BEYOND UNIFORM POOLING: TYPED OBJECT AND RELATION EVIDENCE FOR FROZEN VIDEO LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video language models spend a fixed visual-token budget on uniformly pooled patches or learned latents, so the tokens a model reads carry no record of which object, moment, or relation they describe. We present **StateWeaver**, which replaces part of the uniformly pooled grid with typed evidence that a frozen language model reads within the same token budget. Each video is parsed once, with frozen trackers and a frozen image encoder, into a reusable EvidenceBank of scene features, tracked objects with state timelines, and pairwise relations. For each question, an answer-free retriever takes the dense connector's own 384-token sequence, keeps 256 positions as question-independent context, and writes native scene, object, and relation-selected object features onto the other 128; only the connector is trained. On the official NExT-QA test, StateWeaver reaches **58.1%**, 15.0 points above the 384-token dense connector and 8.7 points above a resampler that reads the same memory and question, and it also improves three further video benchmarks. A factorial ablation finds positive main effects for object evidence, relation-selected object appearance, and the posterior teacher; beyond that teacher, conditioning retrieval on the question adds no separable benefit. All comparisons use one frozen 8B backbone at one 384-token budget, and the resampler contrast changes the layout, the write, and the training objective together.

1 INTRODUCTION

Video multimodal language models must translate hundreds of frames into a small number of language-visible tokens. Token count is a budget; how the budget is allocated is the policy. The dominant policy fixes it in advance: patches are pooled to one or a few preset grids, and every frame and region receives its preset share regardless of the question (Zhang et al., 2024b; Xu et al., 2024b), so object identity, state change, and inter-object interaction stay implicit in a lossy grid. Learned resamplers relax the uniformity with a Q-Former- or Perceiver-style bottleneck (Li et al., 2023a; Jaegle et al., 2021; Alayrac et al., 2022), yet the language model receives exchangeable latents with no record of which object, moment, or relation a token describes, and, instruction-aware variants aside (Dai et al., 2023), both families fix the allocation before the question is read.

We replace part of the uniform grid with typed evidence (fig. 1). A video is parsed once, independently of any question, into the EvidenceBank, a reusable memory organized by scene, object, and relation. Frozen parsers, SAM 2 (Ravi et al., 2024), CoTracker3 (Karaev et al., 2024), and SigLIP2 (Tschannen et al., 2025), propose candidates and store native SigLIP2 values. At question time an answer-free hierarchical retriever reads the memory into exactly 384 tokens: 256 question-independent world positions plus 32 scene, 64 object, and 32 relation positions, each branch competing within its quota. Written values are pooled appearance: a relation slot writes its two endpoint objects, and states enter only the keys. The retriever is distilled from a posterior teacher, the same network run on a query that also contains the answer and used only during training, in the spirit of posterior-guided retriever training (Paranjape et al., 2021). The learned selection is question-specific at test time, but beyond the teacher our controls do not separate a benefit of conditioning on the question.

All connectors share the frozen backbone, features, training data, and evaluation protocol, and the equal-budget control reads the same 24 frames as the parser, so the study ranks interfaces within this frozen, connector-only regime. Our contributions are:

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

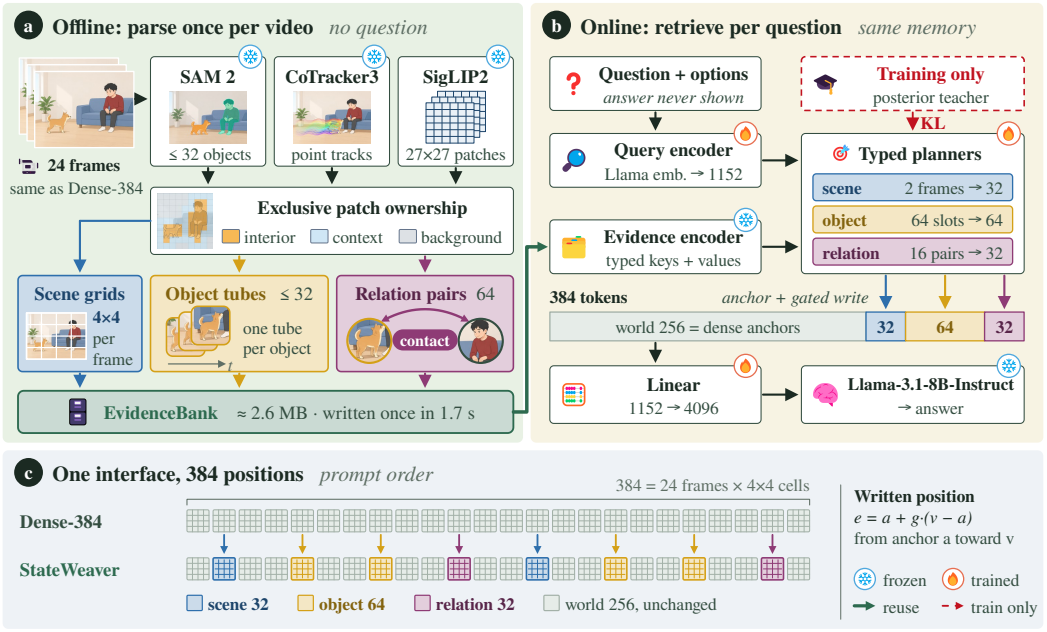

Figure 1: Overview of StateWeaver. **(a)** Each video is parsed once, without the question: frozen SAM 2, CoTracker3, and SigLIP2 assign every patch to one owner and produce scene, object, and relation candidates, which the EvidenceBank stores for every later question. **(b)** For each question, an answer-free query drives three typed planners over the encoded candidates; the posterior teacher is used only in training. **(c)** The 384 tokens start as the Dense-384 sequence; a learned gate writes retrieved values onto 128 positions, interleaved by frame (table 1; the strip in (b) shows shares), and keeps the other 256.

- **The EvidenceBank**, a question-independent memory of scene grids, tracked objects with state timelines, and ordered relation sequences. It is built once per video in 1.7 s, occupies 2.6 MB, and is read by every later question, so the parse is amortized over questions.
- **Typed retrieval into a fixed token budget**. An answer-free retriever writes native scene, object, and relation-selected object features in place onto 128 of the dense connector's own 384 positions, so a closed write is exactly the dense baseline. It reaches 58.1% on NExT-QA, 15.0 points above the equal-budget dense connector and 8.7 above a resampler reading the same memory, improves three further benchmarks, and has positive factorial main effects for the object branch, relation-selected appearance, and the teacher's loss and distillation together.

2 Related Work

Object-centric interfaces. Slot-VLM learns 192 object and event slots on a frozen encoder and LLM (Xu et al., 2024a), VideoOrion tracks objects into a content-dependent token count with a fine-tuned LLM (Feng et al., 2024), and SlotNarrative, the closest contemporaneous work, maintains object narratives in 144 fixed positions with projector-only training (Chen et al., 2026). None maintains a reusable typed memory with scene grids beside object state, writes relation-selected evidence, or reads the bank per question within typed quotas; they are discussed on their published configurations. Structured reasoning has long been used for video question answering: HGA aligns heterogeneous graphs over video shots and question words (Jiang & Han, 2020), VGT builds dynamic graphs over objects and their relations (Xiao et al., 2022), and IGV grounds the question-critical scenes (Li et al., 2022). We differ in where the structure lives: they reason over such structure and emit an answer; we write the endpoint appearance of relation-selected object pairs as native tokens into positions that the relation branch owns inside a fixed budget.

Memory and query-conditioned retrieval. Reusable memory and query-conditioned reading are established. MA-LMM (He et al., 2024), MovieChat (Song et al., 2023), and VideoStreaming (Qian

et al., 2024) keep a bounded or streaming memory of the video for the language model to read, and VideoAgent builds a tracked-object memory once and queries it per question with an LLM agent (Fan et al., 2024). VideoTree (Wang et al., 2024), ViLAMP (Cheng et al., 2025), and LongVU (Shen et al., 2024) adapt which frames or captions reach the model for a given query, and long-context systems such as VideoChat-Flash (Li et al., 2024c), InternVideo2.5 (Wang et al., 2025b), and VideoLLaMA 3 (Zhang et al., 2025) compress video tokens hierarchically or by redundancy; instruction-aware resamplers condition compression on the prompt (Dai et al., 2023). On NExT-QA-style questions, MIST cascades question-conditioned segment and region selection (Gao et al., 2022), TranSTR selects question-critical frames and objects as rationales (Li et al., 2023c), and SeViLA chains a keyframe localizer, trained from the answerer's pseudo-labels, into an answerer (Yu et al., 2023); all three train the module that answers and select raw frames or regions. Here the retained state is the typed EvidenceBank of native values, read by an answer-free retriever into a fixed 384-position layout for a frozen LLM. Distilling an answer-aware module into an answer-free one follows reader-to-retriever distillation (Izacard & Grave, 2020), posterior-guided retriever training (Paranjape et al., 2021), and SeViLA's answer-aware localizer; the distilled quantity is a retrieval distribution over typed visual candidates, not a text-span score or frame index.

Token compression and resampling. Most video LLMs compress a frozen encoder's features into fewer tokens, either through learned bottlenecks such as the BLIP-2 Q-Former (Li et al., 2023a), the Perceiver resampler (Jaegle et al., 2021; Alayrac et al., 2022), TokenPacker (Li et al., 2024b), and xGen-MM-Vid (Ryoo et al., 2024), through fixed per-frame layouts such as SlowFast-LLaVA (Xu et al., 2024b) and LLaMA-VID (Li et al., 2023b), or by clustering, merging, pruning, and progressive or event-based compression, as in Chat-UniVi (Jin et al., 2023), PVC (Yang et al., 2024), METok (Wang et al., 2025a), FrameFusion (Fu et al., 2024), DyCoke (Tao et al., 2024), and VidCom^2 (Liu et al., 2025). None of them assigns tokens to typed scene, object, and relation evidence; StateWeaver keeps typed candidate identities and lets native values compete within typed quotas. Its write rule belongs to the zero-initialized gated-residual family of Flamingo's tanh-gated layers (Alayrac et al., 2022), but the residual carries un-projected native values onto anchors that form a real baseline's token sequence, so the closed gate is itself a scored arm, anchors-only.

Frozen backbones and benchmarks. Connector-only training between a frozen encoder and a frozen language model is the alignment stage of the LLaVA recipe (Liu et al., 2023), whose successors also tune the language model (Li et al., 2024a; Zhang et al., 2024b); here it is a control, not a novelty. Fully trained systems such as Tarsier2 (Yuan et al., 2025) are not matched controls.

3 StateWeaver

Offline, frozen parsers turn a video into the EvidenceBank; online, an answer-free retriever reads it into the 384 language-visible tokens, and a posterior teacher supervises the retriever during training. The backbone, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), the appearance encoder, SigLIP2-SO400M (Tschannen et al., 2025), and the offline trackers are frozen; only the connector is trained, about 99M parameters of which 38M are updated with the language model in the loop. The 384 tokens start from the anchors, the Dense-384 sequence that the dense connector would emit. The 256 world positions keep their anchor; the other 128, the evidence positions, receive a retrieved native value, a pooled SigLIP2 feature rather than a learned latent; anchors-only is StateWeaver with the write switched off. The retriever has one planner per branch (scene, object, relation); each learned slot of a planner selects one candidate and writes fixed positions, and a cell is one location in a frame's 4×4 grid.

3.1 The EvidenceBank: a reusable question-independent memory

Frozen parsing, once per video. Parsing uses no question, answer, or language-model state and reads at most 24 uniformly spaced, endpoint-inclusive frames, the 24 the dense connector pools. A SAM 2-based tracker (Ravi et al., 2024) produces persistent object tracklets, CoTracker3 (Karaev et al., 2024) supplies point tracks and a robust affine estimate of camera motion, and frozen SigLIP2 features from the penultimate layer, 1152-dimensional over 27×27 patches per frame, provide appearance. Every patch of a tracking frame is assigned to at most one owner: object interior, object context ring, or background. Native appearance is pooled per owner into a 4×4 background scene

Table 1: The fixed position layout over the 24 parsed frames: the 8 evidence frames are every third frame from frame 1, and the 16 world frames, which include both endpoints, are the rest. A scene slot writes a 16-cell grid, an object slot one vector, a relation slot two ordered vectors.

Branch	Destination frames	Slots	Cells per slot	Positions
World (question-independent)	all others (16, incl. 0 and 23)	—	—	256
Scene	1, 13	2	16	32
Object	4, 7, 16, 19	64	1	64
Relation	10, 22	16	2	32

grid per frame and, per object, into an interior, a context, and a box view, the last pooling all patches inside the object's box, occluders included. Every candidate is therefore a pooled feature of the same 24×729 patch tensor, with no further encoder pass. Ownership builds the candidates, not the anchors: the same features are also pooled into a plain 4×4 grid over all 729 patches of each frame, and these unfiltered grids, in frame order, are the 384 anchors, the very tensor the dense 24-frame connector sends to the language model.

State and relation candidates. For each object and frame the memory also stores a 32-dimensional state vector and, for each of up to 64 valid ordered object pairs (i, j), a 20-dimensional relation state. Because the pair is ordered, bearing, approach speed, and the other relation components are defined from i to j. No semantic events are hard-selected offline: learned event queries in the encoder summarize these sequences, and state features condition the encoder and the retrieval keys only, never the values. The memory is read unchanged by training and every evaluation.

3.2 Hierarchical native retrieval into 384 tokens

Token layout and anchors. The language-visible sequence is the dense connector's own 384-token tensor, 24 frames $\times$ 16 cells in frame order, each cell a native 1152-dimensional SigLIP2 vector passed through one shared linear readout. The 24 frames are split by index as in table 1: the 8 frames $1, 4, \ldots, 22$, every third frame, are evidence frames, whose 128 positions are question-conditioned, and the other 16, which include both endpoints, are world frames, whose 256 positions are question-independent and always emitted unchanged. Evidence frames go to the branches in the order scene, object, object, relation over each half of the video. The 32 scene positions carry two question-selected frames' grids, cell for cell; the 64 object positions carry object appearance aggregated with event-time attention; the 32 relation positions carry the two endpoint objects' appearance, source first, in adjacent pairs, selected on the relation state, which enters the key and never the value. Every one of the 384 positions starts from its dense anchor, so the 384 anchors together are exactly the Dense-384 token sequence. Every branch retrieves on a typed key but writes a native value on top of its anchor at a fixed place, whatever frame the value was selected from: object slot s writes cell $s \bmod 16$ of the $\lfloor s/16 \rfloor$-th object frame, relation slot s two adjacent cells of the $\lfloor s/8 \rfloor$-th relation frame. For example, given a question about a person approaching a car, an object slot could write one pooled appearance vector for that person. A relation slot could score the ordered person-to-car pair using its motion state and write two adjacent appearance vectors, person first and car second; the relation state selects the pair but is not itself written. A written value carries no timestamp; when an event happened reaches the language model only through what its value shows, and the order of the video through the 256 world positions in frame order.

Evidence encoder. The memory is encoded by 768-wide layers over the 1152-dimensional rows, with hierarchical or sparse attention throughout: a scene encoder emits frame summaries and learned scene-change candidates, an object tube encoder one summary plus four event candidates per object, a relation encoder one summary plus two transition candidates per ordered pair, and a sparse bidirectional fusion layer lets each branch's keys attend to the other two. Retrieval is hierarchical in this sense: candidates exist at two levels, a summary per frame, object, or pair and the scene changes, events, or transitions beneath it, and every key is conditioned by the fusion on the other branches; the flat retriever control scores the raw rows without either (layer shapes in fig. 4).

Answer-free query and typed slots. The query x consists of the system instruction, prior context, current question, and options, never the answer. It is embedded with the frozen Llama embedding table, projected to width $d = 1152$, and encoded by a two-layer query encoder into

$u \in \mathbb{R}^d$. Each branch $b \in \{\text{scene}, \text{object}, \text{relation}\}$ owns 2, 64, and 16 typed learned slots indexed by s, which fill its 32, 64, and 32 positions; slot s carries a learned embedding $z_{b,s}$ and the query $q_{b,s} = W_b^Q[\, z_{b,s};\, u\,]$.

Candidates, keys, and scoring. The evidence encoder gives branch b its candidates $(k_{b,j}, v_{b,j})$, indexed by j: a typed key $k_{b,j}$ built from appearance and state, and the native value $v_{b,j}$ the candidate would write. Let f_b^Q and f_b^K be the branch's learned L_2-normalized 128-dimensional projections, $\tau_b = \exp \rho_b$ a learned temperature, and $m_{b,j} \in \{0, 1\}$ the candidate's validity mask, which covers lifecycle, source validity, and padding, with $\log 0 = -\infty$. Then

$$\ell_{b,sj} = \tau_b^{-1} \big\langle f_b^Q(q_{b,s}),\, f_b^K(k_{b,j}) \big\rangle + \log m_{b,j}, \quad p_{b,s\cdot} = \mathrm{softmax}_j\ \ell_{b,sj}, \quad j^*_{b,s} = \arg\max_j\ \ell_{b,sj}, \tag{1}$$

with the softmax over the branch's candidates j. Slot s writes the native values of its top-scoring candidate, so the per-branch quota fixes the token count; $p_{b,s\cdot}$ is the student's retrieval distribution, supervised by the posterior teacher, so the temperature shapes the distillation target rather than the count. Candidates compete only within their branch, and the 32/64/32 quotas are fixed a priori and varied in a sensitivity sweep (appendix A). Table 5 collects the symbols.

Learned question-conditioned write. With a_i the dense anchor at emitted position i and v_i the native value its slot retrieved, the emitted token is

$$e_i = a_i + w\, g_b(u, a_i)\, \big(v_i - a_i\big), \tag{2}$$

where the gate $g_{b,i} \equiv g_b(u, a_i) = \mathrm{sig}\,\big(\omega_b^\top [\, h_b^u(u);\, h_b^a(a_i);\, h_b^u(u) \odot h_b^a(a_i)\,]\big)$ reads 128-dimensional projections of the question encoding and of the position's own anchor, and w rises linearly from 0 to 1 over the first 3% of training. The gate is evaluated per position, so a scene slot's sixteen positions carry sixteen gate values; a fixed amplitude γ in place of $g_{b,i}$ gives the variant of fig. 3a, and $\gamma = 0$ gives anchors-only. The write is in place: at $t = 0$ the emitted tensor is the Dense-384 token sequence, in the same prompt slot with the same position ids and mask; at convergence the 128 evidence positions deviate from their anchors by a mean $\|e_i - a_i\|_2 / \|a_i\|_2$ of 0.41, while the 256 world positions carry a_i unchanged.

3.3 Training: posterior teacher and answer-free student

Retrieval supervision. No language loss reaches the retrieval scores of eq. (1): the selected index is not differentiated, so only the KL and a small diversity penalty train the metric; the language losses reach the planner through the gate and query encoder (table 8). At inference the planners see only the question and options; during training a posterior query also contains the answer. The posterior teacher is not a separate model but the planner itself, with the same parameters, run on the posterior query and given an auxiliary answer language-model loss; its retrieval distributions are stop-gradient KL targets for the answer-free student, and nothing of it exists at evaluation. The factorial's teacher-off level drops the KL and the teacher's language-model loss together. The signal thus runs through u: the language losses teach the encoder that drives the gate what makes an answer likely, and the KL asks the answer-free encoding to rank candidates as the answer-bearing one does. The teacher distribution comes from the same planner, so this is self-distillation rather than direct supervision of answer-relevant candidates. No loss rewards the selected candidate directly, so whether the ranking carries information is measured by the controls in the analysis.

Question-episode counterfactual supervision (QEC). The supervision averages 4.1 question and answer turns per record, whereas inference retrieves for one question at a time; we align the two by giving each of the 419,488 records exactly one deterministically chosen episode, preserving every row and video. For a multiple-choice episode we build, besides the answer-free query x, a positive posterior query x^+ containing the gold answer and a negative posterior query x^- containing one deterministic wrong option; with token-normalized gold-answer negative log-likelihoods L^+ and L^- and margin μ, the term is $\mathcal{L}_{\mathrm{QEC}} = \mathrm{mean}_{\mathrm{valid}} \max(0,\, L^+ - L^- + \mu)$. Open-ended and caption episodes mask the term; inference is unchanged and answer-free.

Objective and training stages. The planner is trained with

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda\, \mathrm{KL}\big(p_{\text{teacher}} \,\|\, p_{\text{student}}\big) + \mathcal{L}_{\mathrm{aux}}, \tag{3}$$

with the answer-free language-model loss and the distillation term, where $\mathcal{L}_{\text{aux}} = \lambda_{\text{T}}\mathcal{L}_{\text{LM}}^{\text{teacher}} + \lambda_{\text{v}}\mathcal{L}_{\text{video}} + \lambda_{\text{q}}\mathcal{L}_{\text{query}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}$ adds the teacher's own language-model loss; two margin terms on the same path, which require the answer to be likelier with the correct video than with a different one and with the evidence retrieved for its own question than with that retrieved for another question about the same video; and a small penalty on overlapping retrieval distributions of slots in one branch (weights and margins in appendix A). Training has four stages: a short caption alignment of the linear readout, shared by every arm; self-supervised pretraining of the evidence encoder by candidate reconstruction and a shuffled-negative contrastive term, after which it is frozen; retrieval training of the planner and readout with eq. (3), giving StateWeaver (no QEC); and a continuation that adds $\mathcal{L}_{\text{QEC}}$, giving StateWeaver.

4 Experimental Setup

Training data. All connectors are trained on the same 419,488 multiple-choice, open-ended, and caption conversations over 39,901 videos: the YouTube subsets of the four duration splits of LLaVA-Video-178K (Zhang et al., 2024b), 10,000 videos drawn uniformly from each split (table 6). The release's academic subsets, which include NExT-QA's own training videos, are left out to keep the benchmarks' source collections out of training, and the videos were de-duplicated against the evaluation footage by source identifier and perceptual hash, which removed 19 videos.

Benchmarks and denominators. The primary benchmark is NExT-QA multiple choice (Xiao et al., 2021), five options per question. NExT-QA numbers are on the complete official test split of 8,564 questions over 1,000 videos; learning rates and every sweep use the official validation split of 4,996 questions. Transfer is measured without benchmark-specific tuning on MotionBench (Hong et al., 2025), LongVideoBench (Wu et al., 2024), and TVBench (Cores et al., 2024). Every benchmark is evaluated under its protocol in VLMEvalKit (Duan et al., 2024) or, where it is not registered there, lmms-eval (Zhang et al., 2024a), with splits and failure rows in appendix F.

Baselines and controls. The dense, Perceiver, and Memory + resampler arms are trained with the language-model loss alone, except the objective-matched control below; anchors-only and every StateWeaver arm use eq. (3), so the arms differ in objective as well as connector. Dense pooling, LLaVA-Video-style with a linear projection, pools SigLIP2 patches to the budget before one 4.7M linear projection: **Dense-384 (24f)** pools 24 frames at 4×4, the stronger 384-token layout on NExT-QA and therefore our equal-budget control; **Dense-384 (6f)** pools 6 frames at 8×8; **Dense-512** pools 32 frames into 512 tokens. **Perceiver-384** (Jaegle et al., 2021; Alayrac et al., 2022) resamples the same pooled candidates Dense-384 (24f) receives with 384 learned latent queries, 31.7M trainable parameters, and a self-supervised stage matched to StateWeaver's encoder pretraining. **Memory + resampler** reads the same memory: a resampler head reads the identical EvidenceBank through the identical frozen encoder and readout, conditions on the same answer-free query, and fills all 384 positions freely; it does not see the unfiltered anchor grids (appendix E). **Memory + resampler + world** keeps StateWeaver's 256 world positions and lets that head fill only the 128 evidence positions; its **matched-objective** version adds StateWeaver's teacher language-model loss and both ranking terms, so its objective differs from StateWeaver's only by the distillation term, which needs a typed retrieval distribution. **StateWeaver (no QEC) without ranking terms** drops the wrong-video and wrong-query terms from StateWeaver's objective. **Anchors-only** is StateWeaver (no QEC) with the write held at zero, so it emits the Dense-384 token sequence.

Matched training protocol. Every arm starts from the same aligned readout and shares optimizer, schedule shape, batch, clipping, precision, and sequence length (table 7). Every arm trained with the language model in the loop runs one epoch of the training set, 13,109 updates at batch 32, long enough for its validation accuracy to plateau; the QEC continuation and its matched control add 750 updates. Only the learning rate differs, chosen per arm on validation from one six-point grid, $\{1, 2, 5\} \times 10^{-5,-4}$; the QEC continuation and its matched control select none and continue their parent at a fixed 10^{-5}. Seeds are 42, 43, and 44.

Scoring and statistics. Every accuracy is a three-seed mean, and Δs use unrounded means. Intervals are 95% paired bootstrap intervals over video clusters and seeds; the seven factorial effects are judged as one Bonferroni family (appendix F).

Table 2: Accuracy (%, three-seed means) on the official NExT-QA test and three transfer benchmarks, and paired Δs; seed standard deviations and 95% intervals are in appendix F. "Trainable": connector parameters trained while the frozen language model is in the loop; "Connector total": all connector parameters, the frozen evidence encoder included (appendix A); denominators in appendix F.

Method	Tokens	Trainable	Connector total	NExT-QA (test)	MotionBench (val)	LongVideoBench (val)	TVBench
Dense-384 (6f)	384	4.7M	4.7M	41.6	41.1	42.3	36.6
Dense-384 (24f)	384	4.7M	4.7M	43.1	42.8	42.7	36.4
Anchors-only ($\gamma = 0$)	384	37.6M[†]	98.8M	44.3	43.3	43.1	36.7
Dense-512	512	4.7M	4.7M	46.5	43.9	43.9	36.8
Perceiver-384 (pretraining-matched)	384	31.7M	31.7M	48.2	44.6	43.5	37.5
Memory + resampler	384	57.7M	118.9M	49.4	45.4	44.9	37.8
StateWeaver (no QEC)	384	37.6M	98.8M	57.0	48.6	47.5	39.5
StateWeaver (ours)	384	37.6M	98.8M	**58.1**	**49.3**	**48.8**	**40.0**
Δ vs Dense-384 (24f)				+15.0	+6.5	+6.2	+3.6
Δ vs Dense-512				+11.6	+5.4	+5.0	+3.2
Δ vs Memory + resampler				+8.7	+3.9	+4.0	+2.2

[†] At $\gamma = 0$ only the 4.7M readout affects the emitted tokens. At inference every arm needs frozen SigLIP2; Memory + resampler and StateWeaver also need the one-time SAM 2 / CoTracker3 parse (appendix B).

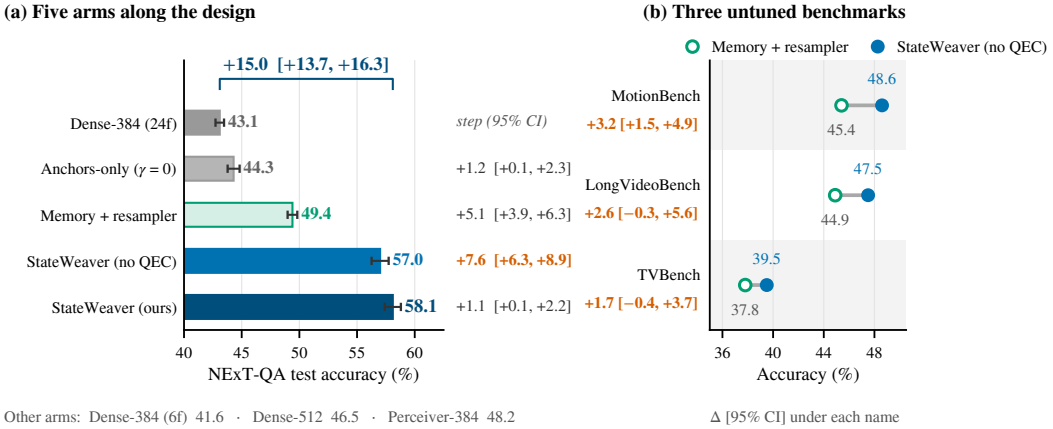

Figure 2: **Five arms ordered along the design.** **(a)** Official NExT-QA test accuracy of each arm with seed standard deviations; the right-hand column gives the paired step from the arm above with its 95% interval. Adjacent arms differ in the whole connector, so the steps order the arms rather than decompose the margin; table 4 splits the +7.6 step. **(b)** Memory + resampler (open) and StateWeaver (no QEC) (filled) on the transfer benchmarks, same checkpoints; under each name, their paired difference with its 95% interval.

5 Main Results

Equal-budget comparison. StateWeaver reaches 58.1% on the official NExT-QA test, 15.0 points ([+13.7, +16.3]) above Dense-384 (24f), the LLaVA-Video-style dense connector, at the same token and frame budget though not the same visual information (table 2). Dense pooling reads each frame's 4×4 grid, whereas the EvidenceBank pools object-owned patches of the full 27×27 grid, a parse only the Memory + resampler controls share. Against anchors-only, which emits the Dense-384 sequence under StateWeaver's own recipe, the evidence write is worth 12.7 points at matched training ([+11.4, +14.0]); the recipe itself accounts for 1.2. The margin shrinks as the baseline strengthens, to **8.7** ([+7.4, +10.0]) over Memory + resampler, which reads the same memory and question but differs in layout, write, and training objective. Figure 2 orders the five arms of the design, and the analysis below examines the 7.6-point no-QEC portion without assigning the full margin to any one component. No external system is scored and every arm trains only a connector, so the absolute level is not calibrated against published results.

Transfer and question types. Without benchmark-specific tuning, StateWeaver improves all three transfer benchmarks over the three baselines with paired rows in table 2, Dense-384 (24f), Dense-512, and Memory + resampler, every interval excluding zero, the TVBench one against Memory + resampler narrowly. Before QEC the lead over Memory + resampler is 3.2, 2.6, and 1.7 points, and

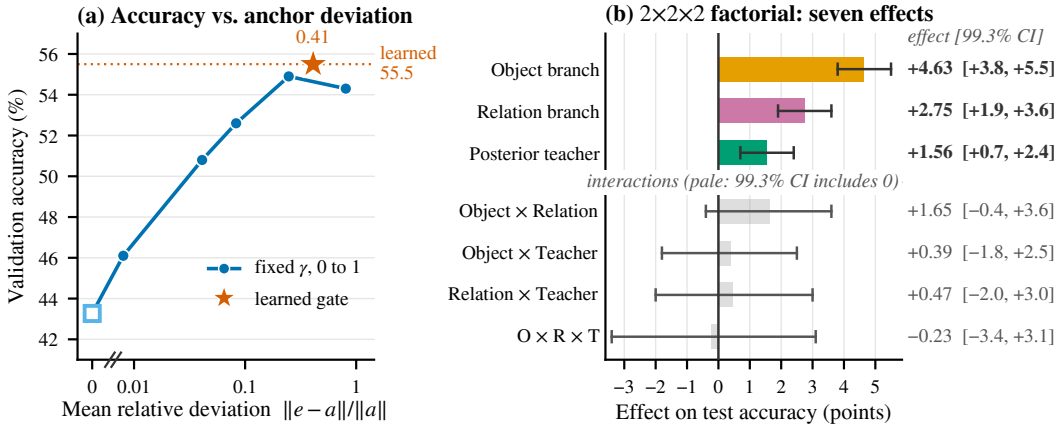

Figure 3: **Write amplitude and factorial effects. (a)** Validation accuracy against the measured mean deviation of the 128 evidence positions from their anchors, for fixed amplitudes γ (curve; $\gamma = 0$, the Dense-384 sequence, left of the axis break) and for the learned gate (star, deviation 0.41). **(b)** The seven effects of the $2\times2\times2$ factorial on NExT-QA test with Bonferroni 99.3% intervals ($\alpha = 0.0071$; 95% intervals in table 16); full-color bars exclude zero, pale bars' intervals include it; interactions are taken with the third factor disabled (appendix D).

only the MotionBench interval excludes zero (table 20). On NExT-QA the gain over Dense-384 (24f) is largest on descriptive questions, 17.9 points, and smallest on temporal ones, 13.4 (appendix C).

Video dependence. Table 3 reverses the frames before parsing or substitutes a different test video. Reversal costs StateWeaver 3.6 points, against 0.4 to 1.3 for every control, and 5.3 on temporal questions against 1.4 on descriptive ones (appendix C). Because the reversed video is parsed again, this measures sensitivity to order, not which component uses it. On TVBench's two order-sensitive tasks StateWeaver leads Dense-384 (24f) by 10.6 points and loses 12.6 under reversal, against 0.7 for the dense connector. A substituted video leaves every arm near the question-only 31.5%, and StateWeaver, which relies most on the video, falls furthest.

6 Ablations and Analysis

What the interface adds over Memory + resampler. Every row of table 4 is measured against one reference, StateWeaver (no QEC), which leads Memory + resampler by 7.6 points ([+6.3, +8.9]). Two trained controls split that step. Giving the resampler head StateWeaver's 256 world positions and only the 128 evidence positions, Memory + resampler + world, closes 2.8 points of it. Giving that control StateWeaver's objective as well, the teacher language-model loss and both ranking terms, closes 1.4 more ([+0.4, +2.5]), and the remaining 3.4 ([+2.1, +4.6]) go with the interface itself: typed quotas, the anchored write, and the retrieval distillation that only a typed retrieval distribution can receive. From the other side, removing the ranking terms costs StateWeaver (no QEC) 1.1 points ([+0.2, +2.1]). The teacher's own contribution, 1.6 points, is a factorial main effect against a different reference (fig. 3b), so it is not an additive share of the 4.8 by which StateWeaver (no QEC) leads Memory + resampler + world.

Table 3: Video dependence on the NExT-QA test (%): real-video accuracy and the change (Δ = ablated minus real) when the video is replaced by a wrong video or reversed in time; question-only accuracy is 31.5 for every method (table 11). StateWeaver's objective contains a wrong-video ranking term and no reversal term; without the ranking terms, StateWeaver (no QEC) still loses 23.1 points to a wrong video (appendix E).

Method	Real video	Δ wrong video	Δ temporal reverse
Dense-384 (24f)	43.1	−9.7	−0.4
Dense-512	46.5	−12.6	−0.7
Perceiver-384	48.2	−14.0	−0.4
Memory + resampler	49.4	−16.6	−1.3
StateWeaver (ours)	**58.1**	−26.9	−3.6

Table 4: Allocation controls on the NExT-QA test (%; Δ vs StateWeaver (no QEC), 95% paired intervals). Rows marked (test time) reuse the trained StateWeaver (no QEC) checkpoints; others are trained under the matched protocol. *No teacher; −0.9 [−2.0, +0.2] vs the matched no-teacher arm.

Control	What changes	Accuracy	Δ
StateWeaver (no QEC)	—	57.0	—
Without ranking terms	wrong-video and wrong-query terms removed	55.9	−1.1 [−2.1, −0.2]
Flat retriever	no summary level, no fusion conditioning	54.6	−2.4 [−3.6, −1.2]
Question-blind planner*	u replaced by a learned constant	54.3	−2.7 [−3.9, −1.4]
Memory + resampler + world, matched objective	+ teacher LM loss and both ranking terms (no KL)	53.6	−3.4 [−4.6, −2.1]
Memory + resampler + world	free resampler fills only the 128 evidence positions	52.2	−4.8 [−6.1, −3.6]
Memory + resampler	free resampler fills all 384 positions	49.4	−7.6 [−8.9, −6.3]
Another question's query (test time)	query of a different question, same video	53.6	−3.3 [−4.1, −2.6]
Random candidates (test time)	a random valid candidate per slot	51.8	−5.1 [−5.9, −4.4]

Question-specific selection. In the two test-time rows the gate keeps the original u: another question's query costs 3.3 points and random candidates 5.1. Object-event selections for two questions on the same video overlap with Jaccard 0.38, against 0.21 for random sets of the same size. These rows measure the checkpoint's reliance on its own selections, not the value of question-conditioned training: a question-blind planner, which can have no teacher, is only 0.9 below its matched no-teacher arm (appendix E), so the question's value arrives mostly through the teacher.

Write amplitude. With a fixed amplitude γ in place of the gate, accuracy on validation rises steeply and saturates above $\gamma = 0.3$ (fig. 3a); the learned gate, at a mean deviation of 0.41, matches the best fixed setting without one having to be selected.

Object branch, relation branch, and posterior teacher. Figure 3b crosses the object branch, the relation branch, and the teacher in all eight combinations on test; a disabled branch returns its positions to the dense anchor (appendix D). Three effects survive Bonferroni correction: the object branch at 4.6 points, the relation branch, which writes relation-selected object appearance, at 2.8, and the teacher at 1.6. The object × relation interaction is positive, +1.7 with 95% interval [+0.2, +3.1], but does not survive correction, its 99.3% interval being [−0.4, +3.6]; the other interactions are near zero. A flat retriever with the same parameters, which scores raw candidate rows without the summary level or fusion conditioning, costs 2.4 points, less than removing either branch.

Question-episode supervision. Against a matched continuation of the same parent, equally long and without the QEC terms (table 18), QEC adds 1.0 point ([0.0, +2.1]); neither episode alignment alone nor the margin's increment over it is separable from zero, so QEC is a small refinement.

Reuse cost. Once a video is parsed, StateWeaver answers in 0.091 s against 0.086 s for the dense connector, since the language model sees the same 384 tokens; with the one-time 1.7 s parse, at NExT-QA's 8.6 questions per video the total cost is 0.29 s per question against 0.16 s.

7 Limitations and Conclusion

Limitations. Our claims are about the ordering of interfaces at one operating point, connector-only training over a frozen 8B backbone after one short shared alignment stage and at one token budget, and absolute accuracies are not comparable to systems that fine-tune a language model. The 7.6-point step from Memory + resampler to StateWeaver (no QEC) is split into the world positions, the training objective, and the interface itself; within the interface, typed quotas and the anchored write have no separate estimates. Inference requires an offline parse with frozen SAM 2, CoTracker3, and SigLIP2, which we treat as inputs like the vision encoder.

Conclusion. StateWeaver combines a reusable EvidenceBank with answer-free hierarchical retrieval into 384 native tokens, written on top of the dense connector's own token sequence, as one interface for a frozen language model. Under the same backbone, training data, alignment, and LLM-visible token and frame budgets it improves official NExT-QA by 15.0 points over the 384-token dense connector and by 8.7 over a resampler reading the same memory, transfers to three further benchmarks, and, after a one-time parse, answers each question within 6% of dense latency.

AI USE STATEMENT

We used generative AI tools only to polish the writing of this manuscript, such as its wording. Every AI-assisted passage was reviewed, edited, and verified by the authors. All research ideas, methods, experiments, and results are the authors' own, and the authors take full responsibility for the content.

REPRODUCIBILITY STATEMENT

All experiments use a frozen Llama-3.1-8B-Instruct backbone and a frozen SigLIP2-SO400M encoder; only the multimodal connector is trained, in four stages of which one is a short caption alignment shared identically by every arm and one pretrains the evidence encoder without the language model (section 3.3, hyperparameters in appendix A). Training uses the YouTube subsets of LLaVA-Video-178K (table 6), de-duplicated against the evaluation footage before training (section 4), with one deterministic frame-sampling policy per method under the matched protocol. The NExT-QA test split is decoded once per trained arm, and every arm decoded on it is reported in this paper (appendix F); learning rates are chosen and every design-stage sweep is reported on the official validation split. Paired statistics and failure rows are in appendix F. Evidence memories are built offline once per video with SAM 2, CoTracker3, and SigLIP2, and are reused unchanged across all questions and all benchmarks.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: A Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198*, 2022.

Junzhe Chen, Siyuan Meng, and Xiaojie Guo. Persistent Object Narratives for Token-Efficient Video Language Models. *arXiv preprint arXiv:2608.04866*, 2026.

Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. Scaling Video-Language Models to 10K Frames via Hierarchical Differential Distillation. *arXiv preprint arXiv:2504.02438*, 2025.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in Time: A New Temporal Benchmark for VideoLLMs. *arXiv preprint arXiv:2410.07752*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*, 2023.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Zerun Ma, Yuxuan Qiao, Mo Li, Tianhao Liang, Lin Zhu, Amit Agarwal, et al. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models. *arXiv preprint arXiv:2407.11691*, 2024.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A Memory-augmented Multimodal Agent for Video Understanding. *arXiv preprint arXiv:2403.11481*, 2024.

Yicheng Feng, Yijiang Li, Wanpeng Zhang, Hao Luo, Zihao Yue, Sipeng Zheng, and Zongqing Lu. VideoOrion: Tokenizing Object Dynamics in Videos. *arXiv preprint arXiv:2411.16156*, 2024.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. FrameFusion: Combining Similarity and Importance for Video Token Reduction on Large Vision Language Models. *arXiv preprint arXiv:2501.01986*, 2024.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. MIST: Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering. *arXiv preprint arXiv:2212.09522*, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. *arXiv preprint arXiv:2404.05726*, 2024.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. MotionBench: Benchmarking and Improving Fine-grained Video Motion Understanding for Vision Language Models. *arXiv preprint arXiv:2501.02955*, 2025.

Gautier Izacard and Edouard Grave. Distilling Knowledge from Reader to Retriever for Question Answering. *arXiv preprint arXiv:2012.04584*, 2020.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General Perception with Iterative Attention. *arXiv preprint arXiv:2103.03206*, 2021.

Pin Jiang and Yahong Han. Reasoning with Heterogeneous Graph Alignment for Video Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11109–11116, 2020.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. *arXiv preprint arXiv:2311.08046*, 2023.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. *arXiv preprint arXiv:2410.11831*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023a.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. TokenPacker: Efficient Visual Projector for Multimodal LLM. *arXiv preprint arXiv:2407.02392*, 2024b.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. VideoChat-Flash: Hierarchical Compression for Long-Context Video Modeling. *arXiv preprint arXiv:2501.00574*, 2024c.

Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An Image Is Worth 2 Tokens in Large Language Models. *arXiv preprint arXiv:2311.17043*, 2023b.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant Grounding for Video Question Answering. *arXiv preprint arXiv:2206.02349*, 2022.

Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. Discovering Spatio-Temporal Rationales for Video Question Answering. *arXiv preprint arXiv:2307.12058*, 2023c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models. *arXiv preprint arXiv:2505.14454*, 2025.

Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D. Manning. Hindsight: Posterior-Guided Training of Retrievers for Improved Open-Ended Generation. *arXiv preprint arXiv:2110.07752*, 2021.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming Long Video Understanding with Large Language Models. *arXiv preprint arXiv:2405.16009*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*, 2024.

Michael S. Ryoo, Honglu Zhou, Shrikant Kendre, Can Qin, Le Xue, Manli Shu, Jongwoo Park, Kanchana Ranasinghe, Silvio Savarese, Ran Xu, et al. xGen-MM-Vid (BLIP-3-Video): You Only Need 32 Tokens to Represent a Video Even in VLMs. *arXiv preprint arXiv:2410.16267*, 2024.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. *arXiv preprint arXiv:2307.16449*, 2023.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models. *arXiv preprint arXiv:2411.15024*, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025.

Mengyue Wang, Shuo Chen, Kristian Kersting, Volker Tresp, and Yunpu Ma. METok: Multi-Stage Event-based Token Compression for Efficient Long Video Understanding. *arXiv preprint arXiv:2506.02850*, 2025a.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. InternVideo2.5: Empowering Video MLLMs with Long and Rich Context Modeling. *arXiv preprint arXiv:2501.12386*, 2025b.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos. *arXiv preprint arXiv:2405.19209*, 2024.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. *arXiv preprint arXiv:2407.15754*, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. *arXiv preprint arXiv:2105.08276*, 2021.

Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video Graph Transformer for Video Question Answering. *arXiv preprint arXiv:2207.05342*, 2022.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Slot-VLM: Object-Event Slots for Video-Language Modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 632–659, 2024a.

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. SlowFast-LLaVA: A Strong Training-Free Baseline for Video Large Language Models. *arXiv preprint arXiv:2407.15841*, 2024b.

Chenyu Yang, Xuan Dong, Xizhou Zhu, Weijie Su, Jiahao Wang, Hao Tian, Zhe Chen, Wenhai Wang, Lewei Lu, and Jifeng Dai. PVC: Progressive Visual Token Compression for Unified Image and Video Processing in Large Vision-Language Models. *arXiv preprint arXiv:2412.09613*, 2024.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-Chained Image-Language Model for Video Localization and Question Answering. *arXiv preprint arXiv:2305.06988*, 2023.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing Large Vision-Language Models from Detailed Video Description to Comprehensive Video Understanding. *arXiv preprint arXiv:2501.07888*, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models. *arXiv preprint arXiv:2407.12772*, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning With Synthetic Data. *arXiv preprint arXiv:2410.02713*, 2024b.

APPENDIX

A EVIDENCEBANK, ENCODER, AND TRAINING DETAILS

Notation. The appendix labels the four training stages A (readout alignment), C (encoder pretraining), R (retrieval training, giving StateWeaver (no QEC)), and Q (the QEC continuation, giving StateWeaver). Table 5 collects the symbols of eqs. (1) and (2) with their role and the stage that trains them. The gate of eq. (2) is

$$g_b(u, a_i) = \mathrm{sig}\Big(\omega_b^\top \big[\, h_b^u(u);\; h_b^a(a_i);\; h_b^u(u) \odot h_b^a(a_i) \,\big]\Big), \tag{4}$$

with h_b^u and h_b^a learned 128-dimensional query and anchor projections, distinct from the retrieval metric f_b^Q, f_b^K, and $w = \min(1, t/0.03\,T)$ at training step t of T.

Table 5: Symbols of eqs. (1) and (2). $d = 1152$ is the SigLIP2 feature width; branch $b \in \{\text{scene}, \text{object}, \text{relation}\}$ owns 2 / 64 / 16 slots (index s) filling 32 / 64 / 32 emitted positions (index i; table 1); j indexes candidates.

Symbol	Shape	Role	Trained in
u	$\mathbb{R}^d$	answer-free query encoding (question and options, never the answer)	Stages R, Q (query encoder)
$z_{b,s}$, W_b^Q, $q_{b,s}$	learned slot embedding ($\mathbb{R}^{128}$); slot query $q_{b,s} = W_b^Q[z_{b,s}; u]$	typed slot s of branch b	Stages R, Q
$k_{b,j}$, $v_{b,j}$	typed key; native value $v_{b,j} \in \mathbb{R}^{c_b \times d}$, $c_b = 16, 1, 2$ vectors for scene, object, relation (pooled SigLIP2 rows)	candidate j of branch b from the evidence encoder	Stage C, then frozen
$m_{b,j}$	$\{0, 1\}$	validity mask (lifecycle, source-valid, ragged padding); $\log 0 = -\infty$	not learned
f_b^Q, f_b^K	$\to \mathbb{R}^{128}$, L_2-normalized	retrieval metric (query and key sides)	Stages R, Q; KL and diversity terms only
$\tau_b = \exp \rho_b$	scalar per branch	retrieval temperature (init. 0.07)	Stages R, Q; KL and diversity terms only
$\ell_{b,sj}$, $p_{b,s\cdot}$, $\hat{j}_{b,s}$	scores; softmax over j; argmax	student retrieval distribution and selected candidate (not differentiated)	—
a_i, v_i, e_i	$\mathbb{R}^d$ each	dense anchor, retrieved native value, emitted token at evidence position i	—
h_b^u, h_b^a, ω_b	$\to \mathbb{R}^{128}$, $\to \mathbb{R}^{128}$; $\mathbb{R}^{3\times 128}$	gate query and anchor projections (not the retrieval metric); gate weights (one head per branch)	Stages R, Q (language, ranking, QEC losses)
$g_{b,i}$, w	$(0, 1)$; $\min(1, t/0.03\,T)$	learned per-position write amplitude; warm-up release schedule	—

Parsing. Tracking frames are endpoint-inclusive: a video with $F \geq 24$ source frames contributes frames $\lfloor i(F-1)/23 \rfloor$ for $i = 0, \ldots, 23$; a video with $F < 24$ contributes all F frames in order, and its last frame fills the remaining $24-F$ slots, which the source-valid mask marks as repeats. No additional appearance keyframes are read: scene-appearance candidates come from the same 24 frames, so the parser's frame budget equals that of the 24-frame dense connector. Object tracks come from a SAM 2 (hiera-small) tracking pipeline with persistent tracklets, backward tracking, and dynamic query points; point tracks from CoTracker3 in offline mode. The candidate cache for one video holds scene features $[24, 16, 1152]$ with timestamps and coverage (ownership, background support, uncertainty, validity), ragged object rows $[M, 3, 1152]$ (interior, context, box) with owner, time, and source-valid indices, per-object state $[O, 24, D_o]$ with lifecycle validity ($D_o = 32$), ordered relation pairs and state $[P, 24, D_r]$ with validity ($D_r = 20$), and camera state $[24, D_c]$ ($D_c = 8$), under the schema caps stated below. All 39,901 training videos have a complete memory (about 105 GB in total); the raw SigLIP2 patch tensor the parser consumes, $24 \times 729 \times 1152$ in FP16, would occupy 40.3 MB per video (1.6 TB for the training set) and is never stored. The cache is identified by a configuration hash and read unchanged by training and by every evaluation.

Architecture constants. Input, value, and query width 1152; evidence-encoder layer width 768 (4.73M parameters per layer at FFN multiplier 2.0); scene encoder 3 layers, object tube encoder 4 layers, relation encoder 2 layers, bidirectional cross-type fusion 2 sparse layers, query encoder 2 layers at width 1152; 16 attention heads; FFN multiplier 2.0; dropout 0.0; retrieval metric dimension 128 with one learned temperature per branch (eq. (1); $\tau_b = \exp \rho_b$, initialized at 0.07, and 0.041 / 0.036 / 0.049 for scene / object / relation at convergence of Stage R); one gate head per branch over the 3×128 concatenation of eq. (4); four learned event queries per object and two transition queries per ordered relation pair. Table 9 sweeps the token split, the branch quotas, the metric dimension, and the event-query count.

Parameter accounting. The connector totals 98,820,322 parameters: the evidence encoder 61,204,480 (the scene, object-tube, and relation encoders and the cross-type fusion; the eleven 768-wide layers hold 52.0M, the 1152-to-768 input projections, state, time, and camera embeddings, and the learned event and transition queries the rest), pretrained in Stage C without any language loss and frozen thereafter; the hierarchical retrieval planner 32,893,154, randomly initialized, of which the question pathway (the 4096-to-1152 projection and the two-layer query encoder) holds 25,981,056 and the three typed branches (slot embeddings, slot-query maps, retrieval metric, and gates) the other 6,912,098; and the language readout 4,722,688, initialized from the shared Stage-A checkpoint. The 37.6M planner and readout parameters are the only ones updated while the language model is in the loop (Stages R and Q), and Stage Q keeps exactly the same trainable set.

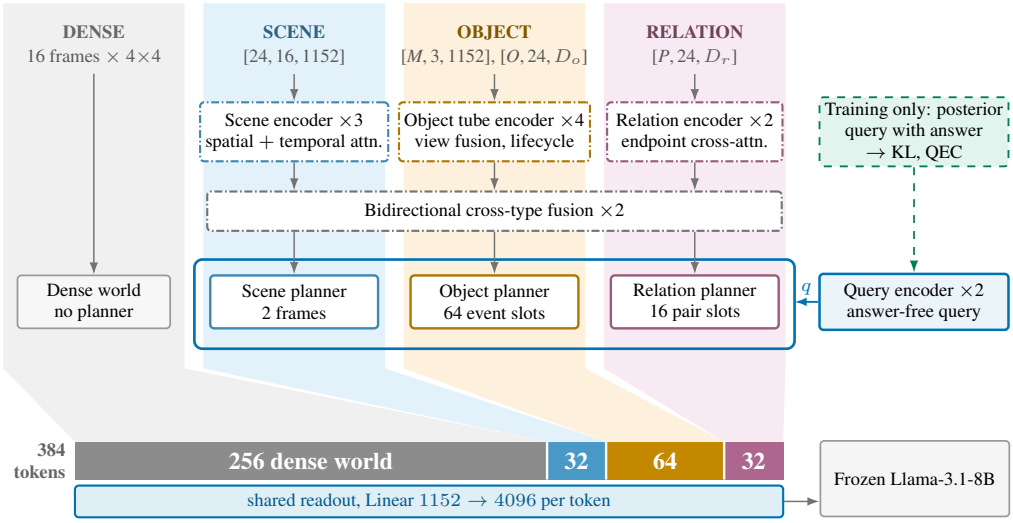

Figure 4: Evidence encoder and planners. Each column follows one branch from its memory tensor through its evidence encoder, the cross-type fusion, and its planner; the dense column passes the 16 world frames' 4×4 grids with no planner. The strip gives each branch's share of the 384 positions (in the sequence they are interleaved in frame order, table 1), which one shared linear readout passes to the frozen language model. Dash-dotted outline: evidence encoder (Stage C, then frozen); blue: trained in Stages R and Q; grey: frozen; green dashed: training-only teacher.

The dense linear connector has 4.7M trainable parameters; Perceiver-384 has 31.7M (4.7M in the linear readout and 27.0M in the resampler head: two cross-attention blocks of 13.3M and 384 latent queries). Memory + resampler has the same head and readout plus its own question pathway of the same shape as StateWeaver's, 26.0M trained from scratch with it, so 57.7M trainable parameters on top of the same frozen 61.2M Stage-C encoder as StateWeaver, a connector total of 118.9M. The language model and vision encoder (8.03B and 0.40B) are never updated. Because the 61.2M encoder is the same frozen tensor set in StateWeaver and in Memory + resampler, it cannot account for any difference between them.

Training data. Training uses only the YouTube subsets of LLaVA-Video-178K (Zhang et al., 2024b), over all four of the release's duration splits (table 6). The release's academic subsets draw on existing video collections, among them NExT-QA's own training videos and questions and collections that several of our transfer benchmarks build on; leaving them out keeps every benchmark's source collection out of training, and re-uploads are caught by the perceptual-hash check of appendix F. The four YouTube pools are unbalanced (79,346 of their 148,265 videos are under 30 s), so 10,000 videos were drawn uniformly at random from each split: every duration band gets the same weight, and the parse and training budget is fixed in advance. Every annotation of a drawn video is kept, so a video contributes all of its caption, open-ended, and multiple-choice rows. Of the 40,000 drawn videos, a quality filter removed 80 that failed to decode, had truncated streams, or had annotations referring to missing footage, and de-duplication removed 19 more. The resulting 419,488 rows train every arm in the same order. Stage A aligns the readout on the same videos' 39,901 caption records, each video's Dense-384 (24f) tokens against its detailed caption, and Stage C reads the videos' memories without any text.

Table 6: Training data: the YouTube subsets of LLaVA-Video-178K, by the release's duration splits. Pool: YouTube videos in the split; drawn: uniform random sample; kept: after the quality filter and de-duplication; conversations: training rows of the kept videos.

Split	Pool	Drawn	Kept	Conversations	Per video
0–30 s	79,346	10,000	9,971	102,041	10.2
30–60 s	19,995	10,000	9,997	105,158	10.5
1–2 min	24,239	10,000	9,938	105,108	10.6
2–3 min	24,685	10,000	9,995	107,181	10.7
Total	148,265	40,000	39,901	419,488	10.5

Language objective and episodes. System, user, and header tokens are masked, assistant content is supervised, sequences are truncated to 2048 tokens, and the 384 projected visual embeddings replace one visual placeholder. Stage R and the matched continuation train on whole records: the planner retrieves once per record from the first turn's question and options, before any answer appears, and every assistant turn is supervised. Stage Q and the episode-alignment arm use a one-episode manifest instead: each of the 419,488 source records contributes one deterministic question episode (turn index $\mathrm{sha256}(\text{row_id}) \bmod \text{turn_count}$), whose earlier turns stay in the prompt as unlabeled context while only that turn conditions retrieval and is supervised, preserving every row and all 39,901 videos; the manifest has 158,405 multiple-choice, 221,182 open-ended, and 39,901 caption episodes, and the multiple-choice episodes are the ones with a parseable counterfactual. Open-ended and caption episodes mask the teacher and QEC terms. The retrieval query length is 512 tokens in Stages R and Q, with no row truncated.

Objective weights. The auxiliary term of eq. (3) is $\mathcal{L}_{\text{aux}} = \lambda_{\text{T}}\mathcal{L}_{\text{LM}}^{\text{teacher}} + \lambda_{\text{v}}\mathcal{L}_{\text{video}} + \lambda_{\text{q}}\mathcal{L}_{\text{query}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}$, and the QEC stage adds $\lambda_{\text{QEC}}\mathcal{L}_{\text{QEC}}$. The weights are $\lambda_{\text{T}} = 0.5$ (teacher language-model loss), $\lambda = 1.5$ for the KL at temperature 1.0 over the scene, object, and relation retrieval distributions, $\lambda_{\text{v}} = 1.0$ with margin 0.1 for the wrong-video ranking term (four rotating deterministic offsets), $\lambda_{\text{q}} = 1.0$ with margin 0.02 and per-example reduction for the same-video wrong-query ranking term, which scores the answer against the evidence retrieved in the same step for another question about the same video, a different turn of the same record chosen deterministically, so the term needs no second sample in the micro-batch (single-turn records mask it), $\lambda_{\text{QEC}} = 1.5$ with margin $\mu = 0.05$, and $\lambda_{\text{div}} = 0.01$. $\mathcal{L}_{\text{div}}$ is a small attention-diversity penalty on the student retrieval distributions $p_{b,s\cdot}$ of eq. (1): it penalizes the agreement between the distributions of different slots of the same branch, and the pathology it prevents is the degenerate solution in which every slot of a branch admits the same candidate, which would spend a 64-position quota on one object event. At weight 0.01 it is a tie-breaker rather than a shaping objective. With $L(x, V, E)$ the token-normalized negative log-likelihood of the answer given query x, video V, and the evidence E written for it, the three terms are

$$\begin{aligned} \mathcal{L}_{\text{video}} &= \text{mean max}\big(0,\ L(x,V,E_x) - L(x,V',E'_x) + 0.1\big), \\ \mathcal{L}_{\text{query}} &= \text{mean max}\big(0,\ L(x,V,E_x) - L(x,V,E_{x'}) + 0.02\big), \\ \mathcal{L}_{\text{div}} &= \text{mean}_b\ \text{mean}_{s\neq s'}\ \langle p_{b,s\cdot},\ p_{b,s'\cdot}\rangle, \end{aligned} \tag{5}$$

where V' is the deterministic wrong video and E'_x the evidence retrieved from its memory for the same query, and $E_{x'}$ the evidence retrieved for another question x' about the same video.

Table 7: Matched training protocol. Every arm shares the Stage-A readout checkpoint, AdamW (weight decay 0.0), a cosine schedule with 3% warm-up, global batch 32 (8 GPUs $\times$ 1 $\times$ accumulation 4), clipping 1.0, BF16, 2048-token sequences, seeds 42/43/44, and the horizon of section 4: one epoch of the training set for every arm with the language model in the loop. Only the learning rate differs, chosen per arm on validation from one six-point grid.

Arm	Init. / trainable set	Params	Learning rate
Stage A: shared readout alignment	random linear; 39,901 video–caption records	4.7M	10^{-3}
Dense-384 (24f $\times$ 4×4)	Stage-A readout	4.7M	2×10^{-4}
Dense-384 (6f $\times$ 8×8)	Stage-A readout	4.7M	2×10^{-4}
Dense-512 (32f $\times$ 4×4)	Stage-A readout	4.7M	2×10^{-4}
Perceiver-384	Stage-A readout + self-sup. stage matched to Stage C	31.7M	10^{-4}
Memory + resampler	Stage-A readout; Stage-C encoder frozen	57.7M	10^{-4}
Memory + resampler + world (and matched objective)	as Memory + resampler, 128 latent queries	57.4M	10^{-4}
Anchors-only ($\gamma = 0$)	as Stage R	37.6M	2×10^{-5}
Stage C: evidence encoder	random; no language loss	61.2M	5×10^{-5}
Stage R: planner + readout (StateWeaver, no QEC)	planner random, readout Stage A; encoder frozen	37.6M	2×10^{-5}
Stage Q: QEC (StateWeaver)	Stage-R checkpoint	37.6M	10^{-5}
Factorial arms, sweeps, question-blind planner	as Stage R	$\leq$ 37.6M	2×10^{-5}

Every arm trained with the language model in the loop runs one epoch of the 419,488 training rows (13,109 updates at batch 32), past the point where its validation accuracy stops improving; Stage A runs one epoch of the 39,901 caption records (1,247 updates) at a fixed 10^{-3}, the usual rate for aligning a linear connector, and Stage C and the Perceiver's self-supervised stage, which read videos rather than conversations, ten epochs of the 39,901 training videos each (12,470 updates); every arm trained from its initialization selects its learning rate from the grid, and the QEC continuation and its matched control continue at a fixed 10^{-5} without selection. The factorial arms, sweeps,

anchors-only, and the question-blind planner therefore share Stage R's 13,109 updates, and the QEC continuation and its matched control run 750 updates each from the Stage-R checkpoint. All training ran on nodes of 8 NVIDIA H200 GPUs. One complete StateWeaver run takes about 120 GPU-hours and a dense or resampler baseline 30 to 40; building the evidence memories for the 39,901 training videos takes about 20, and decoding the official NExT-QA test from the cached memories takes minutes on one GPU.

Parsing rules and schema. Track selection uses only integrated visibility, mask confidence, temporal continuity, and size; no appearance or semantic filter is applied, and the tracker runs with persistent tracklets, backward tracking, and dynamic query points. CoTracker3 seeds query points inside object masks and on a uniform grid, associates them to objects, and supplies object motion relative to a robust affine estimate of camera motion fitted on the unassociated background points. Ownership is exclusive: at each tracking frame the masks are projected onto the 27×27 SigLIP2 patch grid and every patch is assigned to at most one owner, object interior, object context ring, or background; overlaps are resolved in the order mask confidence, point support, temporal continuity, distance to the mask interior, and the residual uncertainty is kept as a coverage channel rather than resolved by summing masks. Appearance is then pooled per owner into a 4×4 background scene grid per frame and, per object, an interior view over the owned patches, a context view over the boundary ring, and a box view over every patch inside the object's box, which remains informative when ownership leaves few interior patches; all three are pooled from the same 27×27 grid, so the parse needs no encoder pass beyond the 24 frames. The schema caps are, in full: at most 24 tracking frames per video; at most 32 persistent objects; at most 8 visual rows per object and at most 192 ragged visual rows per video, whichever of the two binds first, so the two caps do not multiply and a video with 32 objects cannot give all of them 8 rows; 3 visual sources per row (interior, context, box); at most 64 ordered relation pairs; a 4×4 scene grid; and visual dimension 1152. Scene coverage carries $C_s = 6$ channels: ownership as interior, context, and background fractions, background support, uncertainty, and validity. Relation candidates are enumerated over the valid ordered object pairs of a video up to that cap of 64 pairs.

State components. The per-object state vector ($D_o = 32$, one per object and tracking frame) holds the normalized box center, size, and area, visibility, mask confidence, mask entropy, source-valid flags, velocity and acceleration relative to the estimated camera motion, scale change, direction change, curvature, stop/go, appearance change, lifecycle position, and continuity. The per-pair relation state ($D_r = 20$, one per valid ordered pair (i, j) and tracking frame) holds the relative center, distance, bearing, scale ratio, IoU, containment, overlap, relative velocity, approach and separation speed, motion correlation, and contact transitions, each defined from i to j. Camera state ($D_c = 8$) holds the camera-motion estimate and its fit quality.

Encoder layers and value construction. The scene encoder applies per-frame spatial attention within each 4×4 grid and then per-cell temporal attention with a continuous relative-time bias under source-valid masking, emitting frame summaries and learned scene-change candidates. The object tube encoder fuses the interior, context, and box views by source-valid cross-attention, so a missing source is masked rather than zero-averaged, adds state, lifecycle, time, and camera embeddings, applies per-object temporal attention under strict lifecycle masks, and emits one persistent-object summary plus four learned event candidates per valid object. The relation encoder encodes each ordered pair's relation sequence into one pair summary plus two transition candidates, cross-attending to the two endpoint object summaries. The bidirectional cross-type fusion connects object and event candidates to nearby-time scene candidates, scene summaries to active objects, relation candidates to their endpoints only, and object summaries to the relations they participate in; no layer attends densely over all candidates. An object position emits that object's SigLIP2 appearance rows aggregated by the same frozen event-time attention weights that formed the key it was retrieved on, so the value is a convex combination of native rows. A relation slot writes two adjacent positions: the source object's native rows aggregated under the pair's transition-time attention in the first, the target object's in the second, so the value at each position is again a convex combination of native rows and nothing is merged or projected. The direction of the relation is carried by the order of the two positions and by the ordered relation state that formed the key. Stage C pretrains this encoder without any language loss for ten epochs over the memories of the 39,901 training videos (12,470 updates) and freezes it; its objective is self-supervised candidate reconstruction plus a shuffled-negative con-

trastive term. Keeping every value a convex combination of native rows is a design choice made so that the readout aligned in Stage A sees inputs from the distribution it was aligned on; we do not claim that the combined rows are distributed exactly as single rows are, and the write-amplitude sweep of appendix D is the empirical check that the language model tolerates them.

What each loss term updates. Table 8 lists the terms of eq. (3) against the parameters each one reaches. Three properties bound what the retrieval metric can learn: the selected index $j^*_{b,s}$ is not differentiated and only the argmax candidate's value is written, so no language loss reaches the retrieval scores directly; the teacher's retrieval distributions enter the KL with a stop-gradient, so that term trains the student metric only, while the teacher's own language loss does update the shared planner through the teacher's gate and query path; and the Stage-C evidence encoder is frozen in Stages R and Q, so no term in the table updates it.

Table 8: Each term of eq. (3) against the parameter subsets it updates in Stages R and Q. The evidence encoder is frozen throughout both stages. "Through the gate" means the term reaches the planner only via $g_{b,i}$ of eq. (2) and the query encoder, never through the retrieval scores of eq. (1); the linear readout receives every language gradient directly.

Term	Weight	Query text	Parameters updated
$\mathcal{L}_{\mathrm{LM}}^{\mathrm{student}}$	1.0	answer-free	readout (directly); query encoder, gate heads (through the gate)
$\mathcal{L}_{\mathrm{LM}}^{\mathrm{teacher}}$	0.5	posterior	same tensors, evaluated on the posterior query
$\mathrm{KL}(p_{\mathrm{teacher}} \| p_{\mathrm{student}})$	1.5	both	student f_b^Q, f_b^K, τ_b, query encoder; teacher side stop-gradient
$\mathcal{L}_{\mathrm{QEC}}$	1.5	x^+, x^-	readout (directly); query encoder, gate heads (through the gate); masked off-episode
$\mathcal{L}_{\mathrm{video}}$ (wrong video)	1.0	answer-free	readout (directly); query encoder, gate heads (through the gate)
$\mathcal{L}_{\mathrm{query}}$ (wrong query)	1.0	answer-free	readout (directly); query encoder, gate heads (through the gate)
$\mathcal{L}_{\mathrm{div}}$	0.01	answer-free	student f_b^Q, f_b^K, τ_b

Sensitivity to design choices. The token split, the branch quotas, the retrieval metric dimension, and the event-query count are fixed a priori. Table 9 varies each one at a time from the default (StateWeaver (no QEC), final recipe, three seeds, validation split only, since these choices were made before the test split was decoded). The default sits on a plateau: the worst setting in the sweep is 1.3 points below it, and the two "more evidence positions" directions (224/160, 192/192) and the two "more capacity" directions (metric 256, 8 event queries) are each within 0.9 points. Seed standard deviations on this split are 0.4 to 0.8, so a single setting's difference from the default is not resolved at this precision (compare the 0.6 [−0.7, +1.9] of appendix D). Among the settings swept, moving 32 positions from the object to the scene branch costs 1.2 points while doubling the retrieval metric dimension or the event-query count changes accuracy by 0.1, which suggests the design is more sensitive to its quotas than to these two capacity settings.

Table 9: One-factor-at-a-time sensitivity to the design choices, NExT-QA validation accuracy (%, three-seed means); the default is 256 / 128 world / evidence positions, quotas 32 / 64 / 32, metric dimension 128, four event queries per object. Scene quotas are multiples of 16 (one frame) and relation quotas even (one ordered pair); brackets give each world / evidence setting's S / O / R quotas, 1:2:1 where the constraints allow. Δs are computed from unrounded means.

Factor	Setting	Val.	Δ	Factor	Setting	Val.	Δ
World / evidence	320 / 64 [16/32/16]	54.2	−1.3	Quota S / O / R	64 / 32 / 32	54.4	−1.2
	288 / 96 [16/56/24]	55.0	−0.6		**32 / 64 / 32**	**55.5**	—
	256 / 128	**55.5**	—		32 / 32 / 64	54.7	−0.8
	224 / 160 [32/88/40]	55.3	−0.2	Metric dim.	64	55.1	−0.5
	192 / 192 [48/96/48]	54.6	−0.9		**128**	**55.5**	—
Event queries	2	54.9	−0.6		256	55.4	−0.1
	4	**55.5**	—				
	8	55.4	−0.1				

B REUSE EFFICIENCY

All timings are measured on one H200 in BF16 with the production caches on local NVMe and are medians of three runs. Per-question latency includes retrieval (where applicable), language-model prefill over the visual tokens plus the prompt, and an 8-token greedy decode; it is flat in the number of questions asked of one parsed video (measured at 1, 4, and 16). The frozen evidence encoder

runs once per video at parse time and its outputs are part of the cache, so the per-question path adds only the query encoder, the cache read, and the planner to the language model's pass, 5 ms over the dense connector, of which the cache read and the planner take 2.2 ms. The amortized cost at n questions per video is $(\text{parse} + n \cdot \text{latency})/n$ (table 10); NExT-QA asks 8.6 questions per video on average. The 1.7 s parse is 0.6 s of SigLIP2 over the 24 frames, 0.9 s of SAM 2 and CoTracker3, and 0.2 s of the frozen evidence encoder; the 2.6 MB cache is the 0.88 MB dense anchor grid, the 0.88 MB background scene grids, 0.5 MB of object rows, states, and relation states (median video), and 0.3 MB of encoded candidate keys.

Table 10: One-time parse cost, per-question latency, amortized cost, peak GPU memory, cache size, and visual-token count. The raw-patch row is the uncompressed $24 \times 729 \times 1152$ FP16 tensor the parser consumes; its cache and token cells are computed and its parse time measured.

Mode	Parse (s/video)	Latency (s/question)	Amortized s/q at 1 / 4 / 16 q/video	Peak GPU (GB)	Cache (MB/video)	Tokens
Dense-384 (24f)	0.6	0.086	0.69 / 0.24 / 0.12	17.4	0.88	384
Dense-512	0.8	0.103	0.90 / 0.30 / 0.15	17.5	1.18	512
Perceiver-384	0.6	0.089	0.69 / 0.24 / 0.13	17.6	0.88	384
Memory + resampler	1.7	0.092	1.79 / 0.52 / 0.20	18.1	2.6	384
Raw patch features (reference)	0.6	n/a	n/a	n/a	40.3	17,496
StateWeaver (ours)	1.7	0.091	1.79 / 0.52 / 0.20	18.2	2.6	384

C VIDEO DEPENDENCE AND QUESTION GROUPS

Table 11 gives the full video-dependence table behind section 5. The question-only control is one number, 31.5, because with the visual tokens withheld no connector is invoked and every method runs the same frozen language model on the same prompt. A substituted video leaves every arm near that level; for StateWeaver this column partly reads out a training term, since its wrong-video ranking term requires the gold answer to be likelier under the correct video than under a different one. No training term involves reversed frames, so the reversal column and the TVBench order-sensitive subset carry the video-dependence argument. The reversal also changes the motion and lifecycle features in the retrieval keys. By question group (table 12), StateWeaver loses most on temporal and least on descriptive questions, and the controls show the same ordering at a fraction of the size. On TVBench's two order-sensitive tasks, Action Antonym and Object Shuffle (545 of the 2,654 rows), Dense-384 (24f) reaches 33.0, Dense-512 33.7, Perceiver-384 33.5, Memory + resampler 35.2, and StateWeaver 43.7, 10.6 points above Dense-384 (24f) ([+5.4, +15.4]); reversal costs StateWeaver 12.6 points on this subset against 0.7 for Dense-384 (24f).

Table 11: Video-dependence controls on the NExT-QA official test, generation accuracy (%). The last three columns are the change from the real-video column under each substitution, ablated minus real, so every entry is a drop, as in table 3; they are computed from unrounded accuracies.

Method	Real video	Question-only	Wrong video	Temporal reverse	Δ Q-only	Δ Wrong	Δ Reverse
Dense-384 (24f)	43.1	31.5	33.4	42.7	−11.6	−9.7	−0.4
Dense-512	46.5	31.5	33.9	45.8	−15.0	−12.6	−0.7
Perceiver-384	48.2	31.5	34.2	47.7	−16.7	−14.0	−0.4
Memory + resampler	49.4	31.5	32.8	48.1	−17.9	−16.6	−1.3
StateWeaver (ours)	**58.1**	31.5	31.2	54.5	−26.6	−26.9	−3.6

Table 12: Temporal-reverse drop (reversed minus real, points) by NExT-QA question group, from the same per-question predictions as table 11.

Method	Causal (4,502)	Temporal (2,657)	Descr. (1,405)	All
Dense-384 (24f)	−0.3	−0.7	−0.1	−0.4
Memory + resampler	−1.2	−1.9	−0.4	−1.3
StateWeaver (ours)	−3.3	−5.3	−1.4	−3.6

Question groups. Table 13 gives the three official question groups (causal 4,502, temporal 2,657, descriptive 1,405), using the same per-question predictions as table 2. Raising the token budget of a uniform connector does not recover the gain (+12.1 / +7.6 / +17.3 over Dense-512), and neither does giving a resampler the same structured input (+9.1 / +5.7 / +13.1 over Memory + resampler); the Δs are computed from unrounded accuracies.

Table 13: NExT-QA test accuracy (%) by question group, same per-question predictions as table 2.

Method	Causal (4,502)	Temporal (2,657)	Descr. (1,405)	All (8,564)
Dense-384 (24f)	42.9	37.6	54.1	43.1
Dense-512	45.8	43.5	54.6	46.5
Memory + resampler	48.8	45.4	58.8	49.4
StateWeaver (ours)	**57.9**	**51.1**	**72.0**	**58.1**
Δ vs Dense-384 (24f)	+15.0	+13.4	+17.9	+15.0
Δ vs Dense-512	+12.1	+7.6	+17.3	+11.6
Δ vs Memory + resampler	+9.1	+5.7	+13.1	+8.7

D WRITE-AMPLITUDE SWEEP AND FACTORIAL ARMS

Table 14 gives the numbers behind fig. 3a: the learned gate of eq. (2) is replaced by a fixed amplitude γ, $e_i = a_i + w\,\gamma\,(v_i - a_i)$ with no gate head, and everything else is identical. Because the sweep informed the design it is reported on validation, and only its endpoints were decoded on test (44.3 at $\gamma = 0$, 57.0 for the learned gate). The deviation column is measured, not set: under a constant amplitude it equals γ times the relative distance of the retrieved value from its anchor, which averages about 0.81, as the $\gamma = 1.0$ column shows, so the two columns track at roughly 0.8γ. The learned gate's 0.41 thus corresponds to an effective amplitude of about 0.5, spent unevenly across positions (95th percentile 0.78), with per-branch deviations of 0.36 for scene, 0.47 for object, and 0.33 for relation positions. It is 0.6 points above $\gamma = 0.3$ ([−0.7, +1.9]), within seed variation. The smallest amplitude already adds 2.8 points at a deviation of 0.008; the sweep cannot separate the written content from the change in training dynamics that any nonzero write causes.

Table 14: Write-amplitude sweep on the official NExT-QA validation split (three-seed means). Deviation is the mean relative deviation of the 128 evidence positions from their anchors at convergence.

Write rule	$\gamma = 0$	0.01	0.05	0.1	0.3	1.0	learned gate
Mean deviation	0.000	0.008	0.041	0.083	0.247	0.806	**0.410**
Validation accuracy	43.3	46.1	50.8	52.6	54.9	54.3	**55.5**

Factorial arms. Table 15 gives the eight arms behind fig. 3b on the official NExT-QA test with the flat-retriever control, and table 17 the same arms on the transfer benchmarks against the same-recipe full arm (three-seed means; video-clustered bootstrap, 10,000 replicates; failed decodes counted as wrong). Seventeen of the 21 transfer intervals exclude zero and the arm ordering largely reproduces NExT-QA's; these intervals are a descriptive replication, and only the seven NExT-QA effects are judged under Bonferroni correction (appendix F).

The all-off arm and anchors-only differ in their effective trainable sets as well as in the scene branch, so no effect is claimed for the scene branch.

Which estimator each effect is. Main effects are differences of marginal means: the four arms with the factor enabled minus the four without. Interactions are reported as simple interactions at the reference level of the remaining factor(s); the object × relation entry is the difference of differences measured with the posterior teacher disabled, which is the configuration in which both branches are added to the same baseline. The three-way term is the object × relation interaction with the teacher enabled minus the same interaction with it disabled. Every effect is computed from the unrounded arm means, so it can differ in the second decimal from the printed one-decimal arm accuracies. Averaging the difference of differences over both levels of the third factor instead gives +1.54, +0.28, and +0.36 for the three two-way interactions, leaving every sign and every significance verdict unchanged; the arms of table 17 use the same convention (averaged object × relation values +1.68 / +1.27 / +1.21 on MotionBench / LongVideoBench / TVBench).

E LEARNED-RESAMPLER CONTROLS

Two controls, two questions. Perceiver-384 receives the same input as Dense-384 (24f) and a pretraining budget matched to StateWeaver's encoder, and isolates learned resampling from linear pooling. Memory + resampler receives the same memory, encoder, readout, and answer-free query as StateWeaver and is the only control that reads the same parse; it is memory-matched but not

Table 15: The eight factorial arms and the flat-retriever control (section 6) on the official NExT-QA test (accuracy, %; Δ vs the full arm, 95% paired interval; Bench.: unweighted mean of the three per-benchmark Δs of table 17). O: object branch, R: relation branch, T: posterior teacher; the full arm is StateWeaver (no QEC).

O	R	T	NExT-QA	Δ vs full	Bench.
–	–	–	48.1	−8.9 [−10.1, −7.7]	−7.8
✓	–	–	51.7	−5.2 [−6.4, −4.1]	−4.7
–	✓	–	49.8	−7.1 [−8.4, −5.9]	−6.1
–	–	✓	49.3	−7.7 [−9.1, −6.3]	−7.1
✓	✓	–	55.1	−1.8 [−3.0, −0.6]	−1.6
✓	–	✓	53.3	−3.6 [−4.8, −2.5]	−3.4
–	✓	✓	51.5	−5.5 [−6.9, −4.0]	−5.0
✓	✓	✓	**57.0**	—	—
Flat retriever, equal params			54.6	−2.4 [−3.6, −1.2]	—

Table 16: The seven factorial effects on the official NExT-QA test (points), computed from the unrounded arm means, with 95% and Bonferroni 99.3% paired bootstrap intervals ($\alpha = 0.05/7$).

Effect	Estimate	95% interval	99.3% interval	Survives
Object branch	+4.63	[+4.0, +5.2]	[+3.8, +5.5]	yes
Relation branch	+2.75	[+2.2, +3.4]	[+1.9, +3.6]	yes
Posterior teacher	+1.56	[+0.9, +2.2]	[+0.7, +2.4]	yes
Object × relation	+1.65	[+0.2, +3.1]	[−0.4, +3.6]	no
Object × teacher	+0.39	[−1.2, +1.9]	[−1.8, +2.5]	no
Relation × teacher	+0.47	[−1.4, +2.3]	[−2.0, +3.0]	no
Object × relation × teacher	−0.23	[−2.6, +2.2]	[−3.4, +3.1]	no

input-matched, because its head does not receive the 256 unfiltered world grids, a difference that Memory + resampler + world removes.

Fully trained Perceiver-384. Two cross-attention blocks with 384 learned latent queries of width 1152 (16 heads, FFN multiplier 3.0, pre-LayerNorm, dropout 0.0; 27.0M parameters) over the same 24-frame 4×4 SigLIP2 candidates used by Dense-384 (24f), followed by the shared Stage-A $1152 \rightarrow 4096$ readout; 31.7M trainable parameters. Its pretraining is matched to StateWeaver's Stage C: a self-supervised stage of the same 12,470 updates with masked-cell reconstruction plus shuffled-negative contrast over its own candidate grid, the same objective family and the same budget, worth +1.1 points to it (47.1 → 48.2), so the memory comparison is not confounded by pretraining budget. It is then trained under the matched protocol at 10^{-4} (results in table 2). Its input is the pooled $24 \times 4 \times 4$ grid, far less than what StateWeaver's memory is built from, so it is a capacity-and-pretraining control, not an information-matched one.

Memory + resampler. Memory + resampler shares with StateWeaver the offline parse and EvidenceBank, the frozen Stage-C encoder, the Stage-A readout, the answer-free query and its 512-token budget, the 384-token budget, the frozen language model, the training data, the matched protocol, three seeds, and the evaluation protocol; it differs in the allocation policy, the token layout, the write, which has no anchor, and the objective, the language-model loss alone (section 4). A generic Perceiver-style head of the same size as Perceiver-384, 27.0M parameters plus the shared 4.7M readout, cross-attends the encoder's candidate set together with the question encoding u, which its own question pathway (the 4096-to-1152 projection and two-layer query encoder of StateWeaver, 26.0M, trained from scratch with the head) computes from the same query; it has no typing, quotas, anchor, or teacher and fills all 384 positions. Its 57.7M trainable parameters exceed StateWeaver's 37.6M, so capacity favors the control. Memory + resampler + world is the same head with 128 latent queries and 57.4M parameters, whose outputs fill only the 128 evidence positions while the 256 world positions keep their anchors. The question-blind planner replaces u by one learned vector in the slot queries and the gate; with no question there is nothing for a posterior teacher to condition on, so it trains without the teacher terms, and its matched reference is the object + relation arm without teacher (55.1), against which it loses 0.9 ([−2.0, +0.2]). The two test-time controls reuse the trained StateWeaver (no QEC) checkpoints and change only the selection: the query of a different question about the same video, or a uniformly random valid candidate per slot.

Table 17: Factorial arms on the transfer benchmarks: accuracy (%) and Δ vs the full arm with its 95% paired interval.

O	R	T	MotionBench	Δ	LongVideoBench	Δ	TVBench	Δ
–	–	–	39.5	−9.0 [−10.8, −7.2]	39.8	−7.7 [−10.8, −4.5]	32.7	−6.8 [−8.9, −4.8]
✓	–	–	43.4	−5.2 [−6.8, −3.6]	43.2	−4.3 [−7.1, −1.5]	35.0	−4.5 [−6.3, −2.7]
–	✓	–	41.6	−7.0 [−8.7, −5.3]	40.9	−6.6 [−9.3, −3.8]	34.7	−4.8 [−6.8, −2.9]
–	–	✓	40.1	−8.5 [−10.3, −6.7]	40.9	−6.7 [−9.5, −3.9]	33.5	−6.0 [−7.9, −4.0]
✓	✓	–	47.2	−1.4 [−3.0, +0.2]	45.5	−2.0 [−4.6, +0.6]	38.1	−1.4 [−3.5, +0.6]
✓	–	✓	44.5	−4.1 [−5.6, −2.5]	44.7	−2.8 [−5.7, +0.2]	36.1	−3.4 [−5.2, −1.5]
–	✓	✓	42.5	−6.1 [−7.9, −4.3]	42.3	−5.2 [−8.4, −1.9]	35.7	−3.8 [−6.0, −1.6]
✓	✓	✓	**48.6**	—	**47.5**	—	**39.5**	—

Objective-matched controls. Two arms separate the objective from the interface. Memory + resampler + world (matched objective) is Memory + resampler + world trained with StateWeaver's auxiliary terms: the teacher language-model loss, computed by running the same head on the posterior query, and the wrong-video and wrong-query ranking terms with StateWeaver's margins, weights, and deterministic negatives (appendix A). Only the distillation term is absent, since a free resampler emits no retrieval distribution over typed candidates to distill. StateWeaver (no QEC) without ranking terms is Stage R with $\lambda_v = \lambda_q = 0$ and every other setting unchanged. Both are trained under the matched protocol (13,109 updates, three seeds) and decoded once on the official test. The matched objective lifts the resampler control from 52.2 to 53.6% (+1.4, [+0.4, +2.5]), and StateWeaver (no QEC) still leads it by 3.4 points ([+2.1, +4.6]); removing the ranking terms costs StateWeaver (no QEC) 1.1 points ([+0.2, +2.1], 55.9%). On the wrong-video control of table 3, StateWeaver (no QEC) without ranking terms falls from 55.9 to 32.8% (−23.1), against −16.6 for Memory + resampler, so the ranking term sharpens but does not create StateWeaver's dependence on the correct video.

Question-episode supervision against a matched continuation. Table 18 gives the control behind the QEC paragraph of section 6: the same Stage-R checkpoint continued for the same 750 updates at the same learning rate, retrieval query length, and schedule with the QEC terms off. The continuation itself is worth only +0.1, so 1.0 of the 1.1-point difference against Stage R comes from the objective rather than the extra training.

Table 18: QEC against a matched continuation of its Stage-R parent (same 750 updates, learning rate, and query length; QEC terms off): NExT-QA test accuracy (%), Δ vs the control with 95% intervals. Episode alignment switches from whole records to one episode per record (appendix A); the margin then adds $\mathcal{L}_{\mathrm{QEC}}$.

Arm	Accuracy	Δ vs control
StateWeaver (no QEC), Stage R	57.0	−0.1
Matched continuation (control)	57.1	reference
+ episode alignment	57.5	+0.4 [−0.6, +1.3]
+ posterior margin (StateWeaver)	**58.1**	**+1.0** [0.0, +2.1]

F Paired statistics and evaluation details

Procedure. All comparisons pair per-question predictions on the shared rows (the 8,564 official NExT-QA questions; each benchmark's own question set otherwise). Each arm contributes three per-question prediction files, one per seed, over the identical row order, and nothing is averaged before resampling. One bootstrap replicate draws a sample of the video clusters with replacement and, independently per arm, a sample of the three seeds with replacement (arms continued from, or evaluated on, the same checkpoints, namely the QEC stage and its controls and the test-time selection controls, share their parent's seed draw); the replicate's statistic is the paired accuracy difference on the drawn rows under the drawn seeds, averaged over them. We report the 2.5 / 97.5 percentiles of 10,000 replicates at a fixed seed, released with the code, the three-seed mean on all rows as the point estimate, and $P(\Delta > 0)$; McNemar's exact two-sided test is computed on the seed-42 predictions and is labeled as the question-level test it is. Tables print 95% intervals throughout; the Bonferroni decisions for the seven-effect factorial family use the corresponding 99.3% intervals ($\alpha = 0.05/7 = 0.0071$), and every effect marked as surviving does so at that level. The main-table, attribution, supervision, and transfer contrasts of table 19 were fixed before the test split was

decoded and are single contrasts; the question-group splits of appendix C, the TVBench order-sensitive subset, and the 21 per-benchmark factorial intervals of table 17 are descriptive breakdowns of rows that were already decoded. The seed level of the bootstrap resamples three seeds, enough to put training variance inside the interval but too few to estimate it well, so the intervals on the sub-1.5-point contrasts are the least precise numbers here.

Seed variance. Per-arm seed standard deviations (percentage points, three seeds) on NExT-QA / MotionBench / LongVideoBench / TVBench are: Dense-384 (6f) 0.61 / 0.37 / 0.83 / 1.07, Dense-384 (24f) 0.37 / 0.76 / 0.61 / 0.77, anchors-only 0.52 / 0.78 / 0.79 / 0.64, Dense-512 0.71 / 0.34 / 1.23 / 0.89, Perceiver-384 0.42 / 0.48 / 0.71 / 0.98, Memory + resampler 0.42 / 0.46 / 0.99 / 0.57, StateWeaver (no QEC) 0.74 / 0.69 / 1.19 / 1.11, StateWeaver 0.70 / 0.61 / 1.04 / 0.93; the eight factorial arms span 0.36 to 0.89 on NExT-QA.

Table 19: Paired differences (percentage points) with 95% percentile intervals, $P(\Delta > 0)$, and McNemar p; official NExT-QA test (transfer benchmarks in table 20). "no QEC" is StateWeaver (no QEC). Entries left blank (—) are not reported. Every difference is computed from unrounded accuracies, so it can differ by 0.1 from the difference of the printed one-decimal values.

Comparison	Point	95% CI	$P(\Delta > 0)$	McNemar p
StateWeaver − Dense-384 (24f)	+15.0	[+13.7, +16.3]	1.000	$< 10^{-6}$
StateWeaver − Dense-512	+11.6	[+10.2, +12.9]	1.000	$< 10^{-6}$
StateWeaver − Perceiver-384	+9.9	[+8.5, +11.2]	1.000	$< 10^{-6}$
StateWeaver − Memory + resampler	+8.7	[+7.4, +10.0]	1.000	$< 10^{-6}$
StateWeaver (no QEC) − Memory + resampler	+7.6	[+6.3, +8.9]	1.000	$< 10^{-6}$
StateWeaver − anchors-only ($\gamma = 0$)	+13.8	[+12.5, +15.2]	1.000	$< 10^{-6}$
StateWeaver (no QEC) − anchors-only (the evidence write)	+12.7	[+11.4, +14.0]	1.000	$< 10^{-6}$
Anchors-only ($\gamma = 0$) − Dense-384 (24f)	+1.2	[+0.1, +2.3]	0.982	0.043
StateWeaver (no QEC) − Dense-384 (24f)	+13.9	[+12.7, +15.2]	1.000	$< 10^{-6}$
Memory + resampler − anchors-only	+5.1	[+3.9, +6.3]	1.000	$< 10^{-6}$
Perceiver-384 − Dense-384 (24f)	+5.0	[+4.0, +6.1]	1.000	$< 10^{-6}$
Dense-512 − Dense-384 (24f)	+3.4	[+2.2, +4.6]	—	—
Dense-384 (6f) − Dense-384 (24f)	−1.5	[−2.6, −0.3]	—	—
Memory + resampler − Dense-512	+2.8	[+1.6, +4.1]	1.000	$< 10^{-6}$
Memory + resampler − Perceiver-384	+1.2	[+0.1, +2.4]	0.979	0.19
StateWeaver − no QEC	+1.1	[+0.1, +2.2]	0.986	0.005
StateWeaver − matched continuation	+1.0	[0.0, +2.1]	0.975	0.054
Episode alignment − matched continuation	+0.4	[−0.6, +1.3]	0.781	0.27
StateWeaver − episode alignment (the margin)	+0.7	[−0.3, +1.6]	0.918	0.13
Full arm − all-off arm	+8.9	[+7.7, +10.1]	1.000	$< 10^{-6}$
Full arm − flat retriever	+2.4	[+1.2, +3.6]	1.000	0.0005
StateWeaver (no QEC) − question-blind planner	+2.7	[+1.4, +3.9]	1.000	0.0003
StateWeaver (no QEC) − another question's query	+3.3	[+2.6, +4.1]	1.000	$< 10^{-6}$
StateWeaver (no QEC) − random candidates	+5.1	[+4.4, +5.9]	1.000	$< 10^{-6}$
StateWeaver (no QEC) − Memory + resampler + world	+4.8	[+3.6, +6.1]	1.000	$< 10^{-6}$
Memory + resampler + world − Memory + resampler	+2.8	[+1.7, +4.0]	1.000	$< 10^{-6}$
StateWeaver (no QEC) − Memory + resampler + world (matched objective)	+3.4	[+2.1, +4.6]	1.000	$< 10^{-6}$
Memory + resampler + world (matched objective) − Memory + resampler + world	+1.4	[+0.4, +2.5]	0.996	0.004
StateWeaver (no QEC) − without ranking terms	+1.1	[+0.2, +2.1]	0.990	0.019
Question-blind planner − object + relation arm, no teacher	−0.9	[−2.0, +0.2]	0.063	0.61

Table 20: Paired differences on the three transfer benchmarks; columns as in table 19.

Comparison	Point	95% CI	$P(\Delta > 0)$	McNemar p
StateWeaver − Dense-384 (24f), MotionBench	+6.5	[+4.8, +8.2]	1.000	$< 10^{-6}$
StateWeaver − Dense-384 (24f), LongVideoBench	+6.2	[+3.4, +8.9]	1.000	$< 10^{-4}$
StateWeaver − Dense-384 (24f), TVBench	+3.6	[+1.7, +5.6]	0.9998	$< 10^{-4}$
StateWeaver − Dense-512, MotionBench	+5.4	[+3.9, +6.9]	1.000	$< 10^{-6}$
StateWeaver − Dense-512, LongVideoBench	+5.0	[+2.1, +7.8]	0.9997	0.002
StateWeaver − Dense-512, TVBench	+3.2	[+1.2, +5.3]	0.9989	0.003
StateWeaver − Memory + resampler, MotionBench	+3.9	[+2.2, +5.5]	1.000	0.0004
StateWeaver − Memory + resampler, LongVideoBench	+4.0	[+1.2, +6.8]	0.997	0.004
StateWeaver − Memory + resampler, TVBench	+2.2	[+0.3, +4.2]	0.988	0.007
StateWeaver (no QEC) − Memory + resampler, MotionBench	+3.2	[+1.5, +4.9]	1.000	0.002
StateWeaver (no QEC) − Memory + resampler, LongVideoBench	+2.6	[−0.3, +5.6]	0.963	0.035
StateWeaver (no QEC) − Memory + resampler, TVBench	+1.7	[−0.4, +3.7]	0.948	0.51
StateWeaver (no QEC) − Dense-512, MotionBench	+4.7	[+3.1, +6.3]	—	—
StateWeaver (no QEC) − Dense-512, LongVideoBench	+3.6	[+0.7, +6.6]	—	—
StateWeaver (no QEC) − Dense-512, TVBench	+2.7	[+0.5, +4.8]	—	—
StateWeaver − no QEC, MotionBench	+0.7	[−0.6, +2.0]	—	—
StateWeaver − no QEC, LongVideoBench	+1.3	[−0.9, +3.4]	—	—
StateWeaver − no QEC, TVBench	+0.5	[−1.2, +2.3]	—	—
StateWeaver − Dense-384 (24f), TVBench order-sensitive subset (545)	+10.6	[+5.4, +15.4]	0.9999	$< 10^{-3}$

Evaluation details and failure rows.

Denominators and failure rows. NExT-QA official test: 8,564 questions over 1,000 videos, zero skipped; question groups causal 4,502, temporal 2,657, descriptive 1,405. NExT-QA official validation, used for learning-rate selection and for every design-stage result: 4,996 questions over 570 videos. MotionBench validation split: 4,018 questions over 2,706 videos; LongVideoBench validation: 1,337 questions; TVBench: 2,654 questions, of which its two order-sensitive tasks, Action Antonym and Object Shuffle, have 320 and 225. A failed row is one whose EvidenceBank could not be built because SAM 2 found no track to seed (very short, dark, or static clips: MotionBench 10, LongVideoBench 6, TVBench 65); it is counted wrong for every arm that reads the EvidenceBank, identically since they share the memories, and decoded normally by the dense and Perceiver arms, so the rule is conservative against the memory arms and neutral between them.

Arms decoded on the test split. Every arm reported in this paper was decoded once on the official test split; the γ sweep, the sensitivity sweep, and learning-rate selection used validation only. The frozen evaluation protocol will be released with its hash and the date of each decode.

De-duplication. Every training video's source identifier was matched against the identifier lists of NExT-QA test and validation, MotionBench, LongVideoBench, and TVBench, which produced no collision. Because identifier spaces can hide re-uploads, every training and evaluation video was then reduced to a 64-bit perceptual hash of 8 uniformly spaced frames, and pairs within Hamming distance 6 on at least 5 of the 8 frames were inspected manually; 19 training videos were genuine re-uploads or near-duplicate clips of evaluation footage (LongVideoBench and MotionBench). Those 19 videos and their 214 conversations were removed before any arm was trained, so the 419,488-row and 39,901-video figures of table 6 are post-removal counts; no evaluation row was removed and the denominators above are unchanged.