EGOSCOPE: SEEKING WHAT IS MISSING VIA QUERY-SUFFICIENT EVIDENCE ROUTING FOR LONG-HORIZON VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric video understanding supports episodic memory, personal assistance, and embodied reasoning. In long-horizon recordings, answering a query requires combining observations scattered across time, viewpoints, and changing object states under a limited evidence budget. Existing approaches build long-video representations and structured memories or select query-relevant observations through ranking, diversity modeling, and adaptive search. However, individually relevant or diverse observations may still leave essential query relations unsupported, while repeated evidence consumes the remaining budget. Inspired by human problem solving, we propose Egocentric Spatiotemporal Closure over Progressive Evidence (EGOSCOPE), a fixed-budget router that progressively seeks evidence for unresolved query demands. EGOSCOPE represents each query's evidence requirements as a transient Query Demand Graph composed of query-specific relations from explicit semantic categories. A lightweight set encoder recurrently updates relation support and conflict, using marginal closure and necessity–sufficiency signals to prioritize unresolved demands and suppress redundant or conflicting evidence before a single frozen task-head call. Across Ego4D NLQ, EgoLifeQA, and VSI-Bench, EGOSCOPE consistently improves temporal localization, long-term memory, and spatial reasoning over representative task-specific systems and matched evidence-selection baselines, while increasing relation closure and reducing evidence duplication. These results demonstrate that explicitly tracking unresolved query demands enables more sufficient evidence composition under fixed budgets.

1 INTRODUCTION

Egocentric video's moving viewpoint and long horizon scatter decisive evidence across time, object states, and places. Long, unscripted and synchronized first- and third-person recordings reveal this diversity (Damen et al., 2022; Grauman et al., 2024). Ego4D NLQ, EgoLifeQA, and VSI-Bench expose it through temporal localization, multi-day memory, and multi-view spatial reasoning (Grauman et al., 2022; Yang et al., 2025b;a). Context and memory limits constrain how much video a model can process (He et al., 2024; Song et al., 2024), yet long-context tasks require reasoning over scattered details (Wu et al., 2024). The problem is therefore to assemble, within a fixed budget, evidence that jointly supports the queried relations rather than views relevant only in isolation.

Prior work on evidence access in long-horizon egocentric video understanding broadly falls into two paradigms: relevance-based retrieval and diversity-based selection. Relevance-based methods rank candidates by query compatibility, using high-recall proposals, compressed video representations, or persistent memories (Hou et al., 2023; Di & Xie, 2024; Feng et al., 2025; Mangalam et al., 2023; Ye et al., 2025; Sun et al., 2026; Ma et al., 2026; Dong et al., 2026; Taluzzi et al., 2026). They retrieve relevant observations but do not track which query demands are already satisfied, often selecting redundant evidence. Diversity-based methods broaden coverage through combination ranking, redundancy suppression, or adaptive search (Yu et al., 2025; Wang et al., 2024; 2025; Zou et al., 2026; Pan et al., 2026; Ding et al., 2026; Zemskova et al., 2026; Kim et al., 2026). However, diversity does not guarantee the relations or ordering required for a sufficient answer and may still omit decisive evidence. Inspired by human problem solving, we instead identify the temporal events, entity states,

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.

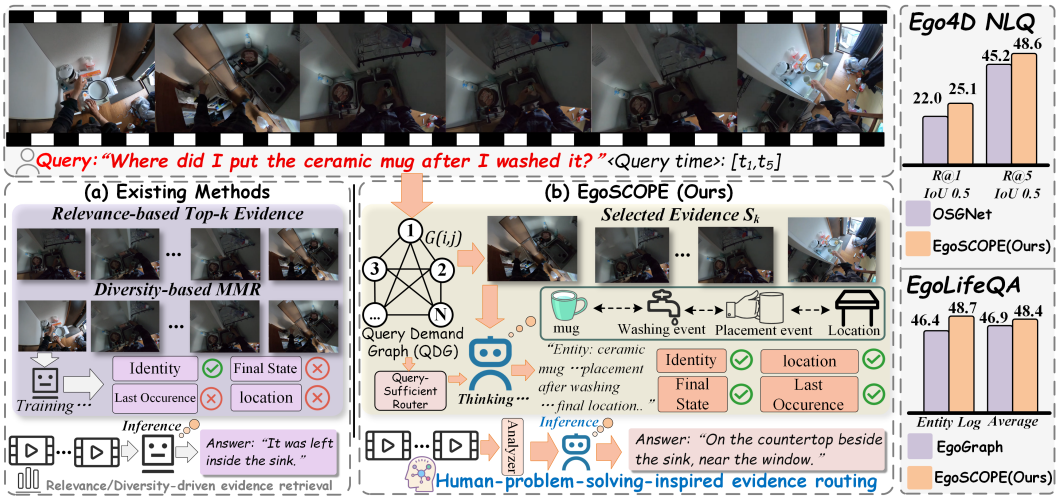

Figure 1: **Relevance-based, diversity-based, and human-problem-solving-inspired evidence routing.** Under the same budget, relevance-based Top-k selects repetitive views, while diversity-based MMR may still miss required relations. EGOSCOPE instead routes complementary evidence toward unresolved query demands. The right panels show its improvements on Ego4D NLQ and EgoLifeQA.

spatial relations, quantities, and goals required by a question, track which are supported, and seek evidence for those that remain unresolved. Each observation updates this state, promoting complementary evidence while suppressing redundancy and conflict. Figure 1 contrasts relevance-based repetition, diversity without guaranteed relation coverage, and our routing toward unresolved query demands.

Following this idea, we propose EGOSCOPE, a fixed-budget router that parses each query into a transient Query Demand Graph (QDG) of typed, instance-specific relations. A lightweight set encoder updates edge support and conflict after every acquisition and recomputes candidate utility from closure, necessity, sufficiency, task gain, redundancy, and contradiction. Satisfied relations lose value while unresolved ones redirect selection. Only the router is trained; candidate encoders and the task head remain frozen, and the answer model is called once after exactly K units are assembled.

Our contributions are fourfold.

- We formulate fixed-budget evidence selection as progressive closure of a transient query-side demand graph that exposes which answer relations remain unsupported.
- We propose a recurrent typed-relation router with explicit necessity, sufficiency, redundancy, and conflict signals to assemble complementary evidence without additional answer-model calls.
- We introduce a necessity–sufficiency objective with relation-level support and conflict supervision to preserve indispensable evidence and suppress redundant or contradictory candidates.
- Under matched budgets and one answer call, EGOSCOPE reaches 25.07 R@1 at IoU 0.5 on Ego4D NLQ, 48.4% macro accuracy on EgoLifeQA, and 44.74 on VSI-Bench.

2 Related Work

2.1 Long-Video Modeling and Structured Egocentric Memory

Video-language pretraining learns transferable egocentric representations from paired clips and text, cross-modal fusion, or automatically generated narrations (Lin et al., 2022; Pramanick et al., 2023; Zhao et al., 2023). Ego4D NLQ localizes natural-language events in first-person clips (Grauman

et al., 2022). GroundNLQ improves multi-scale grounding, GroundVQA couples grounding with long-video QA, and OSGNet adds object and shot structure (Hou et al., 2023; Di & Xie, 2024; Feng et al., 2025); UniversalVTG transfers a lightweight grounding model across datasets, while Hand Trajectory Fusion injects egocentric hand-motion cues (An et al., 2026; Zhong et al., 2026). EgoSchema and MMEgo further expose long-horizon modeling demands (Mangalam et al., 2023; Ye et al., 2025). Memory-based video models compress visual tokens or retain past information to accommodate long inputs (Song et al., 2024; He et al., 2024). Persistent representations retain complementary forms of structure. EgoGraph builds temporal knowledge graphs, R4DSG stores anchor-relative changes, ObjectStream maintains bounded object tracks, Imprint compresses recurring interactions, and EgoSG or action scene graphs serialize long activities (Sun et al., 2026; Ma et al., 2026; Dong et al., 2026; Das & Roy, 2026; Taluzzi et al., 2026; Rodin et al., 2024). Spatial methods reconstruct geometry or organize trajectories and entities for multi-view reasoning (Yang et al., 2025a; Gholami et al., 2026; Hua et al., 2026). These methods determine what a long video memory should retain globally. We reuse their high-recall representations and frozen task heads, but determine which fixed number of candidates jointly supplies the evidence still missing after a query arrives.

2.2 Query-Aware Evidence Selection and Search

Frame-Voyager learns one-shot frame-combination ranking (Yu et al., 2025), whereas A.I.R., TimeSearch-R, and VideoZoomer adaptively propose, validate, or zoom into evidence (Zou et al., 2026; Pan et al., 2026; Ding et al., 2026). FocusGraph selects graph-described clips and keyframes, and ReQuest combines question-aware selection with uncertainty and adaptive suppression (Zemskova et al., 2026; Kim et al., 2026). These methods establish the value of query-conditioned access, but their decision state is a ranking, generic diversity signal, or answer-model judgment rather than an explicit set of typed answer relations. Consequently, an already supported fact can remain highly ranked while a low-similarity but necessary relation is omitted. EgoSCOPE differs in both its state and budget boundary. It recomputes utility from unresolved edges of a transient query graph and never invokes the answer VLM during selection.

3 Method

Inspired by human problem solving, which identifies the information required by a question, assesses what the available evidence has established, and seeks what remains missing, EgoSCOPE assembles a fixed-size evidence set progressively. As shown in Figure 2, the Query Demand Graph makes these requirements explicit, a local state tracks their accumulated support and conflict, and the router redirects each selection toward unresolved relations. The resulting evidence set is then passed to the frozen task head once.

3.1 Problem Formulation

We formulate evidence routing to measure whether a better selection improves the answer under a fixed evidence budget and a single task-head call. Let q be a query, V a temporally ordered egocentric video, and $\mathcal{C} = \{c_i\}_{i=1}^{M}$ a fixed high-recall pool of evidence units. A frozen task head H is a temporal proposal head for Ego4D NLQ benchmark or a video-language answerer for EgoLifeQA and VSI-Bench. The router selects an ordered $S_K = (c_{i_1}, \ldots, c_{i_K})$ under budget K before calling H once:

$$\hat{y} = H(q, S_K), \qquad |S_K| = K. \tag{1}$$

The main protocol fixes K and the candidate pool. Offline feature extraction is excluded from the online call count but shared by every selector. Equation 1 therefore isolates whether a better S_K improves $\hat{y}$ without increasing context or invoking H during selection. The desired set must cover the relations required by q while avoiding redundant or conflicting evidence, so utility depends on what S_t has already resolved rather than on independent relevance alone. We evaluate this query-sufficient behavior through task performance, relation closure, duplication, and matched component, state, and relation controls.

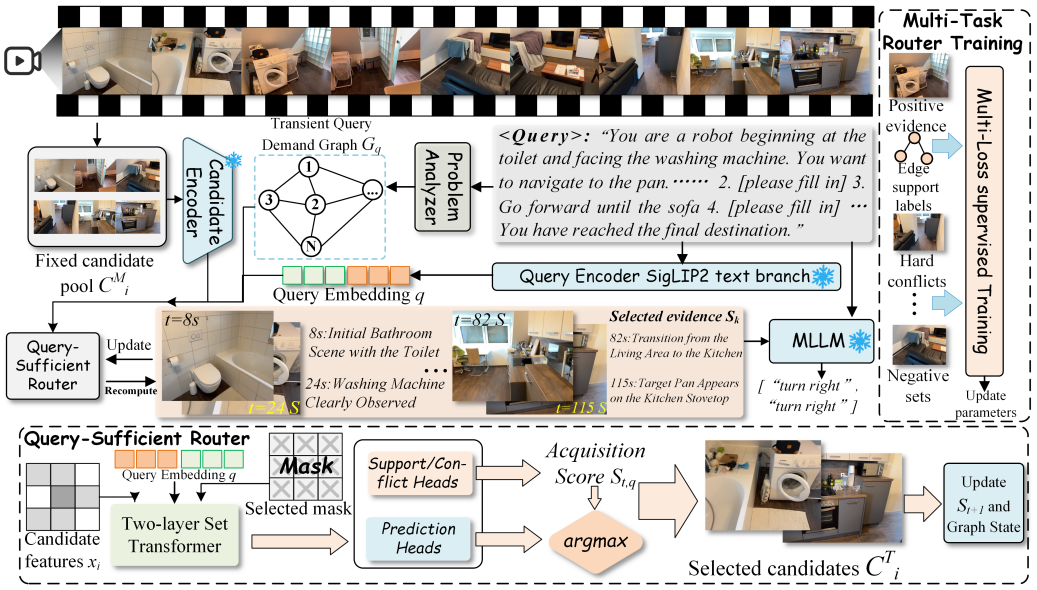

Figure 2: **Overall architecture of EGOSCOPE.** The query is parsed into a transient Query Demand Graph G_q, and a query-sufficient router recurrently selects evidence from a fixed candidate pool for K steps before one frozen MLLM call. Multi-task supervision updates only the router. Encoders and the task head remain frozen.

3.2 QUERY DEMAND GRAPH

The Query Demand Graph makes the evidence requirements of a query explicit so that the router can identify what is still missing. For example, "Where was the cup placed last?" requires cup identity, the last placement event, and its post-event location. We encode these distinct demands by parsing q into $G_q = (\mathcal{V}_q, \mathcal{E}_q)$. An edge

$$e = (u_e, r_e, v_e, a_e, w_e) \tag{2}$$

contains source u_e, predicate r_e, target or answer variable v_e, optional qualifier a_e, and confidence w_e. Predicates belong to temporal/event, entity/state, spatial/view, quantity/metric, or causal/goal families. Finite task templates and lexical rules emit the source span, predicate family and name, target, qualifier, and confidence, making every edge auditable rather than a free-form reasoning trace (Appendix A.4). Frozen before evaluation, the parser reaches 84.8/80.1 edge F1 on held-out Ego4D/EgoLifeQA queries; oracle graphs raise the corresponding task metrics by 1.81/1.96 points (Appendix B.4). Test answers never enter G_q.

3.3 CANDIDATE EVIDENCE AND LOCAL STATE

To decide which observation can satisfy a demand, the router needs both visual content and its temporal and local relational context. We represent a candidate c_i spanning $[t^s_i, t^e_i]$ as

$$\boldsymbol{x}_i = [\boldsymbol{v}_i; \boldsymbol{m}_i; \boldsymbol{t}_i; \boldsymbol{r}_i], \tag{3}$$

where $\boldsymbol{v}_i$ is a frozen SigLIP2 feature (Tschannen et al., 2025), $\boldsymbol{m}_i$ encodes displacement from the preceding view and optional flow, $\boldsymbol{t}_i$ contains normalized start, end, center, and duration, and $\boldsymbol{r}_i$ contains local relation-family priors. The same frozen vision-language encoder represents the query. NLQ units inherit intervals and confidence from mature GroundNLQ/OSGNet proposals (Hou et al., 2023; Feng et al., 2025), while QA pools cover the causal history or camera trajectory before routing (Appendix A.5). This separates pool recall from the paper's focus on evidence composition.

The local state records which requirements the selected set S supports and which it contradicts, allowing later selections to respond to accumulated evidence. It stores support $s_e(S)$ and conflict $k_e(S)$ per required edge. The sigmoid outputs $p_{ie}, \bar{p}_{ie} \in [0, 1]$ are bounded support and conflict scores, not calibrated independent probabilities. We use noisy-OR as a monotone bounded accumulator rather than as an independence assumption. Support accumulates as

$$s_e(S \cup \{i\}) = 1 - (1 - s_e(S))(1 - p_{ie}), \tag{4}$$

Algorithm 1 EGOSCOPE Fixed-Budget Query-Sufficient Routing

Require: Query q, candidate pool $\mathcal{C}$, budget K, frozen task head H
Ensure: Ordered evidence S_K and prediction $\hat{y}$
Parse q into G_q and encode q, candidates, and demand edges
Initialize $S_0 \leftarrow ()$ and edge support/conflict to zero
for $t = 0, \ldots, K-1$ **do**
  Re-encode every $i \notin S_t$ with the selected mask and graph state
  Predict $p_{ie}^t, \bar{p}_{ie}^t, u_i^t, n_i^t, z_i^t$ and compute $A(i \mid S_t, q)$
  Select $i_t \leftarrow \arg\max_{i \notin S_t} A(i \mid S_t, q)$ and append it to S_t
  Update edge support and conflict by the noisy-OR state rule
end for
Call the frozen head once and set $\hat{y} \leftarrow H(q, S_K)$
return $S_K, \hat{y}$

and conflict analogously uses $\bar{p}_{ie}$. Both scores may be high for a mixed observation; in that case support is retained, conflict is accumulated separately, and the conflict state conditions all later scores. The state exposes overall coverage, family-wise missing mass, and family-wise conflict. Because only edges incident to G_q are maintained, its size is bounded by query complexity rather than video duration.

3.4 STATE-CONDITIONED QUERY CLOSURE

The router redirects each selection toward relations that remain unresolved after earlier acquisitions. Because new evidence changes these requirements, it recomputes candidate utility at every step. At step t, a two-layer set Transformer encodes each candidate with the query, position, selected mask, and graph state.

$$\boldsymbol{h}_i^t = F_\theta(\boldsymbol{x}_i, \boldsymbol{q}, \boldsymbol{r}_i, \mathbf{1}[i \in S_t], \boldsymbol{z}_t). \tag{5}$$

An edge encoder maps each family and hashed predicate-argument signature to $\boldsymbol{g}_e$. Pairwise heads on $[\boldsymbol{h}_i^t; \boldsymbol{g}_e; \boldsymbol{h}_i^t \odot \boldsymbol{g}_e]$ predict p_{ie}^t and $\bar{p}_{ie}^t$. Recomputing F_θ after every acquisition lets support already supplied by S_t change all remaining ranks. A decisive early view can suppress its near-duplicates and redirect later steps toward an unresolved relation.

To reward evidence that fills a remaining gap, we measure the additional support a candidate supplies to unresolved edges. The marginal closure of candidate i is

$$\Delta_{\text{close}}(i \mid S_t, q) = \sum_{e \in \mathcal{E}_q} w_e \, \mathbf{1}[s_e(S_t) < \tau] \left(s_e(S_t \cup \{i\}) - s_e(S_t)\right). \tag{6}$$

The indicator removes an edge from marginal utility once its support crosses τ, while w_e retains the parser's confidence in unresolved demands. Consequently, a candidate can be valuable early and redundant later, even though its query similarity is unchanged. Unlike static relevance or pairwise diversity, the score changes after the selected set alters the query's remaining evidence requirements.

Candidate heads additionally estimate task gain u_i^t, deletion-based necessity n_i^t, and marginal sufficiency z_i^t. All three are bounded in $[0, 1]$: u_i^t is candidate–task compatibility, n_i^t predicts whether deleting i removes unique edge support, and z_i^t is the non-negative increase of a pooled-set sufficiency head after adding i. Redundancy $d(i, S_t)$ is the maximum non-negative cosine similarity to selected evidence. The acquisition score is

$$\begin{aligned} A(i \mid S_t, q) = \alpha \Delta_{\text{close}} + \beta n_i^t + \gamma u_i^t + \delta z_i^t \\ - \rho d(i, S_t) - \eta f(i, S_t), \end{aligned} \tag{7}$$

where $f(i, S_t) = \max_{e \in \mathcal{E}_q} \bar{p}_{ie}^t$ penalizes the candidate's strongest immediate edge conflict; accumulated $k_e(S_t)$ enters the graph state $\boldsymbol{z}_t$ and therefore conditions every subsequent head. Parser confidences are normalized over $\mathcal{E}_q$, so closure, the learned heads, redundancy, and conflict all lie in $[0, 1]$ before weighting. Closure rewards new edge support, necessity protects irreplaceable evidence, task gain preserves downstream relevance, and sufficiency scores the assembled set. We greedily choose $i_t = \arg\max_{i \notin S_t} A(i \mid S_t, q)$, update Eq. 4, and repeat to K. Algorithm 1 summarizes this learned recurrent policy; it provides neither a submodularity nor an optimality guarantee.

With $M \leq 64$ and a small number $E = |\mathcal{E}_q|$ of requirements, its recurrent self-attention cost is $O(K(M^2 d + MEd))$.

Table 1: Comparison with prior methods on the Ego4D NLQ v2 validation set. We report recall at ranks 1 and 5 under IoU thresholds 0.3 and 0.5.

Method	R@1, IoU 0.3 ↑	R@1, IoU 0.5 ↑	R@5, IoU 0.3 ↑	R@5, IoU 0.5 ↑
GroundNLQ (Hou et al., 2023)	26.98	18.83	53.56	40.00
GroundVQA (Di & Xie, 2024)	29.68	20.23	52.17	37.83
OSGNet (Feng et al., 2025)	31.63	22.03	57.91	45.19
UniversalVTG (An et al., 2026)	29.06	20.56	53.47	41.52
Hand Trajectory Fusion (Zhong et al., 2026)	26.54	18.50	52.37	–
EGOSCOPE (Ours)	**34.26**	**25.07**	**59.43**	**48.61**

Table 2: Comparison on the Jake-500 evaluation set of EgoLifeQA. We report accuracy for the five question categories and their macro-average. General-model and EgoGPT rows follow the EgoLife evaluation.

System	EntityLog	EventRecall	HabitInsight	RelationMap	TaskMaster	Average
Gemini 1.5 Pro (Yang et al., 2025b)	36.0	37.3	45.9	30.4	34.9	36.9
GPT-4o (Yang et al., 2025b)	34.4	42.1	29.5	30.4	44.4	36.2
LLaVA-OneVision (Li et al., 2025)	36.8	34.9	31.1	22.4	28.6	30.8
EgoGPT (Yang et al., 2025b)	39.2	36.5	31.1	33.6	39.7	36.0
Imprint (Das & Roy, 2026)	–	–	–	–	–	35.8
EgoGraph (Sun et al., 2026)	46.4	46.8	45.9	35.2	**60.3**	46.9
EGOSCOPE (Ours)	**48.7**	**48.2**	**47.4**	**37.6**	60.1	**48.4**

3.5 NECESSITY–SUFFICIENCY TRAINING

The training objective teaches the router to assemble jointly sufficient evidence while retaining indispensable observations and rejecting contradictions. It combines local relevance and edge-support supervision with set-level necessity and sufficiency signals.

$$\mathcal{L} = \lambda_{\text{task}}\mathcal{L}_{\text{task}} + \lambda_{\text{close}}\mathcal{L}_{\text{close}} + \lambda_{\text{nec}}\mathcal{L}_{\text{nec}} + \lambda_{\text{suf}}\mathcal{L}_{\text{suf}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}}. \tag{8}$$

All router targets are constructed automatically from official training annotations; no additional human relation labeling is used. Candidate–edge support follows deterministic temporal, event, or view alignment, while temporal reversals, entity substitutions, and spatial-direction inversions generate hard conflicts when the source annotation permits them. Leave-one-out deletion and complete edge coverage produce necessity and sufficiency targets, respectively; unavailable targets are masked. Training rolls out the model's own greedy route, uses the hard $\arg\max$ mask in the forward pass and a straight-through softmax mask in the backward pass, and applies the five losses at every visited state without teacher forcing. Appendix A.7 gives the loss definitions and weights, Appendix A.8 gives the benchmark mappings, and Appendix A.9 gives the complete training procedure. No VSI-Bench answer label trains the selector.

3.6 BENCHMARK INSTANTIATIONS AND ONLINE BUDGET

To apply the same routing mechanism across temporal localization, episodic memory, and spatial reasoning, we adapt only the evidence units and frozen task head. One router checkpoint, including its vocabulary, state update, and learned parameters, is trained jointly on Ego4D and EgoLifeQA and then deployed on VSI-Bench without VSI answer or relation supervision (Appendix B.3). Ego4D NLQ benchmark reranks interval proposals and returns the top five for the evaluator, EgoLifeQA routes only pre-query evidence to a frozen multiple-choice Video-LLM, and VSI-Bench covers the full trajectory with numeric normalization or choice decoding. LLaVA-OneVision-7B is primary (Li et al., 2025), with Qwen2.5-VL-7B and InternVL3-8B for transfer (Bai et al., 2025; Zhu et al., 2025). All controlled selectors share the cache, K, prompt, backbone, decoding, and one answer call.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

Table 3: Component ablation at $K = 8$. Checkmarks denote active mechanisms. Values are mean±std over three seeds, and closure/duplication are pooled held-out audits.

Variant	Rel.	Nec.	Suf.	Conf.	NLQ R@1@0.5 ↑	EgoLifeQA acc. ↑	VSI-Bench avg. ↑	Closure↑/Dup.↓ (%)
EGOSCOPE w/o closure	–	✓	✓	✓	23.35±0.3	46.79±0.6	43.17±0.6	67.00±1.8 / 24.00±1.5
EGOSCOPE w/o necessity	✓	–	✓	✓	24.24±0.5	47.72±0.7	44.23±0.5	74.00±1.6 / 16.00±1.2
EGOSCOPE w/o sufficiency	✓	✓	–	✓	24.51±0.4	48.14±0.5	44.25±0.6	75.00±1.5 / 15.00±1.1
EGOSCOPE w/o conflict	✓	✓	✓	–	24.88±0.4	48.16±0.6	44.47±0.5	78.00±1.5 / 14.00±1.1
Full EGOSCOPE (Ours)	✓	✓	✓	✓	**25.08±0.3**	**48.32±0.4**	**44.78±0.4**	**79.00±1.4 / 11.00±1.0**

Table 4: State and decision-mechanism ablation at $K = 8$. Entries are mean±std over three seeds.

Router variant	Ego4D R@1, IoU 0.5 ↑	EgoLifeQA acc. ↑	VSI-Bench avg. ↑	Router latency (s) ↓
One-shot ranking	23.55±0.6	47.08±0.8	43.30±0.7	**0.0642±0.0033**
Relation-label permutation	23.16±0.6	46.67±0.8	43.11±0.7	0.0861±0.0042
Random relation-family assignment	23.08±0.7	46.47±0.9	42.68±0.8	0.0859±0.0043
Static state	23.89±0.5	47.47±0.7	43.71±0.6	0.0849±0.0039
Full recurrent state (Ours)	**25.08±0.3**	**48.32±0.4**	**44.78±0.4**	0.0863±0.0041

We form $M = 64$-unit candidate pools and select $K = 8$ units in the main setting (Appendix B.2). Frozen 1,152-dimensional SigLIP2 features feed a two-layer, four-head Transformer router of width 256 (approximately 2.6M trainable parameters). We train with AdamW for at most 10 epochs using learning rate and weight decay 10^{-4}, early-stopping patience five, gradient clipping at 1.0, and seeds $\{42, 43, 44\}$. Training runs use NVIDIA RTX 3090 hardware. Offline features are cached and shared by all selectors. The VSI checkpoint is chosen using only Ego4D and EgoLifeQA supervision.

Table 5: VSI-Bench average over 5,130 questions. Direct and TRACE results are reported by Hua et al. (2026). Selector variants use LLaVA-OneVision with $K = 8$. Uniform, Temporal spread, and Top-k are selection rules evaluated here. Gray marks the matched runner-up/ours.

Method / selector	Average ↑
Gemini 3 Pro	
Direct (Hua et al., 2026)	52.61
+ TRACE (Hua et al., 2026)	**60.15**
Qwen2.5-VL-72B	
Direct (Hua et al., 2026)	36.28
+ TRACE (Hua et al., 2026)	**39.38**
MiMo-VL-7B	
Direct (Hua et al., 2026)	39.79
+ TRACE (Hua et al., 2026)	**41.42**
LLaVA-OneVision-7B (Li et al., 2025)	
Uniform	37.17
+ Temporal spread	38.56
+ Top-k	39.73
+ MMR (Carbonell & Goldstein, 1998)	41.91
+ DPP (Kulesza & Taskar, 2012)	41.46
+ **EGOSCOPE (Ours)**	**44.74**

4.2 Datasets and Metrics

For Ego4D NLQ v2, we use the official splits and annotated clips, reporting R@1 and R@5 at IoU $0.3/0.5$ (Grauman et al., 2022). For EgoLifeQA, following prior work (Yang et al., 2025b; Das & Roy, 2026; Sun et al., 2026), we evaluate the public 500-question Jake set with causal evidence ($t \leq t_q$), five category accuracies, and their macro-average. For VSI-Bench, we evaluate all 5,130 questions from 288 videos using the official aggregate (Yang et al., 2025a); Appendix B.1 gives the full protocols. Published systems provide task context; causal selector comparisons match the candidate pool, modality, $K = 8$ budget, frozen answer model, prompt, decoding, and one answer call. Appendix B.4 reports candidate-pool recall, matched-pool reranking, and the parser audit. Ablation metrics are mean±standard deviation across three seeds (Appendix D).

4.3 Main Results

Tables 1–5 compare EGOSCOPE with representative systems and evidence selectors. With the candidate pool, features, budget, frozen head, prompt, decoding, and answer-call count matched, EGOSCOPE exceeds the strongest selector control by 1.21, 0.84, and 1.39 points on Ego4D, EgoLifeQA, and VSI-Bench, respectively (Appendix B.4).

On Ego4D (Table 1), EGOSCOPE reaches 34.26/25.07 R@1 and 59.43/48.61 R@5 at IoU 0.3/0.5, improving over OSGNet by 2.63/3.04 and 1.52/3.42 points and outperforming UniversalVTG and Hand Trajectory Fusion on R@1. Gains at both thresholds, especially strict overlap, indicate that state-conditioned selection retains decisive occurrences instead of repeating similar intervals.

On EgoLifeQA (Table 2), EGOSCOPE raises the macro-average from EgoGraph's 46.9 to 48.4, with gains of 2.3 on EntityLog and 2.4 on RelationMap while remaining within 0.2 on TaskMaster. The category pattern indicates that the gain comes from joining complementary identity, state, and relation evidence across observations.

On VSI-Bench (Table 5), EGOSCOPE obtains 44.74 and exceeds matched Top-k, MMR, and DPP by 5.01, 2.83, and 3.28 points. With the pool, backbone, budget, and answer call fixed, this gain isolates view composition and reflects the value of unresolved spatial edges for selecting connective and directional observations.

4.4 ABLATION STUDIES

All ablations use the same candidate cache, frozen head, prompt, decoding procedure, and $K = 8$ budget. Trainable variants use three seeds, so their full-row averages may differ slightly from the single evaluations reported in the main tables. Table 3 evaluates relation closure, necessity–sufficiency objectives, and conflict modeling. Removing closure causes the largest degradation, reducing closure from 79% to 67%, increasing duplication from 11% to 24%, and lowering NLQ, EgoLifeQA, and VSI-Bench by 1.73, 1.53, and 1.61 points. The remaining objectives provide smaller complementary gains, while conflict modeling suppresses duplicate or contradictory evidence. The improvements therefore come from covering unresolved relations rather than selecting similar units.

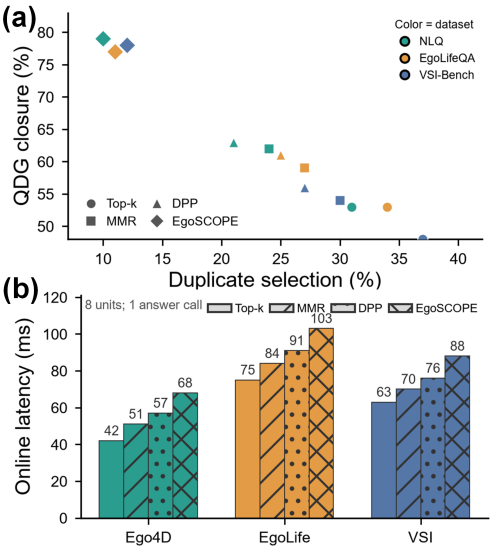

Figure 3: **Routing quality and cost.** QDG closure/duplication (upper-left is better) and router-selection latency at matched $K = 8$.

Table 4 keeps capacity and budget fixed while removing recurrent updates or disrupting typed relation identity. Static routing encodes the pool once, one-shot routing removes sequential acquisition, and permutation controls preserve relation counts while breaking semantic alignment. Table 6 identifies $(\tau, \rho) = (0.5, 0.2)$ as the best setting. Lower or higher τ leaves demands unresolved, whereas $\rho = 0$ removes redundancy suppression and $\rho = 0.4$ over-penalizes complementary views. This follows Eq. 7, where closure removes satisfied demands from marginal utility and redundancy regulates candidate competition.

In Table 4, static and one-shot routing reduce NLQ, EgoLifeQA, and VSI-Bench by $(1.19, 0.85, 1.07)$ and $(1.53, 1.24, 1.48)$ points, respectively. Relation permutations cause larger degradation despite matched capacity and budget, indicating that the state must encode which typed relation remains unresolved. Static routing saves $0.0014\,\mathrm{s}$, whereas one-shot routing saves $0.0221\,\mathrm{s}$ relative to the full $0.0863\,\mathrm{s}$ router. Thus, even the largest latency difference is only about two hundredths of a second.

Table 6: Sensitivity to closure threshold τ and redundancy weight ρ at $K = 8$.

τ	ρ	NLQ R1@0.5 ↑	EgoLifeQA ↑	VSI ↑
0.3	0.2	24.33±0.4	47.54±0.5	43.89±0.5
0.5	0.2	**25.08±0.3**	**48.32±0.4**	**44.78±0.4**
0.7	0.2	24.62±0.4	47.87±0.5	44.19±0.5
0.5	0.0	24.58±0.3	47.73±0.4	44.07±0.4
0.5	0.4	24.91±0.5	48.02±0.5	44.36±0.6

Together, Tables 3 and 4 form a progressive control chain. Removing closure reduces temporal localization, long-term memory, and spatial reasoning by 1.73, 1.53, and 1.61 points, while random relation-family assignment reduces them by 2.00, 1.85, and 2.10 points. The consistent pattern supports a shared missing-evidence bottleneck rather than task-specific gains, since all answer heads remain frozen (Appendix C.4).

Figure 3 connects routing quality with cost. EGOSCOPE achieves 77–79% closure and 10–12% duplicates, compared with 48–63% closure and 21–37% duplicates for Top-k, MMR, and DPP. Its online latency is 0.068, 0.103, and 0.088 s, compared with 0.042, 0.075, and 0.063 s for Top-k. This



Figure 4: **EgoSCOPE versus Top-k and MMR.** Under the same evidence budget, Top-k and MMR miss query-demand relations, whereas EgoSCOPE closes the required temporal and spatial relations and answers correctly on VSI-Bench and EgoLifeQA.

small 0.025–0.028 s overhead comes from K rounds of rescoring; under the same budget and one frozen-head call, the gains reflect complementary relation coverage rather than additional context or answer generation.

4.5 Qualitative Analysis

Figure 4 exposes two failure modes under matched K. On VSI-Bench, Top-k omits the transition that establishes heading and room connectivity, whereas EgoSCOPE suppresses repeated refrigerator views and acquires the turn connecting the printer and recycling bin. On EgoLifeQA, MMR spreads evidence across days but misses that "last" makes the final oven-removal event indispensable. The QDG preserves this unresolved qualifier until evidence supports both the action and its order. With backbone, evidence count, and answer calls fixed, both corrections isolate evidence composition.

5 Conclusion

Inspired by human problem solving—identifying required information, assessing what is known, and seeking what remains missing—we introduced EgoSCOPE, a fixed-budget router that progressively closes typed query demands while tracking accumulated support and conflict. Under matched candidate pools, budgets, frozen heads, and one answer call, it improves temporal localization, EgoLifeQA on the Jake-500 protocol, and spatial reasoning while increasing relation closure and reducing duplication. These results show that recurrent query-demand state converts a fixed evidence budget into a more sufficient set than relevance or generic diversity alone.

AI USE STATEMENT

We used generative AI tools to polish a small portion of the manuscript and to generate parts of the code. Apart from these uses, we did not use generative AI tools for any other tasks. The three authors manually surveyed the literature to check whether research ideas generated by LLMs involved potential plagiarism, and all LLM-generated code was verified and tested for correctness by the three authors. We take responsibility for the final content of this work, including the text and code generated with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Appendices C.2 and C.3 provide the additional qualitative cases. Appendices A–D document technical details, evaluation protocols, qualitative cases, extended results, routing diagnostics, and statistical analysis. The code records run configurations, environment metadata, predictions, route traces, errors, and table-ready aggregates, and generates the statistical plots from saved predictions.

REFERENCES

Joungbin An, Agrim Jain, and Kristen Grauman. UniversalVTG: A universal and lightweight foundation model for video temporal grounding. *arXiv preprint arXiv:2604.08522*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998. doi: 10.1145/290941.291025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella et al. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. doi: 10.1007/s11263-021-01531-2.

Kousik Das and Debaditya Roy. Imprint: Online memory compression for long-horizon egocentric QA. *arXiv preprint arXiv:2607.00696*, 2026.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12934–12943. IEEE, 2024. doi: 10.1109/cvpr52733.2024.01229.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. VideoZoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, 2026.

Mingkang Dong, Muxin Pu, Jie Li et al. ObjectStream: Latent objects as memory anchors for streaming video understanding. *arXiv preprint arXiv:2607.28312*, 2026.

Yisen Feng, Haoyu Zhang, Meng Liu, Weili Guan, and Liqiang Nie. Object-shot enhanced grounding network for egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24190–24200, 2025.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. In *International Conference on Learning Representations*, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01842.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

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

Zhijian Hou, Lei Ji, Difei Gao, Wanjun Zhong, Kun Yan, Chao Li, Wing-Kwong Chan, Chong-Wah Ngo, Nan Duan, and Mike Zheng Shou. GroundNLQ @ Ego4D natural language queries challenge 2023. *arXiv preprint arXiv:2306.15255*, 2023.

Jiacheng Hua, Yishu Yin, Yuhang Wu, Tai Wang, Yifei Huang, and Miao Liu. Unleashing spatial reasoning in multimodal large language models via textual representation guided reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13616–13637. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.620.

Minkuk Kim, Suyong Yun, Young Tae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. ReQuest: Rethinking-based question-aware frame selection for long-form video QA. In *Computer Vision – ECCV 2026*, volume 17052 of *Lecture Notes in Computer Science*, pp. 386–404. Springer, 2026. doi: 10.1007/978-3-032-37235-2_23.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044.

Bo Li, Yuanhan Zhang, Dong Guo et al. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan et al. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7575–7586. Curran Associates, Inc., 2022. doi: 10.52202/068431-0550.

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*, 2026. To appear.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244. Curran Associates, Inc., 2023. doi: 10.52202/075280-2004.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. TimeSearch-R: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. In *International Conference on Learning Representations*, 2026.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5285–5297, 2023.

Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18622–18632. IEEE, 2024. doi: 10.1109/cvpr52733.2024.01762.

Enxin Song, Wenhao Chai, Guanhong Wang et al. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Shitong Sun, Ke Han, Yukai Huang, Weitong Cai, and Jifei Song. EgoGraph: Temporal knowledge graph for egocentric video understanding. *arXiv preprint arXiv:2602.23709*, 2026.

Agnese Taluzzi, Riccardo Santambrogio, Simone Mentasti, Chiara Plizzari, and Matteo Matteucci. Graph it first! enabling reasoning on long-form egocentric videos through scene graphs. *arXiv preprint arXiv:2606.25842*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Computer Vision – ECCV 2024*, pp. 58–76. Springer, 2024. doi: 10.1007/978-3-031-72989-8_4.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3272–3283, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857. Curran Associates, Inc., 2024. doi: 10.52202/079017-0907.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643. IEEE, 2025a. doi: 10.1109/cvpr52734.2025.00994.

Jingkang Yang, Shuai Liu, Hongming Guo et al. EgoLife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28885–28900. IEEE, 2025b. doi: 10.1109/cvpr52734.2025.02690.

Hanrong Ye, Haotian Zhang, Erik Daxberger et al. MMEgo: Towards building egocentric multimodal LLMs for video QA. In *International Conference on Learning Representations*, 2025.

Sicheng Yu, Chengkai Jin, Huanyu Wang et al. Frame-Voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, 2025.

Tatiana Zemskova, Solomon Andryushenko, Ilya Obrubov, Viktoriia Khoruzhaia, Ekaterina Eroshenko, Ekaterina Derevyanka, and Dmitry Yudin. FocusGraph: Graph-structured frame selection for embodied long video question answering. *arXiv preprint arXiv:2603.04349*, 2026.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.

Enmin Zhong, Carlos R. del Blanco, Fernando Jaureguizar, and Narciso García. Hand trajectory fusion for egocentric natural language query grounding. Extended abstract presented at the Third Joint Egocentric Vision (EgoVis) Workshop, held in conjunction with CVPR 2026, 2026. URL `https://egovis.github.io/cvpr26/`. Listed under Accepted Extended Abstracts.

Jinguo Zhu, Weiyun Wang, Zhe Chen et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.I.R.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *International Conference on Learning Representations*, 2026.

APPENDIX

The appendix is organized as follows.

A. Technical Details

1. Notation and data flow
2. Candidate normalization
3. State update and masking
4. Query Demand Graph Schema
5. High-Recall Candidate Pools
6. Acquisition score details
7. Training Targets and Losses
8. Automatic Supervision Construction
9. Sequential Training Rollout

B. Evaluation Protocols and Implementation

1. Dataset and Evaluation Protocols
2. Implementation Details
3. Supervision and Parameter Sharing
4. Parser, Candidate-Pool, and Matched-Selector Audits

C. Additional Analyses

1. Main-Result Visualization
2. First-Appearance Ordering
3. Cross-Episode Habit Reasoning
4. Extended Ablations and Routing Diagnostics

D. Statistical Analysis

A TECHNICAL DETAILS

A.1 NOTATION AND DATA FLOW

For a query q, the candidate cache is constructed once and contains at most $M = 64$ evidence units. An evidence unit is an interval for Ego4D NLQ benchmark and a timestamped frame or short clip for EgoLifeQA and VSI-Bench. The router receives only cached features and metadata online. At step t, it reads the current selected mask and graph state, scores all unselected units, appends one unit, and updates the state. The frozen benchmark head is called once after the last acquisition; no answer-model output is fed back into routing.

A.2 CANDIDATE NORMALIZATION

Visual features are ℓ_2 normalized before concatenation. Start, end, center, and duration are divided by the video duration; displacement is clipped to the 99th percentile of the training set and then standardized. Missing optical-flow or egomotion values are replaced by a zero vector and a binary missingness flag. Candidate ties are broken by increasing center time so that route traces are deterministic.

Table 7: Shared QDG relation families and examples.

Family	Predicate examples	Typical evidence cue
Temporal/event	`before`, `last_event`, `at_interval`	order, duration, action
Entity/state	`identity`, `held_by`, `state_after`	object and state change
Spatial/view	`left_of`, `inside`, `connects`	viewpoint, layout, route
Quantity/metric	`count`, `distance`, `size`	scale and numeric cue
Causal/goal	`enables`, `used_for`, `goal`	purpose and consequence

A.3 State update and masking

Let s_e^t and k_e^t denote support and conflict for edge e after step t. We initialize both to zero and apply the noisy-OR update

$$s_e^{t+1} = 1 - (1 - s_e^t)(1 - p_{i_t e}^t), \tag{A.1}$$

$$k_e^{t+1} = 1 - (1 - k_e^t)(1 - \bar{p}_{i_t e}^t). \tag{A.2}$$

The selected mask is hard during inference and uses a straight-through mask only when computing the training-time ranking surrogate. Once selected, a unit cannot be selected again. The threshold τ for a closed edge is chosen on the validation split and is fixed for all test comparisons.

A.4 Query Demand Graph Schema

Each query is mapped to a small graph whose edges are typed, instance-specific demands. The parser emits the source span, predicate family, predicate name, target variable, optional temporal/spatial qualifier, and confidence. The vocabulary in Table 7 is shared across benchmarks; benchmark-specific predicates are listed as examples rather than exhaustive ontology claims.

For each edge, the automatic target constructor records positive support units, hard conflicts, and whether the edge is uniquely supported. If the official source fields cannot identify a relation, its loss term is masked. All relation targets used for optimization are constructed from official training annotations. The parser audit uses 500 queries per training benchmark drawn from the official training split and held out from parser-rule development. The task-level Oracle-QDG headroom is evaluated separately on the official reported evaluation splits. Predicted QDGs used in the primary evaluations are constructed without evaluation answers; oracle graphs are used only for this diagnostic and never for training or checkpoint selection.

A.5 High-Recall Candidate Pools

For the Ego4D NLQ benchmark, we union the top proposals from the released GroundNLQ and OSGNet systems, remove duplicate intervals using temporal IoU ≥ 0.95, and retain the highest-confidence 64 units. For EgoLifeQA, we sample causal timestamps up to the query time, add annotated event boundaries during training, and apply the same fixed-size cap at evaluation. For VSI-Bench, we cover the camera trajectory with temporally stratified units and retain all units needed to reproduce the official view sequence when the pool is smaller than 64.

A.6 Acquisition score details

The support and conflict heads are separate sigmoid classifiers, so $p_{ie}^t, \bar{p}_{ie}^t \in [0,1]$ but need not sum to one. A candidate may therefore contain both supporting and contradictory cues. Noisy-OR is used only as a bounded monotone state update; it is not interpreted as a calibrated independence model. Parser confidences are normalized within each query, $w_e \leftarrow w_e / (\sum_{e'} w_{e'} + \epsilon)$, so the closure term lies in $[0,1]$.

Let $g_{\text{suf}}(S, q) \in [0,1]$ be the pooled-set sufficiency head. The remaining score terms are

$$u_i^t = \sigma(g_{\text{task}}(\boldsymbol{h}_i^t, \boldsymbol{q})), \qquad n_i^t = \sigma(g_{\text{nec}}(\boldsymbol{h}_i^t, \boldsymbol{z}_t)), \tag{A.3}$$

$$z_i^t = [g_{\text{suf}}(S_t \cup \{i\}, q) - g_{\text{suf}}(S_t, q)]_+, \qquad d(i, S_t) = \max_{j \in S_t} [\cos(\boldsymbol{v}_i, \boldsymbol{v}_j)]_+, \tag{A.4}$$

$$f(i, S_t) = \max_{e \in \mathcal{E}_q} \bar{p}_{ie}^t. \tag{A.5}$$

We set $d(i, \varnothing) = 0$. Thus every component combined in Eq. (7) lies in $[0, 1]$. The instantaneous term f penalizes the candidate's strongest conflict, whereas accumulated $k_e(S_t)$ is part of $\boldsymbol{z}_t$ and changes all subsequent predictions. Conflict does not erase prior support; mixed evidence remains visible through the two state channels. Coefficients $(\alpha, \beta, \gamma, \delta, \rho, \eta)$ and τ are selected on validation data and then frozen. We use $(1.0, 0.3, 0.5, 0.25, 0.2, 0.5)$ and $\tau = 0.5$.

A.7 Training Targets and Losses

We selected the five loss weights by joint grid search on validation data (Table 8) and fixed the best mean-validation setting for all test comparisons.

Table 8: Loss-weight grid and selected values.

Parameter	Search grid	Selected value
λ_{task}	$\{0.5, 1.0, 2.0\}$	1.0
λ_{close}	$\{0.5, 1.0, 2.0\}$	1.0
λ_{nec}	$\{0.1, 0.2, 0.5\}$	0.2
λ_{suf}	$\{0.05, 0.1, 0.2\}$	0.1
λ_{conf}	$\{0.05, 0.1, 0.2\}$	0.1

Positive candidate–edge pairs are intervals or views automatically aligned with an official training annotation and a parsed evidence relation; negatives are sampled from the same video and are matched for temporal distance where possible. Necessity labels are computed by deleting one candidate from an annotated positive set and testing whether the corresponding edge loses unique support. Sufficiency positives are complete annotated sets; hard negatives have equal cardinality and high task score but leave at least one required edge below τ.

The task ranking loss is a pairwise logistic loss over positive and negative units. Closure, conflict, necessity, and sufficiency use masked binary cross-entropy. We normalize each term by the number of observed labels in the mini-batch, and use the weights

$$(\lambda_{\text{task}}, \lambda_{\text{close}}, \lambda_{\text{nec}}, \lambda_{\text{suf}}, \lambda_{\text{conf}}) = (1.0, 1.0, 0.2, 0.1, 0.1). \tag{A.6}$$

No label is silently converted to a negative example. Relation-label permutation and random family assignment are evaluation-only controls.

A.8 Automatic Supervision Construction

All routing targets are generated from official training-split annotations and the frozen QDG parser. We do not collect additional relation annotations. For each query, the constructor first aligns the cached candidate units with the available temporal, event, entity, or view fields. It then emits five masks and targets: task relevance, candidate–edge support, candidate–edge conflict, leave-one-out necessity, and set sufficiency. A target is masked when the source benchmark does not expose the field required to construct it. Table 9 gives the deterministic benchmark mappings.

Table 9: Automatic construction of router supervision. Only official training annotations are used; no validation/test answer or manually added relation label enters target generation.

Benchmark	Source fields	Task/support targets	Conflict targets	Necessity/sufficiency targets
Ego4D NLQ v2	Query text, annotated target interval, proposal intervals and confidence	Candidate relevance follows overlap with the annotated interval; support is assigned to parsed temporal/event edges aligned with that interval	Same-video non-target occurrences and qualifier-violating order swaps form hard conflicts when the parsed edge distinguishes them	Leave-one-out removal marks unique interval support; a set is sufficient when every required edge retains positive aligned support
EgoLifeQA	Training query and answer, query time, event boundaries, timestamps, and available entity/action metadata	Pre-query candidates matching the parsed entity, event, state, or temporal qualifier receive task and edge support	Entity/action substitution and first/last or before/after reversal within the same recording produce matched hard conflicts	Deletion tests unique event support; complete causal evidence sets are positive and equal-size high-score sets missing one edge are negative
VSI-Bench	Query and trajectory candidates at evaluation only	No VSI answer, relation, necessity, or sufficiency target is used. The jointly trained Ego4D/EgoLifeQA router is applied without VSI supervision.		

The constructor is deterministic after candidate normalization and QDG parsing. Candidate–edge labels are stored with an observed-label mask, so an absent field does not become a negative. Table 10 reports the resulting artifact-level audit from the saved training manifests.

Table 10: Coverage audit for automatically constructed supervision from the saved training manifests.

Training source	Queries	QDG edges	Candidate–edge pairs	Support positives	Conflict positives	Sufficient sets	Masked labels (%)
Ego4D NLQ v2	13,849	24,426	1,563,264	45,291	362,038	13,087	38.4
EgoLifeQA	33,352	42,885	2,744,640	185,536	126,982	29,970	51.8
VSI-Bench	0	0	0	0	0	0	100

Across the two training benchmarks, the constructor produces 4.31 million candidate–edge pairs from 47,201 queries. The different support, conflict, and mask rates reflect the official fields available in each benchmark rather than additional relation annotation. VSI contributes no router target, in accordance with the frozen-transfer protocol.

A.9 Sequential Training Rollout

Training follows the model's own route rather than an annotated ordering. At step t, the forward pass uses the hard greedy one-hot mask $\boldsymbol{m}_t^{\mathrm{hard}}$ from Algorithm 1. The backward pass uses the straight-through mask

$$\boldsymbol{m}_t^{\mathrm{ST}} = \mathrm{stopgrad}(\boldsymbol{m}_t^{\mathrm{hard}} - \boldsymbol{\pi}_t) + \boldsymbol{\pi}_t, \qquad \boldsymbol{\pi}_t = \mathrm{softmax}(\{A(i \mid S_t, q)\}_{i \notin S_t}), \tag{A.7}$$

so the realized route remains discrete while gradients reach earlier scoring decisions. No teacher forcing or answer-head output is used.

Algorithm 2 Training-time recurrent route construction

Require: Training query q, official annotations B, candidate pool $\mathcal{C}$, budget K
1: Parse q into G_q and construct masked targets from $(B, G_q, \mathcal{C})$
2: Initialize $S_0 \leftarrow ()$, support/conflict state to zero, and $\mathcal{L} \leftarrow 0$
3: **for** $t = 0, \ldots, K-1$ **do**
4: Predict support, conflict, task, necessity, and sufficiency scores for all $i \notin S_t$
5: Add the masked candidate–edge, task, necessity, and sufficiency losses at state S_t
6: Form $\boldsymbol{\pi}_t$, select the hard greedy unit in the forward pass, and use $\boldsymbol{m}_t^{\mathrm{ST}}$ in the backward pass
7: Append the hard selection to S_t and update support/conflict by noisy-OR
8: **end for**
9: Normalize each loss by its observed targets, combine them using Eq. 8, and update router parameters only
10: **return** Router loss and the realized training route

Local candidate–edge and task losses are evaluated over all unselected units at each visited state. Necessity uses the automatically constructed leave-one-out target. Sufficiency compares the realized prefix augmented by a candidate with the matched equal-size hard-negative sets. Encoders, candidate pools, and answer heads remain frozen throughout this rollout.

B Evaluation Protocols and Implementation

B.1 Dataset and Evaluation Protocols

Table 11: Dataset and protocol summary for the evaluated benchmarks.

Benchmark	Videos	Questions	Candidate cap M	Main K
Ego4D NLQ benchmark v2	2,018	4,552	64	8
EgoLifeQA	1 participant	500	64	8
VSI-Bench	288	5,130	64	8

Table 11 records the evaluated scale and fixes the candidate and evidence budgets used throughout the controlled comparisons. For Ego4D NLQ benchmark, we report R@1 and R@5 at IoU 0.3 and 0.5 using the official NLQ v2 evaluator. For EgoLifeQA, we use the public 500-question evaluation set specified in the main paper and report five category accuracies and their macro-average; all evidence timestamps satisfy $t \leq t_q$. For VSI-Bench, we use the benchmark numeric normalization and choice decoding and report the average over all questions. Every controlled selector sees the same cached candidate pool, frozen backbone, prompt, decoding, and one answer-model call.

B.2 Implementation Details

Table 12: Router hyperparameters and runtime settings.

Item	Value
Visual encoder	SigLIP2, 1,152-dimensional frozen features
Set encoder	2 Transformer layers, width 256, 4 heads
Trainable parameters	approximately 2.6M
Optimizer	AdamW, learning rate 10^{-4}, weight decay 10^{-4}
Epochs / early stopping	10 maximum / patience 5
Seeds	42, 43, 44
Candidate cap / budget	$M = 64$ / $K = 8$ (sweep $K \in \{2, 4, 8, 16, 32\}$)
Hardware	24GB RTX 3090 GPUs

Table 12 lists the shared router configuration used for the three benchmarks. Features and proposal pools are cached once per split. Training uses mixed precision where supported, gradient clipping at norm 1.0, and deterministic candidate tie-breaking. We save the configuration, seed, checkpoint hash, predictions, route traces, per-query metrics, and evaluator version for every run. The anonymous code package includes a script that reconstructs every table from these saved artifacts.

B.3 Supervision and Parameter Sharing

Table 13 makes the cross-benchmark protocol explicit. One router parameter set is optimized jointly on the Ego4D NLQ and EgoLifeQA training splits. Sampling is balanced by dataset within each training batch. The QDG vocabulary and parser rules are frozen after development on training queries. Validation queries contribute no router gradients and are used for checkpoint selection. The reported Ego4D result uses the same official validation split, whereas EgoLifeQA is reported separately on Jake-500. The selected checkpoint is transferred to VSI-Bench without further optimization.

Table 13: Data and supervision used by the shared router. "Auto" denotes targets deterministically derived from official training annotations.

Benchmark split / role	Answer/task target	Relation targets	Parser rules	Router gradients	Checkpoint selection	Reported evaluation
Ego4D NLQ training	Official / Auto	Auto	Training queries only	Yes	No	No
Ego4D NLQ validation	Official (evaluation)	No	Frozen	No	Yes	Yes (Table 1)
EgoLifeQA training	Official / Auto	Auto	Training queries only	Yes	No	No
EgoLifeQA validation	Official (selection)	No	Frozen	No	Yes	No
EgoLifeQA Jake-500 evaluation	Official (evaluation)	No	Frozen	No	No	Yes (Table 2)
VSI-Bench evaluation	No	No	Frozen transfer	No	No	Official test protocol

The benchmark task heads remain separate and frozen because their output spaces differ. "Shared router" therefore means one set of set-encoder, edge-encoder, support/conflict, task, necessity, and sufficiency parameters, rather than only a shared architecture. No VSI example contributes a gradient or checkpoint selection signal.

B.4 Parser, Candidate-Pool, and Matched-Selector Audits

The upper panel of Table 14 evaluates the QDG parser on 500 queries per training benchmark drawn from the official training split and held out from parser-rule development. We canonicalize aliases before exact matching and score the emitted source span, predicate family and name, target, and qualifier. Edge precision, recall, and F1 use exact tuple matching; family and qualifier accuracy are conditional on a matched source–target pair; graph exact match requires the complete edge set to match. The lower panel is a separate task-level headroom analysis on the official evaluation splits used by the main result tables. Its Predicted-QDG row therefore reproduces those results. Oracle-QDG replaces only the predicted graph; candidates, router weights, budget, task head, and evaluator remain fixed. Oracle graphs are unavailable for VSI-Bench.

Table 14: QDG parser audit and oracle-edge headroom. The upper panel reports parser metrics on 500 training-split queries per source benchmark held out from parser-rule development. The lower panel reports task performance on the official evaluation splits used in the main result tables.

Dataset	Queries	Edge P	Edge R	Edge F1	Family acc.	Qualifier acc.	Graph exact match
Ego4D NLQ v2	500	86.8	82.9	84.8	94.1	96.3	68.6
EgoLifeQA	500	83.1	77.4	80.1	91.2	93.5	57.8

QDG source	Ego4D R@1@0.5	EgoLifeQA macro acc.	VSI-Bench avg.
Predicted QDG	25.07	48.4	44.74
Oracle QDG	26.88	50.36	not applicable

On the training-split audit subsets, the parser attains 84.8/80.1 edge F1 and 68.6/57.8 graph exact match on Ego4D/EgoLifeQA, with higher family and qualifier accuracy. On the official evaluation splits, replacing predicted graphs with oracle graphs improves the two task metrics by 1.81 and 1.96 points. This diagnostic estimates parser-related headroom; oracle graphs are not used in the primary results.

For Ego4D, candidate-pool recall is computed before routing. A query is covered when at least one of the retained $M = 64$ intervals overlaps the target at the specified IoU. This distinguishes proposal coverage from reranking quality for the two source pools and their deduplicated union.

Table 15: Ego4D NLQ candidate-pool recall before routing.

Candidate pool	M	Recall@IoU 0.3	Recall@IoU 0.5
GroundNLQ proposals	64	89.2	78.1
OSGNet proposals	64	92.4	82.0
Deduplicated union used by EGOSCOPE	64	**95.3**	**86.1**

The deduplicated union covers 95.3% and 86.1% of queries at IoU 0.3 and 0.5, improving on the stronger individual OSGNet pool by 2.9 and 4.1 points. All selector comparisons below receive this same union, separating reranking quality from proposal coverage.

Table 16 separates published full-system context from causal selector comparisons. Every row uses the same cached pool, features, $K = 8$, frozen answer head, prompt, decoding, and one answer call. The relation-free learned ranker receives the query embedding but no QDG edge or graph state. The recurrent no-QDG control retains recurrent rescoring and parameter count, replacing typed edge state with an untyped pooled state.

Table 16: Matched-pool downstream comparison. Every selector uses the same candidate pool, cached features, $K = 8$ budget, frozen head, prompt, decoding, and one answer call.

Selector	Ego4D R@1@0.5	EgoLifeQA macro acc.	VSI-Bench avg.
Top-k relevance	22.97	47.00	39.73
MMR	23.64	47.42	41.91
DPP	23.78	47.56	41.46
Learned query-only ranker	23.46	47.18	42.62
Recurrent selector without QDG edges	23.86	47.49	43.35
EGOSCOPE	**25.07**	**48.4**	**44.74**

Under this controlled protocol, EGOSCOPE exceeds the strongest alternative on Ego4D, EgoLifeQA, and VSI-Bench by 1.21, 0.84, and 1.39 points, respectively. The learned query-only and recurrent no-QDG controls show that neither learned relevance nor recurrence alone explains the gain; retaining typed unresolved relations is consistently beneficial.

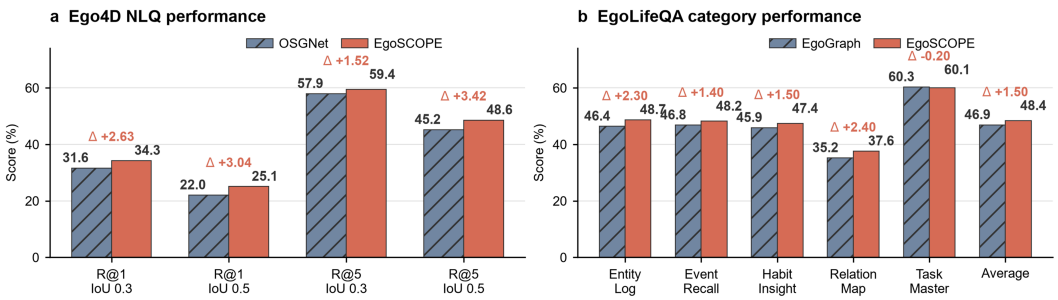

Figure 5: Main-result margins on Ego4D NLQ and EgoLifeQA. Bars reproduce the corresponding main-paper values; annotations show absolute percentage-point changes over OSGNet or EgoGraph.

C Additional Analyses

C.1 Main-Result Visualization

Figure 5(a) shows that the NLQ improvement is not confined to one operating point. Relative to OSGNet, EgoSCOPE improves R@1 by 2.63 and 3.04 points at IoU 0.3 and 0.5, and improves R@5 by 1.52 and 3.42 points. The larger gains at the stricter IoU threshold indicate that routing helps retain evidence for the temporally decisive occurrence rather than only adding loosely overlapping proposals.

Figure 5(b) gives the corresponding category view for EgoLifeQA. EgoSCOPE improves EntityLog, EventRecall, HabitInsight, and RelationMap by 2.3, 1.4, 1.5, and 2.4 points, respectively, while remaining within 0.2 points on TaskMaster; the macro-average rises by 1.5 points. The largest gains on identity and relation questions are consistent with the router allocating its fixed budget to unresolved entity/state and spatial edges. The near-tie on TaskMaster also shows that the average improvement is not produced by uniformly shifting every category.

The following cases complement the aggregate results with query-level traces. They use the same cached candidate pool, $K = 8$ evidence budget, and one frozen answer-model call as the main experiments. The comparisons make explicit why query-conditioned relation closure is different from either similarity-only ranking or generic diversity.

C.2 First-Appearance Ordering

Figure 6 asks for the order in which the bed, shoes, laptop, and toilet first appear. Top-k keeps highly similar bedroom and desk views but misses the first shoe-rack observation, so it relies on a late revisit and predicts the wrong order. MMR covers more places, yet its generic diversity objective also retains that late revisit and does not represent first-time appearance as a typed temporal demand. In contrast, EgoSCOPE creates explicit order edges, selects evidence near 0.00s, 30.52s, 38.15s, and 129.71s, and suppresses redundant later views. Closing the four ordered edges yields the correct answer (A), showing that temporal coverage–not merely more varied views–is the decisive factor.

C.3 Cross-Episode Habit Reasoning

Figure 7 asks what is usually done after finishing in the kitchen. Top-k selects adjacent sink frames from one bathroom episode and confuses the dispenser with dish soap, while MMR spreads views across rooms without linking the kitchen trigger to the subsequent action and misses the repeated habit. EgoSCOPE instead represents the trigger–action edge, chooses separated observations (including DAY4 18:10:38:11, DAY4 21:28:10:19, and DAY4 21:35:01:04), and masks near-duplicate frames from the same episode. Evidence from multiple independent episodes closes the habit relation and supports the correct answer (C), illustrating how recurrent state turns scattered personal memories into a sufficient explanation.

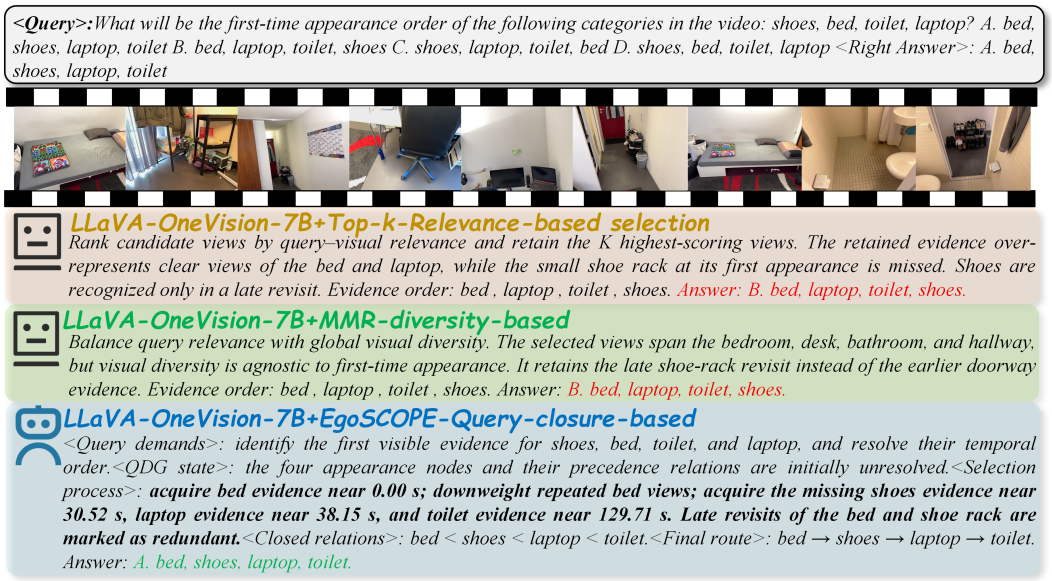

Figure 6: Qualitative case for a first-appearance ordering query.

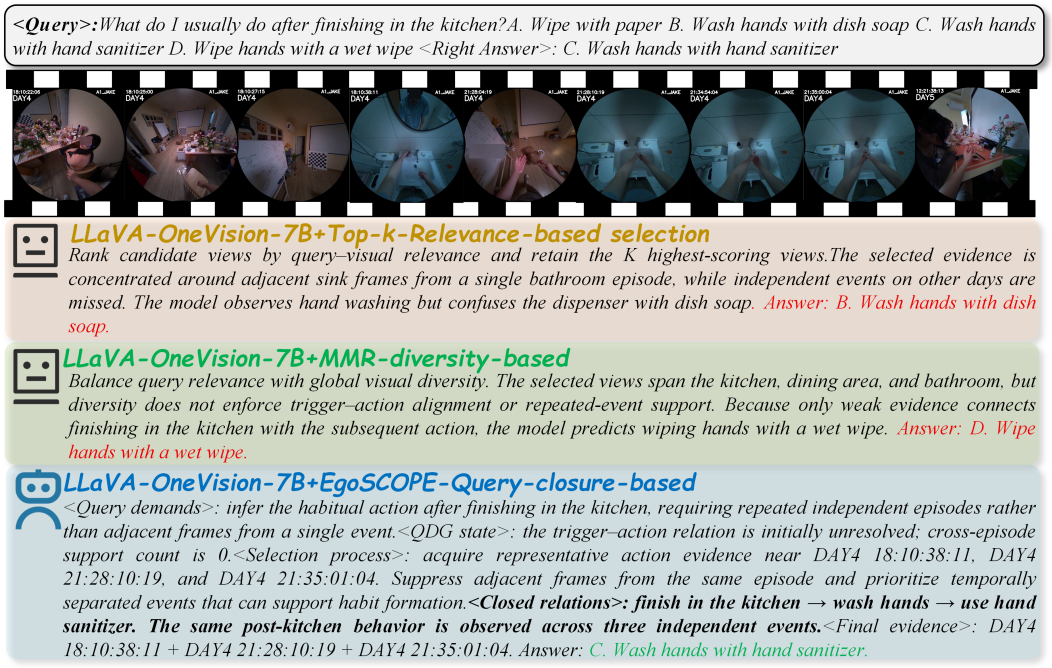

Figure 7: Qualitative case for a cross-episode habit query.

C.4 Extended Ablations and Routing Diagnostics

The main paper reserves two concise tables for the component and state ablations. This section reports the extended grid, including routing diagnostics, budget scaling, paired tests, latency, and backbone transfer.

Table 17 separates accuracy changes from routing behavior and runtime. Removing closure produces the largest joint failure. NLQ, EgoLifeQA, and VSI-Bench decrease by 1.73, 1.53, and 1.61 points, closure falls from 79% to 67%, and duplication rises from 11% to 24%, despite saving only 2.1 ms. Necessity, sufficiency, and conflict modeling provide smaller but consistent improvements, with conflict modeling chiefly reducing the remaining duplicates. Static state and one-shot ranking

Table 17: Extended component and decision-mechanism ablation. QDG closure and duplicate rate summarize the pooled held-out routing audit; all entries are mean±standard deviation over seeds $\{42, 43, 44\}$.

Variant	NLQ R@1@0.5	EgoLifeQA macro acc.	VSI avg.	Closure	Duplicate rate	Router latency (s) ↓
w/o closure	23.35±0.3	46.79±0.6	43.17±0.6	67.00±1.8	24.00±1.5	0.0842±0.0025
w/o necessity	24.24±0.5	47.72±0.7	44.23±0.5	74.00±1.6	16.00±1.2	0.0852±0.0039
w/o sufficiency	24.51±0.4	48.14±0.5	44.25±0.6	75.00±1.5	15.00±1.1	0.0846±0.0021
Static state	23.89±0.5	47.47±0.7	43.71±0.6	72.00±1.3	18.00±1.8	0.0849±0.0039
One-shot ranking	23.55±0.6	47.08±0.8	43.30±0.7	69.00±1.6	22.00±1.4	**0.0642±0.0033**
w/o conflict	24.88±0.4	48.16±0.6	44.47±0.5	78.00±1.5	14.00±1.1	0.0851±0.0026
Full EGOSCOPE	**25.08±0.3**	**48.32±0.4**	**44.78±0.4**	**79.00±1.4**	**11.00±1.0**	0.0863±0.0041

lower closure by 7 and 10 points and raise duplication by 7 and 11 points, respectively. The one-shot variant saves 22.1 ms but loses on every benchmark, quantifying the accuracy–latency trade-off introduced by recurrent rescoring.

Table 18: Evidence-budget sweep from the latest statistical result table. Entries are mean±std over the reported runs.

Benchmark	$K = 2$	$K = 4$	$K = 8$	$K = 16$	$K = 32$
Ego4D NLQ benchmark R@1@0.5	21.34±0.50	23.15±0.40	25.08±0.30	25.56±0.30	25.63±0.40
EgoLifeQA macro acc.	46.42±0.70	47.52±0.60	48.32±0.40	48.73±0.50	48.91±0.60
VSI-Bench avg.	42.19±0.70	43.37±0.60	44.78±0.40	45.09±0.40	45.14±0.50

Table 18 shows a common knee at the main $K = 8$ budget. Increasing K from 2 to 8 improves NLQ, EgoLifeQA, and VSI-Bench by 3.74, 1.90, and 2.59 points, whereas increasing it from 8 to 32 adds only 0.55, 0.59, and 0.36 points. The consistent saturation indicates that the main result is not obtained by exposing the answer model to an unusually large evidence set; most of the useful relations are already closed with eight selected units.

Table 19 resolves the pooled routing summary into individual benchmarks. EGOSCOPE attains 77–79% closure with 10–12% duplicates, compared with 48–63% closure and 21–37% duplicates for Top-k, MMR, and DPP. The separation persists across temporal localization, personal-memory QA, and spatial reasoning, supporting a benchmark-independent effect of state-conditioned relation closure rather than a dataset-specific ranking artifact.

Table 19: Query–relation closure and duplicate-selection rate from the latest statistical result table.

Dataset	Method	Closure (%) ↑	Duplicate rate (%) ↓
Ego4D NLQ benchmark	Top-k	53	31
	MMR	62	24
	DPP	63	21
	EGOSCOPE	79	10
EgoLifeQA	Top-k	53	34
	MMR	59	27
	DPP	61	25
	EGOSCOPE	77	11
VSI-Bench	Top-k	48	37
	MMR	54	30
	DPP	56	27
	EGOSCOPE	78	12

Table 20: Paired per-query gains over Top-k with nonparametric bootstrap 95% confidence intervals.

Dataset	N	Mean gain (pp)	95% CI (pp)
Ego4D NLQ benchmark	500	2.10	[1.00, 3.20]
EgoLifeQA	500	1.40	[0.40, 2.50]
VSI-Bench	1000	4.99	[4.10, 5.88]

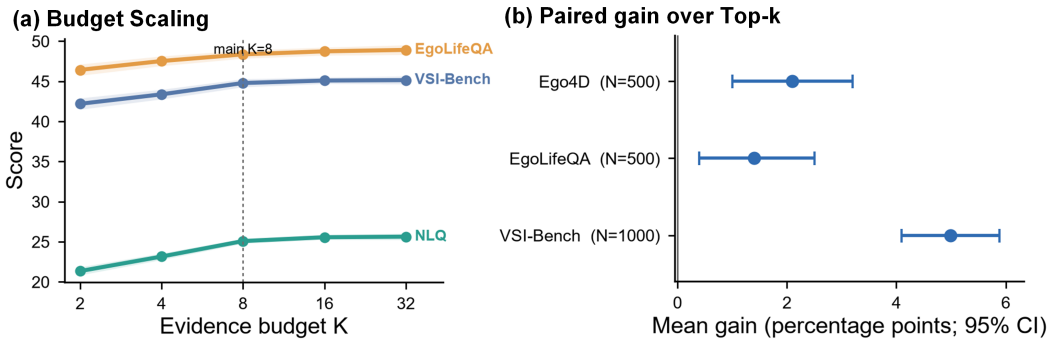

Figure 8: Evidence-budget scaling and paired gains over Top-k. Panel (a) marks the main $K = 8$ operating point; panel (b) shows mean percentage-point gains with query-level bootstrap 95% confidence intervals.

Table 21: Router-selection latency with eight selected evidence units. Times exclude offline feature extraction and frozen-head inference; all downstream evaluations use one answer call.

Dataset	Method	Evidence units	Router latency (ms)	Answer calls
Ego4D NLQ benchmark	Top-k	8	42	1
	MMR	8	51	1
	DPP	8	57	1
	EGOSCOPE	8	68	1
EgoLifeQA	Top-k	8	75	1
	MMR	8	84	1
	DPP	8	91	1
	EGOSCOPE	8	103	1
VSI-Bench	Top-k	8	63	1
	MMR	8	70	1
	DPP	8	76	1
	EGOSCOPE	8	88	1

Figure 8 visualizes the two complementary controls in Tables 18 and 20. Panel (a) makes the diminishing return beyond $K = 8$ visible on every benchmark. Panel (b) shows paired gains of 2.10 points on NLQ, 1.40 on EgoLifeQA, and 4.99 on VSI-Bench; the corresponding intervals $[1.00, 3.20]$, $[0.40, 2.50]$, and $[4.10, 5.88]$ all exclude zero. Thus the aggregate improvements are also supported at the query level and are not explained by a small number of unpaired examples.

Table 21 holds the evidence count and answer-call count fixed. Relative to Top-k, recurrent routing adds 26 ms on NLQ, 28 ms on EgoLifeQA, and 25 ms on VSI-Bench. This bounded overhead is consistent with K rounds of state updates and candidate rescoring. Because no method receives more than eight units or more than one answer call, the accuracy gains cannot be attributed to a larger answer-model context or iterative answer generation.

Table 22: Backbone transfer and matched answer-call control.

Backbone	EgoLifeQA macro acc.	VSI-Bench avg.	Answer calls
LLaVA-OneVision-7B	48.32±0.4	44.78±0.4	1
Qwen2.5-VL-7B	48.90±0.3	45.30±0.4	1
InternVL3-8B	48.70±0.6	45.10±0.5	1

Table 22 shows that the routing result is not tied to the primary LLaVA-OneVision answer model. Under the same single-call constraint, Qwen2.5-VL improves the absolute EgoLifeQA/VSI-Bench scores by 0.58/0.52 points and InternVL3 by 0.38/0.32 points relative to the primary backbone. The small spread across all three backbones is consistent with the router operating as an evidence-composition interface rather than exploiting a model-specific decoding behavior.

D Statistical Analysis

All confidence intervals are query-level, paired, nonparametric bootstrap intervals with 10,000 resamples and a fixed seed. We report the mean gain in percentage points, the 2.5th and 97.5th percentiles, and the number of valid paired queries. A route is counted as a duplicate when its selected unit has temporal IoU or visual cosine similarity above the validation threshold with an earlier selected unit. Run manifests retain dataset versions, split files, configuration and checkpoint hashes, predictions, route traces, per-query metrics, and evaluator versions; the anonymous package reconstructs all tables and figures from these saved artifacts.