_j - v_{j'}\|$, the change of the token relative to the token at the same position in the previous frame; *motion (prior removed)*: the same correlation after regressing out, from both variables, the spatial profile averaged over the other five clips, so that a fixed spatial prior cannot masquerade as a motion preference. *Accel.*: the correlation with the second temporal difference of the token, i.e. changes in motion. *Detail*: the correlation with the local spatial detail of a token, the deviation of its norm from the mean norm of its 3×3 spatial neighbours in the same frame. *Norm*: the correlation with the token's own norm $\|v_j\|$. All correlations are averaged over the six clips.

Three observations follow. First, for the ReCollage encoder, the spatial layout is largely clip-specific at every checkpoint (0.19–0.25), and the temporal layout, which shares a weak prior on short clips (about 0.5), becomes almost entirely clip-specific on 321-frame padded prefixes (0.08), where the informative segment sits at a different position in every clip; the same holds on ARCTIC (0.13 and 0.35). The chunk-wise rule reproduces this behaviour with slightly stronger cross-clip agreement (0.34 and 0.24), since every chunk is scored with the same query set. Second, coverage correlates positively but weakly with motion and acceleration ($+0.15$ to $+0.31$), and the correlation survives removing the spatial prior, so moving regions do receive more slots; the correlations with local detail and token norm are close to zero, so the selector is not simply keeping high-contrast or high-energy patches. Motion is therefore a correlate of what the selector keeps rather than its criterion. Third, the

Table 7: **Selection layout statistics** of the hand-controlled ReCollage encoder (1,620 slots) across training steps and inputs, and of the *collage w/o source positions* variant of Appendix A.3; six clips per row. Cross-clip: mean pairwise Pearson correlation of the spatial and temporal coverage profiles between clips (1 = same layout for every clip). Content: per-token Pearson correlation between coverage and the token's motion, motion with the cross-clip spatial prior regressed out (†), acceleration, local spatial detail and norm, averaged over clips. Chunk-wise: the 321-frame prefix is encoded in four 81-frame chunks with the streaming rule. The deployed checkpoint is step 500.

			Cross-clip		Content (per-token r with coverage)				
Encoder	Step	Input	Spatial	Temporal	Motion	Motion†	Accel.	Detail	Norm
ReCollage	300	EgoDex 81f	0.22	0.53	$+0.24$	$+0.21$	$+0.25$	$+0.07$	$+0.07$
ReCollage	400	EgoDex 81f	0.25	0.52	$+0.22$	$+0.20$	$+0.24$	$+0.05$	$+0.07$
ReCollage	500	EgoDex 81f	0.19	0.58	$+0.19$	$+0.18$	$+0.21$	$+0.05$	$+0.09$
ReCollage	300	EgoDex 321f padded	0.40	0.08	$+0.22$	$+0.18$	$+0.23$	$+0.07$	0.00
ReCollage	300	EgoDex 321f padded, chunk-wise	0.34	0.24	$+0.15$	$+0.13$	$+0.15$	$+0.05$	-0.01
ReCollage	300	ARCTIC 81f	0.13	0.35	$+0.27$	$+0.26$	$+0.31$	$+0.07$	$+0.08$
Collage w/o source positions	800	EgoDex 81f	0.997	0.94	-0.04	-0.03	-0.01	$+0.07$	$+0.26$

collage without source positions gives the opposite picture on every column: cross-clip correlations of 0.997 and 0.94, no relation to motion (-0.04) and a preference for high-norm tokens ($+0.26$), i.e. a fixed layout applied to every clip.

From soft to hard selection. The sharpening of Figure 2b is what makes the argmax conversion cheap. With a median of 2.2 effective source tokens per slot and 68% of slots holding a token with weight above 0.5, replacing the convex assignments by their argmax changes reconstruction by -0.21 dB in PSNR while improving LPIPS and DINO (Appendix A.3); the soft-masked history averages a few neighbouring patches into a smoother token, the hard one keeps a single native patch. Prediction, frame-level conditioning, and streaming all operate on the hard selection, so the encoder is deployed exactly as it was probed.

A.5 Limitations

ReCollage has three main limitations. First, the slot count is fixed during training and cannot automatically expand or shrink with historical content. Second, reconstruction learns to preserve informative history without knowing which evidence future generation will require; query-dependent recall may therefore benefit from combining ReCollage with retrieval. Third, training-free integration with existing autoregressive models remains limited by their ability to exploit sparse history, which they have not been explicitly trained to consume. For example, we observe that LingBot-World v2 can fail to faithfully recover an early view during a later revisit. A promising direction is to fine-tune a bidirectional teacher with ReCollage memory, then adapt and distill it into a memory-conditioned autoregressive generator using approaches such as Self Forcing (Huang et al., 2025) or Causal Forcing (Zhu et al., 2026). We leave adaptive slot budgets, retrieval integration, and memory-conditioned training and distillation to future work.