TRANSPORT WHAT MATTERS: EVIDENCE-DIRECTED OPTIMAL TRANSPORT FOR LONG VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long video understanding often requires reducing thousands of observations to a small frame budget that a vision–language model can process. Existing selectors prioritize query relevance, diversity, or adaptive sampling, but leave implicit how evidence from discarded observations should be represented by the retained set. We introduce Evidence-Directed Optimal Transport (E-DOT), a training-free framework that casts frame selection as transporting task-conditioned evidence mass from dense video observations to a sparse support of frames. Query relevance over dense observations and temporal extent define the source mass. A directed ground cost combines temporal proximity with sparse local-helpfulness estimates from a frozen model: transport to a nearby frame that preserves or improves source evidence is cheap, whereas losses in relevance or helpfulness are penalized. Relaxing the destination marginal lets multiple sources share representatives, allowing the support to concentrate on evidence that remains poorly covered. The resulting support-selection objective admits monotone-submodular greedy optimization and coverage characterizations. Across four long-video question answering benchmarks and three answerers, E-DOT outperforms all compared frame selectors under a fixed answerer budget, while matched ablations isolate the contributions of local helpfulness, directed transport geometry, and coverage-aware support selection.

1 INTRODUCTION

Long video understanding asks multimodal models to reason over minutes or hours of visual content (Fu et al., 2025; Wang et al., 2025a), often far beyond the number of frames that can be processed at once. A practical system must therefore compress the video before answering. Frame selection has become a central way to allocate this limited visual budget: the selector decides which observations are worth passing to the answerer, and the answerer then reasons only over this retained subset.

Recent selectors follow several broad strategies. Relevance-based methods prioritize frames or clips that match the query, for example through query–frame similarity (Liu et al., 2025; Zhang et al., 2025b) or query-aware search (Zhu et al., 2026). Diversity- and coverage-aware methods (Tang et al., 2025; Sun et al., 2025) instead seek to reduce redundant selections while retaining broader information from the video. Other approaches adapt the observation budget to temporal structure, for example through recursive subdivision (Tang et al., 2025), clip-level allocation (Chen et al., 2026), or region-specific selection (Xu et al., 2026). These approaches have substantially improved long-video question answering (QA), but they mostly formulate selection in terms of candidate utility or the diversity of the retained set. Under a fixed budget, however, preserving information from a discarded observation depends on whether the retained set can stand in for it, and this replacement relation is usually left implicit. As a result, two candidates may be treated as merely different even when one is a strictly better substitute for the other for the current question, leaving the choice largely arbitrary.

Figure 1 illustrates why this distinction matters. Nearby frames may preserve much of the same scene while differing in the evidence they expose. If one preserves the useful evidence of another while revealing more, it can replace the weaker observation cheaply, but the reverse need not hold: replacement can be directional. Frame selection is therefore not only about which frames are individually useful, but also about which discarded observations the retained set can safely represent.

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.

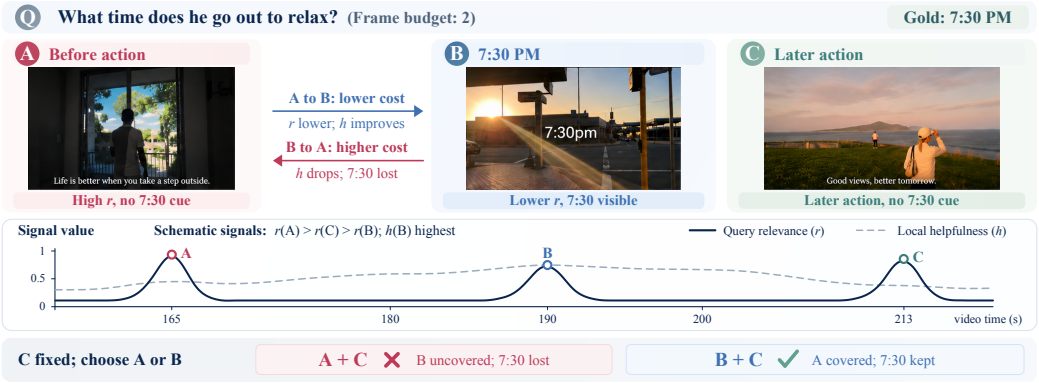


Figure 1: Illustrative selection under a two-frame budget. A and B are depicted as highly relevant, but only B reveals the queried 7:30 PM cue. With C fixed, replacing A with B gains this answer evidence, whereas replacing B with A loses it. Replacement is thus directional, unlike a symmetric distance.

To this end, we propose Evidence-Directed Optimal Transport (E-DoT), a replacement-aware frame selection method that explicitly models how retained observations can stand in for discarded ones under a fixed frame budget. E-DoT captures directional replacement through an evidence-directed ground cost and turns these pairwise relations into a global selection through source-weighted coverage. Each candidate is described by normalized time, query relevance, and a local helpfulness estimate obtained from frozen models. For a source and a candidate representative, the cost is small when the candidate is temporally nearby and preserves or improves the source's evidence signals, while deficits in relevance or helpfulness are penalized. This separates *source importance*, which determines how strongly an observation should be preserved, from *replacement quality*, which measures how well a retained candidate can represent it. Given a selected set, each source is charged only the cost of its best representative. Candidates that mainly serve already well-represented sources therefore add little benefit, whereas those that reduce residual loss elsewhere become more valuable. With a relaxed destination marginal (Gurumoorthy et al., 2021), multiple sources can share a representative, and the transport objective reduces exactly to source-weighted nearest-representative loss. Its complementary benefit is monotone submodular and can be optimized greedily (Nemhauser et al., 1978). We evaluate E-DoT on Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2025), and LVBench (Wang et al., 2025a) under a fixed answerer frame budget. Across three answerers, the full selector outperforms every compared selector in all twelve answerer–benchmark settings, and paired tests confirm three of the four gains for Qwen2.5-VL-7B.

Our contributions are summarized as threefold:

- We identify directional replacement as a missing relation in budgeted long-video frame selection and formulate an evidence-directed ground cost: nearby candidates that preserve or improve task-conditioned evidence signals can replace weaker observations without being penalized for the improvement, whereas the reverse replacement can be much more costly.
- We introduce E-DoT, a training-free selector that combines directed replacement with source-weighted coverage. The objective has an equivalent free-target sparse optimal-transport form, admits standard submodular greedy optimization with a $(1 - 1/e)$ guarantee, and supports coverage and worst-case-gap characterizations over the evidence coordinates.
- We evaluate the method on four long-video QA benchmarks with three answerers and use matched ablations to isolate directionality, local helpfulness, and coverage updating.

2 Related Work

Frame selection for long video understanding. Query-aware selectors prioritize frames or clips that match the question. BOLT (Liu et al., 2025) and Q-Frame (Zhang et al., 2025b) use frame–query relevance to guide selection, AKS (Tang et al., 2025) combines relevance with adaptive temporal partitioning, and Frame-Voyager (Yu et al., 2025) learns to query informative frame combinations.

Other methods explicitly control redundancy or allocate observation effort across time. MDP^3 (Sun et al., 2025) couples a determinantal point process for relevance and list-wise diversity with a Markov decision process for sequential dependence; FOCUS (Zhu et al., 2026) concentrates scoring on promising or uncertain clips; and WFS-SB (Chen et al., 2026) and RIDGE (Xu et al., 2026) exploit structure in the relevance signal to allocate temporal sampling. Agent- and tree-based methods such as VideoAgent (Wang et al., 2024) and VideoTree (Wang et al., 2025c) instead search or hierarchically cluster frames with language-model guidance, while video summarization selects representative and diverse frames without a query, for example through sequential determinantal point processes (Gong et al., 2014) or submodular mixtures (Gygli et al., 2015). These methods improve relevance, coverage, observation efficiency, or query-agnostic representativeness. E-DOT instead asks which discarded observations each retained frame can safely represent under a task-conditioned replacement cost.

Optimal transport and asymmetric representation. Optimal transport (OT; Villani, 2009; Peyré & Cuturi, 2019) provides a principled way to map weighted observations to a compact representative support. SPOT (Gurumoorthy et al., 2021) casts prototype selection as sparse OT with free prototype weights and connects the resulting objective to submodular optimization, while CoverPrune (Ling et al., 2026) applies transport-based coverage to visual token pruning in 3D vision–language models, using feature, spatial, and temporal discrepancies with capacity modeling. Directed relations have also been modeled through order embeddings (Vendrov et al., 2016), which use coordinate-wise order violations to represent visual–semantic hierarchies such as entailment. Building on these ideas, E-DOT defines a temporally constrained one-sided ground cost and jointly chooses a sparse support of real frames. Free destination masses reduce the inner transport problem to weighted nearest-representative coverage, while the asymmetric cost distinguishes evidence gains from losses.

3 EVIDENCE-DIRECTED OPTIMAL TRANSPORT

Given a video V, a question q, and a frame budget K, E-DOT selects K real frames for a fixed answerer. Rather than treating frame selection only as a ranking problem, E-DOT separates two roles. *Source importance* determines how strongly each observation should be preserved under compression, while *replacement quality* measures how safely a retained frame can stand in for it. We encode the first through a task-conditioned source mass and the second through a directed ground cost. These pairwise relations then induce a global sparse-support objective. Figure 2 summarizes the method.

3.1 TASK-CONDITIONED EVIDENCE SIGNALS

Let $0 \le f_1 < \cdots < f_N \le T-1$ denote the candidate frame indices of a video with $T > 1$ native frames, where $1 \le K \le N$. Candidate i is represented by

$$x_i = (t_i, r_i, h_i) \in [0,1]^3, \qquad t_i = \frac{f_i}{T-1}. \tag{1}$$

Here t_i is normalized time, r_i is global query relevance, and h_i is local answer helpfulness. Relevance provides broad query alignment, while helpfulness estimates whether nearby visual context exposes evidence useful for answering the question. These coordinates serve as compact, task-conditioned proxies for replacement quality rather than as a complete semantic description of each video frame.

Global relevance. A frozen BLIP scorer (Li et al., 2022) provides query-relevance measurements over the candidate timeline, using the question and answer choices when available. The resulting coordinate $r_i \in [0,1]$ is a broad search signal rather than a measure of evidence sufficiency: a highly matching observation may still omit the specific event or visual detail needed for a correct answer.

Sparse local helpfulness. Relevance alone does not indicate whether a local region exposes answer-bearing detail. Dense multimodal assessment of every candidate would also be costly, so we complement relevance with sparse helpfulness measurements. Candidate frames are grouped into short temporal blocks, and a deterministic probe schedule selects a limited subset for evaluation. A frozen Qwen2.5-VL-3B observer (Bai et al., 2025) receives a compact three-frame representation of each probed block together with the question and answer choices, without access to the correct answer. If $P_{b,\mathcal{Y}}$ and $P_{b,\mathcal{N}}$ denote its next-token probability masses on fixed positive and negative

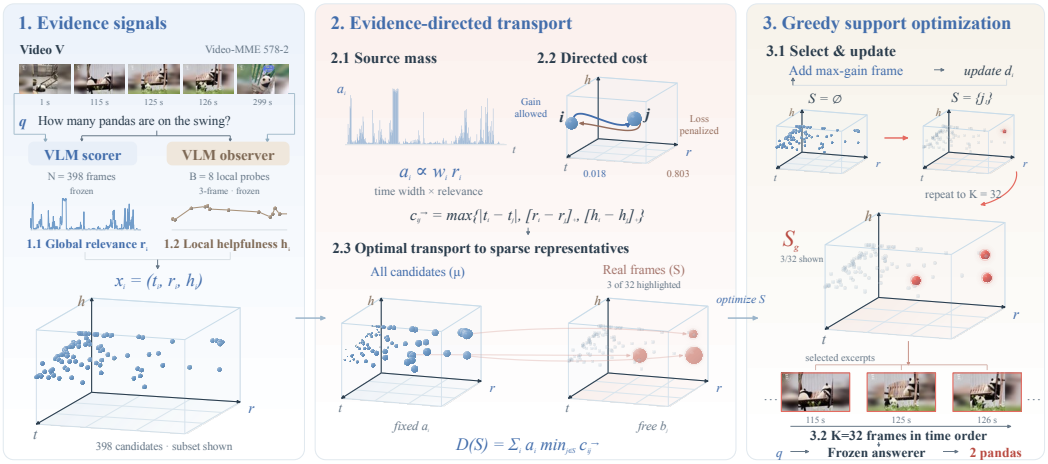


Figure 2: Overview of E-DoT. (1) Frozen models estimate query relevance and sparse local helpfulness. (2) Weighted source observations are assigned to their cheapest selected real frames under directed costs; target masses are free. (3) Greedy selection updates residual losses after each addition; the selected frames reach the fixed answerer in chronological order. Node areas depict their mass.

helpfulness verbalizers (Schick & Schütze, 2021), the block-level readout is

$$u_b = \frac{P_{b,\mathcal{Y}}}{P_{b,\mathcal{Y}} + P_{b,\mathcal{N}}}. \tag{2}$$

The sparse block scores are interpolated over candidate time to obtain $h_{1:N}$. Thus h_i reflects local contextual support rather than the probability that frame i alone answers the question. A separate feedback budget B (the maximum number of observer calls) controls where helpfulness is observed and is independent of the answerer budget K, which counts only the frames the answerer receives.

All observation models are frozen, and all subsequent stages of E-DoT are fully deterministic.

3.2 Evidence Mass over the Timeline

We first determine how strongly each observation should be represented. Let $e_0 = 0$, $e_N = 1$, and $e_i = (t_i + t_{i+1})/2$ for $1 \le i < N$. The temporal Voronoi width $w_i = e_i - e_{i-1}$ measures the portion of the timeline represented by candidate i. We define

$$a_i = \frac{w_i r_i}{\sum_{u=1}^{N} w_u r_u}, \qquad \mu = \sum_{i=1}^{N} a_i \delta_{x_i}, \tag{3}$$

with $a_i = w_i$ when the total relevance is zero. Temporal extent prevents a densely sampled interval from accumulating importance merely because it contains more candidates, while relevance concentrates importance on question-related regions where answer evidence is likely to appear.

Crucially, helpfulness is excluded from the source mass. A relevant region with weak local evidence should not become unimportant simply because its current observation is less informative. Helpfulness instead enters the replacement relation, where preserving or improving it incurs no additional deficit. This cleanly separates the importance of a source from the quality of its potential representative.

3.3 Evidence-Directed Ground Cost

We next determine how safely candidate j can stand in for source i. A suitable representative should remain temporally local while avoiding substantial loss in task-conditioned evidence. For a common tolerance $\epsilon \ge 0$, consider

$$|t_i - t_j| \le \epsilon, \qquad r_j \ge r_i - \epsilon, \qquad h_j \ge h_i - \epsilon. \tag{4}$$

The smallest tolerance satisfying all three requirements gives the directed ground cost

$$c_{ij}^{\rightarrow} = \max\{|t_i - t_j|, [r_i - r_j]_+, [h_i - h_j]_+\}, \qquad [z]_+ = \max(z, 0). \tag{5}$$

Temporal distance discourages substitution across remote events. The positive-part terms measure only evidence deficits, so a nearby destination that preserves or improves relevance and helpfulness incurs no penalty for that improvement. Losing either signal instead increases the replacement cost. Taking the maximum requires temporal proximity and both evidence conditions to hold at the same tolerance, preventing a large deficit in any one coordinate from being averaged away by the others.

This one-sided geometry is central to E-DoT. In general, $c_{ij}^{\rightarrow} \neq c_{ji}^{\rightarrow}$. A stronger nearby observation may therefore replace a weaker source with little loss, while the reverse replacement can be costly. The cost thus treats evidence improvement as a valid substitution rather than as reconstruction error.

For the matched symmetric control used in our ablations, we replace the one-sided deficits with absolute differences:

$$c_{ij}^{\mathrm{sym}} = \max\{|t_i - t_j|, |r_i - r_j|, |h_i - h_j|\}. \tag{6}$$

This control charges evidence changes in either direction, treating improvement as discrepancy.

3.4 Transport to Sparse Frame Support

Pairwise replacement alone does not determine how a fixed frame budget should be distributed across the video. For a candidate support $S \subseteq \{1, \ldots, N\}$, E-DoT transports the source distribution μ only to the selected real frames. We fix the source marginal while leaving the destination marginal relaxed:

$$D(S) = \min_{\pi \geq 0,\ \pi \mathbf{1} = a} \sum_{i=1}^{N} \sum_{j \in S} \pi_{ij} c_{ij}^{\rightarrow}. \tag{7}$$

The induced destination masses are

$$b_j = \sum_i \pi_{ij}. \tag{8}$$

Relaxing the destination marginal, as in SPOT (Gurumoorthy et al., 2021), follows naturally from the replacement view. One retained frame may legitimately represent many redundant source observations, while another may serve only a small distinctive region. Prescribing equal destination masses would instead impose representation responsibilities unrelated to actual replacement quality.

Because no destination marginal is prescribed, each source can assign its mass to a cheapest selected representative. The inner transport problem therefore reduces exactly to

$$D(S) = \sum_{i=1}^{N} a_i \min_{j \in S} c_{ij}^{\rightarrow}, \qquad S^{\star} \in \arg\min_{|S| = K} D(S). \tag{9}$$

This form turns directional replacement into a global coverage problem. Once a source is well represented, another frame serving it offers little additional benefit; a candidate becomes valuable when it reduces residual loss for poorly represented source mass. The selected support therefore favors complementary representatives rather than independently choosing the K strongest candidates.

3.5 Greedy Support Optimization

Finding the optimal size-K support in Eq. (9) is combinatorial. Since $c_{ij}^{\rightarrow} \in [0, 1]$, we set $D(\varnothing) = 1$ and initialize each source's current replacement loss as $d_i = 1$. Given the current support S, the marginal reduction obtained by candidate j is

$$\Delta(j \mid S) = \sum_{i=1}^{N} a_i [d_i - c_{ij}^{\rightarrow}]_+, \qquad d_i = \min_{u \in S} c_{iu}^{\rightarrow}. \tag{10}$$

At each round, E-DoT adds the candidate with the largest marginal gain and updates

$$d_i \leftarrow \min(d_i, c_{ij}^{\rightarrow}). \tag{11}$$

After K rounds, the selected frames are restored to chronological order and passed to the answerer.

Defining $F(S) = 1 - D(S)$ with $F(\varnothing) = 0$ gives, for nonempty S,

$$F(S) = \sum_{i=1}^{N} a_i \max_{j \in S} \left(1 - c_{ij}^{\rightarrow}\right), \tag{12}$$

which is a normalized, monotone, submodular coverage objective. The classical cardinality-constrained greedy result (Nemhauser et al., 1978) therefore gives

$$F(S_g) \geq (1 - 1/e)F(S^\star), \qquad D(S_g) \leq \frac{1}{e} + (1 - 1/e)D(S^\star), \tag{13}$$

where S_g denotes the greedy support. Pairwise costs are evaluated on demand, requiring $O(N + K)$ working memory and $O(KN^2)$ worst-case time once the task-conditioned coordinates are available.

Appendix B further characterizes the resulting objective through integrated replacement coverage and a worst-case support gap over the measured evidence coordinates, with all proofs in Appendix C.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We evaluate on all 2,700 questions of Video-MME (Fu et al., 2025), the 1,337-question LongVideoBench validation set (Wu et al., 2024), the 2,174-question MLVU development set (Zhou et al., 2025), and all 1,549 questions of LVBench (Wang et al., 2025a); Video-MME is evaluated without subtitles. We report accuracy in percent, macro-averaged over the seven tasks for MLVU. All selector comparisons use the same questions within each benchmark.

Models and frame budget. Unless stated otherwise, the answerer is Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and its frame budget is $K = 32$. The main comparison also uses InternVL3-8B (Zhu et al., 2025) and LLaVA-OneVision-7B (Li et al., 2025). E-DoT samples candidates at 1 fps and uses frozen BLIP (Li et al., 2022) for frame–question relevance. At $B = 8$, it also obtains up to eight three-frame neighborhood readouts from a frozen Qwen2.5-VL-3B observer, which sees the question and choices but never the correct label (Appendix A.3). We call $B = 0$ the observer-free variant and $B = 8$ the full configuration; the two differ only in whether local helpfulness is actually observed.

Baselines. We compare with uniform sampling and five frame selectors: AKS (Tang et al., 2025), MDP^3 (Sun et al., 2025), BOLT (Liu et al., 2025), FOCUS (Zhu et al., 2026), and WFS-SB (Chen et al., 2026). Table 1 reports the main comparisons; AKS and BOLT appear in Table 7 (Appendix E).

Evaluation protocols. The main comparisons fix the answerer and frame budget within each model block. The ablations in Table 2 share candidates and evidence scores except for the factor being varied. Appendix E details the protocols, and Appendix G evaluates the selected frames through scene coverage on KFS-Bench (Li et al., 2026) and temporal hits against LVBench reference intervals.

4.2 Main Results

With Qwen2.5-VL-7B, E-DoT ($B = 8$) attains the highest accuracy among the listed selectors on all four benchmarks, exceeding the strongest baseline, WFS-SB, by 1.2, 0.4, 2.2, and 3.2 percentage points (hereafter, points) on Video-MME, LongVideoBench, MLVU, and LVBench, respectively. The observer-free variant ($B = 0$) ties WFS-SB on LongVideoBench and exceeds it by 0.6–0.8 points elsewhere. Paired bootstrap 95% confidence intervals exclude zero for three of the four WFS-SB comparisons (Appendix E.5). E-DoT also outperforms the training-based TSPO (Tang et al., 2026) on Video-MME and LongVideoBench with both tested answerers at $K \in \{32, 64\}$ (Appendix E.2).

With InternVL3-8B and LLaVA-OneVision-7B, the full configuration also outperforms all listed baselines on every benchmark and raises the four-benchmark average over WFS-SB by 0.8 points for both answerers; the smallest margins (0.1–0.2 points) occur on LongVideoBench and InternVL3-8B MLVU. Without any observer calls, the observer-free variant already exceeds WFS-SB on average for all three answerers and is the best selector on LongVideoBench with both of these answerers.

Figure 3 breaks down the Video-MME results by question type, answerer frame budget, and video duration. Appendix F provides the detailed results and run settings. On multi-scene LongVideoBench

Table 1: **Long-video QA accuracy (%).** V-MME: Video-MME; LVB: LongVideoBench. * marks original-paper results under native frame budgets (Appendix E). Other rows use $K = 32$; answerer-only rows use uniform sampling. B caps observer calls. Avg. is the four-benchmark mean where available; bold marks each answerer's best, and shading highlights the full configuration ($B = 8$).

Model / selector	V-MME	LVB	MLVU	LVBench	Avg.
Reported results of foundation models					
LLaVA-OneVision-7B*	58.2	56.3	64.7	–	–
Qwen2.5-VL-7B*	65.1	56.0	70.2	45.3	59.2
InternVL3-8B*	66.3	58.8	71.4	–	–
Frame selectors with a shared answerer and frame budget					
Qwen2.5-VL-7B	61.2	58.9	59.7	38.5	54.6
+ MDP^3 [ICCV'25]	63.8	60.0	66.2	41.3	57.8
+ FOCUS [ICLR'26]	63.4	61.0	68.2	46.3	59.7
+ WFS-SB [CVPR'26]	64.2	64.4	69.9	48.1	61.7
+ E-DoT ($B = 0$)	64.8	64.4	70.6	48.9	62.2
+ **E-DoT** ($B = 8$)	**65.4**	**64.8**	**72.1**	**51.3**	**63.4**
InternVL3-8B	65.6	58.5	68.4	43.5	59.0
+ MDP^3 [ICCV'25]	66.8	60.9	73.9	47.3	62.2
+ FOCUS [ICLR'26]	64.3	61.0	72.0	53.2	62.6
+ WFS-SB [CVPR'26]	66.3	62.5	74.5	53.9	64.3
+ E-DoT ($B = 0$)	67.9	**62.9**	74.3	54.3	64.9
+ **E-DoT** ($B = 8$)	**68.1**	62.6	**74.7**	**54.8**	**65.1**
LLaVA-OneVision-7B	56.5	54.8	63.1	40.4	53.7
+ MDP^3 [ICCV'25]	58.5	59.0	68.2	41.6	56.8
+ FOCUS [ICLR'26]	58.3	60.7	67.1	47.5	58.4
+ WFS-SB [CVPR'26]	61.9	60.4	67.9	48.4	59.7
+ E-DoT ($B = 0$)	61.6	**61.1**	68.4	49.2	60.1
+ **E-DoT** ($B = 8$)	**62.5**	60.9	**69.2**	**49.4**	**60.5**

questions, E-DoT ($B = 8$) leads WFS-SB by 3.3 points, consistent with coverage-aware selection; on single-scene questions, which need less coverage, WFS-SB remains more accurate (Appendix G.2).

4.3 Selection Mechanism Ablations

Table 2 collects the selection and transport controls. Panels (a)–(c) share the candidates, scores, and greedy solver of the main method; panel (d) uses a matched refit solver (Appendix E.6) and is compared only internally. We discuss (b) and (c) here, (a) in Section 4.4, and (d) in Section 4.5.

At $B = 8$ (Table 2(b)), replacing the directed cost with a symmetric one lowers accuracy by 2.3, 1.4, and 1.5 points on Video-MME, LongVideoBench, and MLVU, respectively. One-shot ranking uses the same scores and costs but does not update residual coverage; it trails the full selector by 6.1, 1.9, and 6.1 points. These comparisons thus support both directionality and residual coverage updates.

With feedback disabled (Table 2(c)), the symmetric cost trails the directed greedy reference by 1.4, 1.4, and 0.8 points, while duration-only source mass trails relevance-weighted mass by 1.8, 4.3, and 6.0 points. Directionality and relevance weighting thus help even without any helpfulness readouts.

4.4 Sensitivity to Feedback and Answerer Scale

Feedback budget. Table 2(a) increases the feedback budget from $B = 0$ to 4, 8, and 12 with fixed source masses and coverage updates. $B = 8$ improves over $B = 0$ by 0.6, 0.4, and 1.5 points on Video-MME, LongVideoBench, and MLVU and is the best budget on all three; $B = 12$ adds observer calls without further gain, so a small probe budget of eight calls per question is already sufficient.

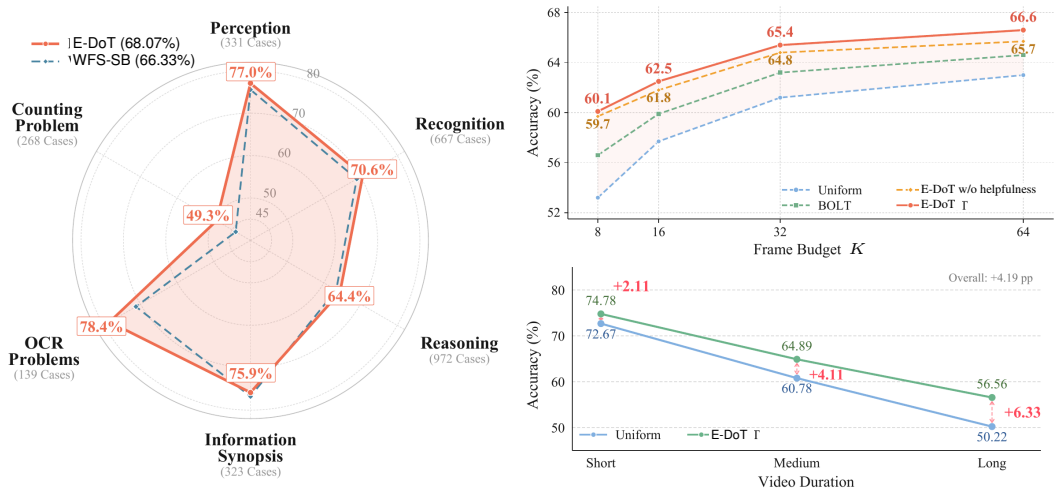


Figure 3: Video-MME diagnostics. Left: accuracy by question type with InternVL3-8B. Right: accuracy versus the answerer frame budget K (top; "w/o helpfulness" denotes $B = 0$) and by video duration (bottom), both with Qwen2.5-VL-7B. The panels come from separate evaluation runs.

Table 2: **Selection and transport ablations** with Qwen2.5-VL-7B ($K = 32$; accuracy, %). Left: (a) feedback budget; (b) cost and coverage updates at $B = 8$; (c) cost and source mass ($a_i \propto w_i$ for duration-only mass) at $B = 0$, all with the greedy solver (one-shot ranking omits the greedy gain updates). Right: (d) transport formulations at $B = 0$ with the matched refit solver of Eq. (39), where λ and ρ weight the entropy (Cuturi, 2013) and target-KL (Chizat et al., 2018) terms, balanced fixes $b_j = 1/K$, and capacity imposes $b_j \le \alpha/K$; $\rho = 0$ unless noted, and the target-KL, balanced, and capacity rows use $\lambda = 0.005$. Bold in (d) marks the best refit score per benchmark (ties included).

Greedy solver	**V-MME**	**LVB**	**MLVU**
(a) Feedback budget			
$B = 0$ (no feedback)	64.8	64.4	70.6
$B = 4$	64.9	64.3	71.5
$B = 8$ (reference)	65.4	64.8	72.1
$B = 12$	65.2	64.6	71.4
(b) Cost and coverage updates ($B = 8$)			
Symmetric cost	63.1	63.4	70.6
One-shot ranking	59.3	62.9	66.0
(c) Cost and source mass ($B = 0$)			
Symmetric cost	63.4	63.0	69.8
Duration-only mass	63.0	60.1	64.6

Refit solver ($B = 0$)	**V-MME**	**LVB**	**MLVU**
(d) Transport formulations			
Free marginal ($\lambda = 0$)	64.8	63.7	70.3
Entropy ($\lambda = 0.005$)	64.5	63.9	70.9
Entropy ($\lambda = 0.020$)	64.8	**64.8**	70.9
Target KL ($\rho = 0.01$)	64.5	64.2	70.6
Target KL ($\rho = 0.1$)	64.1	**64.8**	71.2
Target KL ($\rho = 1$)	**64.9**	64.6	**71.6**
Balanced ($b_j = 1/K$)	64.3	64.2	71.5
Target capacity ($\alpha = 2$)	64.7	64.4	71.3
Target capacity ($\alpha = 4$)	64.6	63.7	70.8

Observer-matched controls. With the same readouts ($B = 8$), E-DoT leads all controls except on LVBench, where direct fusion ($r_i + h_i$ Top-K, no coverage updates) is 0.1 points higher (Table 12).

Feedback model. Table 3 varies the feedback source. InternVL3.5-4B (Wang et al., 2025b) improves over no feedback by 0.2 points on each benchmark. Qwen2.5-VL-7B improves by 0.8 points on each; relative to the default Qwen2.5-VL-3B observer, it is 0.2 and 0.4 points higher on Video-MME and LongVideoBench but 0.7 points lower on MLVU. Every tested observer thus improves all three benchmarks, and the default 3B observer rivals its 7B counterpart, so a large observer is unnecessary.

Answerer scale. Table 4 varies the Qwen2.5-VL answerer size on Video-MME with the default $B = 8$ observer; Δ denotes E-DoT minus Uniform. E-DoT improves over Uniform at every scale, by 3.3–4.2 points from 3B to 32B; the gain narrows to 1.6 points at 72B, where Uniform is already strongest. A single selector configuration with the 3B observer thus helps answerers from 3B to 72B.

Table 3: **Feedback-model sensitivity** with a fixed Qwen2.5-VL-7B answerer ($K = 32$; accuracy, %).

Feedback source	B	V-MME	LVB	MLVU
None (constant helpfulness)	0	64.8	64.4	70.6
Qwen2.5-VL-3B (default)	8	65.4	64.8	72.1
InternVL3.5-4B	8	65.0	64.6	70.8
Qwen2.5-VL-7B	8	65.6	65.2	71.4

Table 4: **Answerer scale** on Video-MME with Qwen2.5-VL ($K = 32$; accuracy, %).

Size	Uniform	E-DoT	Δ
3B	57.4	60.7	+3.3
7B	61.2	65.4	+4.2
32B	63.9	67.3	+3.4
72B	66.8	68.4	+1.6

Table 5: **Shared-scorer accuracy** (%) with Qwen2.5-VL-7B at $K = 32$. E-DoT uses $B = 8$; bold marks the better result within each scorer pair; Avg.: unweighted four-benchmark mean, as in Table 1.

Scorer	Selector	V-MME	LVB	MLVU	LVBench	Avg.
CLIP	BOLT	**63.2**	60.0	68.4	39.3	57.7
	E-DoT	62.9	**61.3**	**68.7**	**47.9**	**60.2**
SigLIP	MDP^3	63.8	60.0	66.2	41.3	57.8
	E-DoT	**64.3**	**63.8**	**72.5**	**49.6**	**62.6**
BLIP-2	WFS-SB	64.2	64.4	69.9	48.1	61.7
	E-DoT	**65.7**	**64.9**	**71.8**	**50.8**	**63.3**

4.5 Alternative Transport Formulations

Table 2(d) compares nine $B = 0$ formulations under the matched refit. Target KL with $\rho = 1$ scores 64.9%, 64.6%, and 71.6%, exceeding the free-marginal refit on all three benchmarks, while the entropy-only variant with $\lambda = 0.020$ and target KL with $\rho = 0.1$ tie for the best LongVideoBench score (64.8%). Target-marginal regularization is thus a strong observer-free alternative, and all nine formulations lie within 1.3 points of one another, so the formulation can be chosen for tractability.

With the same $B = 8$ feedback, target KL reaches 65.2%, 64.8%, and 72.5%, within 0.4 points of greedy E-DoT (65.4%, 64.8%, and 72.1%; Appendix E.8). Greedy free-target selection attains this comparable accuracy at about 1/130 of the selection time, 0.030 versus 3.985 seconds per Video-MME question (Table 10), while retaining the exact reduction and the $(1 - 1/e)$ guarantee; on hour-long LVBench videos, the times are 0.411 versus 33.138 seconds. Both timings are reconstructed $B = 0$ selection-stage estimates that exclude question answering, not end-to-end latency (Appendix E.7).

4.6 Shared Relevance Scorers

Table 5 pairs E-DoT with BOLT under CLIP (Radford et al., 2021), MDP^3 under SigLIP (Zhai et al., 2023), and WFS-SB under BLIP-2 (Li et al., 2023), reusing the WFS-SB results of Table 1. E-DoT raises the four-benchmark average from 57.7% to 60.2%, from 57.8% to 62.6%, and from 61.7% to 63.3%, respectively, and leads in 11 of the 12 scorer–benchmark pairs, trailing only by 0.3 points on Video-MME with CLIP, so its gains do not hinge on the choice of frame–question relevance model.

5 Conclusion

We introduced E-DoT, a training-free approach to long-video frame selection that models how retained frames represent discarded observations. A directed ground cost penalizes evidence deficits but not improvements, and free destination masses reduce sparse-support transport to a monotone submodular coverage objective that greedy selection optimizes with a $(1 - 1/e)$ guarantee. Directed relevance and source-weighted coverage already form an effective observer-free selector; sparse local-helpfulness readouts further refine the replacement cost at the price of additional observer calls. Under a fixed 32-frame answerer budget, the full configuration achieves the highest accuracy among the compared selectors across three answerers and four benchmarks, and matched ablations support directionality, relevance-weighted source mass, and coverage updates. Together, these results demonstrate that directed evidence costs and coverage-aware selection provide an effective approach to selecting informative and complementary frames for long-video QA under a fixed answerer frame budget, with sparse local observer feedback further refining the directed evidence replacement cost.

Reproducibility Statement

Section 3 specifies the method. Appendix A gives implementation details, including candidate sampling, the relevance and helpfulness readouts with the full observer prompt, the probe schedule, candidate-grid augmentation, and the greedy solver with its tie-breaking rule. Appendix B states the theoretical results, Appendix C contains all proofs, and Appendix E describes the datasets, model checkpoints, evaluation protocol, and control experiments. Evaluation uses LMMs-Eval (Zhang et al., 2025a) and, for LVBench, a separate evaluator (Appendix E.1); apart from the resampling-based significance tests (Appendix E.5), the pipeline involves no stochastic operations.

AI Use Statement

Generative AI tools assisted language editing, manuscript organization, LaTeX, and the overview figure. The authors checked the experimental code, data, and manuscript for correctness and consistency and take responsibility for the final work.

Ethics Statement

This work studies frame selection on existing video benchmarks. A selector may omit context or amplify biases present in its relevance scorer and helpfulness observer, and benchmark performance alone does not establish reliability for sensitive decision-making. Use and redistribution of the videos and annotations remain subject to their respective licenses, terms of use, and privacy conditions.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. Wavelet-based frame selection by detecting semantic boundary for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24052–24061, 2026.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025.

Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2069–2077, 2014.

Karthik S. Gurumoorthy, Pratik Jawanpuria, and Bamdev Mishra. SPOT: A framework for selection of prototypes using optimal transport. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 535–551, 2021.

Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3090–3098, 2015.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742, 2023.

Zongyao Li, Kengo Ishida, Satoshi Yamazaki, Xiaotong Ji, and Jianquan Liu. KFS-Bench: Comprehensive evaluation of key frame sampling in long video understanding. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

Peng Ling, Yingda Yin, Lingting Zhu, Weikai Chen, Shengju Qian, Zeyu Hu, Xin Wang, and Wenming Yang. CoverPrune: Coverage-driven token pruning for 3D VLMs via optimal transport. In *European Conference on Computer Vision*, 2026.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. BOLT: Boost large vision-language model without training for long-form video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3318–3327, 2025.

Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*, volume 7 of *Lecture Notes in Control and Information Sciences*, pp. 234–243. Springer, 1978.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294, 1978.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 255–269, 2021.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP^3: A training-free approach for list-wise frame selection in video-LLMs. In *IEEE/CVF International Conference on Computer Vision*, pp. 24090–24101, 2025.

Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Huayu Zhang, Jinglin Xu, and Hao Sun. TSPO: Temporal sampling policy optimization for long-form video language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9368–9376, 2026. doi: 10.1609/aaai.v40i11.37896.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29118–29128, 2025.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations*, 2016.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. LVBench: An extreme long video understanding benchmark. In *IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3272–3283, 2025c.

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

Shanqing Xu, Meng Luo, Mengchen Qian, Yuhui Gao, Siyue Peng, Xiaohan Zhong, Xiaojin Zhang, Zhongyu Wei, Wei Chen, and Xiang Bai. RIDGE: Region-informed derivative-guided evidence selection for long video understanding. *arXiv preprint arXiv:2608.29958*, 2026.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-Voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. LMMs-Eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL*, pp. 881–916, 2025a.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-Frame: Query-aware frame selection and multi-resolution adaptation for video-LLMs. In *IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025b.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations*, 2026.

A Observation and Optimization Details

Table 6 summarizes the notation used throughout the paper.

Table 6: **Notation used throughout the paper.**

Symbol	Meaning
T, N, K	Native video length, number of candidates, and answerer frame budget
f_i, t_i	Native frame index and normalized time of candidate i
r_i, h_i	Query relevance and interpolated local helpfulness
w_i, a_i	Temporal Voronoi width and normalized source mass
B	Feedback budget (maximum number of observer calls)
$c_{ij}^{\rightarrow}$	Cost of using destination j to represent source i
S, π, b	Selected indices, transport coupling, and induced destination masses
$d_i(S), D(S)$	Residual cost of source i and source-weighted transport loss
$F(S), \Delta(j \mid S)$	Coverage benefit and marginal benefit of adding j
Φ	Functions on $[0,1]^3$ nondecreasing in r, h and 1-Lipschitz in ℓ_∞

A.1 Candidate Times and Source Masses

Candidate frames are sampled at 1 fps, with no cap on their number N. Let $f_1 < \cdots < f_N$ be the resulting distinct frame indices in a video of $T > 1$ frames, with $f_i \in \{0, \ldots, T-1\}$ and $1 \le K \le N$. The normalized times are $t_i = f_i/(T-1)$, and relevance scores are used on their unit-interval scale. The temporal Voronoi boundaries are

$$e_0 = 0, \qquad e_N = 1, \qquad e_i = (t_i + t_{i+1})/2 \quad (1 \le i < N), \tag{14}$$

with $w_i = e_i - e_{i-1}$. The widths sum to one, including the intervals before the first and after the last candidate. If all relevance scores are zero, the source masses are $a_i = w_i$; otherwise, Eq. (3) applies. This weighting accounts for irregular spacing, although changing the candidate grid can still change the source approximation and the available representatives.

A.2 Relevance Score Definition

Let o_i denote the scalar output of the frozen BLIP scorer for candidate i and its query input. The relevance coordinate is $r_i = g_{\mathrm{rel}}(o_i) \in [0,1]$. In our experiments, o_i is the positive-class probability of the BLIP image–text matching head, and g_{rel} is the identity. The mapping matters beyond ranking, because r_i determines both the source masses and the numerical deficits in the ground cost. For example, a positive rescaling $r_i' = \kappa r_i$ that remains in $[0,1]$ leaves the normalized source masses unchanged when the total relevance is nonzero, but changes the relevance term in the cost.

A.3 Sparse Helpfulness Observations

A frozen Qwen2.5-VL-3B-Instruct observer receives three frames from a neighborhood together with the question and its answer choices, but not the correct answer. The complete prompt is

```
Question:
{question}
Choices:
(A) {choice_A}
(B) {choice_B}
...
The three images are frames from one short candidate video
segment. Do these frames contain decisive evidence needed to
answer the question and choose among the choices? Answer
only Yes or No.
```

Let $\mathcal{F}_b$ be the three requested frame positions for neighborhood b, and let $\mathcal{I}_b$ be the actual image inputs after decoding and any fallback. The disjoint verbalizer sets $\mathcal{Y}$ and $\mathcal{N}$ contain the first tokens of

the strings in {`Yes`, ␣`yes`, `YES`, ␣`yes.`} and {`No`, ␣`no`, `NO`, ␣`no.`}, respectively, where ␣ marks a leading space. We define

$$P_{b,\mathcal{Y}} = \sum_{v \in \mathcal{Y}} p(v \mid \mathcal{I}_b, q), \qquad P_{b,\mathcal{N}} = \sum_{v \in \mathcal{N}} p(v \mid \mathcal{I}_b, q),$$
$$u_b = \sigma(\log P_{b,\mathcal{Y}} - \log P_{b,\mathcal{N}}) = \frac{P_{b,\mathcal{Y}}}{P_{b,\mathcal{Y}} + P_{b,\mathcal{N}}}. \tag{15}$$

Here p is the observer's next-token distribution and σ the logistic sigmoid. Thus u_b is the normalized probability mass within the Yes/No verbalizers, not a calibrated probability of correctness.

Alignment and interpolation. We assign each readout to the mean of its requested positions, $\bar{f}_b = |\mathcal{F}_b|^{-1} \sum_{f \in \mathcal{F}_b} f$, and average readouts that share a center. After this merge, let $s_1 < \cdots < s_M$ denote the distinct centers normalized by $T - 1$, and let $\bar{u}_1, \ldots, \bar{u}_M$ be the corresponding readouts. For $M \geq 2$ and $s_m \leq t_i \leq s_{m+1}$,

$$h_i = \frac{s_{m+1} - t_i}{s_{m+1} - s_m} \bar{u}_m + \frac{t_i - s_m}{s_{m+1} - s_m} \bar{u}_{m+1}. \tag{16}$$

For $t_i < s_1$ or $t_i > s_M$, we use the nearest endpoint value. If $M = 1$, all candidates receive $\bar{u}_1$; without feedback ($B = 0$), all receive 0.5. Interpolation is a convex combination and hence preserves the unit interval. It transfers an estimate of local context to a candidate position; it neither re-observes that frame nor resolves discontinuities within an unobserved interval.

Decoding failures. If a requested frame fails to decode, the procedure retries neighboring indices within two frames and uses a black image if all attempts fail. The center $\bar{f}_b$ therefore refers to the requested positions, which can differ from the decoded ones; such substitutions were not logged.

A.4 Helpfulness Probe Placement

The probe schedule is fixed before final frame selection. We partition the candidates into consecutive blocks of eight and min–max normalize three block scores: (i) relevance saliency, defined as 0.7 times the maximum plus 0.3 times the mean relevance in the block; (ii) the mean cosine dissimilarity to the available immediate temporal neighbors; and (iii) local relevance contrast, $\max_{i \in \mathcal{G}} r_i - \mathrm{mean}_{i \in \mathcal{G}^- \cup \mathcal{G} \cup \mathcal{G}^+} r_i$ for block $\mathcal{G}$, where $\mathcal{G}^-$ and $\mathcal{G}^+$ are the equal-length blocks immediately before and after $\mathcal{G}$, truncated at the video boundaries. Cosine similarities are computed between block embeddings, each obtained by averaging the L2-normalized cached BLIP vision embeddings of the block's frames and L2-normalizing the mean. The weighted sum of the three scores uses coefficients 0.8, 0.2, and 0.1. Greedy block selection then subtracts 0.2 times the maximum cosine similarity to an already selected block. For a feedback budget B, the observer queries the first $B_{\text{eff}} = \min(B, \lceil N/8 \rceil)$ blocks in this deterministic order. Each queried block is observed through its frames at local indices $\lfloor \mathrm{linspace}(0, 7, 3) \rfloor = (0, 3, 7)$; a block with fewer than eight candidates contributes at most three frames, spread evenly within it. The schedule determines where helpfulness is measured; it does not select the final frames or consume the answerer budget K.

A.5 Candidate-Grid Augmentation

The selection candidates need not coincide with the sparse frames examined by the observer. If the candidate grid contains fewer than K entries, we augment it with the integer positions

$$\big\lfloor \mathrm{linspace}(0, T - 1, \min(2K, T)) \big\rfloor, \tag{17}$$

taking the sorted union and removing duplicate positions. Relevance at the added positions is interpolated from the original candidates with endpoint extension, and helpfulness from the same neighborhood readouts. Times, temporal Voronoi widths, and normalized source masses are then recomputed on the augmented grid before optimization. This adds candidate frames without extra model calls and requires $T \geq K$ for K distinct frames; videos with $T < K$ fall outside this setting.

Algorithm 1 E-DoT greedy support selection

Require: Video length $T > 1$; ordered candidate indices f, relevance r, helpfulness h; budget $1 \le K \le N$

Normalize candidate times and compute source masses a (Appendix A.1)
$S \leftarrow \varnothing$; $d_i \leftarrow 1$ for all i
for $k = 1, \dots, K$ **do**
  $j^\star \leftarrow \arg\max_{j \notin S} \sum_i a_i [d_i - c^{\rightarrow}_{ij}]_+$ ▷ earliest index on ties
  $S \leftarrow S \cup \{j^\star\}$
  $d_i \leftarrow \min(d_i, c^{\rightarrow}_{ij^\star})$ for all i
end for
return $\{f_j : j \in S\}$ in chronological order

A.6 Greedy Optimization

Algorithm 1 specifies exact greedy selection. Its lazy implementation (Minoux, 1978) uses previously computed marginal gains as upper bounds and recomputes a candidate when it reaches the top of a priority queue. A refreshed candidate is accepted only when no remaining bound can outrank it. Ties are resolved in favor of the earlier frame index, including ties at zero gain, and the algorithm continues until K distinct candidates are selected. An equivalent lexicographic priority key is $(\Delta(j \mid S), -f_j)$, ordered from largest to smallest; both gain and index therefore matter when checking acceptance at a tie. All computations use double precision, and ties require exactly equal gains.

Pairwise costs are computed as needed without storing an $N \times N$ matrix. After every nonempty selection, the residuals satisfy $d_i = \min_{j \in S} c^{\rightarrow}_{ij}$, and the reported loss is $D(S) = \sum_i a_i d_i$. Assigning each source to a minimum-cost selected representative yields destination masses b_j; ties in this optional reconstruction can also be resolved by earlier index. These masses can be unequal or zero. They serve only to interpret coverage and are not supplied to the answerer as inputs or attention weights; the answerer receives all K frames in temporal order. Appendix C.3 analyzes the complexity.

B Analysis of the E-DoT Objective

The approximation guarantee in Eq. (13) concerns the greedy optimization of the E-DoT support objective. We next characterize what a small transport loss $D(S)$ means in the task-conditioned coordinates (t, r, h). Two views are useful: coverage across admissible transport tolerances, and a worst-case gap over a class of monotone Lipschitz functions.

Coverage across transport tolerances. For $\epsilon \in [0, 1]$, we call source i covered if at least one selected destination satisfies all three requirements in Eq. (4), and let $A_\epsilon(S)$ denote the set of covered sources.

Proposition 1 (Integrated evidence coverage)**.** *For every nonempty S,*

$$F(S) = 1 - D(S) = \int_0^1 \sum_{i \in A_\epsilon(S)} a_i \, d\epsilon. \tag{18}$$

Source i becomes covered once ϵ reaches $d_i(S) = \min_{j \in S} c^{\rightarrow}_{ij}$, so its coverage indicator integrates to $1 - d_i(S)$ (Appendix C.4). Minimizing $D(S)$ therefore maximizes source-weighted coverage averaged over transport tolerances: a lower loss means that more source mass can reach the selected support under stricter admissibility criteria, that is, at smaller tolerances ϵ.

Worst-case gap in the evidence coordinates. Let Φ be the class of functions $\phi(t, r, h)$ on $[0, 1]^3$ that are nondecreasing in relevance and helpfulness and 1-Lipschitz with respect to the ℓ_∞ norm. Monotonicity encodes the same one-sided preference as the directed ground cost, while the Lipschitz condition limits sensitivity to small coordinate changes; no monotonicity is imposed in time.

For every pair of points $x, y \in [0, 1]^3$,

$$c^{\rightarrow}(x, y) = \sup_{\phi \in \Phi} \{\phi(x) - \phi(y)\}. \tag{19}$$

This relation extends from a source–destination pair to the selected support.

Theorem 1 (Worst-case gap over evidence coordinates)**.** *For every nonempty S,*

$$D(S) = \sup_{\phi \in \Phi} \left\{ \sum_{i=1}^{N} a_i \phi(x_i) - \max_{j \in S} \phi(x_j) \right\}. \tag{20}$$

Hence, for every $\phi \in \Phi$, some selected frame attains the source-weighted average of ϕ up to an additive loss $D(S)$, and no smaller bound holds uniformly over the class. Equality is witnessed by $\phi_S(x) = \min_{j \in S} c^{\rightarrow}(x, x_j)$; Appendix C.5 gives the proof.

These statements concern only the evidence coordinates: event identity and cross-frame relations need not be recoverable from (t, r, h), and an answerer's decision rule need not lie in Φ. Whether this compact geometry is useful for long-video QA is therefore an empirical question (Section 4). Appendix D discusses boundary cases, and Appendix C.7 analyzes helpfulness perturbations.

C Proofs and Objective Properties

C.1 The Directed Ground Cost

Proposition 2 (Minimum common slack)**.** *For any source x_i and destination x_j in $[0,1]^3$, the directed cost in Eq.* (5) *is the minimum tolerance $\epsilon \geq 0$ satisfying all three requirements in Eq.* (4)*.*

Proof. The time inequality requires $\epsilon \geq |t_i - t_j|$. The relevance inequality requires $\epsilon \geq r_i - r_j$ and, together with nonnegativity, $\epsilon \geq [r_i - r_j]_+$. The same argument applies to helpfulness. The maximum of these three lower bounds satisfies all constraints, so it is the smallest feasible slack. □

Proposition 3 (Directed triangle inequality)**.** *The cost is nonnegative, bounded by one, zero on the diagonal, and satisfies the directed triangle inequality $c^{\rightarrow}_{ik} \leq c^{\rightarrow}_{ij} + c^{\rightarrow}_{jk}$ for all i, j, k.*

Proof. The first three properties follow from the unit-interval coordinates. The absolute time difference satisfies the ordinary triangle inequality. For either evidence coordinate $z \in \{r, h\}$,

$$[z_i - z_k]_+ \leq [z_i - z_j]_+ + [z_j - z_k]_+. \tag{21}$$

Hence each coordinate cost from i to k is at most $c^{\rightarrow}_{ij} + c^{\rightarrow}_{jk}$, and taking the maximum proves the claim. Symmetry does not hold in general: a higher evidence coordinate at j contributes no deficit from i to j, but may contribute a deficit in the reverse direction. □

C.2 Free-Target Transport Reduction

For any feasible coupling and nonempty S,

$$\sum_i \sum_{j \in S} \pi_{ij} c^{\rightarrow}_{ij} \geq \sum_i \sum_{j \in S} \pi_{ij} \min_{u \in S} c^{\rightarrow}_{iu} \tag{22}$$

$$= \sum_i a_i \min_{u \in S} c^{\rightarrow}_{iu}. \tag{23}$$

Choose $j(i) \in \arg\min_{j \in S} c^{\rightarrow}_{ij}$ and send the entire mass a_i to $j(i)$. The induced column sums $b_j = \sum_i \pi_{ij}$ define a probability distribution on S, and this coupling attains the lower bound. Hence the free-target transport objective in Eq. (7) equals the nearest-representative loss in Eq. (9). No constraint requires each b_j to be positive. Fixing the destination masses, imposing column capacities, or adding entropy regularization would change this inner problem.

C.3 Submodularity and the Greedy Bound

Write $v_{ij} = 1 - c^{\rightarrow}_{ij} \in [0, 1]$. For nonempty S,

$$F(S) = \sum_i a_i \max_{j \in S} v_{ij}, \qquad F(\varnothing) = 0. \tag{24}$$

For $A \subseteq B$ and $j \notin B$, each source's best value under B is at least its best value under A, so the increase from adding j is no larger under B. Summation with nonnegative weights a_i proves diminishing returns. Normalization and monotonicity follow immediately.

For completeness, let S_m be the greedy set after m additions and $S^\star$ an optimal set of size K. Monotonicity and submodularity give

$$F(S^\star) - F(S_m) \le \sum_{j \in S^\star \setminus S_m} \Delta(j \mid S_m) \le K \max_{j \notin S_m} \Delta(j \mid S_m). \tag{25}$$

Each exact greedy step therefore multiplies this gap by at most $1 - 1/K$. Since $F(S_0) = 0$,

$$F(S_K) \ge \left[1 - (1 - 1/K)^K\right] F(S^\star) \ge (1 - 1/e) F(S^\star). \tag{26}$$

Substituting $F = 1 - D$ yields Eq. (13). This is the classical result for cardinality-constrained submodular maximization (Nemhauser et al., 1978), which E-DoT instantiates with the directed evidence cost. The same argument holds for any fixed costs in $[0, 1]$, including the symmetric control: directionality changes the objective values and selected representatives, not the greedy guarantee.

In exact arithmetic, lazy evaluation (Minoux, 1978) preserves the greedy decisions: by submodularity, a previously computed marginal gain upper-bounds its later value, and a candidate is selected only after its current marginal gain is at least every remaining upper bound.

Evaluating all initial candidate gains requires N^2 pairwise costs. In the worst case, lazy evaluation revisits every candidate in each round, leading to $O(KN^2)$ time; heap operations do not increase this order. The heap, coordinate arrays, and residual array require $O(N)$ storage, and the selected indices and gains require $O(K)$. Reconstructing assignments costs $O(NK)$. The implementation never stores an $N \times N$ matrix, so memory stays linear in N and K.

C.4 Coverage Identity

For $d_i(S) \in [0, 1]$, the layer-cake identity gives

$$1 - d_i(S) = \int_0^1 \mathbf{1}\{d_i(S) \le \epsilon\}\, d\epsilon. \tag{27}$$

The event in the indicator is equivalent to the existence of a selected destination satisfying all three inequalities in Eq. (4). Summing over the finite source set proves Proposition 1. With the symmetric cost, acceptable destinations must additionally satisfy $r_j \le r_i + \epsilon$ and $h_j \le h_i + \epsilon$; these upper bounds are what make the symmetric cost penalize evidence improvements.

C.5 The Monotone Lipschitz Class and the Support Identity

Let $x = (t_x, r_x, h_x)$ and $y = (t_y, r_y, h_y)$, and define

$$v = (t_y, \min(r_x, r_y), \min(h_x, h_y)). \tag{28}$$

For $\phi \in \Phi$, monotonicity gives $\phi(v) \le \phi(y)$, while

$$\|x - v\|_\infty = \max\{|t_x - t_y|, [r_x - r_y]_+, [h_x - h_y]_+\} = c^{\rightarrow}(x, y). \tag{29}$$

The Lipschitz condition therefore yields $\phi(x) - \phi(y) \le c^{\rightarrow}(x, y)$. For fixed y, the function $\phi_y(z) = c^{\rightarrow}(z, y)$ is nondecreasing in the two evidence coordinates and 1-Lipschitz in z. Since $\phi_y(y) = 0$, it attains the upper bound, which proves Eq. (19) for all pairs x, y.

To prove Theorem 1, choose a minimum-cost destination $j(i)$ for every source. Then

$$\phi(x_i) - \max_{j \in S} \phi(x_j) \le \phi(x_i) - \phi(x_{j(i)}) \le c^{\rightarrow}(x_i, x_{j(i)}). \tag{30}$$

Weighted summation gives the upper bound $D(S)$, since $\sum_i a_i = 1$. For the reverse inequality, take

$$\phi_S(x) = \min_{j \in S} c^{\rightarrow}(x, x_j). \tag{31}$$

A finite minimum of functions nondecreasing in an evidence coordinate remains nondecreasing in it. It also remains 1-Lipschitz: choosing a minimizing index at y bounds $\phi_S(x) - \phi_S(y)$ by $\|x - y\|_\infty$,

and exchanging x and y bounds the reverse difference. Thus $\phi_S \in \Phi$. It vanishes on S because every selected point represents itself at zero cost, and its source average is $D(S)$, which proves equality.

The maximizing function may depend on S, as the supremum allows. Eq. (19) and Theorem 1 are one-sided analogues of Kantorovich–Rubinstein duality (Villani, 2009). Appendix D examines properties outside this function class, such as answer loss and uniform frame averages.

C.6 Weighted Functions and the Input-Weight Gap

For a function that is nondecreasing in r, h and L-Lipschitz in ℓ_∞, the construction above gives $\phi(x_i) - \phi(x_j) \le L c^{\rightarrow}_{ij}$. Integrating against an optimal coupling yields

$$\sum_i a_i \phi(x_i) - \sum_{j \in S} b_j \phi(x_j) \le L D(S). \tag{32}$$

An unweighted average over the selected frames is a different quantity. Let $u_j = 1/K$ and let R_S be the range of ϕ on S. Adding and subtracting the weighted target mean gives

$$\sum_i a_i \phi(x_i) - \frac{1}{K} \sum_{j \in S} \phi(x_j) \le L D(S) + R_S \, \mathrm{TV}(b, u), \qquad \mathrm{TV}(b, u) = \frac{1}{2} \sum_{j \in S} |b_j - u_j|. \tag{33}$$

The bound on the remaining term follows by centering ϕ between its extremes on S. We do not assume that weighted or uniform averaging describes how the answerer aggregates frames.

C.7 Sensitivity to Helpfulness

If a constant is added to all h_i and all values remain in $[0, 1]$ without clipping, every helpfulness difference is unchanged; the objective is thus invariant to translations in the probability coordinate. A common logit shift followed by the sigmoid need not preserve differences and is not covered here.

Suppose $\|\widehat{h} - h\|_\infty \le \delta$, with the candidate times, relevance values, and source masses held fixed. Every pairwise helpfulness difference then changes by at most 2δ. Since the positive part and the maximum are nonexpansive in the sup norm, $|\widehat{c}_{ij} - c_{ij}| \le 2\delta$. For any fixed support S, taking the minimum over S and averaging with probability weights preserves the bound:

$$|\widehat{D}(S) - D(S)| \le 2\delta. \tag{34}$$

If $\widehat{S}$ exactly minimizes $\widehat{D}$ and $S^\star$ exactly minimizes D, then

$$D(\widehat{S}) \le \widehat{D}(\widehat{S}) + 2\delta \le \widehat{D}(S^\star) + 2\delta \le D(S^\star) + 4\delta. \tag{35}$$

For greedy selection, the bound additionally includes the greedy optimization gap. These inequalities concern objective values, not the stability of the selected frame identities, and they are conditional on the observation error δ, whose estimation requires separate measurements.

D Examples and Limits of the Proxy Representation

D.1 A Numerical Directionality Example

We use a LongVideoBench question about what happens after a chef pours a yellow liquid into a wok. The video has $T = 1{,}806$ native frames and $N = 63$ candidate frames, and selection uses $K = 32$. Eight three-frame feedback readouts are available. The no-feedback comparison uses the same candidates, relevance scores, and source masses, with $h_i = 0.5$. Frame indices are zero-based, $t = f/(T-1)$, and h is interpolated from the neighborhood readouts.

The cached coordinates of the four frames in Figure 4 are:

Native frame f	t	r	h **(feedback)**
29	0.016066	0.000742	0.437818
58	0.032133	0.972656	0.437818
174	0.096399	0.909668	0.575860
319	0.176731	0.919434	0.834688

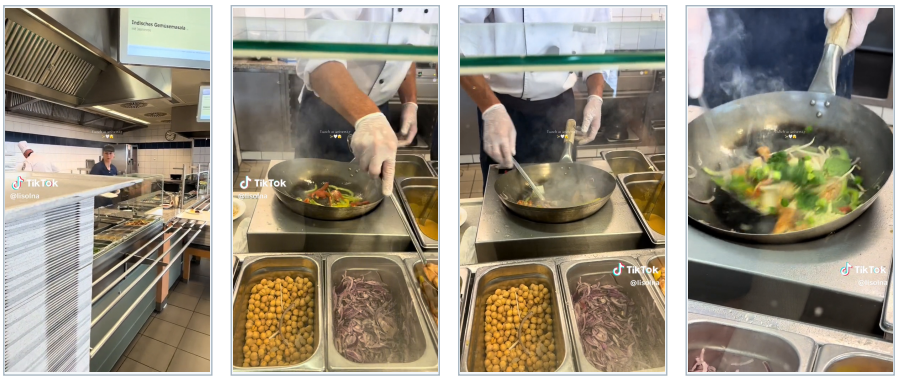


Figure 4: LongVideoBench frames used in the directionality example of Appendix D.1.

Directionality without feedback. Write $c(f \to g)$ for the cost of representing source frame f by destination frame g. For frames 29 and 58, setting $h_i = 0.5$ gives

$$c(29 \to 58) = 0.016066, \qquad c(58 \to 29) = 0.971914.$$

The forward cost is the temporal distance, whereas the reverse cost equals the relevance deficit. The symmetric cost is 0.971914 in both directions. The no-feedback method is thus already directional through relevance. Because the two frames have equal interpolated helpfulness under feedback, their costs do not change when feedback is added, unlike those of the next pair.

Additional directionality from measured helpfulness. For frames 174 and 319, the temporal gap is 0.080332 and the relevance difference is only 0.009766. Without feedback, both directions therefore cost 0.080332. With the observed helpfulness values,

$$c(174 \to 319) = 0.080332, \qquad c(319 \to 174) = 0.258828.$$

The reverse cost now equals the helpfulness deficit, and the symmetric cost becomes 0.258828 in both directions. This pair isolates the asymmetry introduced by the measured helpfulness coordinate.

D.2 A Worked Example of Coverage-Aware Selection

We replay the same question over all 63 candidates, with the same source masses (Eq. (3)) and budget $K = 32$. One-shot ranking orders the candidates once by $\Delta(j \mid \varnothing)$, whereas greedy selection recomputes the gains after each addition. The first two choices already reveal the difference:

Readout	First f	Candidate f	Initial gain	Updated gain	Second choice
No feedback ($B = 0$)	232	203	0.913603	0.006551	One-shot
		406	0.892454	0.027889	Greedy
Feedback ($B = 8$)	319	261	0.902836	0.012432	One-shot
		145	0.835489	0.028348	Greedy

Here, "updated" refers to the gain after the first frame in the same row block is selected. Without feedback, frame 203 initially outranks frame 406, but most of its coverage becomes redundant once frame 232 is selected, and frame 406 then gains more. With feedback, the first choice becomes frame 319, and updating coverage favors frame 145 over the initially higher-ranked frame 261 (Figure 5).

After all 32 additions, the losses are:

Cost used for both supports	One-shot $D(S)$	Greedy $D(S)$
No feedback ($B = 0$)	5.599710×10^{-5}	1.498742×10^{-5}
Feedback ($B = 8$)	5.666272×10^{-5}	1.601195×10^{-5}

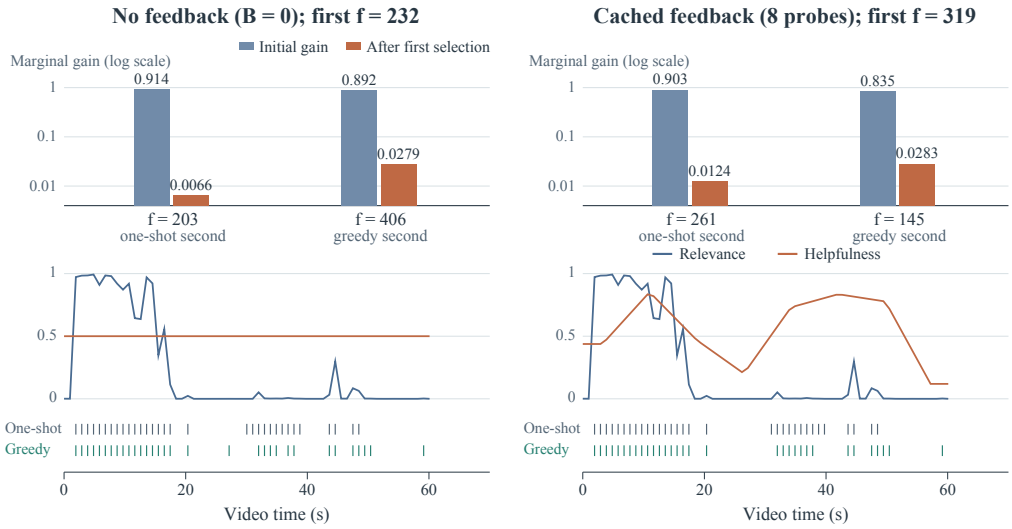


Figure 5: One-shot and greedy selection on the LongVideoBench question of Appendix D.1. Top: initial and updated gains for the second frame. Bottom: the selected positions at $K = 32$.

Under either cost, greedy selection attains a lower transport loss than one-shot ranking. Under the feedback cost, the no-feedback support has a loss of 1.607135×10^{-5}, compared with 1.601195×10^{-5} for the feedback support. Despite different early selection orders, the two greedy supports differ by only one frame (812 versus 1073), consistent with their nearly equal losses.

D.3 An Important Event Can Have Negligible Source Mass

Take $t = (0, 1/2, 1)$, $r = (1, \eta, 1)$, constant helpfulness, and $S = \{1, 3\}$. The middle source mass is $\eta/(1+\eta)$, and it travels a normalized distance of $1/2$. Thus

$$D(S) = \frac{\eta}{2(1+\eta)} \longrightarrow 0 \quad \text{as } \eta \longrightarrow 0. \tag{36}$$

All relevance values are positive when $\eta > 0$. Nevertheless, an answer may depend entirely on an event visible only in the middle frame, and deleting that frame can make a correct answer incorrect. A small loss alone therefore cannot bound the answer loss of an arbitrary answerer.

D.4 Target Responsibilities Need Not Be Uniform

Let $t = (0, 0.001, 1)$, $r = h = (1, 0, 0)$, and $S = \{1, 2\}$. Then $a = (1, 0, 0)$, the induced target masses are $b = (1, 0)$, and $D(S) = 0$. The admissible function $\phi = h$ has a source mean and a weighted target mean of one, but a uniform target mean of $1/2$. The extra term in Eq. (33) is therefore necessary: selecting K input frames does not impose equal transport responsibilities.

E Evaluation Protocol and Additional Comparisons

E.1 Datasets and Models

Video-MME contains 2,700 questions on 900 videos, and LVBench contains 1,549 questions on 103 videos. The MLVU development and LongVideoBench validation sets contain 2,174 and 1,337 questions, respectively. For QA accuracy, we evaluate every question in all four sets.

We evaluate all 1,549 LVBench questions, since the QA instances and their answer labels remain usable. For analyses that additionally require ground-truth temporal evidence annotations (Appendix G.3), we use the 1,476 questions with valid temporal intervals. The remaining annotations contain malformed or unusable intervals, such as missing endpoints (`08:17-None` or `None-None`),

end times preceding start times, or intervals extending beyond the duration of the local video. These temporal-annotation issues do not prevent the associated QA instances from being evaluated.

The starred foundation-model rows in Table 1 are taken from the original model papers and follow their native protocols rather than our $K = 32$ protocol. For example, the Qwen2.5-VL report allows up to 768 frames and 24,576 video tokens on video benchmarks but does not state the frame count behind its LVBench score of 45.3. We therefore exclude them from matched comparisons.

Video-MME, LongVideoBench, and MLVU are evaluated with LMMs-Eval (Zhang et al., 2025a) video tasks (for Video-MME, the `videomme` task without subtitles), and LVBench with a separate evaluator. The reported evaluation pipeline does not involve random sampling or other stochastic operations; therefore, no random seed is applicable. All runs use batch size one, and all selectors share the pixel budget of each answerer; for Qwen2.5-VL-7B, `max_pixels` is 12,845,056. The helpfulness observer is Qwen2.5-VL-3B-Instruct, and the answerers use the following checkpoints:

Answerer	Checkpoint
Qwen2.5-VL-7B (primary)	`Qwen/Qwen2.5-VL-7B-Instruct`
InternVL3-8B	`OpenGVLab/InternVL3-8B`
LLaVA-OneVision-7B	`lmms-lab/llava-onevision-qwen2-7b-ov`

Additional selector baselines. Table 7 reports AKS and BOLT, which are omitted from Table 1 for space. E-DoT ($B = 8$) is more accurate than both in every answerer–benchmark setting.

Table 7: **AKS and BOLT results** under the protocol of Table 1 ($K = 32$; accuracy, %). Avg. is the unweighted mean of the four benchmark scores, as in Table 1.

Answerer	Selector	V-MME	LVB	MLVU	LVBench	Avg.
Qwen2.5-VL-7B	AKS	63.1	63.2	67.2	43.4	59.2
	BOLT	63.2	60.0	68.4	39.3	57.7
InternVL3-8B	AKS	66.3	61.5	74.2	52.4	63.6
	BOLT	66.1	59.2	70.8	44.2	60.1
LLaVA-OneVision-7B	AKS	58.4	59.3	67.5	47.3	58.1
	BOLT	57.9	59.6	66.8	38.2	55.6

E.2 Comparison with a Training-Based Selector

We compare E-DoT with the training-based TSPO (Tang et al., 2026) under matched answerers and frame budgets. Table 8 reports accuracy on Video-MME and LongVideoBench. At $K = 32$, E-DoT exceeds TSPO by 1.9 and 0.9 points with Qwen2.5-VL-7B, and by 1.2 and 1.3 points with LLaVA-OneVision-7B, respectively. At $K = 64$, the corresponding gains are 1.6 and 0.3 points with Qwen2.5-VL-7B, and 1.0 and 0.8 points with LLaVA-OneVision-7B.

Table 8: **Comparison with training-based TSPO.** All entries are accuracy (%). V-MME: Video-MME; LVB: LongVideoBench. Answerers and frame budgets are matched within each block.

	$K = 32$		$K = 64$	
Model / selector	**V-MME**	**LVB**	**V-MME**	**LVB**
Qwen2.5-VL-7B	61.2	58.9	63.0	60.5
+ TSPO [AAAI'26]	63.5	63.9	65.0	65.2
+ **E-DoT** ($B = 8$)	**65.4**	**64.8**	**66.6**	**65.5**
LLaVA-OneVision-7B	56.5	54.8	58.8	57.9
+ TSPO [AAAI'26]	61.3	59.6	60.6	61.5
+ **E-DoT** ($B = 8$)	**62.5**	**60.9**	**61.6**	**62.3**

E.3 Information Access

The relevance scorer, the observer, and the answerer receive the question and its available answer choices, but never the correct-answer label. This inference-time restriction does not by itself rule out benchmark-informed design choices: benchmark results informed the development of E-DoT, and a single configuration of the greedy selector is shared across all four benchmarks.

E.4 Metrics

Video-MME, LongVideoBench, and LVBench use question accuracy. For MLVU, let $\mathcal{Q}_c$ be the questions in task c. The seven-task macro-average is

$$A_{\mathrm{MLVU}} = \frac{100}{7} \sum_{c=1}^{7} \frac{1}{|\mathcal{Q}_c|} \sum_{i \in \mathcal{Q}_c} \mathbf{1}\{\widehat{y}_i = y_i\}, \tag{37}$$

over the tasks topic reasoning, anomaly recognition, needle QA, ego reasoning, plot QA, action order, and action count. Paired differences use the same questions for both selectors; for example, on Video-MME at $B = 8$, 180 questions are answered correctly only with the directed cost and 119 only with the symmetric cost (Table 2(b)). Because differences are computed before rounding, they can differ in the last decimal from differences between rounded table entries.

E.5 Paired Significance against WFS-SB

Table 9 compares the main Qwen2.5-VL-7B results with WFS-SB on the same questions. We use a two-sided paired sign-flip randomization test and 100,000 paired bootstrap resamples with percentile intervals. For MLVU, question pairs are flipped or resampled within each task before macro-averaging the seven task differences. For Video-MME, LongVideoBench, and LVBench, the paired unit is a video: all its questions are flipped or resampled together. The intervals exclude zero on Video-MME, MLVU, and LVBench; the LongVideoBench interval includes zero.

Table 9: **Paired significance against WFS-SB.** Accuracy is in %. Δ is E-DoT minus WFS-SB in percentage points (pp), computed before rounding the displayed accuracies. The 95% CI is the percentile paired-bootstrap interval for Δ; p is from the two-sided paired sign-flip test.

Benchmark	WFS-SB	E-DoT	Δ (pp)	95% CI (pp)	p
Video-MME	64.2	65.4	+1.2	$[+0.296, +2.148]$	0.0124
LongVideoBench	64.4	64.8	+0.4	$[-0.847, +1.744]$	0.5725
MLVU	69.9	72.1	+2.2	$[+0.380, +3.868]$	0.0183
LVBench	48.1	51.3	+3.2	$[+0.673, +5.640]$	0.0176

E.6 Observer-Free Ablation Protocol

At $B = 0$, constant helpfulness leaves only directed relevance and temporal distance:

$$c_{ij}^{\rightarrow,(0)} = \max\{|t_i - t_j|, [r_i - r_j]_+\}\,. \tag{38}$$

The candidate pool, BLIP scores, final frame budget, and answerer are shared across Table 2(c)–(d). Unless ablated, the source masses follow Eq. (3). With $b = \pi^\top \mathbf{1}$ and $u_j = 1/K$, the refit family, which adds entropic regularization (Cuturi, 2013) and a KL penalty on the target marginal, solves

$$\min_{|S|=K,\ \pi \geq 0,\ \pi \mathbf{1} = a} \langle \pi, C_S \rangle + \lambda\, \mathrm{KL}(\pi \| a \otimes u) + \rho\, \mathrm{KL}(b \| u), \tag{39}$$

where C_S is the cost matrix restricted to the columns in S. The balanced and capacity controls replace the target-KL term by the constraints $b = u$ and $b_j \leq \alpha/K$, respectively. These constraints limit the source mass that each frame represents, not the number of selected frames.

The greedy $B = 0$ row in panel (a) uses a different support solver from the free-marginal refit in panel (d). All refits share two initializations (greedy and uniform over candidate indices) and

at most ten outer transport/medoid updates per initialization, with inner feasibility and optimality checks. Comparing the free-marginal refit with the entropy-only rows tests smoothing; comparing the entropy-only row at $\lambda = 0.005$ with the target-KL, balanced, and capacity rows tests target-marginal restrictions; and the $\lambda = 0.020$ row tests the entropy strength. The greedy symmetric-cost and duration-only-mass rows in panel (c) separately test directionality and relevance weighting. Reaching the outer-update limit does not imply convergence, and these local-solver results do not characterize the full family of unbalanced transport methods (Chizat et al., 2018).

E.7 TARGET-KL TRANSPORT

Observer-free accuracy. The target-KL control with $\rho = 1$ in Table 2(d) uses $\lambda = 0.005$, $B = 0$, and the matched refit. Its Video-MME, LongVideoBench, and MLVU accuracies (64.9%, 64.6%, and 71.6%) exceed the greedy $B = 0$ reference by 0.1, 0.2, and 1.0 points and the free-marginal refit by 0.1, 0.9, and 1.3 points. It leads the tested refit variants on Video-MME and MLVU and is 0.2 points below the best tested value on LongVideoBench, making it a competitive observer-free transport control. The tested ρ values and local refit budget do not establish its accuracy ceiling.

Target-mass prior and its limit. For $u_j = 1/K$, the penalty in Eq. (39) satisfies $\mathrm{KL}(b\|u) = \log K - H(b)$, where $H(b) = -\sum_j b_j \log b_j$. It encourages the represented source mass to spread across the selected frames. Free destination masses instead allow many nearby sources to use one strong evidence frame. Target KL therefore changes the objective rather than strictly strengthening the free-target selector: if the answer-bearing evidence is concentrated, the target prior can oppose this many-to-one assignment, although it can also regularize a poor concentration. The strength of this effect can be bounded without any claim about QA accuracy. Let J_S denote the inner objective of Eq. (39) for a fixed support S. The feasible product coupling $\pi = a \otimes u$ makes both KL terms zero and has a transport cost of at most one because $c^{\rightarrow}_{ij} \in [0, 1]$. Thus, an exact inner optimum $\pi^\star$ at $\rho > 0$ satisfies

$$\rho\,\mathrm{KL}(b^\star\|u) \le J_S(\pi^\star) \le J_S(a \otimes u) = \sum_i \sum_{j \in S} a_i u_j c^{\rightarrow}_{ij} \le 1, \qquad e^{H(b^\star)} \ge K e^{-1/\rho}. \tag{40}$$

At $K = 32$ and $\rho = 1$, the entropy-effective number of destinations is therefore at least $32/e \approx 11.8$ for an exact inner solve. This bounds how concentrated the induced transport responsibilities can be; it does not bound the accuracy of the selected frames or of the downstream answerer.

Optimization and runtime. The free-target objective decomposes into independent nearest-representative costs for each source (Appendix C.2), permitting residual updates and the greedy bound (Appendix C.3). The target-KL term couples the source assignments through b, so neither the reduction nor its greedy guarantee applies to the regularized objective. A dense implementation handles an $N \times K$ coupling in each transport solve, which the matched refit (Appendix E.6) repeats.

Table 10 summarizes the recorded selection timings and estimates of the observation cost for eight three-frame calls. The selection columns cover all 2,700, 1,337, 2,174, and 1,549 questions, respectively, from the same no-feedback run, with one OpenMP/BLAS thread per worker on a shared CPU host. Because the free-target greedy solve was performed before the per-variant timers, we reconstruct the shared cost as each question's total time minus the sum of its per-variant times, and add it to the free-target greedy or target-KL variant time. This shared cost includes the historical greedy-baseline replay, construction of the uniform initialization, loop and bookkeeping overhead, and input hashing; each per-variant time additionally includes its diagnostics. The resulting means are reconstructed selection-stage costs, not independently repeated latency benchmarks, and the free-target greedy column differs from the free-marginal refit control. Target KL uses $\rho = 1$, $\lambda = 0.005$, and $K = 32$.

Observer inference versus observation-service cost. The forward-only estimate measures Qwen2.5-VL-3B inference and the Yes/No readout after the input tensors are on the GPU. For each benchmark, a microbenchmark times two neighborhoods from one question, with two repetitions per neighborhood, warm-up, and CUDA synchronization on a shared GPU; we multiply the mean of the two medians by eight. Each call contains three images with a 224^2 pixel budget per image. This estimate isolates the observer's inference cost and excludes decoding, preprocessing, model loading, selection, and answering; it is not a full-dataset measurement. The broader service estimate

Table 10: **Selection and observation cost** in seconds per question. R: reconstructed from full-dataset timing records; E: estimated from separate observation measurements. Both selection columns use $B = 0$. The two observation columns describe alternative timing scopes for eight calls with three frames each and should not be summed. No column includes final question answering.

Dataset	Free-target greedy (R)	Observer forward (E)	Observer service (E)	Target KL $\rho = 1$ (R)
Video-MME	0.030	2.378	15.111	3.985
LongVideoBench	0.027	1.391	15.410	4.061
MLVU	0.049	1.336	18.160	3.375
LVBench	0.411	2.392	12.675	33.138

is eight times the total service duration of earlier observation runs divided by the actual number of three-frame calls, and also includes video I/O, preprocessing, model loading, and shared-service overhead. Because the two estimates come from different runs and sample scopes, their difference is not a measured preprocessing cost, and neither gives the end-to-end latency at $B = 8$.

E.8 Feedback on the Refit Controls

Table 11 crosses the support optimizer with feedback availability. At $B = 8$, the free-marginal and target-KL refits reuse the cached helpfulness values of greedy E-DoT, with the same two initializations, at most ten outer updates, and $\lambda = 0.005$ for target KL. Feedback improves all three optimizers on every benchmark. With feedback, target KL is 0.2 points below greedy E-DoT on Video-MME, tied on LongVideoBench, and 0.4 points higher on MLVU, whereas the free-marginal refit stays below it on all three benchmarks. Greedy E-DoT also benefits most from feedback on Video-MME and MLVU (0.6 and 1.5 points, compared with 0.2–0.3 and 0.5–0.9 points for the refits).

Table 11: **Support optimizer versus feedback availability** ($K = 32$; accuracy, %). Refit variants share the same support-search budget; target KL uses $\rho = 1$ and $\lambda = 0.005$.

Support optimization	B	V-MME	LVB	MLVU
Free-target greedy (E-DoT)	0	64.8	64.4	70.6
Free-target greedy (E-DoT)	8	65.4	64.8	72.1
Free-marginal refit	0	64.8	63.7	70.3
Free-marginal refit	8	65.0	64.2	70.8
Target KL ($\rho = 1$) refit	0	64.9	64.6	71.6
Target KL ($\rho = 1$) refit	8	65.2	64.8	72.5

Why retain the free-target greedy backbone? Target KL is a strong alternative both with and without feedback. We retain the free-target greedy selector because its objective decomposes into independent nearest-representative costs, which permit inexpensive residual updates and the greedy guarantee, and because it lets a strong representative absorb redundant source mass without a uniform target-mass prior. In the recorded $B = 0$ runs, its reconstructed mean selection cost is 0.027–0.049 seconds per question on the first three benchmarks and 0.411 seconds on LVBench, compared with 3.375–4.061 and 33.138 seconds for target KL (Table 10); adding feedback does not remove the iterative refit that target KL requires. These timings support an implementation choice rather than an inherent lower bound on target-KL runtime. Moreover, comparing observer forward time with CPU selection time alone does not establish an end-to-end speed advantage for E-DoT, and we do not report synchronized end-to-end latency at $B = 8$ (Appendix H).

E.9 Observer-Matched Comparisons

Table 12 tests whether the additional Qwen2.5-VL-3B calls benefit E-DoT more than simpler uses of the same observations. All $B = 8$ rows use the same observer, per-question readouts, questions, $K = 32$ final frames, and answerer settings. The direct controls reuse the cached readouts, candidate

Table 12: **Observer-matched comparisons** with Qwen2.5-VL-7B ($K = 32$; accuracy, %). B caps observer calls; bold marks the best result in each column.

Selector	B	V-MME	LVB	MLVU	LVBench
E-DoT	0	64.8	64.4	70.6	48.9
E-DoT	8	**65.4**	**64.8**	**72.1**	51.3
BLIP relevance Top-K	0	62.7	64.0	70.8	50.6
BLIP $r_i + h_i$ Top-K	8	63.6	64.3	71.1	**51.4**
Top-4 helpful neighborhoods + uniform fill	8	62.4	62.8	68.9	47.8
Top-4 helpful neighborhoods + BLIP fill	8	62.9	64.6	70.7	50.9
WFS-SB (native)	0	64.2	64.4	69.9	48.1
WFS-SB + shared-observer reranking (ours)	8	64.6	64.6	70.8	49.7

grid, and BLIP relevance scores of E-DoT. Direct fusion (BLIP $r_i + h_i$ Top-K) ranks candidates by $r_i + h_i$ without coverage updates. The neighborhood controls sort the observed neighborhoods by helpfulness, take the three frames of each of the top four (removing duplicates), and fill the remaining slots from unselected candidates either uniformly in time or by BLIP relevance. WFS-SB retains the same native BLIP-2-ITM scorer and preprocessing in both of its rows. For WFS-SB + shared-observer reranking, our adaptation keeps WFS-SB's wavelet boundary detection on its native BLIP-2-ITM relevance scores r_i, its segment scoring, and its budget allocation, and linearly interpolates the same cached readouts of E-DoT onto its candidate timeline to obtain h_i. Only in WFS-SB's final within-segment maximal-marginal-relevance (MMR) selection (Carbonell & Goldstein, 1998) is the centered relevance $2r_i - 1$ replaced by $(2r_i - 1) + (2h_i - 1)$; the diversity term, MMR coefficient, segment boundaries, and all other hyperparameters are unchanged. It therefore reduces exactly to native WFS-SB when $h_i = 0.5$.

E-DoT ($B = 8$) is the most accurate row on Video-MME, LongVideoBench, and MLVU, whereas direct fusion is 0.1 points higher on LVBench (51.4% versus 51.3%). Feedback also improves both BLIP Top-K and WFS-SB on every benchmark, so the observer signal is useful beyond E-DoT; however, coverage-aware transport uses it more effectively on three of the four benchmarks. Without feedback, BLIP relevance Top-K is slightly more accurate than observer-free E-DoT on MLVU (70.8% versus 70.6%) and more accurate on LVBench (50.6% versus 48.9%).

F Additional Analyses

F.1 Question-Type Analysis

The left panel of Figure 3 compares E-DoT ($B = 8$) and WFS-SB by question type, using InternVL3-8B at $K = 32$ on all 2,700 Video-MME questions. Their overall accuracies, 68.1% and 66.3%, match Table 1. E-DoT is more accurate in most categories, especially counting and OCR.

F.2 Frame Budget

The upper-right panel of Figure 3 varies the answerer frame budget on Video-MME with a Qwen2.5-VL-7B answerer, and Table 13 lists the plotted values. At $K = 8$, Uniform and E-DoT answer 1,436 and 1,624 of the 2,700 questions correctly. E-DoT is more accurate than Uniform and BOLT at every budget, and feedback improves over the observer-free variant at every budget. E-DoT with 8 frames exceeds Uniform with 16 frames by 2.5 points (computed before rounding); the corresponding 16-versus-32 and 32-versus-64 gaps are 1.3 and 2.4 points. These gaps concern only the answerer's input frames, not total observations, model calls, or runtime. The frame-budget, question-type, and duration analyses were conducted in separate evaluation runs.

F.3 Duration-Stratified Comparison

Figure 6 compares E-DoT and Uniform within duration groups, using a Qwen2.5-VL-7B answerer. On Video-MME, the gain grows from 1.8 points on short videos to 3.7 and 6.0 points on medium and

Table 13: **Video-MME accuracy (%)** versus the answerer frame budget K with Qwen2.5-VL-7B, as plotted in Figure 3. E-DoT uses $B = 8$ feedback, and K counts only frames sent to the answerer. Gain is E-DoT minus Uniform in percentage points, computed before rounding.

K	Uniform	BOLT	E-DoT w/o helpfulness	E-DoT	Gain
8	53.2	56.6	59.7	60.1	+7.0
16	57.7	59.9	61.8	62.5	+4.9
32	61.2	63.2	64.8	65.4	+4.2
64	63.0	64.6	65.7	66.6	+3.6

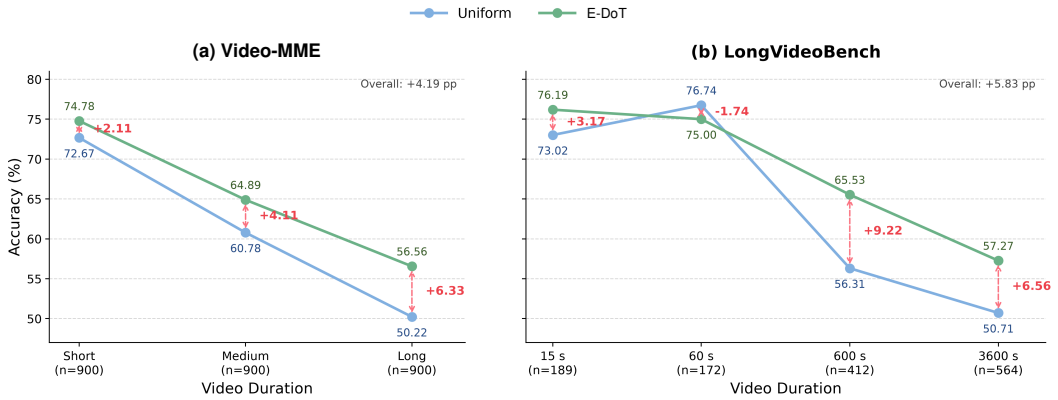


Figure 6: Accuracy of Uniform and E-DoT by video duration with Qwen2.5-VL-7B at $K = 32$. Video-MME has 900 questions per group; LongVideoBench groups have 189, 172, 412, and 564.

long videos. On LongVideoBench, the gains in the four duration groups are 3.2, -1.7, 9.2, and 6.6 points, so the improvement is concentrated in the two longest groups. The LongVideoBench group accuracies aggregate to 58.9% and 64.8%, consistent with Table 1.

G Direct Evaluation of Selected Evidence

We complement QA accuracy with two direct evaluations of the selected frames: scene coverage on KFS-Bench and temporal hits against LVBench reference intervals. All selectors use the answerer budget $K = 32$. The $B = 8$ variant of E-DoT additionally uses up to eight three-frame observer calls; in this section, B denotes the observer-call budget of E-DoT.

G.1 KFS-Bench Scene Evidence

KFS-Bench (Li et al., 2026) annotates answer-relevant scenes for Video-MME and LongVideoBench questions. We evaluate on the intersection of its annotated questions with our evaluation sets, requiring $K = 32$ distinct final frames with valid timestamps. After excluding questions for which the uniform baseline yields fewer than 32 distinct frames (3 in Video-MME and 187 in LongVideoBench), the intersection contains 885 Video-MME and 1,104 LongVideoBench questions. Scene annotations are used only for evaluation and never enter the relevance scorer, observer, selector, or answer prompt.

We report the official per-question Key Frame Rate (KFR), Balanced Scene Recall (BSR), and Balanced Distribution Similarity (BDS), averaged over questions. Following the official aggregation, the Unified Keyframe Sampling Score (UKSS) over a set $\mathcal{E}$ of evaluated questions is

$$\mathrm{UKSS} = \exp\left[\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \log\max\left\{0.01, (\mathrm{KFR}_e\mathrm{BSR}_e\mathrm{BDS}_e)^{1/3}\right\}\right]. \tag{41}$$

Per-question values are computed with the benchmark implementation and are independent of QA correctness. Uniform uses the frames actually selected by our uniform baseline, and all selectors are

evaluated on the same questions. WFS-SB retains its own candidate preprocessing, so it is compared at the level of final frames rather than as a shared-candidate ablation, whereas the two E-DoT rows isolate the effect of helpfulness readouts within the same selector.

Table 14 shows that E-DoT ($B = 8$) attains the highest BSR and BDS on both benchmarks and the highest UKSS on Video-MME. BLIP Top-K attains the highest KFR on both benchmarks and the highest UKSS on LongVideoBench. Feedback improves the UKSS of E-DoT on both benchmarks.

Table 14: **Scene-level evaluation of the selected frames on KFS-Bench** ($K = 32$; 885 Video-MME and 1,104 LongVideoBench questions). Higher is better for every metric, and bold marks the best value in each column; B caps observer calls.

		Video-MME				LongVideoBench			
Selector	B	**KFR**	**BSR**	**BDS**	**UKSS**	**KFR**	**BSR**	**BDS**	**UKSS**
Uniform	0	0.179	0.712	0.755	0.198	0.084	0.483	0.532	0.074
BLIP Top-K	0	**0.308**	0.702	0.755	0.248	**0.203**	0.709	0.757	**0.213**
WFS-SB	0	0.255	0.755	0.802	0.277	0.159	0.698	0.744	0.192
E-DoT	0	0.246	0.798	0.840	0.303	0.150	0.704	0.754	0.188
E-DoT	8	0.246	**0.804**	**0.845**	**0.309**	0.155	**0.728**	**0.777**	0.205

G.2 Single- and Multi-Scene Questions

We split the same questions by the number of distinct annotated scene IDs: questions with one scene ID are single-scene, and those with more than one are multi-scene; disjoint temporal segments with the same ID count as one scene. Within each group, UKSS is recomputed from the per-question scores with Eq. (41), and QA accuracy is computed on the same questions.

Among the selectors in Table 15, E-DoT ($B = 8$) attains the highest UKSS in all four groups. The accuracy gains of E-DoT over WFS-SB, however, appear only on LongVideoBench multi-scene questions (3.6 and 3.3 points at $B = 0$ and $B = 8$); on Video-MME ones, E-DoT ($B = 8$) ties WFS-SB, and on single-scene ones, WFS-SB is more accurate on both benchmarks. The split is descriptive: without paired uncertainty estimates, it cannot attribute these differences to one component.

Table 15: **QA accuracy (%) and UKSS** on single- and multi-scene KFS-Bench questions with Qwen2.5-VL-7B-Instruct ($K = 32$). Video-MME contains 612 single-scene and 273 multi-scene questions; LongVideoBench contains 802 and 302. Questions for which the uniform baseline yields fewer than 32 distinct frames are excluded for all selectors. Bold marks the best value in each column.

		Video-MME				LongVideoBench			
		Single		**Multi**		**Single**		**Multi**	
Selector	B	**Acc.**	**UKSS**	**Acc.**	**UKSS**	**Acc.**	**UKSS**	**Acc.**	**UKSS**
Uniform	0	68.0	0.200	47.6	0.193	60.0	0.061	48.0	0.127
WFS-SB	0	**70.6**	0.279	**55.3**	0.273	**67.3**	0.173	54.0	0.252
E-DoT	0	70.3	0.295	53.5	0.322	65.0	0.165	**57.6**	0.268
E-DoT	8	69.9	**0.302**	**55.3**	**0.324**	65.5	**0.183**	57.3	**0.278**

G.3 LVBench Temporal References

This analysis uses the 1,476 LVBench questions with valid temporal reference intervals (Appendix E.1); QA accuracy is still evaluated on all 1,549 questions. A direct hit means that at least one selected frame's timestamp lies inside a reference interval, endpoints included. For each question, the nearest gap is the minimum distance in seconds from any selected timestamp to any reference interval, which is zero for a hit; we report the hit rate and the mean gap over all 1,476 questions. A temporal hit does not imply that the selected frame alone suffices to answer the question.

As Table 16 shows, E-DoT ($B = 8$) attains the highest hit rate and the smallest mean gap. BLIP Top-K matches the hit rate of observer-free E-DoT but has a much larger mean gap (71.81 versus 9.45 seconds), indicating that its misses lie much farther from the reference evidence.

Table 16: **Temporal localization on LVBench** ($K = 32$; 1,476 questions with valid reference intervals). Higher hit rates and lower mean gaps are better; bold marks the best value in each column.

Selector	B	**Direct hit (%)** ↑	**Mean nearest gap (s)** ↓
Uniform	0	35.1	19.60
BLIP Top-K	0	55.2	71.81
WFS-SB	0	54.3	37.30
E-DoT	0	55.2	9.45
E-DoT	8	**55.4**	**8.94**

H Limitations

E-DoT represents evidence by time, relevance, and, in its full configuration, sparse local-helpfulness estimates. These coordinates do not explicitly encode event identity or relations across frames, and interpolation between sparse readouts may miss brief events. Shared coordinate scales across videos also limit adaptation to tasks that require a different balance between temporal proximity and evidence preservation. The full selector makes up to eight Qwen2.5-VL-3B calls per question that the main-table baselines do not make. We mitigate this with the observer-free variant, which exceeds WFS-SB on average for all three answerers without any observer calls, and with observer-matched controls (Appendix E.9), where simpler uses of the same calls are less accurate on three of four benchmarks. Paired confidence intervals for InternVL3-8B and LLaVA-OneVision-7B, whose smallest margins in Table 1 are 0.1–0.2 points, and synchronized end-to-end latency are left to future work.

I Qualitative Case Studies

Four cases show E-DoT ($B = 0$ or $B = 8$) correcting AKS or WFS-SB; two compare the E-DoT variants when feedback changes the answer. Each panel shows two selected frames, while predictions and counts use all $K = 32$. KFS-Bench bands mark reference scenes; blue and green traces show BLIP relevance and $B = 8$ helpfulness. Dark ticks mark exclusive frames and pale ticks shared frames. Hits use exact timestamps; displayed times are rounded.



Figure 7: **Counting people on the staircase (Video-MME 246-3).** Both selections reach the reference scene, but $B = 0$ retains six frames in it versus three for AKS. AKS predicts four people; E-DoT correctly predicts five.



Figure 8: **Recognizing the first action after the exam (Video-MME 551-2).** Both selections reach S1, but AKS retains three reference frames and predicts cleaning the floor. The $B = 8$ selection retains fourteen reference frames and correctly predicts sorting out clothes.



Figure 9: **Reading the tie color (LongVideoBench).** WFS-SB selects no frame in the annotated intervals and predicts white. E-DoT ($B = 0$) retains two reference frames and correctly predicts black.

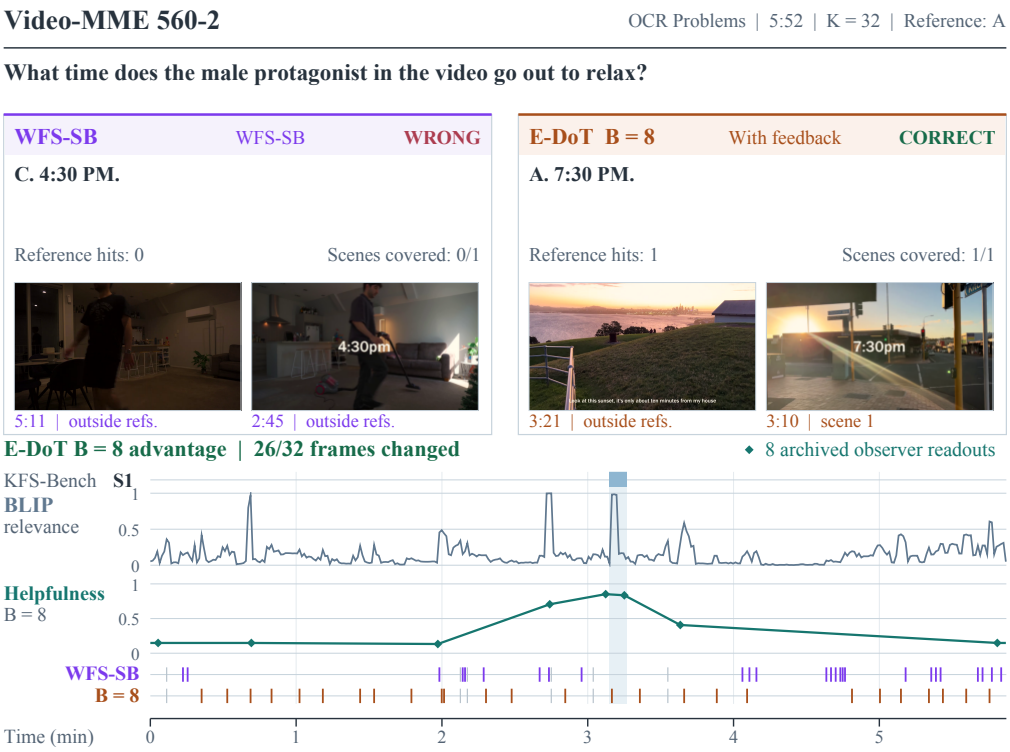


Figure 10: **Reading the time of going out (Video-MME 560-2).** WFS-SB misses the reference interval and predicts 4:30 PM. The $B = 8$ selection includes the displayed 7:30 PM frame at 3:10, retains one reference frame, and predicts the correct answer.

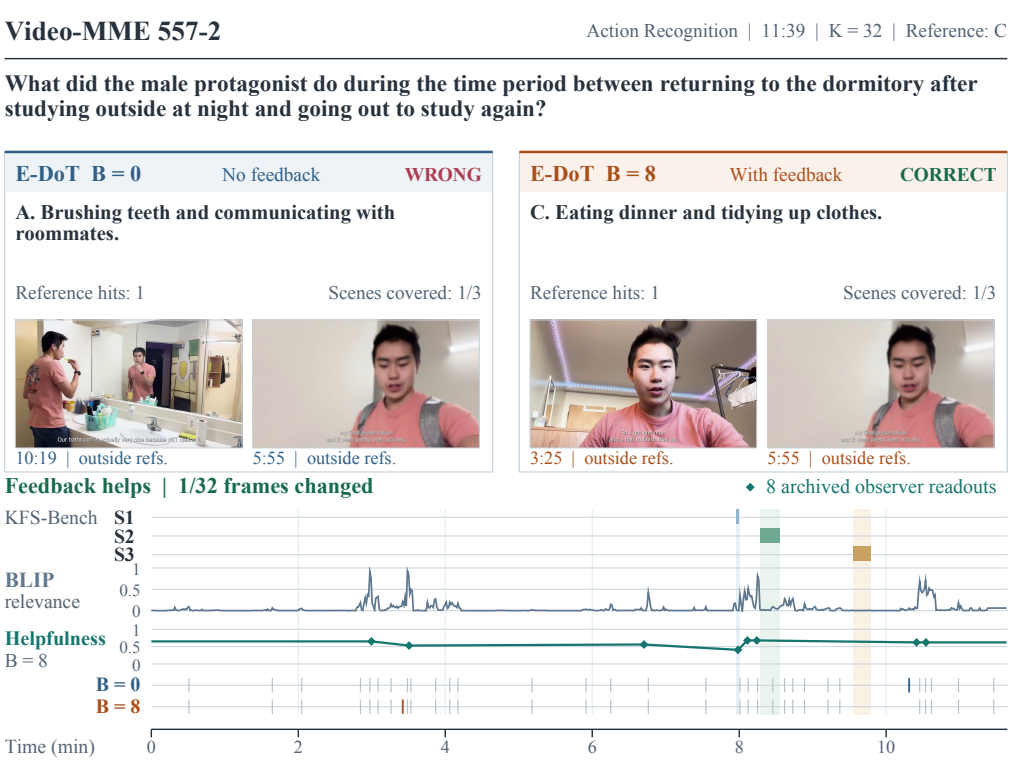


Figure 11: **Feedback changes the answer with one frame replaced (Video-MME 557-2).** The selections share 31 of 32 frames and each covers one of three reference scenes. $B = 0$ predicts brushing teeth and communicating with roommates; $B = 8$ correctly predicts eating dinner and tidying up clothes.



Figure 12: **Recovering regular water breaks (Video-MME 887-3).** The $B = 8$ selection retains 26 reference frames versus 25 for $B = 0$; six of their 32 frames differ. $B = 0$ predicts no regular breaks, whereas $B = 8$ correctly predicts a break every five minutes.