COVERAGE IS UNIQUELY CACHEABLE: CACHE-ONCE COVERAGE COMPRESSION FOR MULTI-TURN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are increasingly served in multi-turn conversation, where an image is encoded once and its visual key-value (KV) cache is reused across every subsequent turn. Visual token compression, however, is designed for single-pass inference. In multi-turn serving, current methods must either keep the full visual KV resident and re-select tokens for every new question, forfeiting the memory saving, or freeze a first-turn selection that is stale for every later question. We present **C^3** (Cache-Once Coverage Compression), a training-free method that addresses this challenge. Because the cached subset must serve questions that are unknown at commitment time, C^3 abandons query relevance and instead spreads the retained tokens across the image feature space. This enables C^3 to preserve a nearby token for whatever visual content a later question may target. The selection depends only on the image, so the cached subset is exactly what any later turn would select, and the whole dialogue runs on one compact visual KV. Across 8 benchmarks, 3 VLMs, and 3 KV budgets, C^3 is the only method that avoids severe accuracy collapse across all benchmarks, models, and turns, cutting the persistent visual KV by $6.7\times$, achieving the lowest cumulative time-to-first-token as dialogues lengthen, and matching or beating the strongest full-KV re-pruning baseline on most benchmarks.

1 INTRODUCTION

Vision-language models (VLMs) now power conversational assistants such as GPT-4V (OpenAI, 2023) and Gemini (Gemini Team, Google, 2023), where a typical session has a user uploading an image and asking a sequence of questions about it. Serving such dialogue is expensive, because the visual encoder turns an image into hundreds to thousands of tokens in the language context (Liu et al., 2024a; Wang et al., 2024; Li et al., 2024), and at high resolution these visual tokens dominate inference cost. Visual token compression, which prunes redundant tokens while preserving accuracy, has become the standard remedy. In the single-pass setting it works well, cutting prefill compute severalfold at little cost in accuracy (Chen et al., 2024; Zhang et al., 2025; Yang et al., 2025; Alvar et al., 2025). Yet this line of work is predominantly designed and evaluated for *single-pass* inference. The model prunes once, answers one question, and the interaction ends. For a conversation, however, this is not optimal, as the next question must be answered from whatever visual tokens survived.

Multi-turn VLM serving requires visual information to remain available across conversation turns. Accordingly, recent efficiency work retains the visual cache so that visual tokens can be accessed at each turn (Khaki et al., 2025). To be precise, the image is encoded at the first turn, its visual KV is cached, and every later turn reuses that cache, so whatever fraction of visual tokens is kept stays resident for the whole conversation. We call this fraction the *persistent* visual-KV footprint, the quantity a compression method actually controls, and at high resolution it spans thousands of tokens per image. On this cost axis, the prevailing methods are caught in a dilemma. FastV, SparseVLM, and PyramidDrop all score visual tokens by their relevance to the current question (Chen et al., 2024; Zhang et al., 2025; Xing et al., 2025), so they must either keep the full visual KV resident to re-select for every new question, or commit to a first-turn selection chosen for a question that is no longer being asked. Figure 1(a) illustrates the dilemma, and Figure 1(b) quantifies it. Frozen to its first-turn

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.

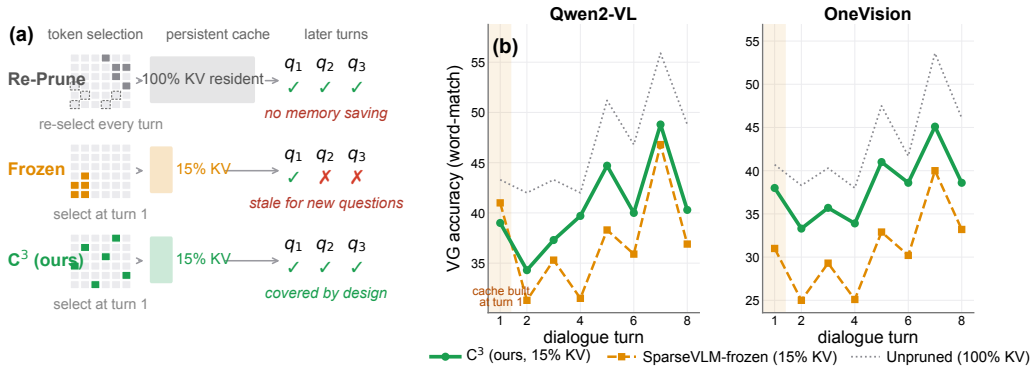


Figure 1: The cache-once dilemma and its per-turn cost. (a) Re-pruning re-selects every turn and keeps the full visual KV; freezing commits a selection tuned to the first question; C^3 commits a query-agnostic covering subset once. (b) Per-turn accuracy on 8-turn VG dialogues at B=15% (word-match); the shaded band marks turn 1, where the frozen cache is built; LLaVA-1.5 in Appendix E.

selection, the strongest query-conditioned baseline wins only the turn its cache was built for and is overtaken from the second turn onward, while at high resolution it trails at every turn of the dialogue.

In this work, we ask which selection criterion remains accurate when the choice of tokens must be made once, at the first turn, and cached for the rest of the dialogue. Whatever is cached must serve questions that do not yet exist, so no query can guide the selection, and the only reliable guide is the image itself. We therefore select a subset that covers the image by distributing the retained tokens across its feature space. This increases the likelihood that, for any visual content targeted by a later question, a representative retained token lies nearby. This coverage principle leads naturally to a worst-case objective. If a future question targets a part of the image, the quality of a cached subset is determined by the visual token it represents least: the token farthest from its nearest retained neighbor. We thus seek the subset that minimizes this maximum distance. This is the classical k-center problem, for which greedy farthest-point sampling provides a constant-factor approximation guarantee (Gonzalez, 1985).

We realize this principle in **C^3** (Cache-Once Coverage Compression). Mechanistically, C^3 applies farthest-point sampling to the input-layer visual token features at the first turn, selects a covering subset under the persistent-KV budget, and caches only the KV states of the retained tokens. Every subsequent turn reuses this compact cache unchanged. The method is training-free, incurs the selection cost only once, and avoids the cache-once dilemma because its selection is independent of the query. Across 8 benchmarks, 3 VLMs, and 3 KV budgets under an equal persistent-KV protocol, C^3 is the only method that avoids severe accuracy collapse across all benchmarks, models, and turns. It reduces the persistent visual KV by 6.7×, achieves the lowest cumulative time-to-first-token as dialogues lengthen, and matches or outperforms the strongest full-KV re-pruning baseline on most benchmarks.

2 Related Work

Visual token compression reduces the visual tokens fed to the LLM by discarding or merging redundant ones, cutting prefill compute and KV memory with little loss in accuracy (Shao et al., 2025). Existing methods fall into two families: query-conditioned ones, scoring tokens by relevance to the current question, and query-agnostic ones, reducing tokens from the visual signal alone.

Query-conditioned selection. FastV (Chen et al., 2024) ranks visual tokens by attention scores from the early LLM layers and prunes those with the lowest scores. SparseVLM (Zhang et al., 2025) rates them by cross-attention with the text, and PyramidDrop (Xing et al., 2025) prunes progressively deeper in the stack, while MustDrop (Liu et al., 2024b), Dynamic-LLaVA (Huang et al., 2025), TRIM (Song et al., 2024), VISA (Jiang et al., 2025), and AutoPrune (Wang et al., 2025a) apply the

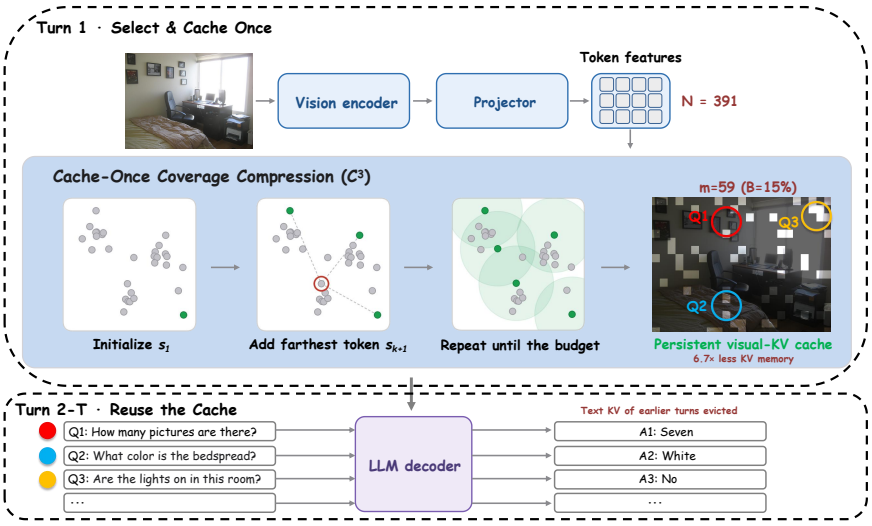


Figure 2: Overview of C^3. The image is encoded once, greedy farthest-point sampling selects the covering subset S_m ($m{=}\lceil BN \rceil$) on the input-layer features, its KV is cached at every layer, and all later turns are served from the same cache while earlier question and answer text is evicted.

same relevance signal at other stages of the pipeline, from the encoder to decoding. All of these methods assume the query is available at selection time, so once the question changes the selection must be recomputed, the root of the dilemma above.

Query-agnostic reduction. The second family selects tokens from the image alone, and its members differ in which signal they use. Saliency methods keep what the vision encoder highlights, with LLaVA-PruMerge (Shang et al., 2024), FasterVLM (Zhang et al., 2024), HiRED (Arif et al., 2025), and HiPrune (Liu et al., 2025) ranking tokens by encoder attention and VisionZip (Yang et al., 2025) keeping the dominant tokens while merging the remainder into contextual ones. Redundancy methods instead keep what is not repetitive, with ToMe (Bolya et al., 2023) and DyMU (Wang et al., 2025b) merging highly similar tokens, LLaVA-Scissor (Sun et al., 2025) merging semantically connected regions, and DivPrune (Alvar et al., 2025) pruning toward a maximally dispersed subset. These signals do not depend on the question, so a selection made once could be reused, yet prior work puts them to the same single-pass use, reducing the cost of answering one question. DivPrune in particular reaches the same farthest-point greedy that C^3 commits, but from a different premise. It maximizes the dispersion of the kept set, an objective in which the discarded tokens never appear, whereas cache-once serving asks the opposite question, how well every discarded token is represented by what remains.

Multi-turn efficiency. SparseVILA (Khaki et al., 2025) is the closest work to ours and treats the conversation as the unit of serving. It prunes lightly at prefill with a query-agnostic criterion, keeps most of the visual cache resident, and retrieves query-relevant tokens from it at every round of decoding (we adopt its text-evict serving protocol in our evaluation). In effect it avoids committing to any small subset, preserving per-question adaptivity at the price of a near-full persistent cache. In video LLMs, query-agnostic merging likewise reuses the reduced KV across questions (Hyun et al., 2025). KV-cache compression for text-only LLMs does commit, with H2O (Zhang et al., 2023) evicting entries by accumulated attention and attention sinks (Xiao et al., 2024) keeping only early anchors and a recent window, but both signals look backward and inherit the staleness of the query-conditioned family above. Existing work thus either refuses the commitment or commits on a backward-looking signal. C^3 commits at the first turn on the signal that makes commitment safe, coverage of the image, the worst-case-optimal cached selection under this constraint.

3 METHOD

Figure 2 provides an overview of C^3. We formalize cache-once serving and its persistent budget, derive the worst-case coverage objective it induces, and instantiate the solution as C^3.

3.1 PROBLEM FORMULATION

Let $V = \{v_i\}_{i=1}^{N}$, $v_i \in \mathbb{R}^D$, denote the visual tokens of an image after the vision encoder and projector. A dialogue about the image is a sequence of questions $q_1, \ldots, q_T$ that arrive one at a time, with T unknown during serving. The image is encoded once; at the first turn, a compression method commits an index subset $S \subseteq [N]$, and the model f_θ caches the key–value (KV) state of the selected tokens at every layer, written $\mathrm{KV}(S)$. Each turn t then generates its answer $a_t = f_\theta(q_t \mid \mathrm{KV}(S))$ conditioned on the cached visual KV and the question, while the text of earlier turns is evicted following streaming practice.

Methods differ in how many tokens they keep at each of the L layers, so we measure them by the layer-averaged fraction of visual KV that stays resident,

$$B = \frac{1}{L} \sum_{\ell=1}^{L} \frac{|S_\ell|}{N}, \tag{1}$$

where $S_\ell \subseteq [N]$ denotes the tokens kept at layer ℓ. We call B the *persistent visual-KV budget*. It is the quantity that scales multi-turn memory, and every comparison in this paper holds it equal. For methods with a single selection point, $S_\ell = S$ for all ℓ and $B = |S|/N$. A committed selection re-gathers any KV written before its selection point to the committed subset, so a frozen method holds $|S_\ell| = |S|$ at every layer, and the one-time compute of the selection is charged to the first turn's latency, never to B.

A selection criterion is a mapping π that returns the kept subset. We call π *cacheable* if its output depends only on the visual tokens, that is, $\pi(V, q) = \pi(V)$ for all questions q. For a cacheable criterion, the subset committed at the first turn coincides with the subset any later turn would select, so freezing it incurs no staleness, and commitment costs nothing relative to re-selecting every turn. A query-conditioned criterion instead sets $S_t = \pi(V, q_t)$, which changes across turns. It must either recompute S_t at every turn, which forces the full visual KV to stay resident, or freeze S_1, which was optimized for q_1 and is stale for every later question. We seek the criterion whose single committed subset best serves an unknown future dialogue at a given budget B.

3.2 WORST-CASE OPTIMAL CACHED SELECTION

We model the visual need of a future question as a target token $j \in [N]$ that is unknown at commitment time. Two assumptions connect this need to the cached subset. First, the evidence a question requires localizes to some token j of the image (Krishna et al., 2017; Hudson & Manning, 2019). Second, feature distance proxies relevance (Bolya et al., 2023; Alvar et al., 2025), so the quality with which a subset S serves a need at j is governed by the *served distance* $\delta(j; S) = \min_{s \in S} d(v_j, v_s)$, where d is a metric on the token feature space. Because nothing is known about which j a later turn will probe, we measure a committed subset by its worst served distance $R(S) = \max_{j \in [N]} \delta(j; S)$.

Proposition (informal). Under the two assumptions and with future needs unknown, the minimax-optimal cached subset $S \subseteq [N]$ at budget B is

$$S^\star = \operatorname*{arg\,min}_{|S|=m} R(S), \qquad m = \lceil BN \rceil, \tag{2}$$

the classical k-center problem.

Finding $S^\star$ exactly is NP-hard, but the greedy farthest-point rule, initialized at an arbitrary token,

$$s_{k+1} = \operatorname*{arg\,max}_{j \in [N]} \min_{s \in S_k} d(v_j, v_s), \qquad S_{k+1} = S_k \cup \{s_{k+1}\}, \tag{3}$$

satisfies $R(S_m) \leq 2\, R(S^\star)$ for the optimum of Eq. 2 (Gonzalez, 1985). Neither the objective nor the solver is new. What the proposition establishes is that they are induced by the setting rather

than chosen, since under query uncertainty worst-case coverage is exactly what a cached selection should optimize.

Cacheability follows from the form of the objective. $R(S)$ involves the visual tokens alone, so the coverage criterion satisfies $\pi(V, q) = \pi(V)$ and the first-turn selection is the selection of every turn.

3.3 Cache-Once Coverage Selection

C^3 instantiates the greedy rule of Eq. 3 at the LLM input. Its features are the projected encoder outputs $V = \{v_i\}_{i=1}^N$, namely the visual token representations consumed by the LLM. We first ℓ_2-normalize each feature as $\hat{v}_i = v_i/\|v_i\|$ and measure the distance between two tokens by the chord distance $d(u, v) = \|\hat{u} - \hat{v}\|$, a metric that responds to the direction of a feature rather than its magnitude. Let $\theta(u, v)$ denote the angle between u and v. Then $d(u, v)^2 = 2\,(1 - \cos\theta(u, v))$. Chord distance and cosine dissimilarity therefore induce the same ranking of token pairs and the same farthest-point selection. In implementation, we use the computationally simpler cosine dissimilarity $1 - \cos\theta(u, v)$.

The selection maintains a distance field over the tokens. Let $\delta_j^{(k)} = \delta(j; S_k)$ denote the served distance of token j under the current selection S_k. The first token anchors the selection at the boundary of the feature distribution, $s_1 = \arg\max_{j\in[N]} d(v_j, \bar{v})$ with $\bar{v}$ the mean token feature; since the bound $R(S_m) \le 2\,R(S^\star)$ holds for an arbitrary initial token (Gonzalez, 1985), this initialization costs nothing in the worst case. Each later step adds the least-covered token and refreshes the field,

$$s_{k+1} = \arg\max_{j\in[N]} \delta_j^{(k)}, \quad \delta_j^{(k+1)} = \min\{\delta_j^{(k)},\, d(v_j, v_{s_{k+1}})\}. \tag{4}$$

Farthest-point selections are nested, $S_1 \subset S_2 \subset \cdots$, so the first $m = \lceil BN \rceil$ selections are identical whether the recursion is run to m or to N. C^3 therefore stops at m. Algorithm 1 in Appendix A summarizes the procedure.

Caching follows the selection. The kept indices are gathered once, the visual KV of every layer is restricted to them,

$$\mathrm{KV}(S_m) = \{(\mathcal{K}_\ell[S_m],\, \mathcal{V}_\ell[S_m])\}_{\ell=1}^L, \tag{5}$$

where $\mathcal{K}_\ell$ and $\mathcal{V}_\ell$ are the key and value tensors of layer ℓ and $[S_m]$ keeps the rows of the selected tokens, and the remainder of the visual cache is freed. At $K{=}0$ the unselected tokens are never prefilled and Eq. 5 is realized directly, the gather freeing existing rows only when the selection follows warm-up layers. The retained tokens keep their original position indices, so rotary position embeddings are untouched. Every later turn is then served from $\mathrm{KV}(S_m)$ exactly as formulated in the problem setup, with no re-selection, no re-encoding, and no growth of the visual cache.

Each step of the recursion updates the field in $O(ND)$, so the whole selection costs $O(mND)$, paid once per image, and the terminal value $\max_j \delta_j^{(m)} = R(S_m)$ certifies the achieved coverage radius. C^3 reads nothing but the input features: it needs no attention maps, so it runs unchanged under fused attention kernels such as FlashAttention (Dao et al., 2022), involves no training or architectural change, and applies to any VLM whose visual tokens enter the LLM.

The recursion of Eq. 4 runs on the projected encoder outputs $V = V^{(0)}$, and it could instead run on the contextualized features $V^{(K)}$ after K decoder layers, at a one-time full-density pass through those layers, with the warm-up KV re-gathered to the committed subset afterward in the manner of Eq. 5, so the persistent footprint is unchanged. Attention-based scores carry little signal in the first layers, which is why such methods prune only after a few layers (Chen et al., 2024), but coverage needs no such warm-up, since the projected encoder features are already as distinct as the tokens will ever be, and further attention mixing only homogenizes them. The design-space study in Section 4 confirms this choice.

4 Experiments

4.1 Experimental Settings

We evaluate three open VLMs, Qwen2-VL-7B (Wang et al., 2024), LLaVA-1.5-7B (Liu et al., 2024a), and LLaVA-OneVision-7B (Li et al., 2024), covering the standard- and high-resolution

Table 1: Main results at B=15% persistent visual KV (word-match) on Qwen2-VL-7B and LLaVA-OneVision-7B. The KV column is the footprint held for the whole dialogue; † marks Re-Prune forms that keep the full visual KV. Shaded bold cells mark the best compression result per column (Unpruned is the reference).

Method	KV	VQAv2	POPE	GQA	TextVQA	MME	VG	ChartQA	Mean
Qwen2-VL-7B									
Unpruned	100%	75.7	87.8	63.7	78.2	83.3	46.7	68.5	72.0
FastV†	100%	43.7	66.2	43.4	35.8	71.8	23.0	14.3	42.6
SparseVLM†	100%	64.4	82.8	53.6	59.7	**79.8**	35.2	22.2	56.8
VisionZip†	100%	37.8	56.6	40.3	48.3	66.7	21.8	14.5	40.9
Frozen-attn	15%	62.2	75.8	53.3	**65.0**	78.8	34.3	20.5	55.7
C^3 (ours)	15%	**69.0**	**83.8**	**60.3**	**65.0**	79.2	**40.5**	**30.8**	**61.2**
LLaVA-OneVision-7B									
Unpruned	100%	77.6	89.0	64.0	70.7	82.5	43.2	64.0	70.1
FastV†	100%	51.8	69.5	47.0	7.2	56.2	18.9	13.5	37.7
SparseVLM†	100%	66.6	80.9	55.3	**50.7**	**80.3**	34.2	27.8	56.5
VisionZip†	100%	47.2	66.8	43.5	7.8	56.2	17.3	11.0	35.7
Frozen-attn	15%	56.9	75.8	52.3	9.8	64.2	24.3	15.7	42.7
C^3 (ours)	15%	**68.9**	**86.9**	**58.9**	50.2	78.0	**38.0**	**29.0**	**58.6**

Table 2: LLaVA-1.5-7B at B=15%, whose CLIP encoder additionally admits the pre-LLM saliency methods HiRED, SparseVILA, and the original encoder-side VisionZip (all cacheable). DocVQA is available on LLaVA-1.5 only. Shaded bold cells mark the best compression result per column (Unpruned is the reference).

Method	KV	VQAv2	POPE	GQA	TextVQA	MME	VG	ChartQA	DocVQA	Mean
Unpruned	100%	66.2	85.7	61.9	42.2	73.3	38.4	19.0	9.2	49.5
FastV†	100%	47.6	63.6	45.3	25.7	59.8	21.9	11.3	2.9	34.8
SparseVLM†	100%	60.5	78.8	55.2	36.2	70.3	32.3	13.7	5.9	44.1
VisionZip†	100%	45.6	61.6	44.4	15.3	56.8	21.4	11.0	2.7	32.4
HiRED	15%	61.2	78.6	58.0	**41.0**	**71.3**	34.5	**16.2**	**7.8**	46.1
SparseVILA	15%	61.1	76.1	56.8	38.7	68.7	32.7	12.7	6.7	44.2
VisionZip (encoder)	15%	62.6	80.2	58.1	40.5	70.0	34.8	15.0	**7.8**	46.1
Frozen-attn	15%	48.6	64.2	50.9	19.7	62.3	25.2	11.0	4.8	35.9
C^3 (ours)	15%	**64.7**	**85.0**	**60.3**	**41.0**	65.8	**36.2**	15.2	6.5	**46.8**

encoding regimes, on eight standard benchmarks: VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), and Visual Genome (Krishna et al., 2017) for general and region-level VQA, POPE (Li et al., 2023) for hallucination, MME (Fu et al., 2023) for perception and cognition, and TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), and DocVQA (Mathew et al., 2021) for text-centric understanding. Following the multi-turn serving protocol of SparseVILA (Khaki et al., 2025), each image's annotated questions form one dialogue, the image is encoded at the first turn only, and the text of earlier turns is evicted after each round, so the visual KV cache is the only state that persists across turns. Models answer with a single word or short phrase and are scored by word match, which removes answer verbosity as a confound between methods. All runs use greedy decoding in bfloat16 on four NVIDIA H100 GPUs, one per run, with up to 300 dialogues per benchmark and turns following the image's annotated questions. Scores under this protocol are not comparable to published single-pass numbers, and the dialogues carry no cross-turn coreference, both inherited from the protocol.

4.2 Main Results

We compare against both families identified in the related work. FastV (Chen et al., 2024), SparseVLM (Zhang et al., 2025), and our in-LLM adaptation of VisionZip (Yang et al., 2025) (VisionZip†) represent query-conditioned selection, and HiRED (Arif et al., 2025) together with the prefill selector of SparseVILA (Khaki et al., 2025) represent pre-LLM saliency, reported on

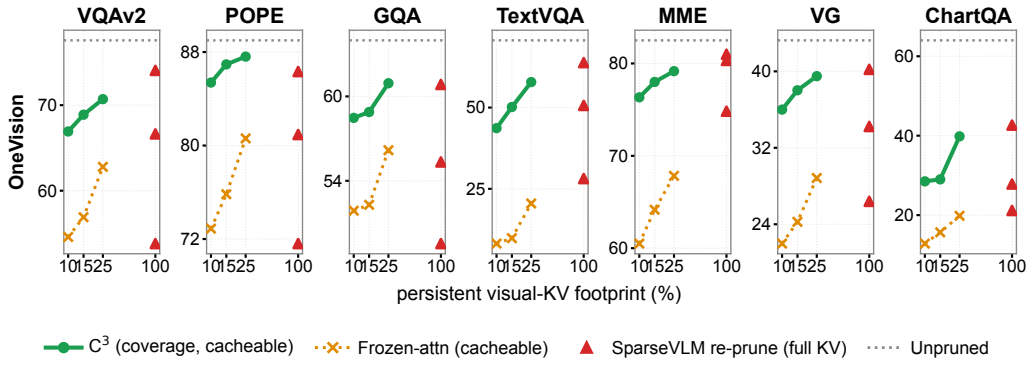


Figure 3: Accuracy versus persistent visual-KV footprint on OneVision across seven benchmarks. Cacheable arms sweep $10/15/25\%$; SparseVLM re-prune sits at a 100% footprint; the Qwen2-VL and LLaVA-1.5 sweeps are in Appendix G.

Table 3: Efficiency on LLaVA-OneVision-7B (high resolution, 3149 visual tokens, $B{=}15\%$). TTFT, decode, and throughput are medians over 40 dialogues on a dedicated GPU; persistent KV and prefill FLOPs are computed from kept-token counts. TTFT_1 includes the one-time selection of C^3, TTFT_{ss} is the steady state, and Σ_8 accumulates eight turns.

Method	Selection	KV (MB)	FLOPs (T)	TTFT_1 (ms)	TTFT_{ss} (ms)	Σ_8TTFT (ms)	Decode (ms)	Tok/s
Unpruned	—	181	48.3	200	200	1602	21.6	46.4
query-conditioned, re-select every turn, full visual KV resident								
SparseVLM	every turn	181	7.7	145	145	1163	21.8	46.0
FastV	every turn	181	7.3	105	105	837	20.1	49.7
VisionZip	every turn	181	7.3	107	107	855	19.8	50.6
query-agnostic, select once, cacheable at the budget								
Random (cached)	once	**27**	7.6	101	101	804	21.7	46.2
SparseVILA	once	**27**	7.6	183	183	1464	43.8	22.8
C^3 (ours)	once	**27**	7.6	137	**99**	831	21.7	46.2

Table 4: Design-space ablation at $B{=}15\%$ on Qwen2-VL (word-match) on VQAv2, POPE, and VG (the general, hallucination, and region families); Mean and $\bar{\Delta}$ are over these three benchmarks. The first row is C^3, and each subsequent row changes the single bolded cell.

Features	Criterion	Top-up	VQAv2	POPE	VG	Mean	$\bar{\Delta}$
$K{=}0$	max–min	–	69.0	83.8	40.5	64.4	—
K=1	max–min	–	68.2	84.8	40.0	64.3	−0.1
$K{=}0$	**grid**	–	67.2	83.3	39.9	63.5	−1.0
$K{=}0$	**norm-bias**	–	68.9	83.8	40.5	64.4	−0.0
$K{=}0$	**random**	–	66.7	84.9	38.7	63.4	−1.0
$K{=}0$	max–min	✓	67.9	82.0	38.7	62.9	−1.6

LLaVA-1.5 whose CLIP encoder they require; the adaptation details appear in Appendix B. The unpruned model is the reference upper bound. DivPrune (Alvar et al., 2025) does not appear as a separate arm because its max–min greedy coincides with the selection C^3 commits at $K{=}0$, so the C^3 rows already report its cached form. The overlap is procedural rather than conceptual: DivPrune maximizes the dispersion of the kept set for a single pass, whereas the cache-once setting induces the same greedy as the worst-case-optimal selection to commit (Section 3), with a guarantee over the discarded tokens that dispersion does not provide.

Under multi-turn serving every query-conditioned method admits two forms. The Re-Prune form re-selects tokens for each question, which requires the full visual KV to stay resident and therefore saves no memory. Each Re-Prune arm prunes at its method's native layer to the per-turn token ratio whose recurring prefill compute equals that of a cacheable arm at budget B. The Frozen form commits to the turn-1 selection and serves all later turns from it. Frozen-attn, the cacheable arm

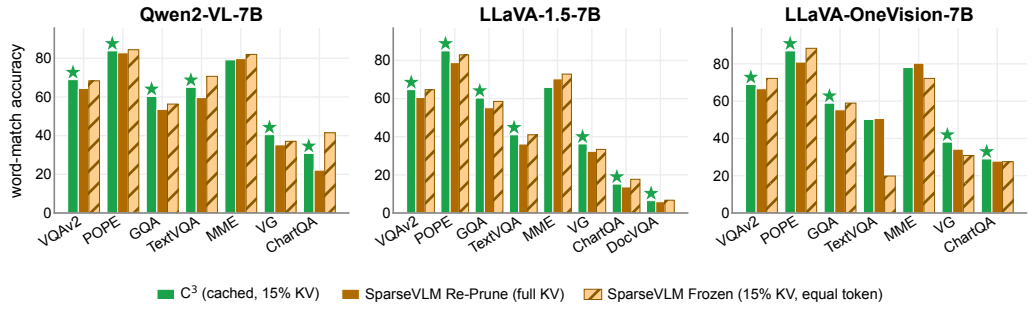


Figure 4: C^3 at 15% persistent KV against SparseVLM in both served forms (word-match; Frozen holds the same equal-token budget). Stars mark cells where C^3 beats SparseVLM even with its full KV resident (18 of 22).



Figure 5: Tokens surviving at 15% kept (real selections; dropped patches dimmed), with the dialogue's first five turns listed. Two further scenes appear in Appendix H.

we construct for the attention criterion, scores visual tokens by the attention they receive from the query at the first decoder layer, commits the top-ranked subset, and re-gathers the warm-up KV to it under the accounting of Eq. 1. Frozen-attn thus differs from C^3 only in the selection criterion, so the gap between the two rows of Table 1 isolates the criterion. All cacheable methods are compared at an equal persistent-KV budget B, with $B{=}15\%$ the main operating point and 10% and 25% in the budget sweep.

At the main operating point of $B{=}15\%$ (Tables 1 and 2), C^3 matches or beats Frozen-attn, its direct equal-budget comparison, on every benchmark and every model, with a single tie on Qwen2-VL TextVQA. The Re-Prune forms of SparseVLM, FastV, and VisionZip keep the full visual KV resident as a stronger reference, and C^3 leads all of them on the mean, trailing SparseVLM only on MME and OneVision TextVQA. In Appendix C, the equal-token frozen baselines grow strong at standard resolution, SparseVLM-frozen even edging past C^3 on the mean, but the lead is confined to the turn the cache was built for and collapses at high resolution (19.8 on OneVision TextVQA). C^3 at 15% persistent KV also matches or beats SparseVLM's full-KV Re-Prune form on 18 of 22 cells (Figure 4; per benchmark in Appendix D). On LLaVA-1.5 the pre-LLM saliency references all sit below C^3 on the mean, their edges confined to OCR and perception tasks. Figure 3 sweeps the budget over 10%, 15%, and 25%, where coverage stays closest to the unpruned reference on all three models. Aggregated over all 66 benchmark, model, and budget combinations, coverage matches or exceeds frozen attention selection in 64 of 66 cases with a mean gain of 10.35 points.

The margin over Frozen-attn grows with the visual token count, from 5.5 points on Qwen2-VL through 11.0 on LLaVA-1.5 to 15.9 on OneVision, because a larger token pool gives a query-conditioned selection more ways to commit to the wrong subset. The failure is most visible on OneVision TextVQA, where frozen attention retains 9.8 against C^3's 50.2, and even the Re-Prune forms of FastV and VisionZip collapse there (7.2 and 7.8), so at high resolution the attention criterion itself, not freezing alone, discards the text regions later questions need. On POPE, C^3 stays within a few points of the unpruned model, since an evenly spread subset retains the distributed evidence hallucination probes require. The residual cost of pruning concentrates on dense-text tasks, with ChartQA losing the most against the unpruned reference on Qwen2-VL and OneVision. The budget sweep reads the same way, as frozen attention needs a 25% footprint (mean 61.6 on Qwen2-VL) to overtake what coverage already delivers at 10% (58.1), a $2.5\times$ budget advantage.

4.3 Efficiency Analysis

Multi-turn serving makes memory the binding resource, because the visual KV persists for the whole dialogue while compute is paid per turn. C^3 commits $m{=}\lceil BN \rceil$ tokens at the first turn, so the saving holds at every turn, cutting the resident visual KV on OneVision from 181 MB to 27 MB, a $6.7\times$ reduction (Table 3).

Wall-clock gains follow the same structure. Pruning to a fraction B cuts per-turn attention roughly by B^2 and the token-wise MLP roughly by B, while the farthest-point selection costs $O(mND)$ at the first turn only, against query-conditioned scoring that recurs every turn. On high-resolution OneVision inputs the per-turn prefill of C^3 is the lowest of any method, $1.46\times$ faster than the unpruned model at the first turn including the one-time selection and $2.0\times$ in steady state, so its cumulative time-to-first-token stays below the unpruned model from the first turn and drops below every re-pruning baseline within about seven turns (Figure 7 in Appendix F). At standard resolution (roughly 400 visual tokens on Qwen2-VL) there is little prefill to save, and the win is memory rather than speed. All latency and throughput numbers are measured on the HuggingFace generation stack without serving optimizations.

Two rows of Table 3 isolate where these gains come from. Random matches every recurring efficiency number of C^3 and differs only in TTFT_1, so the profile comes from the cache-once structure and the criterion makes it affordable in accuracy (Table 4). SparseVILA holds the same 27 MB cache but re-reads it with a per-question retrieval during decoding, doubling its decode latency (43.8 vs 21.7 ms), leaving C^3 the fastest cacheable method.

4.4 Ablation and Qualitative Analysis

We ablate C^3 along the three axes of Table 4, and no variant beats the plain form. Moving the features one decoder layer deeper shifts the mean by less than half a point in either direction (-0.1 on the three benchmarks, $+0.5$ on the seven-benchmark mean), so contextualization adds a full-density pass without buying accuracy, and $K{=}0$ stays the default. On the criterion axis, the spatial grid and the high-norm bias each buy one isolated win (ChartQA and MME respectively) at a loss elsewhere. Replacing the criterion with a uniform random subset, cached in the same way (three seeds), stays within 2.6 points of C^3 on the seven-benchmark mean and above every frozen attention arm, consistent with benchmark findings (Peng et al., 2025), confirming that a cached query-agnostic subset is the right object to serve from, while the structured cover earns its keep where evidence is feature-compact, gaining 13.3 points on TextVQA with POPE marginally favoring random. On the adaptivity axis, topping up the cached base with query-scored tokens helps only on TextVQA, and at a fixed budget where every added token displaces a coverage token the exchange nets out negative. An image-adaptive blend of coverage and saliency is not general either, losing 0.8 in a LLaVA-1.5 screening. Across the search, saliency wins OCR and perception tasks while coverage wins general and reasoning VQA; the two criteria are complementary and composable while remaining cacheable.

Figure 5 overlays the tokens that survive pruning in a real dialogue. Coverage spreads its selections across the whole image, while FastV and VisionZip concentrate on the most salient region and discard the rest. The street dialogue makes this concrete. Its questions name a different object almost every turn (the clock, the two men, a shirt, the white work truck), and the coverage selection keeps tokens on each, whereas the attention selections spend the same budget on the sunlit wall and upper tree line, dropping the work truck almost entirely. A selection that looks reasonable for the first question already misses the evidence for the second, the stale-cache failure of Figure 1 at the token level.

5 Conclusion

We presented C^3, training-free visual token compression for multi-turn VLM serving, where a token subset is committed once and cached for the dialogue. C^3 therefore selects the covering subset, the worst-case-optimal choice under unknown questions, by farthest-point sampling. The selection never reads the query, so first-turn commitment costs nothing over re-selection, and a systematic search of cacheable designs found none better. C^3 avoids severe accuracy collapse across benchmarks, models, and turns, cuts persistent visual KV by $6.7\times$, and serves dialogues at the lowest cumulative time-to-first-token. Composing coverage with saliency and handling cross-turn coreference remain future work.

REPRODUCIBILITY STATEMENT

The serving protocol, the persistent-budget accounting, and the evaluation metric are specified in Sections 3.1 and 4; all benchmarks, model adaptations, and exclusions are listed in Appendix B, and the selection procedure is fully specified by Algorithm 1 (Appendix A). Source code implementing C^3, every baseline arm, the multi-turn serving harness, the word-match scorer, and the efficiency measurements, together with the exact per-arm configurations behind Tables 1–3, is included in the supplementary materials.

AI USE STATEMENT

In this work, we used generative AI tools to assist with text editing and polishing, LaTeX formatting, and the implementation of analysis and plotting scripts. We have not used generative AI tools to produce research ideas, experimental results, or citations: all reported numbers come from experiments run by the authors, and all references were manually verified against their published sources. All AI-assisted text and code were reviewed by the authors, and the code was tested before use. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S. Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1773–1781, 2025.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations (ICLR)*, 2023.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Wenxuan Huang, Zijie Zhai, Yunhang Shen, et al. Dynamic-LLaVA: Efficient multimodal large language models via dynamic vision-language context sparsification. In *International Conference on Learning Representations (ICLR)*, 2025.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Jeongseok Hyun, Sukjun Hwang, Su Ho Han, Taeoh Kim, Inwoong Lee, Dongyoon Wee, Joon-Young Lee, Seon Joo Kim, and Minho Shim. Multi-granular spatio-temporal token merging for training-free acceleration of video llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Pengfei Jiang, Hanjun Li, Linglan Zhao, Fei Chao, Ke Yan, Shouhong Ding, and Rongrong Ji. Visa: Group-wise visual token selection and aggregation via graph summarization for efficient mllms inference. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025.

Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N. Plataniotis, Yao Lu, Song Han, and Zhijian Liu. Sparsevila: Decoupling visual sparsity for efficient vlm inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23784–23794, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Jizhihui Liu, Feiyi Du, Guangdao Zhu, Niu Lian, Jun Li, Bin Chen, Weili Guan, and Yaowei Wang. Hiprune: Hierarchical attention for efficient token pruning in vision-language models. *arXiv preprint arXiv:2508.00553*, 2025.

Ting Liu, Liangtao Shi, Richang Hong, et al. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024b.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics (ACL)*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

OpenAI. GPT-4V(ision) system card. `https://openai.com/index/gpt-4v-system-card/`, 2023.

Tianfan Peng, Yuntao Du, Pengzhou Ji, et al. Can visual input be compressed? a visual token compression benchmark for large multimodal models. *arXiv preprint arXiv:2511.02650*, 2025.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

Kele Shao, Keda Tao, Kejia Zhang, et al. A survey of token compression for efficient multimodal large language models. *arXiv preprint arXiv:2507.20198*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Dingjie Song, Wenjun Wang, Shunian Chen, et al. Less is more: A simple yet effective token reduction method for efficient multi-modal LLMs. *arXiv preprint arXiv:2409.10994*, 2024.

Boyuan Sun, Jiaxing Zhao, Xihan Wei, and Qibin Hou. Llava-scissor: Token compression with semantic connected components for video llms. *arXiv preprint arXiv:2506.21862*, 2025.

Hanshi Wang, Yuhao Xu, Zekun Xu, Jin Gao, Yufan Liu, Weiming Hu, Ke Wang, and Zhipeng Zhang. Autoprune: Each complexity deserves a pruning policy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Zhenhailong Wang, Senthil Purushwalkam, Caiming Xiong, Silvio Savarese, Heng Ji, and Ran Xu. Dymu: Dynamic merging and virtual unmerging for efficient variable-length vlms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Qizhe Zhang, Aosong Cheng, Ming Lu, et al. Beyond text-visual attention: Exploiting visual cues for effective token pruning in VLMs. *arXiv preprint arXiv:2412.01818*, 2024.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A SELECTION ALGORITHM

Algorithm 1 states the first-turn selection of Section 3.3 in full. The recursion maintains one distance field over the visual tokens, adds the least-covered token at each step, and refreshes the field in $O(ND)$, so the $m = \lceil BN \rceil$ selections cost $O(mND)$ in total and run once per image.

Algorithm 1 C^3 selection at the first turn

Require: visual tokens $V = \{v_i\}_{i=1}^{N}$, budget B
Ensure: cached core subset S_m, $m = \lceil BN \rceil$
$\bar{v} \leftarrow \frac{1}{N}\sum_i v_i$; $s_1 \leftarrow \arg\max_j d(v_j, \bar{v})$; $S_1 \leftarrow \{s_1\}$
$\delta_j \leftarrow d(v_j, v_{s_1})$ **for all** $j \in [N]$
for $k = 1$ **to** $m - 1$ **do**
  $s_{k+1} \leftarrow \arg\max_j \delta_j$; $S_{k+1} \leftarrow S_k \cup \{s_{k+1}\}$
  $\delta_j \leftarrow \min\big(\delta_j,\ d(v_j, v_{s_{k+1}})\big)$ **for all** $j \in [N]$
end for
cache $\mathrm{KV}(S_m)$ at every layer, discard the rest
return S_m

B BASELINE ADAPTATIONS AND EXCLUSIONS

The pre-LLM saliency methods rank tokens by the CLS attention of the vision encoder, which only the CLIP encoder of LLaVA-1.5 provides, so HiRED and the prefill selector of SparseVILA are reported on that model. VisionZip originally selects dominant tokens by the same encoder signal, which the SigLIP encoders of Qwen2-VL and OneVision lack, so the main tables evaluate a uniform in-LLM adaptation that applies its dominant-token criterion at an early decoder layer (VisionZip†), and LLaVA-1.5 additionally reports the original encoder-side form, which is text-agnostic and therefore cacheable. PyramidDrop (Xing et al., 2025) is not listed separately in the main tables, because its progressive schedule scores by the same attention signal as FastV, and its turn-1 subset re-gathered to the budget is another frozen attention arm. For the same reason HiRED does not appear in the efficiency table, and the prefill selector of SparseVILA runs query-agnostically on the CLS-free SigLIP encoder of OneVision.

C RE-PRUNE VERSUS FROZEN FORMS AT AN EQUAL BUDGET

The dilemma described in the introduction can be measured directly, by serving each method in both of its forms. Each frozen arm commits its turn-1 selection at its method's native layer and re-gathers the warm-up KV to it, holding $\lceil BN \rceil$ tokens at every layer, exactly as Frozen-attn does. Freezing at the native deeper layer also answers a concern about Frozen-attn, whose layer-1 scores could understate the attention criterion, since attention sharpens only a few layers in; the deeper FastV scores are indeed stronger, yet they still trail C^3 on every model mean and inherit the same high-resolution collapse. Table 5 lists, for every baseline and model, the Re-Prune row above its equal-token Frozen row, with Δ the change in the row mean that freezing brings. At an equal persistent budget the frozen forms are strong on the standard-resolution models, gaining 3 to 16 points over their compute-matched Re-Prune forms, and SparseVLM-frozen edges past C^3 on the Qwen2-VL and LLaVA-1.5 means. Two structural weaknesses accompany those means. First, the advantage is front-loaded. At turn 1, the question the cache was optimized for, SparseVLM-frozen leads C^3 by 6.6 points on Qwen2-VL, and from turn 2 onward the lead is gone (-0.2), the crossover visible in Figure 1(b). Second, the commitment turns catastrophic at high resolution. On OneVision the same arm posts the only negative freeze delta (-3.8) and falls to 19.8 on TextVQA against 50.2 for C^3, less than half of its own Re-Prune form. A query-conditioned commitment also requires the first question before the cache can be built, whereas the coverage selection depends only on the image and can be cached before any question arrives. C^3 itself has no gap by construction, since its selection never reads the query, and re-running the selection at every turn shifts the mean by less than 0.1 points.

Table 5: Re-Prune versus equal-token Frozen forms of each query-conditioned baseline at $B{=}15\%$ (word-match). The Frozen form commits its turn-1 selection and re-gathers the warm-up KV to $\lceil BN \rceil$ tokens at every layer; Δ is Frozen minus Re-Prune on the row mean.

Method	Form	VQAv2	POPE	GQA	TextVQA	MME	VG	ChartQA	DocVQA	Δ
Qwen2-VL-7B										
FastV†	Re-Prune	43.7	66.2	43.4	35.8	71.8	23.0	14.3	–	
	Frozen	66.2	80.8	53.1	71.5	80.3	35.9	24.3	–	+16.3
SparseVLM†	Re-Prune	64.4	82.8	53.6	59.7	79.8	35.2	22.2	–	
	Frozen	68.3	84.4	56.3	70.7	82.0	37.1	41.5	–	+6.1
VisionZip†	Re-Prune	37.8	56.6	40.3	48.3	66.7	21.8	14.5	–	
	Frozen	62.2	78.4	51.2	71.5	78.0	33.9	25.3	–	+16.4
LLaVA-1.5-7B										
FastV†	Re-Prune	47.6	63.6	45.3	25.7	59.8	21.9	11.3	2.9	
	Frozen	58.3	76.2	55.0	36.5	69.8	30.7	12.8	5.6	+8.4
SparseVLM†	Re-Prune	60.5	78.8	55.2	36.2	70.3	32.3	13.7	5.9	
	Frozen	64.6	82.9	58.5	41.0	72.8	33.3	17.7	6.7	+3.1
VisionZip†	Re-Prune	45.6	61.6	44.4	15.3	56.8	21.3	11.0	2.7	
	Frozen	59.6	77.2	56.2	36.2	70.3	30.9	15.2	6.9	+11.7
LLaVA-OneVision-7B										
FastV†	Re-Prune	51.8	69.5	47.0	7.2	56.2	18.9	13.5	–	
	Frozen	66.9	80.3	55.8	15.0	68.3	28.0	20.0	–	+10.1
SparseVLM†	Re-Prune	66.6	80.9	55.3	50.7	80.3	34.2	27.8	–	
	Frozen	72.2	88.3	58.9	19.8	72.2	30.8	27.5	–	−3.8
VisionZip†	Re-Prune	47.2	66.8	43.5	7.8	56.2	17.3	11.0	–	
	Frozen	67.0	77.9	54.1	14.6	67.8	25.2	18.2	–	+10.7

D Cost of commitment per benchmark

Figure 4 breaks the comparison with SparseVLM, the strongest query-conditioned baseline, down to individual benchmarks. Each panel is one model, each group of bars one benchmark, and the three bars are C^3 at 15% persistent KV, SparseVLM Re-Prune with the full KV resident, and SparseVLM Frozen at the same equal-token budget as C^3. Even against the Re-Prune form, C^3 matches or beats it on 18 of 22 benchmark and model cells, the starred cells. The four exceptions are MME on all three models and TextVQA on OneVision, OCR and perception tasks where attention toward salient regions carries genuine signal. Holding less than a sixth of the memory, coverage concedes only these task-specific cells and wins everywhere else, and the frozen bar collapsing on OneVision TextVQA repeats the high-resolution failure of Table 5 in per-benchmark form.

E Per-turn decay on LLaVA-1.5

Figure 6 adds the LLaVA-1.5 panel to the per-turn view of Figure 1(b), plotting VG accuracy against the dialogue turn with the shaded band marking turn 1, where the frozen cache is built. The pattern matches the standard-resolution reading of Table 5, as SparseVLM-frozen leads C^3 only inside the band, is overtaken from turn 2 onward, and Frozen-attn trails at every turn, while the unpruned reference shows that the per-turn variation is question difficulty shared by all methods.

F Cumulative latency over a dialogue

Figure 7 completes the latency picture of the efficiency analysis. Panel (a) accumulates the time-to-first-token over a ten-turn dialogue on OneVision at high resolution, and panel (b) shows the per-turn prefill cost behind those slopes. The cumulative curve of C^3 stays below the unpruned model from the first turn, its one-time selection included, and drops below every re-pruning baseline by about turn seven, while its steady-state prefill is the lowest of any method.

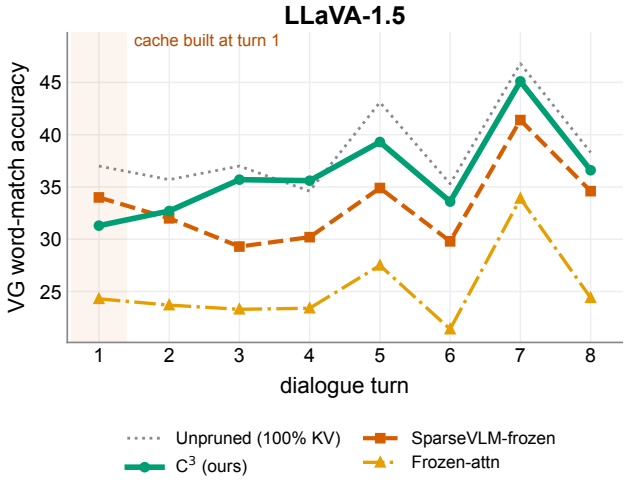


Figure 6: Per-turn accuracy on 8-turn VG dialogues on LLaVA-1.5 (B=15%, word-match), complementing the Qwen2-VL and OneVision panels of Figure 1(b).

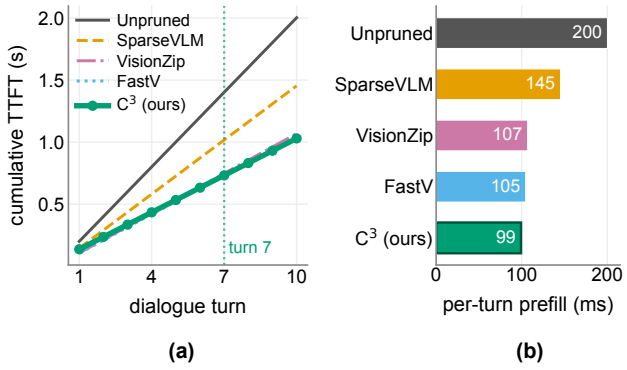


Figure 7: Latency on OneVision-7B at high resolution (measured, median). (a) Cumulative time-to-first-token, with the one-time selection of C^3 included in turn 1. (b) Per-turn prefill cost.

G Budget sweeps on Qwen2-VL and LLaVA-1.5

Figure 8 completes the budget sweep of Figure 3 on the two standard-resolution models. Lines sweep the cacheable arms over 10, 15, and 25% persistent KV, and triangles place SparseVLM re-pruning at its 100% footprint. The Qwen2-VL row carries the budget-advantage reading of the main text, where frozen attention needs a 25% footprint to overtake what coverage delivers at 10%. The LLaVA-1.5 row includes HiRED, the saliency reference its CLIP encoder admits, and coverage leads the general, hallucination, and region benchmarks at every budget while HiRED leads the OCR and perception columns, the same task-axis complementarity discussed in the ablation.

H Token masks for two more scenes

Figure 9 extends the token-mask visualization of Figure 5 to an indoor VQAv2 scene and a GQA snowboarding scene. Coverage spreads its budget across each image, so the chair, the beds, and the wall pictures each retain tokens, while the attention selections leave whole regions uncovered.

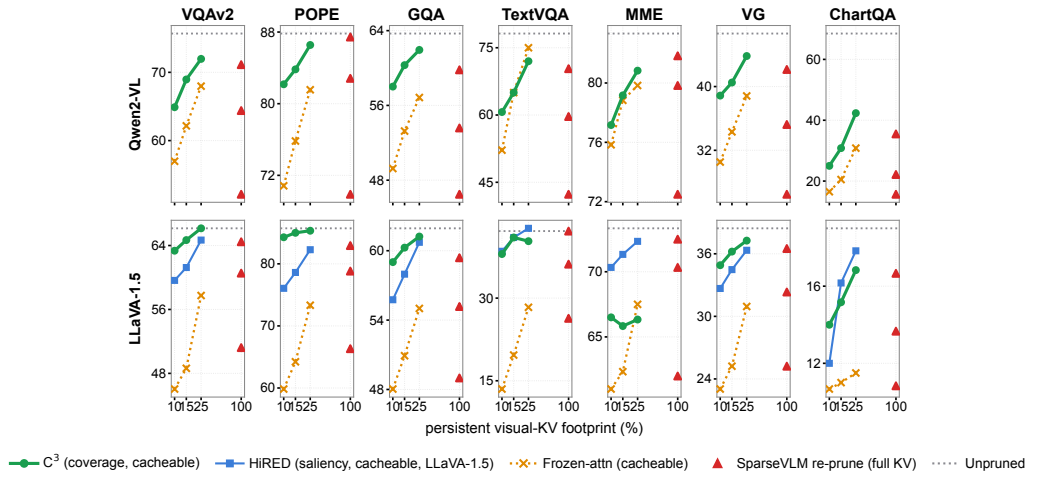


Figure 8: The Qwen2-VL and LLaVA-1.5 budget sweeps completing Figure 3; the LLaVA-1.5 row includes HiRED (blue squares), the cacheable saliency reference available on its CLIP encoder.

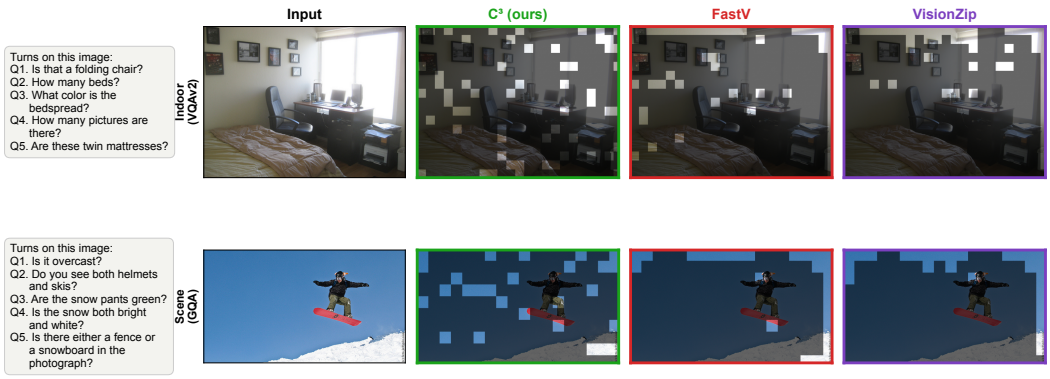


Figure 9: Two further scenes of the token-mask visualization of Figure 5 (15% kept; dropped patches dimmed).