WHERE DOES THE VIDEO GO? PROBING PERSISTENT TEMPORAL MEMORY AFTER VISUAL KV EVICTION IN VIDEO LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video large language models (video LLMs) represent a video as a long multimodal prefix, producing large key–value (KV) caches that can remain resident when the same video is reused for multiple temporal queries. Existing efficiency methods primarily reduce visual computation before or during multimodal processing, or compress the resulting cache by retaining entries selected using importance or relevance scores. What remains unclear is which KV states must stay resident once the complete video prefix has already participated in multimodal prefill. We study this post-prefill setting, in which the complete video prefix is processed once and its KV cache is physically compacted before any downstream query is observed. Across three video LLM families, we find that substantial temporal-grounding ability remains after most visual-payload KV states are removed when the compact cache retains sparse temporal-carrier positions aligned with each model's video serialization. Under identical physical KV budgets, retaining these carrier positions consistently outperforms replacing them with additional visual-payload states, showing that post-prefill memory is functionally organized rather than uniformly distributed across the video prefix. Controlled carrier-content interventions further show that the retained states carry information tied to the processed video, rather than acting only as generic positional scaffolds. Motivated by these findings, we introduce TC-KV, a simple, training-free, and query-independent cache policy that preserves the temporal-carrier structure together with a small visual residual after complete prefill. TC-KV thereby substantially reduces persistent video-prefix KV and enables the same compact memory to serve multiple temporal queries, without retraining or modifying the complete multimodal prefill.

1 INTRODUCTION

Video large language models (video LLMs) encode videos as extended multimodal prefixes to support video understanding and temporal localization (Lin et al., 2024; Ren et al., 2024; Bai et al., 2025). These prefixes produce large key–value (KV) caches that can remain resident when the same video is reused for multiple temporal queries, making cache retention an important efficiency concern (Kim et al., 2025; Yang et al., 2025b). Existing methods reduce visual computation before or during multimodal processing through token pruning or merging (Chen et al., 2024; Zhong et al., 2025; Yang et al., 2025a; Shang et al., 2025), or compress the resulting cache by retaining entries selected using importance or relevance scores (Liu et al., 2023; Li et al., 2024; Cai et al., 2024; Tu et al., 2025). These methods improve efficiency but leave a more basic representational question underexplored: once the complete video prefix has participated in multimodal prefill, which KV states must remain resident for later temporal queries? The states required to process a video need not be identical to those required to remember it.

Streaming video-memory methods maintain bounded caches as frames arrive (Kim et al., 2025; Yang et al., 2025b). We study the post-prefill regime, treating the completed video prefix as reusable memory. The model processes the full prefix once; before any downstream query is observed, we physically compact its KV cache and reuse the retained states across queries about the same video. As illustrated in Figure 1, this setting separates memory formation from later readout and asks three questions: how much temporal capability survives visual-payload KV eviction, whether survival

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.

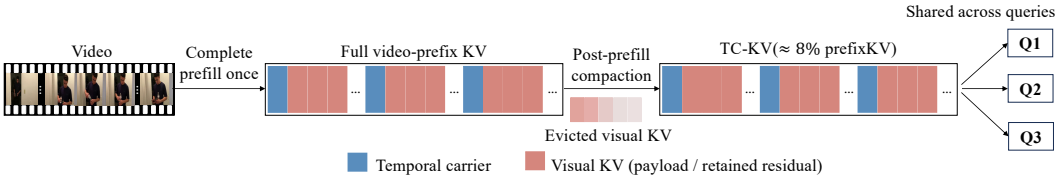


Figure 1: Overview of the post-prefill setting and TC-KV. The complete video prefix participates in multimodal prefill once, producing a full KV cache dominated by visual-payload states. Before any downstream query is observed, TC-KV evicts most visual-payload KV while preserving the model-aligned temporal-carrier topology together with a small visual residual. The resulting query-independent compact cache is shared across multiple temporal queries without re-encoding the video.

depends on the organization of retained states beyond cache size, and whether those states carry information about the processed video rather than serving only as a generic positional scaffold.

Video LLMs differ in how they serialize visual tokens, timestamps, frame markers, and structural boundaries (Bai et al., 2025; An et al., 2025; Wang et al., 2025). We refer to sparse positions associated with temporal units and aligned with each model's serialization as *temporal carriers*, and to their arrangement among visual blocks as *carrier topology*. Our central hypothesis is that post-prefill KV states are functionally heterogeneous: dense visual-payload states participate in processing the video, while sparse carrier states can acquire contextual information during prefill and preserve access to temporal evidence afterward. If so, carrier-aligned retention should support later temporal readout more effectively than retaining the same number of visual-payload states.

Across three video-LLM families, we find that substantial temporal-grounding capability survives visual-payload KV eviction, and carrier-aligned retention outperforms equal-budget visual fill. Further interventions reveal video-dependent carrier content and support a *write–persist–read* account: visual processing forms carrier state during prefill, which remains available for later temporal queries after most visual KV is evicted.

These findings motivate TC-KV, a training-free and query-independent policy that preserves model-aligned temporal carriers and a small visual residual. Retaining only 7.42–8.14% of video-associated prefix KV, TC-KV supports temporal grounding and highlight detection while reducing persistent memory and query-stage cost. The compact cache is reusable across queries without re-encoding the video. Our contributions are threefold:

1. **A post-prefill memory formulation.** We formulate video-prefix KV retention as a persistent-memory problem, separating the states used to process a video from those that must remain resident for later temporal readout.
2. **Causal evidence for carrier-centered temporal memory.** Through matched-budget topology swaps, carrier-content transplantation, depth- and time-resolved interventions, and state restoration, we show that post-prefill temporal memory depends on structured, video-dependent carrier state rather than cache size or positional layout alone.
3. **A minimal operationalization.** We introduce TC-KV, a training-free and query-independent policy that retains model-aligned temporal carriers together with a small visual residual, substantially reducing persistent video-prefix KV across multiple video-LLM families and temporal-understanding tasks.

2 Related Work

Visual-Token Reduction and Analysis. A broad line of work reduces multimodal inference cost through visual-token pruning and merging, including FastV and AIM (Chen et al., 2024; Zhong et al., 2025). VisionZip and LLaVA-PruMerge compress visual representations (Yang et al., 2025a; Shang et al., 2025), while SIEVE uses cross-modal residual guidance to prune visual tokens during language-model computation (Ou et al., 2026). Mechanistic studies further reveal heterogeneous roles among visual tokens, distinguishing input-agnostic sink or inactive tokens from semantically

active representations (Fan et al., 2026). These works study visual representations while they are still participating in model computation. We instead ask which resulting KV states must remain resident after the complete video prefix has already been processed.

KV-Cache Compression. KV-cache compression reduces inference memory through entry selection, budget allocation, or low-precision representation. Scissorhands and SnapKV exploit token importance and attention patterns to select retained entries (Liu et al., 2023; Li et al., 2024), while PyramidKV and VL-Cache adapt cache budgets to layer-wise or modality-specific characteristics (Cai et al., 2024; Tu et al., 2025). KIVI and VidKV instead compress the numerical representation of cached states (Liu et al., 2024; Tao et al., 2025). We examine a complementary structural question: whether components of a serialized video prefix occupy distinct functional roles in persistent memory. Matched-budget topology swaps and carrier-content interventions distinguish the effects of cache quantity, carrier organization, and video-dependent state.

Streaming and Reusable Video Memory. Streaming methods maintain bounded memory as inputs arrive, using persistent attention-sink or query-agnostic states (Xiao et al., 2024; Kim et al., 2025; Yang et al., 2025b). We consider a completed-prefix reuse setting: the full video is processed once, its KV cache is compacted before any downstream query is observed, and the resulting memory is reused across queries. This setting separates memory formation from later readout and allows us to study how previously processed visual evidence remains accessible after most visual-payload KV has been evicted.

3 Post-Prefill Memory and Intervention Framework

Post-prefill memory and cache decomposition. Let $x(v)$ denote the multimodal prefix constructed from a video v. We study the cache state after the model has processed the complete prefix $x(v)$ but before any downstream query is observed. This video-prefix prefill produces a layer-wise KV cache

$$\mathcal{M}^{\text{full}}(v) = \left\{ \left(K_l^{\text{full}}(v), V_l^{\text{full}}(v) \right) \right\}_{l=1}^{L}, \tag{1}$$

where L is the number of transformer layers. We partition the prefix positions into three disjoint operational groups:

$$\mathcal{I}_{\text{pref}} = \mathcal{S} \dot{\cup} \mathcal{C} \dot{\cup} \mathcal{P}. \tag{2}$$

Here, $\mathcal{S}$ contains static prompt and boundary states that remain fixed across interventions, $\mathcal{C}$ contains sparse temporal-carrier positions aligned with the model's serialization of temporal units, and $\mathcal{P}$ contains the dense visual-payload positions produced from the video input. The exact token-level realization of $\mathcal{C}$ is model-dependent; model-specific definitions are provided in the appendix.

Given a retained position set $\mathcal{A} \subseteq \mathcal{I}_{\text{pref}}$, physical cache compaction gathers the corresponding entries from every layer:

$$\mathcal{G}\big(\mathcal{M}^{\text{full}}(v), \mathcal{A}\big) = \left\{ \left(K_l^{\text{full}}(v)[\mathcal{A}], V_l^{\text{full}}(v)[\mathcal{A}] \right) \right\}_{l=1}^{L}. \tag{3}$$

The retained entries preserve their original order and positional metadata. A later query is processed using the compacted cache without re-encoding the video. Because compaction occurs before the query is observed, the resulting cache can be reused across multiple queries about the same video.

Controlled cache interventions. We construct a sequence of interventions that separates persistence, topology, and content effects. The *carrier-only* condition retains the static states and complete carrier topology while removing every visual-payload position. TC-KV additionally retains a small visual residual $\mathcal{R} \subseteq \mathcal{P}$:

$$\begin{aligned} \mathcal{M}^{\text{carrier}}(v) &= \mathcal{G}\big(\mathcal{M}^{\text{full}}(v), \mathcal{S} \cup \mathcal{C}\big), \\ \mathcal{M}^{\text{TC-KV}}(v) &= \mathcal{G}\big(\mathcal{M}^{\text{full}}(v), \mathcal{S} \cup \mathcal{C} \cup \mathcal{R}\big). \end{aligned} \tag{4}$$

The first condition tests whether temporal capability remains accessible without resident visual-payload KV; the second tests whether limited visual support complements the persistent carrier state.

Persistence alone does not establish that carrier organization matters. We therefore construct a matched-budget visual-fill control. The visual residual $\mathcal{R}$ is held fixed, while the carrier positions are replaced by a set of visual-payload positions $\mathcal{U}$ satisfying

$$\mathcal{U} \subseteq \mathcal{P} \setminus \mathcal{R}, \qquad |\mathcal{U}| = |\mathcal{C}|, \qquad \mathcal{M}^{\text{fill}}(v) = \mathcal{G}\big(\mathcal{M}^{\text{full}}(v), \mathcal{S} \cup \mathcal{R} \cup \mathcal{U}\big)\,. \tag{5}$$

Thus, $\mathcal{M}^{\text{TC-KV}}$ and $\mathcal{M}^{\text{fill}}$ retain the same static states, visual residual, and number of physical KV entries. Their paired difference isolates the effect of preserving carrier-aligned organization from the amount of retained cache.

We further intervene on carrier content while holding the recipient positions, positional phases, static states, visual residual, query, and physical KV budget fixed. Replacing carrier K/V states with those produced by another video tests whether carrier utility depends on video-specific content rather than positional layout alone. Independently sourcing carrier keys and values provides a factorial test of their individual and joint effects. Finally, a targeted write-side intervention perturbs visual processing during prefill and restores selected downstream states, testing whether visual evidence causally influences subsequent carrier state. Exact constructions and model-specific controls are described with the corresponding experiments and in the appendix.

TC-KV as reusable compact memory. TC-KV uses a deterministic, query-independent layout that preserves the complete model-aligned carrier topology. The visual residual is model-specific but query-independent: Qwen3-VL and LLaVA-OneVision-1.5 retain fixed first-and-center positions within each visual group, while InternVL3.5 uses a prefix-only successor-checkpoint rule. Its retention ratio over video-associated prefix positions is

$$\rho_{\text{TC-KV}} = \frac{|\mathcal{C}| + |\mathcal{R}|}{|\mathcal{C}| + |\mathcal{P}|}. \tag{6}$$

Static prompt and boundary states are excluded from this ratio because they remain unchanged across cache policies. The layout is determined once from the completed video prefix and reused for all subsequent queries. TC-KV requires neither additional training nor query-dependent attention-score computation. It leaves the complete multimodal prefill unchanged and modifies only which prefix KV states remain resident afterward.

4 Experiments and Analysis

4.1 Experimental Setup

Evaluation protocol. We first perform a complete multimodal prefill for each video. TC-KV constructs its compact cache before any downstream query is observed and reuses it across queries without re-encoding the video. The attention-based selector baselines instead access the full prefix at query arrival to compute their retention scores. Within each controlled comparison, we keep the model, video, query, decoding configuration, and prefill computation fixed, varying only the cache selection or specified intervention. The uncompressed cache, denoted as Full replay, serves as the reference. Baseline implementations and their information access are detailed in Appendix A.3.

Metrics and statistics. We report mIoU and recall for temporal grounding and moment retrieval, and the official mAP and Hit@1 metrics for highlight detection. Retention ratios are computed from the KV entries physically gathered in every transformer layer, excluding static prompt and boundary states that remain unchanged across policies. For paired comparisons, 95% confidence intervals are obtained by video-level bootstrap resampling, keeping all queries from the same video in the same resampled cluster. Model-specific carrier definitions, generation settings, parsing details, and cache-integrity checks are provided in the appendix.

4.2 Post-Prefill Persistence and Carrier-Aligned Topology

We first ask whether dense visual-payload KV states must remain resident after the complete video prefix has participated in multimodal prefill. We compare three post-prefill conditions on the full Charades-STA cohort. Full replay retains the complete prefix cache and provides the reference

Table 1: Post-prefill temporal-grounding performance on the full Charades-STA cohort. The retained KV ratio is measured over video-associated prefix states.

Model	Full	Carrier-only	TC-KV	Carrier KV	TC-KV KV
Qwen3-VL	45.11	37.43	**42.77**	6.30%	8.14%
LLaVA-OV-1.5	21.91	23.04	**23.84**	6.12%	7.42%
InternVL3.5	24.96	15.10	**22.67**	3.42%	8.00%

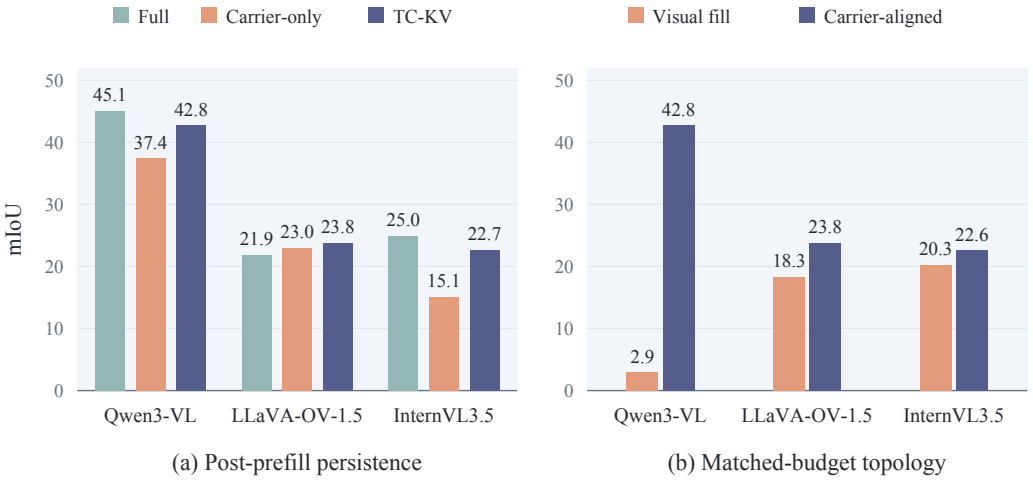


Figure 2: Post-prefill persistence and matched-budget topology comparison on the full Charades-STA cohort. (a) Temporal-grounding performance of Carrier-only and TC-KV after visual-payload KV eviction, with the full cache as a reference. (b) Visual-fill and carrier-aligned caches at the same KV budget.

performance. Carrier-only retains the static and temporal-carrier states while removing all visual-payload KV. TC-KV additionally keeps a small visual residual together with the carrier structure. All conditions use the same completed video prefill and differ only in the KV states retained for subsequent query replay.

As shown in Table 1, removing nearly all visual-payload KV does not eliminate temporal-grounding ability. Qwen3-VL retains 37.43 mIoU with only 6.30% of the video-associated KV, and TC-KV recovers 42.77 mIoU at 8.14%. LLaVA-OV-1.5 and InternVL3.5 also retain substantial capability under compact caches. For LLaVA-OneVision-1.5, Carrier-only slightly exceeds Full replay, indicating that physical compaction can also alter the subsequent cache readout dynamics; Full replay is therefore used as a reference condition rather than a strict upper bound for every compact cache. The magnitude varies across models, but the common observation is that the complete visual prefix is not required to remain resident during later temporal readout.

Carrier-only tests whether the carrier structure alone supports post-prefill temporal access, while TC-KV tests whether a small visual residual adds support. Their difference measures the contribution of limited visual support under the same completed prefill.

Matched-budget topology intervention. The previous comparison does not by itself distinguish the carrier topology from the number of retained KV entries. We therefore construct a matched-budget visual-fill control. For each model, the visual-fill cache retains the same number of retained visual KV entries as the carrier-aligned cache, while replacing the model-aligned carrier positions with visual-payload positions. The visual residual and total physical KV budget are kept unchanged.

Figure 2(a) shows that substantial temporal-grounding capability survives after most visual-payload KV is evicted. At matched budgets, panel (b) further shows that carrier-aligned retention outper-

Table 2: Matched-budget selector-family comparison on the full Charades-STA cohort. Each pair compares a flat eviction rule with its carrier-protected variant at the same physical KV budget. Values report mIoU.

Model	Selector family	Flat	Carrier-protected	Gain
Qwen3-VL	VL-Cache	2.19	**42.20**	+40.01
	SnapKV	13.51	**40.47**	+26.96
	StreamingLLM	8.23	**39.22**	+30.99
LLaVA-OV-1.5	VL-Cache	14.29	**20.70**	+6.41
	SnapKV	16.74	**21.91**	+5.17
	StreamingLLM	16.02	**22.72**	+6.70
InternVL3.5	VL-Cache	20.94	**22.56**	+1.62
	SnapKV	20.27	**22.49**	+2.22
	StreamingLLM	16.66	**19.26**	+2.60

forms visual fill by 39.88, 5.53, and 2.35 mIoU points on Qwen3-VL, LLaVA-OV-1.5, and InternVL3.5, respectively (95% CIs: [38.67, 41.12], [4.57, 6.55], and [1.63, 3.04]). Together, these results show that compact-cache utility depends on which functional roles are preserved, beyond the amount of retained KV.

Topology protection across eviction rules. To test whether carrier protection extends beyond the TC-KV residual policy, we implement VL-Cache, SnapKV, and StreamingLLM within a common physical-position interface while preserving their respective selection rules (Appendix A.3). VL-Cache and SnapKV use the downstream query to compute their retention scores when the query arrives, whereas TC-KV fixes its query-independent layout before query arrival. For each selector family, we compare a flat variant with a carrier-protected variant under the same total physical prefix-KV budget (Table 2). This design separates the effect of explicit carrier protection from the selector used to allocate the remaining payload positions. Carrier protection improves mIoU in all nine model–selector comparisons. These results support carrier topology as a structural constraint on cache eviction, while the visual-payload selector determines how the remaining budget is allocated.

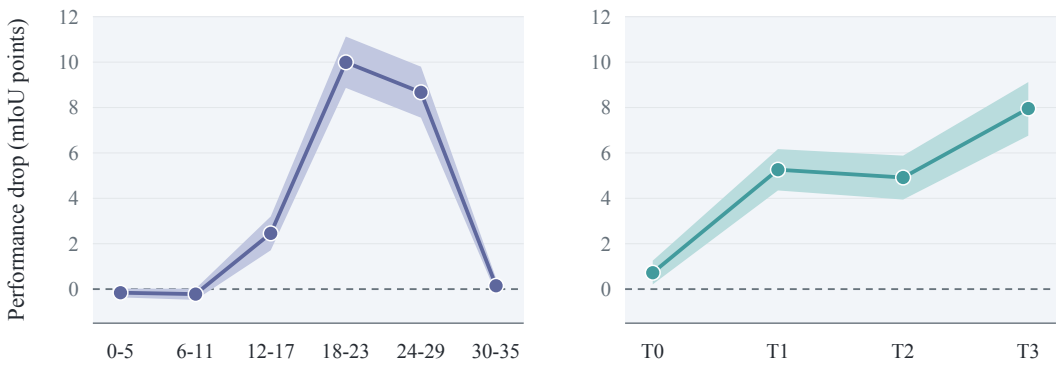
Overall, the results establish the first experimental claim of the paper: after complete video prefill, substantial temporal-grounding capability can persist in a compact cache, and preserving the model-aligned carrier topology is more important than filling the same budget with visual-payload states.

4.3 Mechanistic Analysis of Carrier Memory

The preceding experiments establish the importance of carrier-aligned topology across three video-LLM families. We use Qwen3-VL for detailed mechanistic analysis, examining whether carrier content is video-dependent, where sensitivity is concentrated across layers and temporal regions, and how visual evidence influences carrier states during prefill. We further extend the carrier K/V factorial intervention to LLaVA-OneVision-1.5 and InternVL3.5 to examine shared and model-specific contributions of keys and values.

Carrier content is video-dependent. We replace the carrier content of each recipient video with that from a different donor video while keeping recipient positions, rotary phases, static states, visual residual, and total retained KV budget unchanged. Only the carrier content is replaced; results are reported in Table 3.

The self-carrier condition obtains 42.77 mIoU and 44.41 R@0.5, whereas the donor-carrier condition obtains 19.84 mIoU and 15.51 R@0.5. The video-averaged paired self-minus-donor effect is 22.16 mIoU points, with a 95% confidence interval of [20.63, 23.71]. This effect is computed after averaging queries within each video and therefore need not equal the difference between the query-averaged arm scores. The substantial degradation under unchanged positions, visual support, and physical KV budget shows that the retained carrier contains information associated with the processed video beyond a generic temporal scaffold.



(a) Layer bands (b) Temporal regions

Figure 3: Localization of carrier-content sensitivity in Qwen3. Performance drops caused by replacing carrier content within (a) different layer bands and (b) different native temporal regions. Shaded bands indicate pointwise 95% confidence intervals.

Table 3: Cross-video carrier transplantation on the full Charades-STA cohort. The donor carrier is inserted at the recipient positions while the visual residual, rotary phase, static states, and physical KV budget are preserved. Arm scores are query-averaged; the reported paired effect is video-averaged.

Condition	mIoU ↑	R@0.5 ↑	KV ratio ↓
Official generation	45.24	47.96	100.00%
Full replay	45.11	47.82	100.00%
Self carrier	**42.77**	**44.41**	8.14%
Donor carrier at recipient positions	19.84	15.51	8.14%

Carrier sensitivity varies across depth and temporal regions. We next localize this video-dependent information by replacing carrier content only within selected layer bands or native temporal regions. For a subset S, we define the paired intervention effect as

$$\Delta_S = \mathrm{mIoU}(\mathrm{Self}) - \mathrm{mIoU}(\mathrm{Donor}_S), \tag{7}$$

where a larger positive value indicates greater sensitivity to the replaced carrier content. The complete layer- and region-level results are shown in Figure 3.

The effect is concentrated in depth ranges: replacing carrier content in layers 18–23 and 24–29 reduces mIoU by 9.99 and 8.67 points, respectively, while layers 12–17 show a smaller reduction of 2.46 points and the earliest and final bands have little effect. The temporal profile is also non-uniform, with reductions of 0.72, 5.26, 4.92, and 7.96 points for T0–T3, respectively. Replacing all carrier content produces a 21.94-point reduction. Together, these profiles show that post-prefill readout is most sensitive to carrier states in middle-to-late layers and later temporal regions, while the larger full-transplant effect indicates distributed but non-uniform carrier dependence.

Carrier information uses coupled K/V pathways. To examine how carrier content is represented, we perform a matched 2×2 factorial intervention on carrier keys and values. Each carrier key and value independently comes from either the recipient video (self) or a non-self donor. The four conditions are self/self, donor/self, self/donor, and donor/donor. The arm-level scores are reported in Table 4. Video-level factorial effects and confidence intervals are provided in Table 13.

Qwen3-VL shows the strongest key effect (+16.82), with positive value (+5.12) and interaction (+9.59) effects. LLaVA-OneVision-1.5 exhibits smaller marginal effects but a positive interaction (+3.34), while InternVL3.5 shows a smaller key effect (+1.23) and an interaction (+2.30).

The common result is that carrier keys and values cannot be treated as fully independent components. Their relative contributions vary with the model: Qwen3 shows the strongest key contribution, LLaVA-OneVision-1.5 is primarily characterized by the interaction term, and InternVL3.5

Table 4: Carrier K/V factorial results on Charades-STA. S and D denote self and donor content, respectively. The four source conditions report mIoU; Joint, K, V, and K×V report video-level paired effects. Retained KV ratios are 8.14%, 7.42%, and 8.00% for Qwen3-VL, LLaVA-OV-1.5, and InternVL3.5, respectively.

	Carrier source: K / V				Paired effect			
Model	S/S	D/S	S/D	D/D	Joint	K	V	K×V
Qwen3-VL	**42.77**	20.28	32.32	19.95	+21.94	+16.82	+5.12	+9.59
LLaVA-OV-1.5	**23.63**	21.74	22.16	23.37	+0.30	+0.28	+0.02	+3.34
InternVL3.5	**22.69**	20.49	21.34	21.20	+1.44	+1.23	+0.21	+2.30

Table 5: Visual-to-carrier restoration in Qwen3. Recovery is measured at unpatched downstream carrier positions: early-relay recovery averages layers 8–9, and immediate recovery is measured at layer 8. The wrong-source control applies a norm-matched update from another timestamp at the successor position.

Intervention	Early-relay recovery	Immediate recovery	Valid interventions
Successor-carrier restoration	**3.38%**	**3.57%**	2,272/2,274
Matched visual-state restoration	0.58%	0.46%	2,272/2,274
Norm-matched wrong-source update	−3.67%	−3.50%	2,272/2,274

exhibits a smaller but positive coupled effect. The positive interaction terms further indicate that the usefulness of one pathway depends on the state provided by the other. This motivates describing the retained carrier as a coupled KV state rather than assigning a semantic meaning to keys and values.

Visual processing writes information into successor carriers. The transplantation and factorial experiments establish the role of video-dependent carrier content. To examine how visual evidence influences carrier states during prefill, we perform a state-restoration intervention in Qwen3. This diagnostic traces the write-side pathway; TC-KV itself requires no hidden-state restoration or additional prefill computation. We perturb a frame's visual embeddings at the language-model input, then restore the clean hidden states of its successor carrier at the output of layer 7 and allow subsequent layers to recompute. We compare this intervention with restoration of an equal-sized visual span and a norm-matched update from another timestamp applied at the same successor position. Table 5 reports the results.

Across the full cohort, 2,272 of 2,274 interventions pass the validity checks. Successor-carrier restoration yields 3.38% early-relay recovery, compared with 0.58% for matched visual restoration and -3.67% for the wrong-source update. Recovery is measured at downstream carrier positions that are not directly patched. Together, these interventions support a write–persist–read account in Qwen3: visual evidence influences carrier states during prefill, and video-dependent carrier content supports subsequent temporal readout after visual-payload KV eviction. The cross-model factorial further shows that carrier keys and values interact, while their individual contributions vary across architectures.

4.4 TC-KV as Reusable Compact Memory

We evaluate TC-KV beyond Charades-STA through highlight detection on QVHighlights, then measure its persistent-memory savings and query-stage efficiency under cache reuse.

Transfer to highlight detection. We evaluate TC-KV on QVHighlights, where the model identifies salient temporal segments rather than predicting a single start–end interval. For each model, TC-KV constructs its compact cache before query arrival, while each baseline follows its selection protocol. All compact methods retain the same physical KV budget. Table 6 reports VeryGood highlight mAP, the most selective relevance threshold in the official evaluation protocol.

TC-KV retains 31.96 VeryGood mAP on Qwen3-VL, close to the 33.53 achieved by the full cache and substantially above all matched-budget alternatives. It also provides the strongest compact re-

Table 6: Highlight-detection results on QVHighlights. Values report VeryGood mAP ($\times 100$). All compact methods within each model retain the same physical KV budget. Bold marks the best compact result; Full is the uncompressed reference.

Model	Full	VL-Cache	SnapKV	StreamingLLM	TC-KV	KV
Qwen3-VL	33.53	21.67	26.05	21.01	**31.96**	8.21%
LLaVA-OV-1.5	16.72	12.42	12.21	11.81	**14.74**	7.60%
InternVL3.5	21.34	19.88	**19.98**	18.51	18.92	8.00%

Table 7: Resource measurements on a fixed 100-video Charades-STA cohort. KV denotes the physical cache-byte ratio; query peak excludes model-only allocation. Est. E2E@16 is a component-based estimate using recorded preparation costs and single-query latency (Appendix D.2).

Model	KV	Query peak	Query speedup	Est. E2E@16
Qwen3-VL	8.14%	10.76%	1.13×	0.90×
LLaVA-OV-1.5	7.42%	9.39%	1.79×	0.85×
InternVL3.5	7.99%	14.80%	1.25×	1.21×

sult on LLaVA-OneVision-1.5, reaching 14.74 mAP with 7.60% of the video-associated KV. On InternVL3.5, TC-KV remains competitive, with SnapKV obtaining a 1.06-point higher compact result. Across the three model families, the carrier-centered cache transfers from interval localization to highlight detection while retaining approximately 8% of the video-associated prefix KV. Together with the Charades-STA results, this evaluation shows that a cache formed before query arrival can support two distinct temporal readout formats without query-specific selection or video recomputation. We next measure the physical-memory and query-stage benefits of cache reuse.

Persistent-memory and query-stage efficiency. On a 100-video Charades-STA cohort, we construct TC-KV, release the source full cache, and measure physical KV size, resident allocation, query-stage peak allocation, and query latency (Table 7). TC-KV retains 7.42–8.14% of full-prefix KV bytes and reduces query-stage peak allocation above the model-only baseline to 9.39–14.80% of the full-cache reference. Physical compaction takes 3.8–6.9 ms per video, while query processing accelerates by $1.13\times$, $1.79\times$, and $1.25\times$ on Qwen3-VL, LLaVA-OV-1.5, and InternVL3.5.

We additionally estimate reuse costs by combining one complete video prefill, recorded one-time preparation costs, and repeated single-query latency. At 16 queries, the estimated speedups are $0.90\times$, $0.85\times$, and $1.21\times$, respectively. These estimates use runner-level preparation timing, whose attribution is detailed in Appendix D.2; total reuse cost therefore depends on amortizing cache preparation. Together, the Charades-STA, QVHighlights, and resource results show that carrier-centered temporal memory can be operationalized as a compact reusable cache. TC-KV is training-free and query-independent, retains approximately 8% of the video-associated prefix KV, and serves subsequent temporal queries without re-encoding the video or recomputing a query-specific cache.

5 Conclusion

We studied video-LLM KV retention as a post-prefill persistent-memory problem. Across three video-LLM families, substantial temporal-grounding capability survives after most visual-payload KV has been physically evicted. Matched-budget interventions show that this persistence depends on carrier-aligned organization rather than cache size alone, while carrier-content and state-restoration interventions reveal video-dependent state and a visual-to-carrier write path. Together, these results support a *write–persist–read* account of post-prefill temporal memory.

Based on this finding, we introduced TC-KV, a training-free and query-independent policy that preserves model-aligned temporal carriers together with a small visual residual. TC-KV retains approximately 8% of the video-associated prefix KV, supports multiple temporal readout formats, and reduces persistent memory and query-stage cost. More broadly, our results show that the states required to process a video need not be the states required to remember it.

AI USE STATEMENT

Generative AI tools assisted with literature search, experimental design, code implementation and debugging, preliminary analysis of experimental results, and drafting and editing the manuscript and figures. The authors reviewed the AI-assisted code and text, checked reported measurements against experiment outputs, and made the final decisions about the scientific claims. No synthetic datasets or mathematical proofs were generated using AI tools. The authors take full responsibility for the paper and its supplementary materials.

ETHICS STATEMENT

This work uses existing pretrained models and established public research benchmarks. We do not collect new personal data, recruit human subjects, or deploy the evaluated models in high-stakes settings. All datasets are used solely for their intended research purposes.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.1 specify the post-prefill cache decomposition, controlled interventions, and evaluation protocol used in this work. The appendix documents the model-specific token-role definitions, physical KV gathering, prompt formats, dataset processing, cohort construction, cache budgets, bootstrap procedure, and replay-integrity checks. Expanded full-cohort results and additional controls are also provided in the appendix. All main experiments use fixed cohort manifests and predefined intervention protocols. We will include the implementation, experiment configurations, cohort manifests, and result-aggregation scripts in the anonymized supplementary material.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, et al. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Yingqi Fan, Junlong Tong, Anhao Zhao, and Xiaoyu Shen. What do visual tokens really encode? uncovering sparsity and redundancy in multimodal large language models. *arXiv preprint arXiv:2603.00510*, 2026.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285. IEEE, 2017.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. Infinipot-v: Memory-constrained KV cache compression for streaming video understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=hFxOZjHyTg.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364, 2023.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024.

Congyang Ou, Ruike Song, Yang Zhou, Libo Sun, Haokui Zhang, and Zhenbo Luo. When vision becomes text: Visual token pruning via cross-modal residual guidance in vlms. *arXiv preprint arXiv:2608.10489*, 2026.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867. IEEE, 2025.

Keda Tao, Haoxuan You, Yang Sui, Can Qin, and Huan Wang. Plug-and-play 1. x-bit kv cache quantization for video large language models. *arXiv preprint arXiv:2503.16257*, 2025.

Dezhan Tu, Danylo Vashchilenko, Yuzhe Lu, and Panpan Xu. Vl-cache: Sparsity and modality-aware kv cache compression for vision-language model inference acceleration. In *International Conference on Learning Representations*, volume 2025, pp. 219–239, 2025.

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*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025a.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025b.

Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. Aim: Adaptive inference of multi-modal llms via token merging and pruning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20180–20192. IEEE, 2025.

Table 8: Evaluation cohorts used in this study. Counts refer to distinct videos and annotated queries.

Benchmark	Task	Videos	Queries
Charades-STA (Gao et al., 2017)	Temporal grounding	1,334	3,720
QVHighlights (Lei et al., 2021)	Highlight detection	1,519	1,550
ActivityNet Captions(subset) (Krishna et al., 2017)	Moment retrieval	1,752	3,000

Table 9: Operational carrier definitions used in the post-prefill interventions. The carrier set is model-specific; the table describes the structural role used for cache compaction rather than claiming that the three models expose the same token type.

Model	Serialized temporal structure	Retained carrier positions
Qwen3-VL	Text-number and temporal grouping positions associated with the serialized video units	The complete model-aligned temporal-carrier topology identified from the TextNum-based video serialization
LLaVA-OneVision-1.5	Model-aligned text-number groups associated with the serialized visual units	The complete constructed TextNum topology used by the model-aligned post-prefill intervention
InternVL3.5	Per-frame structural cells containing pre-visual temporal address positions, visual payload positions, and post-visual structural closures	Pre-visual temporal address positions. Static prompt and post-visual closure positions belong to $\mathcal{S}$ and are retained separately.

A Protocol, Serialization, and Reproducibility

A.1 Models and Evaluation Cohorts

We evaluate the 8B instruction-tuned variants of Qwen3-VL, LLaVA-OneVision-1.5, and InternVL3.5 (Bai et al., 2025; An et al., 2025; Wang et al., 2025). All model parameters remain frozen. Within each controlled comparison, we keep the video input, prompt, generation settings, and evaluation queries fixed while varying the retained cache or the specified intervention. Model-specific serialization and carrier definitions are described below.

Table 8 summarizes the evaluation cohorts. Charades-STA is the primary benchmark for temporal grounding and post-prefill cache interventions. QVHighlights evaluates transfer to query-conditioned highlight detection using the full validation annotations. For additional moment-retrieval evaluation, we randomly sample 3,000 queries from ActivityNet Captions, covering 1,752 distinct videos, and evaluate Qwen3-VL and LLaVA-OneVision-1.5 on this fixed subset. Within each benchmark, all compared cache conditions use identical video–query pairs.

Resource measurements use a separate, fixed 100-video Charades-STA cohort stratified by video duration, with one query selected per video. We estimate sequential reuse costs for $N \in \{1, 4, 8, 16\}$ by combining the measured video-prefix prefill, one-time preparation, and single-query replay costs. The timing protocol is detailed in Appendix D.

A.2 Model-Specific Serialization and Carrier Definitions

The three video LLMs serialize temporal units differently. For each model, we define the carrier set $\mathcal{C}$ from temporal-addressing positions in its native or constructed serialization. Dense visual-token positions form the payload set $\mathcal{P}$, while static prompt and boundary positions form $\mathcal{S}$ and remain retained across interventions.

Qwen3-VL. Qwen3-VL serializes the video prefix into temporally ordered visual groups together with text-number positions that provide an explicit model-aligned addressing structure. We treat these text-number positions as temporal carriers and retain all carrier groups in their original order. The visual-token positions within each group form the payload set $\mathcal{P}$, from which TC-KV retains a

small residual subset $\mathcal{R} \subseteq \mathcal{P}$. The carrier definition is obtained from the completed prefix layout and does not depend on the downstream query or its ground-truth temporal interval.

LLaVA-OneVision-1.5. For LLaVA-OneVision-1.5, we construct frame-aligned text-number groups in the multimodal prefix before prefill and designate their positions as $\mathcal{C}$. TC-KV retains this complete ordered carrier set together with a small visual residual. Visual-payload positions belong to $\mathcal{P}$, while static prompt and boundary positions belong to $\mathcal{S}$.

InternVL3.5. InternVL3.5 organizes each visual unit into a structured frame cell. We assign the pre-visual temporal address positions to the carrier set $\mathcal{C}$. Static prompt positions and post-visual closure positions belong to $\mathcal{S}$ and remain retained separately. The remaining visual-token positions form the payload set $\mathcal{P}$, from which TC-KV retains a small visual residual.

Common cache representation. After complete multimodal prefill, the model-specific role definitions are converted into absolute indices of the video prefix. For TC-KV and the carrier-based interventions, we use the partition

$$\mathcal{I}_{\mathrm{pref}} = \mathcal{S} \,\dot\cup\, \mathcal{C} \,\dot\cup\, \mathcal{P}.$$

Physical compaction gathers the retained key and value states from every transformer layer while preserving their original order and positional metadata. The same carrier layout is reused across queries associated with a video. Because the number of carrier and visual-payload positions depends on the serialization scheme, we report retention relative to the video-associated prefix states. Detailed index construction and token counts are given in Appendix A.3.

A.3 Cache Construction and Baseline Implementations

TC-KV cache construction. Following the role definitions above, Carrier-only retains $\mathcal{S} \cup \mathcal{C}$, whereas TC-KV additionally retains a visual residual $\mathcal{R} \subseteq \mathcal{P}$. For Qwen3-VL and LLaVA-OneVision-1.5, the residual consists of the first and central visual-payload positions within each serialized visual group, retaining distinct positions when the group contains at least two tokens. For InternVL3.5, the residual budget is distributed across visual groups, and positions within each group are ranked using a prefix-only successor-checkpoint score. This score measures the leave-one-out change in the local visual attention output at a subsequent checkpoint. The resulting residual is combined with the retained structural states to meet the prescribed cache budget. All three implementations determine the retained positions without using downstream queries, answers, or ground-truth temporal intervals.

Physical compaction and query reuse. The selected indices are sorted in their original sequence order and used to gather both keys and values from every transformer layer. The same position set is used across layers and attention heads; unselected entries are physically removed rather than masked within a full-sized cache. Compaction preserves the retained KV values and their original positional interpretation, including already applied rotary position encodings. Subsequent queries use the corresponding logical position metadata rather than treating the compacted prefix as a newly encoded shorter sequence. For repeated-query evaluation, each query starts from the same compacted video-prefix cache, without inheriting query or answer states from previous evaluations.

Selector-family implementations. We implement three selector families within this shared physical-cache interface. The VL-Cache adapter ranks prefix positions by query-to-prefix attention, averaged over the selected layers, heads, and query tokens. The SnapKV adapter applies one-dimensional max pooling with kernel size five and stride one to these attention scores before selection. The StreamingLLM adapter retains the static prompt states and fills the remaining capacity with the most recent eligible prefix positions. These implementations preserve the respective attention-ranking, pooled-attention, and sink-plus-recency selection mechanisms while using a common position set across layers and heads, rather than reproducing the original layer-wise or per-head cache layouts. The two attention-based adapters obtain their scores from the full prefix when the downstream query arrives. They therefore provide query-aware selection comparisons, whereas TC-KV constructs its reusable cache before query arrival.

Matched-budget carrier protection. For each selector family, we compare a flat variant with a carrier-protected variant at the same physical cache budget. Both retain the static set $\mathcal{S}$. In the flat variant, carrier and payload positions compete for the remaining slots under the selector's ranking. In the protected variant, all positions in $\mathcal{C}$ are retained first, and the same selector fills the remaining slots with payload states. Carrier protection is charged to the total budget: it reduces the number of available payload slots rather than adding extra cache capacity. For each example, the two variants retain exactly the same number of prefix positions as the corresponding TC-KV reference. This comparison tests whether explicit carrier protection improves a given selection mechanism, complementing the role-swap experiment that reallocates carrier capacity to visual-payload states.

A.4 Evaluation, Statistics, and Reproducibility Checks

Task metrics and output processing. For temporal grounding, we compute the temporal intersection-over-union (IoU) between the predicted and ground-truth intervals. We report mean IoU over evaluated queries and $R@\tau$, the percentage of queries whose predicted interval achieves an IoU of at least τ. Missing IoU values are assigned zero rather than excluded from the aggregate. For QVHighlights, generated highlight locations and saliency scores are converted to a dense score vector over the benchmark's two-second clips. We use the same frame-to-time conversion and output parser across cache conditions; scores assigned to the same clip are averaged, and unassigned clips receive zero. We report highlight-detection mAP and Hit@1 under the VeryGood relevance criterion. Within each comparison, the prompt, decoding settings, output parser, and metric implementation are held fixed.

Paired effects and confidence intervals. For the layer-band and temporal-region donor-swap experiments, we compute the IoU difference between the self-content reference and each donor condition for every query. These differences are first averaged within each video, and the reported effect is the mean over videos. A positive effect therefore indicates a performance drop caused by replacing recipient carrier content. We estimate uncertainty using a paired video-level bootstrap with 10,000 resamples, sampling videos with replacement and retaining the within-video query aggregation. The 2.5th and 97.5th percentiles of the bootstrap distribution define the 95% confidence interval. The intervals in Figure 3 are pointwise intervals for individual intervention effects. Unlike the query-averaged task scores, these effect estimates weight each video equally.

Integrity checks and retained records. Experiments use fixed evaluation manifests so that compared conditions share the same videos and queries. For matched-budget selectors, we verify equality of the retained physical prefix-token counts for every example. For carrier-content swaps, we additionally check that the donor differs from the recipient, the retained position indices and cache budget are unchanged, and the recipient's rotary positional encoding is preserved. The layer-band and temporal-region experiments include self-content and all-carrier donor endpoints to check consistency across intervention granularities. Saved experiment records include sample identifiers, generated outputs, parsed predictions, task scores, and cache-selection diagnostics, allowing aggregate results and paired comparisons to be reconstructed.

B Additional Evaluation Results

B.1 Expanded Results on Charades-STA

Table 10 expands the matched-budget selector comparison on the full Charades-STA cohort of 1,334 videos and 3,720 queries. In addition to mIoU, we report recall at three IoU thresholds and output parsing success rates. Each selector is evaluated with and without explicit carrier protection, using the implementations described in Appendix A.3. All compact conditions within each model retain the same number of physical prefix positions for each example.

Carrier protection improves mIoU in all nine model–selector comparisons. For Qwen3-VL and InternVL3.5, these gains also extend to all three recall thresholds. For LLaVA-OneVision-1.5, the two attention-based selectors improve mIoU and R@0.3 but show modest decreases in R@0.5 and R@0.7, indicating that improved average temporal overlap does not necessarily yield more predic-

Table 10: Expanded matched-budget results on Charades-STA. All values are percentages. "+ carriers" reserves the complete carrier set within the same total budget. Parse denotes output parsing success rate.

Method	R@0.3	R@0.5	R@0.7	mIoU	Parse
Qwen3-VL					
VL-Cache	0.54	0.27	0.22	2.19	99.78
+ carriers	66.21	43.04	18.41	42.20	99.95
SnapKV	19.35	8.95	3.17	13.51	99.84
+ carriers	63.55	40.86	17.37	40.47	99.95
StreamingLLM	10.97	4.14	1.05	8.23	98.04
+ carriers	61.94	38.49	16.67	39.22	99.81
TC-KV	66.05	44.35	20.99	42.77	100.00
LLaVA-OneVision-1.5					
VL-Cache	21.08	14.35	6.29	14.29	100.00
+ carriers	30.56	13.36	5.59	20.70	100.00
SnapKV	25.13	17.02	7.50	16.74	100.00
+ carriers	32.63	15.86	6.40	21.91	100.00
StreamingLLM	24.57	14.19	6.10	16.02	100.00
+ carriers	35.78	18.68	7.72	22.72	100.00
TC-KV	37.63	19.76	8.12	23.84	100.00
InternVL3.5					
VL-Cache	28.92	15.19	7.26	20.94	100.00
+ carriers	31.08	17.37	8.47	22.56	100.00
SnapKV	28.04	14.68	6.99	20.27	100.00
+ carriers	31.16	17.26	8.44	22.49	100.00
StreamingLLM	22.80	11.48	4.97	16.66	99.97
+ carriers	26.64	14.22	6.88	19.26	100.00
TC-KV	31.91	17.90	8.20	22.67	100.00

Table 11: Temporal grounding on the ActivityNet Captions subset (1,752 videos, 3,000 queries). Full denotes the uncompressed cache under the same query-replay protocol. All values are percentages.

Method	Retained KV	R@0.3	R@0.5	R@0.7	mIoU
Qwen3-VL					
Full	100.00	54.80	38.10	22.93	40.15
TC-KV	8.18	51.00	34.63	20.93	37.54
LLaVA-OneVision-1.5					
Full	100.00	31.17	17.40	10.27	23.18
TC-KV	7.40	24.60	14.93	9.13	18.77

tions above every fixed threshold. TC-KV achieves the highest mIoU among the listed compact conditions for each model.

The parsing results further distinguish localization quality from output-format validity. All conditions achieve parsing success rates above 98%. In particular, the flat VL-Cache selector on Qwen3-VL achieves 99.78% parsing success despite an mIoU of only 2.19. Its performance collapse therefore cannot be explained by output parsing failures alone.

B.2 Supplementary Results on ActivityNet Captions

We further evaluate TC-KV on a randomly sampled subset of ActivityNet Captions containing 3,000 queries from 1,752 videos. Both models are evaluated on the same fixed subset. TC-KV uses the carrier definitions and first–center visual residual described in Appendix A.3. The full-cache reference and TC-KV use the same prefix-prefill and query-replay procedure.

As shown in Table 11, TC-KV achieves 37.54 mIoU on Qwen3-VL while retaining 8.18% of the prefix KV, compared with 40.15 mIoU for the full cache. On LLaVA-OneVision-1.5, it achieves

Table 12: Expanded QVHighlights results across three relevance criteria. All values are percentages. Full uses the uncompressed cache under the same query-replay protocol.

	Fair		Good		VeryGood	
Method	mAP	Hit@1	mAP	Hit@1	mAP	Hit@1
Qwen3-VL						
Full	65.74	64.84	54.98	62.00	33.53	51.42
VL-Cache	44.51	37.23	36.42	34.84	21.67	29.03
SnapKV	53.22	49.94	43.82	47.29	26.05	38.71
StreamingLLM	44.27	36.65	36.01	34.26	21.01	27.23
TC-KV	63.73	63.23	52.85	60.39	31.96	49.10
LLaVA-OneVision-1.5						
Full	35.29	24.58	28.51	23.23	16.72	19.10
VL-Cache	27.28	21.29	21.52	19.81	12.42	15.16
SnapKV	26.83	21.10	21.21	19.87	12.21	14.77
StreamingLLM	26.13	21.48	20.66	20.06	11.81	15.35
TC-KV	31.72	23.48	25.37	21.94	14.74	17.42
InternVL3.5						
Full	44.46	37.16	35.95	34.39	21.34	28.19
VL-Cache	41.76	30.90	33.76	28.39	19.88	23.42
SnapKV	41.66	30.65	33.81	28.19	19.98	23.35
StreamingLLM	39.06	27.55	31.51	24.97	18.51	20.84
TC-KV	39.75	31.16	32.06	29.29	18.92	23.94

18.77 mIoU with 7.40% retained KV, compared with 23.18 for the full cache. These results extend post-prefill persistence to a second temporal-grounding dataset, while showing that the accuracy cost of the compact cache varies across models and datasets. In particular, the first–center residual preserves performance more closely on Qwen3-VL than on LLaVA-OneVision-1.5 in this evaluation.

B.3 Expanded Results on QVHighlights

Table 12 reports highlight-detection results under the Fair, Good, and VeryGood relevance criteria on the QVHighlights validation set, covering 1,519 videos and 1,550 queries. We evaluate both mAP, which measures clip-ranking quality, and Hit@1, which measures whether the highest-scoring clip is relevant. The compact conditions use the same per-example physical KV budget within each model. The selector baselines are the unprotected variants described in Appendix A.3.

On Qwen3-VL and LLaVA-OneVision-1.5, TC-KV outperforms all three matched-budget selectors on both metrics under every relevance criterion. Its advantage therefore extends beyond the VeryGood results reported in the main text. On InternVL3.5, the attention-based selectors achieve higher mAP, whereas TC-KV obtains the highest Hit@1 among compact conditions under all three criteria. For example, under VeryGood, TC-KV achieves 23.94 Hit@1 compared with 23.35 for SnapKV, despite its lower mAP of 18.92 versus 19.98. The expanded results thus distinguish preservation of overall clip ranking from preservation of the top-ranked highlight.

C Additional Mechanistic Analyses

C.1 Cross-Model Carrier K/V Content Interventions

Intervention design. We evaluate a 2×2 carrier-content intervention on the full Charades-STA cohort of 1,334 videos and 3,720 queries. For each recipient video, carrier keys and values are independently taken from either the recipient itself or a fixed non-self donor video. This yields four conditions: self keys with self values, donor keys with self values, self keys with donor values, and donor keys with donor values. The donor assignment does not use the downstream query, answer, or ground-truth interval. All four conditions preserve the recipient's retained positions,

Table 13: Video-averaged carrier K/V effects on Charades-STA. Entries report the effect in mIoU points followed by the paired 95% bootstrap confidence interval.

Model	Δ_{self}	Δ_K	Δ_V	$\Delta_{K,V}$
Qwen3-VL	21.94 [20.41, 23.47]	16.82 [15.58, 18.04]	5.12 [4.35, 5.88]	9.59 [8.15, 11.05]
LLaVA-OneVision-1.5	0.30 [−0.44, 1.07]	0.28 [−0.28, 0.86]	0.02 [−0.49, 0.52]	3.34 [2.33, 4.37]
InternVL3.5	1.44 [0.76, 2.10]	1.23 [0.75, 1.69]	0.21 [−0.23, 0.66]	2.30 [1.51, 3.12]

rotary positional encoding, static KV, visual residual, and physical cache budget. Only the source of carrier content changes.

Factorial contrasts. Let y_{ab} denote a video's mean query IoU, where a and b specify the sources of carrier keys and values, respectively: s denotes self and d denotes donor. We compute the self-versus-donor contrast, the marginal effects of recipient keys and values, and their interaction as

$$\begin{aligned} \Delta_{\text{self}} &= y_{ss} - y_{dd}, \\ \Delta_K &= \tfrac{1}{2}\big[(y_{ss} - y_{ds}) + (y_{sd} - y_{dd})\big], \\ \Delta_V &= \tfrac{1}{2}\big[(y_{ss} - y_{sd}) + (y_{ds} - y_{dd})\big], \\ \Delta_{K,V} &= y_{ss} - y_{ds} - y_{sd} + y_{dd}. \end{aligned} \tag{8}$$

Here, Δ_{self} is the self/self minus donor/donor contrast, while Δ_K, Δ_V, and $\Delta_{K,V}$ denote the key main effect, value main effect, and K/V interaction, respectively. These contrasts are averaged over videos. Confidence intervals use paired video-level bootstrap resampling, with 10,000 resamples for Qwen3-VL and 4,000 for LLaVA-OneVision-1.5 and InternVL3.5. Because the contrasts weight videos equally, they need not equal differences between the query-averaged arm scores reported in Table 4.

Cross-model findings. Qwen3-VL exhibits a strong recipient-content advantage: the self-versus-donor effect is 21.94 mIoU points with a 95% confidence interval of [20.41, 23.47]. Both key and value main effects are positive, with a larger effect for keys. InternVL3.5 shows a smaller but significant self-versus-donor effect of 1.44 points [0.76, 2.10], together with a positive key main effect. For LLaVA-OneVision-1.5, jointly replacing keys and values produces little change relative to self content, with an effect of 0.30 points [-0.44, 1.07], whereas replacing either component alone lowers performance. Its result therefore primarily supports compatibility between the sources of carrier keys and values. The positive interaction in all three models shows that the effect of replacing one component depends on the source of the other, rather than following a uniformly additive decomposition.

C.2 Carrier Content and Positional Alignment

Intervention design. We complement cross-video transplantation with within-video permutations that separate carrier key content, value content, and positional phase. The experiment uses Qwen3-VL on the full Charades-STA cohort of 1,334 videos and 3,720 queries. Carrier groups are cyclically permuted among groups of equal token width. All conditions retain the same physical positions, static states, visual residual, and cache budget. For this audit, the visual residual is selected once using prefix-only value norms and held fixed across interventions, rather than using the first–center residual of the main TC-KV configuration.

Cached keys already incorporate rotary position encodings. To isolate key content from positional phase, we first invert RoPE and then apply either the original or permuted phase. We compare the intact reference with four interventions: permuting pre-RoPE key content alone, value content alone, positional phase alone, or all three together. The joint intervention applies the same group permutation to all components, preserving their internal correspondence.

Results and interpretation. Permuting key or value content independently causes substantial performance drops of 7.55 and 9.80 mIoU points, respectively. Permuting positional phase alone has a

Table 14: Within-video carrier permutations on Qwen3-VL. All conditions retain 8.14% of video-associated prefix KV. Task scores are query-averaged percentages; Δ is the video-averaged mIoU drop from the intact reference, with a paired 95% confidence interval.

Condition	mIoU	R@0.5	Δ	95% CI
Intact reference	38.31	36.96	–	–
K content permuted	30.36	25.19	7.55	[6.56, 8.57]
V content permuted	28.80	22.88	9.80	[8.75, 10.86]
RoPE phase permuted	37.82	36.05	0.52	[0.12, 0.91]
K/V/phase jointly permuted	38.37	36.99	-0.15	$[-0.45, 0.11]$

Table 15: Detailed results underlying Figure 3. All conditions retain 8.14% of video-associated prefix KV. Task scores are query-averaged percentages; Δ is the video-averaged mIoU drop from the self-content reference.

Condition	Replaced subset	mIoU	R@0.5	Δ	95% CI
Self reference	None	42.77	44.35	–	–
B_0	Layers 0–5	42.87	44.68	-0.16	$[-0.37, 0.02]$
B_1	Layers 6–11	42.98	45.16	-0.22	$[-0.47, 0.02]$
B_2	Layers 12–17	40.22	40.40	2.46	[1.71, 3.20]
B_3	Layers 18–23	32.40	30.05	9.99	[8.87, 11.13]
B_4	Layers 24–29	33.82	31.83	8.67	[7.55, 9.80]
B_5	Layers 30–35	42.62	44.11	0.15	$[-0.18, 0.47]$
T_0	Earliest region	41.99	43.92	0.72	[0.21, 1.25]
T_1	Second region	37.08	37.39	5.26	[4.35, 6.17]
T_2	Third region	37.70	38.36	4.92	[3.95, 5.88]
T_3	Latest region	34.95	34.17	7.96	[6.76, 9.12]

smaller effect of 0.52 points. The joint intervention applies the same group permutation to each carrier key, value, and positional phase, thereby preserving their internal correspondence. Its near-zero performance change serves as a consistency control: the large single-component effects arise when the correspondence among K, V, and position is broken. This within-video result complements the cross-video transplantation experiment by separating state consistency from video-specific carrier content.

C.3 Layer-Wise and Temporal Localization of Carrier Sensitivity

Layer-band and temporal-region interventions. We localize carrier-content sensitivity in Qwen3-VL along transformer depth and native temporal order. Both experiments use the full Charades-STA cohort of 1,334 videos and 3,720 queries. After complete prefill, we construct the compact cache and replace carrier K/V only within the selected layer band or temporal region. The recipient's retained positions, RoPE phases, static states, visual residual, and physical KV budget remain unchanged.

For the depth analysis, the 36 transformer layers are partitioned into six consecutive bands of six layers each, covering layers 0–35. A band intervention replaces carrier content throughout the video, but only within that band. For the temporal analysis, native carrier groups are divided into four contiguous, nonempty bins in serialization order, from the earliest region T_0 to the latest region T_3. A region intervention replaces its carrier content across all layers. These bins are defined by carrier-group order, without using queries or ground-truth intervals. Both experiments use the same frozen non-self donor assignment.

Depth and temporal profiles. The largest single-band effects occur in B_3 and B_4, with drops of 9.99 and 8.67 mIoU points, respectively. B_2 has a smaller positive effect, whereas the intervals for B_0, B_1, and B_5 include zero. Along the temporal axis, T_3 has the largest effect and T_0 the smallest. Direct paired comparisons show that the T_3 effect exceeds those of T_1 and T_2 by 2.70 points [1.28, 4.10] and 3.04 points [1.66, 4.41], respectively. The T_1–T_2 difference is only 0.34 points [-0.79,

Table 16: The two Holm-significant interaction residuals in the 100-video pilot. Residuals and pointwise 95% confidence intervals are in mIoU points. Only one cell also meets the prescribed effect-size threshold.

Cell	Residual	95% CI	Holm p	$\lvert r_{ij}\rvert \geq 1.0$
$B_2 \times T_3$	-2.70	$[-4.88, -0.99]$	0.0368	Yes
$B_5 \times T_0$	-0.80	$[-1.58, -0.24]$	0.0144	No

1.50]. Replacing all carrier content yields a larger 21.94-point drop than any single-band or single-region intervention. These experiments localize sensitivity during post-prefill readout; they do not identify the layer at which carrier memory is first formed.

Temporal-region size check. The temporal partition follows carrier groups rather than enforcing equal carrier-token counts. The mean numbers of carrier tokens in T_0–T_3 are 21.65, 26.44, 26.38, and 29.50. Across videos, the within-region Pearson correlations between token count and intervention effect are weak: 0.061, 0.045, 0.039, and 0.023, respectively. These checks supplement the temporal profile, but the measured effects remain region-level sensitivities rather than estimates of information per carrier token.

C.4 Layer–Temporal Interaction Pilot

Design and interaction criterion. We conduct a joint layer–temporal pilot on 100 videos and 315 queries from Charades-STA. The experiment crosses the six layer bands with the four temporal regions defined in Appendix C.3, yielding 24 intervention cells. Each intervention replaces carrier K/V only within one layer–region intersection, while preserving recipient positions, RoPE phases, static states, visual residual, and cache budget. The corresponding layer-only and region-only effects are reused from the same pilot cohort with identical donor assignments.

A large cell-level performance drop need not indicate an interaction beyond the marginal layer and temporal effects. We therefore evaluate each cell against the prespecified multiplicatively separable reference:

$$\widehat{\Delta}_{ij} = \frac{\Delta_{B_i}\Delta_{T_j}}{\Delta_{\text{all}}}, \qquad r_{ij} = \Delta_{ij} - \widehat{\Delta}_{ij}, \tag{9}$$

where Δ_{ij} is the observed cell-level performance drop, Δ_{B_i} and Δ_{T_j} are the corresponding marginal drops, and Δ_{all} is the all-carrier replacement effect. All effects are averaged over videos. In each paired bootstrap resample, we recompute the marginal means and their ratio before calculating the residual. We apply Holm correction across the 24 tests. The prespecified criterion for extending the experiment to the full cohort requires at least two cells with Holm-adjusted $p < 0.05$ and $\lvert r_{ij}\rvert \geq 1.0$ mIoU point.

Pilot outcome. All 24 intervention cells were completed on the 100-video pilot, with no failures in the position, budget, donor, or recipient-RoPE checks. Two cells remain significant after Holm correction: $B_2 \times T_3$ has an interaction residual of -2.70 mIoU points, while $B_5 \times T_0$ has a residual of -0.80 points. Only the former reaches the prespecified 1.0-point magnitude threshold. The negative residual for $B_2 \times T_3$ indicates a smaller drop than predicted by the separable layer and temporal effects. The pilot therefore does not satisfy the criterion for a full-cohort joint scan. We retain the full-cohort layer and temporal profiles as separate one-dimensional analyses.

C.5 Visual-to-Carrier State Restoration

Write-side intervention. The preceding experiments intervene on cached carrier content after prefill. Here, we examine how visual evidence influences later carrier states during prefill. For Qwen3-VL, we replace the visual embeddings of a query-relevant frame at the language-model input to construct a corrupted video prefix. We then replay this prefix while restoring selected hidden states from the clean computation at the output of decoder layer 7, allowing subsequent layers to recompute. The primary intervention restores the timestamp carrier immediately following the corrupted frame. This is an oracle state-restoration diagnostic using clean activations, rather than a cache retention policy.

Table 17: State restoration at layer 7 on Qwen3-VL. Recovery scores are dimensionless. Early recovery averages layers 8 and 9; final recovery is measured at layer 35. N denotes contributing videos for these readouts.

Restoration condition	N	Immediate	Early	Final
Successor carrier	1334	0.03572	0.03376	0.00306
Predecessor carrier	1289	0.00000	0.00000	0.00000
Matched visual span	1334	0.00462	0.00585	0.01698
Other timestamp at its own position	1305	0.01104	0.01006	−0.00011
Norm-matched wrong update	1334	−0.03500	−0.03674	−0.00351
Wrong state at successor position	1334	−0.33500	−0.55037	−0.17265

We compare successor restoration with five controls: restoring the preceding timestamp, an equal-sized visual span at the end of the corrupted frame, a different timestamp at its own position, a norm-matched update from a different timestamp at the successor position, and a clean state from a different timestamp inserted at the successor position. These controls distinguish temporal placement, restored state volume, update magnitude, and state identity.

Recovery measurement. We measure recovery at downstream timestamp positions, excluding directly patched positions. For matched clean, corrupted, and restored hidden-state tensors, recovery is

$$\mathrm{Rec} = 1 - \frac{\mathrm{RMS}\big(H^{\mathrm{restored}} - H^{\mathrm{clean}}\big)}{\mathrm{RMS}\big(H^{\mathrm{corrupt}} - H^{\mathrm{clean}}\big)}. \tag{10}$$

Positive values indicate movement toward the clean state, while negative values indicate increased deviation. Immediate recovery is measured at layer 8. The primary early-relay score averages recovery over layers 8 and 9; final recovery is measured at layer 35. These quantities measure hidden-state recovery, not mIoU or task-accuracy gains.

Results and scope. The complete run covers 1,334 videos and records 2,274 interventions, of which 2,272 are valid. Successor restoration achieves an early recovery of 0.03376 with a video-level bootstrap 95% confidence interval of [0.03181, 0.03567]. Its paired advantage over matched visual restoration is 0.02792 [0.02598, 0.02995]. The predecessor control yields zero recovery, consistent with its position before the corrupted frame, while the wrong-update and wrong-state controls move downstream states further from the clean trajectory.

The successor advantage is concentrated in early propagation: its final-layer recovery falls to 0.00306, whereas matched visual-state restoration reaches 0.01698 at the final layer. The early successor advantage supports a transient carrier-mediated pathway through which visual perturbations influence later temporal states. Together with the post-prefill content interventions, this result connects visual-to-carrier information transfer during prefill with video-dependent carrier utility during subsequent temporal readout.

D Efficiency and Query-Reuse Analysis

D.1 Measurement Protocol

Benchmark cohort and measurement scope. We measure resource usage on a fixed, duration-stratified cohort of 100 Charades-STA videos, using the first query associated with each video. This systems benchmark is separate from the full-cohort accuracy evaluation. For each model, the uncompressed prefix cache provides the reference for memory and latency measurements. We distinguish video-prefix prefill, cache selection, physical compaction, and subsequent query processing. Because compaction follows complete multimodal prefill, the original video-prefix computation remains part of the measured cost.

Physical cache and CUDA memory. We report three complementary memory quantities: the bytes occupied by retained KV tensors, resident CUDA allocation before query processing, and peak CUDA allocation during query processing. For the compact condition, the source full cache

Table 18: Resource measurements on 100 Charades-STA videos. Memory columns are percentages of the corresponding full-cache measurements; resident and peak allocation exclude the model-only baseline. Query latency is in seconds. StreamMem denotes the implemented flat reproduction.

Model	Method	KV bytes	Resident	Query peak	Query (s)
Qwen3-VL	Full	100.00	100.00	100.00	0.439
	StreamMem	8.14	8.14	10.76	0.384
	TC-KV	8.14	8.14	10.76	0.387
LLaVA-OV-1.5	Full	100.00	100.00	100.00	0.408
	StreamMem	7.42	7.45	9.39	0.231
	TC-KV	7.42	7.45	9.39	0.228
InternVL3.5	Full	100.00	100.00	100.00	0.437
	StreamMem	7.99	7.99	14.80	0.342
	TC-KV	7.99	7.99	14.80	0.349

is released before measuring resident allocation and executing the query. Resident and query-peak measurements subtract the model-only CUDA allocation and are normalized by the corresponding full-cache measurements. The physical KV ratio is computed directly from the retained tensors. Thus, resident and peak ratios describe allocation above the model-only baseline, rather than the fraction of total GPU memory including model weights. The peak incurred during full video prefill is recorded separately from the post-compaction query peak.

Latency components. We record video-prefix prefill time, the selector time exposed by the runner, physical compaction time, and total query-processing time. Where available, query processing is further separated into query prefill and autoregressive decoding. The InternVL runner reports total query latency without this breakdown; unavailable components are left unreported rather than assigned zero. These measurements characterize the executed implementation, including its selection and cache-construction overhead.

Query-reuse cost model. We estimate sequential reuse costs for $N \in \{1, 4, 8, 16\}$ queries using the measured components for each video:

$$\begin{aligned} T_{\text{full}}(N) &= t_{\text{prefill}}^{\text{full}} + N t_{\text{query}}^{\text{full}}, \\ T_{\text{compact}}(N) &= t_{\text{prefill}}^{\text{compact}} + t_{\text{select}} + t_{\text{compact}} + N t_{\text{query}}^{\text{compact}}. \end{aligned} \tag{11}$$

Selection and compaction are charged once per video, while query cost is repeated N times. For each N, we compute the per-video speedup $T_{\text{full}}(N)/T_{\text{compact}}(N)$ and report its mean across the benchmark cohort. The resulting curves are component-based estimates using the measured single-query latency, not measurements of N distinct queries or concurrent serving throughput. They quantify how repeated access amortizes cache-construction overhead while retaining the cost of the initial full prefill.

D.2 Memory Footprint and Query-Reuse Results

Memory footprint and query latency. Table 18 reports results on the 100-video systems benchmark. All measured compact conditions release the source full cache before query execution. TC-KV retains 7.42–8.14% of the physical prefix KV bytes, and the measured resident allocation closely follows these ratios. Its query-peak allocation is 9.39–14.80% of the corresponding full-cache allocation above the model-only baseline. The larger query-peak ratios reflect allocations beyond the retained prefix KV during query processing.

TC-KV reduces mean query latency from 0.439 to 0.387 seconds on Qwen3-VL, from 0.408 to 0.228 seconds on LLaVA-OneVision-1.5, and from 0.437 to 0.349 seconds on InternVL3.5. The two compact methods have similar memory footprints at matched budgets. Their accuracy differences must therefore be assessed using the quality experiments rather than inferred from resource usage alone.

Table 19: Estimated reuse speedups under the recorded runner-level timing. Values above one indicate lower estimated total cost than the full-cache reference. These are component-based estimates, not measured multi-query throughput.

Model	Method	$N = 1$	$N = 4$	$N = 8$	$N = 16$
Qwen3-VL	StreamMem	0.327	0.547	0.703	0.854
	TC-KV	0.396	0.623	0.771	0.903
LLaVA-OV-1.5	StreamMem	0.308	0.431	0.565	0.763
	TC-KV	0.359	0.496	0.641	0.847
InternVL3.5	StreamMem	0.746	0.915	1.023	1.119
	TC-KV	1.045	1.130	1.176	1.212

Reuse estimates and timing attribution. Table 19 gives the component-based reuse estimates defined in Appendix D.1. The estimates improve with query count as one-time preparation costs are amortized. Under the recorded accounting, TC-KV reaches 0.903×, 0.847×, and 1.212× at 16 queries for Qwen3-VL, LLaVA-OneVision-1.5, and InternVL3.5, respectively.

The reuse estimates include the one-time selection and compaction costs recorded by the shared runner. For Qwen3-VL and LLaVA-OneVision-1.5, the recorded selection setup is shared by the compact arms, while the InternVL3.5 runner does not expose a separate selection time. The reuse values therefore characterize the measured runner execution. The directly measured results establish substantial post-prefill memory savings and lower query latency; method-isolated preparation timing is left for future systems benchmarking.