STREAMPILOT: AGENTIC INFERENCE FOR LONG-HORIZON STREAMING VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video generation models produce videos autoregressively. This leads to two challenges that all streaming video generation models share. The first is the *drift*: errors in the generated content accumulate until the picture collapses. The second is long-horizon consistency: as the video grows and the prompt keeps changing, the model does not remember what it generated, and may forget relevant content or recall the wrong content. We propose STREAMPILOT, a training-free agentic inference framework in which a VLM agent that reads frames fills the long-term cache during inference. At each instruction change and at fixed intervals, the agent reads the current instruction and the latest frames, decides which earlier chunk comes back as a reference and checks each new chunk before it is stored. References are stored as latents and recomputed where they are loaded. Without any training, STREAMPILOT achieves better long-horizon consistency than the base models on three streaming video models and four public benchmarks, showing that an agent can manage the memory of a streaming video model at inference time. On Causal Forcing, it raises the share of MemFlow clips showing the requested scene from 29.0% to 73.5% and cuts the videos that collapse after a scene returns from 58% to 13%. It also lets Self Forcing surpass LongLive in memory without any training, although LongLive is tuned for long videos.

1 INTRODUCTION

Streaming video generation is a continual forward process. A streaming video model generates a video as a sequence of chunks, each a short run of frames (Huang et al., 2025), and follows instructions that can change during generation (Yang et al., 2026a; Chen et al., 2026b). It never revises an emitted chunk, so each new chunk must agree with everything before it. This agreement comes from the key-value (KV) cache, which stores the attention keys and values of earlier chunks; we call these their features. Its short-term part, the window, holds the most recent chunks, and a hand-designed or trained rule fills its long-term part.

A long streaming video fails in two ways, drift and the loss of long-term memory, and both follow from how the video model is conditioned. It drifts because each chunk is conditioned on the model's own earlier chunks rather than the real histories it was trained on, so a small error is fed back and compounds (Huang et al., 2025; Zhang et al., 2025b; Li et al., 2026). It forgets because attention cost grows with history length (Shao et al., 2026), so the cache keeps only a short window, and a character or place that leaves it is lost (Huang et al., 2025; Yang et al., 2026a). Solving both at once is hard, because keeping more of the past brings content back but also carries its errors forward, and which earlier chunk matters depends on instructions that change during generation. Existing work either trains the model on long or self-generated histories (Cui et al., 2026; Yang et al., 2026a; Liu et al., 2026c; Li et al., 2026) or keeps selected earlier chunks in the cache: the initial chunk as an attention sink (Xiao et al., 2024; Yang et al., 2026a; Yi et al., 2026a), or chunks retrieved by a score over the instruction, the camera or the frames (Ji et al., 2025; Yu et al., 2025; Xiao et al., 2025; Yi et al., 2026b). These methods have two limitations: (i) one rule fills the long-term cache throughout the video, and (ii) a chunk brought back from stored keys and values keeps its old context and, unless shifted, its old position. LongLive recomputes its cache only at an instruction change, and MaineCoon restores a returning scene at its original positions (Yang et al., 2026a; Bai et al., 2026). Limitation (i) matters because instructions are open-ended: a return, a continued action, a new place or a person called by name each needs different earlier content, so a fixed rule suits some requests

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **An example of StreamPilot on a returning-scene script.** Frames are from LongLive-2.0-5B, one clip per column. A red dashed box marks a clip where the evaluator does not find the requested place, and the number is DINOv2 similarity to the scene returned to. The right panel gives *Correct*, the share of instruction changes at which a method stores a new place or retrieves a returning one, over 192 LongLive-2.0-5B runs; a fixed rule makes at most one kind of change, so it is correct only at changes of that kind. Here fixed rules hold one reference and StreamPilot two; Table 5 compares rules holding two. Appendix G gives the selection rule.

and fails others (Figure 1). A general framework needs a decision-maker that reads each instruction and what has been generated, as language-model agents do for a language model's limited context (Packer et al., 2023), so we let an agent decide what a streaming video model's long-term cache holds.

StreamPilot addresses both limitations during inference: an agent, not one rule, decides what the long-term part of the cache holds, and each chunk loaded there is recomputed at the current position. It is agentic because a pretrained vision-language model (VLM), the agent, makes the memory decisions, and training-free because neither the video model nor the agent is trained or aligned for the task (Liu et al., 2026a). StreamPilot stores earlier chunks in a memory bank as latents, the compact representation from which the video model decodes frames. At each instruction change and at fixed intervals, the agent reads the current instruction and the latest frames. It decides which stored chunks, the references, fill the slots of the long-term cache (Figure 2 compares these decisions with fixed rules). References are generated chunks, not external images, and let a scene shown earlier come back. Before a chunk enters the memory bank, a write check asks the agent whether it shows the requested content and the tracked identity. A chunk that keeps failing the check is stored but marked unreliable. Our contributions are summarized as follows.

- We cast long-term memory in streaming video as an online decision and propose StreamPilot, a training-free agentic inference framework whose VLM agent decides, as instructions change, which earlier chunks return and which new ones are stored.
- We introduce a reference-slot cache connecting these decisions to streaming video models with a sink and a window: slots between the two, filled from a memory bank that stores chunks as latents, 1,440 times smaller than keys and values on LongLive-1.3B; a loaded chunk is recomputed at the current position.
- Extensive experiments on three published streaming video models and four public benchmarks show that StreamPilot improves long-horizon consistency, e.g., raising the share of MemFlow clips with the requested scene from 29.0% to 73.5% on Causal Forcing 1.3B, lets Self Forcing 1.3B surpass LongLive-1.3B in memory without training, and, given the same references, keeps a returning person's identity more often than fixed rules.

2 Related Work

Streaming video models. In the streaming video models below, the long-term part of the cache follows a hand-designed rule. CausVid distills a bidirectional diffusion transformer into a causal student that generates chunk by chunk from a KV cache of earlier chunks (Yin et al., 2025). Self Forcing closes the gap between training and inference by training on histories the model generated itself, with a rolling cache of recent chunks (Huang et al., 2025). Causal Forcing initializes the causal

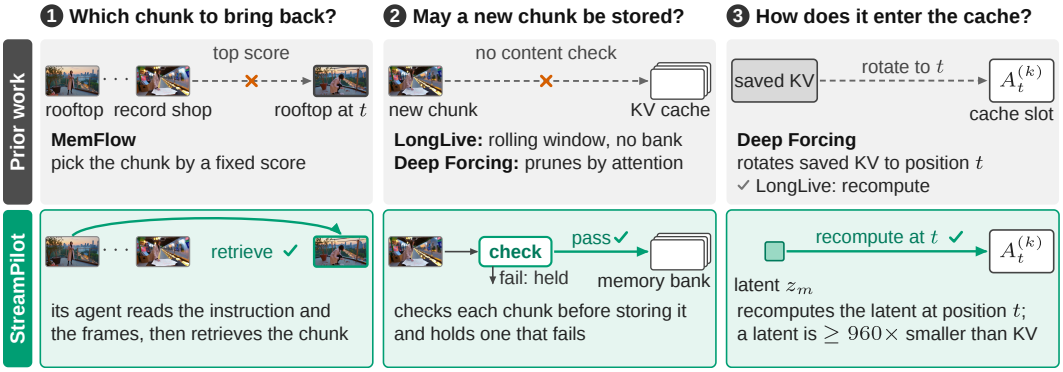

Figure 2: **How existing systems and StreamPilot fill the long-term memory.** Green marks StreamPilot, grey dashes mark fixed rules or rotated features, a grey tick marks a precedent, and red crosses mark failures. The ratio compares a stored latent with one chunk's cached features on LongLive-2.0-5B, the smaller of the two LongLive ratios.

student from an autoregressive teacher instead of the bidirectional one (Zhu et al., 2026). For longer videos, LongLive keeps the initial chunk as a sink beside the recent window, recomputes the cache when the instruction changes and is tuned on its own long rollouts (Yang et al., 2026a). LongLive-2.0 scales it to 5B parameters (Chen et al., 2026b). Deep Forcing keeps, without training, a deep sink of the opening tokens moved to the current position and prunes the rest by recent attention (Yi et al., 2026a). It aims to curb repetition, drift and slowing motion when one prompt runs far past the training length. StreamPilot changes neither their weights nor their attention mechanism.

Memory selection rules. Memories that bring content back select it by a hand-designed or trained score. MemFlow retrieves the chunks the next instruction attends to most and is fine-tuned with them (Ji et al., 2025). Mixture of Contexts selects chunks with a trained router (Cai et al., 2026). EM-Vid and Closing the Loop recall stored latents by entity or camera pose, the latter next to the recent window (Vandersanden et al., 2026; Ma et al., 2026). StreamPilot also stores latents and loads them at the current position, but its agent decides each slot and checks what it stores.

Agents around a streaming model. Language-model agents such as MemGPT manage memory outside a streaming cache, moving information between a limited context and external storage (Packer et al., 2023). Video agents plan scenes (Lin et al., 2024) or keep per-clip memories that they retrieve and write back, some after a check (Zhou et al., 2026; Long et al., 2026a), which a streaming video model cannot do to an emitted chunk. Recent real-time systems put an agent or an interactive loop around a streaming model. In MaineCoon, an agentic streaming inference framework for a 22B audio-visual model that generates past a thousand seconds, an agent plans the prompts and a VLM checks for identity drift (Bai et al., 2026). Its cache follows the plan by fixed policies. Vidu S1 takes voice instructions at any moment of an unbounded stream, and Vidu S2 adds references that can be updated at any moment (Zhang et al., 2026b;a). StreamPilot instead decides each slot while the stream is generated, checks a chunk before storing it and recomputes a loaded latent at the current position. We isolate this cache decision and test it on three published video models against fixed rules at equal capacity. Figure 2 compares earlier systems, and Appendix A surveys the rest.

3 StreamPilot

3.1 Preliminaries

A streaming video model generates chunk t from the current instruction p_t and its KV cache $\mathcal{C}_t$, where t counts chunks, not seconds. The model represents chunk t as a latent z_t, which on the 1.3B models has 3 latent frames that decode to 12 frames. Every video model we run partitions its cache as

$$\mathcal{C}_t = [\,S,\ A_t,\ W_t\,], \qquad A_t = [\,A_t^{(1)}, \ldots, A_t^{(K)}\,], \tag{1}$$

where S holds the features of the initial chunk, the sink, and W_t those of the most recent chunks, the window, which the video model updates. The long-term part A_t has K slots, each holding the per-layer keys and values of one *reference*, an earlier chunk that later chunks are con-

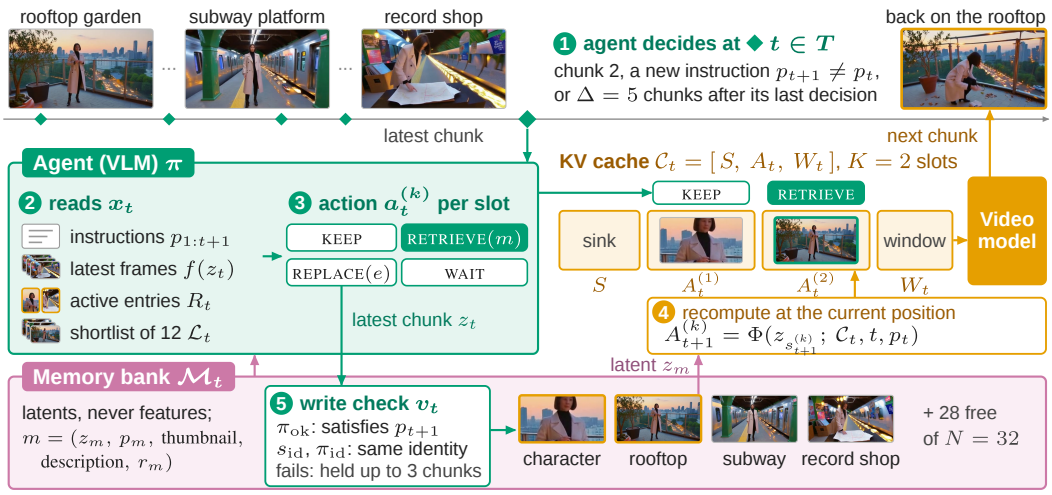

Figure 3: **Overview of StreamPilot.** At a decision chunk $t \in T$ (step 1) the agent reads the instructions, the latest frames and a shortlist of stored chunks (2) and picks an action for each reference slot (3). Retrieved chunks are recomputed at the current position by the video model Φ (4), and a chunk to be stored is checked first (5). The frames are from LongLive-2.0-5B.

ditioned on. References are drawn from the *memory bank* $\mathcal{M}_t$ of at most N entries $m = (z_m, p_m, \text{thumbnail}, \text{description}, r_m)$, where r_m is a reliability flag. Slot k holds the features of entry $s_t^{(k)} \in \mathcal{M}_t$, and $R_t = (s_t^{(1)}, \ldots, s_t^{(K)})$ lists the active entries. On the 1.3B models the sink and window span 12 latent frames as in LongLive, and each slot adds one chunk beside the sink, 18 latent frames with two slots. On LongLive-2.0-5B the first slot replaces a window chunk (Appendix C).

3.2 The Agent Decides What the Long-Term Memory Holds

Which chunk A_t should hold depends on p_t and on what has been generated. StreamPilot therefore asks the agent, denoted π and named in Sec. 4.1, to decide it at the chunks (Figure 3)

$$T = \{2\} \cup \{\, t : p_{t+1} \neq p_t \,\} \cup \{\, t : t - t_{\text{last}} \geq \Delta \,\}, \qquad \Delta = 5, \tag{2}$$

where t_{last} is the chunk of its previous decision and Δ the interval of periodic decisions. A decision at t follows chunk t and applies from chunk $t+1$, at an instruction change the first chunk of the new clip. At $t \in T$ the agent reads

$$x_t = \big(p_{1:t+1},\ f(z_t),\ R_t,\ \mathcal{L}_t\big), \qquad \mathcal{L}_t = \text{first}_{12}\big(R_t \oplus \text{sort}_{m \in \mathcal{M}_t \setminus R_t} \text{CLIP}(m, p_{t+1})\big), \tag{3}$$

where $f(z_t)$ is three frames decoded from chunk t and $\mathcal{L}_t$ is the shortlist, where $\oplus$ concatenates and first_{12} keeps the first 12. The shortlist ranks the other entries by the CLIP similarity (Radford et al., 2021) of their thumbnails to p_{t+1}, recency breaking ties. For each slot the agent answers

$$a_t^{(k)} \in \{\text{KEEP},\ \text{RETRIEVE}(m),\ \text{REPLACE}(e),\ \text{WAIT}\}, \qquad m \in \mathcal{L}_t, \quad k = 1, \ldots, K. \tag{4}$$

These actions keep the reference, put stored entry m into the slot, replace the reference with the new chunk as character or scene e, stored or new, once it passes the write check or was held for 3 chunks, or wait, which keeps the reference and records what is missing. In most 1.3B runs (Appendix C), a slot load decided at a periodic decision is deferred to the next instruction change, whose decision may replace it; chunks to store are checked and stored at once. The agent's prompt states the principles behind these answers: a character keeps its identity unless an instruction replaces it, a character or scene that returns takes its stored entry rather than the latest frames, a new entry is requested only once the frames show the new content, and keeping the references is the default (Appendix B gives all seven). The agent answers in JSON at temperature 0, and a call that still fails after two retries keeps the references. In the ablation of Sec. 4.3, fixed rules act on the same slots in place of π. Prompt only fills the slots with the first chunks of the latest clips, the t with $p_t \neq p_{t-1}$. Similar only fills them every Δ chunks with the chunks of highest $\cos(\bar{z}_m, \bar{z}_t)$ among every fourth

chunk at least 6 chunks old, where $\bar{z}$ is a latent averaged over frames and positions. Initial only keeps the opening chunk and Recent only the latest.

Write check. A chunk that REPLACE would store is written normally when

$$v_t = \pi_{\text{ok}}(f(z_t), p_{t+1}) \;\wedge\; \Big(m^* = \varnothing \;\vee\; \big[s_{\text{id}}(f(z_t), f(z_{m^*})) \geq \tau \wedge \pi_{\text{id}}(f(z_t), f(z_{m^*}))\big]\Big), \quad (5)$$

where m^* is the latest stored entry of e, or $\varnothing$ for a new one. In a separate call the agent gives two answers: π_{ok} that the chunk satisfies p_{t+1} with acceptable quality, so frames of the previous instruction are not stored for a new one, and π_{id} that it keeps that identity. The similarity s_{id} is the ArcFace face similarity (Deng et al., 2019) of decoded frames when both show a face and their DINOv2 similarity (Oquab et al., 2024) otherwise, with threshold τ of 0.15 for faces and 0.25 otherwise. A candidate that fails the check is held, and each of the next chunks is checked in its place, with a slot decision only at an instruction change. If none passes within 3 chunks, the candidate with the highest s_{id}, or the first for a new entry, is stored and loaded with r_m = unreliable, a flag the agent sees and eviction uses, so a check that never passes cannot stall the video. The two similarity thresholds were fixed before the evaluation on videos that no scored run uses. A changed appearance of a character is stored as a new entry beside the earlier one, and when $|\mathcal{M}_t| = N$ an entry that fills no slot is evicted (Appendix B lists every setting of STREAMPILOT).

3.3 MEMORY KEPT AS LATENTS AND RECOMPUTED AT THE CURRENT POSITION

Features computed when a chunk was generated belong to a position and a context that the video moves past, so STREAMPILOT stores latents and never features. When entry $s_{t+1}^{(k)} \in \mathcal{M}_t$ is loaded into slot k after chunk t, Φ runs one forward pass of the video model on its latent without diffusion noise. LongLive likewise recomputes its cache at an instruction change, and MaineCoon rebuilds it from clean latents (Yang et al., 2026a; Bai et al., 2026). Placed at the positions of chunk t and conditioned on p_t, the pass attends to the sink, the slots and the window as the video model's own pass over chunk t does. It writes the keys and values of every layer only to slot k, which holds

$$A_{t+1}^{(k)} = \Phi\big(z_{s_{t+1}^{(k)}};\, \mathcal{C}_t, t, p_t\big), \qquad k = 1, \ldots, K, \quad (6)$$

from chunk $t+1$ until the agent replaces the entry. Its features are thus those of the position, cache and instruction at which it was loaded. On the 1.3B models a reference is recomputed when it is loaded and then kept, and on LongLive-2.0-5B it is also recomputed from its latent every Δ chunks. The memory bank keeps at most $N = 32$ latents, less memory than the cached features of one chunk (Appendix B).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Base models. We run LongLive-1.3B (Yang et al., 2026a), Self Forcing 1.3B (Huang et al., 2025) and Causal Forcing 1.3B (Zhu et al., 2026) from their released weights through one inference harness, which generates 832×480 video with one setting for all three. The agent is Gemma 4 26B-A4B-it (Gemma Team, 2026). The *base model* is a video model run through the harness with no reference added, which is the released inference for LongLive-1.3B and the released checkpoint for the other two (Appendix C). Every comparison pairs the base model and STREAMPILOT on the same scripts and seeds, with one seed per script on the 1.3B models, and Appendix C lists the runs behind each table. Appendix K adds a fourth, larger model, LongLive-2.0-5B (Chen et al., 2026b), and Appendix N compares with Deep Forcing (Yi et al., 2026a).

Benchmarks. Each script is six instructions revealed one at a time, so its video has six clips, one per instruction, of about 9.9 s on the 1.3B models, and the benchmark names which earlier clip, if any, each clip should match. We use the 100 MemFlow scripts (Ji et al., 2025), where every later clip should match the first, and three public 64-script sets (Table C.2). In *continuous scenes* each clip should match the one before it. In *returning scenes* the story returns to a scene shown several clips earlier. In *new scenes* the same subject moves to a new place at every instruction, so no clip is named.

Table 1: **VBench scores of the base models and StreamPilot (%).** Higher is better except for imaging drop. Consistency is to the named clip. Bold marks the better of each pair, and † a significant difference. Table D.1 gives the full results.

			Against the named clip						
Video model	Benchmark	Method	Subject	Background	Flicker	Aesthetic	Imaging	Imaging drop	Worst
LongLive-1.3B (Yang et al., 2026a)	MemFlow	Base model	68.31	80.71	98.36	60.90	**72.82**†	−**2.31**	68.16
		+StreamPilot	**83.08**†	**88.49**†	**98.59**†	**62.03**†	71.65	−1.34	**68.72**
	Continuous	Base model	92.01	93.24	98.63	55.23	73.63	2.53	70.39
		+StreamPilot	**95.19**†	**96.34**†	**98.96**†	**58.11**†	**74.85**†	**0.44**†	**73.88**†
	Returning	Base model	77.77	84.01	98.75	59.96	72.54	−**0.18**	69.01
		+StreamPilot	**93.11**†	**93.72**†	**98.91**†	**61.80**†	**72.61**	0.26	**70.77**†
Self Forcing 1.3B (Huang et al., 2025)	MemFlow	Base model	65.36	77.76	**98.09**†	59.49	68.56	3.10	62.39
		+StreamPilot	**80.49**†	**87.51**†	98.01	**59.66**	**68.92**	**2.35**	**65.23**†
	Continuous	Base model	88.83	91.86	98.20	54.72	65.70	10.15	59.56
		+StreamPilot	**93.03**†	**94.23**†	**98.22**	**56.11**†	**69.94**†	**2.09**†	**67.67**†
	Returning	Base model	69.85	82.08	**98.20**	59.42	68.25	3.08	63.42
		+StreamPilot	**88.97**†	**91.42**†	98.19	**61.33**†	**69.47**†	**1.15**†	**66.50**†
Causal Forcing 1.3B (Zhu et al., 2026)	MemFlow	Base model	37.06	62.23	95.42	54.69	**70.59**	**2.57**	65.19
		+StreamPilot	**67.86**†	**77.68**†	**96.49**†	**58.94**†	70.39	2.96	**65.69**
	Continuous	Base model	71.24	84.28	94.53	49.17	67.60	10.09	60.68
		+StreamPilot	**82.23**†	**87.81**†	**96.78**†	**53.56**†	**72.33**†	**2.02**†	**68.94**†
	Returning	Base model	52.84	76.73	92.54	52.17	67.94	8.66	60.76
		+StreamPilot	**72.11**†	**83.66**†	**94.62**†	**58.87**†	**71.29**†	**3.43**†	**65.77**†

Metrics. We measure memory, scene switching and picture quality, and in the ablation what the agent holds (Table C.4 lists every measure). *Ref.* is the DINOv2 similarity (Oquab et al., 2024) of each frame of a clip to the closest of the last four frames of the named clip, averaged over the three returns on returning scenes. VBench (Huang et al., 2024) subject and background consistency are measured against the named clip, not the first and neighbouring frames (Table 1). A VLM evaluator, the agent's model called separately and shown neither the decisions nor the method, asks two questions per clip. *Scene* asks if the place is as requested (*Right* is *Scene* over the returns only), and *Identity* if the people and objects are those of the named clip or, on new scenes, fit the instruction. *Prev.* is the DINOv2 similarity to the scene just left, 1 and 2 s after a return (lower is better). Picture quality uses VBench's temporal flickering, aesthetic and imaging quality, the *imaging drop* from the first clip to the last, and the *worst clip*, a video's lowest imaging quality. A return *collapses* into noise when saturated pixels jump in its middle frame (Appendix H). *Held* is the share of returns whose first chunk has a slot holding a chunk of the named clip, and *Face* the ArcFace similarity of a returning face to the named person's (Appendix L).

Second evaluator and significance. Because the first evaluator shares the agent's model, a second evaluator from another model family, Qwen3-VL-8B-Instruct (Bai et al., 2025), rescores the 2,008 base-model and StreamPilot videos with the same questions (Table 3). A difference is *significant* († in tables) when a two-sided paired z-test over scripts finds that chance alone would give a difference this large less than 5% of the time. After Holm correction over the 56 pairs of Tables 2 and 4 plus *Prev.* at 4 s, StreamPilot's leads stay significant in 39 of 51 pairs, the base model's in 0 of 4; 1 is a tie (Table E.1).

4.2 Main Results

Video quality. Subject and background consistency with the named clip rise significantly wherever a clip is named (Table 1). Aesthetic quality rises in all nine cells, significantly in eight. On continuous and returning scenes, StreamPilot drifts less, significantly lowering the imaging drop and raising imaging quality and the worst clip on most video models, with no significant base-model lead. Table 4 counts collapses.

Memory. On MemFlow and returning scenes, StreamPilot significantly raises *Ref.*, *Scene*, *Identity* and *Right* (Table 2), and the second evaluator finds the same significant gains on all three 1.3B models (Table 3). On continuous scenes it significantly raises *Ref.*, and nothing in Table 2 significantly favours the base model. Added inside Deep Forcing's own inference, StreamPilot also significantly raises *Ref.*, *Scene* and *Identity* on MemFlow (Appendix N).

Visual results. Figure 4 shows two MemFlow scripts in which the base model loses the person of the first clip within a few instructions, while StreamPilot keeps that person in every later clip. Figure 1 compares StreamPilot with fixed rules on a returning-scene script: StreamPilot finds every requested place, and each rule shown misses at least one.

Table 2: **Does each clip bring back the earlier content its instruction asks for?** *Ref.* is the DINOv2 similarity of a clip to the earlier clip its instruction names. *Scene*, *Ident.* and *Right* are the shares of clips (%) in which the VLM evaluator sees the requested place, the named people and objects, and at a return the right earlier scene. Bold marks the better of each pair, and † a significant difference.

	MemFlow			Continuous scenes			Returning scenes				New scenes	
Method	Ref.	Scene	Ident.	Ref.	Scene	Ident.	Ref.	Scene	Ident.	Right	Scene	Ident.
LongLive-1.3B (Yang et al., 2026a)												
Base model	0.601	84.2	62.7	0.871	99.7	93.0	0.694	26.8	76.8	17.2	**47.7**	**96.9**
+StreamPilot	**0.756**†	**99.3**†	**88.8**†	**0.909**†	**100.0**	**97.9**†	**0.894**†	**35.7**†	**92.4**†	**37.5**†	44.0	95.8
Self Forcing 1.3B (Huang et al., 2025)												
Base model	0.590	75.0	57.7	0.838	99.0	**92.4**	0.612	26.6	68.8	18.8	33.9	93.8
+StreamPilot	**0.726**†	**98.7**†	**87.7**†	**0.877**†	**99.7**	**92.4**	**0.834**†	**36.2**†	**89.8**†	**38.0**†	**38.0**	**96.1**
Causal Forcing 1.3B (Zhu et al., 2026)												
Base model	0.278	29.0	26.3	0.607	54.9	63.5	0.397	39.3	61.5	29.7	54.9	**91.4**
+StreamPilot	**0.599**†	**73.5**†	**57.8**†	**0.731**†	**87.5**†	**65.4**	**0.603**†	**48.2**†	**81.3**†	**54.2**†	**60.7**†	90.1

Table 3: **The questions of Table 2 asked again by a second evaluator.** The first evaluator is the agent's own VLM, so a VLM of another model family (Sec. 4.1) rescores the same videos (%). +SP is StreamPilot. Bold marks the better of each pair, and † a significant difference.

	MemFlow				Returning scenes					
	Scene		Ident.		Scene		Ident.		Right	
Video model	Base model	+SP	Base model	+SP	Base model	+SP	Base model	+SP	Base model	+SP
LongLive-1.3B	91.8	**99.3**†	79.3	**96.0**†	30.2	**36.5**†	83.9	**93.5**†	22.9	**38.0**†
Self Forcing 1.3B	81.0	**99.3**†	70.0	**92.3**†	24.2	**36.5**†	72.9	**90.6**†	12.5	**38.0**†
Causal Forcing 1.3B	33.8	**88.3**†	33.0	**69.2**†	32.0	**37.2**†	64.6	**77.3**†	14.1	**29.2**†

Scene switching, collapse and longer videos. One second after a return, the frames resemble the scene just left less with StreamPilot than with the base model (Table 4, top). Without StreamPilot, 29 and 37 of the 64 videos on Self Forcing 1.3B and Causal Forcing 1.3B collapse into noise after a return; with it, 0 and 8 do. On LongLive-1.3B scripts chained to one, two and four minutes, StreamPilot keeps the higher *Ref.* (Table 4, bottom; chains in Appendix J).

Training-free against trained. Self Forcing 1.3B and LongLive-1.3B are both built on Wan2.1-T2V-1.3B (Wan et al., 2025), and only LongLive-1.3B is tuned on its own 60 s rollouts (Yang et al., 2026a). With StreamPilot and nothing trained, Self Forcing 1.3B significantly surpasses the LongLive-1.3B base model in memory and consistency with the named clip on MemFlow and returning scenes, and in one-second *Prev.* at returns (Tables 1–4; Appendix I).

Table 4: **Scene switching, collapse and longer videos.** Top, returning scenes, where lower is better; collapse counts the videos and returns that turn into noise. Bottom, LongLive-1.3B on 16 chains of one, two and four scripts and 21 of three for the return 180 s back, where higher is better. Bold marks the better of each pair, and † a significant difference (exact tests for collapse).

	Prev. after a return		Collapse at returns	
Method	1 s	2 s	Videos (of 64)	Returns (%)
LongLive-1.3B (Yang et al., 2026a)				
Base model	0.903	0.862	4	3.1
+StreamPilot	**0.863**†	**0.841**	**0**	**0.0**
Self Forcing 1.3B (Huang et al., 2025)				
Base model	0.881	0.829	29	32.3
+StreamPilot	**0.801**†	**0.774**†	**0**†	**0.0**†
Causal Forcing 1.3B (Zhu et al., 2026)				
Base model	0.699	0.603	37	29.7
+StreamPilot	**0.548**†	**0.500**†	**8**†	**4.2**†

	Longer videos, Ref.			Return, 180 s	
LongLive-1.3B	60 s	120 s	240 s	Ref.	Right
Base model	0.712	0.684	0.642	0.458	17.5
+StreamPilot	**0.874**†	**0.838**†	**0.825**†	**0.795**†	**34.9**†

4.3 Ablation Study and Inference Cost

Table 5 fills the same two slots by rules, each built for one kind of request, in place of the agent. On LongLive-1.3B returning scenes, the slots raise *Ref.* and both VBench consistencies with the named clip over the base model even when a fixed rule fills them, and the agent's choice raises all three

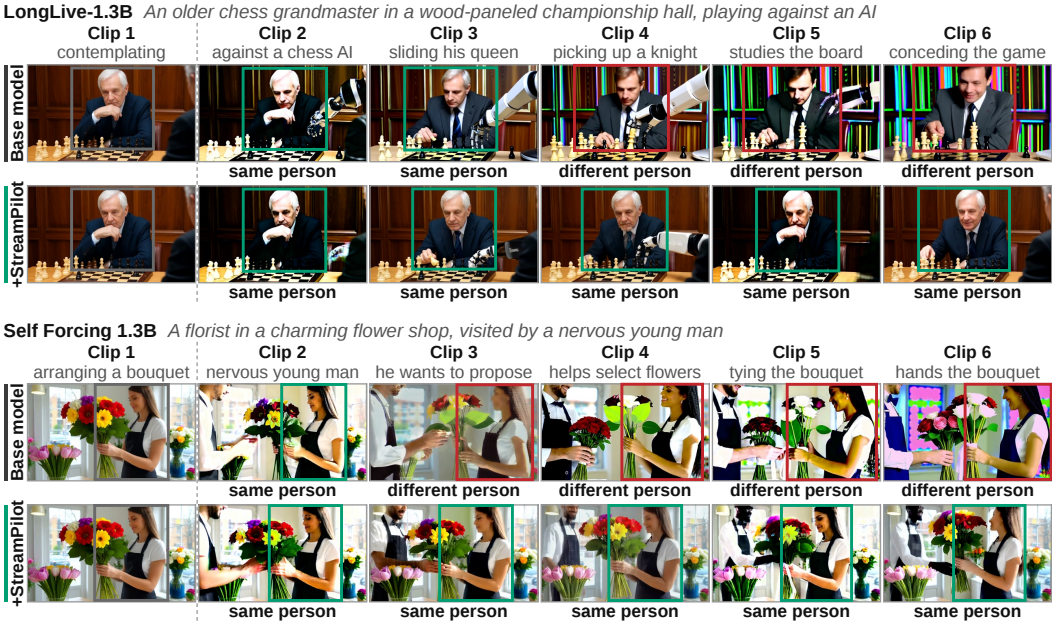

Figure 4: **Two MemFlow scripts with the base model and StreamPilot.** Top, LongLive-1.3B; bottom, Self Forcing 1.3B; both rows of a script share its seed. Frames are clip midpoints, and headings quote each clip's instruction. The grey box marks the person of clip 1; later boxes mark the most similar face, green if the evaluator marks the same person, red if not. Appendix G gives the rule that chose them.

Table 5: **Ablation on LongLive-1.3B.** Top, a fixed rule of Sec. 3.2 or a retrieval rule replaces the agent: the CLIP ranking loads the two stored chunks whose thumbnails best match the instruction, text matching the clips whose instructions BM25 or MiniLM (Wang et al., 2020) ranks closest, and with MiniLM the agent still decides what is stored. *Subj.* and *Backg.* are VBench consistency with the named clip (Table 1), *Ident.*$_2$ is *Ident.* by the second evaluator, *Who* the share at which the evaluator picks the named person among three or none (25% by chance), and *Face* ArcFace similarity ($\times 100$) to that person. Initial only keeps one reference on returning scenes and two on returns by name. N/A means not applicable or not run, and † a significant difference from StreamPilot.

	Returning scenes				Returns by name				
Method	Ref.	Subj.	Backg.	Held	Held	Ident.	Ident.$_2$	Who	Face
Replacing the agent									
Base model	0.694†	77.77†	84.01†	N/A	N/A	6.8†	23.4†	10.4†	16.2†
+StreamPilot	0.894	93.11	93.72	92.2	99.5	79.2	95.3	45.8	34.1
w/o agent (CLIP ranking, $K=2$)	0.849†	89.49†	90.64†	47.9†	30.2†	47.9†	75.5†	27.1†	25.2†
w/o agent (BM25 text match, $K=2$)	0.853†	89.25†	90.51†	100.0†	89.1†	47.4†	81.3†	34.4†	25.8†
agent stores, MiniLM loads ($K=2$)	0.879†	92.28†	92.79†	100.0†	99.0	58.3†	87.5†	33.3†	30.2†
w/o agent (Similar only, $K=2$)	0.845†	89.58†	90.85†	27.1†	37.5†	43.8†	82.8†	26.0†	22.5†
w/o agent (Prompt only, $K=2$)	0.810†	86.15†	88.29†	0.0†	0.0†	32.3†	71.9†	28.1†	22.9†
w/o agent (Initial only)	0.813†	87.70†	89.59†	33.3†	33.3†	29.7†	68.2†	19.8†	24.6†
Removing one part at a time, same scripts and seed									
w/o write check	0.890	92.49†	93.37†	33.9†	N/A	N/A	N/A	N/A	N/A
w/ one reference	0.815†	87.40†	89.95†	62.5†	N/A	N/A	N/A	N/A	N/A

significantly above every rule, text matching included. Text matching holds the named clip at every return yet trails. Other rules reach a *Held* of at most 47.9%, against 92.2% for the agent. Without the write check *Held* falls to 33.9%, as the agent stores the old instruction's last chunk. One reference lowers *Ref.* and *Held*. On LongLive-2.0-5B, recomputing a reference instead of rotating its keys keeps the final face nearer the first in all 8 videos and raises the three VBench scores (Table 6).

Returns by name. We also test returns where a name, not a place, links a return to its clip. On LongLive-1.3B, each of 64 scripts introduces three people of one gender, one per clip, each named once, and returns to each by the first name alone (Appendix L). The agent holds the named person's clip at 99.5% of returns, as often as MiniLM text matching. The returning person then matches that

Table 6: **Recomputing a stored reference against rotating its saved keys, on LongLive-2.0-5B.** The opening chunk is kept as a reference through 120 s videos (4 prompts at 2 seeds each) and is never updated, has its saved keys rotated to the current position, or is recomputed there from its latent, as in StreamPilot. *Final* is ArcFace similarity of the last eighth's faces to the first eighth's, *Drop* its fall, *Time* the share of generation time; bold marks the best, and † significantly below recomputation (sign-flip test over the 8 videos).

Update	Final	Drop	Subject	Background	Imaging	Time (%)
No update	0.727†	0.183	0.967	0.960	67.20	0.1
Rotate saved keys to t	0.761†	0.153	0.965	0.961	66.85	0.1
Recompute at t	**0.831**	**0.083**	**0.974**	**0.965**	**67.86**	1.8

Table 7: **Extra time cost of StreamPilot.** Times are seconds per returning-scene video of 80 chunks, from runs that block on every decision and share the agent among three to five videos.

Measure	LongLive-1.3B	Self Forcing 1.3B	Causal Forcing 1.3B
Base model generation and the agent's decisions			
Base model, per video (s)	89.0	79.4	80.7
Base model, per chunk (s)	1.11	0.99	1.01
Decisions per video	18	18	18
Model calls per video, with write checks	32.0	32.3	32.5
One decision of the agent, mean (s)	2.80	2.25	2.22
Generation between two decisions, mean (s)	4.95	4.41	4.49
Added by StreamPilot *on the generating GPU*			
Reference recomputations per video	7.4	7.7	7.5
One recomputation (s)	0.32	0.22	0.21
Reading the agent's frames, per video (s)	5.9	5.9	5.9
Attention and other overhead, per chunk (s)	0.548	0.358	0.335
Blocked on the agent's answer, per video (s)	55.7	43.1	42.7

clip in 79.2% of returns, against at most 58.3% for the rules not told the named clip, text matching included. The other identity measures agree: 95.3% against at most 87.5% by the second evaluator, 45.8% against at most 34.4% by *Who*, and a *Face* of 34.1 against at most 30.2 (Table 5). On Self Forcing and Causal Forcing 1.3B the agent keeps the named person in 55.2% and 67.2% of returns, against at most 43.2% and 55.7% for the rules, and leads every rule in *Face* (Table L.3).

Inference cost. The agent runs on its own GPU and serves three to five videos at once, in 2.22 to 2.80 s per decision on average, shorter than the 4.41 to 4.95 s of generation between decisions (Table 7). On the generating GPU StreamPilot adds 34.2 to 52.2 s to videos the base models generate in 79.4 to 89.0 s.

5 Conclusion, Limitations and Future Work

StreamPilot makes what a streaming video model remembers an inference-time decision that can be improved separately from the video model. An agent decides it while the video is generated, and latents recomputed where they are loaded bring earlier content back without training either model. One prompted agent serves every benchmark and video model here, with no rule chosen for each. Each fixed rule is built for one kind of request, such as a return to an earlier place or a person called back by name, and even text matching, which holds the named clip, brings it back less faithfully; the agent reads each instruction, which is why a general framework needs it (Figure 1, Table 5).

Limitations. The agent is not always right: when the instruction moves to a new place, its references can keep old content. Each decision takes seconds of a VLM, which can run beside generation, and its request grows with the instructions. Our first evaluator is the agent's model (Panickssery et al., 2024); the second and VLM-free measures confirm its gains on MemFlow and returning scenes.

Future work. As multimodal LLMs improve, agentic inference could become a general paradigm for streaming video, with more accurate memory decisions. StreamPilot gives the community a framework for translating stronger agents into better video consistency without retraining video models. The same agent could also decide when to leave a slot empty, plan upcoming instructions and keep a compact history for hour-long streams. StreamPilot can be added to future streaming video models that keep a sink and a window, adapting its settings to each, as for LongLive-2.0-5B, whose first slot replaces a window chunk and whose references are recomputed every Δ chunks.

Reproducibility Statement

All experiments use publicly released models, with the settings given in Sec. 4.1 and Appendices B and C. The supplementary material contains the code of StreamPilot, the agent's prompts and the scripts that generate and score the benchmark videos.

Use of Language Models

Language models assisted writing and code development. Within the method, a pretrained vision-language model is the agent, and the same pretrained model, with a separate prompt, is the first evaluator (Sec. 4.1). The authors are responsible for the final text, implementation and interpretation of the results.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. doi:10.48550/arxiv.2310.11511.

Lichen Bai, Tianhao Zhang, Shitong Shao, Dingwei Tan, Qiyu Zhong, Zhengpeng Xie, Haopeng Li, Qinghao Huang, Dandan Shen, Tengjiao Ji, Wei Wang, Peicheng Wu, Yuxuan Zhao, Xiangyu Zhu, Welly Luo, Shurui Yang, and Zeke Xie. MaineCoon: Pursuing a real-time audio-visual social world model. *arXiv preprint arXiv:2606.17800*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2407.01392.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026a. arXiv:2602.06028.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, and Weian Mao. LongLive-2.0: An NVFP4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026b.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Yusuf Dalva and Pinar Yanardag. AdaState: Self-evolving anchors for streaming video generation. *arXiv preprint arXiv:2605.30349*, 2026.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. doi:10.1109/cvpr.2019.00482; arXiv:1801.07698.

Qijun Gan, Chenwei Zhang, Meiguang Jin, Junfeng Ma, and Qiu Shen. TaoMate: Anchor-guided memory bridging evolving and reference states for real-time audio-video digital human generation. *arXiv preprint arXiv:2607.24359*, 2026.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. doi:10.1109/cvpr52734.2025.00245.

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 *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.08009.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Jin Qingyang, 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 (CVPR)*, 2024. doi:10.1109/cvpr52733.2024.02060.

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

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. doi:10.18653/v1/2023.emnlp-main.495.

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.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.09212.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent multi-scene video generation via LLM-Guided planning. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024. doi:10.48550/arxiv.2309.15091; arXiv:2309.15091.

Buhua Liu, Shitong Shao, Bao Li, Lichen Bai, Zhiqiang Xu, Haoyi Xiong, James T. Kwok, Sumi Helal, and Zeke Xie. Alignment of diffusion models: Fundamentals, challenges, and future. *ACM Computing Surveys*, 2026a. doi:10.1145/3796982.

Jinzhuo Liu, Jiangning Zhang, Wencan Jiang, Yabiao Wang, Dingkang Liang, Zhucun Xue, Ran Yi, and Yong Liu. Advancing narrative long video generation via training-free identity-aware memory. *arXiv preprint arXiv:2605.18733*, 2026b.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026c.

Do Xuan Long, Yale Song, Min-Yen Kan, Tomas Pfister, and Long T. Le. A^2RD: Agentic autoregressive diffusion for long video consistency. *arXiv preprint arXiv:2605.06924*, 2026a.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan O Arik. VISTA: A test-time self-improving video generation agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b. arXiv:2510.15831.

Jiayi Luo, Qiyan Liu, Tengyang Wang, JunHao Liu, Jiayu Chen, Cong Wang, Hanxin Zhu, Chen Gao, Xiaobin Hu, Qingyun Sun, and Zhibo Chen. Future forcing: Future-aware training-free KV cache policy for autoregressive video generation. *arXiv preprint arXiv:2605.30083*, 2026.

Wenchao Ma, Changran Liu, Sharon X. Huang, and Haomiao Jiang. Closing the loop: Training-free revisit consistency for autoregressive generative rendering. *arXiv preprint arXiv:2607.21848*, 2026.

Yu Meng, Xiangyang Luo, Letian Li, Wenyuan Jiang, Chen Gao, Xinlei Chen, Yong Li, and Xiao-Ping Zhang. TetherCache: Stabilizing autoregressive long-form video generation with gated recall and trusted alignment. *arXiv preprint arXiv:2606.13035*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy 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 Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. arXiv:2304.07193.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2024. oral; doi:10.52202/079017-2197.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, 2023. doi:10.1145/3586183.3606763.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Shitong Shao, Lichen Bai, Pengfei Wan, James Kwok, and Zeke Xie. Efficient video diffusion models: Advancements and challenges. *arXiv preprint arXiv:2604.15911*, 2026.

Jente Vandersanden, Matheus Gadelha, Chun-Hao P. Huang, Hyeonho Jeong, and Yulia Gryaditskaya. EM-Vid: Training-free entity-centric memory for efficient and consistent multi-shot video generation. *arXiv preprint arXiv:2605.23610*, 2026.

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, 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.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020. doi:10.48550/arxiv.2002.10957.

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

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

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.17453.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.12369.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025. doi:10.52202/085713-0593.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yukang Chen, and Song Han. LongLive: Real-time interactive long video generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026a. arXiv:2509.22622.

Yang Yang, Tianyi Zhang, Wei Huang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Anchor forcing: Anchor memory and tri-region RoPE for interactive streaming video diffusion. *arXiv preprint arXiv:2603.13405*, 2026b.

Bo Ye, Xinyu Cui, Jian Zhao, Tong Wei, and Min-Ling Zhang. DySink: Dynamic frame sinks for autoregressive long video generation. *arXiv preprint arXiv:2605.21028*, 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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 *Proceedings of the International Conference on Machine Learning (ICML)*, 2026a. arXiv:2512.05081.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. WorldKV: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.07772; doi:10.1109/CVPR52734.2025.02138.

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 *Proceedings of the SIGGRAPH Asia Conference Papers*, 2025. arXiv:2506.03141; doi:10.1145/3757377.3763833.

Jintao Zhang, Kai Jiang, Jintao Chen, Xu Wang, Deyuan Liu, Jungang Li, Dechuang Chen, Ming Lin, Jingjiang Zhou, Haopeng Jin, Qi Jia, Xiaohang Wang, Yaole Wang, Zhanqiang Zhang, Ran Li, Zhengkun Huang, Shuyue Xiong, Yuji Wang, Zikun Dai, Hui He, Yang Luo, Mang Ning, Weiqi Feng, Chengyang Ye, Xinyue Lin, Min Zhao, Hongzhou Zhu, Hengkai Tan, Zeyuan Wang, Chendong Xiang, Kaiwen Zheng, Zhijie Deng, Fan Bao, Jianfei Chen, and Jun Zhu. Vidu s2: Real-time interactive, editable, and spatial video generation. *arXiv preprint arXiv:2609.11638*, 2026a.

Jintao Zhang, Kai Jiang, Jintao Chen, Xu Wang, Yang Luo, Yuji Wang, Dechuang Chen, Jungang Li, Chengyang Ye, Marco Chen, Hongzhou Zhu, Min Zhao, Yuxuan Jiang, Zhengkun Huang, Chendong Xiang, Kaiwen Zheng, Haoxu Wang, Xiaohang Wang, Qi Jia, Xin Chen, Yimin Chen, Youhe Jiang, Fangcheng Fu, Zhijie Deng, Fan Bao, Jianfei Chen, and Jun Zhu. Vidu s1: A real-time interactive video generation model. *arXiv preprint arXiv:2607.03118*, 2026b.

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

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025b. doi:10.52202/085713-1024.

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*, 2026a.

Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed KV-Memory for consistent long video generation. *arXiv preprint arXiv:2603.21366*, 2026b.

Zhuoran Zhao, Shengju Qian, Tongtong Liang, Xianghao Kong, Songchun Zhang, Junchao Huang, Guian Fang, Xin Wang, Pan Hui, and Anyi Rao. Mask forcing: Improving autoregressive video diffusion distillation via dual-noise masking rollout. *arXiv preprint arXiv:2609.09123*, 2026c.

Jinsong Zhou, Yihua Du, Xinli Xu, Luozhou Wang, Zijie Zhuang, Yehang Zhang, Shuaibo Li, Xiaojun Hu, Bolan Su, and Ying-cong Chen. VideoMemory: Toward consistent video generation via memory integration. *arXiv preprint arXiv:2601.03655*, 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. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.02214.

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, Haoran Li, Haoyang Huang, Zeyue Xue, and Nan Duan. Self gradient forcing: Native long video extrapolation. *arXiv preprint arXiv:2607.20368*, 2026.

APPENDIX
STREAMPILOT: Agentic Inference for Long-Horizon Streaming Video Generation

Contents

A Related Work in Detail . . . 15
B The Agent: Settings, Workflow and Prompts . . . 16
C Benchmark Reproduction Details . . . 18
D Video Quality in Detail . . . 20
E Holm Correction . . . 21
F Content Rate and the Second Evaluator . . . 21
G The Scripts of the Qualitative Figures . . . 21
H Leaving the Previous Scene and Collapse at Returns . . . 22
I Training-Free against Trained LongLive-1.3B . . . 23
J Longer Videos of the Same Scripts in Detail . . . 23
K LongLive-2.0-5B as a Fourth Video Model . . . 24
L What the Agent Holds and Returns by Name . . . 24
M Ablation on All Recorded Metrics . . . 26
N Comparison with Deep Forcing . . . 26

A RELATED WORK IN DETAIL

This section adds detail to the methods of Sec. 2 and covers the rest, giving for each what is stored, how it decides what returns to the cache and whether the weights change.

Streaming video models and their caches. The cache of an autoregressive video diffusion model grows with the video (Yin et al., 2025; Huang et al., 2025). Released designs keep the initial chunk with a window of recent chunks (Xiao et al., 2024; Yang et al., 2026a), prune or merge tokens (Yi et al., 2026a; Luo et al., 2026), summarize older chunks into memory tokens (Kim et al., 2026; Gan et al., 2026), or train the video model to stay stable on its own outputs (Zhuang et al., 2026; Li et al., 2026; Chen et al., 2026a; Zhao et al., 2026c). Diffusion Forcing trains a causal model with a separate noise level per frame (Chen et al., 2024), and Rolling Forcing denoises a rolling window jointly and keeps the opening frames as a sink (Liu et al., 2026c). Self-Forcing++ trains the student with teacher guidance on short spans of its own long videos to reach minutes (Cui et al., 2026), and StreamingT2V keeps scene and object features of the first chunk beside features of the previous one (Henschel et al., 2025). Beyond our video models, only Deep Forcing's memory policy is compared (Appendix N); the peer-reviewed ∞-RoPE, which without training rotates each block relative to the trained horizon, flushes the cache to the sink and the last frame and cuts positions at a scene change (Yesiltepe et al., 2026), Context Forcing, which trains its own long-context video model (Chen et al., 2026a), and Mixture of Contexts, which trains its router (Cai et al., 2026), were not run.

Retrieval of earlier content. These methods save earlier content and put part of it back into the cache before each chunk. LongLive-RAG matches search embeddings of earlier latents, from a separately trained encoder, against that of the latest latent and returns the cached context of the closest to a frozen video model (Hu et al., 2026). Others select earlier frames or cached blocks by attention relevance and temporal diversity (Meng et al., 2026), by agreement with the initial frames (Zhao et al., 2026b), by camera and action correspondence (Yi et al., 2026b), by similarity of visual descriptors with LoRA fine-tuning (Ye et al., 2026), or by compressed attention scores with fine-tuning (Zhao et al., 2026a). StoryMem screens the keyframes of earlier shots by semantic and aesthetic quality and injects them into a video model fine-tuned with LoRA (Zhang et al., 2025a). FramePack compresses earlier frames into a context of fixed length, giving more of it to frames ranked higher by time proximity or feature similarity (Zhang et al., 2025b), and MemoryPack learns to retrieve earlier context with the text and images as guidance (Wu et al., 2025), all in a trained

video model. Echo-Forcing, without training, fuses the cached features of each scene into a compact memory that a tag or prompt similarity recalls (Wu et al., 2026).

Changing the features the cache already holds. Anchor Forcing recomputes the cache under a new instruction from saved anchor caches, with a video model trained on a three-region positional encoding (Yang et al., 2026b). TetherCache adjusts the statistics of recalled features towards those of a trusted context (Meng et al., 2026), and AdaState replaces the fixed initial chunk with a hidden state that a video model trained for it updates every chunk (Dalva & Yanardag, 2026). Others move saved features to a new position: MemRoPE caches keys before positional encoding (Kim et al., 2026), and Relax Forcing places kept frames just before recent ones (Zhao et al., 2026b).

Memories organized around entities or camera poses. EM-Vid keeps a bank of latent patches per entity named in a structured multi-shot script (Vandersanden et al., 2026). IAMFlow extracts the entities of each instruction with a language model, and its greedy recall covers the most entities, so the model scores and the rule decides (Liu et al., 2026b). WorldMem and Context-as-Memory recall earlier frames by camera pose after training (Xiao et al., 2025; Yu et al., 2025), and Closing the Loop recalls stored latents by camera pose without training (Ma et al., 2026).

Agents that plan, check and remember. Generative agents retrieve a record of experience by recency, importance and relevance (Park et al., 2023). In Mem0 the model adds, updates or deletes stored memories or leaves them unchanged (Chhikara et al., 2025), and A-MEM links each new memory to earlier ones and updates them (Xu et al., 2025); Self-RAG critiques the passages it retrieves (Asai et al., 2024), and FLARE retrieves when a sentence it drafts has low-confidence tokens (Jiang et al., 2023). The keep, retrieve, replace and wait actions of Eq. 4 and the write check play these roles for a video cache, and a trigger like FLARE's could replace the fixed longest gap Δ between two decisions. VideoDirectorGPT expands a video prompt into scene descriptions, entity layouts and consistency groupings (Lin et al., 2024), VISTA has several agents assess a video and rewrite its prompt (Long et al., 2026b), VideoMemory keeps a reference image per entity that a memory agent retrieves or creates each shot (Zhou et al., 2026), and A^{2}RD, without training, has a multimodal language model retrieve a text state and frames per clip, check them against a rubric, edit or regenerate, and write the result back (Long et al., 2026a).

B The Agent: Settings, Workflow and Prompts

This section gives the framework of Sec. 3.2 in full: every setting (Table B.1), the workflow of the agent (Figure B.1), its prompts and the memory a latent saves.

The prompts. The three prompts are given verbatim in the supplementary material with the code, and each asks for a fixed JSON answer with short reasons. The decision prompt gives the agent seven principles: a character keeps its identity unless an instruction replaces it; a character or scene that returns takes its stored entry and not the latest frames; no entry whose identity contradicts the instruction is loaded; a new entry is requested only when the frames already show the new content, and otherwise the agent waits and states what is missing; frames with the wrong person, mixed identities, heavy artifacts or a hidden subject are rejected; keeping the references is the default; and when every slot is taken, the slot that matters least for the next chunk is given up. The write-check prompt asks whether a candidate shows the identity of the stored entry, the state the instruction requires, and a clearly visible subject with no two identities blended. The evaluator prompt gives no decision of the agent and no method name, and asks the three questions of Table C.4 as strict verdicts, with an overall score from 1 to 5.

Memory of a stored latent. A stored latent takes 960 times less memory than the cached keys and values of one chunk on LongLive-2.0-5B and 1,440 times less on LongLive-1.3B.

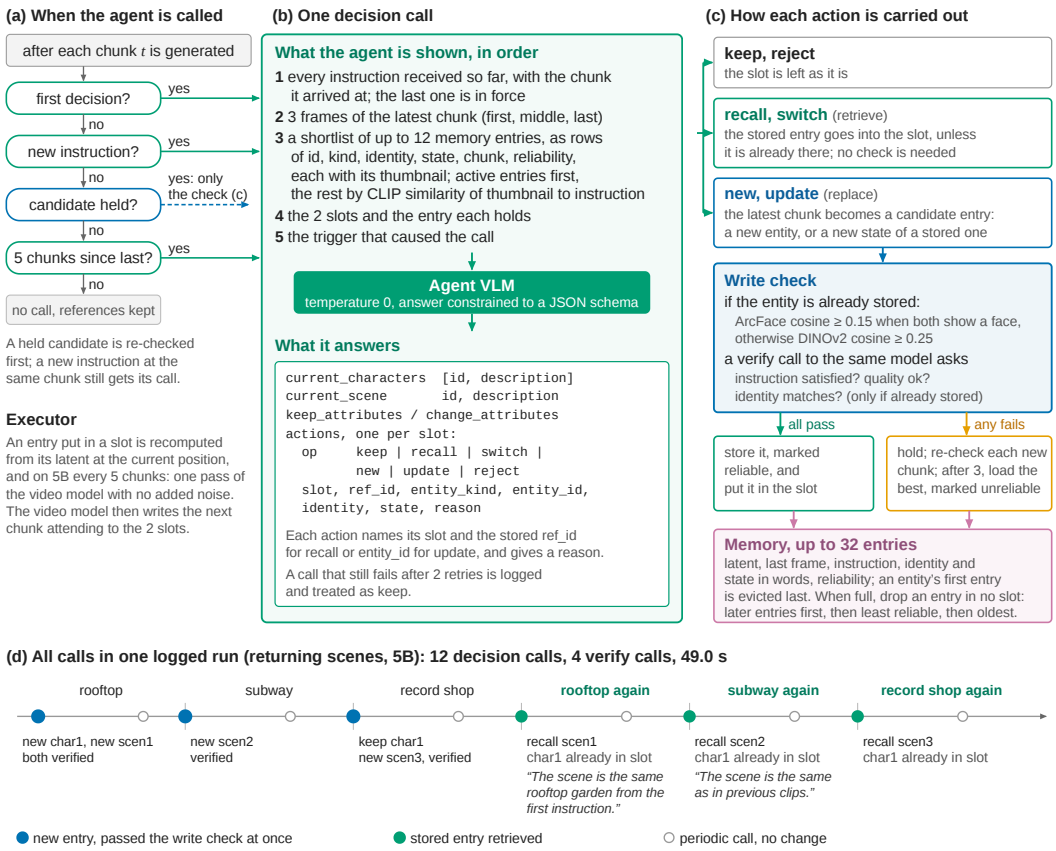

Figure B.1: **When the agent is called and what it does.** (a) The conditions checked after each chunk; in most 1.3B runs (Appendix C) *new instruction* means one for the next chunk, and a periodic slot load waits for the next instruction. (b) One decision call. (c) How each action is carried out. *Retrieve* is `recall` or `switch`, which are carried out alike, and *replace* is `new` or `update`. A `keep` that names what is missing is *wait*. (d) The run of Figure 3, with two of the agent's reasons.

Table B.1: **Settings of StreamPilot.** They are grouped by the decision each controls and hold for every video model, except where a row names LongLive-2.0-5B and as listed in Tables C.1 and C.3. Each ablation row in Tables 5 and M.1 changes one or two of them.

Setting	Value
Deciding which chunk becomes the reference	
Agent model	A VLM (Sec. 4.1), sampling temperature 0
When the agent decides	After chunk 2, before every new instruction (one chunk later on LongLive-2.0-5B and in Appendices J and N) and otherwise 5 chunks after the last decision; in the other 1.3B runs a slot load decided then waits for the next instruction
Frames read per decision	3, decoded from the latest chunk
Failed calls	Retried twice, then the references are kept; the served context is 32,768 tokens
Shortlist shown to the agent	12 stored chunks: the active references, then the others by CLIP similarity of their thumbnails to the instruction, recency breaking ties
Active references	2; 1 for the single strategies of Table K.1 and for Initial only in Table 5's returning scenes
Memory bank	32 chunks; when full, an entry in no slot is removed: an entity's later entries before its first, then the least reliable, then the oldest
Checking whether a newly generated chunk may be stored	
Similarity thresholds	ArcFace (Deng et al., 2019) cosine similarity 0.15, DINOv2 (Oquab et al., 2024) cosine similarity 0.25
Holding a candidate	A candidate that fails is held, each following chunk is checked in its place, and after 3 chunks the checked chunk with the highest identity similarity, or the first for a new entry, is stored, marked unreliable and loaded
Recomputing the features of a stored chunk	
Feature recomputation	One forward pass of the video model per loaded reference, with no added noise, at the positions of the chunk just generated; on LongLive-2.0-5B repeated every 5 chunks while the reference stays in its slot

C BENCHMARK REPRODUCTION DETAILS

This section supplements Sec. 4.1 with the settings of each video model and benchmark (Table C.1), the scripts, seeds and videos (Table C.2), the runs compared (Table C.3), and how the scores and the evaluators are computed (Tables C.4 and C.5). The MemFlow scripts are taken from source revision `6e3fe408e368` and the three Echo-Forcing prompt sets (Wu et al., 2026), there named smooth transition, long-range recall and hard cut, from revision `9e940d3`. A comparison uses only the script and seed pairs available for every setting in it, and scores are averaged over the seeds of a script before scripts are compared, except in Table D.1, which uses one seed of LongLive-2.0-5B, and Table M.2, which uses 4 prompts of our own at 2 seeds on LongLive-2.0-5B. Every scored run used NVIDIA H100 80GB GPUs.

Table C.1: **Settings of each video model.** Each model runs its released weights through one harness, whose sink and window, LongLive's, also serve Self Forcing 1.3B and Causal Forcing 1.3B. STREAMPILOT adds K reference slots, a memory bank of N latents and the longest gap Δ between two decisions (Eq. 2). On LongLive-2.0-5B the first slot takes the place of a window chunk, so the window is 2 chunks with STREAMPILOT and 3 (in parentheses) without it. On LongLive-2.0-5B the sink and window are the settings of the released launch script, which its runs do not record; $N = 32$ is STREAMPILOT's own.

Video model	Resolution	Latent frames per chunk	Recent window (chunks)	Sink (latent frames)	Reference slots K	Memory bank N	Decision gap Δ (chunks)
LongLive-1.3B	832×480	3	3	3	2	32	5
Self Forcing 1.3B	832×480	3	3	3	2	32	5
Causal Forcing 1.3B	832×480	3	3	3	2	32	5
LongLive-2.0-5B	1280×704	8	2 (3)	8	2	32	5

Table C.2: **The four public benchmarks.** MemFlow is from Ji et al. (2025) and the other three from Wu et al. (2026). Each script has six instructions revealed one at a time, and the last column names the clip each clip is compared with. Within a video model, every row of Tables 2 and K.1 runs the same scripts and seeds. Schedules are nominally 64 s on LongLive-2.0-5B and 60 s on the 1.3B models, and saved videos last about 63.6 s and 59.3 s. The model columns count videos of the base model and +STREAMPILOT, on LongLive-2.0-5B also of the four single strategies, at three seeds per script there and one elsewhere.

Benchmark	Scripts	LongLive-2.0-5B	LongLive-1.3B	Self Forcing 1.3B	Causal Forcing 1.3B	Reference for evaluation
MemFlow	100	1,800	200	200	200	the initial clip
Continuous scenes	64	1,152	128	128	128	the previous clip
Returning scenes	64	1,152	128	128	128	an earlier named scene
New scenes	64	1,152	128	128	128	none
Total of these rows		**5,256**	**584**	**584**	**584**	

What each setting holds. All three 1.3B models run through the harness released with LongLive-RAG (Hu et al., 2026), so for Self Forcing 1.3B and Causal Forcing 1.3B we test the released weights rather than their own inference code; all three use Wan-family weights (Wan et al., 2025) and four denoising steps per chunk. On LongLive-2.0-5B the released cache holds 32 latent frames, a sink of 8 and 24 after it; the first slot of 8 takes the place of a chunk after the sink and the second adds 8. For analysis STREAMPILOT is also restricted to each of its four strategies with $K = 1$ (Table C.3), and the *oracle* takes for every script the best of the four, a one-reference comparison that also profits from noise. Deep Forcing's memory policy is also run (Appendix N).

Scores and evaluators. Table C.4 lists the main measures, computed, unless its row says otherwise, on eight frames sampled uniformly per clip and averaged over frames and clips. On returning scenes the named clip is found from the public scene titles and return markers of the script, which are never given to the video model. A difference between two settings is the mean per-script difference, tested by a two-sided paired z-test over scripts and not corrected for the number of comparisons except in Table E.1. The evaluators see the instruction, the eight frames and, when a clip is named, frames of that clip (Table C.5).

How the settings were fixed. The slots ($K = 2$), the memory bank ($N = 32$), the longest gap ($\Delta = 5$), the agent model, its temperature and the prompts were fixed in a pre-registration, and the shortlist of 12 in the first version of the agent's code, before any benchmark video was scored. The

Table C.3: **The four single strategies.** Each replaces the agent by the fixed rule of its row at one reference and keeps the slots and the recomputation of StreamPilot. A citation names the rule a row follows, not a method we reproduce. Tables 5 and L.3 give Similar only and Prompt only two references.

Run	How the reference is chosen	Active references
Base model	Released weights (Chen et al., 2026b; Yang et al., 2026a; Huang et al., 2025; Zhu et al., 2026) run through the harness with no reference added, the released inference for the two LongLive models	none added
StreamPilot *restricted to one strategy*		
Initial only	Keep the initial chunk as the reference, the chunk an attention sink keeps (Xiao et al., 2024; Yang et al., 2026a)	1
Recent only	Replace the reference with the latest chunk, which echoes a window of recent content (Xiao et al., 2024)	1
Similar only	Every 5 chunks, retrieve the stored chunk whose latent, averaged over frames and positions (on LongLive-2.0-5B over frames only), is closest in cosine to that of the latest chunk, from chunks stored every 4 chunks and at least 6 chunks old, following retrieval by latent similarity (Hu et al., 2026)	1
Prompt only	Replace the reference with the current chunk at every instruction change, where the released video model recomputes its cache (Yang et al., 2026a)	1

Table C.4: **The measures.** Where each is defined in full and which way is better.

Measure	What it compares	Better
Ref.	DINOv2 cosine similarity of each frame of a clip to the closest of the last four frames of the named clip; on returning scenes the three returns only	higher
Prev.	The same similarity to the last four frames before each return (returning scenes) or each change of place after the first (new scenes), over the clip or 1, 2 and 4 s after it	lower
Adj.	DINOv2 similarity of neighbouring sampled frames	higher
Txt.	CLIP ViT-B/32 (Radford et al., 2021) similarity of the frames to the instruction in force	higher
Face	*Ref.* with ArcFace, on clips where both show a face; $\times 100$ in Tables 5 and L.3	higher
Subject, background	VBench's DINO and CLIP embeddings of each frame against their mean over the named clip (Table 1), or against neighbouring frames and the first frame (VBench)	higher
Worst clip, imaging drop	Imaging quality of a video's lowest clip; of its first clip minus its last	higher; lower
Scene, *Identity*, *Content*	Share of clips where the evaluator finds the requested place, the people and objects of the named clip (on new scenes, no contradiction of the continuity the instruction states), the requested action and content	higher
Right	*Scene* over the returns	higher
Score	The evaluator's overall rating from 1 to 5	higher
Correct	From the agent's log, share of instruction changes at which the logged operations include the kind of change asked for	higher
Collapse	Videos with a collapsed return (Appendix H)	lower

two thresholds were then set on single-instruction videos that no scored run uses, with two frames of one video 9 s apart as positives and frames of different prompts as negatives: 0.25 for DINOv2 is the point where false accepts and false rejects are equal, 1.6% each, and 0.15 for ArcFace the point with 5% false accepts, faces being fewer and less separable.

Table C.5: **Serving settings of the two evaluators.** Both run on the same server with the same context window and sampling settings. *Revision* is the upstream revision of the weights. The first evaluator's revision was not recorded at serving time; the one listed is the only upstream revision whose files all match the served files in size. *Digests* are the first twelve hex digits of SHA-256. The two are never averaged.

Evaluator	Revision	Context	Server	Digests
First evaluator	`4d7ae4984b7d`	32,768 tokens	vllm 0.29.0	`config.json` ed0c1eb3633d `index.json` 907826a6e46f `config.json` 5cd452860dc1
Second evaluator	`0c351dd01ed8`	32,768 tokens	vllm 0.29.0	`index.json` 520b2e050794

Length ceiling and scoring stability. The released 1.3B implementations build a table of positions for 1,024 latent frames, 256 s of video, and a request beyond it fails inside the rotary embedding; the ceiling is on position and not on memory, one GPU peaking at about 26.9 GiB for a video of one minute and 27.2 GiB for 210 s, the longest length whose memory was sampled. Scoring is stable: over the 936 released videos scored more than once, 98.5% of the later verdicts repeat the first.

Decision timing. In the longer videos of Appendix J and the runs of Appendix N, as on LongLive-2.0-5B, the decision at an instruction change follows the first chunk under it and every load applies from the next chunk; the other 1.3B runs follow Eq. 2.

D VIDEO QUALITY IN DETAIL

This section supplements Table 1 with every VBench dimension on each benchmark (Table D.1). On the three 1.3B video models STREAMPILOT significantly improves aesthetic quality in ten of the twelve cells, temporal flickering and the worst clip in seven and imaging quality in five, against none, one, one and three cells for the base model. It significantly lowers the imaging drop on continuous and returning scenes except LongLive-1.3B at returns, to under half its base-model value in each, and raises dynamic degree and the quality score in three and six cells against one each. Table D.1 also gives VBench's own subject and background consistency, which compare each frame with its neighbours and the first frame, so an instructed change of place counts as inconsistency; on the 1.3B models STREAMPILOT still leads significantly in subject consistency in 6 of the 12 cells and the base model in 2. Against the clip the benchmark names, as in Table 1, STREAMPILOT leads significantly in 9 of the 9 cells that name a clip in subject and in background consistency alike, and the base model in none.

Table D.1: **All VBench dimensions on each benchmark.** Larger is better except the imaging drop. Motion and Dynamic are VBench's motion smoothness and dynamic degree, Quality its weighted quality score, and Worst and Imaging drop are as in Table C.4. Bold marks the better of each pair, and † a significant difference. LongLive-2.0-5B rows use one seed.

	VBench		Against the named clip								Imaging	
Method	Subject	Background	Subject	Background	Motion	Flicker	Dynamic	Aesthetic	Imaging	Quality	drop	Worst
LongLive-1.3B, MemFlow												
Base model	94.17	95.01	68.31	80.71	99.03	98.36	18.00	60.90	**72.82**†	80.34	**−2.31**	68.16
+STREAMPILOT	**94.28**	**95.14**	**83.08**†	**88.49**†	**99.14**†	**98.59**†	**26.00**†	**62.03**†	71.65	**81.15**†	−1.34	**68.72**
LongLive-1.3B, Continuous scenes												
Base model	95.55	95.33	92.01	93.24	99.20	98.63	0.00	55.23	73.63	78.73	2.53	70.39
+STREAMPILOT	**96.20**†	**96.15**†	**95.19**†	**96.34**†	**99.33**†	**98.96**†	0.00	**58.11**†	**74.85**†	**79.85**†	**0.44**†	**73.88**†
LongLive-1.3B, Returning scenes												
Base model	94.11	94.20	77.77	84.01	99.25	98.75	0.00	59.96	72.54	78.87	**−0.18**	69.01
+STREAMPILOT	**95.23**†	**94.98**†	**93.11**†	**93.72**†	**99.31**†	**98.91**†	**1.56**	**61.80**†	**72.61**	**79.75**†	0.26	**70.77**†
LongLive-1.3B, New scenes												
Base model	91.96	93.97	N/A	N/A	98.90	98.06	**28.13**	59.98	**70.55**†	79.83	**−2.15**†	**65.62**
+STREAMPILOT	**92.52**†	**94.35**†	N/A	N/A	**98.97**†	**98.24**†	26.56	**60.23**	69.27	**79.84**	−0.20	64.82
Self Forcing 1.3B, MemFlow												
Base model	**93.57**	94.20	65.36	77.76	**98.77**†	**98.09**†	26.00	59.49	68.56	79.56	3.10	62.39
+STREAMPILOT	93.30	**94.35**	**80.49**†	**87.51**†	98.69	98.01	**35.00**†	**59.66**	**68.92**	**80.24**	**2.35**	**65.23**†
Self Forcing 1.3B, Continuous scenes												
Base model	93.90	94.38	88.83	91.86	**98.88**	98.20	**23.44**	54.72	65.70	78.39	10.15	59.56
+STREAMPILOT	**94.56**†	**94.56**	**93.03**†	**94.23**†	98.84	**98.22**	20.31	**56.11**†	**69.94**†	**79.16**	**2.09**†	**67.67**†
Self Forcing 1.3B, Returning scenes												
Base model	**92.88**	**93.55**	69.85	82.08	**98.88**	**98.20**	**21.88**	59.42	68.25	**79.03**	3.08	63.42
+STREAMPILOT	92.63	93.42	**88.97**†	**91.42**†	98.82	98.19	14.06	**61.33**†	**69.47**†	78.80	**1.15**†	**66.50**†
Self Forcing 1.3B, New scenes												
Base model	**91.53**†	**93.41**	N/A	N/A	**98.48**	**97.53**	**67.19**†	57.57	**66.34**†	**81.18**†	**−0.14**	**60.07**
+STREAMPILOT	90.40	93.27	N/A	N/A	98.43	97.47	51.56	**58.75**†	65.19	79.69	0.46	59.98
Causal Forcing 1.3B, MemFlow												
Base model	88.07	**91.00**†	37.06	62.23	97.44	95.42	50.00	54.69	**70.59**	77.51	**2.57**	65.19
+STREAMPILOT	**88.13**	90.48	**67.86**†	**77.68**†	**97.81**†	**96.49**†	**65.00**†	**58.94**†	70.39	**79.83**†	2.96	**65.69**
Causal Forcing 1.3B, Continuous scenes												
Base model	87.34	90.77	71.24	84.28	96.85	94.53	**76.56**	49.17	67.60	77.39	10.09	60.68
+STREAMPILOT	**88.95**†	**91.10**	**82.23**†	**87.81**†	**98.05**†	**96.78**†	67.19	**53.56**†	**72.33**†	**80.00**†	**2.02**†	**68.94**†
Causal Forcing 1.3B, Returning scenes												
Base model	82.64	**88.76**	52.84	76.73	96.31	92.54	**87.50**	52.17	67.94	76.37	8.66	60.76
+STREAMPILOT	**83.83**†	88.44	**72.11**†	**83.66**†	**96.90**†	**94.62**†	81.25	**58.87**†	**71.29**†	**78.75**†	**3.43**†	**65.77**†
Causal Forcing 1.3B, New scenes												
Base model	**83.16**†	**89.88**†	N/A	N/A	**95.78**†	**92.83**	**98.44**	49.64	**62.76**	**76.19**	**6.91**	**55.54**†
+STREAMPILOT	80.69	88.88	N/A	N/A	94.71	92.21	96.88	**53.42**†	62.72	75.17	7.88	52.96
LongLive-2.0-5B, MemFlow												
Base model	**94.65**†	**95.00**†	70.00	81.30	**99.19**†	**98.66**†	50.00	**58.60**	**69.40**†	82.22	**3.19**	**64.27**†
+STREAMPILOT	93.93	94.49	**71.81**†	**82.26**†	99.07	98.38	**66.00**†	58.09	68.29	**82.79**	4.11	62.71
LongLive-2.0-5B, Continuous scenes												
Base model	**95.17**	**95.10**	**84.92**†	**90.91**†	99.30	**98.81**	**59.38**	54.51	**70.81**†	**82.76**	3.66	**66.27**†
+STREAMPILOT	94.92	94.94	82.32	89.82	99.30	98.76	54.69	**55.02**	70.15	82.28	**2.87**	64.97
LongLive-2.0-5B, Returning scenes												
Base model	**93.55**†	**94.06**†	69.82	82.01	**99.29**†	**98.72**†	31.25	**58.03**	**69.99**†	80.47	**2.67**†	**63.53**†
+STREAMPILOT	92.63	93.64	**72.09**†	**83.98**†	99.18	98.41	**50.00**†	57.81	68.37	**81.18**	5.69	62.13
LongLive-2.0-5B, New scenes												
Base model	**92.24**†	**94.45**†	N/A	N/A	**98.93**†	**98.11**†	81.25	**57.04**†	**66.89**†	**83.09**	**−0.22**	**60.62**†
+STREAMPILOT	91.39	93.85	N/A	N/A	98.74	97.66	**92.19**†	56.52	64.96	82.98	1.02	58.37

E HOLM CORRECTION

This section Holm-corrects the 56 pairs of Tables 2 and 4 plus *Prev.* at 4 s (Table E.1), after which 39 of the 51 leads of STREAMPILOT and 0 of the 4 leads of the base model stay significant.

Table E.1: **Holm-corrected tests for Tables 2 and 4.** Each cell is the probability, corrected over all 56 pairs, that chance alone gives a difference this large between the numbers in the same place in those tables. The test is the paired normal test over scripts, and for the two collapse counts the exact McNemar test on videos and the exact sign test over scripts on returns. Bold marks a difference still significant after correction, and a superscript minus a pair the base model leads.

	MemFlow			Continuous scenes			Returning scenes				New scenes	
Video model	Ref.	Scene	Ident.	Ref.	Scene	Ident.	Ref.	Scene	Ident.	Right	Scene	Ident.
LongLive-1.3B	**<.001**	**<.001**	**<.001**	**<.001**	1	.057	**<.001**	**<.001**	**<.001**	**<.001**	1^-	1^-
Self Forcing 1.3B	**<.001**	**<.001**	**<.001**	**<.001**	1	1	**<.001**	**<.001**	**<.001**	**<.001**	1	1
Causal Forcing 1.3B	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	1	**<.001**	**.017**	**<.001**	**<.001**	.593	1^-

	Prev., return			Collapse		Longer, Ref.			180 s back	
Video model	1 s	2 s	4 s	Videos	Returns	60 s	120 s	240 s	Ref.	Right
LongLive-1.3B	**.003**	1	1^-	1	1	**<.001**	**<.001**	**<.001**	**<.001**	.095
Self Forcing 1.3B	**<.001**	**<.001**	1	**<.001**	**<.001**					
Causal Forcing 1.3B	**<.001**	**<.001**	**.002**	**<.001**	**<.001**					

F CONTENT RATE AND THE SECOND EVALUATOR

This section supplements Table 2 with the content rate and the verdicts of the second evaluator (Table F.1).

Content rate. On returning scenes the first evaluator's content rate is about the same with and without STREAMPILOT on both LongLive models and Self Forcing 1.3B and significantly higher with it on Causal Forcing 1.3B (Table F.1).

The second evaluator. Its 2,008 videos of the base models and STREAMPILOT form 10 studies, a study being one benchmark block of Table F.1 with both the base model and STREAMPILOT. On MemFlow the second evaluator confirms the gains of the first in identity and place on every 1.3B model, each significant, and finds a significantly higher content rate on all three. On returning scenes it confirms the gains in identity, place and *Right* on every 1.3B model and in *Right* on LongLive-2.0-5B. There the two differ on the content rate, which the second evaluator finds significantly higher with STREAMPILOT on Self Forcing 1.3B and the first on Causal Forcing 1.3B.

G THE SCRIPTS OF THE QUALITATIVE FIGURES

This section records how the scripts of Figures 1 and 4 were chosen. The candidates for Figures 1 and 4 were set by rules fixed before any frame was viewed, and every run a rule left was then viewed at the printed size, since a frozen or collapsed row can still be marked as keeping its subject.

Figure 1 is the run, among the 19 of the 192 returning-scene runs on LongLive-2.0-5B in which STREAMPILOT finds the place at all six clips, on which the initial-chunk, latest-chunk and similarity rules miss the most clips, seven in all. Counting the prompt rule as well selects the same run, on which the prompt rule misses only the first visit to the second place.

For Figure 4 a candidate is a MemFlow run on one of the three published 1.3B video models whose opening clip names a person (246 of the 300 paired runs). The evaluator must mark STREAMPILOT as keeping that person in all five later clips and the base model as losing them in at least three (70 runs), and neither video may collapse into noise by the rule of Appendix H, applied to each later clip against clip 1 (30 runs). The two drawn, on LongLive-1.3B and Self Forcing 1.3B, were chosen by viewing all 30 at the printed size; 22 of the 30 have the higher *Ref.* in every later clip. Every box is a detection by a person detector at its default threshold on the frame drawn; clip 1 boxes its largest detection, and each later clip the detected person whose face is most like that person's.

Table F.1: **First and second evaluator.** The second evaluator, from another model family, answers the same questions on the same frames. Differences are in points over the clips both answered, bold where the difference is significant, and n counts the scripts paired. *Agree* is the share of clips with the same verdict, and κ is Cohen's kappa, near zero where both pass almost every clip.

		First evaluator	Second evaluator				
Benchmark	Verdict	Δ	Base model	+StreamPilot	Δ	Agree	κ
LongLive-2.0-5B (Chen et al., 2026b)							
Returning scenes	Ident.	0.0	95.3	95.2	−0.1	91%	0.16
n=64	Scene	+1.4	67.4	68.6	+1.1	89%	0.74
	Content	−1.7	58.6	60.4	+1.8	74%	0.47
	Right	**+9.7**	65.3	72.6	**+7.3**	86%	0.63
New scenes	Ident.	+0.3	99.8	99.5	−0.3	98%	0.09
n=64	Scene	−2.7	74.7	71.9	**−2.9**	81%	0.58
	Content	−1.9	63.4	61.3	−2.1	67%	0.40
LongLive-1.3B (Yang et al., 2026a)							
MemFlow	Ident.	**+26.2**	79.3	96.0	**+16.7**	89%	0.46
n=100	Scene	**+15.2**	91.8	99.3	**+7.5**	99%	0.46
	Content	+1.0	79.5	88.7	**+9.2**	70%	0.27
Continuous scenes	Ident.	**+4.9**	96.6	99.7	**+3.1**	98%	0.43
n=64	Scene	+0.3	100.0	100.0	0.0	100%	0.00
	Content	−0.3	82.3	78.6	−3.6	66%	0.28
Returning scenes	Ident.	**+15.6**	83.9	93.5	**+9.6**	89%	0.33
n=64	Scene	**+8.9**	30.2	36.5	**+6.3**	95%	0.82
	Content	−1.6	26.3	25.3	−1.0	89%	0.65
	Right	**+20.3**	22.9	38.0	**+15.1**	91%	0.67
New scenes	Ident.	−1.0	99.0	99.7	+0.8	96%	−0.01
n=64	Scene	−3.6	57.6	57.0	−0.5	81%	0.62
	Content	−4.2	48.7	48.7	0.0	73%	0.46
Self Forcing 1.3B (Huang et al., 2025)							
MemFlow	Ident.	**+30.0**	70.0	92.3	**+22.3**	90%	0.60
n=100	Scene	**+23.7**	81.0	99.3	**+18.3**	98%	0.67
	Content	**+5.3**	69.2	83.5	**+14.3**	74%	0.42
Returning scenes	Ident.	**+21.1**	72.9	90.6	**+17.7**	89%	0.54
n=64	Scene	**+9.6**	24.2	36.5	**+12.2**	95%	0.87
	Content	−1.0	21.6	27.1	**+5.5**	88%	0.65
	Right	**+19.3**	12.5	38.0	**+25.5**	92%	0.76
Causal Forcing 1.3B (Zhu et al., 2026)							
MemFlow	Ident.	**+31.5**	33.0	69.2	**+36.2**	78%	0.68
n=100	Scene	**+44.5**	33.8	88.3	**+54.5**	83%	0.74
	Content	**+28.0**	32.2	67.0	**+34.8**	71%	0.57
Returning scenes	Ident.	**+19.8**	64.6	77.3	**+12.8**	76%	0.39
n=64	Scene	**+8.9**	32.0	37.2	**+5.2**	79%	0.62
	Content	**+8.6**	27.3	30.5	+3.1	75%	0.48
	Right	**+24.5**	14.1	29.2	**+15.1**	66%	0.38

H Leaving the Previous Scene and Collapse at Returns

This section supplements Table 4 with the paired differences behind its *Prev.* columns (Table H.1) and the definition of its collapse count. *Prev.* asks whether the picture leaves the scene it was in when the instruction moves it elsewhere, which neither *Ref.* nor the evaluators answer. It is read at the three returns of each returning-scene script and at every change after the first on new scenes, on the same 64 scripts and seeds. A broken picture also looks unlike the old scene, so the CLIP similarity of the same frames to the sentence naming the new place, minus that to the sentence naming the old one, is read beside it: at returns it exceeds the base model's significantly on every video model, by +0.004 on LongLive-1.3B, +0.007 on Self Forcing 1.3B and +0.013 on Causal Forcing 1.3B, so the lower *Prev.* is a picture that has moved on. On new scenes, where no earlier place returns, *Prev.* is significantly lower with StreamPilot at one second on every video model and at two seconds on Self Forcing 1.3B, and no difference significantly favours the base model.

Collapse at returns. Some videos of the base models collapse at a return into frames of saturated noise, which VBench imaging quality scores within its range for ordinary frames, so we count collapse from the frames. On the middle frame of each clip we measure the share of pixels whose saturation, in the hue, saturation, value (HSV) representation, exceeds 0.75. A return is collapsed when its share exceeds the mean over the video's clips before the first return by more than 0.20, so a video saturated from its first clip, as a cartoon style is, is not marked, and a video counts as collapsed when any return is. The two collapse thresholds were chosen after the videos were generated and are applied unchanged to every video model and setting. The drop in videos with a collapsed return (Table 4, top) has p <0.0001 on Causal Forcing 1.3B and <0.0001 on Self Forcing 1.3B under McNemar's exact test, and on LongLive-1.3B too few videos collapse to separate the two ($p = 0.125$). Per script, the share of collapsed returns is higher without StreamPilot than with it on 35 scripts against 4 on Causal Forcing 1.3B and on 29 against 0 on Self Forcing 1.3B, with p

Table H.1: **Leaving the previous scene after a change.** *Prev.* is the similarity of the frames after a change to the last four frames before it, and a difference is StreamPilot minus the base model, so a negative one favours StreamPilot. Bold marks a significant difference. At returns the runs are those of Table 4, top, the level columns are *Prev.* over the whole clip, and the Holm-corrected tests are in Table E.1.

	Prev.		Difference			
Video model	Base model	StreamPilot	Clip	1 s	2 s	4 s
Boundaries where the script returns to a place it has already described						
Self Forcing 1.3B	0.769	0.748	−0.021	**−0.080**	**−0.055**	−0.008
LongLive-1.3B	0.810	0.829	+0.019	**−0.040**	−0.020	+0.012
Causal Forcing 1.3B	0.523	0.436	**−0.087**	**−0.151**	**−0.104**	**−0.080**
Boundaries where the script names a place it has not described before						
Self Forcing 1.3B	0.837	0.820	−0.017	**−0.038**	**−0.034**	−0.012
LongLive-1.3B	0.835	0.848	+0.014	**−0.022**	−0.004	+0.005
Causal Forcing 1.3B	0.577	0.568	−0.008	**−0.043**	−0.031	−0.007

<0.0001 and <0.0001 under the exact sign test, and on 4 against 0 on LongLive-1.3B ($p = 0.125$). The per-clip readings are released so that the count can be rebuilt.

I Training-Free against Trained LongLive-1.3B

This section gives every measure behind the training-free against trained comparison of Sec. 4.2 (Table I.1), which pairs Self Forcing 1.3B with StreamPilot against the released inference of LongLive-1.3B on the same scripts and seeds; Self Forcing 1.3B is distilled on 5 s clips (Huang et al., 2025). The content rate, which measures the instruction and not memory, does not differ significantly on either benchmark. LongLive-1.3B, tuned on its own long rollouts, keeps the better picture quality: it flickers less and has a smaller imaging drop and higher imaging quality and worst clip on both benchmarks.

Table I.1: **Self Forcing 1.3B with StreamPilot against LongLive-1.3B.** It covers MemFlow and returning scenes in the runs of Table 2, with the paired difference over scripts, bold where it is significant. Larger is better except *Prev.* and *Imaging drop*.

	MemFlow			Returning scenes		
Measure	LongLive-1.3B base model	Self Forcing 1.3B +StreamPilot	Δ	LongLive-1.3B base model	Self Forcing 1.3B +StreamPilot	Δ
Ref.	0.601	0.726	**+0.125**	0.694	0.834	**+0.140**
Scene (%)	84.2	98.7	**+14.5**	26.8	36.2	**+9.4**
Identity (%)	62.7	87.7	**+25.0**	76.8	89.8	**+13.0**
Content (%)	62.7	62.7	0.0	24.5	23.2	−1.3
Right (%)	N/A	N/A	N/A	17.2	38.0	**+20.8**
Prev. at 1 s	N/A	N/A	N/A	0.903	0.801	**−0.102**
Subject, vs the named clip	68.31	80.49	**+12.18**	77.77	88.97	**+11.20**
Background, vs the named clip	80.71	87.51	**+6.80**	84.01	91.42	**+7.40**
Flicker	98.36	98.01	**−0.35**	98.75	98.19	**−0.56**
Aesthetic	60.90	59.66	**−1.25**	59.96	61.33	**+1.37**
Imaging	72.82	68.92	**−3.90**	72.54	69.47	**−3.07**
Imaging drop	−2.31	2.35	**+4.66**	−0.18	1.15	**+1.33**
Worst clip	68.16	65.23	**−2.93**	69.01	66.50	**−2.52**

J Longer Videos of the Same Scripts in Detail

This section supplements the lower block of Table 4 with how the chains are built (Table J.1), the reach returns split by distance (Table J.2), the capacity of the memory bank and the agent's request.

The chains. The ladder keeps every return where the published script puts it, with the same seed at every length, so a longer row shows that the gain survives a longer stream. In the reach row, which replays the first script's returns at the end of its chain, the gain in *Ref.* is larger at the distant returns, because the base model falls from its near returns to its distant ones while StreamPilot does not, and *Content* separates the two settings in neither group (Table J.2); the distant returns also sit at the end of the chain, so this compares two groups of returns, not the interval alone.

Capacity of the memory bank. No run fills the memory bank within the 240 s tested: no run of any length held more than 21 of the 32 entries.

Table J.1: **Chained scripts for longer videos.** Each chain joins the next few of the 64 published returning-scene scripts in order, every instruction verbatim. Four at a time gives 16 chains, three gives 21, and the ladder rows truncate the chains of four. *Encoded* is below the length asked because the video is decoded in 15 s blocks of 60 latent frames and each block comes out 3 frames short. A return is compared with the earliest clip whose title matches it best, and titles are never shown to the video model.

Length	Encoded (s)	Published scripts	Chains	Clips	Returns	Seeds
60 s	59.25	1	16	6	3	1
120 s	118.5	2	16	12	6	1
240 s	237	4	16	24	12	1
210 s, reach	207.375	3	21	21	12 (3 replayed)	1

Table J.2: **Returns 180 s back against returns 30 s back.** Table 4 gives the replayed returns; this table adds the nine returns inside their own script of the same chains. Each Δ is a mean of per-chain differences over 21 chains, bold where the difference is significant.

Measure	Base model	+StreamPilot	Δ
Replayed returns, 3 per video, the scene 180 s old			
Ref.	0.458	0.795	**+0.337**
Right (%)	17.5	34.9	**+17.5**
Content (%)	17.5	20.6	+3.2
Published returns, 9 per video, the scene 30 s old			
Ref.	0.661	0.796	**+0.135**
Right (%)	25.9	32.3	**+6.3**
Content (%)	24.3	19.0	−5.3

The agent's request. The request lists every instruction so far and grows from 1,519 to 3,675 tokens over a one-minute returning-scene video. A straight line fitted to the largest request of each of the 4,423 released agent runs adds 296 tokens per instruction and meets the served context of 32,768 tokens at about 104 instructions, 17 minutes at one instruction per ten seconds, an extrapolation.

K LongLive-2.0-5B as a Fourth Video Model

This section supplements the LongLive-2.0-5B sentence of Sec. 4.1 with every measure of its runs, including the four single strategies (Table K.1).

***Ref.* and the single strategies.** On continuous scenes every reference strategy lowers *Ref.*, so on this model holding any reference costs similarity to the previous clip. Initial only reaches a higher *Ref.* than StreamPilot on MemFlow and at returns, where every MemFlow clip and the first return of each script name the opening clip that Initial only holds.

L What the Agent Holds and Returns by Name

This section supplements Table 5 with what the agent holds and what the video model shows (Table L.1) and with returns by name, by the person named (Table L.2) and on Self Forcing 1.3B and Causal Forcing 1.3B (Table L.3).

Returns by name. This is the setup of the right block of Table 5, with results by person in Table L.2. Each of the 64 LongLive-1.3B scripts has six clips of 10 s in one place. Clips 1 to 3 each introduce one person, alone in view, with an appearance and a first name. Clips 4 to 6 return to each person once by the first name alone, never to the person of the clip just before, in four orders of 16 scripts each. People and places come from the returning-scene benchmark; the three people of a script share a gender, and names are drawn independently of appearance, so the text links a name to a clip but never to a picture. *Ident.* asks the evaluator whether the return shows the same person as the named clip. *Who* shows it one frame of each of the three people, labelled in a shuffled order, and the frames of the return, and asks which person returned, if any. *Face* ($\times 100$) matches the largest face in each of 8 evenly spaced frames of a return to its closest face in the named clip's last 4 of 8, averaging over frames where both show one (57 scripts for the base model on Causal Forcing 1.3B). The named-clip oracle leads on *Who* mainly for persons 2 and 3, whose returns the agent mostly serves with another person's introduction in its other slot.

Table K.1: **All recorded metrics on LongLive-2.0-5B.** *Txt.* is the CLIP similarity of a clip to its instruction, *Adj.* the DINOv2 similarity of neighbouring frames, and *Face* is *Ref.* with ArcFace on clips with a face. *Right* is the share of returns at which the evaluator finds the requested place. Evaluator rates and *Correct* are percentages, *Score* is its rating from 1 to 5, and *Time* is seconds per video in the scored runs, including time queued behind other videos sharing the agent. Bold marks the better of the base model and +StreamPilot, and † a significant difference (paired test over scripts).

Method	Ref.	Adj.	Txt.	Face	Ident.	Content	Scene	Right	Score	Correct	Time (s)
MemFlow											
Base model	0.587	**0.939**†	**0.302**†	0.278	85.2	57.9	97.6	N/A	3.517	N/A	169.9
StreamPilot *restricted to one strategy and one reference*											
Initial only	0.629	0.931	0.301	0.308	87.7	59.3	98.9	N/A	3.595	N/A	172.9
Recent only	0.575	0.919	0.297	0.243	83.7	57.8	96.1	N/A	3.487	N/A	185.6
Similar only	0.592	0.929	0.300	0.287	85.1	59.9	97.2	N/A	3.571	N/A	175.4
Prompt only	0.574	0.927	0.300	0.258	83.3	59.2	96.9	N/A	3.523	N/A	174.5
Oracle, per script	0.637	0.931	0.300	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
+StreamPilot	**0.605**†	0.924	0.300	**0.299**†	**86.7**	**60.0**	**97.8**	N/A	**3.596**	N/A	254.2
Continuous scenes											
Base model	**0.769**†	**0.936**†	**0.273**	**0.525**†	**95.9**	41.7	**98.4**	N/A	3.212	N/A	173.0
StreamPilot *restricted to one strategy and one reference*											
Initial only	0.746	0.928	0.273	0.457	97.7	47.0	98.4	N/A	3.402	N/A	173.2
Recent only	0.740	0.916	0.273	0.450	95.7	43.8	97.7	N/A	3.260	N/A	185.7
Similar only	0.750	0.926	0.273	0.478	95.9	44.4	98.5	N/A	3.296	N/A	173.4
Prompt only	0.725	0.925	0.273	0.475	94.7	45.1	98.4	N/A	3.297	N/A	175.5
Oracle, per script	0.775	0.927	0.273	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
+StreamPilot	0.744	0.925	**0.273**	0.467	95.4	**47.1**†	98.1	N/A	**3.365**†	N/A	251.1
Returning scenes											
Base model	0.567	**0.923**†	**0.284**†	**0.234**	**92.1**	**51.0**	67.4	68.2	**3.198**	N/A	167.7
StreamPilot *restricted to one strategy and one reference*											
Initial only	0.619	0.908	0.283	0.252	94.4	49.8	68.6	74.7	3.175	0.0	172.3
Recent only	0.543	0.896	0.281	0.224	89.1	49.4	64.3	64.6	3.082	40.0	185.1
Similar only	0.586	0.912	0.282	0.224	92.5	49.8	67.5	74.0	3.173	60.0	172.4
Prompt only	0.556	0.910	0.283	0.206	93.0	51.7	70.4	80.6	3.258	40.0	175.4
Oracle, per script	0.627	0.909	0.284	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
+StreamPilot	**0.598**†	0.905	0.282	0.231	**92.1**	49.3	**68.8**	**78.0**†	3.166	86.1	279.0
New scenes											
Base model	N/A	**0.919**†	**0.304**†	N/A	97.4	**35.2**	**62.4**	N/A	**2.976**	N/A	170.5
StreamPilot *restricted to one strategy and one reference*											
Initial only	N/A	0.900	0.303	N/A	97.7	35.9	62.6	N/A	2.947	0.0	172.8
Recent only	N/A	0.896	0.302	N/A	97.4	36.8	64.8	N/A	3.014	100.0	189.3
Similar only	N/A	0.907	0.302	N/A	97.1	33.4	59.2	N/A	2.873	0.0	172.8
Prompt only	N/A	0.905	0.302	N/A	97.6	36.7	63.9	N/A	3.008	100.0	176.8
+StreamPilot	N/A	0.899	0.301	N/A	**97.7**	33.3	59.7	N/A	2.907	99.4	275.9

Table L.1: **What the agent holds and what the video model shows.** Left, LongLive-1.3B returning scenes: *Held*; *Right* at returns to the opening clip, to later clips and to later clips whose named clip is held (one return for w/o write check); *Scene* at first visits of the second and third places. Right, *Held* on the benchmarks that name a clip, for StreamPilot on each 1.3B model and for two fixed rules on LongLive-1.3B. N/A: no runs, slot logs or such returns.

		Right at returns (%)			First visit (%)			Named clip held (%)		
Method	Held	Opening	Later	Later, held	2nd	3rd	Method	MemFlow	Returning	By name
Base model	N/A	18.8	16.4	N/A	0.0	9.4	+StreamPilot *on*			
+StreamPilot	92.2	96.9	7.8	8.8	0.0	1.6	LongLive-1.3B	79.6	92.2	99.5
CLIP ranking	47.9	93.8	10.9	17.2	0.0	0.0	Self Forcing 1.3B	74.2	96.4	98.4
Similar only	27.1	76.6	10.9	11.1	0.0	6.3	Causal Forcing 1.3B	82.4	91.1	98.4
Prompt only	0.0	68.8	14.1	N/A	0.0	1.6	CLIP ranking	N/A	47.9	30.2
Initial only	33.3	82.8	6.3	N/A	0.0	1.6	Initial only	N/A	33.3	33.3
w/o visual input	100.0	100.0	7.0	7.0	0.0	1.6				
w/o write check	33.9	96.9	7.0	0.0	0.0	3.1				
w/ one reference	62.5	90.6	14.8	14.7	0.0	4.7				

Table L.2: **Returns by name on LongLive-1.3B, by the person named.** Person i is introduced in clip i. *Held* and *Who* are as in Table 5 (25% by guessing among the four answers); Prompt only is the best fixed rule of Sec. 3.2 on *Who*.

	Person 1		Person 2		Person 3	
Method	Held	Who	Held	Who	Held	Who
Base model	N/A	0.0	N/A	0.0	N/A	31.3
+StreamPilot	100.0	73.4	98.4	25.0	100.0	39.1
w/o agent (CLIP ranking, $K=2$)	26.6	26.6	39.1	17.2	25.0	37.5
w/o agent (Prompt only, $K=2$)	0.0	4.7	0.0	21.9	0.0	57.8
Named-clip oracle	100.0	78.1	100.0	60.9	100.0	76.6

Table L.3: **Returns by name on Self Forcing and Causal Forcing 1.3B,** as in Table 5.

Method	Self Forcing 1.3B					Causal Forcing 1.3B				
	Held	Ident.	Ident.$_2$	Who	Face	Held	Ident.	Ident.$_2$	Who	Face
Base model	N/A	7.8†	18.8†	9.4†	13.5†	N/A	25.0†	10.9†	2.1†	11.8†
+StreamPilot	98.4	55.2	76.0	33.9	25.7	98.4	67.2	56.8	40.1	22.0
w/o agent (CLIP ranking, $K=2$)	29.7†	32.3†	51.0†	21.4†	19.8†	40.6†	49.5†	46.9†	14.1†	17.1†
w/o agent (Similar only, $K=2$)	40.6†	43.2†	70.3	21.9†	21.2†	31.3†	55.7†	55.7	18.2†	17.3†
agent stores, MiniLM loads ($K=2$)	100.0	36.5†	63.0†	29.7	23.1†	99.5	44.8†	43.8†	22.4†	18.1†

M Ablation on All Recorded Metrics

This section supplements Table 5 with the measures it omits, *Adj.*, *Txt.*, *Content* and *Score*, on LongLive-1.3B returning scenes (Table M.1), and with a test of recomputation against rotating saved keys (Table M.2).

Table M.1: **Ablation on LongLive-1.3B, returning scenes.** Each row removes or replaces one component, in the runs of Table 5. *No visual input* shows the decision call neither the latest frames nor the stored thumbnails; the write check still reads frames. *No write check* keeps only the ArcFace and DINOv2 thresholds. Columns are those of Table K.1 from Ref. to Score without Face and Right, bold the largest per column, and † a significant difference from two references (paired test over scripts).

Variant	Returning scenes						
	Ref.	Adj.	Txt.	Ident.	Content	Scene	Score
No visual input	0.893	**0.948**	0.279†	91.1	22.7	35.9	2.023
No write check	0.890	**0.948**	0.280†	91.9	23.4	35.7	2.036
+StreamPilot, one reference	0.815†	0.943†	**0.281**†	90.4	**26.6**†	**37.5**	**2.135**†
+StreamPilot, two references	**0.894**	**0.948**	0.278	**92.4**	22.9	35.7	2.013

Table M.2: **Recomputing a stored chunk against rotating its saved keys, on LongLive-2.0-5B.** The opening chunk is held as one fixed reference for a whole 120 s video and never updated, rotated to the current position, or recomputed there from the stored chunk, over 8 paired videos. *Final-face sim.* is the ArcFace cosine of the faces in the last eighth, chunks 79 to 90, to the mean face of the first eighth, and *Drop* its fall from the first eighth. Δ is recomputation minus the row in final-face similarity, bold where the difference is significant. *Subject*, *Background* and *Imaging* are VBench consistency and imaging quality. *Updates* counts computations and rotations of the reference per video, its first computation included: at chunks 11, 16, …, 86 of 90, or once at chunk 9 without updates. *ms each* is their mean time and *Time* their share of generation time.

Update	Final-face sim.			Consistency and quality			Cost per video		
	Final	Drop	Δ	Subject	Background	Imaging	Updates	ms each	Time (%)
No update	0.727	0.183	**+0.104**	0.967	0.960	67.20	1	351	0.1
Rotate saved keys to t	0.761	0.153	**+0.070**	0.965	0.961	66.85	16	28	0.1
Recompute at t	0.831	0.083		0.974	0.965	67.86	16	352	1.8

Recomputation against rotating saved keys. Recomputing a held reference at the current position, as StreamPilot does, keeps the final face closer to the opening one than rotating its saved keys or not updating them, for every prompt and seed, and scores highest on subject and background consistency and on imaging quality (Table M.2). Holding the opening chunk could help by its position rather than its content, so this test keeps the content fixed and varies only how its features reach the current position, over 4 prompts written for it and 2 seeds.

N Comparison with Deep Forcing

This section supplements Sec. 4.1 by pairing Deep Forcing (Yi et al., 2026a), training-free on the Self Forcing 1.3B checkpoint and built to extend one prompt, with StreamPilot on the scripts and seeds both ran (Table N.1), and adds StreamPilot inside Deep Forcing's inference.

Table N.1: **Comparison with Deep Forcing.** Levels and paired differences over the scripts and seeds both ran, bold where significant. Top, the content rate on continuous scenes over 64 scripts; bottom, StreamPilot inside Deep Forcing's inference. Deep Forcing's rows are one run.

Measure	Deep Forcing inference	Harness + StreamPilot	Δ	n
Content (%), Self Forcing 1.3B				
Continuous scenes	28.9	43.0	**+14.1**	64
Inside Deep Forcing's own released inference				
	Deep Forcing	Deep Forcing + StreamPilot	Δ	
Ref., returning scenes	0.874	0.948	**+0.074**	64
Ref., MemFlow	0.819	0.916	**+0.097**	100
Scene, MemFlow (%)	96.7	98.7	**+2.0**	100
Identity, MemFlow (%)	88.8	94.8	**+6.0**	100