WHEN SHOULD THE COUNT CHANGE? LEARNING STATE MAINTENANCE FOR CAUSAL VIDEO COUNTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous video counting requires distinguishing new observations from new objects or completed events. We introduce STAMINA (State Maintenance), which learns to maintain counting state through state-conditioned updates. Recurrent visual context supports recognition; learned transitions maintain visibility, persistent identities, and completed-event records. A differentiable recurrence trains event transitions over legal paths constrained by count endpoints; visibility and association objectives train the object branch. A multi-source pipeline organizes 39.8K spatial queries and complementary event annotations into counting trajectories. On SVCBench, we evaluate counting adaptation with partial video overlap and held-out groups of linked annotations. Under prefix replay (Full) and persistent streaming (Stream), 4B and 8B models reach 41.9/36.4 and 44.9/38.2 Gaussian Precision Accuracy, respectively. The 8B model gains 10.9/3.2 points over Counting-SFT on the same queries. Matched-graph comparisons isolate phase conditioning and trajectory supervision, assessing training objectives alongside hard decisions. Online video benchmarks and count-conditioned decisions assess online understanding and task eligibility.

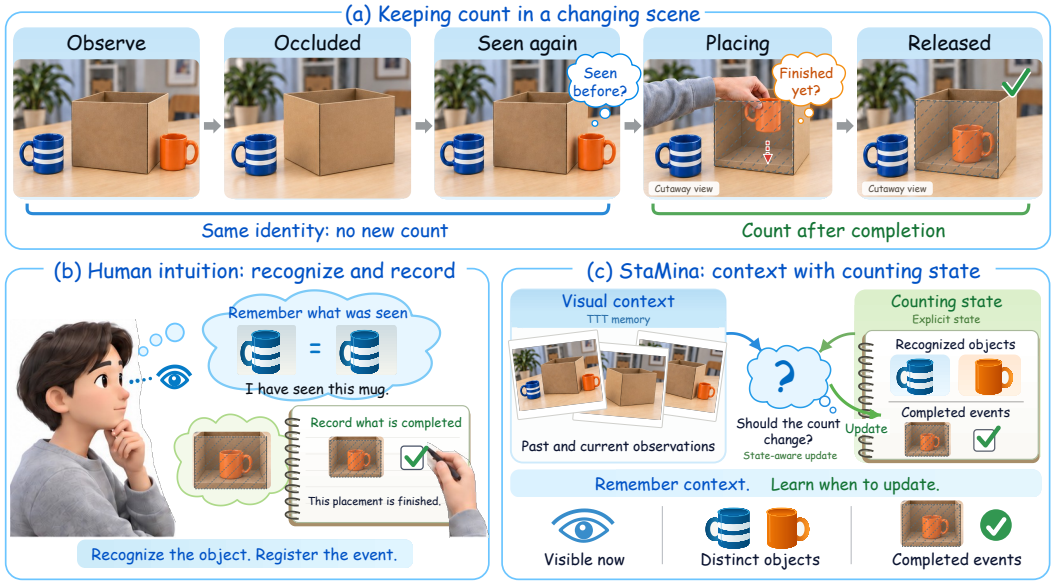


Figure 1: **Counting while packing.** (a) Reappearance preserves object identity; lowering and releasing the orange mug distinguish an ongoing placement from a completed event. Hatched cutaways show the placement stages. (b) The inventory clerk uses remembered appearance to recognize objects and a notebook to track completed placements. (c) Current evidence and previously recorded state jointly determine whether the count should change.

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

1 INTRODUCTION

A packing assistant sees a blue striped mug and an orange mug beside a box (Figure 1a). A mug that disappears from view and reappears is still the same object: visibility changes while the number of distinct mugs stays fixed. The orange mug is then lowered into the box. The placement remains in progress while the hand holds it and completes after release. Observing the placed mug again preserves that completed-event count. These cases distinguish changes in visibility, identity, and completion (Liu et al., 2026c; Yang et al., 2025a; Shangguan et al., 2025).

Long-context and streaming vision-language models retain visual evidence through attention, recurrent summaries, or learned associative memory (Bai et al., 2025a; Xu et al., 2026; Zhang et al., 2025; Yang et al., 2025d; Sun et al., 2025). Hybrid local attention and online learners extend this context along a stream (Chen et al., 2026). The inventory clerk in Figure 1b illustrates the additional role of a record: remembered appearance supports recognition, while the notebook indicates which placements have already been counted. Thus the same observation can warrant an update during an active event and a hold after completion (Figure 1c). The key is to learn when current evidence changes previously recorded state.

We introduce STAMINA (State Maintenance), which learns to maintain the counting state through state-conditioned transitions. Gated delta memory carries causal appearance and motion context; typed state records current visible candidates, accepted identities, and completed events. An event head uses visual context and the current phase to score legal updates. Training sums the probabilities of paths consistent with supported count endpoints, allowing interval counts to supervise completion and persistence without assigning a boundary to every frame. The object branch separately learns visibility and association, so that an identity decision can be deferred while a visible mug remains countable.

This formulation connects data construction to the decisions being learned. The training resource must cover both identity-preserving reappearance and the transition from held to placed. We construct such supervision from visual tracking, independent prefix counting, and verified periodic transformations. The pool contains 39829 spatial queries and complementary event records. Count endpoints constrain the number of accepted updates; reliable identities and boundaries supervise their local meaning. Snapshot and interval queries share the underlying trajectory.

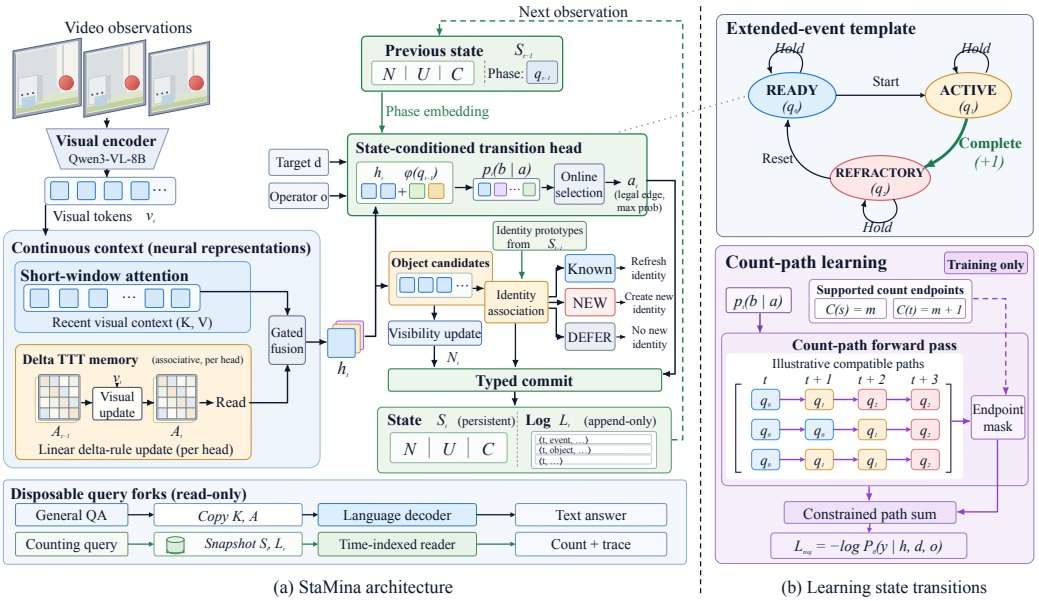
We instantiate STAMINA-4B and STAMINA-8B using the corresponding Qwen3-VL (Bai et al., 2025a) instruction models. Our contributions are:

- **State-conditioned causal counting.** A recurrent visual context and typed transition heads distinguish identity birth, visibility change, and event completion, with numerical answers read from the accepted state.
- **Trajectory-constrained transition learning.** A differentiable forward recurrence trains event transitions from supported count endpoints, while verified boundaries and identities supervise their respective operators.
- **Trajectory data and evaluation.** A multi-source pipeline supplies 39.8K spatial queries and complementary event trajectories. Evaluation at two model scales covers counting adaptation, online-video transfer on disjoint media, and count-conditioned action predicates.

2 RELATED WORK

Streaming visual context. Long-video benchmarks test temporal evidence, retrieval, and spatial reasoning (Fu et al., 2025; Wu et al., 2024; Liu et al., 2026c; Niu et al., 2025; Li et al., 2026; Liu et al., 2026b; Kwan et al., 2026; Fan et al., 2026). Online evaluation also ties answers to the observations available at a decision time (Lin et al., 2024; Niu et al., 2025). Streaming architectures maintain context through memory consolidation, decoupled perception, or selective retention (Qian et al., 2024; 2025; Xu et al., 2026; Zhang et al., 2025; Yang et al., 2025d; Zeng et al., 2025). We study the subsequent decision that converts this context into a persistent count update.

Associative memory and state transitions. Test-time training (TTT) adapts parameters online (Sun et al., 2020; Gandelsman et al., 2022; Wang et al., 2025a; Xie et al., 2025; Nath et al., 2026);



(a) StaMina architecture (b) Learning state transitions

Figure 2: **StaMina architecture and state-transition learning.** (a) Short-window attention and delta memory supply causal features for event transitions and independent visibility/identity updates. DEFER postpones identity association while preserving visible candidates. Accepted updates maintain counting state and an evidence log, with separate readouts for general QA and counting. Previous state conditions transition decisions. (b) The event graph defines legal changes. Training sums the probabilities of paths matching supported count endpoints; online inference selects a legal outgoing edge at each observation.

fast-weight layers use these parameters as recurrent state (Sun et al., 2025; Tandon et al., 2025; Dalal et al., 2025). StreamTTT uses nonlinear fast weights with short-window attention (Chen et al., 2026); Spatial TTT adds geometry-aware prediction (Liu et al., 2026a); Cambrian-S studies persistent spatial recall and counting (Yang et al., 2025b). Our continuous branch uses gated linear delta updates (Yang et al., 2025c; Kimi Team, 2025). The counting mechanism resides in the state-conditioned transition model and its trajectory objective. ActionSwitch learns action-state changes with a conservative transition loss (Kang et al., 2024). Connectionist temporal classification (CTC) marginalizes compatible alignment paths (Graves et al., 2006), while neural video alignment learns from weak action annotations (Richard et al., 2018). We constrain operator-specific paths by multiple count endpoints, carry phase posteriors between them, and evaluate causal integer commitments.

Counting data and embodied state. VSI-Bench, TOMATO, CountLLM, and AV-Reasoner study spatial reasoning, temporal reasoning, and repetitive or clue-grounded counting (Yang et al., 2025a; Shangguan et al., 2025; Yao et al., 2025; Lu et al., 2026). SVCBench distinguishes eight counting operators and their trajectories (Liu et al., 2026c). We construct aligned supervision from ScanNet, ScanNet++, ARKitScenes, and Ego4D (Dai et al., 2017; Yeshwanth et al., 2023; Baruch et al., 2021; Grauman et al., 2022). Count-conditioned predicates connect the resulting state to action eligibility, complementing observation-to-control models (Kim et al., 2024).

3 StaMina: Learning to Maintain Counting State

Figure 2 summarizes the causal context, typed state updates, and count-trajectory learning in StaMina.

3.1 Counting semantics and causal context

For target description d, let $N_d(t)$, $U_d(t)$, and $C_{d,o}(t)$ denote the true visible, first-seen, and completed-event counts, respectively. For $s < t$, O1-Snap/Delta ask for $N_d(t)/N_d(t) - N_d(s)$;

O2-Unique/Gain ask for $U_d(t)$/$U_d(t) - U_d(s)$. Event operators specify which completions count. Hats denote model predictions. Observations have timestamps τ_k; a query at t reads step $k(t) = \max\{k : \tau_k \le t\}$. The target vector e_d is the normalized mean of its pretrained token embeddings.

The runtime maintains recent key–value (KV) context K, associative memory A, typed state S, and an accepted-transition log L. Counting targets are registered at the trajectory origin; later targets require prefix replay. General QA uses the language decoder, whereas counting reads S. Query generation uses disposable forks and leaves the visual stream unchanged.

For visual-token input x_i, each memory head projects $k_i = W_k x_i$, $v_i = W_v x_i$, and $q_i = W_q x_i$ into the same width. Hats on keys and queries denote unit normalization. The gated delta update is

$$\begin{aligned} \bar{A}_i &= \gamma_i A_{i-1}, & A_i &= \bar{A}_i + \eta_i(u_i - \bar{A}_i \hat{k}_i)\hat{k}_i^\top, \\ r_i &= \hat{q}_i + \mathrm{LN}(A_i \hat{q}_i), & u_i &= \mathrm{LN}(v_i - k_i), \end{aligned} \tag{1}$$

where LN is layer normalization (Ba et al., 2016); sigmoid retention γ_i and write η_i gates lie in $(0, 1)$. Slow parameters learn projections and gates; A_i stores stream-specific associations. Gated fusion combines this linear online learner (Yang et al., 2025c; Kimi Team, 2025) with short-window attention to produce causal observation features h_k. Memory continues updating during repeated observations and unfinished actions.

3.2 State-conditioned event transitions

For operator o, a finite-state machine (FSM) defines phases $\mathcal{Q}_o$ and legal edges $\mathcal{E}_o$. A nonlinear head scores visual context, target, operator embedding e_o, phases, and edge role $r(a, b)$ (hold, start, complete, or reset):

$$p_k^o(b \mid a) = \mathrm{softmax}_{b:(a,b)\in\mathcal{E}_o}\, g_\theta(h_k, e_d, e_o, e_a, e_b, e_{r(a,b)}). \tag{2}$$

For an extended event, READY permits a start, ACTIVE permits completion, and REFRACTORY requires reset before another event. Every phase permits hold; only completion increments the count. Other operators use the templates in Appendix A. Causally associated candidates retain separate phases; concurrent counts use their joint graph.

All candidate phases share the same h_k during training. Deployment selects the highest-scoring outgoing edge, including hold. Phase feedback enters the score rather than the continuous-memory writer: the same completion evidence can support a change during an event and a hold after its acceptance.

3.3 Identity and visibility transitions

Let $\mathcal{V}_k$ contain deduplicated visible candidates and $\mathcal{I}_k$ accepted persistent identities. Target-conditioned pooling predicts visibility and association to a retrieved identity, NEW, or DEFER. A visible DEFER candidate remains in $\mathcal{V}_k$ without creating an identity. One-to-one assignment matches existing identities; NEW inserts an identity. Disappearance removes visibility while retaining identity. Predictions are

$$\widehat{N}_d(t) = |\mathcal{V}_{k(t)}|, \qquad \widehat{U}_d(t) = |\mathcal{I}_{k(t)}|. \tag{3}$$

Thus, a reappearing mug can be visible before its identity is resolved. A delayed NEW decision increments $\widehat{U}$ at acceptance, so first-seen timing errors can affect both Unique and Gain. Soft visibility and birth totals receive count supervision; verified identities supervise association. Training rolls the same predicted hard registry as deployment, with discrete assignments detached. Appendix A specifies these losses.

3.4 Learning from count trajectories

Let $\delta_o(a, b) \ge 0$ be an edge's count increment. The probability of legal paths ending at phase b and count n is

$$F_k(b, n) = \sum_{a:(a,b)\in\mathcal{E}_o} F_{k-1}(a, n - \delta_o(a, b))\, p_k^o(b \mid a). \tag{4}$$

At each supported endpoint (k_j, y_j), retain the slice $n = y_j$, sum over phases, and normalize before continuing. The product of normalization factors gives the joint probability of all supported endpoints:

$$\mathcal{L}_{\text{traj}} = -\log P_\theta(\mathbf{y} \mid \mathbf{h}, d, o). \tag{5}$$

An interval increment of two constrains paths to two completions without assigning unverified frame boundaries. Equal cumulative endpoints constrain completion-free paths; equal visible counts can still contain entries and exits.

The objective combines answer-token cross-entropy, event trajectories, soft object counts, verified identity matches, and verified phase-edge cross-entropy:

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \lambda_p \mathcal{L}_{\text{traj}} + \lambda_o \mathcal{L}_{\text{object}} + \lambda_i \mathcal{L}_{\text{id}} + \lambda_b \mathcal{L}_{\text{boundary}}. \tag{6}$$

Each term uses its available supervision sites; positive weights and masks are specified in Appendix A. Phase persists across supervised intervals. Path likelihood rewards mass on correct-count paths, whereas deployment selects one edge per observation. Their counts and completion times can differ; Section 5.3 pairs endpoint fit with hard-prefix correctness and delay, including count-only and verified-supervision strata.

3.5 Online commitment and readout

An accepted update commits to (S, L) with an evidence ID and acceptance timestamp (Fowler, 2005). Repeated delivery of that ID is idempotent; association and phase inference determine whether distinct observations depict the same instance. Observation and effective timestamps both equal τ_k, and queries read the latest committed state at or before their endpoint. Interval answers subtract two versions of the same trajectory. Acceptance never rewrites an earlier answer.

KV and associative matrices have fixed capacity; identity indexes and accepted-event logs grow outside that context. For T observations and count support width B, event training costs $O(T|\mathcal{E}_o|B)$, using the joint graph for concurrent slots. Online decoding scores outgoing edges from the committed phase. Sampling times, rather than token chunks, determine which events are observable.

4 Trajectory-Based Counting Data

The data pipeline provides the update supervision required by Section 3.4. Each trajectory fixes a source segment, target, operator, and count origin; its parent video determines the provenance and split membership. A supervision family links annotations within a parent through shared entities or events, temporal transformations, and count-derived answers. Segment-local counts reset at that segment's origin, while queries within the segment share state.

4.1 Video-to-trajectory construction

A proposal stage discovers countable objects and event descriptions from raw videos. For spatial targets, segmentation and tracking provide candidate visual evidence, while independent VLM prefix views estimate visible and cumulative counts. The acceptance stage checks agreement, supported changes, and conflicting re-identifications. Accepted records store $N_d(t)$ or $U_d(t)$ with their evidence references. These are quality-filtered automatic labels; stable instance IDs are retained where supported, and a fused count is not treated as an identity annotation.

Natural-event records use causal prefix counting and support checks for action, transition, and episode targets. Periodic records apply authorized temporal transformations to annotated source cycles, retaining the frame map and transformed completion boundaries. This gives count-trajectory supervision for all operators and boundary supervision for the subset with verified timing.

4.2 Causal query compilation and auditing

The query compiler samples supported endpoints and emits the target description, query time, causal media scope, and integer answer. O1-Delta is derived from two supported N_d endpoints; O2-Gain is

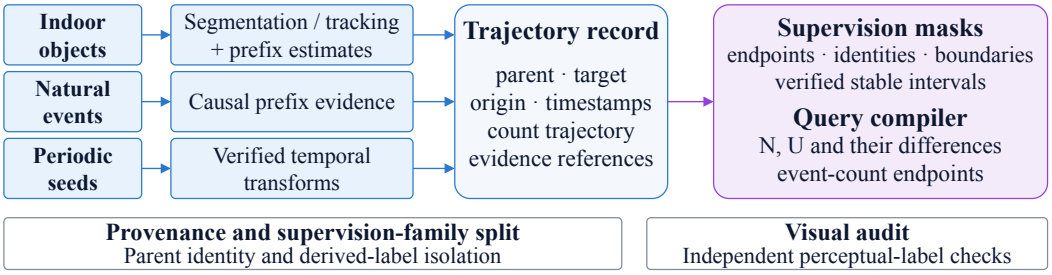


Figure 3: **Counting-data construction.** Source-specific evidence checks produce canonical trajectories and causal queries, with inherited provenance and supervision-family exclusion.

derived from two supported U_d endpoints. Verified stable intervals are retained to supervise persistence; equal visible-count endpoints alone do not establish a stable interval. Rejected or unsupported observations are omitted rather than converted into zero counts.

An independent checker recomputes the differences, validates temporal order, checks duplicate IDs and media hashes, and verifies the parent key of every derived record. This establishes algebraic and provenance consistency. Perceptual label quality is assessed separately through held-out audits and the training ablations in Section 5.4.

4.3 Dataset composition and splits

The multi-source counting pool combines indoor spatial videos, natural action videos, and transformed periodic sequences. The pool contains 15521 grouped records, 53592 query–answer pairs, 383 parent groups, and 1504 referenced media files. The spatial portion contains 39829 queries; natural events contribute 7109 and periodic transformations contribute 6654. The training and development splits of this pool contain 47882 and 5710 queries. Appendix B gives the per-operator inventory.

Training and development splits group records by source video and scene. SVCBench (Liu et al., 2026c) evaluates counting adaptation with partial source-video overlap. Held-out supervision families exclude their counts, derived intervals, and linked identity/boundary evidence across operators and temporal transformations (Appendix B). Both online-video evaluation sets and their derived media are jointly excluded from training and model selection. The inventory describes the source pool; Appendix G gives the selected training membership.

5 Experiments

5.1 Experimental setup

Models and evaluation. We train StaMina-4B and StaMina-8B from Qwen3-VL-4B/8B-Instruct (Bai et al., 2025a). Counting-SFT denotes answer-token supervised fine-tuning on the same selected counting queries, without the added memory or state heads. Controls share visual-token and update budgets within each comparison. SVCBench covers 8 operators, 406 videos, 1000 questions, and 4576 timestamped queries (Liu et al., 2026c), using the supervision-family split in Section 4 for counting adaptation. OVO-Bench and OVO-S-Bench evaluate online understanding and spatial reasoning through the language decoder (Niu et al., 2025; Li et al., 2026). One checkpoint per model size and seed is used across all three benchmarks. Appendix A fixes the model and optimization settings; Appendix B specifies training membership and supervision-family exclusion.

Access and scoring. For our models, Full resets the runtime per query and replays its causal prefix using unwindowed attention over at most 64 frames, or 128 on OVO-S. Stream processes new observations at 1 fps, retaining a 4096-token window, fast weights, and typed state. Both modes run the state interface. Counting targets are registered at the trajectory origin; matched controls receive them at the same time. General QA follows each benchmark's question-reveal schedule. Gaussian Precision Accuracy (GPA) averages query scores within each question, then across questions. Mono-

Table 1: **SVCBench counting** (%, ↑). StaMina adapts with partial source-video overlap and held-out supervision families; references retain their protocols. Groups denote weight availability; Access denotes Full/Stream. Bold/underline mark best/second-best per access mode, separately for reference and adapted rows; ties share rank. Protocols: Appendix C; shared-query controls: Table 18.

Model	Access	O1		O2		E1		E2		Overall		
		Snap	Delta	Unique	Gain	Action	Transit	Episode	Periodic	GPA	MoC	UDA
Human reference												
Human	Human	96.7	100.0	94.5	100.0	94.9	98.3	97.0	93.2	96.1	100.0	99.3
Text-only reference												
GPT-4-Turbo	Text	15.7	19.4	15.8	50.0	13.1	23.6	21.9	0.0	18.7	95.7	4.3
Proprietary VLMs												
Gemini-3-Flash	Full	**44.3**	41.0	36.5	55.3	**28.5**	**49.8**	**41.7**	**3.9**	**37.0**	73.7	73.8
Doubao-Seed-1.8	Full	43.4	**53.3**	32.5	**69.2**	24.0	38.2	40.5	0.8	36.2	77.2	**76.8**
GPT-5.4	Full	26.6	26.5	30.9	35.9	24.2	42.0	37.5	3.2	29.1	71.8	59.3
Open-weight VLMs												
Kimi-K2.5	Full	28.2	18.4	29.9	24.4	21.9	33.5	38.7	0.4	26.4	66.8	73.4
Qwen3-VL-8B	Full	21.1	35.7	34.2	55.1	22.5	36.9	35.6	0.0	31.0	84.3	55.1
Qwen3-VL-30B	Full	19.1	23.5	33.1	26.9	22.5	33.6	35.3	2.5	27.0	**84.6**	56.4
Qwen2.5-VL-7B	Full	29.7	20.4	23.8	46.2	9.5	13.3	11.0	0.0	19.1	68.2	40.8
InternVL-3.5-8B	Full	20.3	6.1	30.5	37.2	22.5	20.0	31.8	1.0	24.2	81.6	55.6
Molmo2-8B	Full	26.5	1.0	13.0	3.8	3.5	2.7	9.4	0.4	8.5	66.2	34.6
Qwen3.5-35B-A3B	Full	20.9	28.8	**37.8**	53.8	19.0	27.7	29.3	0.8	29.3	82.4	59.0
StaMina-4B (Ours, adapted)	Full	45.8	40.1	45.4	60.7	36.4	39.3	41.2	18.8	41.9	**93.5**	82.9
StaMina-8B (Ours, adapted)	Full	**50.2**	**40.8**	**47.6**	**65.9**	**38.7**	**40.0**	**46.3**	**24.2**	**44.9**	93.3	**84.5**
StreamingVLM	Stream	**36.7**	6.1	**26.0**	14.1	15.8	13.8	20.7	0.4	19.1	68.1	**50.3**
Dispider	Stream	11.4	16.3	3.3	3.8	11.8	8.9	7.6	0.0	7.7	34.8	15.5
LiveStar	Stream	24.1	13.3	20.9	26.9	16.1	**22.8**	**23.7**	**11.1**	19.9	**86.4**	36.8
Flash-VStream-7B	Stream	14.8	**45.9**	21.9	**56.4**	**19.0**	11.9	17.2	0.0	**23.4**	78.5	34.7
StaMina-4B (Ours, adapted)	Stream	39.8	33.6	39.3	56.0	28.6	32.6	40.5	15.2	36.4	**100.0**	81.0
StaMina-8B (Ours, adapted)	Stream	**42.8**	**38.1**	**39.5**	**57.3**	**31.1**	**33.8**	**42.3**	**17.4**	**38.2**	**100.0**	**82.5**

tonicity Consistency (MoC) and Update Detection Accuracy (UDA) evaluate cumulative trends and responses at true changes. Our benchmark scores average the unrounded scores of 3 training seeds. Every official query remains in the score, including graph-unreachable endpoints. Appendix C gives metric and reference-protocol details. Throughout, gray identifies the complete method; bold and underline mark the best and second-best values within the indicated comparison groups.

5.2 Counting adaptation

The complete 8B method gains 10.9/3.2 GPA points over Counting-SFT in Full/Stream on shared counting queries (Table 18). This measures the joint contribution of state structure, trajectory objectives, and auxiliary supervision; Section 5.3 isolates transition learning. The 4B/8B models in Table 1 reach 41.9/44.9 in Full and 36.4/38.2 in Stream. Both Full scores exceed the highest proprietary reference, Gemini-3-Flash (37.0).

Scaling adds 3.0/1.8 points. Full exceeds Stream by 5.4 and 6.7 points at the two sizes. These gaps compare complete access settings, including their different observation schedules. A matched-timestamp replay/persistence comparison in Appendix G holds observations and target availability fixed.

5.3 Learning state transitions

Identifying the learned decision. The event-head comparison caches causal features, candidate associations, and lifecycle masks from a common training-only warmup snapshot. With these inputs frozen, it crosses phase input with expected-count or path-likelihood training and shares verified-boundary supervision. A separate annotated diagnostic set supports exact transition errors and paired comparisons; its construction is given in Appendix G. Table 2 shows that phase conditioning and the path objective both improve hard prefix counts. Their combination also reduces false updates

Table 2: **Event-head controls on frozen 8B Stream features.** The diagnostic inputs and legal graph are fixed. GPA, hard prefix exact match, and false-update rate are percentages; a lower false-update rate is better.

Phase input	Endpoint objective	GPA ↑	Hard exact ↑	False updates ↓
Off	Mean	40.8	40.1	1.1
On	Mean	46.3	45.7	0.5
Off	Path	50.7	49.9	0.8
On	Path	**65.1**	**64.4**	**0.4**

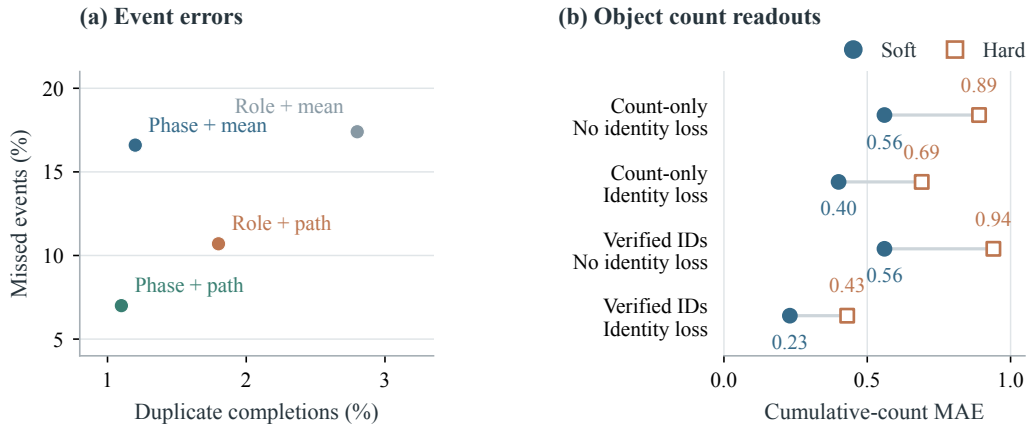


Figure 4: **Hard-decision diagnostics at 8B.** Event errors distinguish repeated or missed completions from stable-interval updates. Object endpoints compare the differentiable cumulative count with the accepted identity count under the same rollout. Detailed denominators and supervision strata are given in Appendix G.

on verified stable intervals. Hard exact match rises from 49.9 to 64.4 when phase conditioning is added to path training. The result concerns which legal transition is selected, since legal eligibility is unchanged across all rows.

From endpoint fit to accepted counts. Each control produces endpoint likelihood and hard prefix accuracy from the same checkpoint and causal rollout. The improvements persist on trajectories from endpoint-only sources, while verified boundaries provide a stronger timing signal. Phase conditioning can improve hard counts even when its mean-objective negative log-likelihood (NLL) changes little; endpoint fit and online accuracy therefore carry complementary information. Duplicate completions, missed increments, and acceptance delay jointly test whether fewer updates reflect more accurate decisions. Appendix G reports both the supervision strata and the paired intervals. The system-level no-FSM comparison, which also changes the legal graph, is reported separately in Figure 5.

Visibility and identity. Object diagnostics compare count supervision with and without the verified identity term, retaining predicted registry rollout in both cases. Soft and hard count errors are evaluated on the count-only and identity-verified source strata. Figure 4 connects the soft/hard object gap to event errors. Identity supervision reduces hard cumulative-count error alongside re-entry double-counting and unresolved identities. Cross-query Gain errors include identities accepted after their first appearance; final totals alone do not assess this timing error. Identity precision and delay use verified tracks, while count-only records remain in the endpoint-error evaluation.

5.4 Training data analysis

Supervision quality and coverage. The selected training set contains 41612 queries from 301 parents, including 11328 identity-supported and 4216 boundary-supported queries. Visual endpoint

correctness ranges from 88.8% on natural events to 96.7% on periodic records. Appendix G reports the selected training queries, overlapping source videos, available supervision masks, and stratified visual audits. Endpoint accuracy and interval accuracy are assessed separately from algebraic consistency. Training exclusions are reported alongside an observation-budget capacity check on official evaluation endpoints. All official scores retain endpoints above this capacity, exposing the effect of sparse observations on dense cycles.

Why trajectories matter. At matched token and update budgets, complete trajectories improve over general video QA by 7.0/3.4 points. Single-query training removes cross-query persistence supervision. The temporal control permutes visual observations over fixed timestamp slots while retaining endpoints, supervision masks, and candidate associations. This training corruption breaks visual–update alignment while preserving graph feasibility; evaluation uses the original sequence. Consistency filtering tests label selection. Table 19 reports these comparisons. Together with the supervision audit, they connect the data contribution to learning updates rather than to the number of QA rows alone.

5.5 Online understanding and efficiency

Capabilities after counting adaptation. The language decoder answers the original OVO-Bench and OVO-S questions. Table 3 summarizes current perception, historical evidence, response timing, and spatial reasoning. The 8B model reaches 55.46/52.08 on OVO-Bench and 50.32/47.27 on OVO-S. Gains concentrate on historical tracing and spatiotemporal context tracking; real-time perception decreases slightly, while spatial reasoning changes are mixed. All comparisons share the checkpoint and training selection for each model and seed. StreamForest and Flash-VStream retain their native protocols (Zeng et al., 2025; Zhang et al., 2024); within-family initial-model controls and full task breakdowns appear in Appendix K.

Table 3: **Online video understanding** (↑). OVO-Bench Avg. averages three task groups; OVO-S L-Avg. averages four spatial levels. Full and Stream are ranked separately; evaluation protocols are specified in Appendix K.

Model	Access	OVO-Bench Avg.	OVO-S L-Avg.
Qwen3-VL-4B	Full	50.81	46.93
Qwen3-VL-8B	Full	<u>55.02</u>	<u>48.58</u>
StaMina-4B (Ours)	Full	51.66	48.57
StaMina-8B (Ours)	Full	**55.46**	**50.32**
StreamForest-7B	Stream	**55.57**	44.13
Flash-VStream-7B	Stream	33.15	24.94
StaMina-4B (Ours)	Stream	47.82	<u>44.74</u>
StaMina-8B (Ours)	Stream	<u>52.08</u>	**47.27**

Cost and count-conditioned decisions. For general QA, mean query latency is 0.48 s in Stream and 1.72 s in Full. Appendix H distinguishes query generation, continuous ingestion, deterministic count readout, and replay for a newly revealed counting target. The fixed GPU context is complemented by a host identity index and event log. On recorded inventory, handover, restocking, and repetition videos, fixed count predicates translate state errors into eligibility decisions. The two 8B conditions share the predicates and balanced trials; exact results and decision definitions appear in Table 20 and Appendix D.

6 Conclusion

StaMina learns to maintain the counting state through state-conditioned transitions and trajectory supervision. Matched-graph and soft/hard diagnostics link these decisions to accepted counts. A multi-source pipeline supplies endpoint, identity, and timing supervision, while the shared language decoder supports online video understanding.

AI USE STATEMENT

We used generative AI tools to assist with language editing and improve the clarity of the manuscript, as well as to identify relevant literature. Vision-language models also assisted automatic annotation in the data-construction pipeline described in Section 4. The authors take responsibility for the accuracy, originality, and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

Section 3 specifies the state representation, causal memory, and learned count-state transitions of STAMINA. Appendix A details the architecture, training objectives, hyperparameters, and runtime procedure. Section 4 and Appendix B describe data construction and supervision-family exclusion. Appendices C, E, and H document the evaluation protocol, comparison settings, and computational accounting. Together, these descriptions support implementation and comparison under the stated settings. We will release the code and model checkpoints to the research community.

ETHICS STATEMENT

This work studies causal video counting and count-conditioned action predicates using existing datasets, benchmarks, and pretrained models. Use of these resources should respect their licenses and access conditions. Because errors in visual evidence and accumulated counts can affect downstream decisions, applications should include task-specific safety evaluation, appropriate operational safeguards, and human oversight suited to the deployment context.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. ARKitScenes: A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/66f041e16a60928b05a7e228a89c3799-Paper-round1.pdf.

ByteDance Seed Team. Official Release of Seed1.8: A Generalized Agentic Model, December 2025. URL https://seed.bytedance.com/en/blog/official-release-of-seed1-8-a-generalized-agentic-model. Published December 18, 2025. Accessed: 2026-09-25.

Joya Chen, Zeyun Zhong, and Mike Zheng Shou. StreamTTT: Reconciling Real-Time Perception and Long-Term Memory in Streaming VLMs. *arXiv preprint arXiv:2608.13416*, 2026. URL https://arxiv.org/abs/2608.13416.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding. *arXiv preprint arXiv:2601.10611*, 2026. URL https://arxiv.org/abs/2601.10611.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17702–17711, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dalal_One-Minute_Video_Generation_with_Test-Time_Training_CVPR_2025_paper.html.

Tulsee Doshi. Gemini 3 Flash: Frontier intelligence built for speed, December 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/. Google, on behalf of the Gemini team. Published December 17, 2025. Accessed: 2026-09-25.

Yingying Fan, Penghui Du, Leyan Zhu, Runze He, Zimeng Wu, Yuxuan Zhang, Liang Chen, Jiahao Xie, Jiangtang Wang, Shuai Shao, et al. Medclaw: Heuristic agent harness for long-horizon surgical video reasoning. *arXiv preprint arXiv:2608.14015*, 2026. URL https://arxiv.org/abs/2608.14015.

Martin Fowler. Event Sourcing, 2005. URL https://martinfowler.com/eaaDev/EventSourcing.html. Published December 12, 2005. Accessed: 2026-09-25.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. In *Advances in Neural Information Processing Systems 35*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/bcdec1c2d60f94a93b6e36f937aa0530-Abstract-Conference.html.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376, 2006. doi: 10.1145/1143844.1143891. URL https://www.cs.toronto.edu/~graves/icml_2006.pdf.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. URL https://arxiv.org/abs/1606.08415.

Hyolim Kang, Jeongseok Hyun, Joungbin An, Youngjae Yu, and Seon Joo Kim. ActionSwitch: Class-agnostic Detection of Simultaneous Actions in Streaming Videos. In *European Conference on Computer Vision*, pp. 383–400, 2024. doi: 10.1007/978-3-031-72684-2_22. URL https://link.springer.com/chapter/10.1007/978-3-031-72684-2_22.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*, 2024. URL https://arxiv.org/abs/2406.09246.

Kimi Team. Kimi Linear: An Expressive, Efficient Attention Architecture. *arXiv preprint arXiv:2510.26692*, 2025. URL https://arxiv.org/abs/2510.26692.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, et al. Kimi K2.5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276*, 2026. URL https://arxiv.org/abs/2602.02276.

Hohin Kwan, Hongyu Li, Ray Zhang, Manyuan Zhang, Xianghao Kong, Anyi Rao, Jiahao Xie, and Si Liu. Video-mme-logical: A controlled diagnostic benchmark for video temporal-logical reasoning. In *Findings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL https://arxiv.org/abs/2606.27828.

Yifei Li, Pengyiang Liu, Yuhang Zang, Zhongyue Shi, Qi Fu, Hongye Hao, and Jiwen Lu. OVO-S-Bench: A Hierarchical Benchmark for Streaming Spatial Intelligence in Multimodal LLMs. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL https://arxiv.org/abs/2606.03890.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the Gap for MLLMs to Achieve Streaming Video Understanding. *arXiv preprint arXiv:2411.03628*, 2024. URL https://arxiv.org/abs/2411.03628.

Fangfu Liu, Diankun Wu, Jiawei Chi, Yimo Cai, Yi-Hsin Hung, Xumin Yu, Hao Li, Han Hu, Yongming Rao, and Yueqi Duan. Spatial-TTT: Streaming Visual-based Spatial Intelligence with Test-Time Training. *arXiv preprint arXiv:2603.12255*, 2026a. URL https://arxiv.org/abs/2603.12255.

Pengyiang Liu, Junbo Niu, Xiaoyang Hu, Zhongyue Shi, Zitian Wang, Linjiang Huang, and Si Liu. TRACE: Temporal Retrieval with Anchored and Convergent Evidence for Long-Horizon Video Understanding. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026b. URL https://arxiv.org/abs/2608.22516.

Pengyiang Liu, Zhongyue Shi, Hongye Hao, Qi Fu, Xueting Bi, Siwei Zhang, Xiaoyang Hu, Zitian Wang, Linjiang Huang, and Si Liu. SVCBench: A Streaming Video Counting Benchmark for Spatial-Temporal State Maintenance. In *European Conference on Computer Vision*, 2026c. URL https://buaa-colalab.github.io/SVCBench/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Lidong Lu, Guo Chen, Zhu Wei, Zhiqi Li, Yicheng Liu, and Tong Lu. AV-Reasoner: Improving and Benchmarking Clue-Grounded Audio-Visual Counting for MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33477–33487, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lu_AV-Reasoner_Improving_and_Benchmarking_Clue-Grounded_Audio-Visual_Counting_for_MLLMs_CVPR_2026_paper.html.

Mriganka Nath, Anurag Das, Jiahao Xie, and Bernt Schiele. Clipttt: Clip-guided test-time training helps lvlms see better. *arXiv preprint arXiv:2603.26486*, 2026. URL https://arxiv.org/abs/2603.26486.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How Far is Your Video-LLMs from Real-World Online Video Understanding? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Niu_OVO-Bench_How_Far_is_Your_Video-LLMs_from_Real-World_Online_Video_CVPR_2025_paper.html.

OpenAI. New models and developer products announced at DevDay. https://openai.com/index/new-models-and-developer-products-announced-at-devday/, 2023. Accessed: 2026-09-25.

OpenAI. GPT-5.4 Model, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.4. Model snapshot: gpt-5.4-2026-03-05. Accessed: 2026-09-25.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming Long Video Understanding with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 119336–119360, 2024. doi: 10.52202/079017-3792. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d7ce06e9293c3d8e6cb3f80b4157f875-Abstract-Conference.html.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling Video LLMs with Active Real-Time Interaction via Disentangled Perception, Decision, and Reaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Qian_Dispider_Enabling_Video_LLMs_with_Active_Real-Time_Interaction_via_Disentangled_CVPR_2025_paper.html.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-09-25.

Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. NeuralNetwork-Viterbi: A Framework for Weakly Supervised Video Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7386–7395, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Richard_NeuralNetwork-Viterbi_A_Framework_CVPR_2018_paper.html.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. TOMATO: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models. In *International Conference on Learning Representations*, volume 2025, pp. 7593–7734, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/16ba99f25a235f1100a4014d71d34ad8-Abstract-Conference.html.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (Learn at Test Time): RNNs with Expressive Hidden States. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 57503–57522. PMLR, 2025. URL https://proceedings.mlr.press/v267/sun25h.html.

Arnuv Tandon, Karan Dalal, Xinhao Li, Daniel Koceja, Marcel Rød, Sam Buchanan, Xiaolong Wang, Jure Leskovec, Sanmi Koyejo, Tatsunori Hashimoto, et al. End-to-end test-time training for long context. *arXiv preprint arXiv:2512.23675*, 2025. URL https://arxiv.org/abs/2512.23675.

Renhao Wang, Yu Sun, Arnuv Tandon, Yossi Gandelsman, Xinlei Chen, Alexei A. Efros, and Xiaolong Wang. Test-Time Training on Video Streams. *Journal of Machine Learning Research*, 26 (9):1–29, 2025a. URL https://jmlr.org/papers/v26/24-0439.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. URL https://arxiv.org/abs/2508.18265.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857, 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Jiahao Xie, Alessio Tonioni, Nathalie Rauschmayr, Federico Tombari, and Bernt Schiele. Test-time visual in-context tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19996–20005, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xie_Test-Time_Visual_In-Context_Tuning_CVPR_2025_paper.html.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. StreamingVLM: Real-Time Understanding for Infinite Video Streams. In *International Conference on Learning Representations*, volume 2026, pp. 61463–61475, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6445dd88ebb9a6a3afa0b126ad87fe41-Abstract-Conference.html.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards Spatial Supersensing in Video. *arXiv preprint arXiv:2511.04670*, 2025b. URL https://arxiv.org/abs/2511.04670.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. In *International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=r8H7xhYPwz.

Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. LiveStar: Live Streaming Assistant for Real-World Online Video Understanding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 31266–31304, 2025d. doi: 10.52202/085713-1050. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2ce4f0b8e24c45318352068603153590-Abstract-Conference.html.

Ziyu Yao, Xuxin Cheng, Zhiqi Huang, and Lei Li. CountLLM: Towards Generalizable Repetitive Action Counting via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19143–19153, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yao_CountLLM_Towards_Generalizable_Repetitive_Action_Counting_via_Large_Language_Model_CVPR_2025_paper.html.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yeshwanth_ScanNet_A_High-Fidelity_Dataset_of_3D_Indoor_Scenes_ICCV_2023_paper.html.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, Yi Wang, and Limin Wang. StreamForest: Efficient Online Video Understanding with Persistent Event Memory. In *Advances in Neural Information Processing Systems*, volume 38, pp. 75804–75835, 2025. doi: 10.52202/085713-2547. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6dd91fec726dbed8915a1fbadd91d1d2-Abstract-Conference.html.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-VStream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024. URL https://arxiv.org/abs/2406.08085.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient Real-Time Understanding for Long Video Streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21059–21069, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Flash-VStream_Efficient_Real-Time_Understanding_for_Long_Video_Streams_ICCV_2025_paper.html.

A TRAINING AND RUNTIME SPECIFICATION

STAMINA-4B and STAMINA-8B initialize from Qwen3-VL-4B/8B-Instruct. Their processor, visual tokenization, and pretrained widths are unchanged. A delta branch parallels attention in every 4th decoder block, before the feed-forward sublayer. Each branch concatenates 8 heads of width 64, projects their readouts to the decoder width, and adds a sigmoid-gated residual to the attention output. Each square A_0 is a shared trainable parameter initialized to zero and copied at a new trajectory. Retention and write gates are sigmoids of separate linear projections of the token input, scalar per head. The channel-wise retention comparison changes this gate alone (Kimi Team, 2025). Slow weights are fixed at deployment. Scoring heads use the Gaussian error linear unit (GELU) activation (Hendrycks & Gimpel, 2016).

Table 4: Default complete-system configuration for both model scales. Pretrained dimensions follow Qwen3-VL; the frozen-input event-head controls use the training protocol in Appendix E.

Setting	Value
Memory heads $\times$ width; block stride	8×64; 4
Transition/association hidden width	256 (GELU)
Pooling queries; entity-event slots	32; 4
Identity shortlist; visibility threshold	8; 0.5
Deduplication/track cosine thresholds	0.95/0.5
Token chunk; continuous truncation	128; 8 chunks
Warmup; joint epochs; trajectory batch	256; 3; 8
Decoder/projector; new-module learning rates	`0.00002`; `0.0001`
AdamW (β_1, β_2); ϵ	$(0.9, 0.95)$; `0.00000001`
Weight decay; gradient-norm clip	0.01; 1.0
$(\lambda_p, \lambda_o, \lambda_i, \lambda_b)$	$(1.0, 1.0, 0.2, 0.2)$

Features and nonlinear scores. The target vector e_d is the unit-normalized mean of target-token input embeddings; e_o, phase embeddings, and edge-role embeddings are learned. Final decoder features of the current observation provide a mean visual vector and target-conditioned candidate vectors. For an associated candidate, their concatenation is h_k; a global target uses the mean vector in both positions. Equation 2 is instantiated as

$$z_k = W_h h_k + W_d e_d + W_o e_o, \tag{7}$$

$$s_k(a, b) = w^\top \mathrm{GELU}(z_k + W_a e_a + W_b e_b + W_r e_{r(a,b)} + c). \tag{8}$$

Operator, phase, and role embeddings and all score projections have width 256. A legal-edge softmax produces $p^o_k(b \mid a)$. Zeroing e_a, e_b retains edge roles and the legal mask. Projection matrices in different branches are independent even when their subscripts coincide. Features, target encoding, and candidate association are independent of the event path being summed.

Candidate lifecycle and event graphs. E2-Periodic always uses a single global cycle stream. For the other event operators, the target schema fixes global or entity-associated counting before observations arrive: global actions use one stream and entity-associated actions use at most 4 slots. Learned target-conditioned pooling queries attend to current visual features. Candidates are ordered by visibility, with query index breaking ties; cosine duplicate suppression retains distinct feature vectors. Candidate–slot association maximizes summed cosine similarity subject to one-to-one matches above 0.5. Equal-score assignments use candidate order and then slot-creation order. An unmatched visible candidate takes the first unused slot; excess candidates are recorded as overflow. Slot prototypes are replaced by detached matched features. Slots retain their phase and prototype through absence and are not recycled within a segment. An absent slot has only hold enabled. Association and creation are causal, detached, and identical across mechanism controls.

New slots start READY. E1-Action completes on $\mathrm{READY} \to \mathrm{REFRACTORY}$ and re-arms on observed release/renewed eligibility. E1-Transit uses the same graph: reaching the requested destination completes; re-establishing its source resets. E2-Episode uses $\mathrm{READY} \to \mathrm{ACTIVE} \to \mathrm{REFRACTORY} \to \mathrm{READY}$, with completion only on $\mathrm{ACTIVE} \to \mathrm{REFRACTORY}$. E2-Periodic uses $\mathrm{READY} \to \mathrm{AWAY}$ for departure and

AWAY $\rightarrow$ READY for the specified return, which completes one cycle. Every phase allows hold, and there is no extra time guard. Initial phases never use annotation labels; incomplete observation of a left-censored event is evaluated under the same initialization.

For concurrent slots ℓ, the joint transition is

$$p_k(\mathbf{b} \mid \mathbf{a}) = \prod_{\ell} p_{k\ell}(b_\ell \mid a_\ell), \qquad \delta(\mathbf{a}, \mathbf{b}) = \sum_{\ell} \delta_o(a_\ell, b_\ell). \tag{9}$$

The forward state retains the joint phase tuple after conditioning on a shared count, so posterior dependencies between slots are preserved. Candidate association is fixed within this sum. Hard decoding maximizes this product from the committed tuple, equivalently selecting each slot's best outgoing edge. Ties prefer hold, then the template edge order. Same-time completions across slots are summed. With m allocated slots, these templates have at most 3^m phase tuples and 2^m outgoing edges per tuple.

Sampled time and reachable counts. Let source-frame times be ρ_j. If a prefix contains M frames and its budget is B_f, Full takes $m = \min(M, B_f)$ frames at indices $\lfloor r(M-1)/(m-1) \rfloor$, $r = 0, \ldots, m-1$; for $m = 1$, it takes index zero. Stream takes the latest source frame at each grid time from the segment origin at 1 fps, suppressing duplicate frame indices. Observation timestamps τ_k are the selected source-frame times. Query and interval boundaries introduce no extra frames. A read at s uses $k(s)$, or the empty initial state if no observation precedes s. Full interval endpoints are read from the same replay used for that query. Labels retain their physical query times: a change between the last sample and a query remains a scored error if unobserved. Accepted changes have observation and effective time τ_k; later evidence cannot backdate an answer.

Each slot traverses at most one edge per observation. The global periodic stream therefore completes at most $\lfloor T/2 \rfloor$ cycles in T observations. A single E1 slot completes at most $\lceil T/2 \rceil$ events, and an episode slot at most $\lfloor (T+1)/3 \rfloor$. Allocating all 4 entity slots from the first observation gives the corresponding aggregate upper ceilings. The run checks reachable support using actual timestamps, slot creation/absence, and initialization. Graph-incompatible training endpoints are logged and excluded. Every official evaluation query remains scored, including overflow and unreachable endpoints. Increasing the count array cannot recover unobserved transitions.

Endpoint and boundary supervision. Initially $F_0(a, n) = \mathbf{1}[a = a_{\mathrm{READY}}, n = 0]$ for the joint ready state; inactive slots are constrained to hold until creation. Endpoint j supplies $Z_j = \sum_a F_{k_j}(a, y_j)$. We retain $F_{k_j}(a, y_j)/Z_j$ and minimize $-\sum_j \log Z_j$, carrying its phase posterior onward. After conditioning, subtracting y_j from the count coordinate preserves its offset and phase posterior. Count support spans all reachable interval increments and computations use log probabilities. For an interval-only label, the count coordinate starts at zero at the interval's left endpoint while inheriting the marginal phase distribution from the preceding observations. Interval-only chains are chronological and non-overlapping; overlapping alternatives use separate causal replays and share their supervision-family weight. Interval labels directly supervise count increments. Shared endpoints and derived differences form one supervision family rather than repeated independent losses.

Verified phase-edge records form $\mathcal{M}_B = \{(k, \ell, a^*, b^*)\}$. A physical transition at τ^* maps to the first sampled observation with $\tau_k \geq \tau^*$. A site is used only if the associated instance and both phases are verified and $(\tau_{k-1}, \tau_k]$ contains exactly that legal edge, or verified persistence for a hold. Multiple hidden transitions, uncertain phases, and unmatched instances mask out the site. The loss is conditional edge cross-entropy,

$$\mathcal{L}_{\mathrm{boundary}} = -\frac{1}{|\mathcal{M}_B|} \sum_{(k,\ell,a^*,b^*) \in \mathcal{M}_B} \log p^o_{k\ell}(b^* \mid a^*). \tag{10}$$

It uses causal features at τ_k; boundary times and phases are supervision, not runtime inputs. No observation after a query is inserted to satisfy a boundary label. Verified hold and change sites receive equal sampling weight. An empty mask contributes no loss. Count-only records use endpoint slices. The recurrence is exact for the given causal features, association, and event graph; endpoint fit and greedy commitment are evaluated separately.

Visibility, association, and delayed identity acceptance. The object branch uses 32 target-conditioned pooling queries. With current visual vectors v_{kj} and learned queries q_i, its candidate

is $x_{ki} = \sum_j \alpha_{kij} v_{kj}$, where $\alpha_{kij} \propto \exp[(q_i + W_d e_d)^\top W_v v_{kj}/\sqrt{D}]$ and $D = 256$ is the attention width. Visibility is $\nu_{ki} = \sigma(w_v^\top x_{ki} + b_v)$. A candidate is suppressed when its cosine similarity to an earlier survivor is at least 0.95, before the visibility threshold 0.5. Hard visibility uses the surviving candidates above that threshold; soft losses sum over all deduplicated candidates before thresholding. Entity-event pooling uses the same construction.

For each candidate, the 8 nearest historical prototypes by cosine similarity form $\mathcal{R}_{ki}$ (all available identities if fewer). The association multilayer perceptron (MLP) has the same hidden width as the transition head. The operator sg preserves its input value and stops its gradient:

$$p_{ki}(c) = \mathrm{softmax}_{c \in \mathcal{R}_{ki} \cup \{\mathrm{NEW}, \mathrm{DEFER}\}}\, g_{\mathrm{id}}(x_{ki}, e_d, \bar{q}_c),$$
$$\bar{q}_c = \begin{cases} \mathrm{sg}(q_c), & c \in \mathcal{R}_{ki}, \\ q_c, & c \in \{\mathrm{NEW}, \mathrm{DEFER}\}. \end{cases} \tag{11}$$

The score is $g_{\mathrm{id}} = w_{\mathrm{id}}^\top \mathrm{GELU}(W_x x + W_d e_d + W_q \bar{q}_c + c_{\mathrm{id}})$, with independent projections. NEW and DEFER have trainable class embeddings with the same width as candidate features and stored prototypes. For visible candidates, one-to-one assignment maximizes the sum of log probabilities, with a private NEW/DEFER alternative per candidate. Each existing identity receives at most one match. Equal-score assignments are resolved in candidate order, preferring DEFER, then the oldest identity, then NEW. Matches replace prototypes with detached features; NEW appends an identity; DEFER keeps visibility without insertion. Persistent identities are never evicted within a trajectory. A deferred candidate is reconsidered from current evidence when observed again.

A visible unresolved object immediately contributes to $\widehat{N}$. An old-identity resolution leaves $\widehat{U}$ unchanged; NEW increments it at acceptance. The target U counts physical first appearances, so delayed acceptance can yield $\widehat{U}(t) - \widehat{U}(s) \neq U(t) - U(s)$ even after the final unique count becomes correct. We measure both acceptance delay and cross-query Gain errors without rewriting earlier prefixes.

Object losses and gradient boundaries. Soft visible and birth totals are

$$\widetilde{N}_k = \sum_i \nu_{ki}, \quad \widetilde{b}_k = \sum_i \nu_{ki} p_{ki}(\mathrm{NEW}), \quad \widetilde{U}_k = \sum_{j \le k} \widetilde{b}_j. \tag{12}$$

The birth accumulator resets at the count origin. It is a surrogate conditioned on predicted identity history, not a marginalization of identity paths. Let $\mathcal{J}_N, \mathcal{J}_U$ index supported visible and unique endpoints and $\mathcal{M}_{\mathrm{id}}$ verified candidate/identity labels. Then

$$\mathcal{L}_{\mathrm{object}} = \mathrm{mean}_{j \in \mathcal{J}_N} (\widetilde{N}_{k_j} - y_j^N)^2 + \mathrm{mean}_{j \in \mathcal{J}_U} (\widetilde{U}_{k_j} - y_j^U)^2, \tag{13}$$
$$\mathcal{L}_{\mathrm{id}} = -\,\mathrm{mean}_{(k,i) \in \mathcal{M}_{\mathrm{id}}} \log p_{ki}(c_{ki}^*). \tag{14}$$

Empty sets contribute no term; interval labels use corresponding soft differences. An identity label is admitted only for a unique verified match inside the shortlist or a verified first appearance eligible for NEW. Ambiguity, missing identities, and retrieval misses mask out identity CE and never become NEW or DEFER labels. Retrieval misses are recorded separately in the error log. $\mathcal{L}_{\mathrm{ans}}$ is the mean answer-token cross-entropy on supported causal QA records. Counting answers also supervise this decoder during training; deployed counting still reads typed state.

Training advances the same predicted hard registry as deployment. Selection, deduplication, retrieval, assignment, insertion, and historical writes are detached. Gradients reach retained current features and visibility/association heads, including the NEW and DEFER embeddings; soft accumulators remain connected until supervised endpoints. No count label substitutes for a predicted identity. Soft/hard count error is measured in both count-only and identity-verified source strata. NEW precision/recall, re-entry double counting, unresolved rate, and acceptance delay use verified identities; unresolved cases remain failures rather than disappearing from the denominator.

Optimization and resets. The vision encoder stays frozen. During 256 warmup updates, only new memory branches and pooling/transition/association heads are trained with available answer, object, identity, and boundary terms. Joint training follows the Stream observation and KV schedule and updates those modules, the decoder, and visual projector for 3 epochs with the full objective, cosine

learning-rate decay to zero, and the AdamW (Loshchilov & Hutter, 2019) settings in Table 4. A batch contains 8 trajectories; token, object, identity, and boundary losses average over available sites within each trajectory, while the trajectory term retains its joint NLL. Each term is then averaged over contributing trajectories. Each complete-system final checkpoint is used across all benchmarks. Appendix E defines the separate frozen-input event-head controls.

Continuous-memory backpropagation spans 8 token chunks while forward state persists. The event forward graph and soft object accumulators remain connected to their supervised trajectory losses; optimizer updates follow the whole trajectory batch. At a visual step, context updates, scores are computed, and hard changes commit before query readout. Query forks are discarded. A new segment/count origin resets KV, slots, identities, logs, and matrices to their initial states; queries inside it do not reset state. Full reconstructs the prefix with unwindowed causal attention; Stream persists bounded context and typed state.

B Data Inventory and Quality Semantics

Table 5: Composition of the multi-source pool by counting operator. Entries count query–answer pairs, not independent source videos.

Operator	Queries
O1-Snap	16379
O1-Delta	6746
O2-Unique	10627
O2-Gain	6077
E1-Action	4606
E1-Transit	885
E2-Episode	1618
E2-Periodic	6654
Total	53592

A grouped record contains a question and a sequence of supported query times and counts. Expanding it into SFT rows changes the physical row count but not the number of underlying trajectories. Delta and Gain reuse endpoints from their source trajectory. Multiple media files can be crops or temporal transformations of one parent. We therefore retain parent groups, grouped records, query points, and referenced media as separate inventory units.

Spatial and natural-event labels are automatically supported count trajectories. They provide supervision for count changes and persistence. Periodic transformations retain verified frame maps from annotated seeds and support temporal boundary training. Algebraic replay checks derived answers, while visual audits assess whether the source evidence supports the counts. Unsupported targets and failed generation attempts remain outside the accepted training pool.

Supervision acceptance and selection. Every accepted endpoint links to its causal media scope and source evidence. Finite integer counts, ordered local timestamps, consistent endpoint differences, and operator-valid trends are required. Conflicting visual evidence disables the affected supervision mask; algebraic agreement alone does not create identity or boundary labels. Periodic boundaries additionally require the transformed frame map. Visual audits separately assess exact endpoint and interval-increment accuracy, stratified by source and operator.

A trajectory key contains parent, derived segment/frame map, canonical target, operator, and count origin. Exclusion operates on supervision families before SFT expansion. For one parent, aliases and overlapping entity/event scopes are linked across every timestamp, crop, speed change, and repeated-cycle derivative. The dependency graph links N to Delta, U to Gain, shared identity tracks to both object families, and event counts to their supporting phase/boundary annotations. Overlapping operator definitions and any QA answer derived from these nodes join the same connected component. If a component intersects a held-out benchmark target, all its count, interval, identity, boundary, and answer-token supervision is excluded from training and model selection. Consequently, hiding Gain alone cannot leave its two Unique endpoints available. Uncertain semantic matches are excluded rather than separated by wording.

SVCBench evaluates adaptation to held-out supervision families with partial source-video overlap. The family exclusions apply to all project-specific training stages and model selection, including warmup, answer fine-tuning, and transition training. Training and development records are grouped by source scene and video before these exclusions. For transfer, the union of OVO-Bench and OVO-S evaluation media, original source videos, and all derived media is removed from every training and development selection. One final checkpoint per scale and seed is then shared across all three benchmarks. Each run records family IDs, ancestry, exclusions, selected queries, and supervision-mask counts; the source-pool inventory is reported separately.

Target-reveal protocol. For counting, the target description and operator are registered at the trajectory origin; future answers and completion boundaries are not runtime inputs. Matched controls receive the same target information. Each target/operator trajectory is evaluated separately, so a resident-target timing measurement does not imply an unbounded number of concurrent targets. A late target replays the available prefix to construct its state, and that replay cost is included in its end-to-end latency. General QA follows each benchmark's original question-reveal schedule.

C Reference Models and Scoring

Table 1 includes the complete reference roster from SVCBench's camera-ready object and event tables (Liu et al., 2026c). Human performance and the text-only GPT-4-Turbo control (OpenAI, 2023) provide context. The proprietary group contains Gemini-3-Flash (Doshi, 2025), Doubao-Seed-1.8 (ByteDance Seed Team, 2025), and GPT-5.4 (OpenAI, 2026). The open-weight group contains Kimi-K2.5 (Kimi Team et al., 2026), Qwen3-VL-8B/30B and Qwen2.5-VL-7B (Bai et al., 2025a;b), InternVL-3.5-8B (Wang et al., 2025b), Molmo2-8B (Clark et al., 2026), Qwen3.5-35B-A3B (Qwen Team, 2026), the streaming models StreamingVLM, Dispider, LiveStar, and Flash-VStream-7B (Xu et al., 2026; Qian et al., 2025; Yang et al., 2025d; Zhang et al., 2025), and StaMina. Model groups reflect weight availability; the Access column specifies Full or Stream. Within each access mode, reference models and StaMina adaptation rows are ranked separately across weight-availability groups.

Qwen3.5-35B-A3B retains its published aggregate over 998 scored questions, including 54 periodic questions rather than the nominal 56. Human performance also retains its published aggregate, rather than a reconstruction from rounded categories.

Full denotes per-query access to a causal video prefix; Stream denotes sequential ingestion with persistent state. Each model retains its input sampling protocol. Most Full references use the benchmark's 64-frame sampling protocol, while Gemini-3-Flash uses native 1 fps input. The Stream references follow their native 1 fps inference. Reference scores come from the original SVCBench evaluations; Table 18 compares system variants on shared counting queries.

$$\mathrm{GPA}_q = \frac{1}{n_q} \sum_i \exp\left[-\frac{(\hat{y}_{qi} - y_{qi})^2}{2\sigma_{qi}^2} \right], \qquad \sigma_{qi} = 0.05 \max(y_{qi}, 1). \tag{15}$$

For GPA, first average query-point scores within each question, then average across questions. Overall is consequently weighted by category question counts, not an equal mean of eight category cells. MoC applies to O2-Unique and the four cumulative event categories. UDA applies to O1-Snap, O2-Unique, and the four event categories. Single-query differences do not have trajectory metrics. UDA evaluates changes on ground-truth change steps; it does not measure false updates during unchanged steps. Stable-interval false updates are evaluated in the event diagnostics.

Our SVCBench results use the arithmetic mean of unrounded per-seed question scores over 3 seeds; all own SVCBench conditions score the common 1000 questions, with category weights determined by their question counts. Means and differences are computed before display rounding. The confidence intervals shown for the two event-head contrasts use 10000 parent-level resamples and paired predictions. Parent/query IDs, invalid flags, actual frame times, checkpoint identities, and scoring code identify each comparison. Published reference aggregates retain their source evaluation populations.

D Intervention and Embodied Evaluation Details

The no-FSM condition is retrained with an independent increment head and commits above the fixed confidence threshold 0.5. It removes phase conditioning and the legal-transition constraint while retaining the same continuous memory, supported intervals, and verified-boundary records. Count-state readout controls use the complete model's fixed rollout. Clear replaces its count projections and readout log with the empty state. Delay maps each requested endpoint to the previous distinct read endpoint for the same target, including interval starts, or to the empty origin when none exists. Both interval endpoints use that same mapping. Swap fixes a different-video donor with a compatible object or event schema, preferring an identical target when available. Stream reads the donor's persistent state at the recipient's elapsed times; Full uses the latest donor prefix no later than the recipient query, or the empty state if unavailable. Both interval endpoints use one donor history. These probes test the indexed count payload, including its availability; visual processing and neural state updates remain fixed.

Each routine uses 40 independent recorded videos with 5 decisions per video, giving 200 decisions balanced between eligible and ineligible cases. These parents are disjoint from training, all three benchmarks, and the core diagnostic collections. The fixed predicates are inventory $N \geq 4$, handover $C \geq 1$, restocking $N \leq 2$, and repetition completion $C \geq 5$. Verified video counts determine eligibility at the fixed thresholds above. Counting-SFT and the complete 8B model both use Stream access and answer at the same timestamps; their integer counts enter the same predicate function. Invalid or unavailable counts are scored as incorrect. For each model, accuracy pools one prediction per decision within each routine, dividing the number of correct eligibility decisions by 200.

E Transition-Learning Comparisons

Table 2 crosses phase conditioning with the endpoint objective on fixed causal inputs; detailed results appear in Appendix G. For each 8B seed, the common training-only warmup snapshot is frozen before the four heads diverge. Its chronological Stream pass caches observation and target features, candidate associations, slot creation/absence masks, and thus the actual legal graph and count support. Training and diagnostic caches use the same frozen snapshot; diagnostic labels never enter this pass. The recurrent memory still updates causally while creating the cache.

All heads start from the same warmup scorer parameters. Only the transition scorer and its operator, phase, and role embeddings are optimized; the decoder, memory, pooling, and association parameters remain frozen. The four conditions use the same ordered batches of 8 training trajectories and the same cached observation/token inputs, endpoint labels, boundary masks, initial phases, and greedy decoder. Each receives 1000 updates at initial learning rate `0.0001`, with fresh AdamW states and cosine decay to zero; other optimizer settings follow Table 4. Phase-off zeroes e_a, e_b and retains semantic edge roles and legal eligibility. The expected-count control uses unconditioned forward marginals $\bar{F}$, without endpoint-label slicing:

$$\mu_j = \sum_{a,n} n \bar{F}_{k_j}(a, n), \qquad \mathcal{L}_{\text{mean}} = \text{mean}_{j \in \mathcal{J}} (\mu_j - y_j)^2. \tag{16}$$

For interval labels, the expected increment is the difference of the two unconditioned means. Head optimization uses $\lambda_p \mathcal{L}_{\text{mean}} + \lambda_b \mathcal{L}_{\text{boundary}}$ or $\lambda_p \mathcal{L}_{\text{traj}} + \lambda_b \mathcal{L}_{\text{boundary}}$, with the same weights; answer, object, and identity terms have no trainable parameters in this control. The phase-conditioned path head instantiates the complete event-transition rule under fixed inputs. Its control checkpoint is distinct from the jointly trained benchmark model, so this comparison identifies transition learning conditional on the shared visual evidence. Independent increments, fixed phase filters, and memory-cell comparisons remain separate complete-system controls.

Diagnostic populations and commitment. Diagnostic trajectories are independent of the training and model-selection parents. Endpoint-only and boundary-supervised strata describe the source's available training annotations; all diagnostic endpoints are verified for scoring and are never optimization inputs. Within each event-head condition, the same control checkpoint and cached causal inputs produce endpoint NLL and greedy hard-prefix predictions. Reported NLL divides the joint conditional negative log-likelihood by the number of verified endpoints. Predictions are matched chronologically within a slot to the nearest unmatched true completion in the preceding 3 seconds.

Additional predictions within a true-completion window are duplicates; unmatched true completions are misses and predictions outside all windows are false updates. Stable false-update rate counts verified sampling intervals containing any commit, excluding intervals within 3 seconds on either side of a true completion. False-update rate divides marked stable intervals by all eligible stable intervals; duplicate and miss rates divide their counts by ground-truth events. These ratios pool counts across trajectories and seeds. Delay averages prediction time minus matched completion time over all matched events; unmatched events remain misses. Parent-level paired intervals quantify the factorial contrasts. Low NLL and poor hard counts identify a remaining surrogate/decoder gap; delay is reported alongside accuracy so conservative holds cannot appear successful merely by suppressing updates.

Object controls follow complete-system joint training, setting λ_i to zero or retaining the complete objective. Registry-update rules are shared, and each condition rolls its own predicted identity history. Both count-only-source and identity-verified-source records contribute soft/hard Unique error and supported Gain error. NEW precision/recall, re-entry duplication, unresolved identities, and acceptance delay use the verified subset, and retrieval misses are identified in the error log. Soft/hard Unique MAE uses supported endpoints; Gain MAE uses consecutive endpoint differences, including the empty pre-observation origin before the first query. NEW precision counts the first accepted birth of a true identity as a true positive and repeated or unmatched births as false positives; recall divides true positives by the number of true identities. Re-entry duplication is measured per re-entry encounter. Unresolved rate sums seen-but-unaccepted identities over queries and divides by all seen-identity exposures at those queries. Identity ratios pool their numerators and denominators across trajectories and seeds. Delay averages acceptance minus first-seen time over all resolved identities, with unresolved cases retained in the unresolved rate. Cross-query errors retain the original first-seen target time even when identity is accepted later. Soft birth fitting and hard identity acceptance are compared empirically rather than equated.

Observations and temporal controls. The matched-access experiment uses the jointly trained benchmark checkpoints. Replay and persistence receive the same causal frames, target revelation, and queries. It measures the access implementation, including replay versus persistence and unwindowed versus bounded attention, after removing the standard sampling difference. The capacity diagnostic compares each ground-truth endpoint with an observation-budget ceiling: $\lfloor T/2 \rfloor$ for the global periodic stream, $4\lceil T/2 \rceil$ for E1, and $4\lfloor (T+1)/3 \rfloor$ for episodes. Full uses its frame-budget ceiling; Stream has at most $T = \lfloor (t - t_0)f \rfloor + 1$ observations for origin t_0 and sampling rate f. Coverage below these ceilings is an upper bound on feasible endpoint coverage; actual slot availability, discarded duplicate frames, and joint endpoint constraints can reduce it. Speed/pause transformations preserve event identity and map boundaries through their frame ancestry. With the target and observation schedule fixed, Stream state updates are independent of when that state is queried. Standard Full resamples and replays each queried prefix, so changing query times can change its visual processing. Whole-action duplication increases the annotated count. Prefix decoding, target replay, and reset rules follow Appendix A in every condition.

F Scope and Failure Modes

Candidate overflow, missed observations, and identity association errors can precede transition scoring. Fixed event slots preserve phase through absence but bound the represented entity streams. The accepted-identity index and event log grow with distinct objects and accepted events, even while neural context stays bounded. Endpoint likelihood and soft births are training objectives; online integer correctness and acceptance timing are measured on hard rollouts. The transfer benchmarks test the shared checkpoint on excluded media, while SVCBench measures the declared same-video new-question adaptation.

G Mechanism and Data Diagnostics

Diagnostic collections. The event collection contains 160 trajectories from 40 parent videos, balanced across the four event operators. The object collection contains 160 trajectories from 40 parents. Each trajectory has 5 endpoint queries. These additional videos come from the source collections used by the pipeline and are excluded from training and development before model fitting; they

Table 6: Event hard decisions and trajectory fit. GPA, exact match, duplicates, and misses are percentages. NLL is normalized by the number of supervised endpoints and delay is in seconds.

Phase	Objective	GPA ↑	Exact ↑	NLL ↓	Duplicate ↓	Miss ↓	Delay ↓
Off	Mean	40.8	40.1	0.93	2.8	17.4	1.11
On	Mean	46.3	45.7	0.94	1.2	16.6	0.93
Off	Path	50.7	49.9	0.67	1.8	10.7	1.15
On	Path	**65.1**	**64.4**	**0.58**	**1.1**	**7.0**	**0.89**

are not counted in the source-pool inventory. Every condition uses the same parents, observations, targets, query times, and 3 training seeds.

Event strata identify whether the source pipeline supplies count endpoints alone or also verified boundaries. Dense temporal annotations are added to both diagnostic strata for evaluation and are never used to fit their models. Object count-only sources have verified visible and cumulative endpoints; identity-supported sources additionally have verified physical tracks. Identity metrics use the latter group, while both groups contribute count errors. Each collection is balanced between its two source strata. This grouping distinguishes available training supervision from evaluation annotations.

Paired reporting. GPA, hard exact match, endpoint-normalized NLL, and mean absolute error (MAE) average within trajectories and then across trajectories and seeds. Ratio metrics instead pool numerator and denominator across trajectory–seed records: false updates use eligible stable intervals; duplicates and misses use ground-truth events; identity ratios use their specified eligible identity units. Delay pools all matched completions or resolved identities. Parent-level resampling retains all trajectories and all seed predictions belonging to a selected parent. The two event-head contrasts use the same resampled parent indices for both conditions and 10000 replicates. Their intervals describe variation across parents conditional on the evaluated seed set. The factorial event study trains separate head-only control checkpoints on common frozen inputs. The complete object and matched-access evaluations use the end-to-end benchmark checkpoint; the identity-loss ablation is trained separately. With path likelihood fixed, phase conditioning adds 14.4 GPA points (95% interval: 11.4–17.3); with phase conditioning fixed, path likelihood adds 18.8 points over expected-count training (15.9–21.7). These are conditional contrasts on the common frozen inputs.

Event decisions and endpoint fit. Table 6 complements Table 2 with endpoint negative log-likelihood (NLL), duplicate completions, missed events, and acceptance delay. NLL is computed with the same edge probabilities that drive greedy decoding. Event matching is performed chronologically within an instance; a matched completion contributes a delay, an unmatched target remains a miss, and excess accepted completions remain errors. Verified intervals with no event provide the false-update denominator. Delay is reported together with misses, so postponing every update cannot produce an apparently good result. Table 7 separates endpoint-only and boundary-supported sources.

Object counts and identity acceptance. The object ablation sets $\lambda_i = 0$ and keeps all other training terms, masks, budgets, and registry operations fixed. Table 8 reports mean absolute errors for soft and hard cumulative counts and their interval differences. Table 9 evaluates NEW decisions and re-entry behavior on verified identities. An unresolved identity stays in the recall and unresolved-rate denominators. Gain is evaluated between adjacent endpoints, including the empty pre-observation origin before the first query, so a late birth can incur an interval error even when the final unique count matches. Both conditions use the same candidate budget and selection rules.

Matched observation schedules. Table 10 supplies identical timestamps and target availability to Full replay and Stream persistence on the event diagnostic set. Full uses unwindowed prefix attention while Stream retains its fixed window and recurrent context. This comparison measures the remaining access differences with sampling controlled; it is distinct from the official-budget results in Table 1.

Table 7: Event diagnostics by source supervision. Evaluation annotations are held out; each source stratum is ranked separately.

Phase	Objective	Source stratum	GPA ↑	Exact ↑	NLL ↓
Off	Mean	Endpoint only	38.1	37.4	0.95
Off	Mean	Boundary supported	43.5	42.8	0.90
On	Mean	Endpoint only	41.6	41.0	0.98
On	Mean	Boundary supported	51.0	50.3	0.90
Off	Path	Endpoint only	45.3	44.5	0.71
Off	Path	Boundary supported	56.2	55.3	0.62
On	Path	Endpoint only	**57.9**	**57.2**	**0.63**
On	Path	Boundary supported	**72.3**	**71.6**	**0.54**

Phase	Objective	Source stratum	Duplicate ↓	Miss ↓	Delay ↓
Off	Mean	Endpoint only	2.8	19.7	1.13
Off	Mean	Boundary supported	2.9	15.0	1.10
On	Mean	Endpoint only	**1.0**	19.5	0.92
On	Mean	Boundary supported	1.4	13.7	0.93
Off	Path	Endpoint only	2.0	13.3	1.15
Off	Path	Boundary supported	1.7	8.0	1.14
On	Path	Endpoint only	1.1	**9.7**	**0.88**
On	Path	Boundary supported	**1.2**	**4.2**	**0.90**

Table 8: Soft and hard object errors. Each column reports MAE (↓) on the same endpoint or interval queries; ranks compare the two training conditions within each source stratum.

Identity loss	Source stratum	Soft U	Hard U	Soft Gain	Hard Gain
$\lambda_i = 0$	Count only	0.56	0.89	0.55	0.68
$\lambda_i = 0$	Identity verified	0.56	0.94	0.56	0.69
Complete	Count only	**0.40**	**0.69**	**0.39**	**0.50**
Complete	Identity verified	**0.23**	**0.43**	**0.25**	**0.31**

Training selection and supervision coverage. Table 11 separates the source pool from the selected training membership. The selected data span 301 parents; 76 overlap SVCBench and contribute 9216 retained queries from new supervision families. Selection removes evaluation-media descendants, held-out supervision families and graph-incompatible training endpoints in that order, recording each removal once. Parent overlap with SVCBench describes visual exposure; family-level exclusion removes linked counts, intervals, identities and boundaries. A Gain interval cannot be held out while its determining Unique endpoints remain in training. Counts linked by overlapping intervals or the same physical identity/event evidence belong to one exclusion component, including transformed time coordinates.

All selected queries carry supported count labels. Table 12 partitions them into endpoint-only, identity-supported object, and boundary-supported event records. Verified stability overlaps these strata and is reported separately, so its count is not added to the training-set size. Auxiliary labels require the corresponding verified evidence.

Visual quality and reachability. The audit samples grouped records by source group, restricted to records with both supported endpoints and positive-duration intervals. Each sampled record contributes one endpoint and one interval, drawn independently from its eligible sets without reference

Table 9: Identity decisions on verified tracks. Precision, recall, re-entry double-counting and unresolved rates are percentages; delay is in seconds.

Identity loss	NEW precision ↑	NEW recall ↑	Re-entry ↓	Unresolved ↓	Delay ↓
$\lambda_i = 0$	88.1	87.9	11.8	18.3	3.04
Complete	**96.3**	**95.1**	**3.7**	**7.1**	**1.08**

Table 10: Matched-timestamp access on the event diagnostic collection. Both conditions use the complete 8B model.

Access	GPA ↑	Hard exact ↑
Full	**63.2**	**62.4**
Stream	60.9	60.1

Table 11: Training selection from the source pool. Counts refer to queries unless a row specifies parents.

Selection statistic	Count
Source train pool	47882
OVO media exclusion	2710
Dependency-family exclusion	2316
Graph-incompatible exclusion	694
Other unsupported records	550
Selected training queries	41612

to answer correctness; zero-count increments remain eligible. Two annotators inspect causal clips and reconcile disagreements before assigning the binary correctness label. Identity and boundary masks additionally require the physical track or transition time to be supported. Table 13 reports sample counts and correctness rates; it evaluates perception rather than algebraic self-consistency. Table 14 checks the necessary count-capacity bound before prediction, using the maximum observation budget and ignoring candidate failures. Full permits at most 64 observations; Stream permits at most $\lfloor t\,1 \rfloor + 1$ from the zero origin. A global periodic stream completes at most $\lfloor T/2 \rfloor$ cycles. Other event bounds allow all 4 entity slots to be present from the origin. Passing this bound does not establish joint-path or perceptual reachability. Every official query remains in the main score.

H Complete-System Runtime

The runtime evaluation uses the complete 8B model on the same NVIDIA H200 host, BF16 arithmetic, and batch size 1. General-QA latency averages requests generating 32 output tokens. Stream begins with the visual prefix already processed and takes 0.48 seconds per request; Full includes its 64-frame causal prefix replay and takes 1.72 seconds. These are mean request latencies, with no seed-level confidence interval attached.

Chronological visual ingestion runs at 9.6 frames per second, including the vision encoder, causal memory, and typed-state update for a registered counting target. A typed-count read takes 1.8 milliseconds on average, including CPU as-of lookup and answer assembly; it performs neither neural decoding nor visual replay. This read uses the already maintained target state.

A late-target diagnostic retains exactly 1800 observations from a 30-minute history. Replay uses these retained frames without adding an observation at the zero-time boundary. On target revelation, those cached frames are processed chronologically through the counting stream, using observations no later than the query cutoff. Replay processing time is the retained-frame count divided by 9.6 frames per second; adding the typed-read cost gives 187.50 seconds after rounding. This cost calcu-

Table 12: Supervision coverage in the selected training set. The first three strata are mutually exclusive; verified stability overlaps them.

Supervision	Selected queries
Endpoint only	26068
Verified identity	11328
Verified boundary	4216
Verified stable interval	12894

Table 13: Stratified visual-label audit. Each record contributes one endpoint and one interval judgment; both rates use the audited-record count as denominator.

Source	Audited records	Endpoint correct (%)	Increment correct (%)
Objects	320	93.1	89.4
Natural events	240	88.8	85.4
Periodic	240	96.7	93.8

Table 14: Observation-budget capacity coverage on official event endpoints (%). This is an upper bound on reachable coverage; all endpoints remain in the official score.

Operator	Endpoint queries	Full upper coverage	Stream upper coverage
Action	1281	100.0	100.0
Transit	205	100.0	100.0
Episode	513	100.0	100.0
Periodic	280	44.3	58.2

lation uses the explicit retained trace and differs from the Full general-QA request above. The host ledger occupies 84 and 1240 kilobytes after 1 and 30 minutes, respectively, using decimal kilobytes.

I Precursor Component Evaluation

The precursor continuous-memory experiments use Qwen3-VL-8B with parallel MLP-TTT branches and the language-model answer decoder. Full replays each causal prefix independently. Persistent Stream uses a local KV window, carries visual state between questions, and discards each QA fork after decoding. These experiments use nonlinear MLP fast weights and precede the delta-memory transition model. They identify effects of that continuous component.

Continuous-memory adaptation. The continuous branch admits a separate intervention before an event reader is attached. On the adaptation split, two-stage training increases Full GPA from 42.01 to 55.61. Setting the learned branch gates to zero reduces it to 30.65; the paired difference is 24.96 points, with a 95% interval of [22.45, 27.38]. This establishes that the adapted model uses the branch, including its interaction with the jointly adapted language model.

On the development split, the dynamically updated streaming model scores 31.91 GPA, compared with 26.12 for a fixed fast-weight state. Initializing from the Full-adapted checkpoint raises the score to 36.80, a paired gain of 4.89 points [1.77, 7.96]. Further streaming training reaches 39.05; its additional GPA interval [-1.65, 6.35] crosses zero. These results motivate supervising state transitions directly rather than relying on additional recurrent training alone. The component settings and reader control are given below.

Table 15: Continuous-component diagnostics. Full rows evaluate the adaptation split; Stream rows evaluate the development split. Each block retains its precision and training setup and is ranked separately.

Access / population	Configuration	GPA ↑
Full / 3706 queries	First-stage adaptation	42.01
Full / 3706 queries	Two-stage adaptation	**55.61**
Full / 3706 queries	Two-stage, branch gates zero	30.65
Stream / 866 queries	Recurrent initialization	31.91
Stream / 866 queries	Fixed fast weights	26.12
Stream / 866 queries	Full-adapted initialization	36.80
Stream / 866 queries	Further streaming adaptation	**39.05**
Stream / 866 queries	Matched-precision reader control	35.94
Stream / 866 queries	Reader adaptation	**37.11**

Zeroing the trained branch gates measures dependence on the branch in the jointly adapted network. Freezing fast weights stops the update, momentum, and decay together. These interventions characterize branch use and recurrent dynamics; the discrete-state interface is evaluated separately.

The reader experiment freezes the continuous-memory parameters and fits a last-layer MLP adapter. Exact answers increase from 231 to 244 out of 866, an EM gain of 1.50 points [0.13, 2.77]. The GPA interval crosses zero. This distinction is expected because query-point EM and question-averaged GPA assign different weights to trajectories.

General video understanding is checked on 2700 Video-MME questions without subtitles. Full-adapted initialization scores 55.70%, while further streaming count adaptation scores 54.30%. The task-specific adaptation therefore changes the accuracy tradeoff beyond counting.

Paired state-summary readout. A later paired control evaluates an external state-summary reader on the same 866 queries. The memory condition answers 241 correctly (27.83%), while zeroing its memory input yields 244 (28.18%): a difference of -0.35 points. Across 26 training-cache records, 100% of the inspected write activations are saturated and the minimum cosine similarity between read directions is 0.99999988. This adapter behaves almost like a fixed readout direction; its result does not establish content-specific long-term retrieval. It is distinct from the fast-weight gate test and the typed event-state interface.

Precursor execution. In a precursor MLP-memory execution test, the 8B engine processes 8899 frames and 581709 video tokens across 15 queries. It sustains 21.99 fps with 36.40 GiB GPU allocation, 2.73–2.78 GiB host RSS, and 0.300 seconds p95 ready-prefix first-token latency.

J Count-State Readout

Table 16: Structured updates and count-state readout on SVCBench (GPA). The threshold variant is retrained; the remaining controls modify only the count projection read from a fixed complete-model rollout.

Condition	Full ↑	Stream ↑
Complete model	**44.9**	**38.2**
Threshold updates in place of FSM	<u>38.2</u>	<u>34.4</u>
Clear count projection	6.1	6.1
Cross-video count projection	12.7	15.6
Previous-endpoint projection	16.1	15.8

The fixed-sensor assay uses a causal pixel recurrence estimator on 25 periodic videos. Prefix estimates are computed at each query time, and the append-only ledger commits supported nonnegative increments. The direct estimates and the ledger both achieve 80.0% exactness; the ledger prevents downward revisions and changes MAE from 3.256 to 3.248. Clearing and delaying the ledger each yield 0.0% exactness; cross-video swapping yields 31.2%. This assay holds perception fixed and measures the dependence of the readout on correctly indexed state.

K Online and Spatial Transfer Protocols

Shared checkpoints and causal input. The transfer tests use the same counting-trained 4B and 8B checkpoints as SVCBench. No OVO-Bench or OVO-S questions, answers, or evaluation videos are used for task adaptation or checkpoint selection. A single training selection jointly excludes both benchmarks' matching source videos and transformed descendants. Full samples at most 64 prefix frames on OVO-Bench and 128 on OVO-S. Stream ingests at 1 fps with a 4096-token window. The pretrained Stream controls retain the same observation rate and KV budget without the learned fast-weight or counting-state modules. Runtime state resets by video, while question decoding uses disposable forks. Benchmark prompts determine the original answer format; these general questions do not supply the typed counting interface.

OVO-Bench. The fixed released annotation set contains 1640 records from 644 videos, yielding 3035 scored decisions. We retain real-time visual perception (RT), backward tracing (BT), and forward active responding (FA) (Niu et al., 2025). Accuracy is computed over queries within each task, then averaged over tasks within a group and over the three groups. FA retains every scheduled trigger in the accuracy track, including repeated timestamps belonging to separate entries. A Full decision replays its available prefix; a Stream decision advances through newly arrived frames. Trigger schedules and answer extraction are shared across models; invalid responses count as incorrect.

OVO-S-Bench. The released OVO-S snapshot contains 1695 items from 334 videos; expanding multi-time questions gives 1722 scored queries. Its four levels cover instantaneous egocentric perception, spatiotemporal context tracking, generative spatial reasoning, and allocentric mapping (Li et al., 2026). We retain the multiple-choice prompts, greedy decoding, and 1024-token output cap. Answer extraction accepts the labels in each question's option set; invalid or failed responses remain in the denominator. We compute query accuracy within each task family, average the three families in each level, and give the four levels equal weight in L-Avg. Query-pooled accuracy is retained separately. The same frozen query list and aggregation apply to every initial-model and adapted-model comparison.

Reference protocols. Scores for StreamForest-7B and the memory-based Flash-VStream-7B (Zhang et al., 2024) on OVO-Bench follow the three-group results in Zeng et al. (2025), with native sequential ingestion. Their spatial scores follow Li et al. (2026). Both are grouped under Stream, with their original evaluation releases and input rates: StreamForest uses one frame per second on both benchmarks; Flash-VStream uses one frame per second on OVO-Bench and two on OVO-S. The matched Qwen controls and StaMina share the frozen evaluation releases, observation budgets, and answer extraction described above.

On the original 1680-item OVO-S evaluation set, Qwen3-VL-4B uses the same prefix-frame cap as Full and scores 43.25/48.18/54.54/41.35 on L1–L4, with L-Avg. 46.83 (Li et al., 2026). The Qwen3-VL-4B comparison in Table 17 instead uses the frozen evaluation release specified above.

Table 17: Task-group and spatial-level breakdown (↑). RT: real-time perception; BT: backward tracing; FA: forward active responding. L1–L4 follow the spatial benchmark hierarchy. Group and level aggregates reproduce Table 3. Results are ranked within each access mode.

		OVO-Bench			OVO-S			
Model	Access	RT	BT	FA	L1	L2	L3	L4
Qwen3-VL-4B	Full	57.32	46.38	48.75	42.01	49.01	53.67	43.05
Qwen3-VL-8B	Full	**62.16**	50.56	52.35	42.48	50.68	**55.20**	45.98
StaMina-4B (Ours)	Full	56.11	49.63	49.23	41.91	54.67	53.72	43.96
StaMina-8B (Ours)	Full	60.71	**52.97**	**52.68**	**42.68**	**56.93**	54.91	**46.77**
StreamForest-7B	Stream	**61.20**	**52.02**	**53.49**	**46.64**	45.20	49.75	34.92
Flash-VStream-7B	Stream	29.86	25.35	44.23	18.65	29.93	22.53	28.67
Qwen3-VL-4B	Stream	52.42	41.66	46.62	37.55	46.35	47.87	40.44
Qwen3-VL-8B	Stream	57.76	46.59	49.38	39.72	48.29	50.50	42.96
StaMina-4B (Ours)	Stream	51.11	45.15	47.21	37.59	52.31	48.15	40.90
StaMina-8B (Ours)	Stream	56.34	49.92	49.98	39.80	**54.90**	**50.53**	**43.84**

The two transfer benchmarks probe different question families and are not combined into a single score. Their source domains can overlap. The breakdown distinguishes improvements in historical and spatial context from changes in immediate perception and general reasoning; it does not assume uniform transfer across tasks. The Video-MME control in Appendix I additionally records the behavior of an earlier component-only adaptation recipe.

L Additional Ablations and Model Scaling

Figure 5 summarizes system composition and supervision controls. These controls change modules or supervision, whereas Table 2 retains the legal graph to isolate transition learning. The data

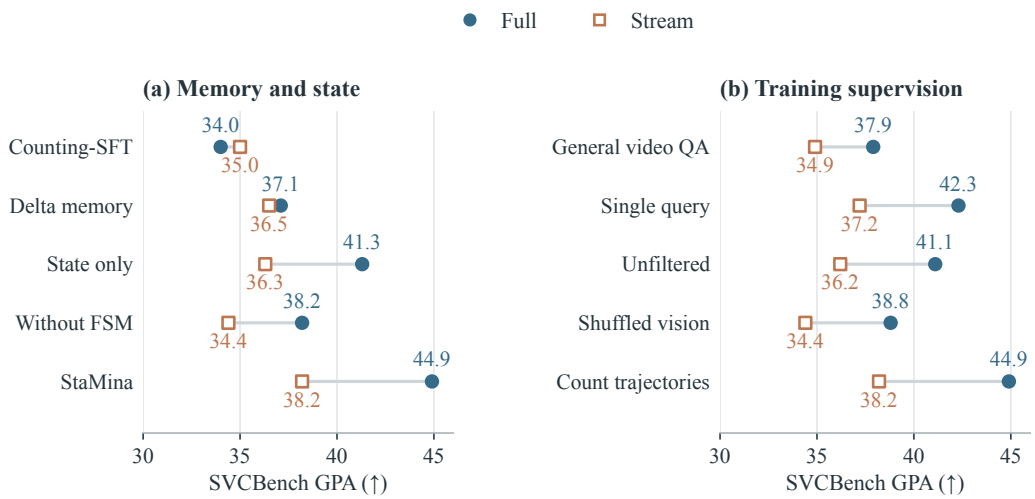


Figure 5: System and supervision ablations at 8B on SVCBench. State only removes delta memory; without FSM replaces the finite-state machine with independent thresholded increments.

Table 18: SVCBench system variants at 8B. Adapted rows share counting queries and update budgets; auxiliary objectives follow their components. Full and Stream share a checkpoint within each row. The two single-component variants are independent controls.

Model	Full ↑	Stream ↑
Qwen3-VL-8B	31.0	28.0
Counting-SFT	34.0	35.0
Counting-SFT + delta memory	37.1	<u>36.5</u>
Counting-SFT + explicit state	<u>41.3</u>	36.3
StaMina-8B (Ours)	**44.9**	**38.2**

controls use the same 8B checkpoint family, training updates, and visual-token budget. The visual-alignment control keeps each trajectory's timestamp slots, legal count endpoints, and supervision masks fixed while permuting its visual observations. The candidate association graph is held fixed in this control, so count reachability and the number of supervised endpoints are unchanged. Observations are processed causally in the resulting perturbed sequence; this is a training corruption, not an alternative evaluation input. Filtering changes label selection. Figure 6 summarizes the access gap at both model scales, and Table 20 gives the exact predicate accuracies underlying Figure 7.

Table 19: Supervision ablations at matched token and update budgets.

Supervision	Full ↑	Stream ↑
General video QA	37.9	34.9
Single query per trajectory	<u>42.3</u>	<u>37.2</u>
Before consistency filtering	41.1	36.2
Shuffled visual alignment	38.8	34.4
Complete count trajectories	**44.9**	**38.2**

Table 20: Action-predicate accuracy on recorded videos (%, ↑). Each cell pools 200 paired decisions; ranking is within each routine.

Routine	SFT	StaMina-8B (Ours)
Inventory	<u>61.0</u>	**74.0**
Handover	<u>65.0</u>	**78.0**
Restocking	<u>58.0</u>	**72.0**
Repetition completion	<u>60.0</u>	**75.0**

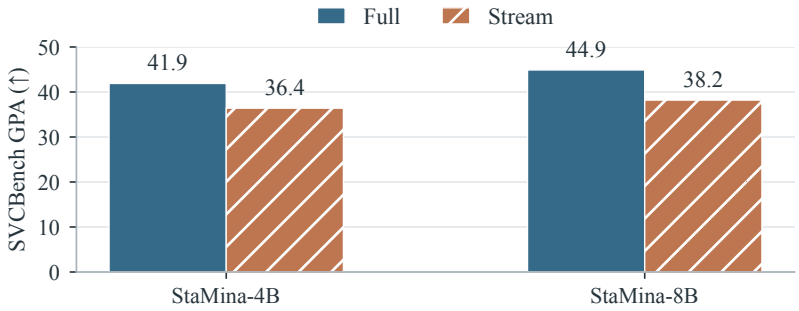


Figure 6: Model scale and access on SVCBench. Each bar is the corresponding official-question aggregate in Table 1.

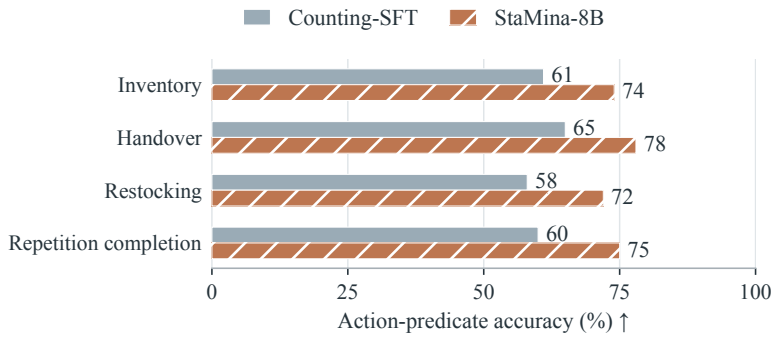


Figure 7: Action-predicate accuracy on recorded task videos. Both 8B conditions use Stream with identical predicates and decision times.