LEARNING TO REASON WITH PERSISTENT OBJECT STATES FOR VIDEO INSTANCE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video segmentation models preserve object identities by carrying information about each instance across frames. They typically update this history as new observations arrive. Under prolonged occlusion, reappearance, or interactions between similar instances, however, an incorrect update can overwrite a valid history and turn a local mismatch into persistent identity drift. This raises a question that current pipelines largely leave implicit: when should an observation be allowed to change an object's state? We introduce **POSReasoner**, a trainable, plug-and-play framework that makes this decision explicit by reasoning over persistent object states. For each object, POSReasoner carries a compact state that records identity, geometry, visibility history, and confidence. A sparse interaction graph links these states to candidate observations. Over several reasoning steps, POSReasoner compares each candidate with the object's history, resolves conflicting associations, and determines whether the state should be retained, updated, or reactivated. Intermediate supervision guides association and presence prediction, while a learned gate blocks unreliable updates. The resulting state is carried forward, so each decision affects subsequent association and segmentation rather than merely reranking a fixed set of predictions. POSReasoner is trained with standard video annotations while the base model remains frozen. This lightweight design allows it to augment diverse VOS and VIS architectures. Across long-term VOS and VIS benchmarks, POSReasoner consistently improves strong baselines, with the largest gains under occlusion and object reappearance.

1 INTRODUCTION

Video segmentation must determine not only what occupies each frame, but also whether masks separated in time belong to the same object. This temporal identity requirement is shared by video object segmentation (VOS), where target objects are specified, and video instance segmentation (VIS), where they must also be discovered and classified (Yang et al., 2019; Qi et al., 2022). Modern segmentation backbones and video-level object representations produce increasingly accurate masks (Cheng et al., 2022; 2021a; Ravi et al., 2025). Identity continuity, however, remains brittle under prolonged occlusion, reappearance, and interactions between similar instances, when current observations are least reliable and a mistaken association can affect the rest of the video.

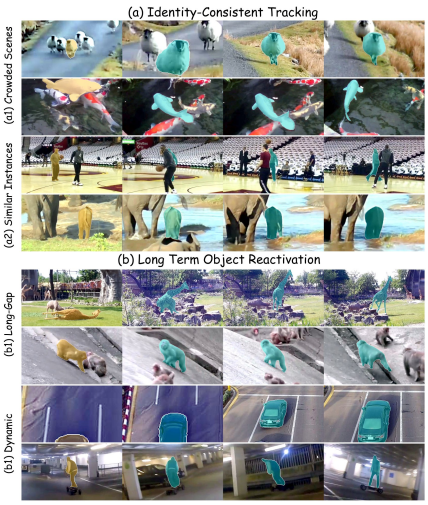

Figure 1: POSReasoner preserves identity under crowding and similar-instance interactions (a), and reactivates returning objects after long gaps or dynamic motion (b). Yellow denotes references; teal denotes predictions.

Existing methods preserve temporal continuity by carrying features or object queries across frames. Memory-based VOS systems retrieve features from prior frames (Cheng & Schwing, 2022; Cheng et al., 2024; Ravi et al., 2025), while query-based and decoupled VIS systems propagate queries, link predictions, or refine tracks (Wu et al., 2022a;b; Huang et al., 2022; Heo et al., 2022; 2023; Zhang et al., 2023a;b). Recent designs strengthen video-

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.

frame interaction and contextual association (Zheng et al., 2024; Lee et al., 2025b), manage object memory explicitly (Lee et al., 2025a). DAM4SAM filters distractors, whereas SAM2Long retains multiple segmentation pathways (Videnovic et al., 2025; Ding et al., 2025). Despite this progress, three coupled limitations remain. First, filtering and pathway search reduce bad evidence, but selected observations still condition later predictions; once an error is admitted, subsequent matching uses an altered identity reference (Videnovic et al., 2025; Ding et al., 2025). Second, a missing match may indicate occlusion, true absence, or a missed detection; suppressing low-confidence evidence alone does not determine whether an old identity should persist or reactivate. LOMM models object presence and LTMU predicts update readiness (Lee et al., 2025a; Dai et al., 2020), but these signals are not jointly reasoned with multi-object association and state admission. Third, synchronized, contextual, and decoupled designs improve candidate association (Zhang et al., 2023a;b; Zheng et al., 2024; Lee et al., 2025b), but do not revisit provisional matches with lifecycle state before write-back; competing identities may therefore be locally plausible yet mutually inconsistent. These limitations share a root cause: the current observation is treated as the next state rather than evidence for a state transition.

To address these limitations, we introduce **POSReasoner**, a trainable, plug-and-play framework that reasons over persistent, identity-indexed object states. To protect reliable history, each state records identity evidence, geometry, visibility history, and confidence, and persists by default until a transition is verified. To model absence and reappearance explicitly, a sparse state–observation graph links active and absent states to the current candidates and a learned null observation. Reasoning then follows a Propose–Verify procedure. *Propose* recalls the object history and forms provisional state–observation associations, while *Verify* revisits them using predicted object presence and excess candidate demand. Gated residual corrections allow verification to resolve duplicate claims without discarding a competent proposal. A transition head finally retains, updates, reactivates, or suppresses each state. Association, presence, and transition estimates receive intermediate supervision, but the internal steps do not advance video time: only the final verified transition writes candidate evidence into the persistent state. POSReasoner therefore turns association, lifecycle inference, and memory admission into a single recurrent state transition rather than a sequence of disconnected decisions. The verified transition then conditions the next frame, allowing reasoning to shape future association and segmentation rather than merely revise the current output. All reasoning targets come from standard video annotations, while the host model remains frozen.

By jointly leveraging persistent object states, conflict-aware verification, and selective write-back, POSReasoner maintains reliable identities through occlusion and reappearance, as shown in Figure 1. The resulting lightweight framework augments structurally different host architectures without redesigning their segmentation or tracking components. Experiments demonstrate consistent improvements over strong hosts, with the clearest gains under severe occlusion and object reappearance.

In summary, our main contributions are as follows:

- We formulate long-horizon identity maintenance as reasoning over persistent object states and introduce POSReasoner, a trainable framework for both VOS and VIS that leaves the host model frozen.
- We develop recurrent reasoning on a sparse state–observation graph to jointly infer object presence, resolve competing associations, and selectively commit verified state transitions.
- Experiments across long-term VOS and VIS benchmarks show consistent improvements over strong host architectures. Occlusion-focused analyses and controlled interventions further isolate the effects of persistent state reasoning and learned transitions.

2 Related Work

2.1 Video Instance Segmentation

Video instance segmentation (VIS) jointly predicts object categories, masks, and identities (Yang et al., 2019; Qi et al., 2022). Video-level methods aggregate frame or clip queries (Cheng et al., 2021a; Wu et al., 2022a; Heo et al., 2022), whereas online methods associate frame predictions through query propagation or learned embeddings (Wu et al., 2022b; Huang et al., 2022). GenVIS

and CTVIS strengthen temporal association with propagated prototypes and memory banks (Heo et al., 2023; Ying et al., 2023); DVIS and DVIS++ decouple segmentation, tracking, and temporal refinement (Zhang et al., 2023a;b). Recent methods further introduce dynamic anchors, synchronized queries, and contextual matching (Zhou et al., 2024; Zheng et al., 2024; Lee et al., 2025b). LOMM maintains a presence-aware latest-object memory and separates existing-object association from new-object allocation (Lee et al., 2025a). Tracking instability nonetheless remains pronounced in long, crowded, and heavily occluded videos (Hamdi et al., 2026). Together, these designs make associations more reliable, yet association and state update are usually coupled: once an observation is matched, it becomes part of the reference used in later frames. A local association error can therefore alter the identity evidence on which subsequent decisions depend.

2.2 Memory-Based Video Object Segmentation

Memory-based VOS propagates reference-frame objects through past observations. STM and STCN establish dense space–time correspondence (Oh et al., 2019; Cheng et al., 2021b), while AOT and DeAOT propagate multiple objects through identification embeddings (Yang et al., 2021; Yang & Yang, 2022). XMem separates sensory, working, and long-term memory (Cheng & Schwing, 2022), and Cutie combines pixel memory with object queries (Cheng et al., 2024). Foundation models retain a similar recurrent interface: SAM 2 uses streaming memory attention (Ravi et al., 2025), while SAM 3 combines image-level detection with a memory-based tracker and an object-presence head (Carion et al., 2025). Extensions for long videos retain alternative mask pathways, suppress distractor-contaminated memory, retrieve memories using motion and spatiotemporal cues, or make SAM 3's memory selection object-specific (Ding et al., 2025; Videnovic et al., 2025; Yang et al., 2025; Shen et al., 2026). These methods improve which observations are stored or retrieved within a particular memory design. Our focus is complementary: given the candidates exposed by an existing segmenter, we model their competing claims on persistent identity states and determine the resulting state revision.

2.3 Object-Centric Video Learning

Object-centric models decompose a video into recurrent object slots. SAVi propagates slots using motion and initialization cues (Kipf et al., 2022). Dual-State Slot Attention separates transient appearance from persistent identity and filters identity updates through a learned transition (Tran et al., 2026), while Temporal Slot Activation predicts whether a slot is active and gates both its update and decoding during invisibility (Nguyen et al., 2026). Recent open-world segmentation also combines hierarchical mask discovery, deferred object admission, and track consolidation for long-range identity maintenance (Su et al., 2026). The connection to our setting lies in their persistent representation of identity. Their learning objectives, however, concern scene decomposition or open-world discovery rather than the states maintained by an existing segmentation model. Our work addresses this missing state-revision step: it starts from candidates supplied by a frozen host, reasons over their claims on persistent object states, and carries only the verified transitions forward.

3 Method

Figure 2 summarizes the transition from frame-local association and direct memory updates to our persistent-state formulation, which verifies competing hypotheses before admitting evidence to memory.

Given a video $\mathcal{X} = \{I_t\}_{t=1}^{T}$, a frozen host model produces a set of object observations at every frame. They correspond to propagated queries in VIS (Zhang et al., 2023a;b; Lee et al., 2025a) or memory readouts in VOS (Cheng & Schwing, 2022; Ravi et al., 2025; Carion et al., 2025). A match does not by itself determine a memory update. Depending on the object's history and competing identities, the matched observation may update a visible state, revive an absent one, or be rejected. POSReasoner treats each association as a **state-transition hypothesis**. The problem is thus not only which observation matches, but whether and how that match should alter the persistent state.

As illustrated in Figure 3(a), POSReasoner pairs frozen-host observations with persistent states that retain identity evidence and lifecycle history. Propose forms provisional associations, while Verify uses object competition and state history to refine both identity assignments and transition actions

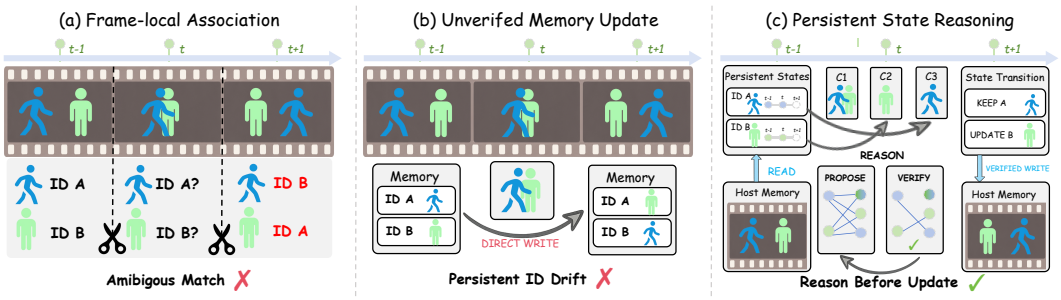

Figure 2: From association to persistent-state reasoning. (a) Frame-local association leaves ambiguous matches unresolved across time. (b) Direct write-back can propagate one unreliable observation through later states. (c) POSReasoner keeps identity-indexed states separate from host memory, verifies competing hypotheses, and applies an explicit state transition.

(b). These decisions gate memory write-back, admitting candidate evidence in proportion to the support for an update or reactivation (c). Intermediate hypotheses remain in a temporary workspace: only the final verified transition updates the persistent state carried to the next frame. Thus, a proposal can be revised without prematurely altering persistent memory.

3.1 Persistent Object States

For frame I_t, the host $\mathcal{H}$ returns object features, class probabilities, and masks as $[\widetilde{\mathbf{Q}}_t, \mathbf{P}_t, \mathbf{M}_t] = \mathcal{H}(I_t)$. The adapter represents row j by a host feature q_t^j and appends cues already available from the host: foreground confidence, class confidence, class margin, and predicted mask quality. VIS queries and VOS memory readouts thus share the same state–observation interface.

Before processing I_t, the state bank $\mathcal{S}_{t-1} = \{s_{t-1}^i\}_{i=1}^{N_{t-1}}$ stores one entry for every admitted identity. Each entry keeps a latent identity feature together with state confidence and time since the last reliable observation. VOS states are initialized at the annotated first appearance. In VIS, a foreground observation that is not assigned to an existing identity opens a free state slot. During absence, the lifecycle record advances while the identity feature is left intact. State confidence is updated only for a visible observation; otherwise its previous value is retained and the absence duration is incremented.

Let η_i collect the state metadata and μ_j the observation metadata. We suppress time indices below, writing $z_i = z_{t-1}^i$ and $q_j = q_t^j$. Two lightweight projections map them to the common reasoning dimension: $h_i^0 = f_s([z_i; \eta_i])$ and $u_j^0 = f_o([q_j; \mu_j])$. Here z_i is the persistent identity feature and q_j is the current frozen-host feature. The projections are trained with POSReasoner, while the host features remain frozen. In our implementation, η_i contains state confidence and normalized absence duration, and μ_j contains visual confidence, class confidence, class margin, and mask quality.

We connect the state and observation sets with a sparse bipartite graph. Its edges retain globally confident observations and the top-K identity-compatible observations for each state. A learned null observation is connected to every state, so occlusion need not be explained by a forced match. Every real edge is treated as a candidate association whose effect on the persistent state is determined jointly with the transition action; the null edge represents leaving the state unmatched.

3.2 Propose–Verify State Reasoning

A locally plausible association may be inconsistent with the object set, as when two identities claim the same observation. POSReasoner therefore runs two association steps with shared parameters. At step r, self-attention contextualizes the states and cross-attention reads adjacent observations (Vaswani et al., 2017). Let h_i^r and u_j be the resulting features, and let z_i and q_j denote the persistent identity feature and current frozen-host feature, respectively. Before the first step, observations are also contextualized by self-attention. A learned step embedding distinguishes Propose from Verify while the attention and prediction parameters remain shared. Invalid graph edges are masked

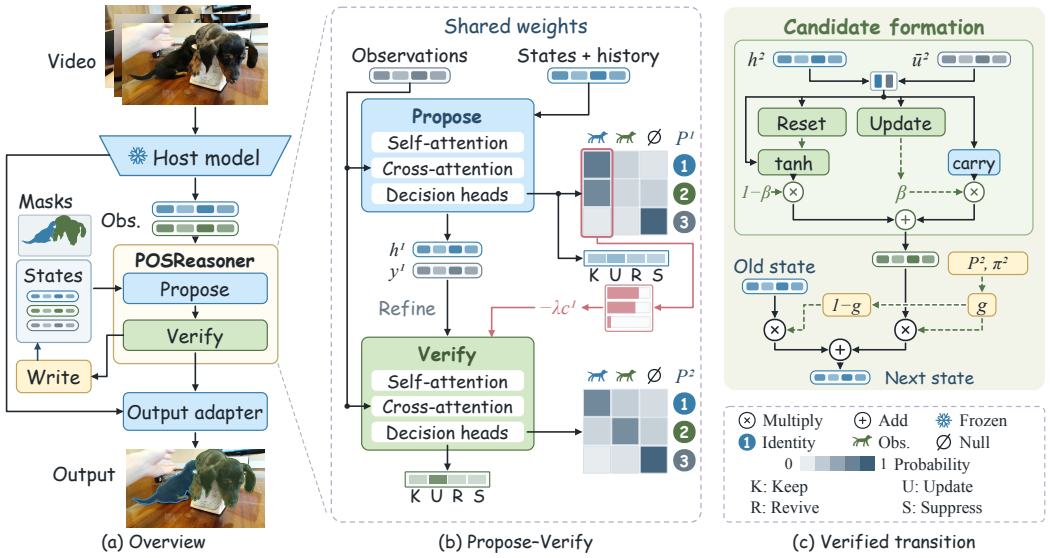

Figure 3: Overview of POSReasoner. (a) Frozen-host observations are paired with persistent object states. (b) Propose–Verify reasoning refines associations and actions using identity history and object competition. The shaded matrices associate identities with observations or null. (c) Verified decisions gate a single state update. Internal values are schematic; DAVIS frames with ground-truth masks illustrate the interface, not model predictions.

throughout cross-attention and association. Writing $\widetilde{\mathcal{N}}_i = \mathcal{N}_i \cup \{\varnothing\}$, we compute

$$e_{ij}^r = \frac{\cos(W_s h_i^r, W_o u_j)}{\tau} + \lambda_{\text{id}} \cos(z_i, q_j),$$
$$p_{ij}^r = \underset{j \in \widetilde{\mathcal{N}}_i}{\text{softmax}} \left(e_{ij}^r - \lambda_{\text{cmp}} c_j^{r-1} \right), \qquad c_j^r = \left[\sum_i p_{ij}^r - 1 \right]_+ . \tag{1}$$

Here $\mathcal{N}_i$ is the sparse neighborhood of state i, and $\varnothing$ denotes the learned null observation, whose identity-prior term is explicitly defined as zero.

At the Propose step, we set $c_j^0 = 0$ in Equation 1. The score e_{ij}^1 combines compatibility in the contextualized feature space with the original host identity similarity. Normalization over $\widetilde{\mathcal{N}}_i$ gives a distribution over the retained candidates and the null observation. The column sum $\sum_i p_{ij}^1$ then measures how much total probability the state set assigns to observation j, and c_j^1 retains only the probability mass above one. It therefore acts as a soft duplicate-demand penalty rather than enforcing a hard one-to-one assignment.

For state i, the matched evidence is $\bar{u}_i^r = \sum_j p_{ij}^r u_j$. We compare it with the reasoned state using their concatenation, element-wise product, and absolute difference. Association entropy and probability-weighted candidate occupancy are appended to this comparison. The decision feature is

$$r_i^r = \left[h_i^r; \bar{u}_i^r; h_i^r \odot \bar{u}_i^r; |h_i^r - \bar{u}_i^r|; \mathcal{E}(p_i^r); \rho_i^r \right],$$
$$\rho_i^r = \sum_j p_{ij}^r \sum_k p_{kj}^r, \tag{2}$$

where $\mathcal{E}$ is association entropy and ρ_i^r measures the demand on the candidates preferred by state i. Three prediction heads applied to r_i^r estimate visibility and existence, a transition distribution π_i^r, and **transition advantage** v_i^r. Together with the association logits, these predictions form a decision tuple $\mathbf{y}_i^r = [\boldsymbol{\ell}_i^r; \mathbf{b}_i^r; \boldsymbol{a}_i^r; v_i^r]$, where $\mathbf{b}_i^r$ contains visibility and existence beliefs, and $\boldsymbol{a}_i^r$ contains the four transition logits, with $\pi_i^r = \text{softmax}(\boldsymbol{a}_i^r)$; $\boldsymbol{\ell}_i^r = [e_{ij}^r]_j$ contains the association logits. The tuple records a candidate identity, its current presence, the proposed state change, and its predicted benefit relative to retaining the state. We supervise this advantage with the corresponding quality difference,

$$\Delta_i = Q(\hat{z}_t^i, \mathcal{Y}_{t:t+H}^i) - Q(z_{t-1}^i, \mathcal{Y}_{t:t+H}^i), \qquad v_i^r \approx \Delta_i, \tag{3}$$

where Q is the task quality over a short annotated continuation $\mathcal{Y}^i_{t:t+H}$ when the object is decoded from the candidate or retained state. At inference, the reasoner predicts v^r_i directly from the current hypothesis and persistent state.

At the Verify step, the association is repeated with c^1_j. Observation competition penalizes contested real observations relative to null and allows their competing states to move to alternative candidates or the null observation. We keep $c^r_\varnothing = 0$, because multiple absent states may select null simultaneously. State confidence and absence duration remain part of h^r_i, so Verify can also change the action without changing the selected observation: the same match may imply UPDATE, REVIVE, or KEEP for different state histories. Verify can therefore revise either the proposed ownership or the proposed state transition. This coupled revision distinguishes state-transition reasoning from independently reranking a fixed set of predictions.

The second-step tuple is obtained by gated interpolation, $\mathbf{y}^2_i = \mathbf{y}^1_i + \boldsymbol{\alpha} \odot (\widetilde{\mathbf{y}}^2_i - \mathbf{y}^1_i)$, where $\widetilde{\mathbf{y}}^2_i$ is predicted from the updated matched evidence, association entropy, and observation competition. Learned coefficients separately refine association, action, and the belief/advantage residuals. They are initialized near zero, so Verify begins from the proposed tuple and learns how strongly each output group should change. Both reasoning steps receive intermediate supervision.

3.3 VERIFIED STATE TRANSITION AND LEARNING

The verified action probabilities determine the state transition. KEEP retains a reliable state, UPDATE incorporates a visible observation, REVIVE reconnects a previously absent object, and SUPPRESS rejects inconsistent evidence. A gated recurrent unit (Cho et al., 2014) first forms a candidate state $\hat{z}^i_t = \mathrm{GRU}(h^2_i, \bar{u}^2_i)$. In Figure 3(c), Reset gates the features entering the nonlinear branch, while Update produces the carry weight β that blends the carry and nonlinear branches. These internal gates form the candidate; a separate write gate controls its admission to persistent memory:

$$g_i = \big(\pi^2_{i,\mathrm{update}} + \pi^2_{i,\mathrm{revive}}\big) \max_{j \neq \varnothing} p^2_{ij} \big(1 - p^2_{i\varnothing}\big),$$
$$z^i_t = (1 - g_i) z^i_{t-1} + g_i \hat{z}^i_t. \tag{4}$$

The gate combines the probability of a writing transition, the strongest real association, and the total non-null mass. Otherwise, the persistent state is carried forward. Probability assigned to KEEP or SUPPRESS does not contribute to g_i; as either action dominates, the state write approaches zero. Both reasoning steps operate on a temporary workspace anchored to z^i_{t-1}; intermediate proposals are never committed to temporal memory. This prevents the same frame from updating an identity multiple times. Only the final transition advances the state at frame t. Appendix A derives the properties of association, verification, and memory admission.

At inference, conflicting non-null claims are ranked using association confidence and transition advantage. The advantage estimates whether a proposed write is preferable to retaining the current state; it does not alter the sparse candidate graph. For assignment-based hosts, the output adapter retains only the highest-ranked claim to each real observation. The verified state and its visibility record are carried to $t + 1$ and become the reference for the next association.

The verified decision is also returned through a host-specific output adapter. Residual hosts blend a query toward the proposed or matched state according to the UPDATE and REVIVE probabilities, while assignment-based hosts transfer the selected observation to its state slot. After existing states have been assigned, valid foreground observations that remain unclaimed are allocated to free slots. The null observation never consumes a slot and may be selected by multiple absent states.

Training targets are constructed from frozen-host predictions and annotations on the training split. The association target is the observation with the same identity, or the null observation when the object is absent. A visible match is labeled UPDATE, and a match after absence is labeled REVIVE. An observed state without a valid identity receives SUPPRESS; an unobserved state without a reliable candidate receives KEEP. The advantage target is the signed quality difference in Equation 3.

$$\mathcal{L}_{\mathrm{POSR}} = \sum_{r=1}^{2} \omega_r \big(\mathcal{L}^r_{\mathrm{assoc}} + \lambda_{\mathrm{tr}} \mathcal{L}^r_{\mathrm{tr}} + \lambda_{\mathrm{bel}} \mathcal{L}^r_{\mathrm{bel}}\big)$$
$$+ \lambda_{\mathrm{state}} \mathcal{L}_{\mathrm{state}} + \lambda_{\mathrm{adv}} \mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{ref}} \mathcal{L}_{\mathrm{ref}}. \tag{5}$$

$\mathcal{L}_{\text{assoc}}$ and $\mathcal{L}_{\text{tr}}$ are cross-entropy losses, while $\mathcal{L}_{\text{bel}}$ sums binary cross-entropy terms for visibility and existence. $\mathcal{L}_{\text{state}}$ is the cosine distance to a stop-gradient target: the matched observation embedding for a writing transition and the retained identity feature otherwise. $\mathcal{L}_{\text{adv}}$ is a smooth-L_1 loss on transition advantage, and the refinement loss is $\mathcal{L}_{\text{ref}} = |\mathcal{V}|^{-1} \sum_{i \in \mathcal{V}} [p^1_{i,y_i} - p^2_{i,y_i}]_+$, where y_i is the target association and $\mathcal{V}$ is the valid-state set. The verified step receives the larger weight. At validation and test time, POSReasoner uses only host outputs, its persistent state, and task-provided prompts.

4 Experiments

4.1 Experimental Settings

We evaluate VIS on YouTube-VIS 2019, 2021, and 2022 (Yang et al., 2019) and on OVIS (Qi et al., 2022), reporting mask AP, AP50, and AP75. YouTube-VIS covers diverse video lengths and category distributions, while OVIS emphasizes crowded scenes and inter-object occlusion. For long-term VOS, we use LVOS v1 and v2 (Hong et al., 2023; 2026), with region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their mean $\mathcal{J}\&\mathcal{F}$ (Perazzi et al., 2016). These benchmarks evaluate identity maintenance both with and without object discovery and category prediction.

For VIS, we integrate POSReasoner with CTVIS (Ying et al., 2023), DVIS++ (Zhang et al., 2023b), DVIS-DAQ (Zhou et al., 2024), and LOMM (Lee et al., 2025a), representing different association and memory designs. GenVIS (Heo et al., 2023) and other published methods provide additional comparisons. For VOS, we use SAM3 (Carion et al., 2025). Paired comparisons keep the host checkpoint, candidate masks, input resolution, and official evaluator fixed. Only POSReasoner is trained, using training annotations; the host remains frozen.

Training uses video-grouped folds to keep each video within one split. Out-of-fold predictions select the readout before model ensembling. For hosts without frame-wise identity sequences, we use the available state and relation descriptors under the same protocol. The VOS readout uses disjoint fitting, calibration, and holdout videos. All architecture, optimization, and candidate-budget choices are fixed before evaluation.

4.2 Comparison with State-of-the-Art Methods

Results on YouTube-VIS. POSReasoner consistently improves AP across the three benchmarks (Table 1). With ResNet-50, CTVIS gains 1.0, 1.1, and 2.1 AP on YouTube-VIS 2019, 2021, and 2022, respectively; DVIS++ and DAQ also improve on each benchmark. With ViT-L, LOMM gains 0.7, 0.8, and 1.0 AP. These paired improvements, obtained with fixed host weights and candidate masks, show that state reasoning complements different association and memory designs. Figure 4 illustrates identity preservation under occlusion, similar-object interactions, and reappearance.

Table 1: YouTube-VIS validation. Published methods provide context; each highlighted row adds POSReasoner to the host immediately above. Bold and underline mark the best and second-best result within each backbone.

Backbone	Method	YTVIS 2019			YTVIS 2021			YTVIS 2022		
		AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
R50	GenVIS (Heo et al., 2023)	50.0	71.5	54.6	47.1	67.5	51.5	37.5	61.6	41.5
	CTVIS (Ying et al., 2023)	51.8	74.1	55.2	49.5	72.5	53.6	45.4	67.5	48.9
	+ POSReasoner	52.8	75.2	56.5	50.6	**73.7**	55.1	**47.5**	**69.7**	**51.4**
	DVIS++ (Zhang et al., 2023b)	<u>55.7</u>	**80.8**	59.8	50.4	71.4	55.2	46.3	66.6	50.5
	+ POSReasoner	**56.0**	<u>80.6</u>	**60.9**	**51.1**	72.6	<u>55.8</u>	<u>47.0</u>	<u>68.2</u>	<u>51.0</u>
	DAQ (Zhou et al., 2024)	54.6	78.2	<u>60.3</u>	50.1	71.6	55.1	45.9	66.7	50.0
	+ POSReasoner	55.5	79.2	**60.9**	<u>51.0</u>	<u>72.9</u>	**55.9**	46.7	67.9	50.9
ViT-L	DVIS++ (Zhang et al., 2023b)	67.7	88.8	75.3	62.3	82.7	70.2	37.5	53.7	39.4
	LOMM (Lee et al., 2025a)	<u>69.1</u>	<u>89.3</u>	<u>76.5</u>	<u>65.0</u>	<u>85.8</u>	<u>72.7</u>	<u>60.4</u>	<u>82.7</u>	<u>67.1</u>
	+ POSReasoner	**69.8**	**89.5**	**77.0**	**65.8**	**86.0**	**73.3**	**61.4**	**83.4**	**67.6**

Results on OVIS. The gains are larger on occlusion-heavy OVIS (Table 2). With ResNet-50, CTVIS, DVIS++, and DAQ improve by 2.7, 2.1, and 1.6 AP, respectively. Each host gains more than on any YouTube-VIS benchmark, consistent with the value of retaining identity history when current observations are ambiguous. Improvements of 3.2, 1.9, and 1.8 AP75 also show that the benefits extend to a stricter mask-overlap threshold.

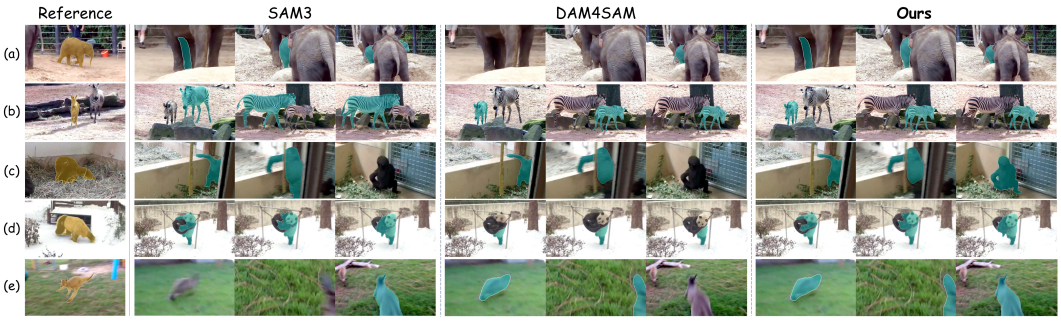

Figure 4: Qualitative comparison under occlusion, identity ambiguity, and object reappearance.

With larger backbones, DAQ improves from 49.6 to 50.6 AP with Swin-L and from 53.9 to 55.2 AP with ViT-L, with gains in both AP50 and AP75. State reasoning thus remains beneficial alongside stronger visual representations in crowded, occluded scenes.

Table 2: OVIS validation. Highlighted rows add POSReasoner to the matched host; ΔAP is relative to that host. Bold and underline mark the best and second-best results within each backbone.

Method	ΔAP	AP	AP50	AP75
ResNet-50				
CTVIS (Ying et al., 2023)	–	34.6	59.9	33.8
+ POSReasoner	**+2.7**	37.3	63.2	37.0
DVIS++ (Zhang et al., 2023b)	–	36.8	61.3	37.5
+ POSReasoner	**+2.1**	<u>38.9</u>	64.3	**39.4**
DAQ (Zhou et al., 2024)	–	38.3	<u>65.2</u>	37.3
+ POSReasoner	**+1.6**	**39.9**	**66.7**	<u>39.1</u>
Swin-L				
GenVIS (Heo et al., 2023)	–	45.2	69.1	48.4
LOMM (Lee et al., 2025a)	–	47.8	73.6	51.4
DAQ (Zhou et al., 2024)	–	<u>49.6</u>	<u>76.0</u>	<u>52.8</u>
+ POSReasoner	**+1.0**	**50.6**	**77.0**	**53.9**
ViT-L				
DVIS++ (Zhang et al., 2023b)	–	49.6	72.5	55.0
LOMM (Lee et al., 2025a)	–	51.7	73.9	57.5
DAQ (Zhou et al., 2024)	–	<u>53.9</u>	<u>78.9</u>	<u>58.5</u>
+ POSReasoner	**+1.3**	**55.2**	**79.7**	**59.9**

4.3 Long-Horizon State Analysis

We study how state reasoning supports identity maintenance over long videos. Starting from the frozen SAM3 host, we progressively add state reasoning, reactivation, and cross-trajectory reasoning on LVOS v1, and evaluate the complete model on LVOS v2 (Table 3).

Persistent states and reactivation. State reasoning improves $\mathcal{J}\&\mathcal{F}$ by 2.1 points. Its central role is to separate an object's identity history from its current visibility: the identity feature remains available during absence, providing a reference when observations become reliable again. Building on these states, reactivation yields a further 1.2-point improvement. This progression highlights two complementary requirements of long-term segmentation: preserving an identity through absence and reconnecting the returning object to that identity. The former retains the information needed for later association; the latter uses that information to resume the object's trajectory. Thus, the cumulative improvement supports treating absence and return as explicit state transitions.

Cross-trajectory reasoning. Adding cross-trajectory reasoning gives the best results across all three metrics, bringing the full model 3.5 points above SAM3 in $\mathcal{J}\&\mathcal{F}$. This component extends the decision from an individual state–observation match to competition among identities. When several states favor the same observation, verification can redistribute their support before the persistent states are updated. It therefore complements state retention and reactivation by considering whether an association is consistent with the other objects in the scene. Together, these components connect identity history, lifecycle transitions, and object competition within the same reasoning process.

Table 3: Long-term VOS and cumulative component analysis with frozen SAM3 predictions.

Variant	State	React.	Cross	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
LVOS v1 cumulative components						
SAM3 host (Carion et al., 2025)	✗	✗	✗	79.5	90.6	85.1
+ state reasoning	✓	✗	✗	82.1	92.4	87.2
+ reactivation	✓	✓	✗	83.3	93.6	88.4
Full POSReasoner	✓	✓	✓	**83.4**	**93.7**	**88.6**
LVOS v2 transfer						
SAM3 host (Carion et al., 2025)	✗	✗	✗	83.4	91.0	87.2
Full POSReasoner	✓	✓	✓	**85.3**	**92.7**	**89.0**

Generalization and qualitative analysis. On LVOS v2, the complete model improves $\mathcal{J}\&\mathcal{F}$ from 87.2 to 89.0, with gains in both region similarity and contour accuracy. The improvements across both benchmarks support the effectiveness of the combined state-reasoning design. Figure 5 further illustrates the distinct roles of state retention and reactivation. In the upper example, the state-based variant recovers the target missed by the host. In the lower example, reactivation moves the prediction from the foreground railing back to the returning target. These examples illustrate how stored identity evidence supports object recovery when observations become ambiguous or an object reappears.



Figure 5: LVOS v1: persistent state preserves identity through absence (top); reactivation reconnects it on return (bottom).

4.4 Further Analysis

The largest observed post-reappearance gains occur after gaps of 11–30 frames: 13.1 and 4.5 points in $\mathcal{J}$ on LVOS v1 and v2, respectively (Table 6). This pattern is consistent with retaining an identity reference through occlusion to support association when the object returns. The OVIS controls further highlight the role of state history: the full model reaches 37.31 AP, compared with 34.67 for shuffled states and 34.82 for host cues alone (Table 4). The contrast supports using identity-specific histories beyond current host evidence.

These benefits come with a compact reasoner: the five-member VIS ensemble has 0.90M parameters and takes 5.64–5.80 ms per inference on an A100 for 10–40 candidates (Table 5). The similar timings across these budgets indicate that increasing the candidate count has little effect on the measured reasoner latency.

5 Conclusion

We present POSReasoner, a trainable, plug-and-play framework for maintaining object identities in video segmentation. The framework represents each object with a persistent state that preserves its identity history through occlusion and absence. A Propose–Verify procedure jointly reasons about candidate associations, object presence, and state transitions, allowing reliable observations to update the state and returning objects to recover their identities. With a shared state–observation interface, POSReasoner can be integrated into different VOS and VIS architectures while keeping the host models frozen. Experiments on YouTube-VIS, OVIS, and LVOS demonstrate consistent improvements across the evaluated hosts, with particularly strong gains under occlusion. Ablation studies and qualitative comparisons highlight how state persistence and reactivation contribute to long-term identity maintenance.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and the creation or modification of scientific figures. They were not used to generate datasets, conduct experiments, formulate or prove mathematical claims, or determine the scientific conclusions of this work. All AI-assisted text and figures were reviewed and revised by the authors for accuracy, consistency with the method, and appropriate attribution. The authors take responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2Former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021a.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, 2022.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. *arXiv preprint arXiv:2207.07115*, 2022.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021b.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734. Association for Computational Linguistics, 2014.

Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6297–6306, 2020. doi: 10.1109/CVPR42600.2020.00633. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Dai_High-Performance_Long-Term_Tracking_With_Meta-Updater_CVPR_2020_paper.html.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. SAM2Long: Enhancing SAM 2 for long video segmentation with a training-free memory tree. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13614–13624, 2025. doi: 10.1109/ICCV51701.2025.01264. URL https://doi.org/10.1109/ICCV51701.2025.01264.

Danial Hamdi, Fardin Ayar, and Mahdi Javanmardi. Mind the gap: Disentangling performance bottlenecks in video instance segmentation. *arXiv preprint arXiv:2606.07394*, 2026. URL https://arxiv.org/abs/2606.07394.

Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. VITA: Video instance segmentation via object token association. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. LVOS: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13434–13446, 2023. doi: 10.1109/ICCV51070.2023.01240.

Lingyi Hong, Zhongying Liu, Wenchao Chen, Chenzhi Tan, Yuang Feng, Xinyu Zhou, Pinxue Guo, Jinglun Li, Zhaoyu Chen, Shuyong Gao, Wei Zhang, and Wenqiang Zhang. LVOS: A benchmark for large-scale long-term video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):946–961, 2026. doi: 10.1109/TPAMI.2025.3611020.

De-An Huang, Zhiding Yu, and Anima Anandkumar. MinVIS: A minimal video instance segmentation framework without video-based training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Thomas Kipf, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=aD7uesX1GF_.

Seunghun Lee, Jiwan Seo, Minwoo Choi, Kiljoon Han, Jaehoon Jeong, Zane Durante, Ehsan Adeli, Sang Hyun Park, and Sunghoon Im. LOMM: Latest object memory management for temporally consistent video instance segmentation. *arXiv preprint arXiv:2507.19754*, 2025a. URL https://arxiv.org/abs/2507.19754.

Seunghun Lee, Jiwan Seo, Kiljoon Han, Minwoo Choi, and Sunghoon Im. Context-aware video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4507–4517, 2025b.

Duc Nguyen, Sieu Tran, Hao Vo, Khoa Vo, Duy Minh Ho Nguyen, Nghi D. Q. Bui, Anh Nguyen, Long Mai, and Ngan Le. TSA: Temporal slot activation for persistent object-centric video representation. *arXiv preprint arXiv:2606.13714*, 2026. URL https://arxiv.org/abs/2606.13714.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9226–9235, 2019.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 724–732, 2016. doi: 10.1109/CVPR.2016.85.

Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip H. S. Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Ruiqi Shen, Chang Liu, and Henghui Ding. SAM3-DMS: Decoupled memory selection for multi-target video segmentation of SAM3. *arXiv preprint arXiv:2601.09699*, 2026. URL https://arxiv.org/abs/2601.09699.

Qing Su, Kaiyang Li, Yuan Zhuang, Fei Miao, and Shihao Ji. Open-world video segmentation. *arXiv preprint arXiv:2606.15632*, 2026.

Sieu Tran, Duc Nguyen, Hao Vo, Khoa Vo, and Ngan Le. Dual-state slot attention: Decoupling appearance and identity for video object-centric learning. *arXiv preprint arXiv:2606.12601*, 2026. URL https://arxiv.org/abs/2606.12601.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with SAM2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24255–24264, 2025. doi: 10.1109/CVPR52734.2025.02259. URL https://openaccess.thecvf.com/content/CVPR2025/html/Videnovic_A_Distractor-Aware_Memory_for_Visual_Object_Tracking_with_SAM2_CVPR_2025_paper.html.

Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 553–569, 2022a.

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 588–605, 2022b.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5188–5197, 2019.

Qiushi Yang, Yuan Yao, Miaomiao Cui, and Liefeng Bo. MoSAM: Motion-guided segment anything model with spatial-temporal memory selection. *arXiv preprint arXiv:2505.00739*, 2025. URL https://arxiv.org/abs/2505.00739.

Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. CTVIS: Consistent training for online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 899–908, 2023.

Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. DVIS: Decoupled video instance segmentation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a.

Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. DVIS++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023b.

Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. SyncVIS: Synchronized video instance segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yikang Zhou, Tao Zhang, Shunping Ji, Shuicheng Yan, and Xiangtai Li. DVIS-DAQ: Improving video segmentation via dynamic anchor queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

A Mathematical Properties of State Reasoning

This section analyzes the association, verification, and state-transition rules in Section 3, together with their refinement supervision. The results characterize how the rules handle competing and uncertain evidence; segmentation performance is evaluated empirically in Section 4.

A.1 Sparse Association and Competition

Notation and assumptions. For a fixed frame, let $A_i = \mathcal{N}_i \cup \{\varnothing\}$ be the finite legal candidate set of state i. The null candidate is always present. We assume finite legal logits, $\tau > 0$, and $\lambda = \lambda_{\mathrm{cmp}} \geq 0$. Illegal edges have probability zero. For the conditional analysis below, hold A_i and the compatibility logits e_{ij} fixed, and write

$$p_{ij}(c) = \frac{\exp(e_{ij} - \lambda c_j)}{\sum_{k \in A_i} \exp(e_{ik} - \lambda c_k)}, \qquad j \in A_i, \quad c_\varnothing = 0. \tag{A.1}$$

The positive denominator ensures $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$. If $\mathcal{N}_i$ is empty, $p_{i\varnothing} = 1$: the rule remains defined without forcing a real association. Consequently, the matched evidence $\bar{u}_i = \sum_j p_{ij} u_j$ lies in the convex hull of the legal observation features, including the learned null feature.

Proposition 1 (effect of competition). For any $j, k \in A_i$, the association odds satisfy

$$\frac{p_{ij}(c)}{p_{ik}(c)} = \frac{p_{ij}(0)}{p_{ik}(0)} e^{-\lambda(c_j - c_k)}. \tag{A.2}$$

In particular, a real candidate's odds relative to null are multiplied by $e^{-\lambda c_j}$. Let $M_j(c) = \sum_i p_{ij}(c)$. Increasing only the penalty of real candidate j gives

$$\frac{\partial M_j}{\partial c_j} = -\lambda \sum_{i : j \in A_i} p_{ij}(1 - p_{ij}) \leq 0, \qquad \frac{\partial p_{i\varnothing}}{\partial c_j} = \lambda p_{i\varnothing} p_{ij} \geq 0. \tag{A.3}$$

Thus, componentwise increases in real-candidate penalties cannot decrease the null probability when the compatibility logits are fixed.

Proof. Taking the ratio of two terms in Equation A.1 cancels their common normalizer and yields Equation A.2. Differentiation with respect to a legal real-candidate penalty gives $\partial p_{ik} / \partial c_j = -\lambda p_{ik}(\mathbf{1}\{k = j\} - p_{ij})$. Taking $k = j$ and summing over states proves the first derivative; taking $k = \varnothing$ proves the second. An illegal edge contributes zero. Integrating the nonnegative null derivatives along a componentwise increasing penalty path proves the final statement. □

This result isolates the contribution of the competition term. Verify also changes the contextualized logits, so it does not imply that every column mass decreases between the two reasoning steps. Soft competition shapes relative preferences; the assignment-based output adapter separately enforces at most one retained claim per real observation by selecting a single highest-ranked claimant (with ties resolved to one claimant). Null is exempt from both the excess-demand penalty and this exclusivity rule.

Decision features. The entropy and occupancy terms in Equation 2 describe different aspects of an association. With natural logarithms and $0 \log 0 = 0$, $0 \leq \mathcal{E}(p_i) \leq \log |A_i|$. The lower bound follows from $-p \log p \geq 0$; the upper bound uses KL nonnegativity (Boyd & Vandenberghe, 2004, Example 3.19): $\mathrm{KL}(p_i \| \mathrm{Unif}(A_i)) = \log |A_i| - \mathcal{E}(p_i) \geq 0$. Padding illegal edges with zeros puts all rows in the same candidate space, where the occupancy feature has the exact decomposition

$$\rho_i = \|p_i\|_2^2 + \sum_{k \neq i} \langle p_i, p_k \rangle, \qquad \sum_i \rho_i = \sum_{j \neq \varnothing} M_j^2 + M_\varnothing^2. \tag{A.4}$$

Indeed, substituting $M_j = \sum_k p_{kj}$ into $\rho_i = \sum_j p_{ij} M_j$ and separating $k = i$ gives the first equality; exchanging the two sums gives the second. Occupancy therefore combines within-state concentration with cross-state overlap, whereas entropy depends only on the individual row. The null contribution provides context about unmatched states; it is not a real-observation conflict, consistent with $c_\varnothing = 0$.

A.2 Controlled Refinement of Transition Decisions

Proposition 2 (action-distribution perturbation). Let a be the proposed action logits and $d = \alpha_a \odot (\widetilde{a}^2 - a)$ their gated correction, both finite. Set $\pi^1 = \mathrm{softmax}(a)$, $\pi^2 = \mathrm{softmax}(a + d)$, and $D = \|d\|_\infty$. Then

$$\|\pi^2 - \pi^1\|_1 \le \min\{2, D\}, \qquad \mathrm{KL}(\pi^1 \| \pi^2) \le \tfrac{1}{2} D^2. \tag{A.5}$$

No sign or interval constraint on α_a is needed. In particular, $D \le \|\alpha_a\|_\infty \|\widetilde{a}^2 - a\|_\infty$.

Proof. Consider $\pi(t) = \mathrm{softmax}(a + td)$ for $t \in [0, 1]$. Writing $\mu_t = \sum_k \pi_k(t) d_k$, differentiation gives $\pi_k'(t) = \pi_k(t)(d_k - \mu_t)$. Since $|d_k| \le D$,

$$\|\pi'(t)\|_1 = \mathbb{E}_{\pi(t)} |d - \mu_t| \le \sqrt{\mathrm{Var}_{\pi(t)}(d)} \le D. \tag{A.6}$$

Integration yields the D bound, while two probability distributions are always at most 2 apart in ℓ_1. For the log-sum-exp function (Boyd & Vandenberghe, 2004, Section 3.1.5), $f(t) = \log \sum_k \exp(a_k + t d_k)$, we have $f'(t) = \mu_t$ and $f''(t) = \mathrm{Var}_{\pi(t)}(d) \le D^2$. Hence

$$\mathrm{KL}(\pi^1 \| \pi^2) = f(1) - f(0) - f'(0) = \int_0^1 (1 - t) f''(t)\, dt \le \tfrac{1}{2} D^2. \tag{A.7}$$

This proves both claims. □

The bound explains how small gated logit corrections preserve a proposal's action distribution. It concerns the size of a correction, not its accuracy. For association probabilities, the change in the competition penalty must also be included in the effective logit correction.

A.3 Selective Admission to Persistent Memory

Proposition 3 (quadratic attenuation of uncertain evidence). For a single state, let $a = \pi^2_{\text{update}} + \pi^2_{\text{revive}}$ and $s = 1 - p^2_\varnothing$ denote the write-action probability and total real association mass. Define the conditional concentration

$$\kappa = \begin{cases} \max\limits_{j \in \mathcal{N}} p_j^2 / s, & s > 0, \\ 0, & s = 0. \end{cases} \qquad g = a \kappa s^2, \quad 0 \le g \le a s^2 \le s^2 \le 1. \tag{A.8}$$

Here the maximum over an empty real-candidate set is defined as zero. For any norm, the state transition in Equation 4 satisfies

$$\|z_t - z_{t-1}\| = g \|\hat{z}_t - z_{t-1}\| \le a s^2 \|\hat{z}_t - z_{t-1}\|. \tag{A.9}$$

In particular, $p^2_\varnothing \ge 1 - \varepsilon$, with $\varepsilon \in [0, 1]$, implies $g \le \varepsilon^2$.

Proof. Both action and association probabilities are normalized, so $a, s \in [0, 1]$. For $s > 0$, the maximum real probability is κs with $\kappa \in [0, 1]$; substituting this into the write rule gives $g = a \kappa s^2$. If $s = 0$, the real probabilities and the gate are zero, so the same identity holds. Subtracting z_{t-1} from the update and using norm homogeneity proves Equation A.9. Finally, $s \le \varepsilon$ gives the claimed attenuation. □

The factors isolate action preference, conditional concentration, and real-evidence mass. With $K \ge 1$ real candidates and $s > 0$, $1/K \le \kappa \le 1$ because the conditional probabilities sum to one. The update is convex: $g = 0$ preserves the identity feature, though lifecycle metadata may advance. For example, $p^2_\varnothing \ge 0.9$ gives $g \le 0.01$, limiting feature movement to one percent of the candidate displacement. This is a relative bound; an absolute bound additionally requires bounding $\|\hat{z}_t - z_{t-1}\|$.

Corollary 3.1 (retention along a state trajectory). Consider one identity over L frames without slot reinitialization. Define $R_{u:v} = \prod_{r=u}^{v} (1 - g_r)$, with an empty product equal to one. For the realized gates and candidate states,

$$z_L = R_{1:L} z_0 + \sum_{t=1}^{L} g_t R_{t+1:L} \hat{z}_t, \qquad R_{1:L} + \sum_{t=1}^{L} g_t R_{t+1:L} = 1. \tag{A.10}$$

All coefficients are nonnegative. If $p^2_{t,\varnothing} \ge 1 - \varepsilon_t$ with $\varepsilon_t \in [0, 1]$ at each frame, then

$$1 - R_{1:L} \le 1 - \prod_{t=1}^{L} (1 - \varepsilon_t^2) \le \min\left\{1, \sum_{t=1}^{L} \varepsilon_t^2\right\}. \tag{A.11}$$

Proof. Repeated substitution of the single-write recurrence gives the first identity. The second follows by telescoping $g_t R_{t+1:L} = R_{t+1:L} - R_{t:L}$. Proposition 3 gives $g_t \le \varepsilon_t^2$, hence $R_{1:L} \ge \prod_t (1-\varepsilon_t^2)$. The final inequality follows by induction from $1-(1-x)(1-y) = x+y-xy \le x+y$ for $x, y \in [0,1]$, together with $0 \le \prod_t (1-\varepsilon_t^2) \le 1$. □

For a uniform bound $\varepsilon_t \le \varepsilon \le 1$, the retained coefficient is at least $(1-\varepsilon^2)^L$. Moreover, if $\|z_0\|, \|\hat{z}_t\| \le B$ throughout the interval, convexity gives $\|z_L\| \le B$. These statements hold for the realized trajectory even when gates and candidates depend on earlier states. The coefficients are algebraic mixture weights, not derivatives of the trajectory with respect to z_0; a contraction claim would require additional control of those dependencies. Propose and Verify contribute only through the final gate and candidate, since neither intermediate step writes to persistent memory.

A.4 Meaning of Refinement Supervision

For a nonempty valid-state set $\mathcal{V}$, put $d_i = [p^1_{i,y_i} - p^2_{i,y_i}]_+$ and $n = |\mathcal{V}|$. Since every d_i is nonnegative, $\mathcal{L}_{\text{ref}} = 0$ if and only if $p^2_{i,y_i} \ge p^1_{i,y_i}$ for every $i \in \mathcal{V}$. More generally, for any $\epsilon > 0$, let $B_\epsilon = \{i \in \mathcal{V} : p^1_{i,y_i} - p^2_{i,y_i} > \epsilon\}$. Then

$$\frac{|B_\epsilon|}{n} \le \min\{1, \mathcal{L}_{\text{ref}}/\epsilon\}. \tag{A.12}$$

To see this, sum $\epsilon \mathbf{1}\{i \in B_\epsilon\} \le d_i$ over the valid states and divide by $n\epsilon$. Thus, the refinement loss controls the proportion of states whose target-association probability deteriorates by more than a prescribed amount on the supervised samples. The transition-advantage target in Equation 3 separately specifies whether a candidate improves the chosen continuation quality relative to retaining the state. Neither the joint objective nor the algebraic properties above assume exact optimization or perfect advantage prediction; their empirical effect is assessed by the reported comparisons and controlled analyses.

B Implementation Details

B.1 Host Interface

POSReasoner consumes frozen trajectory scores, masks, identity descriptors, and their temporal summaries. It does not change host weights or request additional image-level proposals. For every matched comparison, the host checkpoint, input resolution, candidate export, and evaluator are held fixed. We preserve each host's native output budget unless a larger candidate export is explicitly reported as part of that host configuration.

B.2 Training and Evaluation

VIS reasoner. The VIS instantiation uses a 96-dimensional hidden space, four attention heads, and two relation layers, trained for 180 epochs with AdamW, a learning rate of 1.5×10^{-3}, and weight decay of 2×10^{-4}. Five video-grouped folds select the executor from out-of-fold predictions, and their models are averaged at inference. Candidate budgets are fixed before validation. For hosts without frame-wise identity sequences, a compact readout uses their state and relation descriptors under the same protocol.

VOS reasoner. The VOS instantiation replays frozen identity-indexed masks and constructs events from visibility, geometry, and cross-trajectory changes. Persistent-state, reactivation, and sparse multi-trajectory policies are fitted or calibrated on video-disjoint training splits, then frozen before validation. Readout videos remain disjoint from calibration and holdout evaluation.

Selection and evaluation. No validation annotation is read during inference; only the official evaluator and the post-hoc analysis in Table 6 load it. Main benchmarks are single runs of a fixed host–reasoner pair, while the OVIS mechanism study repeats seeds 20260728, 20260729, and 20260730. Hyperparameters and executor choices use grouped out-of-fold training performance and are fixed before validation.

B.3 Controlled Comparisons and Efficiency

The controls use the same frozen candidates and training protocol.

Host cues alone and shuffled states remain near the frozen host (Table 4). Intact states reach 36.55 AP even without a graph; the full model reaches 37.31 AP, supporting structured interaction beyond state retention. Shared two-step reasoning also exceeds one-step reasoning (36.97 versus 36.13 AP), consistent with the benefit of revisiting initial decisions. These controls support using identity-specific history together with structured interaction. Across three seeds, the full model achieves 37.203 ± 0.099 AP, a 2.570 ± 0.099 gain.

Table 4: Causal controls on OVIS with the same frozen candidates and training protocol.

Variant	AP	ΔAP	AP50	AP75
Frozen host	34.63	–	59.94	33.80
Host cues only	34.82	+0.18	60.21	34.09
Shuffled state	34.67	+0.04	59.95	33.94
State, no graph	36.55	+1.92	62.88	35.61
State, one step	36.13	+1.50	62.01	35.27
State, shared two-step	36.97	+2.33	63.00	36.17
Full POSReasoner	**37.31**	**+2.68**	**63.20**	**36.98**

Table 5: Five-member VIS reasoner efficiency on one A100, batch size 1.

Candidates	Params	Latency	Memory
10	0.90M	5.80 ms	12.71 MB
20	0.90M	5.66 ms	12.99 MB
40	0.90M	5.64 ms	14.03 MB

The five-member VIS reasoner uses 0.90M parameters across budgets of 10–40 candidates (Table 5). Median latency remains within 5.64–5.80 ms, while peak allocated CUDA memory increases from 12.71 to 14.03 MB. These measurements cover the reasoner, with latency measured over 500 runs after 100 warm-ups.

Experiments use NVIDIA A100-SXM4 80 GB GPUs and Intel Xeon Platinum 8369B CPUs under Ubuntu 24.04, with Python 3.10.20, PyTorch 2.4.1, CUDA 12.1, and NumPy 1.26.4. Independent folds and dataset–host pairs run in parallel; each frozen replay needs one GPU.

C Reappearance Analysis

Figure 6 illustrates identity recovery, while Table 6 groups reappearance events by absence duration. These post-hoc diagnostics distinguish long-term recovery from ordinary frame-to-frame matching and do not affect policy selection. Intermediate gaps show the clearest gains when local identity is ambiguous but persistent evidence remains.

Gaps are measured in absent frames. The table lists region similarity, event and video counts, and paired bootstrap intervals over videos. H/T/D counts events that improve, remain unchanged, or degrade.



Figure 6: LVOS v1 reappearance, similar-object interaction, and crowded motion. Yellow: references; teal: unmodified POSReasoner outputs.

Table 6: Post-reappearance results by absence duration, with paired differences and 95% bootstrap confidence intervals.

Dataset	Gap	Events	Videos	Host $\mathcal{J}$	POSR $\mathcal{J}$	$\Delta\mathcal{J}$ [95% CI]	H/T/D
LVOS v1	5–10	45	16	70.6	74.9	+4.3 [−2.3, 16.3]	22/9/14
	11–30	34	15	51.1	64.2	+13.1 [−3.1, 29.1]	20/6/8
	31–100	14	8	72.2	78.0	+5.9 [−2.7, 19.1]	7/3/4
	> 100	6	4	23.6	30.6	+7.1 [−10.1, 46.9]	1/2/3
LVOS v2	5–10	114	42	66.9	65.8	−1.1 [−7.7, 5.3]	53/17/44
	11–30	83	39	56.6	61.1	+4.5 [−2.3, 11.6]	38/15/30
	31–100	33	22	67.1	69.6	+2.5 [−4.5, 9.7]	17/5/11
	> 100	9	7	26.2	24.1	−2.0 [−5.5, 0.1]	1/7/1

D Additional Benchmark Results

Tables 7 and 8 extend the OVIS and YouTube-VIS comparisons; paired runs share the host, candidate budget, and evaluator.

Table 7: Complete OVIS evaluation across three backbones. Cited rows are published results; each uncited host is paired with POSReasoner under the same frozen host and candidate set.

Method	Backbone	AP	AP50	AP75
GenVIS (Heo et al., 2023)	ResNet-50	35.8	60.8	36.2
CTVIS (Ying et al., 2023)	ResNet-50	35.5	60.8	34.9
DVIS++ (Zhang et al., 2023b)	ResNet-50	37.2	62.8	37.3
DAQ (Zhou et al., 2024)	ResNet-50	38.7	65.5	37.6
CTVIS	ResNet-50	34.6	59.9	33.8
CTVIS + POSReasoner	ResNet-50	**37.3**	**63.2**	**37.0**
DVIS++	ResNet-50	36.8	61.3	37.5
DVIS++ + POSReasoner	ResNet-50	**38.9**	**64.3**	**39.4**
DAQ	ResNet-50	38.3	65.2	37.3
DAQ + POSReasoner	ResNet-50	**39.9**	**66.7**	**39.1**
MinVIS (Huang et al., 2022)	Swin-L	39.4	61.5	41.3
IDOL (Wu et al., 2022b)	Swin-L	40.0	63.1	40.5
GenVIS (Heo et al., 2023)	Swin-L	45.2	69.1	48.4
DVIS (Zhang et al., 2023a)	Swin-L	45.9	71.1	48.3
CTVIS (Ying et al., 2023)	Swin-L	46.9	71.5	47.5
LOMM (Lee et al., 2025a)	Swin-L	47.8	73.6	51.4
DAQ	Swin-L	49.6	76.0	52.8
DAQ + POSReasoner	Swin-L	**50.6**	**77.0**	**53.9**
DVIS++ (Zhang et al., 2023b)	ViT-L	49.6	72.5	55.0
LOMM (Lee et al., 2025a)	ViT-L	51.7	73.9	57.5
DAQ	ViT-L	53.9	78.9	58.5
DAQ + POSReasoner	ViT-L	**55.2**	**79.7**	**59.9**

Table 8: Complete YouTube-VIS plug-in evaluation; 2022 baselines follow LOMM (Lee et al., 2025a). Each POSReasoner row retains its frozen host; * denotes offline refinement and † provisional results.

Method	Backbone	YTVIS 2019			YTVIS 2021			YTVIS 2022		
		AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
Query-based Association										
MinVIS (Huang et al., 2022)	R50	47.4	69.0	52.1	44.2	66.0	48.1	23.3	47.9	19.3
VITA* (Heo et al., 2022)	R50	49.8	72.6	54.5	45.7	67.4	49.5	32.6	53.9	39.3
GenVIS (Heo et al., 2023)	R50	50.0	71.5	54.6	47.1	67.5	51.5	37.5	61.6	41.5
Memory and Decoupled Tracking										
DVIS++ (Zhang et al., 2023b)	R50	55.5	80.2	60.1	50.0	72.2	54.5	37.2	57.4	40.7
LOMM (Lee et al., 2025a)	R50	55.7	79.8	61.4	50.7	72.9	56.9	41.1	62.4	46.2
DVIS++ (Zhang et al., 2023b)	ViT-L	67.7	88.8	75.3	62.3	82.7	70.2	37.5	53.7	39.4
LOMM (Lee et al., 2025a)	ViT-L	69.1	89.3	76.5	65.0	86.0	72.7	48.2	70.5	53.2
Persistent Object-State Reasoning										
CTVIS	R50	51.8	74.1	55.2	49.5	72.5	53.6	45.4	67.5	48.9
CTVIS + POSReasoner	R50	**52.8**	**75.2**	**56.5**	**50.6**	**73.7**	**55.1**	**47.5**	**69.7**	**51.4**
DVIS++	R50	55.7	80.8	59.8	50.4	71.4	55.2	46.3	66.6	50.5
DVIS++ + POSReasoner	R50	**56.0**	80.6	**60.9**	**51.1**	**72.6**	**55.8**	**47.0**	**68.2**	**51.0**
DAQ	R50	54.6	78.2	60.3	50.1	71.6	55.1	45.9	66.7	50.0
DAQ + POSReasoner	R50	**55.5**	**79.2**	**60.9**	**51.0**	**72.9**	**55.9**	**46.7**	**67.9**	**50.9**
LOMM†	R50	55.8	80.1	61.5	50.9	73.3	56.6	47.2	68.8	51.9
LOMM + POSReasoner†	R50	**56.1**	80.0	**62.1**	**51.3**	**73.6**	**56.9**	**47.7**	**69.2**	**52.4**
LOMM	ViT-L	69.1	89.3	76.5	65.0	85.8	72.7	60.4	82.7	67.1
LOMM + POSReasoner	ViT-L	**69.8**	**89.5**	**77.0**	**65.8**	**86.0**	**73.3**	**61.4**	**83.4**	**67.6**

Reading the extended comparisons. Tables 7 and 8 provide complementary views of the evaluation. Cited rows place the tested systems within the published benchmark landscape, whereas adjacent host and POSReasoner rows isolate the contribution of the added reasoner. Each pair shares the host checkpoint, candidate masks, input resolution, and evaluator; only POSReasoner is trained. The appropriate reference for a plug-in gain is therefore the paired host, rather than another published implementation of the same architecture. This distinction is particularly relevant to YouTube-VIS 2022, whose published comparison rows follow LOMM. The offline refinement and provisional-result markers are retained so that the evaluation setting remains explicit for every row.

On OVIS, all five evaluated host–backbone pairs improve in AP, AP50, and AP75. With ResNet-50, the AP gains are 2.7 points for CTVIS, 2.1 for DVIS++, and 1.6 for DAQ. These hosts use different temporal designs, yet each benefits from reasoning over persistent states. Moreover, each of these OVIS gains exceeds the corresponding host's gain on any of the three YouTube-VIS versions. This pattern is consistent with the motivation for retaining identity evidence in crowded videos: when current observations are ambiguous, state history supplies a reference that can be revisited before an association is committed. The benefit is thus observed across the evaluated hosts, not only with one association architecture.

The DAQ results further separate the choice of visual backbone from the addition of state reasoning. AP rises from 38.3 to 39.9 with ResNet-50, from 49.6 to 50.6 with Swin-L, and from 53.9 to 55.2 with ViT-L. These paired improvements show that the reasoner remains useful across different feature representations and host performance levels. The gains also persist at the stricter AP75 threshold: DAQ improves by 1.8, 1.1, and 1.4 points, respectively. These results are consistent with a complementary benefit from state reasoning across the evaluated backbones: a higher host score does not eliminate the observed improvement from adding POSReasoner.

Across YouTube-VIS 2019, 2021, and 2022, AP and AP75 improve in all 15 reported host–version comparisons, including the provisionally marked LOMM/ResNet-50 configuration. CTVIS gains 1.0, 1.1, and 2.1 AP across the three versions, while DAQ gains 0.9, 0.9, and 0.8. LOMM with ViT-L also improves on every version, by 0.7, 0.8, and 1.0 AP. The repeated direction of these paired changes supports the applicability of the state interface across the evaluated VIS systems. Comparing the versions separately also preserves their distinct evaluation settings, rather than combining their scores into a single aggregate.

AP50 and AP75 help interpret these results alongside overall AP. AP75 requires closer mask overlap, and its improvements show that the benefit is visible beyond the more permissive AP50 threshold. Since the candidate masks remain fixed within each pair, the comparison concerns how available predictions are selected, associated, and carried through time, rather than training a new mask generator. Video-instance AP jointly reflects these decisions and the resulting track quality. The per-threshold columns provide complementary measurements, while the state and component analyses examine the temporal behavior more directly.

In particular, the controls in Table 4 connect the benchmark improvements to identity-specific history and structured interaction. Host cues alone and shuffled states remain near the frozen host; intact states without a graph reach 36.55 AP, and the full model reaches 37.31 AP. The LVOS study in Section 4.3 separately examines retention, reactivation, and cross-trajectory reasoning, while Appendix C groups recovery events by absence duration. Together, these comparisons give the extended tables a behavioral context: the benchmark pairs establish where the reasoner helps, and the targeted analyses examine the roles of persistent history, renewed observations, and competing identities within that reasoning process.