SEKAI: INCREMENTAL EVENT MEMORY FOR GENERALIZED REASONING VIDEO SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Online reasoning video segmentation (Online RVS) requires a model to segment query-referred objects as a video streams in, using only the frames observed so far. Existing formulations largely assume a fixed target set throughout the video, overlooking a fundamental property of streaming reasoning: as new events occur, the objects that satisfy a query may change over time. To address this, we propose a new task, **Generalized Reasoning Video Segmentation (GRVS)**, where the target set depends on query-relevant events accumulated over the observed video history. As new evidence accumulates, the target set can evolve from empty to one or multiple objects, requiring a model to continuously maintain target validity while localizing the valid objects currently in view. To evaluate GRVS, we build **Isekai**, a diagnostic benchmark of 60 videos and 197 queries across 11 temporal query families, with 12,569 scoring points whose target sets are causally annotated over time. We further propose **Sekai**, an online framework with two complementary components: an **Event Memory Reasoner (EMR)** that maintains query-relevant events and evolving target states in structured memory, and a **Memory-Grounded Decoder (MGD)** that identifies valid targets from this memory and localizes them in the current frame. Experiments show that Sekai significantly outperforms existing online RVS methods, achieving 50.33 J&F and improving over the strongest baseline by 3.2 points. Code, example annotations, and evaluation scripts are available at `https://anonymous.4open.science/r/Sekai`.

1 INTRODUCTION

Online Reasoning Video Segmentation (Online RVS) brings language-guided segmentation to streaming video, where the system can only access frames observed so far and must produce segmentation results on the fly. Unlike offline settings with the complete video available, the model in online RVS cannot access future context or defer its decision to the end of the sequence. Online segmentation is thus essential for real-world applications that must continuously adapt to an unfolding visual stream.

However, the online setting exposes a fundamental yet largely overlooked challenge: the segmentation target can evolve as the video unfolds. Consider the example in Figure 1, where the query asks, "Which dogs have stood up?". Before either dog stands up, there is no valid target; after the first dog stands up, only that dog should be segmented; once the second dog also stands up, both dogs become valid targets. More generally, the target set can grow, shrink, or switch between instances as new evidence arrives. Two common online paradigms struggle with such evolving targets. Reference-tracking methods ground the referred object once and then follow it, regardless of whether other instances later satisfy the query, and thus miss the second dog. Window-based methods ground the query only within the current observation window; once the first dog's standing-up event leaves the window, they output an empty mask, even though the dog remains a valid target.

Yet existing online methods are largely built around a fixed-target formulation and cannot adapt to evolving target sets. Both paradigms above share a common cause: they do not explicitly record which objects have already satisfied the query and what events support that decision. Window-based methods discard evidence once it leaves the current window. Reference-tracking methods use temporal memory mainly to propagate the identified target across frames, through visual features or cached tokens. Such memory captures what the target looks like, but not how the target set changes

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

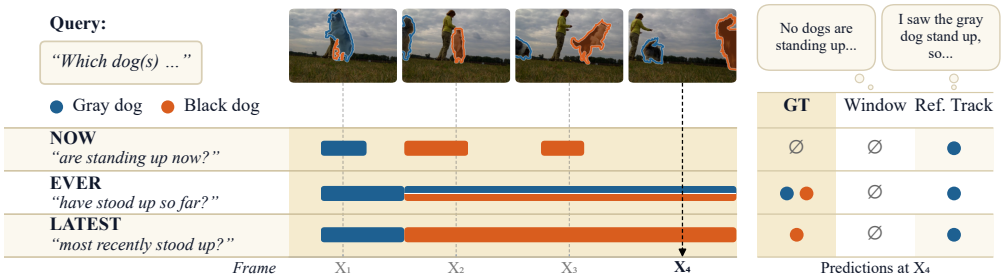


Figure 1: **Generalized Reasoning Video Segmentation (GRVS).** Target identities dynamically evolve under causal temporal contexts, leading to varying targets that standard sliding-window or tracking-based baselines fail to capture.

as new events occur. The same limitation appears at the evaluation level: existing benchmarks typically assign a fixed target set to the entire video, and thus cannot directly test whether a model correctly follows changes in the valid target set over time.

Therefore, we formalize a new task, **Generalized Reasoning Video Segmentation** (GRVS). In GRVS, the model processes the video as a stream. At each moment, the valid target set is determined by the query-relevant events observed so far, rather than by a one-time reference or the current window alone. As the stream evolves, this set may change in both size and membership. Importantly, target validity is also distinct from instantaneous visibility: an object may remain a valid target even when temporarily occluded or outside the frame, and should be segmented again once it becomes visible. GRVS thus generalizes reasoning video segmentation in three aspects: it operates causally on a stream, allows the target set to vary in size and over time, and separates target validity from visibility. These properties give rise to two fundamental requirements: maintaining the evolving target state over time and localizing the currently visible valid targets at each moment.

To evaluate GRVS, we construct **Isekai**, a diagnostic benchmark of 60 real videos, 197 queries, and 12,569 scored query-frame instances. Unlike existing benchmarks, Isekai annotates targets causally: the valid target set at each frame depends only on what has been observed up to that frame, and thus changes over time. The queries span 11 temporal query families, ranging from current states, past occurrences, and most recent events to absent events, event counts, and temporal order. Instances are further stratified by how long ago the supporting evidence occurred, which allows us to measure how well a model retains past evidence.

To meet these two requirements, we introduce **Sekai**, an online framework whose name summarizes its design: knowledge of query-relevant events is accumulated in increments, one arriving chunk at a time, in a structured text memory, and segmentation is decoded from this knowledge rather than from the current window alone. As each video chunk arrives, the Event Memory Reasoner (EMR) maintains the evolving target state by recording query-relevant events in this memory, keeping, for each object, what it has done and when. The Memory-Grounded Decoder (MGD) then reasons over the memory to decide which objects currently satisfy the query, and localizes those visible in the current frame. Experiments on Isekai show that Sekai consistently outperforms existing online methods, demonstrating its effectiveness in reasoning over dynamically evolving targets.

Our contributions are:

- We formalize GRVS, a streaming task in which query-relevant events observed so far determine a time-varying target set of variable size, and target validity is decoupled from instantaneous visibility.
- We introduce Isekai, a diagnostic benchmark for GRVS spanning 11 temporal query families. To the best of our knowledge, it is the first benchmark to provide causally defined, time-varying target annotations.
- We propose Sekai, an online segmentation framework for GRVS that records query-relevant events in a structured text memory, and decodes the resulting variable-sized target set into current-frame masks.

Table 1: **Benchmark annotation.** Comparison of annotations in existing benchmarks and Isekai.

Benchmark	Causal labels	Empty	Multi	Dynamic membership	Evidence age	Temporal operators	Temporal grounding
ReasonVOS (Bai et al., 2024)	✗	✗	✗	✗	✗	✗	Object tracks
ReVOS (Yan et al., 2024)	✗	✓	✓	✗	✗	✗	Object tracks
MeViS v2 (Ding et al., 2025)	✗	✓	✓	✗	✗	✗	Object tracks
GroundMoRe (Deng et al., 2025)	✗	✓	✓	✓	✗	✗	Motion intervals
ORVOSB (Liu et al., 2026)	✓	✓	✗	✗	✗	✗	Validity intervals
Isekai	✓	✓	✓	✓	✓	✓	Event records

2 Related Work

Reasoning Video Segmentation. Referring video object segmentation grounds language expressions to video masks, with motion-centric expressions making temporal dynamics an explicit part of the query (Ding et al., 2023). Reasoning variants extend expressions to implicit descriptions that require world knowledge or multi-step inference, typically by decoding multimodal language model representations into masks or tracks (Bai et al., 2024; Yan et al., 2024; Zheng et al., 2025; Munasinghe et al., 2025; Gong et al., 2025; Yuan et al., 2026). A second line decouples semantic selection from spatial-temporal propagation, either by prompting propagation from reasoned keyframes (Kao et al., 2026) or by pairing candidate selection with dedicated tracking stages (Jin et al., 2026; Zhu et al., 2026). Generalized referring segmentation introduced empty- and multi-target outputs for images (GRES (Liu et al., 2023), GSVA (Xia et al., 2024)), but video formulations still assume static instance identities across the sequence. GRVS instead makes eligibility a function of prefix-confirmed events.

Memory in Video Segmentation. Existing streaming trackers and reasoning frameworks propagate implicit states across time: target queries for selected instances (Wu et al., 2023), visual and mask memories built on SAM2 (Cuttano et al., 2025; Ravi et al., 2024), recurrent language embeddings or token reservoirs (Zhang et al., 2026b; Liu et al., 2026), or re-evaluated keyframe contexts (Kao et al., 2026). Long-video models similarly retain visual memory banks or compressed KV caches (He et al., 2024; Song et al., 2024; Di et al., 2025; Zhang et al., 2026a; Xiao et al., 2026; Shen et al., 2025). However, these approaches rely on continuous visual features or implicit latent tokens, which struggle to preserve symbolic event histories once visual evidence exits the local temporal window. Sekai instead carries an explicit, editable text memory of objects and prefix-confirmed events (Section 4.1).

Benchmarks. Referring and reasoning VOS benchmarks, including ReVOS (Yan et al., 2024), MeViS v2 (Ding et al., 2025), and ReasonVOS (Bai et al., 2024), associate each expression with static instance identities across the video; recent extensions add motion intervals (Deng et al., 2025) or causal referent shifts (Liu et al., 2026), but allow at most one target at a time, and none annotates the events that establish eligibility. Isekai annotates these events directly and derives time-varying, zero-to-many target sets from them (Table 1).

3 Isekai: Diagnostic Benchmark

3.1 Construction

Sources and scale. Isekai builds on videos and instance masks from MeViS v2 (Ding et al., 2025), GroundMoRe (Deng et al., 2025), and DAVIS 2017 (Pont-Tuset et al., 2017): 60 videos, 197 fixed queries, and 12,569 scoring points (Table 2). Isekai is evaluation-only; Sekai is neither trained nor tuned on it (Section 5.1).

Reviewed event timelines. Starting from automatically proposed object–event intervals, annotators verify each event's participants and boundaries and mark the first frame at which the prefix confirms it. An event is marked ENDED only with observed ending evidence, and events whose order cannot be resolved are excluded from scoring. The target set at a frame is the eligible instances that are also visible in the official masks (Figure 2; Appendix A).

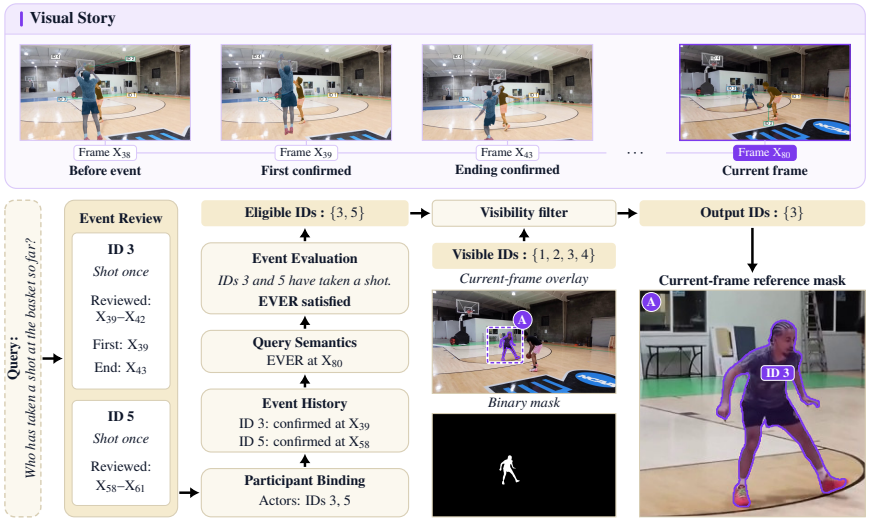


Figure 2: **Isekai annotation example.** A reviewed event is bound to its participants; the EVER predicate selects the eligible IDs, whose intersection with the visible IDs gives the current-frame reference masks.

Annotation agreement. A second annotator, without access to the original labels, re-annotated the events in 20 videos. On the 2,307 resulting NOW/EVER/NEVER scoring points (Section 3.2), the two annotations give identical target sets on 93.0% (mean Jaccard 0.940), although 80.6% of individual events match one-to-one (Appendix A.3). Target sets are thus more stable than the event segmentations that produce them.

Dataset landscape. Figure 3 shows how scoring points distribute over query group, evidence age, and target count; Order pools LATEST, FIRST, and BEFORE/AFTER, and Other pools the remaining compound families. Of the 12,569 points, 5,379 have an empty target set and 1,028 have more than one target, so both departures from the single-target setting are common.

3.2 QUERY FAMILIES

Isekai has eleven query families (Table 6). The five primary families apply one predicate each to the event history: NOW selects the current participants, EVER all participants so far, NEVER their complement, and LATEST and FIRST the most recent and the earliest one.

The six compound families add counting, ordering, role chains, and duration; "Who was the first player to take a shot after dribbling the ball themselves?" must bind the shooter to an earlier dribbling event by the *same* instance. Appendix A.1 gives the formal predicates and boundary cases. Results pool the 38 queries outside NOW, EVER, LATEST, and NEVER as *Other*.

3.3 DIAGNOSTIC PROTOCOL AND METRICS

Evidence age and history subset. Evidence age is the number of frames from the supporting event's confirmation to the scoring frame, in four strata: in window (confirmed within the current chunk), 1–8, 9–24, and 25+. A scoring point is in the *history subset* if its target set changes when recomputed from only the events confirmed in the current chunk, i.e., its answer depends on evidence outside the window. The subset is defined on NOW, EVER, and NEVER, whose targets follow from the event rules, and contains 4,133 of its 8,980 points.

Metrics. We report region IoU $\mathcal{J}$ and boundary $\mathcal{F}$, computed at each scoring point between the union of predicted masks and the union of reference masks; two empty unions score $\mathcal{J} = \mathcal{F} = 1$, one empty union scores zero. Scores are averaged within each video and then across videos; subgroup results are filtered to their points before averaging. Every frame is scored, so a system that updates its targets only at chunk endpoints is penalized on each frame between a membership change and its next update.

Table 2: **Statistics of Isekai.**

Source	Videos	Queries	Frames	Points
GroundMoRe	41	143	2,550	9,010
MeViS v2	7	23	450	1,448
DAVIS 2017	12	31	823	2,111
Total	60	197	3,823	12,569

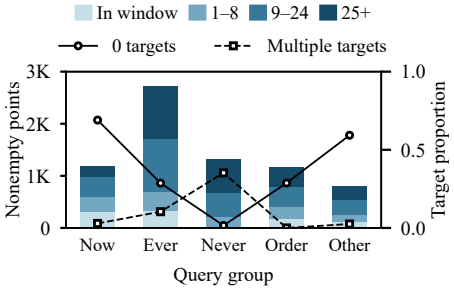


Figure 3: **Landscape of Isekai.** Bars count nonempty points by evidence age; the lines show zero-/multiple-target proportions over all points in each group. Appendix A.5 gives timing and grouping rules.

4 METHOD

GRVS Task Definition. A fixed query q is given before the video stream $x_{0:T}$. At time $t \in \{0, \dots, T\}$ the model sees only the prefix $x_{0:t}$ and cannot revise masks it has already emitted. Let $\mathcal{I}_t$ be the instances observed so far, $s_{i,t} \in \{0, 1\}$ indicate whether i satisfies q given $x_{0:t}$, and $v_{i,t} \in \{0, 1\}$ whether i is visible at t. The target set and the required output are

$$\mathcal{A}_t = \{i \in \mathcal{I}_t : s_{i,t} = 1\}, \qquad \mathcal{Y}_t = \{(i, m_{i,t}) : i \in \mathcal{A}_t,\ v_{i,t} = 1\}, \tag{1}$$

where $m_{i,t} \in \{0, 1\}^{H_t \times W_t}$ is a nonempty mask of the visible region. Eligibility and visibility are decoupled: an event can flip $s_{i,t}$ while i stays in view, and occlusion can flip $v_{i,t}$ while i stays eligible. Hence $|\mathcal{Y}_t|$ may be zero, one, or many, and an empty output states that no *visible* instance currently qualifies.

Because q refers to events rather than appearance, $s_{i,t}$ depends on history. An event $e = (P_e, [a_e, b_e], \tau_e)$ has participants $P_e \subseteq \mathcal{I}_t$, an interval $[a_e, b_e]$, where b_e is the last active frame, set only once the ending is observed at $b_e + 1$ ($b_e = \infty$ if no ending is observed), and a confirmation frame $\tau_e \geq a_e$, the first frame at which the prefix suffices to confirm e. The causal timeline $\mathcal{E}_t = \{e : \tau_e \leq t\}$ grows monotonically, and every query family is a predicate over it (Appendix A.1).

Framework Overview. Sekai splits the two requirements of Equation 1, historical eligibility and current localization, between two modules (Figure 4). The Event Memory Reasoner (EMR) is a frozen writer W_{θ_W} that maintains a structured text memory M_k, a bounded approximation of $\mathcal{E}_t$, across chunks. The Memory-Grounded Decoder (MGD) is a LoRA-adapted locator L_{θ_L} followed by a frozen segmentor S_Φ; it reads M_k against the current chunk and produces the target masks.

After an initialization call on x_0, the stream is consumed in chunks of 8 frames, $X_k = x_{a_k:b_k}$ (indexing in Appendix B.1). At call k,

$$\begin{aligned} &\text{EMR:} \quad M_k = W_{\theta_W}(X_k, q, M_{k-1}), \\ &\text{MGD:} \quad P_k = L_{\theta_L}(X_k, q, M_k) = \{p_{k,j}\}_{j=1}^{n_k}, \quad \widehat{\mathcal{Y}}_{b_k:b_{k+1}-1} = S_\Phi(x_{b_k:b_{k+1}-1}, P_k), \end{aligned} \tag{2}$$

with $M_0 = \emptyset$; P_k holds $n_k \geq 0$ boxes on the endpoint frame x_{b_k}, and the last propagation interval runs to the end of the video. Writer and locator start from a fresh context at every call, with no KV cache or visual features carried over, so M_k is the only semantic state passed between the modules or across chunks; the segmentor alone keeps a visual propagation state.

4.1 EVENT MEMORY REASONER

The EMR records event history in an editable text memory so that it survives after the supporting frames have left the window.

Memory structure. Each object entry holds a stable local ID, an appearance phrase, its last visible location and frame, and a list of events; each event holds an action label, a state in

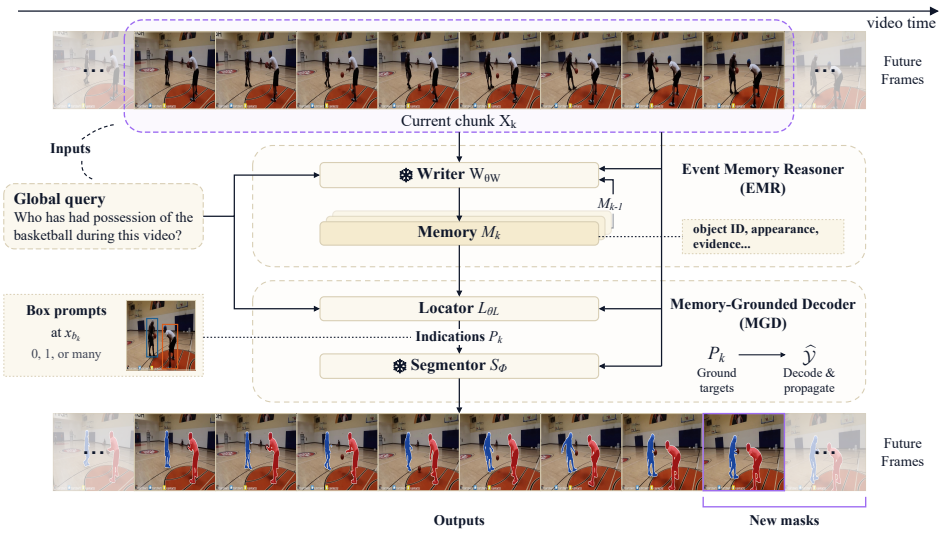


Figure 4: **Sekai.** The EMR maintains event memory; the MGD emits zero, one, or many boxes P_k and decodes and propagates their masks. Snowflakes mark frozen components.

{ACTIVE, ENDED, UNKNOWN}, and its first and most recent evidence frames. The rendered memory is capped at 128 tokens, so the carried state is $O(1)$ in stream length. For the video in Figure 1, the memory at frame x_4 reads

```
{"objects":[
 ["o1","gray dog","left of frame",53,[["stand-up","ended",5,11]]],
 ["o2","black dog","right image boundary",53,
   [["stand-up","ended",15,22],["stand-up","ended",31,37]]]]}
```

where 11, 22, and 37 are the frames at which each rearing was first seen to have ended.

Reasoning operations. At each call the writer performs three prompt-specified operations. (i) *Identity binding*: keep local IDs stable across calls and never transfer one object's history to another. (ii) *State transition*: an event is ACTIVE while visibly ongoing, ENDED only once its ending has been observed, and UNKNOWN otherwise; disappearance alone does not end an event, matching the Isekai annotation rule (Appendix A.1). (iii) *Consolidation*: keep supported history and uncertainty, and do not invent events.

Relation to the causal timeline. Each field approximates one quantity of an event (owner $\leftrightarrow P_e$, entry presence $\leftrightarrow \tau_e \leq t$, ACTIVE/ENDED $\leftrightarrow t \in [a_e, b_e]$, first evidence $\leftrightarrow a_e$), so every primary predicate (Eq. 3, Appendix A.1) has a direct counterpart on M_k.

4.2 MEMORY-GROUNDED DECODER

The MGD turns the EMR's memory into masks on the current frames. We call this memory-grounded decoding: the query is applied to the memory, not to the chunk alone.

Memory readout and grounding. The locator reads q, M_k, and X_k and returns boxes on the endpoint frame x_{b_k} for every visible instance that satisfies q. This involves two steps: applying the query's temporal condition to the memory entries, which determines who qualifies, and matching the qualifying entries' appearance phrases and last locations against x_{b_k}, which determines where they are. An empty list is a valid output and means no visible instance currently qualifies.

Mask decoding and propagation. S_Φ converts the $n_k \geq 0$ boxes into masks on x_{b_k} and propagates them over arriving frames until the next endpoint, where new prompts replace them. Because prompts are recomputed from M_k at every endpoint, a change in the target set, such as a new EVER participant or an ended NOW event, takes effect at the next endpoint without revising earlier outputs.

Table 3: **Online methods on Isekai under identical causal inputs.** Family columns report J&F and mix empty and nonempty frames in different proportions (Figure 3). *Other* pools the families outside Now/Ever/Latest/Never; *Hist.* is the history subset (Section 3.3).

Method	J	F	J&F	Now	Ever	Latest	Never	Other	Hist.
OnlineRefer (Wu et al., 2023)	28.08	33.38	30.73	20.92	39.47	34.75	29.10	25.93	48.87
SAMWISE (Cuttano et al., 2025)	33.52	37.07	35.30	22.30	46.94	40.17	34.07	27.52	57.93
CoT-RVS (Kao et al., 2026)	46.43	47.90	47.16	**59.31**	39.52	25.81	18.00	**53.63**	22.15
StreamingRVOS (Zhang et al., 2026b)	39.34	43.23	41.28	29.40	55.18	47.26	**38.17**	33.28	**68.71**
Sekai w/o memory	44.33	46.89	45.61	34.60	59.88	**50.73**	37.51	36.43	63.93
Sekai	**49.09**	**51.58**	**50.33**	42.04	**64.46**	44.76	27.00	48.42	65.97

Table 4: **Results on existing benchmarks.** MeViS v2 results are in Appendix D.5.

Method / Configuration	**Online**	**ReasonVOS**	**ReVOS (Overall)**
VISA (13B, IT) (Yan et al., 2024)	✗	–	50.90
GLUS (7B) (Lin et al., 2025)	✗	49.90	54.90
CoT-RVS (GPT-4o, offline) (Kao et al., 2026)	✗	65.50	55.90
Refer-Agent (9B) (Jiang et al., 2026)	✗	69.80	61.30
VideoSEG-O3 (4B) (Dai et al., 2026)	✗	62.90	67.70
OnlineRefer (Wu et al., 2023; Bai et al., 2024)	✓	38.70	–
CoT-RVS (LLaVA-1.5-7B, online) (Kao et al., 2026)	✓	52.00	46.20
ORVOS (Liu et al., 2026)	✓	–	54.30
Sekai (ours)	✓	59.00	57.29

Training. We freeze the writer, the vision backbone, and the segmentor, and adapt only the locator with LoRA under response-only cross-entropy on the serialized boxes, $\mathcal{L} = -\frac{1}{|y^*|}\sum_r \log p_{\theta_L}(y_r^* \mid X_k, q, M_k, y_{<r}^*)$. Training pairs combine writer-generated memories with annotated endpoint boxes. They are supplemented by synthetic memories with substituted event states and query families, which are intended to make the target depend on the memory rather than on the chunk alone. Data and hyperparameters are given in Appendix B.3.

5 Experiments

5.1 Experimental Setup

Data and splits. We evaluate on Isekai (Section 3) and, with their original queries and masks, on MeViS v2 (Ding et al., 2025), ReasonVOS (Bai et al., 2024), and ReVOS (Yan et al., 2024). MeViS v2 is in-domain for the locator; ReVOS uses its full validation set.

Implementation. The EMR's writer and the MGD's locator are both Qwen3-VL-8B-Instruct (Bai et al., 2025); only the locator is LoRA-adapted, on MeViS v2 training data disjoint from Isekai (Appendix B.3). SAM 3.1 (Carion et al., 2025) is the segmentor. Isekai is scored as in Section 3.3; the other benchmarks use their official metrics on causal predictions.

Baselines and controls. We adapt four online methods to the same arrived-frame schedule and audit every input against future frames (Appendix C): OnlineRefer (Wu et al., 2023) propagates a target query, SAMWISE (Cuttano et al., 2025) a causal visual window, and StreamingRVOS (Zhang et al., 2026b) semantic tokens; our CoT-RVS (Kao et al., 2026) adaptation shares Sekai's Qwen3-VL-8B and SAM 3.1 and rereads a compressed grid of the arrived prefix at each refresh. Two controls bracket Sekai: *Sekai w/o memory* removes the EMR, so the locator sees only q and X_k; *GT boxes + SAM 3.1* feeds ground-truth boxes to the segmentor at the same refresh points and bounds what mask decoding alone can reach.

5.2 Main Results

Isekai. Sekai obtains the best overall J&F and the best score on Ever (Table 3). The remaining family columns are split among the baselines, and Table 11 shows why: CoT-RVS is strong on empty points but collapses on single-target points, StreamingRVOS the reverse, so each leads the families whose empty-frame proportion (Figure 3) matches its bias toward withholding or emitting

a mask. Sekai is the only method without a collapsed regime across empty, single-, and multi-target points, which is what the overall score rewards.

Where the gain comes from. The gain is concentrated on EVER, where Sekai leads the strongest baseline by 9.28 and the memory-free control by 4.58: an EVER target stays eligible after its event ends and its evidence leaves the window, which is exactly the state the EMR carries. The control also trails on NOW and the compound families but is higher on LATEST and NEVER, where the same memory becomes a liability: a stale or spurious entry is never revisited, so on LATEST an older participant is not superseded and on NEVER one wrong entry excludes an instance for the rest of the video (deltas in Appendix D.3). On the history subset Sekai is above the control and within 3 points of StreamingRVOS. With ground-truth boxes the same segmentor reaches 87.80 (Table 5), so the remaining error lies mostly in target selection, not mask decoding.

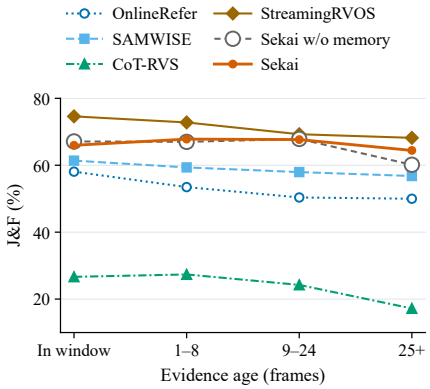


Figure 5: **Performance by evidence age on Isekai.** J&F over nonempty scoring points in the four strata, averaged within videos then across videos. All methods use the same points in each stratum.

Existing benchmarks. On ReasonVOS and ReVOS, which fix target identity per video and so test referring and segmentation quality, Sekai is the highest online entry (Table 4).

5.3 ABLATIONS AND ANALYSIS

Evidence age. Figure 5 scores every method on the same nonempty points per stratum (values in Table 10). Every baseline declines with evidence age; the memory-free control holds through 24 frames and drops 7.7 points at 25+, while Sekai is the only method whose curve is flat. The crossovers have a mechanical reading: for two or three chunks the qualifying instance is usually still the salient actor in X_k, so an appearance prior recovers it without any record of the event; beyond that, only the record identifies who acted. Two caveats bound this. The strata exclude empty frames and so reward emitting a mask: StreamingRVOS, which almost never withholds one, leads every stratum, and even the GT-box oracle scores 41.43 in window because a target confirmed mid-chunk is not prompted until the next endpoint, so the in-window stratum measures anticipation as much as reasoning. Conversely, Table 11 shows the memory is neutral on nonempty points and gains 11.11 on empty ones; the strata therefore show only half of its contribution, retaining old-evidence targets, and cannot register the other half, withholding a mask when nothing visible qualifies.

Module ablations. Table 5 ablates one module at a time. *EMR.* Removing the memory costs 4.72 J&F. Removing only the state labels costs 1.76 overall, since wording and evidence frames still encode most of the state, but on the history subset a memory without states falls below no memory: an entry that no longer says whether its event ended reads as still active, so a record that omits state can be worse than none.

Table 5: **Module ablations on Isekai.**

Module	Condition	All J&F	Hist. J&F
—	Sekai	50.33	65.97
EMR	w/o memory	45.61	63.93
EMR	No state labels	48.57	62.87
MGD	Unadapted locator	27.66	40.34
MGD	SAM2.1 segmentor	47.92	60.26
GT boxes + SAM 3.1		87.80	90.85

MGD. The unadapted locator loses 22.67 on the same memory, and its two scores locate what the adaptation teaches. The base model finds the instance the memory names but cannot decline to emit a box; the LoRA's first effect is abstention, the same regime where the memory contributes most. The segmentor matters least: SAM2.1 (Ravi et al., 2024) costs 2.41, whereas backends that hold an anchor mask without propagating fall to 24–29 (Appendix B.5). With prompts only at endpoints and every frame scored, the segmentor buys coverage of the frames between endpoints, not a better endpoint mask.



Figure 6: **Qualitative examples on Isekai.** Orange dashed outlines mark the sixth-frame endpoint; yellow boxes are the MGD locator's saved outputs.

5.4 Qualitative Results

Case Study. Figure 6 shows two correct selections and one failure from the run in Table 3, all COUNT queries ("at least two separate occasions"), the family that exercises both modules: the writer must keep the occurrences as separate closed entries, and the locator must count entries rather than read the chunk, since an eight-frame chunk typically shows at most one occurrence and reveals that it is the second. In the first two rows the MGD selects the runner and the breakdancer at frame 56; a chunk-only reader can make these selections only by guessing the salient actor, the behavior the 9–24 stratum rewards and the 25+ stratum exposes.

Failure Analysis. In the third row, Sekai selects the dancer at frame 64, although the label turns positive only at frame 79; the error is one of timing, not identity. Both modules err. The EMR stores the arm-raising as a single ACTIVE entry with evidence frames 0 and 64: because an entry closes only on observed ending evidence, a missed ending lets the writer extend one entry rather than open a second, merging the "separate occasions" the query asks about. COUNT on this memory gives one, so a faithful readout would have withheld a box; that the MGD emitted one indicates it fell back to the chunk, where the raised arms are visible. Since the locator does not emit the IDs it selects, the two errors cannot be separated in the output; the case instantiates the pattern behind the LATEST and NEVER deficits in Table 3, a record wrong once and never revisited, and motivates both the ID readout and the re-verification step of Section 6.

6 Limitations and Conclusion

Limitations. Sekai's carried state is a text memory written by a frozen model, and the design accepts a trade-off: what the writer records stays available indefinitely, but what it misrecords is not revisited, since later calls cannot re-examine frames that have left the window. Knowledge accumulated in increments also accumulates its errors. Table 3 shows both sides: the same memory that carries EVER targets past their evidence can also hold a stale entry that hurts LATEST and NEVER, so a verification step that re-checks entries against new evidence is the natural next component. The locator, adapted partly on synthetic memories, selects from the memory without emitting the IDs it selects; an explicit ID readout would make cases such as the third row of Figure 6 diagnosable at the module level.

Conclusion. GRVS treats the streaming target set as a function of the observed event history that may be empty, single, or plural at any frame. Isekai measures this frame by frame with event-derived targets and evidence-age strata. Sekai shows that a bounded text record of who did what, and when, retains targets after their evidence leaves the window, with its clearest margin over prior methods on EVER queries and, relative to a memory-free control, on the oldest evidence. Keeping that record correct, and reading it for recency and negation, remain open.

Generative AI Usage Statement

We used generative AI tools to help organize and edit the manuscript, search the literature, and write plotting code. The authors designed and ran all experiments and checked all technical content and results.

Ethics Statement

The work uses existing video benchmarks and their annotations, with data sources and usage conditions recorded alongside the assets. Models that bind past events to visible people or objects can make consequential identity errors; deployment in sensitive settings requires evaluation beyond the present research benchmarks.

Reproducibility Statement

Appendix B summarizes the prompts (full text released with the code), locator training data and hyperparameters; Appendix C describes the causal inputs for every baseline. The code, including the full prompt text, evaluation scripts, and example Isekai annotations, is available at the anonymous repository linked in the abstract. The full Isekai annotations will be released upon publication.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2409.19603.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. SAMWISE: Infusing wisdom in SAM2 for text-driven video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3395–3405, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cuttano_SAMWISE_Infusing_Wisdom_in_SAM2_for_Text-Driven_Video_Segmentation_CVPR_2025_paper.html.

Ming Dai, Sen Yang, Boqiang Duan, Boyuan Tong, Jiedong Zhuang, Wankou Yang, and Jingdong Wang. VideoSEG-O3: A Multi-turn Reinforcement Learning Framework for Reasoning Video Object Segmentation. *arXiv preprint arXiv:2606.06819*, 2026. URL https://arxiv.org/abs/2606.06819.

Andong Deng, Tongjia Chen, Shoubin Yu, Taojiannan Yang, Lincoln Spencer, Yapeng Tian, Ajmal Saeed Mian, Mohit Bansal, and Chen Chen. Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8625–8636, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Deng_Motion-Grounded_Video_Reasoning_Understanding_and_Perceiving_Motion_at_Pixel_Level_CVPR_2025_paper.html.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video KV-cache retrieval. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.00540.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://henghuiding.com/MeViS/.

Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. MeViS: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. URL https://henghuiding.com/MeViS/.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The Devil is in Temporal Token: High Quality Video Reasoning Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2501.08549.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13504–13514, 2024. URL https://arxiv.org/abs/2404.05726.

Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13332–13341, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/He_Decoupling_Static_and_Hierarchical_Motion_Perception_for_Referring_Video_Segmentation_CVPR_2024_paper.html.

Haichao Jiang, Tianming Liang, Wei-Shi Zheng, and Jian-Fang Hu. Refer-Agent: A collaborative multi-agent system with reasoning and reflection for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2602.03595.

Woojeong Jin, Jaeho Lee, Heeseong Shin, Seungho Jang, Junhwan Heo, and Seungryong Kim. AgentRVOS: Reasoning over object tracks for zero-shot referring video object segmentation. *arXiv preprint arXiv:2603.23489*, 2026. URL https://arxiv.org/abs/2603.23489.

Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. CoT-RVS: Zero-shot chain-of-thought reasoning segmentation for videos. In *International Conference on Learning Representations*, 2026. URL https://danielshkao.github.io/cot-rvs.html.

Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. GLUS: Global-Local Reasoning Unified into A Single Large Language Model for Video Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2504.07962.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23592–23601, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Liu_GRES_Generalized_Referring_Expression_Segmentation_CVPR_2023_paper.pdf.

Jinyuan Liu, Yang Wang, Zeyu Zhao, Weixin Li, Song Wang, and Ruize Han. Online reasoning video object segmentation. *arXiv preprint arXiv:2604.11411*, 2026. URL https://arxiv.org/abs/2604.11411.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. VideoGLaMM: A Large Multimodal Model for Pixel-Level Visual Grounding in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2411.04923.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. URL https://arxiv.org/abs/1704.00675.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Yiqing Shen, Bohan Liu, Chenjia Li, Lalithkumar Seenivasan, and Mathias Unberath. Online reasoning video segmentation with just-in-time digital twins. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24698–24706, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Shen_Online_Reasoning_Video_Segmentation_with_Just-in-Time_Digital_Twins_ICCV_2025_paper.html.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2307.16449.

Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. OnlineRefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2307.09356.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3858–3869, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xia_GSVA_Generalized_Segmentation_via_Multimodal_Large_Language_Models_CVPR_2024_paper.html.

Junbin Xiao, Jiajun Chen, Tianxiang Sun, Xun Yang, and Angela Yao. MuKV: Multi-grained KV cache compression for long streaming video question-answering. *arXiv preprint arXiv:2605.22269*, 2026. URL https://arxiv.org/abs/2605.22269.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/02307.pdf.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with MLLM for Dense Grounded Understanding of Images and Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. URL https://arxiv.org/abs/2501.04001.

Haowei Zhang, Shudong Yang, Jinlan Fu, See-Kiong Ng, and Xipeng Qiu. HERMES: KV cache as hierarchical memory for efficient streaming video understanding. *arXiv preprint arXiv:2601.14724*, 2026a. URL https://arxiv.org/abs/2601.14724.

Wenkang Zhang, Kaicheng Yang, Xiang An, Qiang Li, Ziyong Feng, Wankou Yang, and Jiankang Deng. Towards streaming referring video segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_Towards_Streaming_Referring_Video_Segmentation_via_Large_Language_Model_CVPR_2026_paper.html.

Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, and Hengshuang Zhao. ViLLa: Video reasoning segmentation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23667–23677, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zheng_ViLLa_Video_Reasoning_Segmentation_with_Large_Language_Model_ICCV_2025_paper.html.

Zhengtong Zhu, Jiaqing Fan, Zhixuan Liu, and Fanzhang Li. Training-Free Spatio-temporal Decoupled Reasoning Video Segmentation with Adaptive Object Memory. *arXiv preprint arXiv:2603.01545*, 2026. URL https://arxiv.org/abs/2603.01545.

Table 6: **Isekai query families.** Primary families apply a single predicate to the reviewed event history; compound families compose temporal conditions with counting, ordering, negation, or duration. n: number of fixed queries; scoring points count query–frame pairs marked for scoring.

Family	Meaning	n	Scoring points	Example
NOW	Participates now	60	3,821	"Which dog is rearing up?"
EVER	Participated so far	60	3,820	"Which dogs have reared up so far?"
LATEST	Participant of the most recently begun occurrence	18	1,151	"Which dog most recently reared up?"
NEVER	No occurrence confirmed so far	21	1,339	"Which dog has not reared up?"
FIRST	Participant of the earliest-begun occurrence	5	306	"Which dog reared up first?"
PAST-NOT-PRESENT	Participated before, but not now	4	246	"Which dogs have reared up but are not rearing now?"
COUNT	Number of occurrences	9	676	"Which dog reared up at least twice?"
TWO-HOP / OBJECT-HISTORY	Participant roles across events	7 / 6	460 / 372	"first player to shoot after dribbling"
BEFORE/AFTER / CONTINUOUS	Event order / duration	4 / 3	186 / 192	"Who reared up before the other dog?" / "Who kept moving?"

A ISEKAI CONSTRUCTION AND REVIEW

This appendix gives the release details, review safeguards, and scoring conventions for Isekai. The main text reports the 60 videos, 197 fixed queries, and 12,569 scored query–frame records; this appendix records the construction and scoring details.

A.1 PER-FAMILY DEFINITIONS AND SAFEGUARDS

NOW selects currently participating instances. EVER retains every participant confirmed so far. LATEST selects the participant of the most recently begun occurrence, and FIRST selects the participant of the earliest-begun occurrence. PAST-NOT-PRESENT combines a past event with a known current nonparticipation state. COUNT, OBJECT-HISTORY, TWO-HOP, BEFORE/AFTER, and CONTINUOUS add counting, participant roles, relation chains, ordering, or interval constraints. NEVER is the prefix complement of EVER, $s^{\text{NEVER}}_{i,t} = 1 - s^{\text{EVER}}_{i,t}$, so its set contracts as qualifying events are confirmed.

Using the event notation in Section 4, the temporal predicates are

$$s^{\text{NOW}}_{i,t} = \mathbf{1}\{\exists e \in \mathcal{E}_t :\ i \in P_e,\ a_e \le t \le b_e\}, \qquad s^{\text{EVER}}_{i,t} = \mathbf{1}\{\exists e \in \mathcal{E}_t :\ i \in P_e\},$$
$$s^{\text{LATEST}}_{i,t} = \mathbf{1}\{i \in P_{e^\star_t}\}, \quad e^\star_t = \arg\max_{e \in \mathcal{E}_t} a_e, \tag{3}$$
$$s^{\text{FIRST}}_{i,t} = \mathbf{1}\{i \in P_{\underline{e}_t}\}, \quad \underline{e}_t = \arg\min_{e \in \mathcal{E}_t} a_e, \qquad s^{\text{NEVER}}_{i,t} = 1 - s^{\text{EVER}}_{i,t}.$$

Crucially, while typically $\tau_e \in [a_e, b_e]$, fast transitional events (e.g., a brief jump confirmed only upon landing) can have $\tau_e > b_e$; in strict causal streaming, such an event never enters NOW but commits directly to EVER and LATEST once confirmed at τ_e. When $\mathcal{E}_t$ is empty, LATEST and FIRST are false for every instance. The remaining families (Section 3.2) compose these predicates with counting, ordering, or duration. $\mathcal{E}_t$ changes only at confirmation frames τ_e, and NOW additionally at ending frames $b_e + 1$, so $\mathcal{A}_t$ changes only at event boundaries.

For the compound families, PAST-NOT-PRESENT is $s^{\text{EVER}}_{i,t} \wedge \neg s^{\text{NOW}}_{i,t}$, and COUNT tests $|\{e \in \mathcal{E}_t : i \in P_e\}|\ \{=, \ge\}\ c$. TWO-HOP and OBJECT-HISTORY follow a role chain across events; BEFORE/AFTER and CONTINUOUS impose order and interval constraints, respectively.

We require prefix-confirmed evidence for every positive event and observed evidence of ending before marking an event ended. Disappearance alone cannot certify an ending. We distinguish

Table 7: **Independent annotation audit.** Event matching uses same video and event type, overlapping intervals, and exact participant–role sets. Participant agreement uses the relaxed same-type, overlapping-interval pairing.

Measure	Value
Event annotations (20 audited videos; 36 independent, 34 prior events)	
Strictly matched events	29
Independent-event match rate	80.6%
Prior-event coverage	85.3%
Mean interval IoU among matched events	0.916
Matched intervals with IoU ≥ 0.5	100%
Matched pairs crossing a chunk endpoint boundary	9
Primary causal labels (2,307 scoring points)	
Exact output-set agreement	93.0%
Mean output-set Jaccard	0.940
Empty/nonempty accuracy	94.9%
Empty/nonempty Cohen's κ	0.962

known absence, qualified-but-invisible instances, insufficient evidence, and missing annotation. The last two are withheld from determinate scoring.

Counterparts on the EMR memory. The predicates have direct counterparts on M_k (Section 4.1): NOW checks for an ACTIVE entry, EVER for any entry, NEVER for none, LATEST and FIRST pick the owner of the entry with the latest or earliest first evidence, and COUNT counts entries.

A.2 REVIEW PROCEDURE AND RELEASE GATE

Machine proposals provide candidate object–condition intervals. Reviewers inspect continuous frames, identify participants, correct inclusive boundaries, and inspect every object's complete video, including objects that never qualify. They record participant roles, first prefix-confirmable evidence, endings, evidence frames, and uncertainty. Reviewers may inspect the full video to locate an event, but each label must be justified from the observed prefix. Concurrent or unresolved order remains marked unresolved.

For each scored (q, t), we derive `eligible_ids` from reviewed events and `visible_ids` from official masks, then store their intersection as `output_ids`. Reference IDs link event participants to masks. The release gate checks answer transitions, empty and multiple targets, ordinary sampled episodes, source hashes, and mask availability.

A.3 INTER-ANNOTATOR AGREEMENT

A blind independent audit compares annotations on 20 sampled videos. Table 7 reports the event-level comparison and the recompiled scoring-point comparison. The scoring-point audit covers NOW, EVER, and NEVER only. Latest and compound query families are omitted because the independent export does not separately record the required answer and evidence states or query-specific intervals.

A.4 SCORING DETAILS

All main results use the 12,569 scoring points. At each point, reference masks for the selected visible instances are unioned, as are predicted masks. Region IoU and boundary F are evaluated on these binary foregrounds. Two empty unions score one and a one-sided empty union scores zero; invalid or missing predictions count as empty. We average within each video and then equally over videos. This score does not assess instance separation.

A.5 EVIDENCE AGE

For a scoring frame t, let $c(t) = 0$ at initialization and $c(t) = 1+8\lfloor(t-1)/8\rfloor$ otherwise. The current observed chunk is $x_{c(t):t}$. Restrict the reviewed event history to $\mathcal{E}_t^{\text{win}} = \{e \in \mathcal{E}_t : \tau_e \geq c(t)\}$. A point belongs to the replayable history subset when

$$\{i \in \mathcal{A}_t(\mathcal{E}_t) : v_{i,t} = 1\} \neq \{i \in \mathcal{A}_t(\mathcal{E}_t^{\text{win}}) : v_{i,t} = 1\}.$$

The full-history replay agrees with all 8,980 labels for NOW, EVER, and NEVER. For an ongoing NOW event, its first confirmation can precede the chunk even when the action remains visible; the history criterion concerns event replay rather than visual impossibility.

For a supporting event e, evidence age is $t - \tau_e$. An event confirmed in the current chunk is assigned to *in window*; remaining points fall into 1–8, 9–24, or 25+ frames. FIRST uses the earliest qualifying event and LATEST the most recent. Repeated EVER events use the latest qualifying occurrence for each participant. A multi-target point uses the greatest participant age and is counted once.

Dataset-landscape timing. Figures 3 and 5 use the same timing assignments. Temporal order combines LATEST, FIRST, and BEFORE/AFTER; Other history combines COUNT, TWO-HOP, OBJECT-HISTORY, PAST-NOT-PRESENT, and CONTINUOUS. Target counts use distinct visible reference IDs. The bars and performance curves contain 7,190 nonempty points, including 1,028 multi-target points; the landscape proportions use all 12,569 points. Event-age strata group nonempty labels by evidence timing, whereas the history subset includes empty references whenever window-only event replay changes the target set.

Where an explicit supporting-event timing is unavailable, the current answer interval's onset supplies a timing proxy (1,621 points). For NEVER, the anchor is the earliest confirmed exclusion event (918 points) or the video start when no such event exists (400 points). Temporary invisibility does not reset an answer interval. The four strata contain 972, 1,211, 2,522, and 2,485 points; Appendix D.1 reports their scores.

A.6 TEASER PROVENANCE

The teaser uses DAVIS 2017 (Pont-Tuset et al., 2017) *dogs-jump* and official visible masks for gray dog G and black dog B.

B SEKAI IMPLEMENTATION

This appendix specifies the reference-system implementation and response contracts. It records the active locator adaptation and inference procedure.

B.1 SCHEDULE AND MODULE INTERFACE

The EMR's writer and the MGD's locator use Qwen3-VL-8B-Instruct (Bai et al., 2025); SAM 3.1 (Carion et al., 2025) is frozen. Calls follow the first-frame schedule in Section 4: initialize on x_0, then process $X_k = x_{a_k:b_k}$ with $a_1 = b_1 = 0$ and $a_k = b_{k-1} + 1$. A final partial chunk is processed if the video ends off-schedule. The writer reads q, X_k, and M_{k-1} and replaces the EMR memory. The locator reads q, X_k, and the new memory in a fresh context. Both the EMR memory and locator rendering are validated against the 128-token budget. Event indices record first and most recent evidence, not guaranteed event boundaries.

B.2 EVENT MEMORY REASONER

The active writer instruction requires stable local IDs, observed appearance and locations, event labels, event states, evidence indices, complete replacement output, and a 128-token limit. It forbids inventing events or treating disappearance as an ending.

An over-budget output is rejected and regenerated, with at most three identical attempts. Both the memory and its locator rendering must fit within 128 tokens. In the 1,770 writer-call logs, 11 calls (0.7%) encountered the budget limit. After repeated failure, the previous accepted memory and segmentation state are carried forward; the failure remains in the evaluation denominator.

Table 8: **Segmentor replacement on Isekai.** Video-equal union-mask scores (%) on all released scoring points. The four replacement backends share locator boxes. Hold copies the anchor mask until the next refresh; Propagate updates masks on arriving frames. The GT-box row supplies privileged prompts; its family and history-subset J&F scores are shown below.

Backend	Between anchors	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Filled rectangle	Hold	28.39	20.57	24.48
GrabCut	Hold	30.00	27.35	28.67
SEEM (FocalT-v0)	Hold	29.73	28.78	29.26
SAM2.1 (Hiera-Large)	Propagate	46.47	49.36	47.92
GT boxes + SAM 3.1	Propagate	87.28	88.31	87.80

Prompts	Now	Ever	Latest	Never	Other	Hist.
GT boxes	84.53	89.27	82.64	91.53	91.12	90.85

B.3 Memory-Grounded Decoder

Grammar-constrained decoding validates the single top-level JSON response. The locator always returns a list of boxes, including an empty list. Coordinates use normalized $[0, 1000]$ xyxy values, and malformed, truncated, or degenerate boxes are invalid.

The active reference run freezes the writer, vision backbone, token embeddings, and segmentor, and adapts the locator with response-only cross-entropy. It uses 3,579 real samples and 2,592 synthetic samples from the released MeViS v2 training corpus. Real samples pair queries and current chunks with writer-generated memory states and annotated endpoint boxes; synthetic samples pair object descriptions and endpoint geometry with hypothetical temporal states. The main locator uses LoRA rank 16, scaling 32, dropout 0.05, learning rate 2×10^{-5}, and effective batch size 4 (two real and two synthetic samples).

The active locator instruction asks for every visible instance satisfying the original query in the last frame and permits an empty box list. Full prompt text, validators, and exact serialization are included in the anonymous repository.

Video-level separation. We manually checked the complete source-video manifests for Isekai and locator training, and reviewed all candidate overlaps identified by the available-frame duplicate checks. To the best of our knowledge, none of the 60 Isekai videos overlaps with the source videos used for locator training, including the sources of real samples and the video geometry used in synthetic samples. This check concerns locator-training data; it does not establish separation from foundation-model pretraining data.

B.4 Inference Procedure

1. Initialize empty semantic memory and segmentation state.
2. At each reasoning endpoint, run the EMR (writer) on the arrived chunk, the query, and the previous memory; validate the replacement memory.
3. In a fresh context, run the MGD's locator on the new memory and the same chunk; validate its endpoint boxes.
4. Use the MGD's segmentor to apply valid boxes to the endpoint frame. A valid empty list clears active segmentation state while preserving semantic memory.
5. Use the MGD's segmentor to propagate accepted masks on intervening arrived frames and carry the memory to the next endpoint.

B.5 Segmentor replacement and GT-box oracle

The four locator-prompted branches reuse saved outputs with the saved writer memories.

As in the main evaluation, the scorer unions instance masks within each frame, computes region IoU and boundary F, averages scored frames within each video, and then gives all 60 videos equal weight. Two empty masks score one; a one-sided empty match scores zero. Family scores use

Table 9: **Shared-backbone and shared-segmentor comparison.** Scores on Isekai (%).

Method	J	F	J&F
CoT-RVS, causal adaptation	46.43	47.90	47.16
Sekai	**49.09**	**51.58**	**50.33**

the same video-equal aggregation after filtering; Other pools the compound families, excluding Now/Ever/Latest/Never before aggregation. Union-mask scoring does not measure instance separation.

Filled rectangles, GrabCut, and SEEM generate a mask at each locator anchor and hold it until the next anchor. GrabCut falls back to a filled rectangle on an OpenCV error; the receipts do not count these internal fallbacks separately. SEEM uses the official FocalT-v0 weights and `transformer_encoder_fpn` configuration. Each locator box is rasterized into the official spatial-stroke prompt interface. SAM2.1 Hiera-Large instead propagates within arrived segments; only the audited causal implementation is included. Thus the comparison changes propagation as well as the anchor-frame segmentor.

The SAM 3.1 oracle receives ground-truth boxes at the same causal refresh schedule and propagates through arriving frames; it does not read ground-truth masks during inference. Its 87.80 score measures performance with privileged target prompts.

C Baseline Adaptations and Causal Audit

We adapt each baseline to the same arrived-frame schedule before comparing it on Isekai. The audit records the frames, features, caches, and prompts available at every call and rejects any future-derived input.

C.1 OnlineRefer and CoT-RVS

OnlineRefer (Wu et al., 2023) receives the fixed query and newly arrived frames while retaining its target representation. CoT-RVS (Kao et al., 2026) receives a causal prefix grid at each refresh and uses its native reasoning and segmentation interface. The online branch follows a single object of interest and can switch targets. OnlineRefer uses the official Swin-L checkpoint trained on Refer-YouTube-VOS; our CoT-RVS adaptation uses Qwen3-VL-8B-Instruct and SAM 3.1.

C.2 StreamingRVOS and SAMWISE

StreamingRVOS (Zhang et al., 2026b) carries semantic tokens between refreshes and is evaluated with its causal streaming schedule. SAMWISE (Cuttano et al., 2025) receives only its recent visual context and segmentation memory. We keep their native mask decoders. StreamingRVOS uses the official 4B checkpoint and SAMWISE the official MeViS HSA+CME checkpoint; both include MeViS training.

C.3 Controls

Sekai w/o memory removes the EMR, so the locator receives only the query and current chunk. GT boxes + SAM 3.1 supplies reference boxes at the endpoint and measures mask generation after target selection.

C.4 Comparison with a Shared Backbone and Segmentor

Our causal adaptation of CoT-RVS (Kao et al., 2026) and Sekai both use Qwen3-VL-8B-Instruct (Bai et al., 2025) and SAM 3.1 (Carion et al., 2025). They follow the same arrival schedule and are evaluated on the same 12,569 scoring points. CoT-RVS reads a causal-prefix grid and supplies text prompts to the segmentor. Sekai reads incremental memory with an adapted locator and supplies box prompts. This comparison holds the backbone family and segmentor fixed; locator adaptation and prompting differ.

Table 10: J&F by evidence age. Counts are scoring points.

Method	In window	1–8	9–24	25+
Points	972	1,211	2,522	2,485
OnlineRefer	58.08	53.47	50.37	49.99
SAMWISE	61.38	59.35	57.95	56.78
CoT-RVS	26.64	27.36	24.23	17.19
StreamingRVOS	74.61	72.79	69.27	68.19
Sekai w/o memory	67.11	66.99	67.86	60.14
Sekai	65.96	67.81	67.67	64.42
GT boxes + SAM 3.1	41.43	91.88	90.26	92.51

Table 11: **J&F by target count on Isekai.** Counts refer to visible reference instances. Best per column in bold, excluding the privileged GT-box reference.

Method	0 targets	1 target	≥ 2 targets
Points	5,379	6,162	1,028
OnlineRefer (Wu et al., 2023)	4.60	53.32	34.82
SAMWISE (Cuttano et al., 2025)	4.92	59.13	44.66
CoT-RVS (Kao et al., 2026)	**75.76**	21.45	30.07
StreamingRVOS (Zhang et al., 2026b)	5.63	**70.74**	**50.91**
Sekai w/o memory	17.14	65.39	49.44
Sekai	28.25	65.22	49.60
GT boxes + SAM 3.1	93.38	80.17	89.11

D Supplementary Diagnostic Records

D.1 Evidence-age summaries

Table 10 gives the values plotted in Figure 5 and the GT-box reference. Each stratum uses the same nonempty reference points for every method. Scores are averaged within videos and then across videos represented in that stratum.

D.2 Target-count summaries

Table 11 groups scoring points by the number of distinct visible reference instances. We filter to each group, average scores within each video, and then average over videos represented in that group, using the same saved predictions as Table 3.

D.3 Per-Family Comparison with the Memory-Free Control

Table 3 gives the absolute scores; here we list the differences that Section 5.2 summarizes. Relative to Sekai w/o memory, Sekai changes J&F by $+7.44$ on Now, $+4.58$ on Ever, -5.97 on Latest, -10.51 on Never, $+11.99$ on the pooled OTHER, and $+2.04$ on the history subset. Relative to the strongest baseline in each column, Sekai is $+9.28$ on Ever (StreamingRVOS), -17.27 on Now and -5.21 on Other (CoT-RVS), and -2.50 on Latest, -11.17 on Never, and -2.74 on the history subset (StreamingRVOS). The two families where the memory hurts share one failure mode: a stale or spurious entry that cannot be corrected once its frames have left the window. Latest requires the newest event to supersede an older participant, and Never excludes an instance permanently once any entry records the event, so a single false positive in the memory costs every later frame.

Uncertainty of the overall comparison. We estimate uncertainty by paired bootstrap over the 60 videos. Each replicate samples 60 videos with replacement and uses the same sampled videos for both methods, retaining all queries and scoring points within each video. With 100,000 replicates , Sekai improves over the adapted CoT-RVS baseline by 3.163 J&F points, with a 95% percentile confidence interval of $[0.05, 6.27]$.

Table 12: **Results on MeViS v2.** J&F (%). Reference results are as reported by Ding et al. (2025).

Method / Configuration	Online	J&F
LMPM (Ding et al., 2023)	✗	38.30
DsHmp (He & Ding, 2024)	✗	40.80
LMPM++ (Ding et al., 2025)	✗	43.90
VISA (Yan et al., 2024) (MeViS v2 evaluation)	✗	40.70
Sekai (ours)	✓	50.45

D.4 HISTORY-SUBSET COVERAGE

The 4,133 replayable history points comprise 866 NOW, 2,401 EVER, and 866 NEVER points. We first replay each full-history query and check equality with its visible reference set, then retain only events first confirmed within the current chunk and compare the two sets. Appendix A.5 defines the criterion. No LATEST or compound point is assigned to this subset without a replayable query rule.

D.5 MEVIS V2 RESULTS

Table 12 reports the MeViS v2 results. Sekai uses the official complete validation-set evaluation and the locator configuration described in Appendix B.3.