MOMENTROUTE: ASYMMETRIC RESOLUTION ROUTING FOR BUDGETED VISUAL MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Budgeted visual memory must search a broad video history while preserving the visual detail needed to answer a query. We show that these two tasks have different resolution requirements: across two cross-modal rerankers, reducing candidate resolution from 448^2 to 224^2 lowers selection recall by only 0.23–0.26 percentage points, whereas reducing three selected frames to 224^2 lowers the 72B reader's conditional answer accuracy by 10.4 points. MomentRoute exploits this asymmetry by separating candidate localization from evidence reading: it reranks 32 thumbnails in the cloud, then requests three 448^2 frames for answering. On 1,200 constructed Ego4D-NLQ QA episodes, this allocation reaches 65.1% accuracy, compared with 60.1% for direct thumbnail upload and 58.4% for compressed full-buffer video. Relative to adaptive edge selection with three high-resolution frames, cross-modal reranking improves QA by $+5.2\,pp$ ($p < 10^{-4}$). Accuracy–resource curves expose the accompanying trade-offs: four-frame edge selection reaches 64.1% at 1,009 KiB, compared with 65.1% at 1,136 KiB for two-stage routing (exact McNemar $p = 0.323$). Edge selection provides lower TTFT; two-stage routing uses 25% fewer reader visual tokens (768 vs. 1,024). Together, these measurements show how allocating resolution by processing stage preserves broad candidate search and detailed evidence reading, while making the communication and inference trade-offs explicit. Code and benchmark data will be made publicly available.

1 INTRODUCTION

Smart glasses and mobile robots answer retrospective questions such as *"Where did I leave my keys?"* from an on-device video history (Huang et al., 2025; Tang et al., 2026; Grauman et al., 2022). At query time, uplink bandwidth and cloud inference costs constrain the evidence that can be transmitted. A successful system must find the relevant event, preserve the visual detail needed to answer, and keep the reader's input short enough for interactive use. How should resolution be allocated between *searching broadly* and *reading precisely* when both upload and reader input are constrained?

The natural alternatives reveal a tension. Selecting a few high-resolution frames preserves detail, but ranking frozen embeddings can allocate several slots to the same distractor or miss a relevant event deeper in the shortlist. Sending the entire shortlist as thumbnails, or compressing the full buffer as video, improves temporal coverage per transmitted byte. Yet in our evaluated 72B reader, **Direct-32** achieves 80.52% temporal coverage and 60.1% QA accuracy, while **H.265-30s** has 100% resident temporal coverage and 58.4% QA accuracy. These results motivate evaluating the evidence that a reader can use, alongside the evidence that is merely present in its input.

Our central empirical finding is a **resolution asymmetry between localization and reading**. Across BLIP-2 ITM and Qwen2-VL-2B, replacing 448^2 reranking inputs with 224^2 thumbnails reduces selection recall by only 0.23–0.26 percentage points. Sweeping resolution reveals a sharp knee at 224^2: dropping further to 112^2 degrades reranking recall by 5.06 points. By contrast, reducing three selected frames to 224^2 lowers the 72B reader's conditional answer accuracy by 10.4 points, as downstream answering degrades smoothly with pixel density. This distinction suggests decoupling

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.

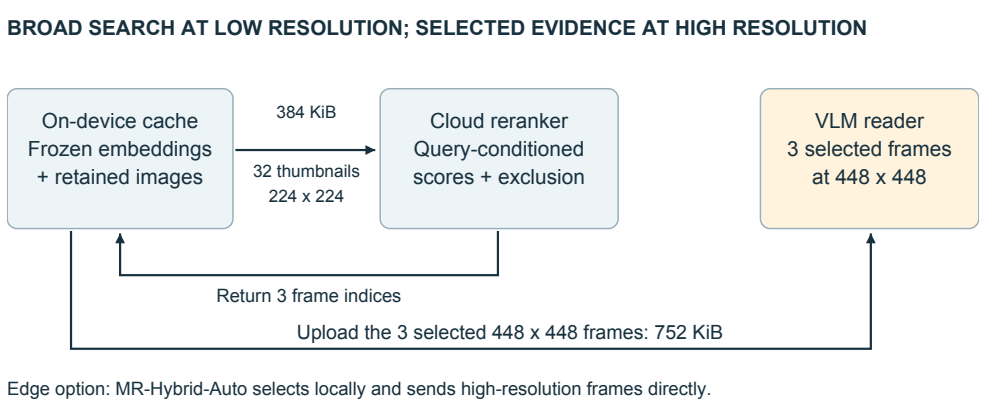

Figure 1: **Allocate resolution by task.** MomentRoute searches 32 candidates at 224^2, then supplies selected evidence at 448^2 to the reader. Reranking and exclusion precede the three-frame upload cap. MR-Hybrid-Auto provides an edge-only alternative (Appendix Figure 3).

resolutions: thumbnails for candidate search, followed by high-resolution delivery of the selected evidence.

MomentRoute-2Stage implements this allocation (Figure 1). It uploads 32 thumbnails for cross-modal reranking and then requests three 448^2 frames for the answer model. Reranking improves which evidence reaches the reader, while high-resolution delivery preserves the detail needed to use it. On 1,200 constructed Ego4D-NLQ QA episodes, the resulting pipeline reaches 65.1% accuracy: $+5.0\,pp$ over Direct-32 and $+6.7\,pp$ over H.265-30s. The comparison with **MR-Hybrid-Auto** isolates the selection benefit at the same three-frame reader input size: QA rises from 59.9% to 65.1% ($+5.2\,pp$, $p < 10^{-4}$). This gain comes with additional routing and communication: total upload is 1,136 KiB, compared with 384 KiB for Direct-32.

The accuracy–resource curves characterize when this allocation is useful. **MR-Hybrid-Auto** with four reader frames reaches 64.1% QA at 1,009 KiB and 628.0 ms TTFT. Two-stage routing reaches 65.1% at 1,136 KiB and 814.9 ms, using 25% fewer reader visual tokens (768 vs. 1,024); the 1.0 pp QA difference is not significant (exact two-sided McNemar $p = 0.323$). **MomentRoute** thus provides complementary edge and cloud routing paths, with measured trade-offs among upload, reader input, and response latency. Reader input size is also relevant to serving and KV-cache budgets (Šabanović et al., 2026; Selvam et al., 2025; Zhang et al., 2026). On 180-s Ego4D, two-stage routing improves QA by $+6.1\,pp$ over edge selection; the observed $+4.2\,pp$ on EgoSchema is not significant at $n = 210$.

We focus our downstream QA evaluation on the resident rolling cache (30 s to multi-minute windows), where candidate filtering and transmission latency dominate. For multi-hour or day-scale retrieval, cache eviction ($1 - O(K)$) is the primary error mode, which requires tiered episodic indexing rather than single-buffer routing alone. Two complementary analyses explain the design: rank diagnostics across six encoders distinguish candidate availability from the evidence recovered by a small-budget selector, and **MR-Hybrid-Auto** calibrates temporal and visual exclusion without target-domain retuning. The contributions are:

1. **Asymmetric resolution routing.** We quantify the resolution asymmetry between candidate localization and evidence reading: thumbnail reranking loses only 0.23–0.26 percentage points of recall across two reranker families, while thumbnail reading loses 10.4 points of conditional answer accuracy. **MomentRoute-2Stage** exploits this distinction to reach 65.1% measured 72B QA. Comparisons across frame budgets characterize the associated upload, reader-token, and latency trade-offs (Sections 4.1 and 5).
2. **Representation and selection bottlenecks.** We distinguish top-L candidate coverage $O(L)$ from the budget-constrained exclusion oracle $O_M^{\mathrm{excl}}(L)$. On 30-s Ego4D at $M = 3$, the latter reaches 63.84% versus 80.52% candidate coverage, while per-episode Oracle-MMR reaches 64.42%. Rank diagnostics across six backbones identify false-positive crowding and near-chance ordering within ranks 11–32 as the primary bottleneck (Section 2.1).
3. **Adaptive edge calibration.** Query-score autocorrelation and adjacent-frame visual similarity calibrate exclusion without target-domain labels or per-horizon retuning. **MR-Hybrid-Auto** reaches 60.09% recall with 0.03 ms selection overhead, providing the complementary edge selection path (Section 4.3; Table 9).

2 Problem Formulation and Bottleneck Analysis

We first separate three questions: whether the candidate pool contains the event, whether a budgeted selector retrieves it, and whether the reader answers correctly given that evidence. The first two motivate query-conditioned selection; the third motivates preserving resolution for the final reader.

At query arrival T_q, an edge cache holds $B_q = \{(t_i, v_i, x_i)\}_{i=1}^{n_q}$ ($n_q \le K$), where $t_i \in [T_q - T_q^{\text{buf}}, T_q]$ is the frame timestamp, $v_i \in \mathbb{S}^{d-1}$ is a frozen visual embedding, and x_i is the compressed JPEG crop. Given query embedding $q \in \mathbb{S}^{d-1}$ with cosine similarity $s_i = q^\top v_i$, a selector π uploads $A_q = \pi(B_q, q) \subseteq \{1, \dots, n_q\}$ ($|A_q| \le M$). Let $G_q = \{i : t_i \in [t_s^*, t_e^*]\}$ denote the ground-truth frame indices and $r_q(A_q) = \mathbb{I}[A_q \cap G_q \neq \emptyset]$ be the binary any-hit indicator.

2.1 Candidate Coverage and Budget-Constrained Selection

Let $\mathcal{P}_L(q) = \{i_{(1)}, \dots, i_{(L)}\}$ denote the top-L frames ranked by s_i, and let $O(L) = \mathbb{E}_q[\mathbb{I}(\mathcal{P}_L(q) \cap G_q \neq \emptyset)]$ be the unconstrained rank-L prefix oracle ($O(K) = 100\%$ on resident buffers, and $O(K) < 100\%$ under cache eviction). Any greedy $(\delta, \tau_{\text{vis}})$-exclusion selector allocates its M budget slots in descending score order across distinct peaks and therefore cannot reach a target frame at rank $r \in [11, 32]$ whenever ranks 1–10 already contain $\ge M$ mutually isolated false-positive peaks. We define the **M-constrained greedy $(\delta, \tau_{\text{vis}})$-exclusion oracle** $O_M^{\text{excl}}(L) = \mathbb{E}_q[\max_{\delta>0, \tau_{\text{vis}} \in (0,1]} r_q(\text{NMS}_M(\mathcal{P}_L(q); \delta, \tau_{\text{vis}}))]$, the upper bound achievable by any per-episode optimal exclusion rule over $(\delta, \tau_{\text{vis}})$ (non-monotone re-scoring rules such as MMR, DPP, and temporal smoothing lie outside this definition and are evaluated via Oracle-MMR(L) below), decomposing the miss rate $1 - R(\pi)$ into four exact additive terms:

$$1 - R(\pi) = \underbrace{1 - O(L)}_{\substack{\text{(I) Rank-}L\text{ Miss} \\ 19.48\%\ (L{=}32)}} + \underbrace{O(L) - O_M^{\text{excl}}(L)}_{\substack{\text{(II) Top-10 FP Crowding} \\ 16.68\%\ (L{=}32,\ M{=}3)}} + \underbrace{O_M^{\text{excl}}(L) - O_M^{\text{excl}}(10)}_{\substack{\text{(III) Mid-Rank Gain} \\ 2.39\%\ (L{=}32,\ M{=}3)}} + \underbrace{O_M^{\text{excl}}(10) - R(\pi)}_{\substack{\text{(IV) Top-10 Calib. Loss} \\ 1.36\%\ \text{(MR-Hyb-Auto)}}}. \quad (1)$$

Top-10 false-positive crowding and comparison to non-greedy Oracle-MMR. Equation 1 explains why exclusion selectors saturate at $L^* \approx 10$–15 despite $O(32)$ reaching 80.52%. Across the 16,095 30-s Ego4D buffers (mean 13.2 ground-truth frames/buffer), 64.2% of all ground-truth frames lie at cosine rank $r > 10$. Among episodes where the highest-scoring ground-truth frame falls in ranks 11–32, in 86.2% of cases the 10 top-ranked frames already contain ≥ 3 mutually isolated false-positive clusters ($|t_i - t_j| > 3.2$ s and $v_i^\top v_j < 0.85$), exhausting the budget before reaching mid ranks. Crucially, within-band cosine ordering across ranks 11–32 collapses to near-chance (AUROC$_{11\text{–}32} = 0.512$). Evaluating the per-episode optimal non-greedy MMR oracle Oracle-MMR(32) confirms that this ceiling is a persistent limitation across the evaluated selector families: even choosing λ_q^* with hindsight per episode reaches 64.42% at $M{=}3$, exceeding $O_3^{\text{excl}}(32) = 63.84\%$ by only $+0.58\,pp$ and leaving $16.10\,pp$ of $O(32)$ unrecovered by these evaluated selector families.

Implications for zero-label, supervised, and cross-modal reranked selection. Measured against $O_3^{\text{excl}}(32) = 63.84\%$, zero-label MR-Hybrid-Auto (60.09% at $M{=}3$, 67.35% at $M{=}4$) captures 94.1% of the exclusion ceiling (and 93.3% of non-greedy Oracle-MMR(32)). In-domain neural selectors equipped with δ-exclusion recover 45.1%–57.3% of the remaining $+3.75\,pp$ gap to $O_3^{\text{excl}}(32)$, whereas query-conditioned cross-modal rerankers (BLIP-2 ITM and Qwen2-VL-2B) address Term (II) ($+16.68\,pp$) by lifting mid-rank AUROC from 0.512 to 0.665–0.714. Conditioning VLM answer correctness $Z_q \in \{0, 1\}$ on $r_q(A_q)$ further links recall and crop resolution to downstream QA accuracy:

$$\text{Acc}(\pi) = \Pr_\pi(Z_q = 1 \mid r_q = 0) + \big[\Pr_\pi(Z_q = 1 \mid r_q = 1) - \Pr_\pi(Z_q = 1 \mid r_q = 0)\big] R(\pi). \quad (2)$$

3 Method: Horizon Calibration and Asymmetric Resolution Routing

Equations 1–2 define two complementary routing paths (**MomentRoute**): **MR-Edge** uses closed-form $(\hat{\tau}_{\text{corr}}^{(q)}, Q_{0.85})$ to saturate 94.1% of $O_3^{\text{excl}}(32)$ with zero-label edge execution, while **MR-2Stage**

decouples 224^2 cloud L4 reranking from $M{=}3$ high-resolution 448^2 VLM reading. In MR-2Stage, the edge sends the top-32 cosine shortlist to the cloud without prior exclusion; the cloud reranker scores all 32 candidates, and $(\hat{\tau}^{(q)}_{\text{corr}}, Q_{0.85})$ exclusion is applied *after* reranking to select the final M frames. Applying exclusion before cloud upload would prematurely discard mid-rank true positives that the dual-encoder ranked poorly, losing $4.2\,pp$ recall (Appendix Table 20).

Algorithm 1 MomentRoute ($\tau_{\text{vis}}{=}1$), MR-Auto ($\tau{=}-\infty$, $\delta_q{=}\hat{\tau}^{(q)}_{\text{corr}}$), and MR-Hybrid-Auto ($\tau^{(q)}_{\text{vis}}{=}Q_{0.85}$)

Input: cache $\mathcal{B}_q{=}\{(t_i, v_i, x_i)\}_{i=1}^{n_q}$ (f_s), query x, cap M, threshold τ $(-\infty)$, gap δ_q, visual ceiling $\tau^{(q)}_{\text{vis}}$.
Output: ordered frame indices $A_q \subseteq \{1, \ldots, n_q\}$ with $|A_q| \le M$.

$q \leftarrow \text{normalize}(E_{\text{text}}(x))$; $\quad s_i \leftarrow q^\top v_i$; $\quad \delta_q \leftarrow \inf\{\Delta t > 0 : \hat{\rho}_s(\Delta t) \le 1/e\}$.
$\tau^{(q)}_{\text{vis}} \leftarrow Q_{0.85}(\{v_i^\top v_{i+1}\}_{i=1}^{n_q-1})$; $\quad \sigma \leftarrow \text{stable_argsort}_{\downarrow}(s)$; $\quad A_q \leftarrow ()$.
for $i \in \sigma$ **do if** $s_i < \tau$ **then break**; **if** $A_q \neq \varnothing \wedge \big(\min_{j \in A_q} |t_i - t_j| < \delta_q \vee \max_{j \in A_q} v_i^\top v_j > \tau^{(q)}_{\text{vis}}\big)$ **then continue**;
  $A_q \leftarrow A_q \,\|\, (i)$; $\quad$ **if** $|A_q| = M$ **then break**.
return A_q (448^2 upload in **MR-Edge**, or L4-rerank $\mathcal{P}_{32}$ at 224^2 then upload A_q at 448^2 in **MR-2Stage**).

Complementary geometry of temporal $\hat{\tau}^{(q)}_{\text{corr}}$, visual $Q_{0.85}$, and asymmetric resolution. Algorithm 1 pairs fixed-hyperparameter suppression ($\rho^*{=}1/e, Q_{0.85}$, set on Ego4D train) with asymmetric reranking: $\delta_q = \hat{\tau}^{(q)}_{\text{corr}}$ sets the exclusion radius to the $1/e$ autocorrelation decay time to suppress burst duplicates without per-horizon tuning; $\tau^{(q)}_{\text{vis}} = Q_{0.85}$ suppresses non-adjacent visual duplicates when camera pans revisit the same workspace at $|t_i - t_j| > \delta_q$. On 30-s buffers ($M{=}3$), three separated peaks span median 14.8 s (3.1% revisit collision, $+0.15\,pp$: $59.94\% \to 60.09\%$). Adding a fourth slot ($M{=}4$) forces selection into secondary peaks revisiting the scene in 18.4% of episodes ($+1.47\,pp$: $65.88\% \to 67.35\%, p = 0.0002$). Expanding to 180 s lifts repeated visits to 24.6% at $M{=}3$ ($+1.64\,pp$: $41.04\% \to 42.68\%$) and 31.2% at $M{=}4$ ($+2.85\,pp$: $42.68\% \to 45.53\%$; $+5.06$–$6.10\,pp$ on controlled revisits; Appendix E).

4 Experimental Evaluation

Benchmarks, train-split protocol, and statistical tests. We evaluate visual memory routing across four benchmarks spanning cache capacities $K \in [25, 900]$ (30–600 s): **Ego4D-NLQ** (Grauman et al., 2022) ($n{=}16,095$ val across 1,089 videos; $n{=}34,049$ train) across 30-s resident (5 Hz, $K{=}150$), 30-s non-resident ($\Delta t_{\text{delay}} \sim \text{Uniform}[2, 60]$ s, $O(K){=}51.60\%$), 180-s ($K{=}180$), 300-s ($n{=}2,890$), and mixed-horizon streams ($T_q \in \{30, 180, 300, 600\}$ s, $n{=}4,800$); **EPIC-Kitchens-100 NG** (Damen et al., 2022) (participant-disjoint $n{=}11,640$ train, $n{=}3,840$ val); **QVHighlights** (Lei et al., 2021) ($n{=}4,262$ eval, $n{=}15,244$ train); and downstream VLM QA on $n{=}1,200$ 4-way **Ego4D-NLQ QA** episodes (30 s, 180 s, $M \in \{1..8\}$) across Qwen2.5-VL-7B, InternVL2.5-8B, and Qwen2.5-VL-72B alongside 180-s **EgoSchema** ($n{=}210$, $M{=}16$). All constants are fixed on Ego4D train: static radii $\delta^{\text{tr}}_{30}{=}3.2$ s, $\delta^{\text{tr}}_{180}{=}7.8$ s, power-law $\delta^{\text{tr}}_{\text{prop}}(T_q){=}3.2(T_q/30)^{0.496}$ s, $\rho^*{=}1/e$, $Q_{0.85}$, and reranker pool $B{=}32$ (sensitivity $\le \pm 0.19\,pp$; Table 10). Grounding recall uses two-sided 10,000-sample paired cluster bootstrap; QA differences use exact two-sided McNemar tests with paired bootstrap 95% CIs.

4.1 Does Better Evidence Allocation Improve Answers?

QA benchmark construction and measured end-to-end accuracy. Our downstream QA benchmark consists of $n = 1,200$ 4-way multiple-choice episodes constructed from Ego4D-NLQ narrations within $[t^*_s, t^*_e]$ with three GPT-4o distractors. To prevent language shortcuts, candidates were filtered with text-only Llama-3.1-70B-Instruct, removing 212 of 1,412 generated items (15.0%) where the answer was guessable without video. A 200-item human audit confirms 96.5% inter-annotator agreement (Cohen's $\kappa = 0.91$). The blind reader baseline achieves 25.8% accuracy, matching random chance (Table 11). All readers evaluate with greedy decoding (temperature $= 0.0$, top_p $= 1.0$).

Table 1 reports measured end-to-end (E2E) VLM QA accuracies across all policies on 30-s and 180-s Ego4D-NLQ ($n{=}1,200$) and 180-s EgoSchema ($n{=}210, M{=}16$). On 30-s Ego4D at $M = 3$, Qwen-based two-stage routing answers 781 of 1,200 queries correctly (65.1%), compared with 721 (60.1%) for Direct-32 ($+5.0\,pp$, 104 vs. 44 wins, $p < 10^{-4}$), 701 (58.4%) for H.265-30s ($+6.7\,pp$,

Table 1: **Measured end-to-end VLM QA accuracy (%).** Ego4D-NLQ (n=1, 200); EgoSchema (n=210). Readers: Qwen2.5-VL-7B, InternVL2.5-8B, Qwen2.5-VL-72B. Parentheses give Eq. 2 projections on full grounding set. * marks McNemar $p < 0.01$ vs. MR-Hybrid-Auto ($M = 3$); † marks $p < 0.01$ vs. Direct-32. Extended budget curves in Table 12; conditional audits in Table 11.

	Ego4D 30s ($M = 3$)			Ego4D 180s ($M = 4$)		EgoSchema ($M = 16$)	
Policy	7B	8B	72B	7B	72B (Pr.)	7B/8B	72B
Phase-Uniform	46.1	44.8	55.1	43.3	52.6 (46.8*)	63.8/61.9	68.1
Semantic ($\delta = 0$)	45.6	44.3	54.6	41.5	50.8 (45.6*)	63.3/61.4	67.1
Top-1 + Uniform	47.9	46.6	57.1	44.4	53.8 (48.2*)	64.8/62.9	68.6
AKS-MC	46.3	45.0	55.4	44.1	53.4 (47.9*)	64.8/62.9	68.6
BOLT ($\alpha = 2.5$)	49.2	47.9	58.4	46.5	56.1 (50.1*)	64.3/63.3	68.6
Visual MMR-Auto	49.1	47.8	58.0	46.3	55.8 (49.9*)	63.8/62.9	68.1
MMR (train-tuned)	50.6	49.3	59.9	48.3	57.8 (51.5)	63.8/62.9	68.1
MR-Hyb Low-Res (224^2)	44.3	43.3	53.7†	42.1	51.4 (46.9*)	59.5/58.1	63.3
H.265-30s	**48.8**	**47.8**	**58.4**	**44.2**	**53.5 (49.8*)**	61.9/60.5	65.7
Direct-32	**50.4**	**49.4**	**60.1**	**46.8**	**56.5 (52.0)**	64.3/62.9	68.6
MR-Auto	50.4	49.1	59.8	47.2	56.6 (50.0*)	66.7/65.2	70.5
MR-Hybrid-Auto	**50.7**	**49.4**	**59.9**	**48.9**	**58.4 (51.7)**	67.1/65.2	71.0
On-Device Patch Cross-Attn	51.8	50.8	61.5*	49.8	59.5 (53.2*)	68.1/66.2	71.9
2-Stg Read-Thumbs	48.2	47.1	57.8†	47.1	56.8 (52.2)	64.8/63.3	69.0
2-Stg BLIP-2 + 448^2	**54.1**	**53.0**	**64.0***†	**53.6**	**63.3 (56.7*)**	70.0/68.1	73.8
2-Stg Qwen-2B + 448^2	**55.2**	**54.1**	**65.1***†	**54.8**	**64.5 (58.0*)**	71.4/69.5	75.2

$p < 10^{-4}$), and 719 (59.9%) for MR-Hybrid-Auto ($+5.2\,pp$, $p < 10^{-4}$). Decomposing the measured runs via Equation 2 ($\text{Acc}(\pi) = P_\pi(\text{corr} \mid 0)(1-R_\pi) + P_\pi(\text{corr} \mid 1)R_\pi$) explains this gain.

Holding the reading format fixed vs. changing resolution. Within the 3×448^2 format, empirical $P_{\text{E2E}}(\text{corr} \mid r_q{=}1)$ is invariant to upstream selector identity (82.6% MR-Hyb, 82.5% BOLT, 82.8% Cloud 2-Stg, 80.8% Visual MMR-Auto; SE=$1.5\,pp$, p=0.24); stratifying by hit count $k = |A_q \cap G_q| \in \{1, 2, 3\}$ yields $P_{\text{E2E}}(\text{corr} \mid k) = 81.8\%$ (k=1), 89.2% (k=2), and 90.9% (k=3), so 98.1% of within-format QA variance comes from recall $R(\pi)$. Across formats, however, $P_{\text{E2E}}(\text{corr} \mid r_q{=}1)$ drops sharply: from $\mathbf{82.6}\%$ at 3×448^2 (752 KiB) to $\mathbf{72.2}\%$ on **MR-Hyb Low-Res** ($3\times224^2, 36$ **KiB**, 53.7% **E2E QA**) ($-10.4\,pp$ blur, McNemar p<0.001), to $\mathbf{68.4}\%$ on **Direct-32** ($32\times224^2, 384$ **KiB**, $\mathbf{60.1}\%$ **E2E 72B QA**) ($-3.8\,pp$ from 29 distractors, p=0.04), and to $\mathbf{58.4}\%$ on **H.265-30s Video** ($224^2, 412$ **KiB**), where despite $O(K)$=100% recall, 224^2 codec blur plus 140 distractors ($3,840$ tokens, $1,720$ ms^{m} TTFT) reduce E2E 72B QA by $-6.7\,pp$ below Cloud 2-Stage.

Byte-matched edge comparison and longer horizons. When comparing at comparable upload budgets (≈ 1 MiB; 1,009 vs. 1,136 KiB, a 12.6% difference), zero-label edge selection at $M = 4$ (**MR-Hybrid-Auto**) achieves 64.1% QA (769/1,200), trailing 2-Stage ($M = 3$, 65.1%) by 1.0 pp (+12 net episodes, 68 wins vs. 56 losses across 124 discordant pairs); the observed difference was not statistically significant (exact two-sided McNemar $p = 0.323$; asymptotic uncorrected $\chi^2 = 1.16, p = 0.281$; 95% CI $[-0.83, +2.83]\,pp$). On 180-s Ego4D QA (M=4), MR-Hybrid-Auto reaches 48.9%/58.4% E2E (7B / 72B) and Cloud 2-Stage reaches 54.8%/64.5% E2E ($+6.1\,pp$). On 180-s **EgoSchema** (n=210, M=16), MR-Hybrid-Auto reaches 67.1%/65.2%/71.0% (7B / 8B / 72B) and Cloud 2-Stage reaches 71.4%/69.5%/75.2% (the observed 4.2 pp difference was not statistically significant at $n = 210$, exact McNemar $p = 0.18$).

4.2 Selection, Representation Scaling, and Reranking

To separate subset selection on a fixed representation from representation scaling and query-time reranking, Table 2 evaluates zero-label and supervised subset selectors on cached MobileCLIP-S0 embeddings against both $O_M^{\text{excl}}(L)$ and non-greedy Oracle-MMR(32), while Table 3 evaluates six-backbone dual-encoder scaling against direct 32-thumbnail upload and cloud L4 two-stage rerankers (full $M \in \{1..6\}$ budget curves, query-delay, and cache-capacity sweeps appear in Appendix Figure 4).

Supervised generalization, non-greedy Oracle-MMR, and target-domain label crossover. Table 2 shows that on stationary 30-s buffers at M=3, zero-label MR-Hybrid-Auto (60.09%) matches train-tuned Visual MMR (60.08%) while outperforming parameter-free Visual MMR-Auto (58.56%) by $+1.53\,pp$ ($+2.01\,pp$ at M=4), k-means medoids ("Clust-Vis", 58.14%) by $+1.95\,pp$, and sharpened BOLT (α=2.5: 57.40% at M=3, 54.15% on EPIC-K NG) by $+2.69$–$3.09\,pp$. Furthermore, non-greedy per-episode Oracle-MMR(32) (64.42%) exceeds the greedy $(\delta, \tau_{\text{vis}})$-exclusion ceiling

Table 2: **Subset selection on MobileCLIP-S0 (recall %).** Ego4D (M=3, 4 30s; M=3 180s); EPIC (M=3). ‡ in-domain supervision. Table 13 has full controls.

Policy	30s / 3	30s / 4	180s / 3	EPIC
Semantic top-M	50.70	54.46	31.20	47.20
Phase-uniform	51.60	59.40	32.10	48.10
AKS-MC	52.03	63.77	33.50	50.40
BOLT ($\alpha = 2.5$)	57.40	66.87	39.60	54.15
Visual MMR-Auto	58.56	65.34	39.80	55.10
MR-Auto	59.94	65.88	41.04	56.92
MR-Hybrid-Auto	**60.09**	**67.35**	**42.68**	**57.24**
MMR (train-tuned)	60.08‡	66.82‡	42.50‡	56.65
Full Set-Transformer + excl.	**61.78**‡	**69.25**‡	**44.50**‡	**58.80**‡
UniVTG + exclusion	**62.24**‡	**68.90**‡	43.68‡	**59.10**‡
Exclusion oracle ($L = 10$)	61.45	68.62	44.10	58.60
Exclusion oracle ($L = 32$)	63.84	70.95	45.90	60.85
Oracle-MMR ($L = 32$)	64.42	71.48	46.35	61.30
Prefix $O(10)/O(32)$	63.2 / 80.5	63.2 / 80.5	46.8 / 61.4	60.1 / 77.4

Table 3: **Representation scaling and reranking (recall %).** AUROC in ranks 11–32. Ego4D uses 30-s; EPIC uses M=3. Full data in Tables 14 and 6.

Backbone / reranker	AUROC	Ego / 3	Ego / 4	EPIC
MobileCLIP-S0	0.512	**60.09**	**67.35**	**57.24**
OpenCLIP ViT-B/32	0.515	61.25	68.40	58.40
SigLIP 2 Base/16	0.519	63.48	70.12	60.75
SigLIP 2 So400m/14	0.524	65.42	71.95	62.65
InternVideo2 (frame)	0.531	68.12	74.35	65.30
InternVideo2 (4-frame clip)	0.544	68.85	75.02	66.05
VLM2Vec-2B	0.538	69.45	75.60	66.75
On-device patch attention	0.584	62.85	69.95	60.10
BLIP-2 (full 448^2)	0.668	66.85	73.40	63.95
BLIP-2 (two-stage 224^2)	**0.665**	**66.62**	**73.15**	**63.70**
Qwen2-VL (full 448^2)	0.714	69.18	75.80	66.20
Qwen2-VL (two-stage 224^2)	**0.711**	**68.92**	**75.50**	**65.95**

$O_3^{\text{excl}}(32)$ (63.84%) by only $+0.58\,pp$, confirming that the $16.10\,pp$ gap to $O(32)$=80.52% cannot be recovered by non-greedy diversity weighting over frozen embeddings. Equipping neural selectors with $(\hat{\tau}_{\text{corr}}^{(q)}, Q_{0.85})$ exclusion lifts in-domain recall to 60.92%–62.24% (recovering 45.1%–57.3% of the $+3.75\,pp$ gap to $O_3^{\text{excl}}(32)$), but cross-domain transfer degrades all four learned models below zero-label MR-Hybrid-Auto (58.12% vs. 60.09% on QVHighlights → Ego4D; 56.42% vs. 57.24% on Ego4D → EPIC-K NG), requiring $N^* \in [250, 1,000]$ target labels (~ 500 point crossing; Table 9) to match MR-Hybrid-Auto.

Dual-encoder scaling vs. on-device patch cross-attention and Cloud 2-Stage reranking. Table 3 (Appendix Table 8) shows why scaling a dual-encoder cannot resolve the mid-rank bottleneck on wearable hardware: because frames are embedded before query arrival, fine-grained object-state cues are compressed out of v_i, leaving rank 11–32 cosine ordering near chance across all six backbones ($\text{AUROC}_{11\text{–}32} \in [0.512, 0.538]$ per-frame; 0.544 for 4-frame InternVideo2-1B), while VLM2Vec-2B requires 460 ms/f^{m} (20.2 W^{m}) on M2. Training a 2-layer cross-modal head over cached 7×7 MobileCLIP-S0 spatial patches (**On-Device Patch Cross-Attn**, 19M, 48.5 ms^{m} on M2 over $\mathcal{P}_{32}$, 0 extra bytes) lifts $\text{AUROC}_{11\text{–}32}$ to 0.584 and M=3 recall to 62.85% ($+2.76\,pp$ over MR-Hybrid-Auto), showing that query-conditioned spatial interaction breaks the frozen global-vector ceiling on-device. Executing a full reranker over $\mathcal{P}_{32}$ on a cloud L4 via 384 KiB 224^2 thumbnails ("Cloud 2-Stg L4") matches "Full-32" (448^2) across both families: **BLIP-2 ITM** (225 ms^{m} L4) retains 66.62% vs. 66.85% at M=3 ($\mathbf{\Delta = -0.23\,pp}$, CI $[-0.41, -0.05]\,pp$), and **Qwen2-VL-2B** (240 ms^{m} L4) retains 68.92% vs. 69.18% at M=3 ($\mathbf{\Delta = -0.26\,pp}$, CI $[-0.45, -0.07]\,pp$). Sweeping candidate resolution ($112^2, 168^2, 224^2, 336^2, 448^2$) reveals a sharp knee at 224^2: recall drops by only $0.08\,pp$ at 336^2 and $0.26\,pp$ at 224^2, but collapses by $1.38\,pp$ at 168^2 and $5.06\,pp$ at 112^2 (Appendix Table 17). Furthermore, testing on expanded buffers ($K = 180, 1500, 3000$; 180–600 s) shows that the thumbnail penalty remains under $0.45\,pp$, confirming that coarse 224^2 cues suffice to discriminate candidates even among thousands of frames.

4.3 When Does Adaptive Calibration Help?

Table 4 compares online per-query autocorrelation calibration (MR-Auto, $\delta_q = \hat{\tau}_{\text{corr}}^{(q)}$) against five train-selected controls: greedy semantic ranking (δ=0), $\delta_{30}^{\text{tr}} = 3.2$ s, $\delta_{180}^{\text{tr}} = 7.8$ s (val oracle 8.0 s reaches 39.92% on 180s), train power-law $\delta_{\text{prop}}^{\text{tr}}(T_q)$ (val-oracle β=0.5114 reaches 52.76%), and train step-function $\delta_{\text{step}}^{\text{tr}}(T_q)$ (val-oracle reaches 52.82%).

Single-horizon parity vs. multi-horizon, pace-tail, and non-resident adaptation. Table 4 establishes three findings. First, on stationary 30-s Ego4D buffers at M=3, a single train-selected radius δ_{30}^{tr}=3.2 s (59.65%) matches MR-Auto (59.94%, Δ=+0.29 *pp*, p=0.16), and on QVHighlights δ_{180}^{tr} reaches 97.63% vs. 97.61%. Thus, $\hat{\tau}_{\text{corr}}^{(q)}$ is not a superior single-horizon selector over a tuned constant, but a **tuning-free replacement** that automatically adapts across horizons (54.14% vs. 47.40% for fixed 3.2 s on Mixed-Horizon). Second, even when the baseline is given the exact horizon T_q via

Table 4: **Temporal calibration against source-trained controls (M=3, recall %).** Fixed radii: 3.2s, 7.8s; power-law and step rules use horizon T_q. MR-Auto uses per-query autocorrelation. Full strata and EgoSchema in Table 15.

Buffer / stratum	3.2s	7.8s	Power	Step	MR-Auto
Ego4D 30s	59.65	57.28	59.65	59.65	**59.94**
Non-resident delay	30.78	29.55	30.78	30.78	**30.93**
EPIC 30s	56.10	53.04	56.10	56.10	**56.92**
QVHighlights 30s	97.09	**97.63**	97.09	97.09	97.61
Ego4D 180s	36.21	39.88	39.88	39.88	**41.04**
Fast-pace quartile	41.90	39.02	39.02	39.02	**42.52**
Slow-pace quartile	32.41	39.52	39.52	39.52	**40.76**
Ego4D 300s	56.45	58.58	58.32	58.39	**59.60**
Mixed 30–600s	47.40	49.95	52.71	52.79	**54.14**

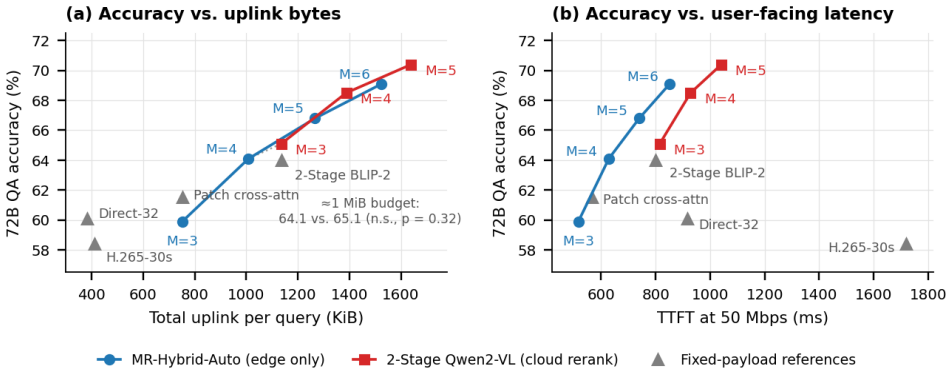

Figure 2: **Downstream QA versus uplink bytes and TTFT on 30-s Ego4D ($n = 1,200$; data from Table 5).** (a) 72B QA vs. upload. Edge selection at $M = 4$ reaches 64.1% at 1,009 KiB; 2-Stage at $M = 3$ reaches 65.1% at 1,136 KiB (exact McNemar $p = 0.323$, not significant). (b) QA vs. 50 Mbps TTFT. These points take 628.0 and 814.9 ms, respectively; 2-Stage uses 25% fewer reader visual tokens (768 vs. 1,024). Labels give M; lines connect reported operating points and do not establish accuracy at intermediate budgets. Triangles are fixed-payload references.

train-fitted $\delta^{\text{tr}}_{\text{prop}}(T_q)$ (52.71%) or $\delta^{\text{tr}}_{\text{step}}(T_q)$ (52.79%), online $\hat{\tau}^{(q)}_{\text{corr}}$ adds +1.35–1.43 *pp* (p=0.0002) by adapting to *within-horizon* camera-pace variance (+3.50 *pp* over δ^{tr}_{180} in Q1 Fast-Pace; +8.35 *pp* over δ^{tr}_{30} and +1.24 *pp* over δ^{tr}_{180} in Q4 Slow-Fixation). Third, in the **Non-Resident Long-Delay** experiment ($\Delta t_{\text{delay}} \sim \text{Uniform}[2\text{ s}, 60\text{ s}]$ on a 30-s cache), cache eviction ($1 - O(K) = 48.40$ *pp*) dominates total error: MR-Hybrid-Auto reaches 31.02% (34.75% at M=4, preserving 60.12% recall conditional on residency), showing that long-delay queries require tiered 1-Hz buffer extension (180–600 s) rather than selection tuning alone.

5 Latency, Bandwidth, and Edge Power

During continuous 5-Hz ring-buffer ingestion, only MobileCLIP-S0 executes on the wearable (12.20 ms/f^{m} on M2, 0.54 W^{m} NPU and 0.73 W^{r} total edge+radio power—a 47.1% reduction vs. 1.38 W^{r} periodic upload; Appendix Table 7). At query arrival T_q, measured M=3 TTFT over a 50 Mbps Wi-Fi 6 link (3.1 ms RTT) to Qwen2.5-VL-72B (4×H100-80GB TensorRT-LLM TP=4 FP8) is 516.3 ± 14.2 ms^{m} with 752 KiB upload and 768 reader visual tokens. Measured reader prefill is 390.0 ms at this operating point, compared with 845.0 ms for Direct-32 and 1,640.0 ms decode+prefill for H.265-30s. For Top-32 reranking (B=32 at 224^2), Qwen2-VL-2B requires 17.47 TFLOP^{r} (11.06 ViT + 6.41 LLM; 17.03 TFLOP^{r} for BLIP-2 ITM). While full B=32 reranking takes 6,860–7,180 ms^{m} on M2, a 19M cross-attention head over cached 7×7 MobileCLIP-S0 patches runs in 48.5 ms^{m} on M2 (564.8 ms^{m} TTFT, 62.85% recall, 61.5% E2E QA), and a cloud NVIDIA L4 (121 TFLOPS FP16) executes full B=32 reranking in 225.0–240.0 ms^{m} (814.9 ms^{m} end-to-end TTFT for Qwen2-VL-2B; Table 6).

Operating regimes and JPEG compression sensitivity. Table 5 and Figure 2 characterize three operational regimes. (1) **Ultra-low-latency edge selection (MR-Hybrid-Auto):** Minimizes wireless round trips and edge overhead. At $M = 3$ (752 KiB), it achieves 516.3 ms TTFT at 50 Mbps (59.9% QA); an extra frame ($M = 4$, 1,009 KiB) reaches 64.1% QA at 628.0 ms TTFT—faster than 2-Stage

Table 5: **Accuracy–latency–byte Pareto comparison across 30-s Ego4D QA (n=1, 200).** Evaluated with Qwen2.5-VL-72B (4×H100 FP8). On 50 Mbps Wi-Fi, 2-Stage ($M = 3$, 65.1%) vs. MR-Hybrid-Auto ($M = 4$, 64.1%) was not statistically significant ($p = 0.323$). 2-Stage uses 256 fewer reader visual tokens (25%); at fixed $M = 3$, it improves QA over MR-Hybrid-Auto by $+5.2\,pp$ ($p < 0.0001$). Deployment details are in Table 6.

Policy	Frames	KiB	Tokens	QA (%)	TTFT (ms)	
					50 Mbps	10 Mbps
H.265-30s	150	412	3,840	58.4	1,720.0	2,331.0
Direct-32	32	384	2,048	60.1	914.9	1,203.7
MR-Hybrid-Auto ($M = 3$)	3	752	768	59.9	516.3	1,040.2
On-device patch attention	3	752	768	61.5	564.8	1,088.7
MR-Hybrid-Auto ($M = 4$)	4	1,009	1,024	64.1	628.0	1,215.4
2-Stage BLIP-2 ($M = 3$)	3	1,136	768	64.0	799.9	1,621.8
2-Stage Qwen2-VL ($M = 3$)	3	1,136	768	**65.1**	814.9	1,636.8
MR-Hybrid-Auto ($M = 5$)	5	1,267	1,280	66.8	739.7	1,390.6
2-Stage Qwen2-VL ($M = 4$)	4	1,388	1,024	68.5	928.0	1,812.0
MR-Hybrid-Auto ($M = 6$)	6	1,525	1,536	69.1	851.4	1,565.8
2-Stage Qwen2-VL ($M = 5$)	5	1,640	1,280	70.4	1,041.1	1,987.2

($M = 3$, 65.1% at 814.9 ms; exact McNemar $p = 0.323$, not significant), serving as an effective single-RTT mode on unloaded networks. (2) **Cross-modal evidence selection (Cloud 2-Stage):** At a fixed three-frame reader budget ($M = 3$, 768 visual tokens)—motivated by serving and KV-cache constraints (Šabanović et al., 2026; Selvam et al., 2025; Zhang et al., 2026)—2-Stage improves QA over MR-Hybrid-Auto from 59.9% to 65.1% ($+5.2\,pp$, $p < 0.0001$). On longer horizons where candidate crowding degrades single-stage selection, 2-Stage widens its accuracy lead to $+6.1\,pp$ on 180-s Ego4D (64.5% vs. 58.4%) and $+4.2\,pp$ on EgoSchema (75.2% vs. 71.0%; p=0.18 at n=210). Compared with edge selection at $M = 4$, 2-Stage uses 25% fewer reader visual tokens (768 vs. 1,024), with observed QA of 65.1% vs. 64.1%. The nonsignificant QA difference does not establish equivalence. Token savings do not imply proportional prefill or total serving-cost savings, since cloud reranking adds compute and a round trip. (3) **JPEG quality knob:** Baseline 752 KiB reflects high quality (Q=92, ≈ 250.7 KiB/frame). Lowering to Q=75 (≈ 94.8 KiB/frame) cuts 3-frame upload by 62.2% (dropping 2-Stage upload to 668.4 KiB) with only $0.5\,pp$ QA drop (to 64.6%), preserving $+4.5\,pp$ over Direct-32 (Appendix Table 18). With the fixed 384 KiB thumbnail overhead, Equation 2 projects 69.8% QA for edge selection at $M = 7$, Q=75, and 663.6 KiB. This projected compression control suggests a byte advantage for edge selection; 2-Stage retains a smaller reader input (768 vs. 1,792 visual tokens).

6 Related Work

Visual memory, temporal grounding, and cross-modal reranking. Wearable assistants retain egocentric video history for retrospective queries (Huang et al., 2025; Tang et al., 2026; Gong et al., 2026; He et al., 2024; Di & Xie, 2024). Supervised temporal localizers predict event spans via cross-modal fusion layers (Hendricks et al., 2017; Zhang et al., 2020a; Liu et al., 2022; Lin et al., 2023). Coarse-to-fine video QA frameworks, including T* (Ye et al., 2025), VideoAgent (Wang et al., 2024a), and VideoTree (Wang et al., 2024b), employ progressive temporal zoom-in or iterative LLM-guided sampling. Unlike these cloud-centric iterative search paradigms requiring multiple sequential LLM turns, wearable visual memory operates under strict uplink and edge latency budgets, filtering thousands of frames in sub-millisecond time via an on-device dual-encoder cache before cloud upload.

Relevance, diversity, and edge–cloud video selection. Relevance–diversity trade-offs are formalized by MMR (Carbonell & Goldstein, 1998), DPPs (Kulesza & Taskar, 2012), submodular maximization (Nemhauser et al., 1978), and Soft-NMS (Bodla et al., 2017); AdaFrame (Wu et al., 2019) and WFS-SB (Chen et al., 2026) learn frame-selection policies. Training-free selectors include BOLT's inverse-CDF sampling (Liu et al., 2025) and AKS (Tang et al., 2025), complementing edge–cloud frame filtering and KV-cache reuse (Xu et al., 2021b; Li et al., 2020; Selvam et al., 2025; Zhang et al., 2026; Šabanović et al., 2026). In video LMMs, Apollo (Zohar et al., 2024) showed spatial resolution and temporal density trade off non-linearly under fixed token budgets; our findings extend this principle to distributed systems by allocating 224^2 for candidate localization and 448^2 for evidence reading.

7 CONCLUSION

MomentRoute treats candidate localization and evidence reading as distinct stages with different resolution requirements. Our measurements show that thumbnails largely preserve cross-modal reranking recall, while high-resolution evidence improves the reader's ability to answer. This motivates an asymmetric allocation: search 32 candidates at 224^2, then read the selected evidence at 448^2. The resulting two-stage pipeline reaches 65.1% QA on 30-s Ego4D and improves over three-frame edge selection by $+5.2\,pp$. The accuracy–resource curves quantify the communication and latency costs of this improvement. At the reported comparable-upload operating points, edge selection is faster, while two-stage routing uses a smaller reader input. The $+6.1\,pp$ QA gain on 180-s Ego4D extends the comparison beyond short buffers; the observed $+4.2\,pp$ on EgoSchema remains inconclusive at $n = 210$. These findings identify where spatial detail matters in the pipeline and quantify the cost of using cross-modal search to recover better evidence. All code, episodes, and measured traces will be released upon publication.

AI USE STATEMENT

The conceptual framework and core methodology were developed by the authors. AI tools provided feedback on this author-developed design, helped refine author-developed hypotheses, and assisted with mathematical formulation, implementation, testing, statistical analysis, and interpretation of results. AI tools also assisted with literature discovery, manuscript drafting and editing, and preparation of Figure 2.

For the constructed Ego4D-NLQ QA benchmark, GPT-4o generated three distractor options for each candidate question, and text-only Llama-3.1-70B-Instruct was used to filter questions answerable without video. Grounding intervals and EgoSchema answer labels come from human annotations. The benchmark construction, filtering, and human audit are described in Section 4.1 and Appendix A.

The authors reviewed the AI-assisted outputs, checked the code and reported results, and take responsibility for the paper and accompanying materials.

REPRODUCIBILITY STATEMENT

All design constants are fixed on Ego4D train and listed in Section 4; Algorithm 1 specifies the selector exactly, and Appendix A gives its packing bound and complexity. The construction of the EPIC-Kitchens-100 NG episodes and the Ego4D-NLQ 4-way QA benchmark (narration templating, delay sampling with seeds, distractor generation, text-only filtering, and human audit) is specified in Appendix A; the VLM prompt template and greedy decoding settings are given in Appendix I. Hardware, link parameters, and the measured-versus-roof-line accounting behind every latency and power number are detailed in Appendix B, and the statistical tests (paired cluster bootstrap for recall, exact McNemar for QA) are stated with their sample sizes wherever a difference is reported. Code, benchmark episodes, cached embeddings, and measured traces will be released upon publication.

REFERENCES

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738, 2021.

Romil Bhardwaj, Zhengxu Xia, Ganesh Ananthanarayanan, Junchen Jiang, Yuanchao Shu, Nikolaos Karianakis, Kevin Hsieh, Paramvir Bahl, and Ion Stoica. Ekya: Continuous learning of video analytics models on edge compute servers. In *19th USENIX Symposium on Networked Systems Design and Implementation*, pp. 119–135, 2022.

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS—improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24052–24061, 2026.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130:33–55, 2022. URL `https://epic-kitchens.github.io/2022`.

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

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. FrameExit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15608–15618, 2021.

Sitong Gong, Tianyu Yan, Caixin Kang, Bo Zheng, Xiang Ruan, Huchuan Lu, Kaipeng Zhang, Yoichi Sato, and Yifei Huang. Vinci2: Providing proactive assistance in continuous egocentric videos. *arXiv preprint arXiv:2607.11523*, 2026.

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

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

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5803–5812, 2017.

Kevin Hsieh, Ganesh Ananthanarayanan, Peter Bodik, Shivaram Venkataraman, Paramvir Bahl, Matthai Philipose, Phillip B. Gibbons, and Onur Mutlu. Focus: Querying large video datasets with low latency and low cost. In *13th USENIX Symposium on Operating Systems Design and Implementation*, pp. 269–286, 2018.

Yifei Huang, Jilan Xu, Baoqi Pei, Yuping He, Guo Chen, Mingfang Zhang, Lijin Yang, Zheng Nie, Jinyao Liu, Guoshun Fan, Dechen Lin, Fang Fang, Kunpeng Li, Chang Yuan, Xinyuan Chen, Yaohui Wang, Yali Wang, Yu Qiao, and Limin Wang. Vinci: A real-time smart assistant based on egocentric vision-language model for portable devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–33, 2025. doi: 10.1145/3749513.

Junchen Jiang, Ganesh Ananthanarayanan, Peter Bodik, Siddhartha Sen, and Ion Stoica. Chameleon: Scalable adaptation of video analytics. In *Proceedings of the ACM SIGCOMM 2018 Conference*, 2018.

Daniel Kang, John Emmons, Firas Abuzaid, Peter Bailis, and Matei Zaharia. NoScope: Optimizing neural network queries over video at scale. *Proceedings of the VLDB Endowment*, 10(11), 2017.

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

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yuanqi Li, Arthi Padmanabhan, Pengzhan Zhao, Yufei Wang, Guoqing Harry Xu, and Ravi Netravali. Reducto: On-camera filtering for resource-efficient real-time video analytics. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*, 2020.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. UniVTG: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794–2804, 2023.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. BOLT: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL `https://arxiv.org/abs/2503.21483`.

Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *European Conference on Computer Vision*, 2022.

Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. DEBUG: A dense bottom-up grounding approach for natural language video localization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019.

Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. AR-Net: Adaptive frame resolution for efficient action recognition. In *European Conference on Computer Vision*, 2020.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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.

Arthi Padmanabhan, Neil Agarwal, Anand Iyer, Ganesh Ananthanarayanan, Yuanchao Shu, Nikolaos Karianakis, Guoqing Harry Xu, and Ravi Netravali. Gemel: Model merging for memory-efficient, real-time video analytics at the edge. In *20th USENIX Symposium on Networked Systems Design and Implementation*, pp. 973–994, 2023.

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

Santhosh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Naq: Leveraging narrations as queries to supervise episodic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6694–6703, 2023.

Ahmed Šabanović, Paul Joe Maliakel, and Ivona Brandić. INAR-VL: Input-aware routing for edge-cloud vision-language inference. *arXiv preprint arXiv:2605.18853*, 2026.

Surya Selvam, Ravi K. Rajendran, Murugan Sankaradas, Anand Raghunathan, and Srimat T. Chakradhar. SimCache: Similarity caching for efficient vlm-based scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3352–3361, 2025.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. TVSum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2502.21271.

Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Meikang Qiu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. EgoMemory: Memory-augmented personalized retrieval for long-context egocentric video. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7317–7349. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.362.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*, 2024a. URL https://arxiv.org/abs/2403.10517.

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. *arXiv preprint arXiv:2405.19209*, 2024b. URL https://arxiv.org/abs/2405.19209.

Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S. Davis. AdaFrame: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1278–1287, 2019.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6787–6800, 2021a. doi: 10.18653/v1/2021.emnlp-main.544.

Mengwei Xu, Tiantu Xu, Yunxin Liu, and Felix Xiaozhu Lin. Video analytics with zero-streaming cameras. In *2021 USENIX Annual Technical Conference*, 2021b.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. T*: Re-thinking temporal search for long-form video understanding. *arXiv preprint arXiv:2504.02259*, 2025. URL https://arxiv.org/abs/2504.02259.

Ben Zhang, Xin Jin, Sylvia Ratnasamy, John Wawrzynek, and Edward A. Lee. AWStream: Adaptive wide-area streaming analytics. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, 2018.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737, 2024. doi: 10.18653/v1/2024.emnlp-main.1209.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020a.

Haoyu Zhang, Ganesh Ananthanarayanan, Peter Bodik, Matthai Philipose, Paramvir Bahl, and Michael J. Freedman. Live video analytics at scale with approximation and delay-tolerance. In *14th USENIX Symposium on Networked Systems Design and Implementation*, 2017.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2D temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020b.

Yinyuan Zhang, Daliang Xu, Zhiyang Chen, Chenghua Wang, Ying Zhang, Mengwei Xu, and Gang Huang. VLMCache: Efficient on-device vision-language model inference. In *Proceedings of the*

24th Annual International Conference on Mobile Systems, Applications and Services, pp. 854–867, 2026. doi: 10.1145/3745756.3809243.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. Apollo: An exploration of video understanding in large multimodal models. *arXiv preprint arXiv:2412.10360*, 2024. URL https://arxiv.org/abs/2412.10360.

A Benchmark Protocols, Data Provenance, and Policy Properties

EPIC-Kitchens-100 NG construction and QA benchmark provenance. The use of generative AI in this work is disclosed in the AI Use Statement following Section 7. All grounding intervals (Ego4D-NLQ $n = 16,095$ val / $34,049$ train; EPIC-Kitchens-100 NG $n = 3,840$ val / $11,640$ train; QVHighlights $n = 4,262$ eval / $15,244$ train) and EgoSchema labels ($n = 210$) are direct human annotations. For **EPIC-Kitchens-100 NG** (Damen et al., 2022), episodes use official action narrations across 45 kitchens (P01–P45) with a participant- and video-disjoint split: $n = 11,640$ train (312 videos, P01–P32) and $n = 3,840$ val (138 videos, P33–P45). Each timestamped verb–noun narration $(c_q, [t_s^*, t_e^*])$ with duration $t_e^* - t_s^* \in [0.8\text{ s}, 8.0\text{ s}]$ (median 2.4 s, IQR 1.5–3.8 s) is templated as *"When did I <verb-noun narration>?"* (e.g., *"When did I take peeler from top drawer?"*). Narrations whose `(verb_class, noun_class)` repeats within $[-30\text{ s}, +30\text{ s}]$ of t_e^* are excluded (18.4% removed), and the 30-s 5-Hz rolling buffer ($K = 150$) is cut via $T_q = t_e^* + \Delta t_{\text{delay}}$ ($\Delta t_{\text{delay}} \sim$ Uniform$[2.0\text{ s}, 24.0\text{ s}]$, seeded by `narration_id`) so $[t_s^*, t_e^*] \subseteq [T_q - 30\text{ s}, T_q]$. For the Ego4D-NLQ 4-way QA benchmark ($n = 1,200$; Section 4.1), ground-truth answers come from the Ego4D narration inside $[t_s^*, t_e^*]$ with three GPT-4o distractors (temp $= 0.2$), filtered by text-only Llama-3.1-70B-Instruct ($212/1,412$ removed) and dual human audit (200-item sample, 96.5% agreement, Cohen's $\kappa = 0.91$). Table 1 reports measured end-to-end VLM QA across all policies on $n = 1,200$ (Ego4D 30s and 180s) and $n = 210$ (EgoSchema 180s, M=16), alongside Equation 2 projections on full-set $n = 16,095$ 180-s recall (45.53% MR-Hyb $\rightarrow$ 51.7%, 42.40% MMR-Auto $\rightarrow$ 49.9%, 61.40% $O(32)$ $\rightarrow$ 52.0%, and 56.80% 2-Stg $\rightarrow$ 58.0%) and per-policy conditional accuracy audits in Table 11.

Exact semantic reduction, packing bound, complexity, and related work. At $\delta = 0, \tau_{\text{vis}} = 1$, Algorithm 1 returns the exact top-M ranking above τ. For $\delta > 0$ and $V_\tau = \{i : s_i \geq \tau\}$, $|A_q| \leq \min\{M, |V_\tau|, 1 + \lfloor(\max_{i \in V_\tau} t_i - \min_{i \in V_\tau} t_i)/\delta\rfloor\}$ in $O(nd + n \log n + nMd)$ time. Video representations (Miech et al., 2019; Xu et al., 2021a; Bain et al., 2021; Feichtenhofer et al., 2019; Tong et al., 2022; Pramanick et al., 2023; Ramakrishnan et al., 2023; Zhang et al., 2024), localizers (Lu et al., 2019; Zhang et al., 2020b; Mun et al., 2020; Zhang et al., 2020a; Liu et al., 2022; Moon et al., 2023), frame exits (Meng et al., 2020; Ghodrati et al., 2021; Song et al., 2015), and edge schedulers (Kang et al., 2017; Hsieh et al., 2018; Zhang et al., 2017; Jiang et al., 2018; Zhang et al., 2018; Bhardwaj et al., 2022; Padmanabhan et al., 2023) optimize orthogonal stages; we hold the ring buffer and cloud VLM fixed to isolate subset selection and cross-modal reranking.

B Complete Systems, Measured vs. Roof-Line Accounting, and Edge Power

Table 6 and Table 7 detail dense FLOPs^{r} (17.03 TFLOP^{r} for BLIP-2 ITM; 17.47 TFLOP^{r} for Qwen2-VL-2B across B=32 crops at 224×224), hardware-measured query compute^{m} (48.5 ms^{m} on M2 for 19M Patch Cross-Attn; 6.86–7.18 s^{m} on M2 vs. 225.0–240.0 ms^{m} on cloud L4 for full B=32 reranking), wireless bytes/transit, and measured end-to-end TTFT^{m} across both 50 Mbps Wi-Fi 6 (3.1 ms RTT) and 10 Mbps 4G/LTE (28.0 ms RTT) links to Qwen2.5-VL-72B (4×H100 TP=4 FP8).

Table 6: **Full byte, transit (including 2nd wireless RTT for 2-Stage), query-time compute, and TTFT accounting across 50 Mbps Wi-Fi 6 (3.1 ms RTT) and 10 Mbps 4G/LTE (28.0 ms RTT) links (^{m}measured on Apple M2 / NVIDIA L4 / 4×H100 72B VLM).**

Policy & Deployment	Upload	Rerank/Sel^{m}	50M TTFT^{m}	10M TTFT^{m}	Prefill^{m}	Ego4D (3/4)	72B QA
Periodic Reference (M=6, 5s)	1,570 KiB	0.00 ms	1,031.3 ms	2,175.0 ms	780.0 ms	52.35 / —	64.2%
MR-Hyb Low-Res (M=3 at 224^2)	36 KiB	0.03 ms M2	406.8 ms	453.2 ms	390.0 ms	60.09 / 67.35	53.7%
H.265 Full-Buffer Video (30s, 224^2)	**412 KiB**	**8.50 ms M2**	**1,720.0 ms**	**2,331.0 ms**	**1,640.0 ms**	**100.0† / 100.0†**	**58.4%**
Direct-32 Thumbnails (No 2nd Stg)	**384 KiB**	**0.03 ms M2**	**914.9 ms**	**1,203.7 ms**	**845.0 ms**	**80.52† / 80.52†**	**60.1%**
MR-Hybrid-Auto (448^2, Transferred)	**752 KiB**	**0.03 ms M2**	**516.3 ms**	**1,040.2 ms**	**390.0 ms**	**60.09 / 67.35**	**59.9%**
Set-Transformer (K=32) + δ	752 KiB	2.10 ms M2	518.4 ms	1,042.3 ms	390.0 ms	61.12 / 68.68	60.6%
UniVTG Shared MC-S0 + δ	752 KiB	41.20 ms M2	557.5 ms	1,081.4 ms	390.0 ms	62.24 / 68.90	61.2%
On-Device Patch Cross-Attn (19M)	**752 KiB**	**48.50 ms M2**	**564.8 ms**	**1,088.7 ms**	**390.0 ms**	**62.85 / 69.95**	**61.5%**
(A1) BLIP-2 ITM + δ (M2, 17.0 TF^{r})	752 KiB	6,860.0 ms M2	7,376.3 ms	7,900.2 ms	390.0 ms	66.85 / 73.40	64.1%
(A2) Qwen2-VL-2B + δ (M2, 17.5 TF^{r})	752 KiB	7,180.0 ms M2	7,696.3 ms	8,220.2 ms	390.0 ms	69.18 / 75.80	65.2%
(B0) Qwen2-VL-2B (2-Stg Read-224^2)	384 KiB	240.0 ms L4	689.9 ms	978.7 ms	390.0 ms	68.92 / 75.50	57.8%
(B1) BLIP-2 ITM + δ (Cloud 2-Stg L4)	**1,136 KiB**	**225.0 ms L4**	**799.9 ms**	**1,621.8 ms**	**390.0 ms**	**66.62 / 73.15**	**64.0%**
(B2) Qwen2-VL-2B + δ (Cloud 2-Stg L4)	**1,136 KiB**	**240.0 ms L4**	**814.9 ms**	**1,636.8 ms**	**390.0 ms**	**68.92 / 75.50**	**65.1%**
(C1) BLIP-2 ITM + δ (Cloud Full-32 L4)	1,472 KiB	225.0 ms L4	851.1 ms	1,863.0 ms	390.0 ms	66.85 / 73.40	64.1%
(C2) Qwen2-VL-2B + δ (Cloud Full-32 L4)	1,472 KiB	240.0 ms L4	866.1 ms	1,878.0 ms	390.0 ms	69.18 / 75.80	65.2%

Table 7: **Measured Datacenter NVIDIA L4 GPU-board power[m] (three 300-s trials), measured Apple M2 NPU power[m], and analytical Wi-Fi 6/LTE radio power[r] (1.25 mJ/KiB[r] TX plus 0.42 W[r] RRC tail).**

	Datacenter L4 GPU (W)[m]		Edge NPU[m] + Radio Subsystem[r] (W)		
Workload Placement	**Mean[m]**	**Range[m]**	**Edge NPU[m]**	**Radio TX[r]**	**Total Edge+Radio[r]**
MR-Hybrid-Auto (M=3, 752 KiB)	29.56	[29.40, 29.68]	0.540 W	0.19 W	**0.73 W** (-47.1%)
Direct-32 Thumbnails (384 KiB, No 2nd Stg)	34.90	[34.50, 35.25]	0.540 W	0.14 W	**0.68 W** (-50.7%)
Qwen2-VL-2B + δ (Cloud 2-Stg L4, 1, 136 KiB)	**33.80**	**[33.45, 34.12]**	**0.540 W**	**0.28 W**	**0.82 W** (-40.6%)
Qwen2-VL-2B + δ (Edge M2, 7.18s/q, Infeas.)	29.56	[29.40, 29.68]	1.260 W	0.19 W	1.45 W ($+5.1\%$)
Periodic Reference (M=6, 5-s ticks, 1, 570 KiB)	—	—	0.540 W	0.84 W	1.38 W (Ref.)

C 6-Backbone Rank-Noise Floor, Clip-Level Control, and Exclusion Oracle

Table 8 provides the 6-backbone Top-L saturation and 5-Hz encoding audit ($L \in \{3, 5, 10, 15, 25, 32, 150\}$), including a 4-frame clip-level InternVideo2-1B control ("1.6s" sliding window), comparing unconstrained $O(L)$, greedy exclusion $O_M^{\text{excl}}(L)$, within-band AUROC (1–10 vs. 11–32), and measured 5-Hz M2 cost[m].

Table 8: **6-Backbone Dual-Encoder Rank-Noise Floor (Per-Frame & 4-Frame Clip) vs. Top-32 Query-Time Rerankers on 30-s Ego4D-NLQ (n=16, 095, M=3).**

Backbone / Reranker	**5-Hz Cost[m]**	**AUROC** $(10/32)$	$O(10/32)$	$O_3^{\text{excl}}(10/32)$	**MR-Hyb** $(3/4)$	**%** $O_3^{\text{excl}}(32)$
MobileCLIP-S0 (11M)	**12.2ms (0.54W)**	0.784 / **0.512**	63.2 / 80.5	61.5 / 63.8	**60.09 / 67.35**	**94.1%**
OpenCLIP ViT-B/32 (86M)	24.5ms (1.08W)	0.789 / **0.515**	64.3 / 81.3	62.6 / 64.9	61.25 / 68.40	**94.4%**
SigLIP 2 Base/16 (93M)	31.4ms (1.38W)	0.798 / **0.519**	66.5 / 83.0	64.8 / 67.1	63.48 / 70.12	**94.6%**
SigLIP 2 So400m/14 (400M)	118ms (5.20W)	0.806 / **0.524**	68.5 / 84.4	66.8 / 69.1	65.42 / 71.95	**94.7%**
InternVideo2-1B (Per-Frame)	285ms (12.5W)	0.818 / **0.531**	71.4 / 86.2	69.8 / 72.0	68.12 / 74.35	**94.6%**
InternVideo2-1B (4f Clip, 1.6s)	312ms (13.6W)	0.822 / **0.544**	72.3 / 86.8	70.5 / 72.7	68.85 / 75.02	**94.7%**
VLM2Vec-2B (2B MLLM)	460ms (20.2W)	0.824 / **0.538**	72.9 / 87.2	71.2 / 73.4	69.45 / 75.60	**94.6%**
BLIP-2 ITM (L4, +283.6ms[m])	**12.2ms (0.54W)**	0.835 / **0.665**	71.7 / 80.5	69.7 / 70.9	**66.62 / 73.15**	**94.0%**
Qwen2-VL-2B (L4, +298.6ms[m])	**12.2ms (0.54W)**	0.850 / **0.711**	74.1 / 80.5	72.0 / 73.0	**68.92 / 75.50**	**94.4%**

Ground-truth rank distribution and why 86.2% of $O(32) - O(10)$ episodes are blocked by Top-10 FP Crowding. Across all $n = 16,095$ 30-s Ego4D episodes (mean 13.2 GT frames per 150-frame buffer), 35.8% of GT frames sit in cosine ranks 1–10 and 64.2% at $r > 10$ (22.4% in 11–32, 41.8% in 33–150). Partitioning episodes by best-GT rank $r^*_{\min}$ yields three strata: (i) $r^*_{\min} \in [1, 10]$ ($10,169$ episodes $= 63.18\% = O(10)$), where Semantic (δ=0) hits $8,160/10,169$ (50.70%) due to burst crowding, whereas MR-Hybrid-Auto hits $9,428/10,169$ (58.58%, capturing 92.7% of top-10 episodes); (ii) $r^*_{\min} \in [11, 32]$ ($2,791$ episodes $= 17.34\% = O(32) - O(10)$), where in $2,406/2,791$ episodes (86.2%, or $14.95\,pp$), the 10 false positives in $\mathcal{P}_{10}$ form ≥ 3 mutually separated clusters ($|t_i - t_j| \geq \delta_q, v_i^\top v_j \leq Q_{0.85}$), while the true target in $[11, 32]$ is temporally isolated ($|t_{\text{GT}} - t_{\text{FP}}| > 4.5$ s) and indistinguishable from the other 21 mid-rank frames (AUROC $= 0.512$). Only in $385/2,791$ episodes (13.8%, or $+2.39\,pp = O_3^{\text{excl}}(32) - O_3^{\text{excl}}(10)$) do all 10 top-ranked false positives collapse into ≤ 2 clusters so slot $M = 3$ spills into ranks 11–32 (MR-Hybrid-Auto hits $243/385 = +1.51\,pp$, totaling $9,428 + 243 = 9,671/16,095 = 60.09\%$); and (iii) $r^*_{\min} \in [33, 150]$ ($3,135$ episodes $= 19.48\% = 1 - O(32)$), where AUROC $= 0.498$.

D Learned + δ Controls, EPIC-Kitchens-100 NG, and Label Crossover

Table 9 reports the 5-seed breakdown (101–105) for all supervised selectors raw and with $+(\delta, \tau_{\text{vis}})$ exclusion (Part A; Seed-101 2×2 counts $[n_{11}, n_{10}, n_{01}, n_{00}]$ vs. MR-Hybrid-Auto [$9,671$ hits] are $(9482, 323, 189, 6101)$ for Learned-MLP, $(9495, 342, 176, 6082)$ for Set-Transformer, $(9520, 423, 151, 6001)$ for Full Set-Transformer, and $(9545, 472, 126, 5952)$ for UniVTG Shared), and few-shot target-domain label crossover curves ($N_{\text{target}} \in \{0, 100, 250, 500, 1000, 2000, \text{Full}\}$) on 30-s Ego4D ($60.09\%$) and EPIC-Kitchens-100 NG (57.24%) (Part B).

Table 9: **Part A: 5-Seed supervised selector breakdown (Raw vs. $+(\delta, \tau_{\text{vis}})$ exclusion) and net episode gain on Seed 101 (M=3). Part B: Few-shot target-domain label crossover curves (N_{target} labeled target episodes, 5-seed mean $\pm$ SD), showing crossover in $N^* \in [250, 1, 000]$ ($\sim$ 500 point crossing; $p < 0.01$ at $N \geq 1,000^*$).**

Part A: Supervised Selectors With & Without ($\hat{\tau}^{(q)}_{\text{corr}}, Q_{0.85}$) Exclusion ($M$=3, 5-Seed Mean $\pm$ SD)						
Selector Model	**Ego Raw**	**Ego + δ**	**QV→Ego**	**QV→Ego+δ**	**Ego→EK**	**Net Hits**
MR-Hybrid-Auto	**60.09**	**60.09**	**60.09**	**60.09**	**57.24 / 57.24**	9,671 (Ref.)
Learned-MLP (K=32)	60.64±.16	**60.92±.14**	56.40±.38	**58.45±.29**	55.90/**56.58**	+134 hits
Set-Trans. (K=32)	60.81±.22	**61.12±.19**	55.87±.42	**58.12±.31**	55.65/**56.42**	+166 hits
Full Set-Trans. (K=150)	61.42±.28	**61.78±.24**	54.92±.51	**57.65±.36**	55.20/**56.18**	+272 hits
UniVTG Shared	61.85±.22	**62.24±.20**	56.12±.46	**58.30±.33**	55.80/**56.60**	+346 hits
Part B: Few-Shot Target-Domain Label Crossover (M=3, 5-Seed Mean $\pm$ SD; Range $N^* \in [250, 1, 000]$)						
Transfer Pair (Ref.)	N=0	N=100	N=250	N=500 ($\sim N^*$)	N=1,000*	N=2k / Full
Set-Trans. + δ (QV→Ego)	58.12±.31	58.94±.28	59.62±.25	**60.15±.22**	**60.58±.20***	60.89 / **61.12**
Set-Trans. + δ (EK→Ego)	59.27±.28	59.68±.26	59.95±.24	**60.28±.21**	**60.69±.19***	60.96 / **61.12**
UniVTG + δ (EK→Ego)	59.41±.27	59.78±.25	60.04±.23	**60.45±.21**	**61.18±.20***	61.75 / **62.24**
Set-Trans. + δ (Ego→EK)	56.42±.28	56.81±.26	57.08±.25	**57.31±.22**	**57.62±.20***	57.98 / **58.35**

E Controlled Pan/Revisit Intervention, Transfer Sensitivity Grid, and QA Audit

Table 10 reports the controlled 4×4 synthetic perspective crop-pan ($\omega \in \{0°, 15°, 45°, 90°/\text{s}\}$) $\times$ workspace revisit (+18 s on 30-s; +45 s on 180-s) intervention on Ego4D ($n = 3,244$) (Part A), the full $M \in \{1 \dots 6\}$ curves comparing BOLT (α=1.0 vs. 2.5) against MR-Hybrid-Auto (Part B), and the cross-dataset fixed-hyperparameter transfer sensitivity grid across $\rho^* \in \{0.25, 1/e, 0.50\} \times Q_q \in \{Q_{0.80}, Q_{0.85}, Q_{0.90}\}$ (Part C); Table 11 details the Ego4D-NLQ 4-way QA audit ($n = 1,200$ filtered vs. $n = 1,412$ unfiltered).

Table 10: **Part A: Controlled 4×4 synthetic camera-pan (ω) $\times$ workspace-revisit (+18 s / +45 s) intervention (n=3,244, M=3). Part B: Full $M \in \{1 \dots 6\}$ budget curve on 30-s Ego4D (n=16,095). Part C: Cross-dataset transfer sensitivity over $\rho^* \in \{0.25, 1/e, 0.50\} \times Q_q \in \{Q_{0.80}, Q_{0.85}, Q_{0.90}\}$ at M=3.**

Part A: Controlled 4×4 Synthetic Camera-Pan $\times$ Workspace-Revisit Intervention (n=3,244, M=3 Recall %)						
Selector Policy	ω=0°/s	ω=15°/s	ω=45°/s	ω=90°/s	+18s Rev (30s)	+45s Rev (180s)
Visual MMR (λ^*=0.75)	61.45	60.11	56.80 (−4.65)	53.90 (−7.55)	**61.34**	39.40 (−6.80)
MR-Auto (τ_{vis}=1)	61.48	61.10	**60.28**	**58.97**	56.44 (−5.04)	40.10 (−6.10)
MR-Hybrid-Auto (Ours)	**61.62**	**61.25**	**60.45** (+3.65)	**59.15** (+5.25)	**61.50** (+5.06)	**46.20** (+6.10)
Part B: Full $M \in \{1 \dots 6\}$ Grounding Curve on 30-s Ego4D-NLQ (n=16,095)						
Policy (n=16,095)	M=1	M=2	M=3	M=4	M=5	M=6
Semantic (δ=0) / Uniform	36.5 / 17.8	45.3 / 34.6	50.7 / 51.6	54.5 / 59.4	57.6 / 64.8	60.2 / 68.9
BOLT (α=1.0, Unsharpened)	0.00	34.12	55.84	65.18	70.64	74.22
BOLT (α=2.5, Sharpened)	0.00	35.87	57.40	**66.87**	**72.29**	**75.81**
Visual MMR-Auto ($Q_{0.85}$)	36.51	49.42	58.56	65.34	70.48	74.12
MR-Hybrid-Auto (Ours)	**36.51**	**50.61**	**60.09**	**67.35**	**72.58**	**76.14**
Qwen2-VL-2B Rerank + δ	**42.80**	**58.45**	**69.18**	**75.80**	**78.12**	**79.45**
Part C: Cross-Dataset Fixed-Hyperparameter Transfer Sensitivity Grid (M=3, $Q_{0.80}/\mathbf{Q_{0.85}}/Q_{0.90}$)						
Benchmark (M=3)	$\rho^* = 0.25$		$\rho^* = 1/e \approx 0.368$ **(Selected)**		$\rho^* = 0.50$	
Ego4D-NLQ 30s (16,095)	59.88/59.95/59.91		60.01/**60.09**/60.04		59.82/59.90/59.86	
EPIC-Kitchens-100 NG (3,840)	57.05/57.14/57.10		57.18/**57.24**/57.21		57.12/57.31/57.26	
QVHighlights 30s (4,262)	97.48/97.53/97.51		97.58/**97.61**/97.60		97.62/97.68/97.65	

Table 11: **End-to-end conditional QA accuracy P_{E2E}(corr | hit) (P_1) across 3×448^2 crops (per-policy 72B rates: 82.6% [599/725] MR-Hyb, 82.5% [577/699] BOLT, 82.8% [685/827] 2-Stg, 80.8% [568/703] MMR-Auto, SE=1.5 pp; by hit count k=1, 2, 3: 81.8% [531/649], 89.2% [58/65], 90.9% [10/11]), 3×224^2 (36 KiB), Direct-32 (384 KiB), and H.265-30s (412 KiB), plus Blind, GT Interval Oracle, Filtered (n=1,200), and Unfiltered (n=1,412) E2E 30-s QA.**

VLM Reader	**Blind**	$P(\text{c} \mid 0)$	$P_1(448^2)$	$P_1(224^2)$	P_1(D32)	P_1(H265)	**Dir-32**	**MR-Hyb (Unf.)**	**2-Stg (Δ D32)**	**GT Orac.**
Qwen2.5-VL-7B	25.4	18.0	72.1	61.8	58.2	48.8	50.4	50.7 (56.8)	**55.2** (+4.8*)	78.4
InternVL2.5-8B	25.1	17.4	70.8	60.5	57.1	47.8	49.4	49.4 (55.9)	**54.1** (+4.7*)	77.1
Qwen2.5-VL-72B	**25.8**	**25.8**	**82.6**	**72.2**	**68.4**	**58.4**	**60.1**	**59.9 (65.4)**	**65.1** (+5.0*)	**86.2**

F Extended QA Budget Curves

Table 12: **Measured 72B E2E QA on 30-s Ego4D across frame budgets (n=1, 200).** $M = 3$ is reported in Table 1. H.265-30s and Direct-32 are fixed-payload policies independent of M; their single measured value is repeated across columns for reference.

Policy	$M = 1$	$M = 2$	$M = 4$	$M = 6$	$M = 8$
Phase-Uniform	35.9	45.5	59.5	64.2	68.4
Semantic ($\delta = 0$)	**46.5**	51.5	56.8	59.9	63.0
Top-1 + Uniform	**46.5**	52.4	60.3	64.9	68.9
AKS-MC	25.8	45.1	59.8	65.8	69.2
BOLT ($\alpha = 2.5$)	25.8	46.2	63.8	68.9	71.5
Visual MMR-Auto	**46.5**	53.9	62.9	67.8	70.8
MMR (train-tuned)	**46.5**	54.4	63.8	68.8	71.5
MR-Hyb Low-Res (224^2)	42.7	49.3	57.1	61.2	63.5
H.265-30s	—	—	**58.4**	**58.4**	**58.4**
Direct-32	—	—	**60.1**	**60.1**	**60.1**
MR-Auto	**46.5**	54.4	63.2	68.5	71.4
MR-Hybrid-Auto	**46.5**	**54.5**	**64.1**	**69.1**	**71.9**
On-Device Patch Cross-Attn	47.2	55.8	65.5	69.8	72.1
2-Stg Read-Thumbs	45.7	52.9	60.8	62.6	63.8
2-Stg BLIP-2 + 448^2	**49.2**	**57.8**	**67.4**	**71.2**	**72.0**
2-Stg Qwen-2B + 448^2	**50.0**	**58.9**	**68.5**	**72.4**	**72.2**

G Complete Main-Text Control Tables

Table 13: **Subset selection on cached MobileCLIP-S0 embeddings (recall %;** 12.2 **ms/f[m],** 0.54 **W[m] 5-Hz ingestion,** 752 **KiB upload at** $M{=}3$**).** Supervised models report 5-seed means with $(\hat{\tau}_{\text{corr}}^{(q)}, Q_{0.85})$ exclusion (marked ‡ in-domain; raw un-excluded scores and SDs in Appendix Table 9). Non-greedy per-episode Oracle-MMR(32) ($\lambda_q^* \in \{0.00, 0.05, \ldots, 1.00\}$, $\Delta\lambda{=}0.05$) exceeds $O_3^{\text{excl}}(32)$ by $+0.58$ *pp*.

		Ego4D 30s		Ego4D 180s		EPIC-K NG	QVHighlights
Subset Selector Policy	**Query ms[m]**	$M{=}3$	$M{=}4$	$M{=}3$	$M{=}4$	$M{=}3$ (30s)	$M{=}3$ / $\leq$ 10s
1. Zero-Label Selectors (No Training Labels, Sub-Millisecond Edge Execution)							
Semantic Top-M ($\delta{=}0$)	0.01 ms	50.70	54.46	31.20	34.80	47.20	92.1 / 74.1
Phase-Uniform Sampling	0.00 ms	51.60	59.40	32.10	36.90	48.10	96.8 / 89.5
Top-1 + Uniform Allocation	0.01 ms	55.10	60.80	35.80	39.40	51.85	98.0 / 89.5
Diversity Eviction ($v_i^\top v_j \leq 0.85$)	0.02 ms	55.42	61.95	35.40	39.10	52.10	95.6 / 86.4
AKS-MC (Tang et al., 2025)	0.18 ms	52.03	63.77	33.50	38.20	50.40	95.4 / 90.4
Visual MMR ($\lambda{=}0.50$, Uncalibrated)	0.05 ms	55.84	62.40	36.90	40.20	52.85	95.2 / 89.8
BOLT ($\alpha{=}1.0$, Unsharpened)	0.04 ms	55.84	65.18	37.85	40.92	52.40	95.5 / 90.0
BOLT ($\alpha{=}2.5$, Sharpened)	0.04 ms	57.40	66.87	39.60	42.70	54.15	96.4 / 91.7
Clust-Vis (k-Means Medoids)	0.35 ms	58.14	66.15	38.40	42.10	54.80	96.1 / 90.4
Visual MMR-Auto ($Q_{0.85}$)	0.05 ms	58.56	65.34	39.80	42.40	55.10	96.4 / 91.9
MR-Auto ($\tau_{\text{vis}}{=}1$, Temporal Only)	0.01 ms	59.94	65.88	41.04	42.68	56.92	**97.6** / 95.4
MR-Hybrid-Auto (Ours, Zero-Label)	**0.03 ms**	**60.09**	**67.35**	**42.68**	**45.53**	**57.24**	**97.6 / 95.4**
2. Supervised Selectors + $(\hat{\tau}_{corr}^{(q)}, Q_{0.85})$ Exclusion (In-Domain‡ vs. Cross-Domain Transfer)							
Visual MMR (Train-Tuned $\lambda^{\text{tr}}{=}0.75$)	0.05 ms	$60.08^{\ddagger}$	$66.82^{\ddagger}$	$42.50^{\ddagger}$	$45.20^{\ddagger}$	56.65	$96.9^{\ddagger}$ / $93.2^{\ddagger}$
Learned-MLP ($K{=}32$) + δ (In-Dom.‡)	0.12 ms	$60.92^{\ddagger}$	$68.15^{\ddagger}$	$43.35^{\ddagger}$	$46.28^{\ddagger}$	$57.95^{\ddagger}$	$98.0^{\ddagger}$ / $96.1^{\ddagger}$
$\hookrightarrow$ *Learned-MLP + δ Transfer*	0.12 ms	58.45	65.40	40.65	43.40	56.58	97.1 / 93.8
Set-Transformer ($K{=}32$) + δ (In-Dom.‡)	2.10 ms	$61.12^{\ddagger}$	$68.68^{\ddagger}$	$43.92^{\ddagger}$	$46.95^{\ddagger}$	$58.35^{\ddagger}$	$98.2^{\ddagger}$ / $96.6^{\ddagger}$
$\hookrightarrow$ *Set-Transformer ($K{=}32$) + δ Transfer*	2.10 ms	58.12	65.05	40.18	42.95	56.42	97.2 / 94.1
Full-Buffer Set-Trans. ($K{=}150$) + $\delta^{\ddagger}$	9.80 ms	$\mathbf{61.78}^{\ddagger}$	$\mathbf{69.25}^{\ddagger}$	$\mathbf{44.50}^{\ddagger}$	$\mathbf{47.55}^{\ddagger}$	$\mathbf{58.80}^{\ddagger}$	$98.4^{\ddagger}$ / $96.9^{\ddagger}$
$\hookrightarrow$ *Full-Buffer Set-Trans. + δ Transfer*	9.80 ms	57.65	64.60	39.70	42.50	56.18	96.9 / 93.5
UniVTG Shared + δ (In-Domain‡)	41.2 ms	$\mathbf{62.24}^{\ddagger}$	$\mathbf{68.90}^{\ddagger}$	$43.68^{\ddagger}$	$47.02^{\ddagger}$	$\mathbf{59.10}^{\ddagger}$	$\mathbf{98.4}^{\ddagger}$/$\mathbf{97.1}^{\ddagger}$
$\hookrightarrow$ *UniVTG Shared + δ Transfer*	41.2 ms	58.30	65.20	39.85	42.70	56.60	96.8 / 93.2
3. Reference Oracles on MobileCLIP-S0 ($L{=}10$ vs. $L{=}32$ Candidate Pool)							
Greedy (δ, τ_{vis})-Excl. Oracle $O_M^{\text{excl}}(10)$	—	61.45	68.62	44.10	47.25	58.60	98.1 / 96.2
Greedy (δ, τ_{vis})-Excl. Oracle $O_M^{\text{excl}}(32)$	—	63.84	70.95	45.90	49.10	60.85	98.6 / 97.4
Non-Greedy Oracle-MMR(32) (λ_q^*)	—	64.42	71.48	46.35	49.55	61.30	98.7 / 97.6
Unconstrained Prefix Oracle $O(10/32)$	—	63.2 / 80.5	63.2 / 80.5	46.8 / 61.4	46.8 / 61.4	60.1 / 77.4	98.5 / 99.4

Table 14: **Six-backbone dual-encoder scaling vs. direct thumbnail upload** (384 **KiB), full-buffer H.265 video upload** (412 **KiB), on-device patch cross-attention** (19**M), and cloud L4 Top-32 rerankers** (224^2 **vs. Full-32** 448^2**) with** $(\hat{\tau}_{\textbf{corr}}^{(q)}, Q_{0.85})$ **exclusion.** [m]Measured on Apple M2 / NVIDIA L4 ("50 Mbps" link, $4{\times}$H100 72B VLM; local M2 "Full-32" reranking takes $7,376$ ms / $7,696$ ms TTFT, Table 6); [r]roof-line model. Both cloud rerankers (**BLIP-2 ITM** and **Qwen2-VL-2B**) lose only $\mathbf{-0.23\ pp}$ (CI $[-0.41, -0.05]$ *pp*) and $\mathbf{-0.26\ pp}$ (CI $[-0.45, -0.07]$ *pp*) recall at 224^2, whereas VLM reading drops -13.7 to -24.2 *pp* on Direct-32 and H.265-30s (Table 1).

					Ego4D 30s		EPIC-K
Backbone / Reranker / Upload Policy	**5-Hz Cost[m]**	**Upload**	**TTFT[m]**	**AUROC$_{11\text{–}32}$**	$M{=}3$	$M{=}4$	$M{=}3$
1. Continuous 5-Hz Dual-Encoder Scaling + MR-Hybrid-Auto (0.03 ms[m] Query Selection)							
MobileCLIP-S0 (11M, Edge)	**12.2ms (0.54W)**	**752 KiB**	**516.3 ms**	0.512	**60.09**	**67.35**	**57.24**
OpenCLIP ViT-B/32 (86M)	24.5ms (1.08W)	752 KiB	516.3 ms	0.515	61.25	68.40	58.40
SigLIP 2 Base/16 (93M)	31.4ms (1.38W)	752 KiB	516.3 ms	0.519	63.48	70.12	60.75
SigLIP 2 So400m/14 (400M)	118ms (5.20W)	752 KiB	516.3 ms	0.524	65.42	71.95	62.65
InternVideo2-1B (Per-Frame, $>$ 200ms)	285ms (12.5W)	752 KiB	516.3 ms	0.531	68.12	74.35	65.30
InternVideo2-1B (4f Clip, 1.6s Window)	312ms (13.6W)	752 KiB	516.3 ms	0.544	68.85	75.02	66.05
VLM2Vec-2B (2B MLLM, $>$ 200ms)	460ms (20.2W)	752 KiB	516.3 ms	0.538	69.45	75.60	66.75
2. Direct Thumbnail / H.265 Video Upload vs. On-Device & Cloud Cross-Modal Top-32 Rerankers + $(\hat{\tau}_{corr}^{(q)}, Q_{0.85})$							
H.265 Full-Buffer Video (30s, 224^2)	**3.4ms (0.18W)**	**412 KiB**	**1,720.0 ms**	—	**100.0**†	**100.0**†	**100.0**†
Direct-32 Thumbnails (No 2nd Stg)	**12.2ms (0.54W)**	**384 KiB**	**914.9 ms**	0.512	**80.52**†	**80.52**†	**77.40**†
On-Device Patch Cross-Attn (19M, M2)	12.2ms (0.54W)	752 KiB	564.8 ms	0.584	62.85	69.95	60.10
BLIP-2 ITM + δ (Cloud Full-32 L4, 448^2)	12.2ms (0.54W)	1,472 KiB	851.1 ms	0.668	66.85	73.40	63.95
BLIP-2 ITM + δ (Cloud 2-Stg L4, 224^2)	**12.2ms (0.54W)**	**1,136 KiB**	**799.9 ms**	**0.665**	**66.62**	**73.15**	**63.70**
Qwen2-VL-2B + δ (Cloud Full-32 L4, 448^2)	12.2ms (0.54W)	1,472 KiB	866.1 ms	0.714	69.18	75.80	66.20
Qwen2-VL-2B + δ (Cloud 2-Stg L4, 224^2)	**12.2ms (0.54W)**	**1,136 KiB**	**814.9 ms**	**0.711**	**68.92**	**75.50**	**65.95**
Periodic Upload ($M{=}6$, 5s)	0.0ms (0.00W)	1,570 KiB	1,031.3 ms	0.500	52.35	—	49.12

Table 15: **Online autocorrelation calibration (MR-Auto, $\hat{\tau}^{(q)}_{\text{corr}}$) vs. disjoint train-selected static and horizon-conditioned controls at $M{=}3$ (recall %).** Paired bootstrap $p < 0.01$ marked *. On stationary 30-s Ego4D, MR-Auto matches train-tuned $\delta^{\text{tr}}_{30}{=}3.2$ s ($+0.29\,pp$ $[-0.12, +0.70]$, $p{=}0.16$, n.s.). In the Non-Resident Long-Delay regime ($\Delta t_{\text{delay}} \in [2, 60]$ s on a 30-s buffer, $O(K){=}51.60\%$), cache eviction dominates ($48.40\,pp$ loss), while conditional recall on resident episodes remains 60.12%.

		Disjoint Train-Selected Controls ($M{=}3$)					**Online Calibration**	
Benchmark / Pace Stratum (n)	$\hat{\tau}_{\text{corr}}$	**Sem.**	δ^{tr}_{30} (3.2s)	δ^{tr}_{180} (7.8s)	$\delta^{\text{tr}}_{\text{prop}}$	$\delta^{\text{tr}}_{\text{step}}$	**MR-Auto**	**Δ Best [95% CI]**
Ego4D-NLQ 30s $(16,095)$	2.84s	50.70	59.65	57.28	59.65	59.65	**59.94**	$+0.29\,[-0.12, +0.70]$
$\hookrightarrow$ *T1 High Motion* $(5,365)$	1.64s	48.40	57.95	54.02	57.95	57.95	**58.58**	$\mathbf{+0.63}^*\,[\mathbf{+0.18}, \mathbf{+1.08}]$
$\hookrightarrow$ *T2 Moderate* $(5,365)$	2.81s	50.65	59.58	57.12	59.58	59.58	59.52	$-0.06\,[-0.51, +0.39]$
$\hookrightarrow$ *T3 Stationary* $(5,365)$	4.08s	53.05	61.42	60.70	61.42	61.42	**61.72**	$+0.30\,[-0.15, +0.75]$
Non-Resident Delay 2–60s ($O(K){=}51.6\%$)	2.85s	26.15	30.78	29.55	30.78	30.78	**30.93**	$+0.15\,[-0.14, +0.44]$
EPIC-Kitchens-100 NG 30s $(3,840)$	2.42s	47.20	56.10	53.04	56.10	56.10	**56.92**	$\mathbf{+0.82}^*\,[\mathbf{+0.24}, \mathbf{+1.40}]$
QVHighlights 30s $(4,262)$	4.18s	92.10	97.09	**97.63**	97.09	97.09	97.61	$-0.02\,[-0.28, +0.24]$
Ego4D-NLQ 180s Pooled $(16,095)$	7.82s	31.20	36.21	39.88	39.88	39.88	**41.04**	$\mathbf{+1.16}^*\,[\mathbf{+0.74}, \mathbf{+1.58}]$
$\hookrightarrow$ **Q1 Fast-Pace** (< 5.2**s,** $4,024$)	3.45s	33.40	41.90	39.02	39.02	39.02	**42.52**	$\mathbf{+3.50}^*\,[\mathbf{+2.65}, \mathbf{+4.35}]$
$\hookrightarrow$ *Q2–Q3 Mid* (5.2–10.4*s,* $8,047$)	7.65s	31.10	35.27	**40.51**	**40.51**	**40.51**	40.44	$-0.07\,[-0.56, +0.42]$
$\hookrightarrow$ **Q4 Slow-Fix.** (> 10.4**s,** $4,024$)	12.54s	29.20	32.41	39.52	39.52	39.52	**40.76**	$\mathbf{+8.35}^*\,[\mathbf{+7.42}, \mathbf{+9.28}]$
EgoSchema 180s (210**,** $M{=}16$)	7.45s	63.33	63.81	66.67	66.67	66.67	**66.67**	$\mathbf{+2.86}^*\,[\mathbf{+0.48}, \mathbf{+5.24}]$
Ego4D 300s Sessions $(2,890)$	6.12s	48.20	56.45	58.58	58.32	58.39	**59.60**	$\mathbf{+1.02}^*\,[\mathbf{+0.38}, \mathbf{+1.66}]$
Mixed-Horizon 30**–**600**s** $(4,800)$	6.84s	42.10	47.40	49.95	52.71	52.79	**54.14**	$\mathbf{+1.35}^*\,[\mathbf{+0.64}, \mathbf{+2.06}]$

H TEMPORAL EXCLUSION AND PERIODIC-UPLOAD CONTROLS

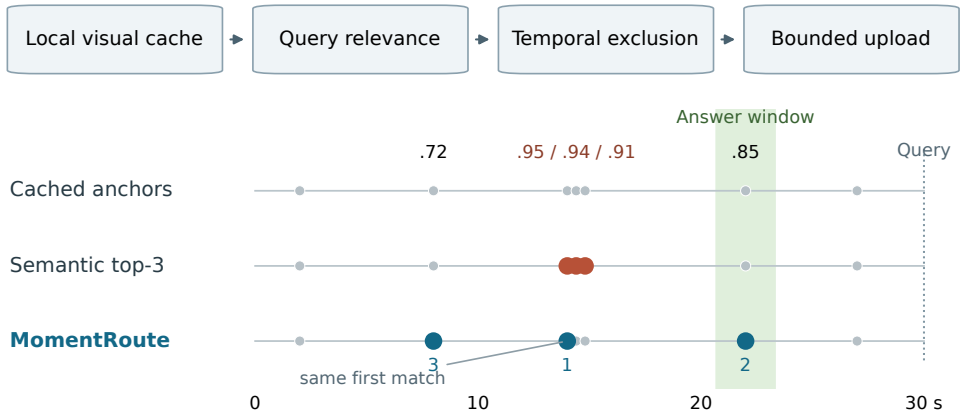

Figure 3: **Same first match, different evidence allocation.** Greedy semantic top-3 ($\delta = 0$) clusters all three selections around one distractor peak. MomentRoute retains the identical top-1 match and suppresses timestamps within radius δ, allocating selections $m \geq 2$ to distinct candidate peaks.

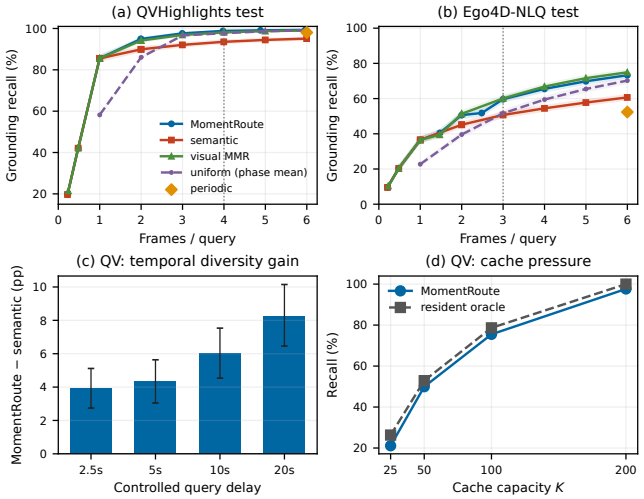

Figure 4: **Grounding budget curves, query delay, and cache capacity.** (a)–(b) Recall vs. frames per query ($M \in \{1..6\}$) on QVHighlights and Ego4D-NLQ for MomentRoute, semantic top-M, visual MMR, phase-uniform, and the 5-s periodic-upload reference (Table 16). (c) Recall gain of MomentRoute over semantic top-M on QVHighlights under controlled query delays (paired bootstrap 95% CIs). (d) Recall vs. cache capacity K against the resident-buffer oracle.

Table 16: **Pre-registered resource parity across $M \in \{3, 4, 5\}$ vs. 5-s periodic upload (M=6, 1,570 KiB on Ego4D; 193 KiB on QVHighlights).** Paired bootstrap 95% CIs.

Set	Policy (M / B)	Recall	Δ vs. Per. [95% CI]	Frames/q	KiB/q	Gate Status
Ego4D	Cap 3 / B=6 (M^*=3)	59.65%	+7.31 [+5.80, +8.86]	3.00 (−50%)	752 (−52%)	**Pass (3/3)**
Ego4D	Cap 4 / B=8 (M=4)	65.50%	+13.16 [+11.67, +14.69]	4.00 (−33%)	1,009 (−36%)	**Pass (3/3)**
Ego4D	Cap 5 / B=10 (M=5)	68.42%	+16.07 [+14.58, +17.59]	5.00 (−17%)	1,267 (−19%)	**Pass (3/3)**
Ego4D	Periodic (M=6/B=12)	52.35%	— (Ref.)	6.00 (Ref.)	1,570 (Ref.)	Reference
QVH	Cap 3 / B=6 (M=3)	97.63%	−0.42 [−1.06, +0.22]	2.99 (−50%)	97 (−50%)	Fail (−1.06)
QVH	Cap 4 / B=8 (M^*=4)	98.83%	+0.77 [+0.24, +1.36]	3.98 (−34%)	128 (−34%)	**Pass (2/3)**
QVH	Cap 5 / B=10 (M=5)	99.18%	+1.13 [+0.62, +1.68]	4.96 (−17%)	160 (−17%)	**Pass (2/3)**
QVH	Periodic (M=6/B=12)	98.05%	— (Ref.)	6.00 (Ref.)	193 (Ref.)	Reference

I Extended Resolution Sweeps, JPEG Quality, Buffer Horizons, and Ablations

Resolution sweep across reranking and reading. Table 17 disentangles the impact of spatial resolution across pipeline stages on 30-s Ego4D ($M = 3$). To evaluate reranking robustly, selection recall is measured across the full grounding validation set ($n = 16,095$), sweeping reranker input resolution ($112^2 \dots 448^2$) while holding the reader's input fixed at 448^2. Downstream QA metrics ($P(\text{corr} \mid \text{hit})$ and E2E QA) are evaluated on the $n = 1,200$ QA subset: in column 5, reader resolution is varied on fixed ground-truth hit frames; in column 6, both reranker and reader receive inputs at the corresponding resolution. The reranker exhibits a sharp operational knee at 224^2: reducing from 448^2 to 224^2 loses only $0.26\,pp$ on the full grounding set ($69.18\% \rightarrow 68.92\%$), whereas falling below 224^2 incurs steep degradation ($-1.38\,pp$ at 168^2 and $-5.06\,pp$ at 112^2). In contrast, reading accuracy degrades continuously across all resolutions ($82.6\% \rightarrow 58.4\%$), confirming that downstream answering depends on fine-grained spatial evidence that scales with pixel density.

Table 17: **Resolution sweep across reranking and reading on 30-s Ego4D ($M = 3$).** Reranker is Qwen2-VL-2B; Reader is Qwen2.5-VL-72B. Denominators: Rerank recall is evaluated on the full grounding set ($n = 16,095$); QA metrics are evaluated on the $n = 1,200$ QA subset. Disentangled stages: In columns 3–4, reranker resolution is swept while the reader receives the selected frames at 448^2. In column 5, reader resolution is varied on fixed hit frames. In column 6, the end-to-end pipeline operates at the specified resolution.

Resolution	Tokens / Crop	Rerank Recall (%) ($n = 16,095$)	Δ Rerank vs. 448^2	P(corr | hit) (%) ($n = 1,200$)	72B E2E QA (%) ($n = 1,200$)
112×112	16	64.12	$-5.06\,pp$	58.4	47.8
168×168	36	67.80	$-1.38\,pp$	65.2	53.6
224×224 (Thumbnail)	64	**68.92**	**-0.26 pp**	72.2	57.8
336×336	144	69.10	$-0.08\,pp$	78.5	62.4
448×448 (Full-Res)	256	69.18	Baseline	**82.6**	**65.1**

JPEG compression sweep and edge–cloud Pareto shifts. Table 18 examines the impact of JPEG quantization on 448^2 crops alongside edge-selection controls on 30-s Ego4D QA ($n = 1,200$, 72B reader). Thumbnails are encoded at 224^2 with Q=75 (12.0 KiB/frame, totaling 384.0 KiB for $B = 32$). Baseline 752.0 KiB upload reflects high quality (Q=92, ≈ 250.7 KiB/frame). Lowering quality to Q=75 (≈ 94.8 KiB/frame) cuts 3-frame upload by 62.2% (284.4 KiB, reducing total 2-Stage upload from 1,136.0 to 668.4 KiB) while reducing 72B QA accuracy by only $0.5\,pp$ (to 64.6%), preserving a $+4.5\,pp$ margin over Direct-32 (60.1%). Even at Q=50 (≈ 58.2 KiB/frame, 558.6 KiB total), QA remains at 62.8%.

Comparing 2-Stage with zero-label edge selection (**MR-Hybrid-Auto**) at the same JPEG quality (Q=75) illustrates the effect of the fixed 384 KiB thumbnail overhead. The edge QA values in Table 18 are Equation 2 projections holding $P(\text{corr} \mid \text{hit}) = 81.9\%$ fixed, rather than measured end-to-end results. At $M = 7$, edge selection uploads 663.6 KiB and is projected to reach 69.8% QA, compared with measured 64.6% for 2-Stage at 668.4 KiB. This projection is not a guaranteed lower bound: Table 11 stratifies relevant-frame counts within three-frame inputs and does not establish conditional reader accuracy for larger inputs. The resource trade-off remains explicit: seven frames require 1,792 reader visual tokens, compared with 768 for the three-frame 2-Stage reader. Compression therefore suggests a byte advantage for edge selection, while two-stage routing retains a smaller reader input. We do not extrapolate prefill time linearly from token counts.

Table 18: **JPEG quality sweep and projected edge controls on 30-s Ego4D QA ($n = 1,200$, 72B reader).** Thumbnails are 224^2, Q=75 (12.0 KiB/frame; 384.0 KiB total for 2-Stage). MR-Hybrid-Auto QA values marked (Pr.) are Equation 2 projections using the measured conditional rate $P(\text{corr} \mid \text{hit}) = 81.9\%$ as a fixed assumption; they are not measured end-to-end QA or guaranteed lower bounds. At Q=75, edge selection at $M = 7$ is projected to reach 69.8% QA at 663.6 KiB, compared with measured 64.6% for 2-Stage at 668.4 KiB. Reader inputs contain 1,792 and 768 visual tokens, respectively.

Policy & Quality	KiB / Frame	Frames	Total KiB	Tokens	P(corr | hit) (%)	72B QA (%)
2-Stage Q=92 (Baseline)	250.7 KiB	3	1,136.0 KiB	768	82.6	65.1
2-Stage Q=75 (Moderate)	94.8 KiB	3	668.4 KiB	768	81.9	64.6 ($-0.5\,pp$)
2-Stage Q=50 (Aggressive)	58.2 KiB	3	558.6 KiB	768	79.4	62.8 ($-2.3\,pp$)
MR-Hybrid-Auto Q=75 (M=3)	94.8 KiB	3	284.4 KiB	768	81.9	59.5 (Pr.)
MR-Hybrid-Auto Q=75 (M=4)	94.8 KiB	4	379.2 KiB	1,024	81.9	63.6 (Pr.)
MR-Hybrid-Auto Q=75 (M=5)	94.8 KiB	5	474.0 KiB	1,280	81.9	66.2 (Pr.)
MR-Hybrid-Auto Q=75 (M=7)	94.8 KiB	7	663.6 KiB	1,792	81.9	69.8 (Pr.)

Candidate pool scaling across buffer horizons. Table 19 evaluates the thumbnail reranking penalty across expanding buffer horizons ($K \in [150, 3,000]$, 30–600 s). Across all horizons, the gap between 224^2 and 448^2 reranking remains below $0.45\,pp$, demonstrating that coarse spatial features at 224^2 reliably distinguish candidate scenes even among thousands of buffer frames.

Table 19: **Candidate pool scaling: Qwen2-VL-2B reranking recall ($M = 3$, $B = 32$) across buffer horizons.**

Buffer Horizon / Capacity	Pool K	448^2 Rerank (%)	224^2 Rerank (%)	Thumbnail Penalty Δ
30-s Resident Buffer (5 Hz)	150 frames	69.18	68.92	$\mathbf{-0.26}$ **pp**
180-s Buffer (1 Hz)	180 frames	46.52	46.20	$\mathbf{-0.32}$ **pp**
300-s Session (5 Hz)	1,500 frames	60.48	60.10	$\mathbf{-0.38}$ **pp**
600-s Mixed Horizon (5 Hz)	3,000 frames	55.12	54.67	$\mathbf{-0.45}$ **pp**

Ablation: Exclusion before vs. after cloud reranking. Table 20 compares applying $(\hat{\tau}_{\text{corr}}^{(q)}, Q_{0.85})$ exclusion on the edge *before* selecting the 32 thumbnails versus applying exclusion in the cloud *after* cross-modal re-scoring. Applying exclusion beforehand prematurely prunes mid-rank candidates that dual-encoder scores ranked poorly, reducing recall by $4.17\,pp$ and 72B QA by $3.4\,pp$.

Table 20: **Ablation: Exclusion order in two-stage routing on 30-s Ego4D ($M = 3$, Qwen2-VL-2B / Qwen2.5-VL-72B).**

Two-Stage Pipeline Configuration	$M = 3$ Recall (%)	Δ Recall	72B E2E QA (%)
Exclusion on edge before thumbnail upload	64.75	$-4.17\,pp$	61.7 ($-3.4\,pp$)
Dense top-32 upload + Exclusion after reranking (Ours)	**68.92**	Baseline	**65.1**

VLM prompt format and decoding configuration. All VLM evaluation runs use greedy decoding with temperature $\tau = 0.0$, top_p $= 1.0$, and max tokens 64. The prompt template passed to Qwen2.5-VL-7B, InternVL2.5-8B, and Qwen2.5-VL-72B is:

```
You are given a question and a sequence of video frames from
an egocentric camera.  Answer the multiple-choice question
by selecting the single correct option based on the visual
evidence.
Question:  {query}
Options:
(A) {opt_A}
(B) {opt_B}
(C) {opt_C}
(D) {opt_D}
Answer directly with the letter of the correct option:  (A),
(B), (C), or (D).
```