SAME FRAMES, EXPLICIT REFERENTS: IDENTITY-LINKED EVIDENCE BUNDLES FOR LONG-VIDEO QA

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video QA can retrieve every decisive frame yet still lose the identity relation that makes those frames jointly interpretable. To measure this post-retrieval failure directly, we introduce an exact-visual-parity test in which paired VLM inputs share every frame and crop byte, timestamp, item identifier, position, prompt, and visual order; only the answer-time organization changes from window-local records to a cross-window referent bundle. ReferentWeave realizes the tested bundle by composing typed temporal patches into persistent identity chains with relation and provenance fields, deterministic serialization, and uncertainty-triggered evidence expansion. On 3,842 Identity VideoQA queries, this intervention raises exact-match accuracy from 56.1% to 59.4% (+3.3 points; 95% CI [1.4, 4.9], video-cluster paired p=0.019) while serving the same 151 images per query. The exact-parity effect remains +2.5 points with InternVL2.5 and +2.3 under human-linked identities; when selection and decoding are restored under a common ceiling, ReferentWeave reaches 59.7% accuracy, +5.6 over Track-RAG and +8.3 over Dense-frame, while reducing identity swap from 7.8% to 6.0% and median served evidence from 186 to 148 images. Evidence access is therefore not a complete control for long-video QA: the referential structure attached to that evidence is an independently testable part of the model input.

1 INTRODUCTION

Consider a blue cup seen on a table, absent for several minutes, and later surrounded by similar cups beside a sink. The frames needed to answer "was the first cup used again?" can be present in the VLM input in two semantically different forms. A window-local input presents two timestamped cup records; a referent-linked input states that the records belong to one continuing object and gathers their evidence pointers under that object. The pixels are identical, but the proposition made legible to the model is not.

This counterfactual isolates a failure that retrieval recall cannot express. Long-video systems increasingly improve decoding, multimodal search, streaming memory, and graph construction (Ren et al., 2025; Cao et al., 2025; Li & Liu, 2026; Arnab et al., 2025; Di et al., 2025; Zhang et al., 2025; Yeo et al., 2026; Rege et al., 2026); entity-centric memories additionally preserve objects across time (Fan et al., 2026; Zhao et al., 2026; Ma et al., 2026b;a). End-to-end scores entangle those choices with candidate access, selection, and presentation. They therefore cannot reveal whether a model fails because the decisive observation is missing or because observations of the same entity arrive as unrelated local records.

We make this distinction measurable with an exact-visual-parity intervention. A condition-blind ranker first selects one canonical sequence of frames and crops. Both paired conditions receive the same bytes, timestamps, item identifiers, positions, prompt, and visual order; one serializer preserves window-local groups, while the other emits a cross-window referent bundle. Figure 1 depicts this controlled input transformation. ReferentWeave is the instrumented system that realizes the bundle with persistent identities, relation and provenance state, typed temporal patches, and continuity-aware evidence expansion.

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.

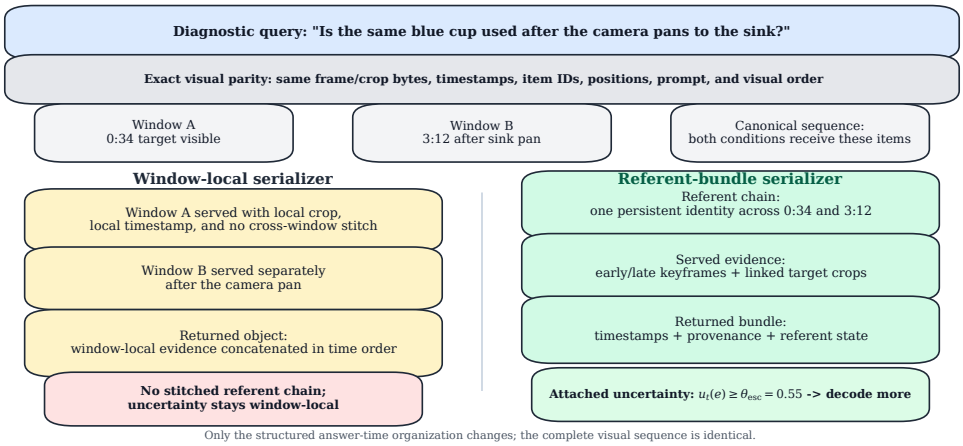


Figure 1: Exact visual parity, different referential organization. A condition-blind selector fixes every served image, crop, timestamp, identifier, and position. The paired serializers either preserve two window-local records or bind them into one referent bundle with relation and provenance state.

The visual-parity test is the primary hypothesis. Across 3,842 queries, changing only the answer-time bundle raises exact-match accuracy from 56.1% to 59.4% (+3.3; 95% CI [1.4, 4.9]). The effect remains positive with InternVL2.5 and independently human-linked identities. Restoring evidence selection from one pool yields a +5.1-point system difference; restoring query-time decoding under a common ceiling yields +5.6 points over Track-RAG while serving 38 fewer images per query. Public retrieval and memory policies, a rebuilt perception stack, Video-MME, and cross-camera chains test whether the system-level ordering survives changes in policy, ontology, footage, and viewpoint.

Contributions.

- **Exact-parity diagnostic.** The paired test fixes the complete visual sequence and estimates the joint post-selection effect of cross-window identity links, relation/provenance state, grouping, and deterministic serialization.
- **Referent-bundle construction.** ReferentWeave composes typed temporal patches into deterministic query-specific bundles and expands evidence according to calibrated continuity uncertainty.
- **An evidence ladder.** Exact-item, shared-pool, and common-decode-ceiling protocols separate the bundle-level effect from selection and end-to-end system behavior, followed by model, annotation, stack, domain, public-policy, and camera-topology checks.
- **Mechanism diagnostics.** Component removals and continuity-burden slices associate the gains with persistent identity, composition, and targeted allocation rather than visual volume alone.

The evaluated regime is identity-centric QA over bounded VLM inputs, where observations recur across temporal windows and the answer depends on their continuity. In this regime, referential organization is a measurable input variable rather than an incidental by-product of retrieval.

2 Related Work

Long-video access and memory. Long-video QA benchmarks (Mangalam et al., 2023; Yu et al., 2019; Xiao et al., 2021; Sadhu et al., 2021; Patraucean et al., 2023; Fu et al., 2025) and video-language backbones (Luo et al., 2026; Zhang et al., 2023; Bertasius et al., 2021; Arnab et al., 2021; Tong et al., 2022) test reasoning beyond short clips. Retrieval and streaming systems extend usable context through multimodal search, temporal weighting, frame reasoning, KV retrieval, or synopsis/detail memories (Ren et al., 2025; Cao et al., 2025; Arnab et al., 2025; Di et al., 2025; Zhang et al., 2025).

WorldMM combines episodic, semantic, and visual memories, while VAM couples online indexing with hierarchical visual memory and agentic retrieval (Yeo et al., 2026; Li et al., 2026). These methods optimize which content is retained or retrieved. Our question begins after that choice: what contribution remains when candidate access, and ultimately the visual items themselves, are fixed?

Entity-aware video memory. Recent systems make entities central to long-horizon reasoning. EGAgent queries a temporally annotated entity scene graph; FOLIO maintains entity-centered semantic records linked to visual evidence; ViSAGE uses cross-modal entity binding and retrospective memory refinement; ReflectWorld-MM organizes open-ended streams around resolved entities; and R4DSG represents persistent objects through anchor-relative state changes (Rege et al., 2026; Fan et al., 2026; Zhao et al., 2026; Ma et al., 2026b;a). These advances are closest in representational spirit to ReferentWeave. Their reported end-to-end comparisons vary memory construction, retrieval, and presentation together; our diagnostic instead turns answer-time referential organization into a controlled input variable. ReferentWeave supplies a concrete realization whose same-item intervention measures the joint bundle effect without changing visual inputs.

Entity alignment versus persistent visual reference. FastV-RAG addresses entity recognition in retrieval-augmented video QA through similarity-based filtering; it is not presented as a cross-window visual identity partition or a serializer for one continuing referent (Li & Liu, 2026). Tracking, re-identification, and video-object memories provide the upstream ingredients for such a partition (Liu et al., 2024; Zhou et al., 2022; Cheng et al., 2022; Ravi et al., 2025; Zhang et al., 2022; Cao et al., 2023; He et al., 2021; Cheng et al., 2021; Cheng & Schwing, 2022). Our contribution is to make their answer-time use both operational and experimentally isolatable.

Table 1: Closest design families. "Persistent visual ID" means an identity chain over nonadjacent observations, not answer–frame similarity. "Same-item test" means that every served visual item and its position are identical across paired conditions.

System family	Answer-time organization	Persistent visual ID	Same-item test
VideoRAG / TV-RAG / Flash-VStream	Retrieved items or streaming memories	no	no
FastV-RAG	Candidate-answer entity verification	no	no
WorldMM / VAM	Multiscale multimodal memory	not explicit	no
EGAgent / FOLIO	Entity graph or semantic records	entity-level	no
ViSAGE / ReflectWorld-MM / R4DSG	Identity- or object-centered memory	yes	no
ReferentWeave	Serialized query-specific referent bundle	yes	yes

3 METHOD

For query q, let A_q denote the accessible candidate set. We call the explicit answer-time evidence interface a *referent contract*:

$$\mathcal{E}_q = (M_q, \Pi_q, R_q, \sigma_q),$$

where $M_q \subseteq A_q$ is the served visual-item multiset, Π_q contains identity partitions and cross-window links, R_q stores relation and provenance state, and σ_q is the deterministic serialization. Matched information fixes A_q; the exact same-item control also fixes M_q and its order. ReferentWeave intervenes on Π_q, R_q, and σ_q by composing observations before the image list and structured text are emitted.

Typed state and bundle generation. ReferentWeave constructs the interface from a deterministic typed store

$$G = (V_{\text{obj}} \cup V_{\text{rel}} \cup V_{\text{evt}},\ E_{\text{temp}} \cup E_{\text{rel}} \cup E_{\text{ptr}}),$$

The upstream tracker assigns a `local_track_id` within each window. ReferentWeave never treats this label as globally persistent: an observation key is the pair (window, `local_track_id`), while a composed object node receives a `referent_id` after accepted cross-window links are

resolved. The answer-time serializer exposes that persistent key as `track_id` for the VLM and retains every contributing local key in provenance. Each object node records its confidence, last-seen time, boxes/masks, and attributes; each relation or event node records evidence pointers, time ranges, and predicates. Object attributes are copied from the frozen perception outputs and OCR rather than generated at answer time. Spatial predicates are deterministic functions of boxes and masks; temporal predicates and event boundaries are changes in tracked state over adjacent observations. Every field carries the producing module, confidence, timestamp range, and frame/crop pointer, so provenance resolves to a served visual item. The serializer admits only this finite typed schema; it does not insert free-form summaries. Representation determines the fields exposed by $\mathcal{E}_q$; policy chooses M_q, allocates serializer budget, and invokes additional decoding. Supplementary Table S6 gives the grammar, ordering, and field-level truncation rule.

Patches and retrieval. Scene boundaries define contiguous windows, clamped to 5–20 s; adjacent windows share their boundary keyframe rather than using an independently tuned sliding stride. Each window becomes a patch $P_w = (G_w, K_w, \mathbf{p}_w, u_w)$. G_w records identity and relation state, K_w provides reusable visual pointers, $\mathbf{p}_w$ supports query retrieval, and u_w drives uncertainty-aware allocation. K_w contains 4–12 keyframes selected greedily for track/event coverage; a frame covers a track at detection confidence ≥ 0.6 and visible mask area $\geq 40\%$. We mean-pool CLIP ViT-L/14 embeddings over keyframe crops for $\mathbf{p}_w$, following the contrastive matching family of CLIP, ALIGN, and SigLIP (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023). The uncertainty state stores observation gaps, boundary stitch confidence, occlusion indicators, and relation/event entropy.

Given a query q, we retrieve the top $K{=}8$ patches by cosine similarity between its CLIP text embedding and $\mathbf{p}_w$, breaking equal scores by earlier timestamp. Retrieved patches are ordered by time and aligned through their local observation keys. Composition collects keyframes, crops, timestamps, provenance, uncertainty, and relations under each accepted `referent_id`. Track-RAG instead preserves the local keys and concatenates independently selected window objects. The exact same-item control tests the complete pre-answer bundle as one intervention: visual inputs remain identical while identity partitions, relation/provenance organization, grouping, and deterministic text fields change together.

Cross-window alignment. Cross-window track alignment uses

$$s(i,j) = \beta_1 \cdot \mathrm{IoU}(i,j) + \beta_2 \cdot \cos(\mathbf{e}_i, \mathbf{e}_j) + \beta_3 \cdot m(i,j), \tag{1}$$

using default weights $\beta_1{=}0.3$, $\beta_2{=}0.5$, $\beta_3{=}0.2$. Boundary overlap is denoted by IoU; the CLIP re-ID similarity, after rescaling to $[0,1]$, is $\cos(\mathbf{e}_i, \mathbf{e}_j)$; motion consistency $m(i,j) \in [0,1]$ results from applying $\exp(-d^2/2)$ to the normalized Kalman innovation. The re-ID term follows the same broad intuition as DeepSORT-, TransReID-, and re-ranking-based association pipelines Wojke et al. (2017); He et al. (2021); Zhong et al. (2017). Candidate pairs must share a detector category and occur in temporally ordered patches. Adjacent patches use all three terms; for a nonadjacent reappearance, boundary IoU is zero and the score is carried by appearance and motion consistency. We solve one-to-one assignments with the Hungarian algorithm and reject scores below 0.40. Accepted links are processed in temporal order. A merge is rejected if it would place two observations from the same window in one referent, give an observation competing predecessors or successors, or join chains with conflicting accepted assignments; rejected and unmatched observations remain singleton referents. These rules make chain formation deterministic under score and timestamp ties. All reported results use this procedure.

Continuity uncertainty and escalation. For each decision-critical entity e, the query-time escalation score is

$$u_t(e) = \alpha_1 \cdot \Delta_{\mathrm{time}}(e) + \alpha_2 \cdot (1 - s_{\mathrm{stitch}}(e)) + \alpha_3 \cdot p_{\mathrm{occ}}(e) + \alpha_4 \cdot H_{\mathrm{rel}}(e), \tag{2}$$

with defaults $\alpha_1{=}0.30$, $\alpha_2{=}0.35$, $\alpha_3{=}0.20$, $\alpha_4{=}0.15$. $\Delta_{\mathrm{time}} = \min((t - t_{\mathrm{last}})/180\mathrm{s}, 1)$; s_{stitch} is the accepted alignment score above; and p_{occ} is the fraction of the latest track interval below the 40% visibility criterion. For C active predicates, H_{rel} is normalized predicate entropy when $C \geq 2$ and is defined as zero for $C \in \{0,1\}$. Thus every term and $u_t(e)$ lie in $[0,1]$. A referent is decision-critical when a query noun phrase retrieves one of its crops or typed attributes among the selected patches; ties follow patch score and timestamp order. On the source-disjoint validation partition, each query is answered once from the unexpanded top-eight bundle and once after dense fallback up to $B_{\max}$. The

query-level calibration target is one exactly when the unexpanded answer is wrong and the fallback answer is correct; reliability bins the maximum $u_t(e)$ over decision-critical entities against this target. Accordingly, θ_{full} controls full fallback, whereas the lower θ_{esc} is a validation-fixed staging point for entity-only expansion. Features, weights, and thresholds are frozen before final evaluation. Reliability and shift results appear in the supplement; the construction complements established calibration and uncertainty estimators (Guo et al., 2017; Lakshminarayanan et al., 2017). If $u_t(e) \geq \theta_{\text{esc}}$=0.55, we decode additional raw frames for that entity only; if $u_t(e) \geq \theta_{\text{full}}$=0.78, we fall back to dense retrieval up to $B_{\max}$.

4 Experimental Setup

The evaluation answers three questions in order: **Q1**, does the referent bundle change answers when every visual input is identical; **Q2**, what accuracy, identity stability, and evidence cost does the complete system achieve when selection and decoding are restored; and **Q3**, which components and continuity regimes are associated with the observed gains, and does the ordering persist under changed models, policies, annotations, stacks, domains, and cameras? All primary experiments use Qwen2.5-VL-32B-Instruct in bfloat16 on one 4×NVIDIA A100 80 GB node (CUDA 12.4, driver 535.129.03, PyTorch 2.3.1, FlashAttention-2). The frozen perception stack is GroundingDINO-T, SAM2-Large, ByteTrack, and CLIP ViT-L/14; PaddleOCR v2.7 is enabled when text is present. The independent rebuild uses Detic, Mask2Former, OCSORT, and TransReID. Seeds $\{13, 37, 61, 71, 97\}$ govern stochastic VLM decoding and deterministic tie streams for keyframe and retrieval ranking; each paired comparison reuses the same seed on both methods, while suite membership and cached perception outputs remain fixed.

Three control regimes. The matched-budget protocol fixes query timestamps, prompts, perception outputs, and the raw-frame decode ceiling $B_{\max}$=512 across all methods. For each query, matched information provides all methods with one pre-decoded pool comprising 2 fps samples from the ±180 s band centered on its timestamp and scene-change keyframes within that band. Evidence selection and formatting are restricted to this common pool. For the exact same-item protocol, a condition-blind canonical ranker scores every shared-pool window by CLIP query–patch cosine, takes the top eight with timestamp tie-breaking, collects its frames, scene keyframes, and associated crops, deduplicates by item ID, and orders items by timestamp, item type, and ID. This rule uses no cross-window links, relation state, or output from either presentation condition. The harness copies the resulting frame and crop bytes, timestamps, item identifiers, and canonical order to both conditions and verifies their checksum; query-time escalation cannot add, remove, or reorder an item. The window-local and referent-linked serializers then populate their respective structured fields under the same T_{ser}=768 ceiling.

Evidence and runtime accounting. Each raster image delivered to the VLM contributes one to Fr/q, whether it is an object crop, a decoded full frame, or a cached keyframe. Only query-time decoding of raw frames contributes to $B_{\max}$. Every reported wall-clock latency is measured on the 4×A100 node at batch size 1, with each process handling one active query. For every query, cached crops and keyframes are served again, requiring the VLM image encoder to encode them again; neither visual tokens nor image-encoder activations are cached, and KV caching does not extend across queries. Measured end-to-end latency and conversion to raw-decode units are reported separately in supplementary Tables S10 and S16.

Evaluation population and scoring. The primary exact-parity comparison applies one frozen perception output and one canonical visual sequence to both presentation conditions. Identity VideoQA contains 3,842 queries over 1,200 videos drawn from the official EgoSchema and ActivityNet-QA sources (Mangalam et al., 2023; Grauman et al., 2022; Yu et al., 2019). "EgoSchema-Long v1.2" and "ActivityNet-QA-Long v2.0" name the frozen conversion specifications used in this study rather than source benchmark titles: source evaluation videos are partitioned by video, the long clip is retained, and no final-evaluation video is used for calibration or hyperparameter selection. The supplement retains two secondary suites: Embodied refs, with 1,416 queries over 600 videos, and Occlusion stress, with 1,064 queries over 400 videos. We use four main metrics: citation IDF1, Fr/q, swap rate, and exact-match QA accuracy. Multiple-choice sources are scored by option identifier. Free-form answers are Unicode-normalized, lowercased, stripped of punctuation and articles, and whitespace-collapsed

before exact match; no semantic judge is used. Citation IDF1 treats each parsed timestamp–identity pair as an identity observation, matches it to the annotated identity active at that timestamp, and computes $2\,\mathrm{IDTP}/(2\,\mathrm{IDTP}+\mathrm{IDFP}+\mathrm{IDFN})$. An unparseable answer is incorrect; a missing or unparseable citation list contributes no IDTP and counts as a swap for identity-bearing queries. A query enters Identity VideoQA only when source annotations identify an answer-critical chain in at least two temporal segments and a category-matched distractor carries a different source-annotated answer attribute or action, so substituting that chain changes the normalized answer label. This eligibility test is applied before any method output; ambiguous target mappings, one-segment references, and unchanged labels are excluded. Supplementary Table S12 specifies timestamp assignment and the identity source used for membership, system input, and scoring in each evaluation. Suite membership is fixed before scoring and independent of all method outputs. The human-linked audit replaces membership and scoring chains while retaining tracker-derived system inputs; the independent rebuild replaces membership, system inputs, and scoring mappings together (supplementary Tables S27 and S30).

Comparators. The main end-to-end table compares Track-RAG, Dense-frame, Keyframe-only, and BM25 summary + keyframes. For BM25 summary + keyframes, BM25 searches text summaries to retrieve 60 s windows; the VLM still receives scene-change keyframes from those windows, so served images determine Fr/q, not the summary tokens. Our primary in-house comparator, Track-RAG, retains retrieval and serialization within individual windows but uses the same tracker outputs, serializer budget $T_{\mathrm{ser}}{=}768$, truncation policy, crop accounting, cache footprint, and retrieval embeddings as ReferentWeave. Under matched budget, their retrieval and escalation policies remain part of each end-to-end system; the shared-pool and same-item protocols provide the tighter access and representation controls. Public adaptations preserve VideoRAG's multimodal retrieval, TV-RAG's temporal re-ranking and entropy-weighted frame choice, and Flash-VStream's synopsis/detail memory (Ren et al., 2025; Cao et al., 2025; Zhang et al., 2025). We replace only the backbone, prompt, frozen perception inputs, and accounting interface required for parity; author defaults are selected once on the shared validation partition and never tuned on final queries. Under matched information, Dec/q is zero because these policies read the already decoded pool. Supplementary Tables S26 and S31 give the policy-level mapping; Cache-only, Long-video VLM, and Learned Controller add breadth in Tables S20 and S21.

Inference. For each seed we retain paired query outcomes. Reported accuracy means and standard deviations summarize the five paired seed runs. Confidence intervals resample videos, carrying all of their queries and seed outcomes together, for 10,000 cluster-bootstrap replicates. Two-sided permutation tests flip the sign of each video's mean paired difference; query-level wins/ties/losses are descriptive and are not the permutation units. The single confirmatory hypothesis is the exact-same-item accuracy contrast between the window-local and referent-bundle serializers. All end-to-end, comparator-transfer, component, and slice tests are secondary; their displayed p-values are raw, two-sided cluster-paired values.

5 Results

5.1 Q1: Exact visual parity and the referent-bundle effect

The confirmatory comparison begins after selection. On every query, the condition-blind ranker supplies the same 151-image median sequence to both serializers, and all 3,842 pairs pass the exact byte-and-order checksum. Window-local organization reaches 56.1% accuracy; the referent bundle reaches 59.4%, a +3.3-point difference (95% CI [1.4, 4.9], $p{=}0.019$), while swap falls from 7.0% to 6.2% and IDF1 rises from 70.0 to 72.3. This estimate belongs to the complete structured bundle—identity partition, relation/provenance organization, grouping, and serialization—rather than to any field in isolation.

Allowing both systems to select from the same candidate pool increases the difference to +5.1 points and gives ReferentWeave 60.3% accuracy versus 55.2% for Track-RAG and 52.8% for Dense-frame. This shared-pool result combines selection with bundle construction, whereas the exact-item pair above isolates the joint post-selection bundle effect. Varying pool density and temporal extent retains positive system differences of 3.4–5.1 points (supplementary Table S17).

Table 2: Evidence ladder and changed-condition checks. "Comp./RW" gives absolute comparator and ReferentWeave accuracy. The exact-item row is the primary paired intervention; subsequent rows measure shared-pool system differences or transfer the complete system to another policy, identity source, stack, domain, or camera topology.

Setting	Comparator	n	Comp./RW Acc	ΔAcc	95% CI
Exact same items	Window-local	3842	56.1 / 59.4	+3.3	[1.4, 4.9]
Shared candidate pool	Track-RAG	3,842	55.2 / 60.3	+5.1	[2.7, 7.4]
Public matched-info	Flash-VStream	3,842	57.4 / 60.3	+2.9	[1.0, 4.8]
Human-linked labels	Track-RAG	612	53.7 / 56.8	+3.1	[0.8, 5.0]
Independent stack	Track-RAG	3517	53.8 / 58.1	+4.3	[1.9, 6.6]
Cross-camera chains	Flash-VStream	936	45.1 / 47.8	+2.7	[0.5, 4.9]
Human-linked Video-MME	Track-RAG	728	47.8 / 51.2	+3.4	[1.0, 5.6]

Table 3: Identity VideoQA under matched-budget $B_{\max}$=512. ReferentWeave improves by +5.6 points over Track-RAG and +8.3 points over Dense-frame while serving fewer evidence images.

Method	Acc (%)↑	Swap (%)↓	Fr/q med↓
BM25 summary + keyframes	45.8±1.7	13.2±1.5	418
Keyframe-only	50.1±1.4	10.4±1.1	257
Dense-frame	51.4±1.6	9.6±1.2	482
Track-RAG	54.1±1.6	7.8±1.0	186
ReferentWeave	59.7±1.8	6.0±0.9	148

5.2 Q2: End-to-end accuracy and evidence cost

Restoring each method's complete selection and decoding policy raises ReferentWeave accuracy from 54.1% to 59.7% relative to Track-RAG, lowers swap from 7.8% to 6.0%, and reduces the median evidence load from 186 to 148 images. Against Dense-frame, the +8.3-point gain accompanies a 69% reduction in served images (482 to 148). IDF1 rises from 68.2 to 72.4; the secondary raw cluster-paired p-values are 0.003 against Dense-frame and 0.011 against Track-RAG, with 95% intervals [5.8, 10.6] and [5.1, 6.0] (supplementary Tables S20 and S23). These end-to-end contrasts quantify system behavior under a common decode ceiling; they do not replace the exact-item estimand.

5.3 Q3: Public policies and changed ontologies

The remaining controls change one source of external validity at a time. VideoRAG, TV-RAG, and Flash-VStream replace the retrieval or streaming policy; human links replace membership and scoring chains; the Detic/Mask2Former/OCSORT/TransReID rebuild replaces the complete perception ontology; Video-MME changes the footage domain; and synchronized cameras change the viewpoint topology. InternVL2.5-38B-Instruct also preserves the ranking at 57.1% accuracy (+4.8 over Track-RAG; Table S18), while its exact same-item extension retains a 2.5-point gain (Table S33).

Table 2 reports positive margins from +2.7 to +5.1 points as comparator family, annotation source, stack, domain, and camera topology change. The strongest public shared-pool comparator, Flash-VStream, reaches 57.4%; ReferentWeave reaches 60.3% while serving 151 rather than 158 median images (Table S31). This comparison extends the finding beyond the window-local in-house null.

5.4 Continuity burden and execution cost

Continuity slices exhibit a diagnostic rise-and-compress profile: the advantage is largest for 60–180 s gaps (+6.0), 4–12 s occlusions (+6.3), and moderate ambiguity (+6.0), then narrows when observations become too sparse to link reliably (supplementary Fig. S5 and Table S34). In separate matched-budget, single-removal ablations, omitting persistent IDs, re-ID, uncertainty allocation, referent composition, and graph-backed patches is associated with accuracy drops of 6.8, 4.7, 3.9, 3.1, and 2.0 points, respectively (Table S22); these effects are not an additive decomposition of the

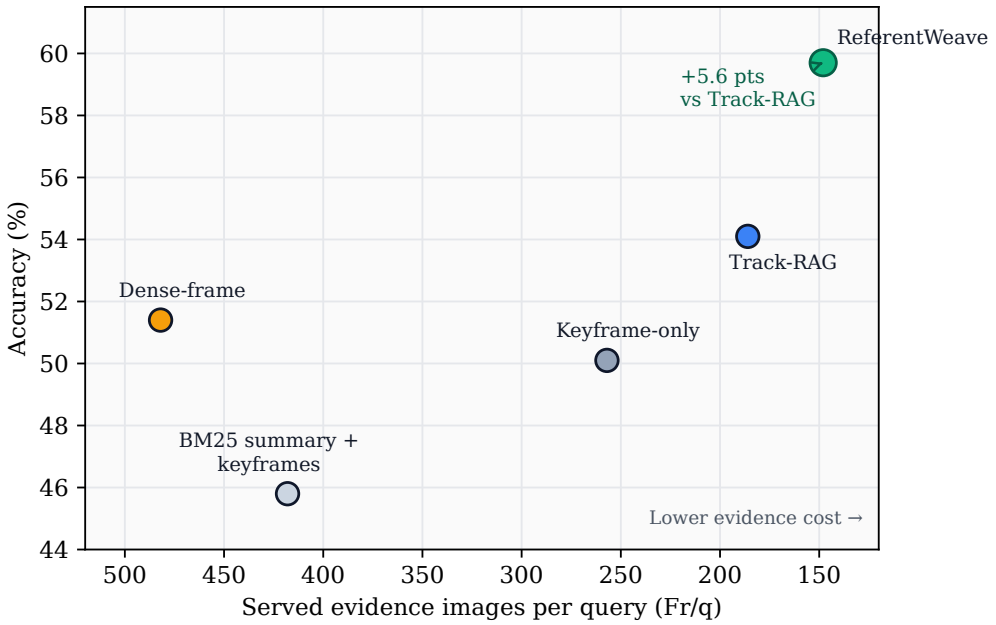


Figure 2: Observed accuracy and median served evidence under the common $B_{\max}$=512 raw-decode ceiling. ReferentWeave improves both axes over Track-RAG. Fr/q is measured rather than equalized; Table 2 separates the exact-item, shared-pool, and changed-condition comparisons.

same-item margin. Under batch-1 execution, within-system query p50 spans 0.8–1.4 s and p95 spans 2.4–4.2 s across videos and query types. Escalation is invoked on 17.4% of queries; its stage timings are conditional on invocation, while cached images require no new decoding but are re-encoded by the VLM (Tables S10 and S11).

6 Discussion

The exact-item intervention establishes that visual evidence can be sufficient in content yet incomplete in reference. Its 3.3-point gain identifies the joint effect of replacing window-local records with the complete referent bundle while holding pixels, timestamps, item positions, prompt, and visual order fixed. This is the paper's causal bundle-level estimate. The shared-pool and matched-budget comparisons answer a different question—how the complete systems behave when selection and decoding are active—and the public-policy, rebuilt-stack, human-linked, domain, and cross-camera results test the stability of that system ordering under changed conditions.

The component ordering supplies mechanism-consistent diagnostics rather than a factorial decomposition of the primary effect. Identity persistence establishes the chain, re-identification bridges discontinuities, calibrated allocation acquires evidence where that chain is uncertain, and composition communicates the chain to the VLM. Several removals serve more images while losing accuracy, separating the result from visual volume alone. At the main operating point, 148 referent-consistent images outperform both 186-image window-local retrieval and 482-image dense exposure.

Failure traces remain actionable. Patch-boundary splits are detected through low target coverage and retrieval margins; near-duplicate re-ID confusion produces small association margins; occlusion-driven drift combines long observation gaps with elevated uncertainty. Adaptive overlap, conservative identity matching, and calibrated escalation address these signals directly. The descriptive slice profile is largest at moderate continuity burden, where evidence is present but must be joined, and compresses at the most ambiguous extremes, where usable identity observations become sparse.

Operating regime and responsible use. The evidence covers identity-centric questions with bounded VLM inputs, two perception stacks, two VLM backbones, tracker- and human-linked labels,

and single- and multi-camera video. Extreme occlusion and ambiguity reduce the available continuity evidence, while the exact-item intervention measures the bundle jointly rather than assigning its 3.3 points to an individual field. Persistent identity state also requires governance beyond storage accounting: deployments should use footage and annotations under valid consent and license terms, document human-linking instructions, audit identity errors across demographic and visual conditions, and prevent surveillance reuse outside the authorized task. Feature-only state, bounded retention, conservative matching, and access-controlled review of uncertain links provide concrete safeguards while preserving the evaluated use case.

7 Conclusion

Identical frames support different long-video answers when their cross-window reference is represented differently. ReferentWeave makes that reference explicit through a typed answer-time bundle, yielding a 3.3-point exact-parity effect and a 5.6-point end-to-end gain over Track-RAG while reducing median served evidence from 186 to 148 images. Positive margins under changed VLMs, identity labels, public policies, perception stacks, domains, and camera topologies connect the controlled finding to complete-system behavior. Long-video evaluation should therefore control and test not only which evidence a model receives, but also the referential structure through which that evidence becomes interpretable.

Reproducibility Statement

Sections 3 and 4 specify the representation, alignment equations, normalization, thresholds, models, hardware, prompts, budgets, cache policy, seeds, metrics, scoring, and clustered inference. The appendix provides suite rules, baseline specifications, bundle grammar, runtime accounting, uncertainty analyses, and paired denominators.

AI Use Statement

Generative AI tools assisted manuscript preparation, including language editing, formatting, and bibliographic verification. The authors reviewed the manuscript and take responsibility for its final content, including all claims and reported results.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *ICCV*, pp. 6836–6846, 2021. doi: 10.1109/ICCV48922.2021.00676. URL https://doi.org/10.1109/ICCV48922.2021.00676.

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d285ab51e167a923f07e6137d802305a-Abstract-Conference.html.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824. PMLR, 2021. URL https://proceedings.mlr.press/v139/bertasius21a.html.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *CVPR*, pp. 9686–9696, 2023. doi: 10.1109/CVPR52729.2023.00934. URL https://doi.org/10.1109/CVPR52729.2023.00934.

Zongsheng Cao, Yangfan He, Anran Liu, Jun Xie, Feng Chen, and Zhepeng Wang. TV-RAG: A temporal-aware and semantic entropy-weighted framework for long video retrieval and understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9071–9079. ACM, 2025. doi: 10.1145/3746027.3755873. URL https://doi.org/10.1145/3746027.3755873.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pp. 1290–1299, 2022. doi: 10.1109/CVPR52688.2022.00135. URL https://doi.org/10.1109/CVPR52688.2022.00135.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson–shiffrin memory model. In *Computer Vision – ECCV 2022*, volume 13688 of *Lecture Notes in Computer Science*, pp. 640–658. Springer, 2022. doi: 10.1007/978-3-031-19815-1_37. URL https://doi.org/10.1007/978-3-031-19815-1_37.

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*, volume 34, pp. 11781–11794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/61b4a64be663682e8cb037d9719ad8cd-Abstract.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://openreview.net/forum?id=8g9fs6mdEG.

Haoyang Fan, Dhruv Parikh, Anvitha Ramachandran, Sameh Gobriel, Nilesh Jain, Rajgopal Kannan, and Viktor Prasanna. FOLIO: Focused semantic memory for streaming video understanding. *arXiv preprint arXiv:2607.13298*, 2026. URL https://arxiv.org/abs/2607.13298.

Chaoyou Fu et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *CVPR*, pp. 24108–24118, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://doi.org/10.1109/CVPR52734.2025.02245.

Kristen Grauman et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, pp. 18995–19012, 2022. doi: 10.1109/CVPR52688.2022.01842. URL https://doi.org/10.1109/CVPR52688.2022.01842.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based object re-identification. In *ICCV*, pp. 15013–15022, 2021. doi: 10.1109/ICCV48922.2021.01474. URL https://doi.org/10.1109/ICCV48922.2021.01474.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Aiden Yiliu Li, Nels Numan, and Anthony Steed. Visual agentic memory: Enabling online long video understanding via online indexing, hierarchical memory, and agentic retrieval. *arXiv preprint arXiv:2605.16481*, 2026. URL https://arxiv.org/abs/2605.16481.

Gen Li and Peiyu Liu. FastV-RAG: Towards fast and fine-grained video QA with retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 879–889, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.38. URL https://aclanthology.org/2026.acl-long.38/.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, volume 15105 of *Lecture Notes in Computer Science*, pp. 38–55. Springer, 2024. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3.

Ruipu Luo, Ziwang Zhao, Min Yang, Zheming Yang, Minghui Qiu, Zhongyu Wei, Yanhao Wang, and Cen Chen. Valley: Video assistant with large language model enhanced ability. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 22(7):1–25, 2026. doi: 10.1145/3796716. URL https://doi.org/10.1145/3796716.

Ke Ma, Yamin Mao, Weiming Li, Shuai Tan, Yijie Zhong, Hao Chen, Haofen Wang, and Meng Wang. R4DSG: Relative 4d scene graph memory for object-centric question answering in long egocentric video. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026a. doi: 10.1145/3767308.3835995. URL https://arxiv.org/abs/2608.11017. In press.

Xiaokang Ma, Yifan Sun, Zhihong Jin, Jie Gu, Yudong Luo, Shenyi Shao, Chu Tang, Jingmin Chen, and Li Pu. ReflectWorld-MM: An entity-oriented multimodal memory system for open-ended video streams. *arXiv preprint arXiv:2607.09759*, 2026b. URL https://arxiv.org/abs/2607.09759.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244, 2023. doi: 10.52202/075280-2004. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

Viorica Patraucean et al. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42748–42761, 2023. doi: 10.52202/075280-1852. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8540fba4abdc7f9f7a7b1cc6cd60e409-Abstract.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nikhila Ravi et al. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46575–46602, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.2161. URL https://aclanthology.org/2026.acl-long.2161/.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. VideoRAG: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025. URL https://arxiv.org/abs/2502.01549.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *CVPR*, pp. 5589–5600, 2021. doi: 10.1109/CVPR46437.2021.00554. URL https://doi.org/10.1109/CVPR46437.2021.00554.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022. doi: 10.52202/068431-0732. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pp. 3645–3649. IEEE, 2017. doi: 10.1109/ICIP.2017.8296962. URL https://doi.org/10.1109/ICIP.2017.8296962.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, pp. 9777–9786, 2021. doi: 10.1109/CVPR46437.2021.00965. URL https://doi.org/10.1109/CVPR46437.2021.00965.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. WorldMM: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yeo_WorldMM_Dynamic_Multimodal_Memory_Agent_for_Long_Video_Reasoning_CVPR_2026_paper.html.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019. doi: 10.1609/aaai.v33i01.33019127. URL https://doi.org/10.1609/aaai.v33i01.33019127.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11975–11986, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://doi.org/10.1109/ICCV51070.2023.01100.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49/.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21059–21069, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Flash-VStream_Efficient_Real-Time_Understanding_for_Long_Video_Streams_ICCV_2025_paper.html.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *Computer Vision – ECCV 2022*, volume 13682 of *Lecture Notes in Computer Science*, pp. 1–21. Springer, 2022. doi: 10.1007/978-3-031-20047-2_1. URL https://doi.org/10.1007/978-3-031-20047-2_1.

Xinkui Zhao, Enbo Chen, Yifan Zhang, Chang Liu, Guanjie Cheng, Naibo Wang, and Yueshen Xu. ViSAGE: Constructing self-correcting memories for long-form video understanding. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026. doi: 10.1145/3767308.3835852. URL https://arxiv.org/abs/2607.28678. In press.

Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pp. 3652–3661, 2017. doi: 10.1109/CVPR.2017.389. URL https://doi.org/10.1109/CVPR.2017.389.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision – ECCV 2022*, volume 13669 of *Lecture Notes in Computer Science*, pp. 350–368. Springer, 2022. doi: 10.1007/978-3-031-20077-9_21. URL https://doi.org/10.1007/978-3-031-20077-9_21.

SUPPLEMENTARY MATERIAL

ROADMAP: WHERE TO FIND WHAT

The supplement first specifies calibration, accounting, runtime, suite construction, identity-source interventions, and comparator implementations. It then reports the exact-visual-parity intervention, public-policy comparisons, human-linked evaluations, the independent perception stack, cross-camera continuity, and burden-conditioned diagnostics. Tables S24 and S25 give the complete primary comparison.

S1 ESCALATION RELIABILITY AND ROBUSTNESS BREAKDOWN

This section reports escalation trigger calibration and per-condition robustness (Table S1). "Viewpoint / shot change" is a within-video re-linking stress slice under the primary stack; the cross-camera evaluation is reported separately in Table S32.

On Identity VideoQA, the combined time-gap, stitch-confidence, occlusion, and relation-entropy trigger achieves ECE = 0.06 and AUROC = 0.89 over five seeds. On the source-disjoint validation split, each query is answered from the unexpanded top-eight bundle and from dense fallback up to $B_{\max}$. The binary target is one only when the first answer is wrong and the fallback answer is correct; reliability bins the maximum entity score for that query against this target. The 0.78 threshold controls full fallback against this target. The lower 0.55 threshold is a validation-fixed staging point for entity-only expansion, not a second calibrated event. Features, weights, and both thresholds are then frozen. Track-RAG has no uncertainty trigger under matched budget, where allocation remains part of each end-to-end method; the exact same-item control prevents escalation from changing either paired item list.

Table S1: Per-condition robustness with a shared VLM, prompt, and $B_{\max}$=512; values are mean±std over 5 seeds at fixed query timestamps. Synthetic motion blur and 360p downsampling are applied to source frames before perception and VLM evidence construction, so the stack is rerun on each transformed input. Dense-scene and viewpoint/shot-change rows are naturally occurring subsets. ReferentWeave leads Track-RAG by at least 3.6 points across all conditions.

Condition	Dense-frame	Track-RAG	ReferentWeave	Δ vs Dense	Δ vs TR
Clean	51.4±1.6	54.1±1.6	59.7±1.8	+8.3	+5.6
Motion blur	49.8±1.8	52.9±1.8	57.4±2.0	+7.6	+4.5
Low-res (360p)	48.9±1.9	52.0±1.9	56.1±2.1	+7.2	+4.1
Dense (>15 ent.)	47.6±2.0	50.9±2.1	54.6±2.3	+7.0	+3.7
Viewpoint / shot change	46.9±2.1	50.2±2.2	53.8±2.2	+6.9	+3.6

Applying synthetic corruptions before detection and tracking prevents clean perception metadata from leaking into the robustness conditions. The smallest ReferentWeave-minus-Track-RAG margin is +3.6 points under naturally occurring viewpoint/shot-change stress.

S2 SENSITIVITY ANALYSIS DETAILS

This section reports sensitivity to uncertainty and stitch weights (Fig. S1).

The broad plateau shows that the accuracy advantage persists across a neighborhood of uncertainty and stitch settings. This stability supports one fixed operating point across suites and rules out a knife-edge tuning explanation.

S3 CALIBRATION DRIFT UNDER DOMAIN SHIFT

This section reports calibration under stress and lightweight recalibration (Fig. S2).

Figure S2 separates ranking quality from probability calibration. The uncertainty signal remains informative under shift, while lightweight recalibration restores the probability scale used for compute budgeting without changing the representation.

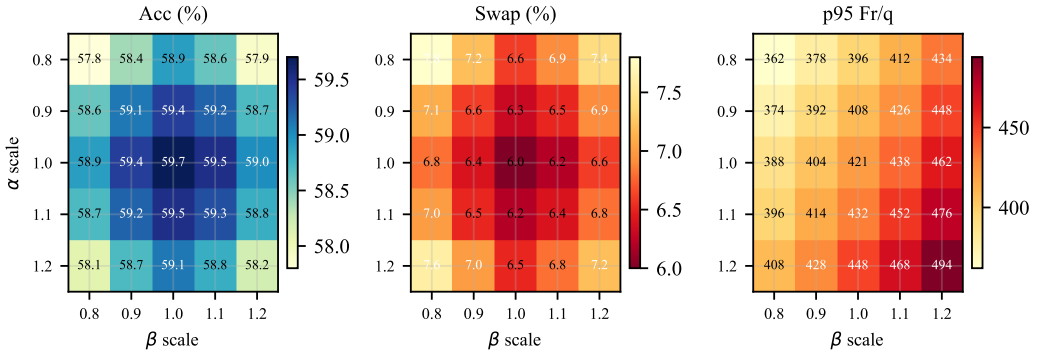


Figure S1: Sensitivity to multiplicative uncertainty (α) and stitch (β) weight scales under matched budget, with all other settings fixed. Accuracy and swap form a broad plateau around the default $(1.0, 1.0)$ and degrade smoothly toward the sweep boundaries. Higher α raises escalation and p95 Fr/q, exposing a controlled tail-compute trade-off. The plateau supports one stable operating point across suites.

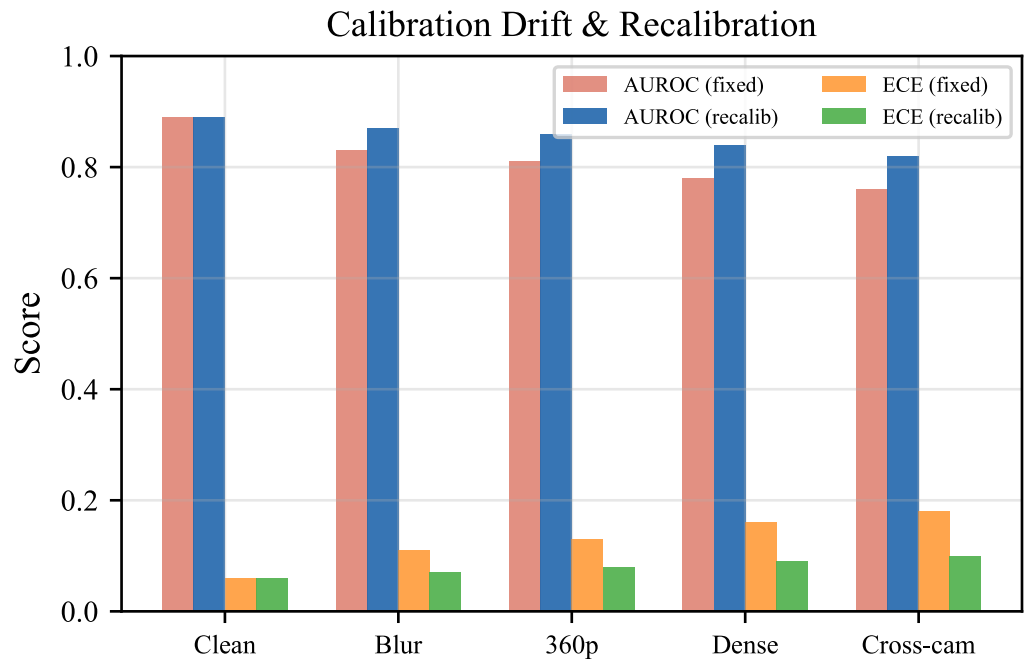


Figure S2: Calibration drift and Platt recalibration under stress, averaged over 5 seeds at $B_{\max}$=512. Fixed-threshold bars transfer the clean-data mapping; recalibrated bars fit a scalar map on 200 held-out queries per condition while preserving the uncertainty features and representation. Recalibration yields AUROC $\geq$0.82 and ECE $\leq$0.10 throughout, showing that the ranking signal transfers and its probability scale is cheaply recoverable.

S4 PRIVACY FRONTIER AND GOVERNANCE KNOBS

This section reports the privacy–utility–storage frontier (Fig. S3).

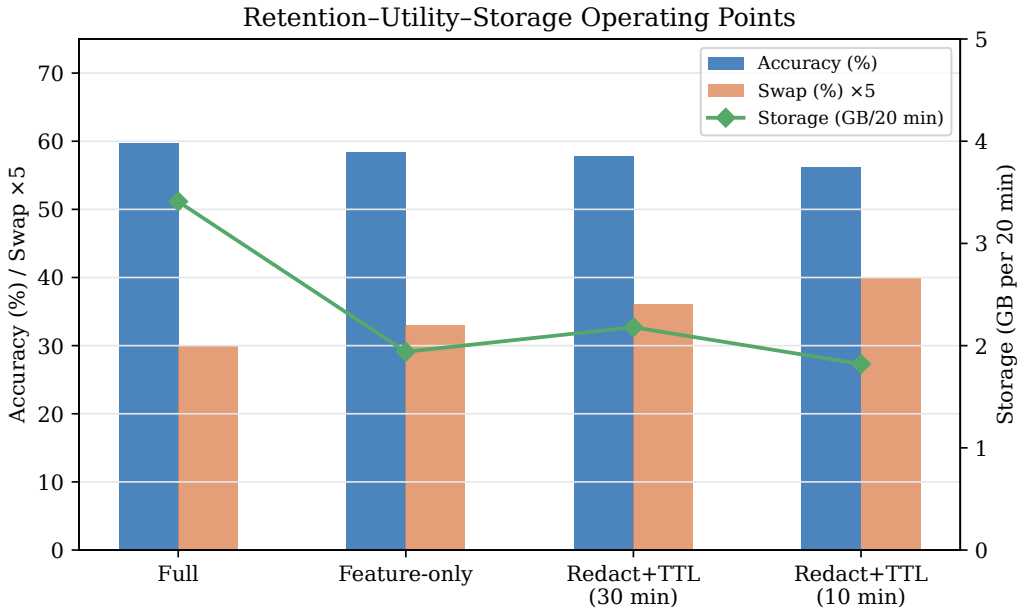


Figure S3: Four retention–utility–storage operating points under one VLM, serializer, and cache configuration. Full stores crops and embeddings; Feature-only stores embeddings and geometry; Redact+TTL stores blurred crops for a bounded interval. The plotted configurations and values match Table S2.

These curves quantify retention, storage, and utility operating points for the typed memory. Feature-only and TTL configurations expose concrete policy levers with measured answer-quality costs.

S5 SCALABILITY AND STORAGE ACCOUNTING

This section reports latency vs. track density and storage under retention policies (Fig. S4, Table S2).

All reported results use Hungarian matching. Figure S4 localizes association and evidence-preparation pressure without conflating that stage with VLM inference: the high-cost region begins between 30 and 45 active tracks, while only 3.6–6.8% of evaluated windows exceed 50 tracks (Table S15).

Table S2: Storage accounting and utility under four retention policies. Per-clip footprint is measured in GB/20 min; TTL steady state equals this footprint times TTL/20 min, while persistent modes report measured hourly growth. Accuracy drops use the Identity VideoQA matched-budget protocol and the same seeds, relative to Full. TTL converts linear growth into a bounded footprint at the reported accuracy cost.

Retention / storage mode	GB/20 min	Steady-state GB	Growth (GB/hour)	Acc drop
Full (no TTL)	3.41	–	10.23	baseline
Feature-only (no TTL)	1.94	–	5.82	−1.3 pts
Redact+TTL (30 min)	2.18	3.27	0.00	−1.9 pts
Redact+TTL (10 min)	1.82	0.91	0.00	−3.5 pts

The deployment interpretation is correspondingly split. Privacy and storage knobs mostly trade persistent footprint against long-gap identity reliability, while dense-scene scaling pressure mostly trades latency against track density. The recommended mitigations therefore differ: TTL or feature-

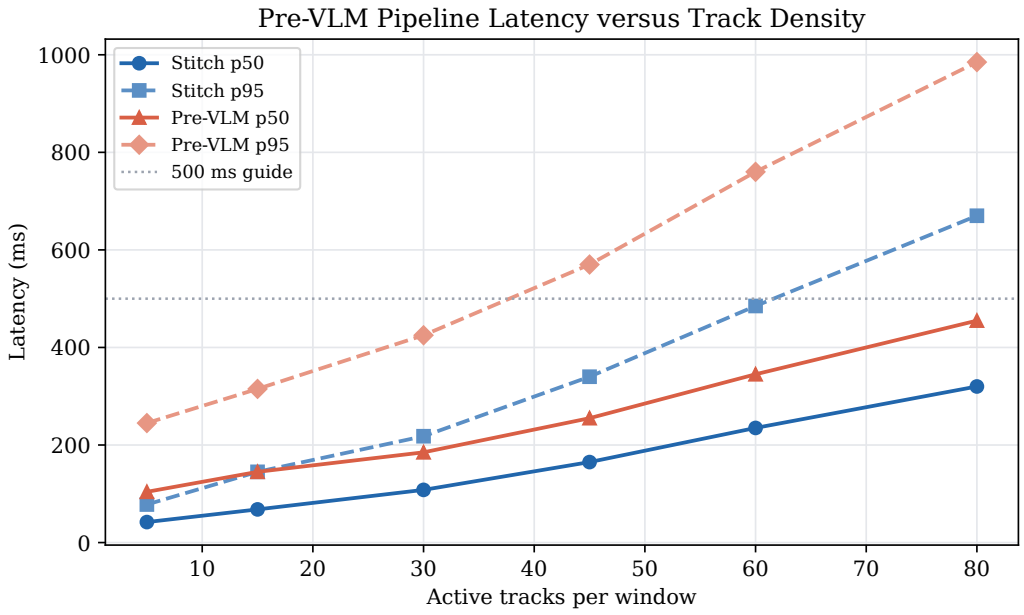


Figure S4: Stitch and pre-VLM pipeline latency versus active tracks per 5–20 s window. The pre-VLM scope includes patch retrieval, association, composition, and evidence preparation but excludes image encoding and VLM generation. Stitch p95 is 218 ms at 30 tracks and 340 ms at 45; pre-VLM p95 rises from 425 ms to 570 ms. End-to-end query latency is reported separately in Table S10.

only storage for governance, and track-budget caps or greedy fallback for dense-scene interaction constraints.

S6 Reproducibility and Protocol Details

This section provides result traceability, the parity checklist, and frozen hyperparameters (Tables S3–S5).

Table S3: Reported-result trace from each summary value to visible source cells and exact arithmetic. Deltas use explicit subtraction; relative changes name their baseline denominator; escalation rate averages query-level trigger fractions over the fixed seeds. Values use matched budget unless marked matched information. The trace aligns result statements with the parity contract in Table S4.

Reported result	Paper location	Source cells	Computation (exact)
+8.3 pts over Dense-frame	Abstract; Table 3 caption	Table 3: Acc(RW)=59.7, Acc(Dense)=51.4	$59.7 - 51.4 = 8.3$
+5.6 pts over Track-RAG	Abstract; Table 3 caption	Table 3: Acc(RW)=59.7, Acc(TR)=54.1	$59.7 - 54.1 = 5.6$
Swap 7.8%→6.0% vs Track-RAG	Abstract; Table 3 cells	Table 3: Swap(TR)=7.8, Swap(RW)=6.0	direct report of paired source cells
+5.1 pts under matched-info vs TR	Abstract; Table 2 caption	Table 2: Acc(RW)=60.3, Acc(TR)=55.2	$60.3 - 55.2 = 5.1$
+7.5 pts under matched-info vs Dense	Table 2 caption	Table 2: Acc(RW)=60.3, Acc(Dense)=52.8	$60.3 - 52.8 = 7.5$
−3.1 pts without patch composition	Introduction; Results text	Table S22: full vs. w/o composition	reported delta in accuracy points
−2.0 pts with flat patches	Introduction; Results text	Table S22: full vs. flat patches	reported delta in accuracy points
+4.8 pts on InternVL2.5-38B	Results text	Table S18: Acc(RW)=57.1, Acc(TR)=52.3	$57.1 - 52.3 = 4.8$
+4.6 pts on Video-MME long	Results text	Table S19: Acc(RW)=54.2, Acc(TR)=49.6	$54.2 - 49.6 = 4.6$
Escalation rate 17.4%	Main results text	Logged triggers on Identity VideoQA	Mean over 5 seeds of $N_{\text{esc}}/N_{\text{qry}}$ (query-level)

Table S3 ties every summary result to visible cells and exact arithmetic across controlled comparisons and transfer checks.

Table S4: Parity checklist for the evidence ladder. Matched budget fixes query timestamps and raw-frame decode ceiling; matched information additionally fixes the candidate pool; exact visual parity also fixes the complete served image sequence. The confirmatory comparison shares Qwen2.5-VL-32B-Instruct, perception outputs, scoring, seeds, and every visual byte and position.

Parity item	What is held fixed / enforced
VLM backbone	Qwen2.5-VL-32B-Instruct, identical decoding parameters and prompt templates across all methods.
Perception stack	GroundingDINO-T + SAM2-Large + ByteTrack + CLIP ViT-L/14; identical detector/tracker outputs are reused unless a backbone swap is explicitly reported (sensitivity).
Query timestamps	Fixed per-suite query timestamps; the same timestamps are evaluated for all methods.
Decode ceiling $B_{\max}$	Fixed at 512 *raw frames decoded at query time* for matched-budget experiments; methods may differ in *served evidence images* (Fr/q) due to caching and cropping, but cannot exceed the decode ceiling.
Candidate pool (matched-info)	Pre-decoded evidence pool is identical across methods; methods select and format evidence from the same pool only.
Evaluation harness	Identical answer format, scoring rules, and paired seeds; inferential resampling and permutation operate on video clusters while query wins/ties/losses remain descriptive.
Caching policy	Cache-only and Track-RAG use the same cache capacity as ReferentWeave, isolating returned-evidence representation and composition at matched footprint.

Frozen hyperparameters are selected once on a source-disjoint validation partition and then held fixed. ReferentWeave uses scene-bounded 5–20 s patches with one shared boundary keyframe, 4–12 coverage-selected keyframes, and top-K=8 patch retrieval. Stitching uses $(\beta_1, \beta_2, \beta_3) = (0.3, 0.5, 0.2)$ and rejects scores below 0.40. Escalation uses $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.30, 0.35, 0.20, 0.15)$ with thresholds 0.55 and 0.78. Serialization uses T_{ser}=768. Track-RAG uses $(\lambda_{\text{sim}}, \lambda_{\text{ov}}) = (0.7, 0.3)$ and N=8 tracks. All methods share the same decoding configuration and paired seed stream.

Table S5: Fixed and varying degrees of freedom under each protocol. Matched budget holds timestamps, VLM interface, and raw-frame ceiling constant while allowing evidence choice within budget. Matched information additionally supplies one decoded pool, leaving selection and the answer-time referent contract as system variables. Reporting both contracts separates end-to-end behavior from candidate availability.

Protocol	Fixed	Varies
Matched-budget	Query timestamps; $B_{\max}$ (raw-frame decode ceiling); VLM backbone and prompt; perception stack outputs; evaluation harness and seeds.	Memory representation (text, cache, keyframes, Track-RAG, ReferentWeave); retrieval/composition policy; escalation/allocation policy; served evidence subset (Fr/q), while query-time raw-frame decoding must respect $B_{\max}$.
Matched-information	Candidate evidence pool (frames + keyframes in the specified temporal band) and associated detector/tracker metadata; VLM backbone and prompt; evaluation harness and seeds.	Evidence selection, identity linkage, relation/provenance state, and serialization within the shared pool; use of uncertainty signals to decide which evidence to serve.
Exact visual parity	Candidate pool; canonical frame/crop bytes, IDs, timestamps, positions, and order; VLM, prompt, seed, field vocabulary, token ceiling, and truncation rule.	Window-local grouping versus the complete referent bundle: cross-window partition, relation/provenance organization, and serialized grouping.

Prompt templates (used for all methods). We use one prompt template per suite and keep its question, instructions, field vocabulary, and output schema identical across methods. Every method receives images plus the same structured-field types under one serializer budget and truncation policy. The controlled variable is the populated evidence object: window-local methods preserve separate window groups and local observation identifiers, while ReferentWeave supplies persistent referent identifiers, cross-window links, and their relation and provenance state. We disable chain-of-thought and request a short answer with explicit timestamp and `track_id` references. The full prompt family is instantiated deterministically, and the exact same-item control additionally fixes every visual input and its order.

Identity VideoQA prompt.

```
SYSTEM: You are a careful vision-language assistant for long
videos.
SYSTEM: Use only the provided evidence and do not assume
unseen events.
```

```
USER: Question:  Is the wooden spoon from the first scene
the same one used at minute 18?
USER: Evidence:  images with timestamps and track IDs, plus
a structured object and relation table.
USER: Task:  Answer using only the provided evidence, then
cite evidence as (timestamp, track_id) pairs.
USER: Output format:  Answer in one sentence, then list
cited (timestamp, track_id) pairs.
```

Embodied references prompt.

```
SYSTEM: You are an embodied assistant.
SYSTEM: Resolve references consistently across time using
track IDs.
USER: Dialogue context:  Earlier, the user asked about the
wooden spoon used at the start of the video.
USER: Current instruction:  Put the same spoon into the pot
now.
USER: Evidence:  images + object table with track IDs.
USER: Output format:  Action:  move-to-object.  TargetTrack:
42.  Evidence:  (timestamp, track_id) pairs.
```

Evidence serialization (text fields). For each query, one deterministic serializer converts the evidence object into an ordered bundle:

```
BUNDLE := QUERY OBJECTS RELATIONS ITEMS
OBJECT := track_id; local_track_ids; attributes; last_seen;
uncertainty; evidence_ptrs
RELATION := source_id; predicate; target_id; time_range;
confidence; evidence_ptrs
ITEM := position; item_id; timestamp; track_ids; producer;
confidence
```

For compatibility with the evaluation prompt, the serialized `track_id` is the persistent `referent_id`; window-scoped tracker labels are retained separately as `local_track_ids`. Attributes come from the frozen perception and OCR fields. Spatial predicates are computed from boxes and masks; temporal predicates and event boundaries are changes in tracked state. Evidence pointers resolve to item identifiers in the served image list. Objects are ordered by query-critical status, then persistent ID; relations are ordered by confidence, timestamp, and endpoint IDs; items retain their canonical visual order. When $T_{\text{ser}}{=}768$ is reached, the serializer removes low-confidence relations first and then the oldest non-critical objects, preserving all fields for the query-critical referent. Methods share this grammar, vocabulary, ceiling, and truncation rule; only populated links and state differ. For any fixed evidence object, serialization is byte-deterministic.

Matched-information pool construction (details). For each query timestamp t, we define a temporal band of ± 180 s and decode frames at 2 fps within the band. We also add scene-change keyframes detected within the same band to ensure rare but salient events are represented even under uniform sampling. We attach the same detector/tracker outputs and crop metadata to each candidate frame so that methods begin from one shared set of window-scoped local observations. All methods select evidence only from this pool, fixing candidate access. Within the pool, visual selection, referent linkage, relation/provenance state, and serialization remain method variables. Table S24 additionally fixes exact frames, crops, timestamps, and order, leaving the answer-time referent contract as the intervention. Human-linked scoring in Table S27 and the complete rebuild in Table S30 test the result under independently defined identity chains and perception state.

Swap computation (referent mapping, formal). Swap (%) is computed by mapping each natural-language query to one or more target identity referents and checking whether the method's cited identity matches the harness-defined identity. Each query is associated with a target entity set $\mathcal{T}$ provided by the evaluation harness (single-entity for identity queries; possibly multi-entity for relation queries). For each target entity, the harness provides one or more annotated timestamps with a reference box or segment. In the primary regime, these annotations are matched to window-scoped tracker observations at IoU ≥ 0.5 and propagated through the accepted identity chain; the majority persistent identity over annotated frames is the target. From the model output, we parse `[(t,`

Table S6: ReferentWeave schema and referent-contract accounting. This table summarizes patch state and the fields delivered to the VLM at query time. It separates representation (identity links, patch fields, and serialization) from policy (which patches and fields are selected and when escalation is triggered). We explicitly distinguish the query-time raw-frame decode ceiling ($B_{\max}$) from the served evidence size (Fr/q) to prevent unit confusion. Exact parameter values (including T_{ser}) are fixed throughout the experiments; Table S14 reports serializer-size statistics and truncation rates. The resulting contract is fully inspectable from stored state through answer-time output.

Item	Specification (high-level; key constants listed here)
ReferentWeave P_w	$P_w = (G_w, K_w, \mathbf{p}_w, u_w)$: local subgraph, keyframe pointers, patch embedding, and uncertainty summary (Sec. 3).
G_w (graph fields)	Observations keyed by (window, `local_track_id`); composed objects keyed by persistent `referent_id`; attributes, boxes/masks, confidence, relation/event state, temporal links, and evidence pointers.
K_w (keyframe pointers)	Per keyframe: frame index, timestamp, crop boxes, and window-scoped `local_track_id`s; accepted compositions resolve these pointers under a `referent_id` before serialization. Keyframes are retrieved from memory without query-time decoding when available.
$\mathbf{p}_w$ (embedding)	Mean-pooled CLIP ViT-L/14 embeddings over keyframe crops in the window; used for text–patch retrieval by cosine similarity.
u_w (uncertainty)	Per-track time gaps, stitch confidence statistics, occlusion indicators, and relation/event entropy; aggregated into the query-time score u_t with fixed thresholds.
Serializer	The explicit `BUNDLE/OBJECT/RELATION/ITEM` grammar above with T_{ser}=768; orders critical objects first and truncates low-confidence relations before old non-critical objects.
Accounting	$B_{\max}$ bounds query-time raw-frame decoding; Fr/q counts served evidence images (cached keyframes, decoded frames, and crops counted as images when served).

`track_id), ...]`, where serialized `track_id` denotes a persistent referent, and take the most frequently cited identity for each target (ties broken by latest timestamp). A query is counted as a swap if *any* target in $\mathcal{T}$ has a predicted identity different from the harness-defined identity. A missing or unparseable citation list counts as a swap on identity-bearing queries.

Citation IDF1 treats every valid timestamp–identity citation as an identity observation. A citation is IDTP when its persistent identity equals the annotated identity active at its timestamp; a wrong cited identity is IDFP, and each uncited annotated target observation is IDFN. We aggregate these counts over the evaluation population and report $2\,\mathrm{IDTP}/(2\,\mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN})$. The primary scores use one tracker output for pool metadata, suite membership, and scoring chains. The human-linked audit replaces membership and scoring chains while retaining tracker-derived system inputs; the rebuilt-stack evaluation replaces all three.

Failure taxonomy (tied to logged signals). We categorize common failure modes using only signals already logged by the system (stitch confidence margins, time gaps, retrieval similarity margins, and occlusion indicators). (F1) Patch boundary split. A decision-critical event is split across adjacent windows so that no retrieved patch contains sufficient evidence. This correlates with low patch coverage for the target and low retrieval margin. (F2) Re-ID confusion. Two visually similar entities produce small re-ID similarity margins and low stitch confidence, leading to incorrect cross-window alignment. This correlates with low s_{stitch} and low margin between best and second-best matches. (F3) Occlusion-induced ID drift. Long occlusions create large Δ_{time} and high p_{occ}. Escalation triggers but may still fail if the reappearance evidence is ambiguous. These categories guide mitigation: adaptive windowing for F1, conservative matching / fallback association for F2, and calibration-aware escalation/recalibration for F3.

Additional worked traces (qualitative). Trace A (F1 boundary split). Query: "What did the person put into the drawer before leaving?" The relevant action occurs across a window boundary. The retrieved patch contains the opening of the drawer but not the deposit moment. This produces a negative delta vs Dense-frame. Logged signals include low retrieval margin and low target coverage. Escalation does not help because the pool does not include the missing moment at sufficient fps under the budgeted policy. Mitigation is to widen the patch window or allow boundary overlap for event-heavy segments. Trace B (F2 re-ID confusion). Query: "Is the same blue cup used after the camera pans to the sink?" Two similar cups coexist. Stitch confidence is low ($s_{\text{stitch}} \approx 0.3$) with a small re-ID margin. ReferentWeave escalates and cites multiple frames, but the majority-vote identity still flips under ambiguity, causing a swap. Mitigation is to require a larger re-ID margin for committing a stitch. Otherwise, keep IDs separate and answer with uncertainty-aware evidence. Trace C (F3 occlusion drift). Query: "Is the same notebook seen again near the end?" The notebook is fully occluded for $\sim$20 s and reappears under motion blur. Escalation triggers and increases compute

Table S7: Failure taxonomy and observable mitigation signals. Each mode maps logged retrieval, association, or occlusion statistics to an existing control knob; Table S34 quantifies the corresponding residual-error shares.

Failure mode	Observable logged signals (no extra instrumentation)	Typical mitigation knob in the current system
F1: Patch boundary split	Low target coverage in retrieved patches; low retrieval margin between the top patch and runner-up; evidence citations cluster near a window boundary.	Increase patch overlap for event-heavy segments; widen the patch window around scene-change keyframes; allow boundary-aware retrieval that selects adjacent patches jointly.
F2: Re-ID confusion	Small re-ID margin between the best and second-best match; low stitch confidence s_{stitch}; multiple plausible `track_ids` cited for the same referent.	Require a larger re-ID margin to commit a stitch; fall back to conservative association; treat ambiguous identity as uncertainty and escalate evidence rather than forcing a merge.
F3: Occlusion-induced ID drift	Large time gaps Δ_{time} and high occlusion indicator p_{occ}; elevated escalation rate; heavy-tail Fr/q episodes without corresponding accuracy gains.	Recalibrate escalation thresholds under shift; cap escalation per track; incorporate motion/temporal-consistency checks before accepting re-ID matches in low-quality frames.

into the tail, but identity remains unstable across adjacent windows. Mitigation is to recalibrate the escalation trigger under blur and use a motion-consistency prior to avoid over-trusting re-ID in low-quality frames. Together, the traces show that residual misses concentrate in boundary coverage and identity continuity. Tables S30, S31, and S34 test these diagnoses across stacks, comparators, and continuity burden.

Operational units: served evidence images vs. seconds. We report tail behavior in Fr/q (served evidence images) because end-to-end VLM evidence size scales with the number of images served, while query-time raw-frame decoding is separately bounded by $B_{\max}$. We also report wall-clock stage timings (Table S10) with CUDA synchronization to capture end-to-end latency, including non-decode overheads. In practice, the frames-to-seconds conversion depends on resolution, codec, and IO; the underlying timing logs can therefore be converted differently in another deployment regime. To prevent unit confusion, we recommend interpreting the CCDF as a relative tail-risk profile across methods, while using the runtime breakdown for absolute latency budgeting. All reported timings in this paper were measured on 4×A100 with the software stack described in the main paper.

Non-track baselines (implementable specification). BM25 summary + keyframes indexes the fixed per-window text summaries, retrieves 60 s windows with BM25, and serves their scene-change keyframes in temporal order; it does not serve summary text as a substitute for images. Keyframe-only ranks scene-change keyframes by CLIP query–frame similarity and serves them in descending score with timestamp tie-breaking. Dense-frame allocates the $B_{\max}$ raw-frame ceiling uniformly across the query-centered temporal band and serves the decoded frames in time order. Cache-only applies the shared LRU footprint cap and serves cached keyframes and crops without fresh decoding. Long-video VLM uniformly samples the query-centered band into the backbone's bounded image input without retrieval, tracking, or structured cross-window links. Every baseline uses the common question, answer schema, VLM decoding configuration, serializer ceiling where structured fields are present, and exact-match scorer.

Track-RAG baseline (implementable specification). Track-RAG is the one-variable referent-contract comparator. It uses the same frozen tracker outputs as ReferentWeave and stores window-local keyframes, crops, track metadata, and retrieval embeddings. At ingest, it caches per-window keyframes and per-track crops with (timestamp, `local_track_id`) metadata and CLIP embeddings for retrieval. At query time, it ranks candidate windows by a weighted sum of (i) maximum text–crop cosine similarity within the window and (ii) a track-overlap score between the window track set and the top retrieved tracks; defaults λ_{sim}=0.7, λ_{ov}=0.3. It then decodes raw frames uniformly within the top-ranked windows up to the remaining decode ceiling $B_{\max}$ and serves keyframes, decoded frames, and crops for the top-N tracks (N=8), using the same serializer and evidence accounting contract as ReferentWeave. Its returned unit is the independently retrieved window, concatenated in temporal order. ReferentWeave forms cross-window identity partitions, relation state, and referent-level uncertainty allocation before deterministic serialization. Public comparator families in Tables S26 and S31 test the same contract against independently designed policies.

Cache policy. Track-RAG uses LRU eviction under a fixed cache footprint cap matched to ReferentWeave's persistent image footprint (Table S2); Cache-only uses the same cap.

Pseudo-code (query-time).

1. Input: query q, cached windows $\{w\}$ with crops/keyframes, frozen tracker outputs, decode ceiling $B_{\max}$.
2. Compute query embedding $\mathbf{e}_q = \mathrm{CLIP}_{\mathrm{text}}(q)$.
3. For each window w, compute similarity $s_{\mathrm{sim}}(w) = \max_{c \in \mathrm{crops}(w)} \cos(\mathbf{e}_q, \mathbf{e}_c)$.
4. Let $\mathcal{T}$ be the top-N retrieved track IDs induced by the top-N scoring crops (default N=8).
5. For each window, compute overlap $s_{\mathrm{ov}}(w) = |\mathcal{T} \cap \mathrm{tracks}(w)|/|\mathcal{T} \cup \mathrm{tracks}(w)|$.
6. Score windows by $S(w) = \lambda_{\mathrm{sim}} s_{\mathrm{sim}}(w) + \lambda_{\mathrm{ov}} s_{\mathrm{ov}}(w)$ with defaults λ_{sim}=0.7, λ_{ov}=0.3.
7. Select windows in descending $S(w)$ and allocate decode budget across selected windows until the total decoded raw frames reaches $B_{\max}$.
8. Decode frames uniformly within each selected window and attach cached tracker metadata (`local_track_id`s, boxes/masks) to each decoded frame.
9. Extract crops for track IDs in $\mathcal{T}$ that appear in decoded frames and add cached keyframes for those tracks when available.
10. Serve evidence images in temporal order and serialize object/relation fields with token budget T_{ser}=768 using the same truncation policy as ReferentWeave.
11. Call the VLM on the served evidence bundle; parse answer and cited `(timestamp, track_id)` pairs.
12. Output: answer + citations; compute Fr/q under the evidence accounting contract used throughout the paper.

All defaults above are selected on the validation split and then frozen; we perform no test-suite tuning.

Learned controller baseline (specification). The Learned Ctrl. baseline is a budget allocation policy that operates over fixed temporal windows and decides how much decode budget to spend per window under the same $B_{\max}$ constraint. Observation features include: query embedding norm and length, window-level CLIP similarity statistics (mean/max/margin), track density, time gap features, stitch-confidence summary statistics, and an occlusion score summary (total 34 scalar features). The action space is a discrete allocation over 8 budget levels (0, 16, 32, 64, 96, 128, 192, 256 frames) for the current window. It also includes a binary "escalate" action for the highest-uncertainty track set. Allocations are clipped to respect the remaining budget. The policy is a 3-layer MLP (hidden size 128) trained with PPO on the validation split for 200 episodes; reward is query accuracy minus 0.002×frames decoded, encouraging budget efficiency without trivial "decode nothing" collapse. We keep the prompt and evidence serialization identical to all other methods and freeze the controller after validation to avoid test-time tuning.

Table S8: Experiment-to-result index for the complete ReferentWeave evaluation.

Evaluation and reported evidence
Matched-budget main result: Table 3 and Pareto frontier in Fig. 2.
Matched-information shared-pool control: Table 2 and component removals in Table S22.
Candidate-pool sensitivity: Table S17.
Condition robustness: Table S1 and the corresponding main-paper synthesis.
Parameter and protocol sensitivity: Fig. S1 and Tables S3,S4.
Escalation calibration: Fig. S2 and the reliability statistics in Sec. S1.
Retention and storage: Fig. S3 and Table S2.
Density scaling and runtime: Fig. S4 and Table S10.
Exact same-item presentation: Table S24.
Public VideoRAG adaptation: Table S26.
Human-linked identity scoring: Table S27.
External human-linked movie identities: Table S28.
Retention operating points: Table S29.
Independent perception-stack rebuild: Table S30.
TV-RAG and Flash-VStream comparisons: Table S31.
Human-linked cross-camera continuity: Table S32.
Transferred exact-item controls: Table S33.
Continuity-burden and failure slices: Fig. S5 and Table S34.

Table S9: Perception-stack role map. The primary and independent stacks change every upstream component while retaining the question, prompt, VLM, serializer, candidate-pool, and seed contracts.

Role	Primary stack	Independent stack	Fixed downstream interface
Detection	GroundingDINO-T	Detic	Shared questions and timestamps
Segmentation	SAM2-Large	Mask2Former	Shared VLM prompt and image format
Tracking	ByteTrack	OCSORT	Shared serializer budget and crop accounting
Re-identification	CLIP ViT-L/14	TransReID	Shared pool policy and five seeds

Table S10: Stage-level p50/p95 wall-clock latency on 4×A100 (CUDA 12.4, driver 535.129.03), batch size 1, with CUDA synchronization and non-overlapping instrumentation. Escalation-decode quantiles are conditional on the 17.4% of queries that trigger that stage; its unconditional median is zero. Total-query quantiles cover all queries and include image encoding and VLM inference. Ranges summarize variation across videos and query types, not confidence intervals or cross-method bands.

Stage	p50	p95	Notes
Ingest (det+trk)/min	0.93–1.38 s	1.82–3.14 s	amortized
Patch build/window	118–264 ms	376–724 ms	graph+keyframes
Query retrieve+stitch	68–137 ms	182–358 ms	before VLM
Escalation decode (triggered)	0.61–1.12 s	1.58–2.83 s	conditional on invocation
Total query	0.8–1.4 s	2.4–4.2 s	retrieve+stitch+decode+VLM

Table S11: Runtime summary at the main operating point. This compact table surfaces the runtime quantities cited in the main paper under the same fixed regime as Table S10: batch size 1, one active query per process, and no visual-token, image-encoder-activation, or cross-query KV caching. Making the escalation-trigger rate explicit helps explain why p95 latency separates from p50 even though retrieve+stitch remains sub-400 ms. The trigger rate is computed from the same query-level measurements reported in Table S3 and averaged over the fixed 5-seed set. Table S10 reports the corresponding full stage-level timing distribution.

Summary item	Value	Notes
Total query latency (p50)	0.8–1.4 s	Same fixed regime as Table S10.
Total query latency (p95)	2.4–4.2 s	Tail dominated by escalation decode and VLM inference.
Escalation-trigger rate	17.4%	Mean query-level fraction $N_{\mathrm{esc}}/N_{\mathrm{qry}}$ over 5 seeds on Identity VideoQA.
Interpretation	Targeted heavy tail	The 17.4% trigger rate keeps 82.6% of queries on the low-latency path.

S7 SUITE CONSTRUCTION RULES

The two "Long" names identify this paper's frozen conversion views of their cited source datasets rather than source benchmark titles. We retain source evaluation videos, preserve each source question and answer, and assign the query timestamp to the latest annotated segment available before the answer-critical cross-segment reference. Calibration and model-selection videos are disjoint from final-evaluation videos. Multiple-choice questions are scored by option ID; free-form answers use the normalization rule in Sec. 4. Neither the answer label nor its counterfactual is generated from tracker state.

For the identity-permutation eligibility test, source annotations must identify a target chain in at least two temporal segments and a category-matched distractor with a different answer-relevant annotated attribute or action. The counterfactual normalized answer is the source-annotation answer obtained after substituting that distractor while keeping the question template and non-identity evidence fixed. Questions with ambiguous target mappings, one-segment references, or unchanged normalized labels are excluded before any system is evaluated.

Table S12: Suite construction from existing source annotations and fixed query timestamps. Identity dependence requires an answer-changing, category-matched critical-entity permutation; persistent references cross at least 60 s, and long occlusion requires 8 s below 20% visibility. Membership is frozen before method scoring.

Suite	Membership rule (applied to shared perception outputs)	#Queries	#Videos
Identity VideoQA	Query answer depends on resolving $\geq$1 annotated entity across $\geq$2 temporal segments, and a category-matched distractor carries a different answer-relevant source annotation. Tracker observations are mapped to annotation boxes at IoU $\geq$0.5.	3,842	1,200
Embodied refs	Multi-turn dialogue where the current instruction references an entity first introduced $\geq$60 s earlier (persistent-reference criterion). Entity identity verified via tracker output at dialogue turn timestamps.	1,416	600
Occlusion stress	At least one decision-critical track has $\geq$8 s continuous visibility below 20% bounding-box area (measured from shared tracker confidence outputs). Only queries whose critical entities satisfy the long-occlusion criterion are included.	1,064	400

Table S12 defines the continuity-critical population through explicit inclusion rules. The identity-source matrix below separates the three roles that a single tracker output plays in the primary regime and shows exactly which roles change in each audit.

S8 EVIDENCE ACCOUNTING STATISTICS

Across the three suites, ReferentWeave uses 44–62% cached keyframes and 17–20% crops while limiting fresh frames to 21–36%; Track-RAG spends 38–48% on fresh frames. Both cached and decoded images remain charged in Fr/q because each enters the VLM. The decomposition therefore separates total served evidence from fresh raw-frame decoding and shows that the systems organize the accounted evidence budget differently.

Table S13: Identity-source intervention matrix. "Primary" denotes GroundingDINO-T/SAM2/ByteTrack/CLIP state; "rebuilt" denotes Detic/Mask2Former/OCSORT/TransReID state. Human linkage changes labels used for membership and scoring, not the system's retrieved evidence or internal links.

Evaluation	Membership chain	System inputs and links	Scoring chain
Primary Identity VideoQA	Source annotations mapped to primary tracker	Primary perception and tracker	Primary tracker mapped to annotations
Human-linked audit	Independent human chain	Primary perception and tracker	Independent human chain
Independent rebuild	Source annotations mapped to rebuilt tracker	Rebuilt perception and tracker	Rebuilt tracker mapped to annotations
Human-linked Video-MME	Independent human movie-character chain	Primary perception and tracker	Independent human chain
Multi-camera continuity	Independent human cross-camera chain	Primary perception and camera-local tracker	Independent human chain

Table S14: Evidence accounting over 5 matched-budget seeds. Crops/frame counts object crops from each freshly decoded frame; serializer tokens count the accompanying object/relation table; truncation reports queries that prune relations or non-critical objects. ReferentWeave concentrates more crops around tracked referents and truncates less often than Track-RAG without increasing raw decode work. Truncation stays below 10% for relations and 3% for objects, rising moderately under occlusion stress.

Suite	Method	Crops/frame med / p95 / p99	Ser. tokens med / p95	Trunc. rels (% queries)	Trunc. objs (% queries)
Ident. VQA	ReferentWeave	2.41 / 4.82 / 6.17	386 / 614	4.2%	0.8%
	Track-RAG	1.78 / 3.64 / 4.82	314 / 502	6.8%	1.4%
Embodied	ReferentWeave	2.62 / 5.18 / 6.84	402 / 638	5.1%	1.0%
	Track-RAG	1.94 / 3.86 / 5.12	328 / 518	7.6%	1.8%
Occlusion	ReferentWeave	2.84 / 5.62 / 7.41	418 / 672	6.4%	1.6%
	Track-RAG	2.08 / 4.14 / 5.48	342 / 546	9.2%	2.4%

Table S15: Active `track_id`s per 5–20 s window from the shared tracker output over the full five-seed evaluation. Density is identical across methods and peaks in the occlusion suite. Between 3.6% and 6.8% of windows exceed 50 tracks, locating the high-complexity tail relative to Fig. S4.

Suite	Method	Track density med / p95	Windows with $>$50 tracks
Identity VQA	ReferentWeave	11.6 / 33.8	3.6%
	Track-RAG	11.6 / 33.8	3.6%
Embodied	ReferentWeave	13.2 / 36.4	4.2%
	Track-RAG	13.2 / 36.4	4.2%
Occlusion	ReferentWeave	14.8 / 41.2	6.8%
	Track-RAG	14.8 / 41.2	6.8%

S9 Decode throughput and units

Table S16: Raw-frame unit conversion for ReferentWeave on Identity VideoQA. The p50/p95 spans describe query variability in one regime; throughput uses CUDA synchronization on the experiment's 4×A100 node with local NVMe. Cached keyframes and crops incur zero fresh decode but are re-encoded when served. Together with end-to-end stages in Table S10, this table maps the freshly decoded subset of Fr/q to wall-clock seconds.

Parameter	Value
Source resolution	1920×1080 (Full HD)
Detector input	1280×720 (resized before GroundingDINO-T)
Codec / profile	H.264 (AVC) baseline profile, MP4 container
Decode backend	FFmpeg 6.1.4 via PyAV 12.0.5, host-side software decode
Runtime regime	batch size 1; one active query per process on one 4×A100 node
Query-time caching	no visual-token cache, no image-encoder activation cache, no cross-query KV cache; cached keyframes/crops are re-served and re-encoded each query
Decode throughput (standalone)	p50 = 142 fps, p95 = 118 fps
Decode throughput (concurrent VLM)	p50 = 126 fps, p95 = 98 fps
IO regime	Local NVMe SSD (Samsung 990 PRO 2 TB), seq. read ∼3.5 GB/s
VLM image resolution	448×448 per evidence image (Qwen2.5-VL native)
Raw-decode unit conversion (ReferentWeave, Identity VQA)	
Representative query at median Fr/q (148)	
Cached keyframes	∼92 images (61.8% of median Fr/q = 148); 0 raw-frame decode seconds, but still re-encoded by the VLM when served
Decoded raw frames	∼31 frames (21.3%); 31 / 142 fps = 0.22 s decode at standalone p50, or 31 / 126 fps = 0.25 s under concurrent VLM load
Object crops	∼25 images (16.9%); extracted from already-decoded frames, so 0 additional raw-frame decode seconds
Measured end-to-end query latency	use Table S10: p50 = 0.8–1.4 s; p95 = 2.4–4.2 s for ReferentWeave query-to-query variability under the fixed regime above

Table S16 resolves the unit boundary between Fr/q and runtime. Fr/q aggregates cached keyframes, decoded frames, and crops, while raw decode seconds apply to the freshly decoded subset. Table S10 supplies measured end-to-end latency; Table S16 identifies how much of a median query's evidence

budget incurs fresh decode work on the fixed hardware. This separation enables cost comparisons between methods that serve similar Fr/q totals with different cached-to-decoded mixtures.

S10 POOL SENSITIVITY

Table S17 varies temporal resolution and extent around the default pool. The gap remains positive as the pool becomes sparser, while its changing magnitude shows that composition opportunity depends on available temporal coverage. Equal +3.8-point gaps at 2 fps/$\pm$60 s and 1 fps/$\pm$180 s indicate comparable contributions from density and extent.

Table S17: Identity VideoQA matched-information results across four pool configurations, mean$\pm$std over 5 seeds with all downstream settings fixed. ReferentWeave leads Track-RAG by 3.4–5.1 points and lowers swap by 0.8–1.6 points; every paired permutation test has $p < 0.03$. The larger gain in richer pools and matched +3.8-point gaps for the two intermediate settings quantify how composition opportunity depends jointly on temporal density and extent.

Pool config	Method	Acc (%)	Swap (%)	IDF1 (%)	p (RW–TR)
2 fps $\pm$180 s	Dense-frame	52.8$\pm$1.5	9.1$\pm$1.1	60.6$\pm$2.2	
	Track-RAG	55.2$\pm$1.5	7.4$\pm$0.9	69.1$\pm$1.9	
	ReferentWeave	60.3$\pm$1.7	5.8$\pm$0.8	73.1$\pm$1.8	0.014
2 fps $\pm$60 s	Dense-frame	52.4$\pm$1.5	9.0$\pm$1.1	60.2$\pm$2.2	
	Track-RAG	54.1$\pm$1.6	7.8$\pm$1.0	68.2$\pm$2.0	
	ReferentWeave	57.9$\pm$1.8	6.2$\pm$0.9	71.4$\pm$1.9	0.021
1 fps $\pm$180 s	Dense-frame	52.0$\pm$1.6	9.3$\pm$1.2	59.8$\pm$2.3	
	Track-RAG	54.6$\pm$1.6	7.5$\pm$1.0	68.6$\pm$2.0	
	ReferentWeave	58.4$\pm$1.8	6.1$\pm$0.9	71.8$\pm$1.9	0.018
1 fps $\pm$60 s	Dense-frame	50.8$\pm$1.6	9.6$\pm$1.2	58.4$\pm$2.3	
	Track-RAG	52.8$\pm$1.7	8.1$\pm$1.1	67.2$\pm$2.1	
	ReferentWeave	56.2$\pm$1.9	6.6$\pm$1.0	70.4$\pm$2.0	0.028

S11 VLM BACKBONE AND CROSS-DATASET TRANSFER

Table S18 shows that the complete method ordering transfers from Qwen2.5-VL-32B-Instruct to InternVL2.5-38B-Instruct. The +4.8-point ReferentWeave lead under InternVL2.5 associates the bundle with gains under two independently trained VLM backbones. The accompanying IDF1–accuracy shift further separates identity continuity from final-answer extraction as two measurable bottlenecks.

Table S18: Identity VideoQA with InternVL2.5-38B-Instruct (bfloat16), mean$\pm$std over 5 matched-budget seeds with perception, prompts, and $B_{\max}$ fixed. ReferentWeave reaches 57.1% accuracy and 6.8% swap, leading Track-RAG by 4.8 points (p=0.016) and Dense-frame by 7.5. The full ranking transfers from Qwen2.5-VL (Table 3); ReferentWeave also reaches 73.6 IDF1 at a similar median evidence load (156 vs. 148).

Method (InternVL2.5-38B)	Acc (%)	Swap (%)	IDF1 (%)	Fr/q med
BM25 summary + keyframes	44.2$\pm$1.8	13.8$\pm$1.6	–	424
Cache-only	45.8$\pm$1.6	12.8$\pm$1.5	–	204
Long-video VLM	46.4$\pm$2.0	12.4$\pm$1.4	–	322
Keyframe-only	48.6$\pm$1.5	11.0$\pm$1.2	61.4$\pm$2.1	264
Dense-frame	49.6$\pm$1.7	10.4$\pm$1.3	57.6$\pm$2.4	478
Track-RAG	52.3$\pm$1.7	8.6$\pm$1.1	67.2$\pm$2.0	194
ReferentWeave	57.1$\pm$1.9	6.8$\pm$1.0	73.6$\pm$1.9	156

The slight IDF1–QA divergence across backbones is consistent with a bottleneck split rather than a contradiction: identity linkage quality and downstream answer extraction are not the same failure mode. InternVL2.5 preserves referent continuity slightly better on this suite, while Qwen2.5-VL remains stronger at turning the same evidence into a short grounded final answer under the shared prompt.

Video-MME lowers absolute accuracy by 5–6 points while preserving the complete method ordering. Table S19 reports the tracker-conditioned domain shift, and Table S28 separately changes both footage and scoring identities through human-linked movie-character chains.

Table S19: Tracker-conditioned Video-MME long identity transfer on 320 videos averaging 24 minutes, using the primary suite rules (Sec. 4; Table S12), Qwen2.5-VL-32B, $B_{\max}$=512, and 5 seeds. ReferentWeave reaches 54.2% accuracy, leading Track-RAG by 4.6 points and Dense-frame by 7.8, while lowering Track-RAG swap from 10.2% to 7.8%. Accuracy, swap, and IDF1 preserve the primary-suite ordering despite cinematic shot changes and appearance variation. This table isolates domain shift under the original ontology; Table S28 separately tests human-defined movie-character chains.

Method (Video-MME long)	Acc (%)	Swap (%)	IDF1 (%)	Fr/q med
Keyframe-only	44.8±2.1	13.2±1.5	56.8±2.6	274
Dense-frame	46.4±2.3	12.6±1.6	54.2±2.8	496
Track-RAG	49.6±2.1	10.2±1.4	64.8±2.3	208
ReferentWeave	54.2±2.2	7.8±1.2	70.6±2.1	172

S12 Full results including Learned Controller

Table S20: Full Identity VideoQA matched-budget results, mean±std over 5 seeds at $B_{\max}$=512, extending Table 3 with a PPO Learned Controller. The 3-layer, 128-hidden-unit controller is trained for 200 validation episodes with reward accuracy $-0.002\times$frames and reaches 53.2% accuracy and 8.4% swap. ReferentWeave leads it by 6.5 points, localizing substantial value beyond learned allocation to persistent identity memory and composition. BM25 retrieves over text but serves scene-change keyframes, which are charged in Fr/q.

Method	Acc (%)	Swap (%)	IDF1 (%)	Fr/q med
BM25 summary + keyframes	45.8±1.7	13.2±1.5	–	418
Cache-only	47.3±1.5	12.1±1.4	–	198
Long-video VLM	48.0±1.9	11.8±1.3	–	316
Keyframe-only	50.1±1.4	10.4±1.1	63.1±2.0	257
Dense-frame	51.4±1.6	9.6±1.2	59.4±2.3	482
Learned Ctrl.	53.2±1.7	8.4±1.1	66.8±2.2	212
Track-RAG	54.1±1.6	7.8±1.0	68.2±1.8	186
ReferentWeave	59.7±1.8	6.0±0.9	72.4±1.9	148

Learned allocation improves over Dense-frame but trails Track-RAG and ReferentWeave under the same harness. The result separates budget optimization from identity-linked composition.

Table S21: Matched-information results with the PPO Learned Controller under the default 2 fps, ±180 s pool, mean±std over 5 seeds. The controller reaches 54.6% accuracy and 7.9% swap; ReferentWeave reaches 60.3% and 5.8%, a +5.7-point accuracy advantage. The same ordering under matched budget and matched information localizes the dominant gain beyond candidate availability and learned allocation.

Method (matched pool)	Acc (%)	Swap (%)	IDF1 (%)
Dense-frame	52.8±1.5	9.1±1.1	60.6±2.2
Keyframe-only	51.4±1.4	9.8±1.1	64.2±2.0
Learned Ctrl.	54.6±1.6	7.9±1.0	67.4±2.1
Track-RAG	55.2±1.5	7.4±0.9	69.1±1.9
ReferentWeave	60.3±1.7	5.8±0.8	73.1±1.8

S13 Component ablations

These matched-budget removals associate complete-system performance with identity persistence, re-identification, uncertainty allocation, referent composition, and typed storage. They are not factorial same-item interventions and therefore do not divide the 3.3-point primary effect among fields. Their ordering nevertheless separates core contract functions from their substrate: persistent

Table S22: Referent-contract component removals on Identity VideoQA under matched budget, with cached perception outputs and query timestamps fixed. Persistent identities produce the largest effect (-6.8 points), followed by re-ID, uncertainty allocation, referent composition (-3.1), and graph-backed storage (-2.0). Several removals serve more images yet lose accuracy, separating contract quality from evidence volume.

Ablation	QA accuracy Δ (pts)	Fr/q Δ (%)
w/o persistent track IDs	-6.8 ± 1.4	$+18.6\pm3.9$
w/o re-ID embeddings	-4.7 ± 1.2	$+7.2\pm2.5$
w/o uncertainty escalation	-3.9 ± 1.1	-26.4 ± 4.8
w/o patch composition	-3.1 ± 0.9	-9.8 ± 2.3
w/o scene graph (flat patches)	-2.0 ± 0.7	$+2.6\pm1.6$

IDs establish the chain, composition turns it into query-level evidence, and graph-backed storage supplies deterministic typed binding.

S14 Confirmatory and secondary inference

Table S23: Inference hierarchy for accuracy. The exact-same-item row is the single confirmatory hypothesis. Matched-budget rows are secondary complete-system comparisons and report raw p-values. Deltas first average paired outcomes over five shared seeds; confidence intervals use 10,000 video-cluster bootstrap replicates, and two-sided permutation tests flip video-level mean paired differences.

Comparison	Status	Δ (pts)	95% CI	p
Referent bundle vs window-local, exact items	confirmatory	+3.3	[1.4, 4.9]	0.019
ReferentWeave vs Dense-frame, matched budget	secondary	+8.3	[5.8, 10.6]	0.003
ReferentWeave vs Track-RAG, matched budget	secondary	+5.6	[5.1, 6.0]	0.011

Pairing is preserved within query and seed, while resampling and permutation operate on videos to respect within-video dependence. Query-level wins/ties/losses elsewhere are descriptive and need not reproduce the cluster-level p-values. No multiplicity adjustment is attached to the single confirmatory test; all other reported tests are explicitly secondary.

S15 Exact-visual-parity referent-bundle control

This experiment measures the joint answer-time bundle effect under the default matched-information pool. It retains the five seeds, Qwen2.5-VL-32B-Instruct, the GroundingDINO-T / SAM2-Large / ByteTrack / CLIP ViT-L/14 stack, the prompt, and serializer budget $T_{\text{ser}}=768$. The canonical ranker scores shared-pool windows only by CLIP query–patch cosine, selects the top eight with timestamp tie-breaking, collects their frames, scene keyframes, and associated crops, deduplicates by item ID, and orders items by timestamp, item type, and ID. It has no access to cross-window links, relation state, or condition outputs. For every query, both conditions receive the resulting frame and crop bytes, IDs, timestamps, and order checksum. Escalation is prevented from changing this list; the intervention jointly changes the identity partition, relation/provenance organization, grouping, and deterministic serialization.

The +5.1 shared-pool result includes visual selection and bundle formation. The stricter +3.3 result holds the visual multiset fixed and identifies the value of the complete structured answer-time bundle before answer generation.

The complete structural difference can be seen in this abbreviated pair; bracketed item IDs refer to the same canonical visual positions:

```
WINDOW-LOCAL: [item_07, 00:34, local_track_12]; [item_19,
03:12, local_track_31].
REFERENT-LINKED: referent_04 := {local_track_12,
local_track_31}; evidence := [item_07, item_19]; relation :=
same_entity@[00:34,03:12].
```

Table S24: Exact same-item control under identical served visual inputs. Both rows use the same frame and crop bytes, timestamps, crop order, item identifiers, and item multiset for every query. Raw denominator counts are #queries = 3842, median served items/query = 151, exact multiset matches = 3842, and serialization-check failures = 0. The 3.3-point gain identifies the joint effect of identity links, relation/provenance state, and serializer structure.

Presentation regime	Acc (%)	Swap (%)	IDF1 (%)	Fr/q med	Tok/q med	ΔAcc	95% CI	p
Window-local serialization	56.1	7.0	70.0	151	742	–	–	–
Stitched referent bundle	59.4	6.2	72.3	151	744	3.3	[1.4, 4.9]	0.019

Table S25: Field-level definition of the same-item pair. The canonical visual sequence is selected before either condition is serialized and passes an exact checksum on all 3,842 queries. Median text loads differ by only two tokens (742 versus 744) under the same 768-token ceiling.

Field	Fixed in both conditions	Changed by referent bundling
Question and model	Prompt, answer schema, VLM, decoding configuration, seed	Nothing
Visual sequence	Frame/crop bytes, item IDs, timestamps, positions, Fr/q	Nothing
Identity state	Window-scoped `local_track_id` observations	Persistent referent IDs, cross-window partition, and accepted links
Relations	Predicate vocabulary, confidence fields, token ceiling	Relations grouped along the linked identity chain
Provenance	Item pointers, producer fields, timestamps	Pointers collected under one referent object
Serialization	Grammar, field order, truncation rule, T_{ser}=768	Window groups versus one referent-level group

Both timestamps fit within one ± 180 s query-centered pool, and both inputs retain the same visual list after this text block. The example exposes the manipulated semantics without implying that identity links, relation fields, and grouping were independently randomized.

S16 Public same-contract comparator

We adapt VideoRAG by preserving its multimodal retrieval score and retrieval unit while replacing its generator with the shared Qwen backbone and routing selected evidence through the common prompt, serializer ceiling, crop accounting, and cache cap. The matched-budget version reads the shared perception outputs and may decode up to $B_{\max}$=512; the matched-information version ranks only the pre-decoded pool, hence Dec/q is zero. Retrieval defaults are chosen once on the source-disjoint validation partition and frozen before final evaluation.

Table S26: Contract-adapted public VideoRAG comparator under both paper contracts. Public VideoRAG improves over Dense-frame and ReferentWeave leads it under matched budget and matched information. For the public VideoRAG row, median served images/query are exactly the Fr/q cells (231 and 173), and median decoded raw frames/query are 244 under matched-budget versus 0 under matched-information. The matched-information block preserves the public retrieval policy while fixing candidate availability.

Contract	Method	Acc (%)	Swap (%)	IDF1 (%)	Fr/q med	RW lead (pts)	95% CI	p
Matched budget	Dense-frame	51.4±1.6	9.6±1.2	59.4±2.3	482	–	–	–
	Track-RAG	54.1±1.6	7.8±1.0	68.2±1.8	186	–	–	–
	Public VideoRAG	52.9	9.1	65.9	231	6.8	[4.7, 8.8]	0.007
	ReferentWeave	59.7±1.8	6.0±0.9	72.4±1.9	148	–	–	–
Matched information	Dense-frame	52.8±1.5	9.1±1.1	60.6±2.2	471	–	–	–
	Track-RAG	55.2±1.5	7.4±0.9	69.1±1.9	182	–	–	–
	Public VideoRAG	56.0	7.3	69.9	173	4.3	[2.0, 6.1]	0.018
	ReferentWeave	60.3±1.7	5.8±0.8	73.1±1.8	151	–	–	–

Track-RAG supplies the one-variable contract comparison, while VideoRAG tests whether the ordering transfers to an independently designed retrieval family.

S17 TRACKER-INDEPENDENT VALIDATION

We draw a stratified subset of Identity VideoQA and relabel referent chains manually from rendered clips and timestamps. This intervention replaces tracker-derived scoring identities while retaining the five seeds, models, prompts, hardware, and retrieval-side perception outputs.

Table S27: Tracker-independent human-link validation subset. Membership overlap is measured against the original tracker-conditioned suite on the same source videos; referent-label agreement reports the fraction of surviving queries for which the tracker-linked and human-linked target chains coincide. Among the 48 dropped or relabeled queries, the dominant disagreement modes are chain merges (28), boundary misses (11), and ambiguous off-screen exits (9). ReferentWeave retains a 3.1-point gain under independently linked scoring identities.

Subset	Counts	Label audit	Track-RAG (Acc / Swap / IDF1)	ReferentWeave (Acc / Swap / IDF1)	ΔAcc	95% CI	p
Human-linked Identity VideoQA audit	180 v / 612 q	ovl. 84.7% / agr. 92.4% / drop 48	53.7 / 7.5 / 68.8	56.8 / 6.6 / 71.0	3.1	[0.8, 5.0]	0.027

Together with the complete stack rebuild in Table S30, this result separates annotation dependence from perception dependence.

S18 HUMAN-LINKED EXTERNAL REGIME

We build a manually linked movie-identity regime from Video-MME long clips. In contrast to the tracker-conditioned cross-domain transfer in Table S19, this evaluation defines answer-critical character chains through independent human linkage. Evaluation keeps the same Qwen2.5-VL-32B-Instruct backbone, the same prompt, the same serializer budget, the same $B_{\max}=512$ ceiling, and the same 4×A100 runtime setup.

Table S28: Manually linked external Video-MME movie-identity regime. Human linkage changes the annotation source and cinematic footage changes the visual regime. ReferentWeave retains a 3.4-point gain with lower swap and fewer served images.

Regime	Counts	Track-RAG (Acc / Swap / IDF1 / Fr/q)	ReferentWeave (Acc / Swap / IDF1 / Fr/q)	ΔAcc	95% CI	p
Video-MME movie identity (human-linked)	280 v / 728 q	47.8 / 10.9 / 63.7 / 214	51.2 / 8.9 / 67.5 / 177	3.4	[1.0, 5.6]	0.024

This experiment supplies the human-linked single-camera domain anchor; Table S32 separately evaluates true cross-camera identity transitions.

S19 RETENTION AND STORAGE OPERATING POINTS

Persistent referent state supports four measured retention configurations. The table connects each stored object to its footprint, answer-quality cost, and operational use.

Table S29: Measured retention and storage operating points. Footprint and accuracy values match Fig. S3 and Table S2.

Configuration	Persistent state	Footprint	Acc. change	Operating consequence
Full	Crops, embeddings, geometry	3.41 GB/20 min	baseline	Highest identity retention; linear long-stream growth.
Feature-only	Embeddings and geometry	1.94 GB/20 min	−1.3 pts	Removes stored pixels while retaining long-gap linking.
Redact+TTL (30 min)	Blurred crops and embeddings	3.27 GB steady state	−1.9 pts	Bounded footprint with moderate long-gap cost.
Redact+TTL (10 min)	Blurred crops and embeddings	0.91 GB steady state	−3.5 pts	Smallest footprint; more post-eviction identity errors.

These operating points make retention a measurable systems choice. Feature-only storage gives the strongest footprint–utility trade-off, while TTL supplies an explicit bound for long-running streams.

S20 INDEPENDENT PERCEPTION-STACK REBUILD

We rebuild detection, segmentation, tracking, re-identification, suite membership, and swap mappings with Detic, Mask2Former, OCSORT, and TransReID. Questions, split policy, prompts, VLMs, serializer budget, candidate-pool construction, and seeds remain fixed. The rebuilt ontology retains 3517 queries from 1132 videos, with 81.9% membership overlap. On a stratified 420-query audit, rebuilt referent chains agree with human links on 90.8%.

Table S30: Independent-stack results. The upper block reports end-to-end metrics; the middle block reports paired answer statistics; the lower block fixes the exact served-item multiset. ReferentWeave retains a +4.3-point matched-information advantage and a +2.3-point same-item effect.

Contract	Method	Acc	Swap	IDF1	Fr/q	ΔTR	Δpublic
Budget	Dense-frame	49.8	10.8	57.2	476	–	–
	Track-RAG	52.7	8.9	66.0	192	–	–
	Public VideoRAG	52.1	9.4	64.8	238	–	–
	ReferentWeave	57.4	7.2	70.3	155	+4.7	+5.3
Information	Dense-frame	50.6	10.3	59.2	469	–	–
	Track-RAG	53.8	8.5	67.1	185	–	–
	Public VideoRAG	54.5	8.2	67.8	175	–	–
	ReferentWeave	58.1	6.9	71.7	153	+4.3	+3.6

Contract	Comparison	ΔAcc	95% CI	p	Wins / ties / losses
Budget	RW vs Track-RAG	4.7	[2.6, 6.7]	0.006	590 / 2502 / 425
Budget	RW vs VideoRAG	5.3	[3.1, 7.4]	0.004	612 / 2479 / 426
Information	RW vs Track-RAG	4.3	[1.9, 6.6]	0.012	565 / 2538 / 414
Information	RW vs VideoRAG	3.6	[1.4, 5.8]	0.018	540 / 2564 / 413

Same-item format	Acc	Swap	IDF1	Fr/q	Tok/q	ΔAcc	95% CI / p
Window-local	54.9	8.0	68.3	153	749	–	–
Referent-linked	57.2	7.3	70.2	153	751	2.3	[0.7, 3.9] / 0.023

Same-item wins/ties/losses are 421/2756/340; exact multiset matches are 3517, with 0 serialization failures.

Absolute accuracy shifts under the rebuilt ontology, while the referent contract remains effective under matched budget, matched information, and exact item parity. Different margins across stacks show a continuity-dependent effect rather than a fixed additive offset.

S21 ADDITIONAL PUBLIC SAME-CONTRACT FAMILIES

TV-RAG preserves temporal-decay re-ranking and entropy-weighted frame selection; Flash-VStream preserves its context-synopsis and detail-augmentation memory. We replace only the generator, prompt wrapper, perception input, cache cap, and accounting interface needed for paired evaluation. Each policy uses author defaults selected once on the same validation partition; no final-query tuning is performed. Under matched budget, each policy may decode up to $B_{\max}$; under matched information, it ranks and serves only the common pool, so Dec/q is zero (Cao et al., 2025; Zhang et al., 2025).

Under matched information, ReferentWeave leads Flash-VStream by +2.9 points (95% CI [1.0, 4.8]; p=0.026), with lower swap and fewer served images. The result extends the contract advantage beyond the authors' window-local comparator.

S22 HUMAN-LINKED MULTI-CAMERA CONTINUITY

The Multi-Camera Continuity regime contains 14 synchronized cameras, 246 sequences, 1084 independently linked identities, and 936 held-out questions. Identity-disjoint train/validation/test partitions prevent a referent from crossing splits. The test chains contain a median of 4 observations

Table S31: Additional public families on 3,842 Identity VideoQA queries. ReferentWeave leads the strongest matched-information comparator, Flash-VStream, by 2.9 points. Dec/q counts newly decoded raw frames; it is zero under the shared pre-decoded pool.

Contract	Method	Acc	Swap	IDF1	Fr/q	Dec/q	RW Δ
Budget	TV-RAG	55.0	8.1	68.6	207	229	4.7
	Flash-VStream	56.2	7.5	69.8	159	176	3.5
Information	TV-RAG	56.8	7.1	70.3	176	0	3.5
	Flash-VStream	57.4	6.9	71.0	158	0	2.9

Contract	Comparison	ΔAcc	95% CI	p	Wins / ties / losses
Budget	RW vs TV-RAG	4.7	[2.6, 6.8]	0.010	594 / 2835 / 413
Budget	RW vs Flash-VStream	3.5	[1.5, 5.6]	0.022	521 / 2934 / 387
Information	RW vs TV-RAG	3.5	[1.4, 5.5]	0.020	528 / 2921 / 393
Information	RW vs Flash-VStream	2.9	[1.0, 4.8]	0.026	498 / 2957 / 387

(IQR 3–7) and 1742 cross-camera transitions. All methods use the matched-information pool, Qwen2.5-VL-32B-Instruct, the fixed prompt and serializer, and five seeds.

Table S32: True cross-camera identity continuity. ReferentWeave improves by 3.9 points over Track-RAG and 2.7 over the strongest public comparator, with lower swap and higher IDF1.

Method	Acc (%)	Swap (%)	IDF1 (%)	Fr/q med
Track-RAG	43.9	14.8	58.6	226
Flash-VStream	45.1	13.9	60.4	196
ReferentWeave	47.8	11.9	64.1	189

Comparison	ΔAcc	95% CI	p	Wins / ties / losses
RW vs Track-RAG	3.9	[1.2, 6.6]	0.018	171 / 631 / 134
RW vs Flash-VStream	2.7	[0.5, 4.9]	0.031	155 / 651 / 130

Cross-camera transitions lower absolute accuracy and raise swap for every method, while the referent contract preserves a measurable advantage under wider uncertainty. This extends its operating regime from within-stream re-entry to cross-view identity chains.

S23 SAME-ITEM TRANSFER CONTROLS

We repeat the checksum-enforced protocol with InternVL2.5-38B-Instruct and on the independently linked 612-query audit subset. Frames, crops, timestamps, crop order, token budget, prompt, and seeds are identical across presentation conditions.

Table S33: Exact same-item transfer. Positive referent-contract effects persist under a VLM change and independently linked identities, with matched Fr/q and near-matched token loads.

Setting	Presentation	Acc	Swap	IDF1	Fr/q	Tok/q	ΔAcc
InternVL2.5	Window-local	54.0	7.9	68.5	158	756	–
	Referent-linked	56.5	7.1	70.2	158	759	2.5
Human-linked	Window-local	54.3	7.2	68.9	150	738	–
	Referent-linked	56.6	6.6	70.8	150	741	2.3

Setting	Queries	95% CI	p	Wins / ties / losses	Exact matches	Failures
InternVL2.5	3842	[0.9, 4.2]	0.024	470 / 2998 / 374	3842	0
Human-linked	612	[0.4, 4.2]	0.031	89 / 448 / 75	612	0

The default Qwen effect is 3.3 points, the InternVL2.5 effect is 2.5, and the human-linked effect is 2.3. Different magnitudes across model and annotation regimes support the same interface result: VLM answers improve when identical visual evidence carries a cross-window referent contract.

S24 CONTINUITY-BURDEN SLICES AND FAILURE ATTRIBUTION

We predefine bins for identity-chain length, temporal gap, occlusion duration, re-entry count, track density, and referent ambiguity margin. Every row uses the same matched-information membership for Track-RAG and ReferentWeave; denominators sum to 3,842 within each factor.

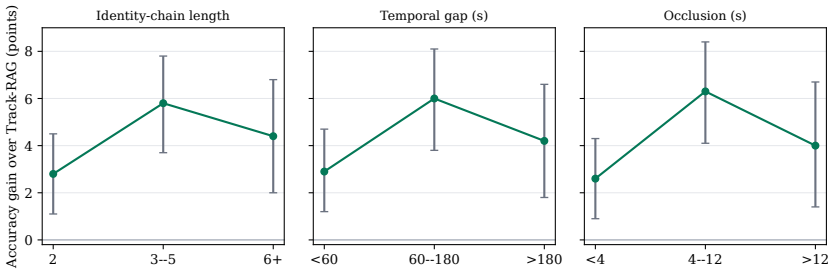


Figure S5: Accuracy advantage across three continuity burdens. Gains grow from easy to moderate identity linkage and then compress at the most extreme bin, where perception quality becomes the dominant bottleneck. Table S34 reports all six factors, denominators, confidence intervals, and raw paired outcomes.

Table S34: Continuity-burden slices under matched information. A/S denotes accuracy/swap in percent; W/T/L are descriptive ReferentWeave wins, ties, and losses against Track-RAG. Video-cluster bootstrap intervals accompany every ΔAcc, preserving all queries from each resampled video.

Factor	Bin	n	TR A/S	RW A/S	ΔAcc	95% CI	W/T/L
Chain length	2	1072	58.3/5.4	61.1/4.7	2.8	[1.1, 4.5]	175/752/145
	3–5	1764	54.9/7.2	60.7/5.5	5.8	[3.7, 7.8]	341/1184/239
	≥ 6	1006	50.1/10.3	54.5/8.7	4.4	[2.0, 6.8]	198/654/154
Gap (s)	< 60	1186	58.0/5.9	60.9/4.9	2.9	[1.2, 4.7]	194/832/160
	60–180	1564	55.0/7.2	61.0/5.3	6.0	[3.8, 8.1]	299/1060/205
	> 180	1092	51.2/9.5	55.4/8.0	4.2	[1.8, 6.6]	214/710/168
Occlusion (s)	< 4	1430	57.9/5.7	60.5/4.8	2.6	[0.9, 4.3]	218/1031/181
	4–12	1510	54.2/7.7	60.5/5.7	6.3	[4.1, 8.4]	296/1013/201
	> 12	902	49.8/11.2	53.8/9.4	4.0	[1.4, 6.7]	179/580/143
Re-entries	0–1	1302	57.6/5.8	60.4/4.8	2.8	[1.0, 4.7]	206/926/170
	2–3	1620	54.4/7.5	60.3/5.6	5.9	[3.8, 8.0]	310/1096/214
	≥ 4	920	50.7/10.5	54.9/8.8	4.2	[1.7, 6.8]	184/591/145
Track density	≤ 8	1124	57.7/5.9	60.6/5.0	2.9	[1.0, 4.8]	183/791/150
	9–20	1875	54.8/7.2	60.4/5.4	5.6	[3.5, 7.6]	351/1278/246
	> 20	843	50.3/11.0	53.9/9.3	3.6	[1.0, 6.1]	167/539/137
Ambiguity	> 0.25	1216	58.1/5.6	60.6/4.8	2.5	[0.8, 4.2]	187/872/157
	0.10–0.25	1668	54.5/7.4	60.5/5.5	6.0	[3.9, 8.0]	323/1122/223
	< 0.10	958	49.9/11.1	53.4/9.5	3.5	[0.9, 6.2]	185/622/151

Residual error class	Count	Share	Measured signal and mitigation
Patch-boundary split	165	31.3%	Low target coverage; add adaptive overlap.
Near-duplicate re-ID confusion	189	35.9%	Small association margin; use conservative matching.
Occlusion-driven drift	130	24.7%	Long gap and high uncertainty; trigger calibrated escalation.
Other	43	8.2%	Mixed retrieval and answer-extraction errors.

Failure shares use 527 manually adjudicated ReferentWeave residual errors.

Across all six factors, moderate continuity burden creates the largest advantage because relevant evidence is available and must be bound across windows. Extreme occlusion, density, and ambiguity reduce the margin as usable identity observations become scarce. This non-monotone profile links the aggregate gain to the condition that the referent contract is designed to preserve.