QUERY-AGNOSTIC VISUAL KV COMPRESSION AT MATCHED MEMORY BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prefix caching allows vision-language models (VLMs) to reuse the key-value (KV) cache of an image across questions, which requires compression to be completed before the question is known. Most existing visual KV compression methods are query-aware and therefore cannot be applied in this scenario; for the query-agnostic setting, there is a lack of a systematic and fair comparison under the same memory allocation. In this paper, we investigate query-agnostic visual KV compression under matched memory budgets, across two VLM families and four visual question answering benchmarks. On our fixed paired evaluation manifests, retaining deterministically spaced evicted tokens has higher mean scores than merged tokens in all eight model–task comparisons and matches or exceeds tiled moment summaries in all comparisons. Besides, pre-registered falsification tests and an exploratory mixture study do not establish a summary advantage. Meanwhile, we provide an attention-mass bound on the maximum contribution of an eviction summary and analyze three mechanisms that limit summary methods: dilution of individual value outliers, instability of the dispersion correction, and per-block RoPE interactions in merged keys. Finally, quantization behavior depends on the grouping axis and on metadata cost. Together, these controlled comparisons clarify how byte allocation, retained-token fidelity, and task structure affect query-agnostic visual KV compression.

1 INTRODUCTION

Vision-language serving increasingly uses a shared-prefix pattern. An image (or video frame, or document page) is seen first and prefilled once, and its KV cache is then reused across the many questions that users ask about it (Kwon et al., 2023a;b). This pattern makes compression *between* questions useful, but no question is available at that point. Therefore, the eviction decisions, budget splits, and summaries of evicted content must be *query-agnostic*. Existing works provide little guidance for this problem setting because most visual KV compression methods are query-aware (Tu et al., 2025; Wan et al., 2024) and are evaluated under different regimes. Under a one-image-many-questions amortization, time-to-first-token (TTFT) is reduced by 6.5–8.3× with a compressed prefix, and peak KV memory is reduced by 2.75–3.4× relative to per-question full prefill (as measured in this paper). Table 2 reports the corresponding latency and memory measurements.

A natural use of the bytes freed by eviction is to summarize the evicted content. Tile means, variances, and merged centroids can represent dozens of evicted tokens in a few hundred bytes, motivating work on merging and moment summaries (Bolya et al., 2023; Zhang et al., 2024; Wan et al., 2025; Jana et al., 2026; Zhan et al., 2026). In this paper, we evaluate spatially tiled moment summaries over the visual grid, read out through the cache's exact attention mixture and byte-matched against every alternative under a single memory-accounting basis. Proposition 1 (Section 4) bounds the value of an eviction summary by the attention mass on evicted tokens. Meanwhile, query-agnostic selection cannot make that mass small in our measurements. It raises that mass 3.1× above the query-aware level (median 7.4% vs. 2.3%). This leaves damage that a summary could in principle repair, but does not establish that a compact summary can do so. We test summaries in three pre-registered falsification rounds and report all failures: **Round 1:** Global moment summaries over the evicted set fail in 4/4 cells. **Round 2:** On a byte-matched set of tiled-moment configurations and sampled tokens, moment summaries improve no cell; both the prespecified superiority and mono-

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.

tonicity criteria fail. **Round 3:** The tested mixture of sampling and summaries does not outperform token retention alone in any of the four cells.

The continuum comparison agrees with these bf16-core rounds. Sampled tokens have higher mean scores than merged tokens in 8/8 cells and match or exceed per-tile moment summaries in 8/8 (one exact tie; only the Qwen ChartQA interval clears zero) and mixtures in 4/4. The analysis examines whether value-outlier dilution, dispersion-correction instability, and per-block RoPE interactions contribute to the observed lack of a summary advantage. In summary, we have made the following contributions:

1. **A controlled map of query-agnostic visual KV compression.** We compare quantization, pruning, retained tokens, moment summaries, and merged tokens under a shared per-sample byte ledger across two vision-language model families and four visual question-answering benchmarks. The resulting map shows that the appropriate method depends on compression regime, task structure, and the grouping used by quantization.
2. **Matched-memory evidence for retaining tokens rather than summarizing them.** Across the controlled comparisons, deterministically spaced retained tokens are the more reliable use of evicted-cache bytes than merged tokens or tiled moment summaries. This conclusion is scoped to bf16 retained cores; the corresponding quantized-core comparison is reported separately rather than folded into the central verdict.
3. **A bound and mechanism analysis for eviction summaries.** We bound the potential value of a summary by the attention mass on evicted content, and connect observed failures to dilution of individual value outliers, instability of the dispersion correction, and per-block RoPE interactions. These analyses identify conditions under which a compact summary may be useful and conditions under which it cannot recover lost token-level information.

2 Related Work

Visual token pruning. FastV (Chen et al., 2024) observes visual-token redundancy in shallow layers, and later work such as SparseVLM (Zhang et al., 2025) improves the importance signal. These methods use attention at decode time, making their decisions question-conditioned. We adapt the idea to a query-agnostic form.

KV eviction in text LLMs. H2O (Zhang et al., 2023) evicts low-contribution entries during decoding using accumulated attention. SnapKV (Li et al., 2024) and PyramidKV (Cai et al., 2024) compress the prompt cache during prefill using observation-window attention. Their decision signals depend on the question text or on content already generated, so they are unavailable in the shared-prefix setting, where compression precedes the question.

Query-aware caching for VLMs. VL-Cache (Tu et al., 2025) and Look-M (Wan et al., 2024) compress the cache by exploiting question and modality sparsity, and both require access to the question. PrefixKV (Wang et al., 2025) adapts retention per layer for the visual prefix, but its importance signal comes from prefill attention over a prompt that already contains the question. VisCache (Wang et al., 2026) likewise prunes conditioned on the prompt. We instead ask what the evicted bytes should hold when the question is unavailable.

Query-agnostic compression. StreamMem (Yang et al., 2025) introduces a query-agnostic memory but targets streaming video understanding. Kamera (Ma et al., 2026) repairs KV reuse for multimodal agents with a training-free low-rank conditional patch. KVzip (Kim et al., 2025) is closest to our problem setting. It reconstructs the context from synthetic questions to score eviction decisions, establishing the value of query-agnostic compression for text. KVzip addresses which tokens to evict, whereas we address what the evicted bytes should hold for visual KV.

Merging and moment summaries. ToMe (Bolya et al., 2023) and CaM (Zhang et al., 2024) reduce token count by merging. D2O (Wan et al., 2025) retains outlier tokens discriminatively after eviction, but its retention decision remains query-conditioned. MomentKV (Jana et al., 2026) and ResKV (Zhan et al., 2026) compress KV with moment statistics, but for text only. We compare these

approaches in the query-agnostic visual setting and provide theory linking their failures to structural limitations.

KV quantization. KIVI (Liu et al., 2024b) and KVQuant (Hooper et al., 2024) establish KV quantization baselines, while QuaRot (Ashkboos et al., 2024) and related work spread outliers with rotations. Our Hadamard rotation configuration faithfully implements that method, so its collapse on Qwen2.5-VL measures the sensitivity of visual KV to error *direction* rather than a weakness of the implementation. The KIVI-faithful per-channel-K / per-token-V grouping configuration measures the cost of quantizing a static visual prefix, so our quantization conclusions depend jointly on the grouping axis and the memory-accounting basis.

Contribution beyond individual components. Quantization, pruning, merging, and sampling are standard components. Our contribution is to identify which to use at each compression ratio in the query-agnostic regime and explain why. The best mixture of sampling and summaries does not improve on sampling in 4/4 cells at matched bytes (with bf16 retained cores; Section 6), providing no evidence of an advantage from this combination in our setting.

3 SETTINGS

In this paper, we investigate the query-agnostic compression scheme of visual prefixes, and reveal the findings that are most relevant to the stated claims.

Segmented prefill. In order to compress the visual key-value state before any question is known, we cut the fully formatted prompt at the end of the image prefix. The image prefix is prefilled first and its visual KV is compressed. The question segment is then prefilled and the answer decoded on the compressed cache, so the first answer token uses compressed state. Besides, we verify the agreement between the segmented and unsegmented inference before any experiments (Section 6.1). The full cache follows the identical segmented path and serves as the upper-bound reference.

Matched memory budget. Every comparison uses a per-sample memory accounting. We therefore require the comparator's persistent bytes to be no smaller than those of the method under test (a single-sided assertion). The absolute gap must not exceed 512 bytes per (layer, KV-head), equivalent to one bf16 token of K and V, and the per-cell mean relative gap on the persistent memory accounting must not exceed 1%. In our experiments, only the disclosure-only rows (the question-conditioned reference) are exempt. The relative ceiling holds at the retention ratio $\rho = 0.2$ (the fraction of visual tokens kept) but is exceeded at the deeper $\rho = 0.05$ budget, where ten of the twelve cell–configuration combinations land between 1.04% and 1.31%. The compression ratio is full visual KV bytes divided by persistent visual KV bytes after compression, computed per sample from the memory accounting and reported as the per-cell median.

Models and benchmarks. We use LLaVA-1.5-7B (Liu et al., 2024a) (multi-head attention, 1D RoPE, 576 visual tokens) and Qwen2.5-VL-7B (Bai et al., 2025) (grouped-query attention, mRoPE, dynamic resolution capped at 602,112 pixels) as our VLMs. The primary comparisons use fixed source-order paired manifests of 150 examples per model–task cell with greedy decoding; they provide controlled within-cell comparisons rather than a claim of complete report-split evaluation. For VRSBench, the reported comparisons use ANLS; the GPT fallback of its public evaluator was not executed, so these results are not presented as completed official GPT evaluation.

Standard measurement methods Quantization uses group-wise symmetric absmax INT4 and INT2 over per-(layer, KV-head, token) rows of 128 dimensions. Pruning uses two query-agnostic selectors: grid-uniform geometric anchors and FastV-blind rankings from second-layer visual self-attention. A question-conditioned selector is disclosure-only and enters no criterion. Sampled-token retention appends 17 evenly spaced evicted tokens with their original K and V, without changing positions. Moment summaries partition the image-token grid into $K = 1, 4, 16$ equal-area tiles and store per-tile counts, K/V means, and low-order moment factors. A log-normal readout operates at question prefill and every decode step, matching the timing at which retained tokens are physically present. Merged tokens use the same 16 tiles but append post-RoPE K/V means directly, without

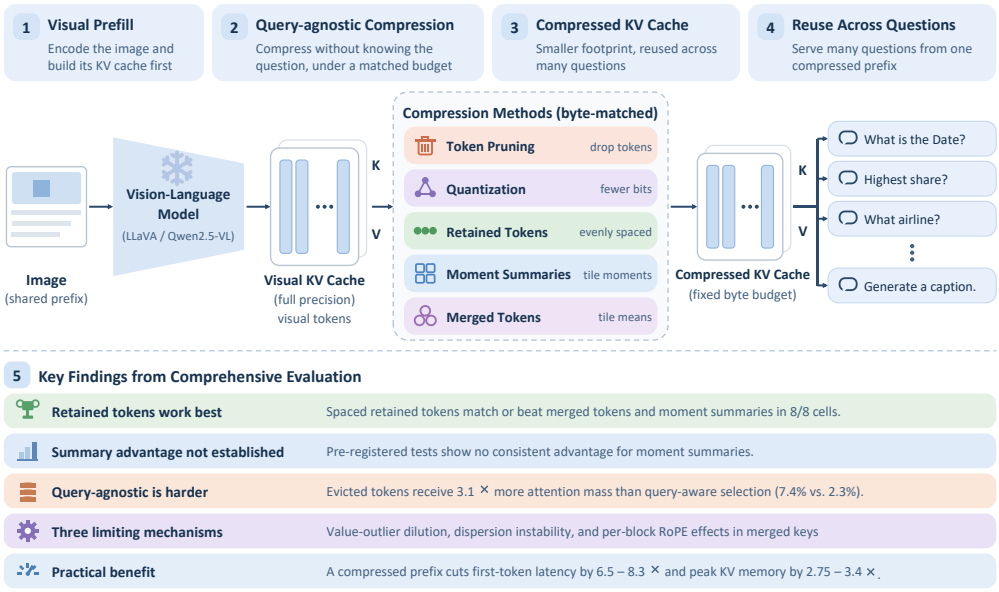

Figure 1: The diagram of our proposed query-agnostic visual KV compression scheme: prefill once, compress five byte-matched method families to a fixed budget, then reuse the cache across questions.

a count, dispersion (second-moment, variance) correction, or value term. These families share a budget ladder of about 16 token-equivalents (8196–8256 bytes, within 1% of one another). Figure 1 traces the deployment path: visual prefill, compression without a question under a fixed byte budget, reuse of the compressed cache, and answering many questions.

Pre-registration and prespecified evaluation. Before running experiments, we fixed the decision thresholds, memory constraints, maximum permitted numerical-fallback rate, sample sizes, and statistical protocol (paired bootstrap, 2000 resamples, seed 42). Evaluation applies these prespecified rules uniformly and discloses all covered results. Failure of a prespecified criterion may be addressed only by an amendment, not an ad hoc exemption. The appendix records the corresponding protocol details.

4 Theory: When Will Evicted-KV Summaries Be Useful?

This section derives the mass bound and analyzes limits of moment and merging methods, and relates each mechanism result to a measurement.

Model. Fix one query q with keys $\{k_i\}$ and values $\{v_i\}$, and partition the index set into a retained core C and an evicted set D. Then $x_i = q \cdot k_i$, $z_i = e^{x_i}$, $Z_S = \sum_{i \in S} z_i$, and $m_S = Z_S^{-1} \sum_{i \in S} z_i v_i$ for the mass-weighted mean over a set. The quantity that governs everything below is the mass ratio carried by the evicted set, $r = Z_D / Z_C$. Full attention and core-only attention yield

$$o_{\text{full}} = \frac{Z_C m_C + Z_D m_D}{Z_C + Z_D} \quad \text{and} \quad o_{\text{core}} = m_C. \tag{1}$$

A residual summary, the estimator form of the eviction summary, is any pair of estimators $(\hat{S}, \hat{Z})$ for the evicted numerator and the evicted mass, read out as

$$\tilde{o} = \frac{Z_C m_C + \hat{S}}{Z_C + \hat{Z}}. \tag{2}$$

The moment-summary readout is one instance of this form. Sampled-token retention lies outside this estimator class; this distinction is central to the byte-matched comparison.

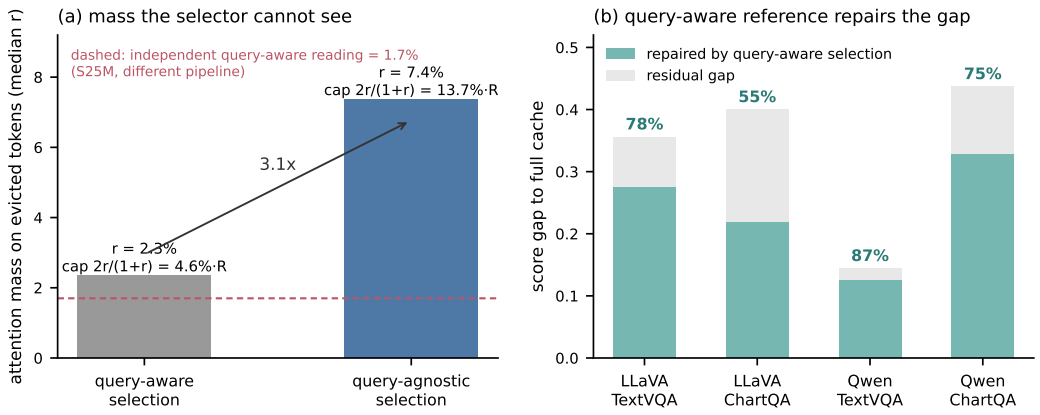

Figure 2: Evicted-token attention mass and its repairable damage. (a) Query-aware and query-agnostic selection; (b) recovered gap across anchor cells.

Proposition 1 (Mass bound). *Let* $R = \max_{i \in C \cup D} \|v_i - \bar{v}\|$ *for any reference point* $\bar{v}$. *Then* (a) $\|o_{\text{full}} - o_{\text{core}}\| = \frac{r}{1+r}\|m_D - m_C\| \le \frac{2r}{1+r}R$; *and* (b) *writing* $E_S = \hat{S} - Z_D m_D$ *and* $E_Z = \hat{Z} - Z_D$, *every estimator with* $\hat{Z} \ge Z_D$ *obeys*

$$\|\tilde{o} - o_{\text{core}}\| \;\le\; \frac{2r}{1+r}R \;+\; \frac{\|E_S\| + |E_Z|\,\|o_{\text{full}}\|}{Z_C + Z_D}. \tag{3}$$

Under the weaker hypothesis $\hat{Z} \ge 0$ *the same bound holds with* Z_C *in place of* $Z_C + Z_D$, *that is, the estimator-error term is inflated by the factor* $1 + r$.

Part (a) is independent of the estimator. Even an oracle given the exact (Z_D, m_D) repairs at most $\frac{2r}{1+r}R$. Part (b) adds an estimator's moment-level error to that ceiling. The two parts distinguish *whether a summary could help* from *whether a suitable summary can be constructed.*

Corollary 1 (The query-aware limit). *If question-conditioned selection guarantees an evicted-mass ratio* $r \le \varepsilon_m$, *then the upper bound on recoverable error obeys* $\frac{2r}{1+r}R \le \frac{2\varepsilon_m}{1+\varepsilon_m}R \le 2\varepsilon_m R$. *Thus even an oracle residual correction has at most this much headroom over core-only attention, and the headroom vanishes as* $\varepsilon_m \to 0$.

The corollary has a measurable consequence. Under question-conditioned selection the median mass ratio is 2.3%, compared with a query-agnostic median of 7.4% in the same-pipeline paired measurement. This 3.1× rise increases the ceiling from $\approx 0.046R$ to $\approx 0.137R$. An independent pipeline reproduces the query-aware side at 1.7%; this is reported only as a cross-check and is not included in the paired ratio. Across the four selected model–task pairs, the query-agnostic core is 0.14–0.44 below the full cache, and the question-conditioned reference repairs 55–87% of that gap. Most damage is therefore attributable to selection, consistent with Corollary 1 and Proposition 1. Prefix caching leaves potential value for eviction summaries, but the next three results explain why the moment family cannot recover it.

Three mechanisms, one cause. Three further mechanisms identify distinct limitations of moment and merging methods with a common cause: a summary records which values occurred but not which token carried them. Under 1D RoPE, averaging a block over positions can attenuate its key norm when the pre-RoPE key content is shared; this is a conditional per-block mechanism, not a universal attention-logit result. The outlier analysis demonstrates the first effect directly: summary approximation errors are two-sided, whereas token retention preserves selected tokens exactly. The dispersion analysis demonstrates the rank-dependent instability, since increasing rank introduces the dispersion term, which explains why rank > 0 can perform worse than rank $= 0$. The comparison across model families demonstrates the RoPE mechanism directly: one mechanism, rather than a difference in implementation quality, explains why merging causes severe performance degradation in one family and weaker effects in the other.

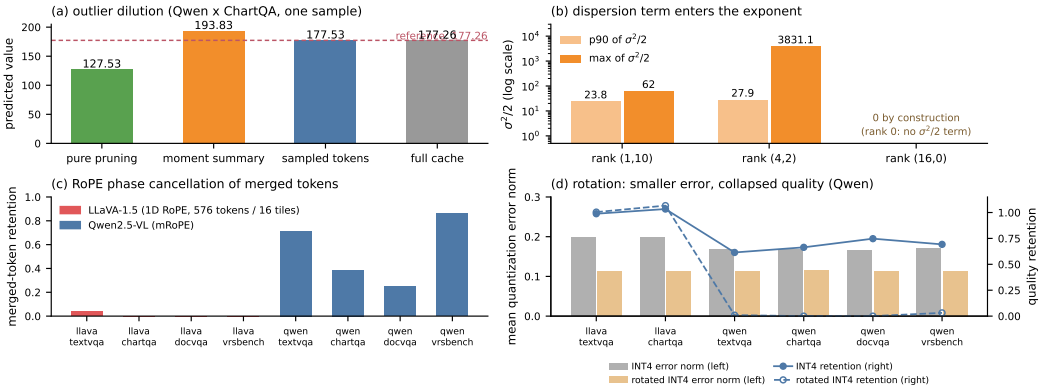

Figure 3: Comparison of the measured correspondences for outlier dilution, dispersion instability, RoPE phase cancellation, and quantization-error direction.

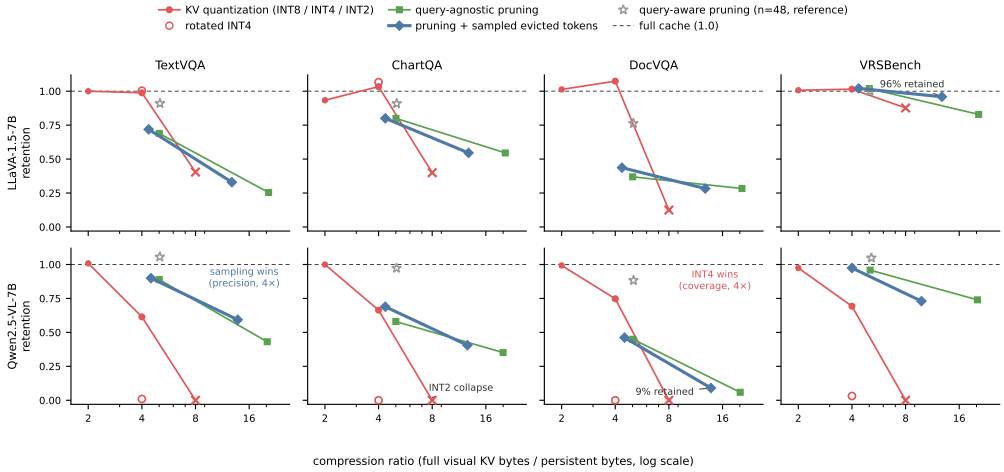

Figure 4: Quality retention, normalized to the full cache, versus median compression ratio on a logarithmic axis. Curves show quantization; squares, diamonds, and stars denote pruning, sampled-token retention, and the question-conditioned reference.

Error direction as a fourth correspondence. Rotation tests the same mechanism in quantization. Downstream quality is sensitive to error *direction*, not just magnitude, as the value-outlier mechanism also predicts. Figure 3 summarizes the four measured correspondences.

Why sampling avoids these limitations. Moment summaries cover all of D but are structurally limited in precision, whereas sampling preserves precision but has limited coverage.

5 The Compression-Regime Map

In this section, we draw the compression-regime map: we compare the five query-agnostic method families on one common pair of axes, quality retention against compression ratio, and locate the regime in which each family is the right tool. Figure 4 shows the map for the eight cells (model $\times$ benchmark). The map has three regimes: a model- and task-dependent $4\times$ INT4 regime, an $8\times$ per-row-INT2 regime where pruning has the highest observed quality retention at its lower $5\times$ operating point, and an extreme $\geq 10\times$ regime where sampled-token retention is the strongest tested eviction approach.

Table 1: Query-agnostic visual KV compression.

	TextVQA		ChartQA		DocVQA		VRSBench	
Method	L	Q	L	Q	L	Q	L	Q
Full cache	0.633	0.873	0.200	0.793	0.172	0.925	0.439	0.425
INT8 (2×)	0.633	0.881	0.187	0.793	0.174	0.918	0.442	0.414
INT4 (4×)	**0.626**	0.536	**0.207**	0.527	**0.185**	**0.691**	0.446	0.294
Rotated INT4 (4×)	0.636	0.008	0.213	0.000	—	0.000	—	0.013
INT2 (8×)	0.256	0.000	0.080	0.000	0.021	0.000	0.385	0.000
Pruning (5.0×)	0.436	0.776	0.160	0.460	0.064	0.416	0.448	0.407
FastV-blind (5.0×)	0.548	0.842	0.193	0.493	0.128	0.512	0.393	0.377
+ moment summaries (4.4×)	0.428	0.775	0.160	0.460	0.065	0.415	0.422	0.381
+ merged tokens (4.4×)	0.026	0.623	0.000	0.307	0.000	0.232	0.000	0.366
+ sampled tokens (4.4×)	0.455	**0.785**	0.160	**0.547**	0.075	0.427	**0.448**	**0.414**
Question-conditioned†	0.609	0.884	0.208	0.750	0.165	0.841	0.508	0.449
+ sampled tokens (≥ 10×)	0.221	0.497	0.125	0.312	0.061	0.086	0.487	0.312

Table 2: First-token latency and peak KV memory of the full and compressed prefix caches.

	First token (ms)				Peak KV memory (MiB)		
Model	Prefill	Full cache	Compressed	Speedup	Full	Compressed	Ratio
LLaVA-1.5	97.0	20.6	**14.9**	6.5×	313.5	91.5	3.4×
Qwen2.5-VL	133.4	19.9	**16.1**	8.3×	20.9	7.6	2.75×

Table 1 gives the numerical compression map behind Figure 4: absolute scores for LLaVA-1.5-7B (L) and Qwen2.5-VL-7B (Q), where bold marks the stronger of the two 4×-class options in each cell. Those two options divide the eight cells unevenly. INT4 quantization wins the three LLaVA VQA cells and Qwen DocVQA, and it is nearly free there: within 0.007 of the full cache on LLaVA TextVQA, and above it on LLaVA ChartQA and DocVQA. Sampled-token retention wins the remaining four cells, which include both VRSBench cells, and by clear margins (Qwen TextVQA 0.785 against 0.536). Summarizing the evicted bytes is the weakest equal-byte option of all: merged tokens collapse to 0.026 and 0.000 on LLaVA TextVQA and ChartQA. No single family dominates the map, which is the paper's central claim stated as a measurement: the right method depends on the model, the task, and the compression regime. The question-conditioned reference is disclosure-only; ratios are per-cell median memory-accounting values, with the metadata-inclusive ratios for per-channel quantization in Appendix F.

Table 2 gives the deployment payoff of the compressed prefix: the 6.5× and 8.3× first-token latency reductions and the 2.75–3.4× peak-memory reductions quoted in Section 1. The uncompressed full cache is already cheaper than per-question prefill, and compression widens that advantage by a further 1.4× and 1.2×; the latency gain therefore comes together with a persistent-memory gain rather than by trading one against the other.

Sampled-token retention at ≥ 10×. The extreme budget retains about 5% of visual tokens and adds 17 evenly spaced evicted tokens, giving about 12.8× compression. Its margin over blind pruning is non-negative in 7/8 cells and significant on Qwen TextVQA and LLaVA VRSBench; Qwen VRSBench is the sole negative cell and its interval straddles zero. A per-channel INT4 core admits 1.6–2.4× more retained tokens at the same effective budget, but its exploratory comparisons are inconclusive. The byte-matched conclusion is therefore scoped to bf16 retained cores.

6 Allocating the Evicted Byte Budget

In this section, we put the bounded summaries of Section 4 to a byte-matched test against retained tokens: we run three pre-registered falsification rounds and read the continuum comparison over all eight cells.

Table 3: Byte-matched continuum with bf16 retained cores ($\rho = 0.2$).

	TextVQA		ChartQA		DocVQA		VRSBench	
Method (equal bytes)	L	Q	L	Q	L	Q	L	Q
Full cache	0.633	0.873	0.200	0.793	0.172	0.925	0.439	0.425
Pruning only	0.436	0.776	0.160	0.460	0.064	0.416	0.448	0.407
+ moment summaries	0.428	0.775	0.160	0.460	0.065	0.415	0.422	0.381
+ merged tokens	0.026	0.623	0.000	0.307	0.000	0.232	0.000	0.366
+ sampled tokens	0.455	0.785	0.160	0.547	0.075	0.427	0.448	0.414
Mixture (exploratory)	0.395	0.756	0.167	0.553	—	—	—	—
Sampling control (exploratory)	0.479	0.771	0.188	0.600	—	—	—	—

6.1 Implementations

All rounds use one segmented pipeline. Empty-summary injection reproduces the unpatched attention output, and segmented inference matches the unsegmented path apart from one disclosed score-equivalent degenerate sample. The log-space readout has zero fallback in the main matrix and continuum study.

6.2 Falsification rounds

Round one: global moment summaries. A global rank-8 summary (about 13 token-equivalents) is compared with 14 retained tokens and the full cache in four cells. No cell favors the summary: paired differences are -0.014, -0.063, -0.259, and -0.104.

Round two: spatial-tiling configurations. We test equal-byte tiled summaries with $(K, \ell) = (1, 10)$, $(4, 2)$, and $(16, 0)$, each at about 16 token-equivalents. No configuration establishes superiority over retained tokens: paired differences range from -0.087 to $+0.013$, and the sole interval excluding zero is negative (Qwen2.5-VL ChartQA, $(16, 0)$). The LLaVA results across configurations are non-monotone, with a trough at $K = 4$.

Round three: the tested mixture. This exploratory study compares a 16-tile summary plus 17 retained tokens with 34 retained tokens at equal bytes. Three cells do not pass the prespecified expansion criterion; the expanded Qwen ChartQA comparison reverses to -0.047 with interval $[-0.120, +0.033]$. No tested mixture establishes superiority over plain sampling.

6.3 The continuum comparison

Point estimates order the methods as sampled-token retention, moment summaries, then merged tokens. Sampled-token retention's paired margin over moment summaries is non-negative in 8/8 cells. Seven of the eight intervals straddle zero; the one that clears it is Qwen ChartQA at $+0.087$ with an interval of $[+0.033, +0.147]$, and LLaVA ChartQA is an exact tie. Moment summaries outperform merged tokens in 8/8 point estimates.

Table 3 reports the byte-matched absolute scores. Its first five rows (Full cache, Pruning only, and the three equal-byte variants) use bf16 retained cores; its last two rows are exploratory and cover only the four TextVQA and ChartQA cells; the corresponding quantized-core comparison is reported separately in Appendix G. Sampled-token retention is non-negative against moment summaries in 8/8 grid-uniform cells and 6/8 FastV-blind cells. Figure 5 shows the paired margins from adding sampled tokens on both selector bases.

7 Discussion and Limitations

In this section, we make explicit what the preceding evidence does and does not cover.

Scope. Our measurements hold under greedy decoding, short-answer VQA and two model families, with the model-side claim pinned to the 1D RoPE/mRoPE structural difference. The full cache

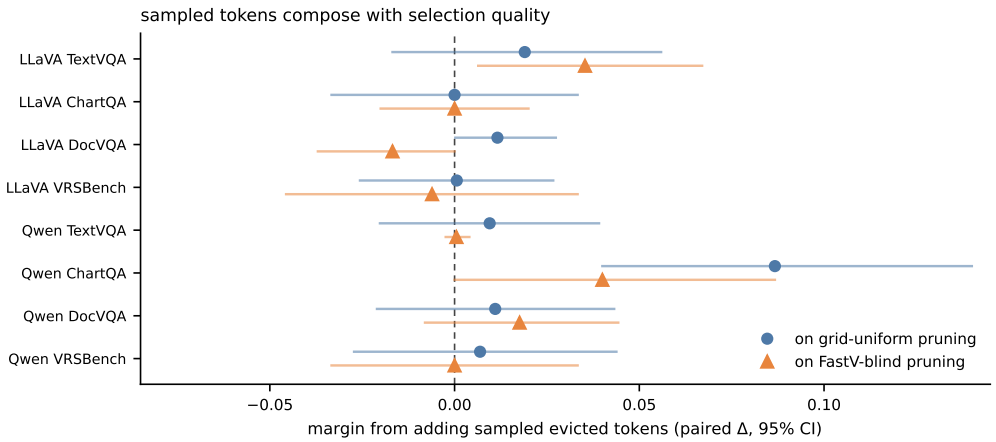

Figure 5: Paired margins (95% intervals) from adding sampled tokens to grid-uniform (circles) or FastV-blind (triangles) pruning.

on LLaVA DocVQA reaches only 0.172, so that cell has limited discriminating power. At $\geq 10\times$ compression, every tested method incurs substantial absolute loss on text-dense tasks. The setting is a single-image prefix; multi-image and video prefixes are not covered.

Quantization scope. Beyond basic per-row INT4 and INT2, we evaluate rotation for spreading outliers, asymmetric scaling, and KIVI-style grouping, although coverage is uneven (this paper). The asymmetric and per-channel INT4 configurations cover all eight cells. Per-channel INT2 covers four without a pre-registered adjudication rule, and the 2×2 grouping factorial isolating K covers three. KIVI's residual full-precision recent window is inapplicable to a static prefix and was not implemented, a disclosed deviation omitting a component that could only strengthen that configuration. Asymmetric INT2 uses four levels against three for ternary symmetric INT2, so its greater representational capacity makes the collapse of the per-row asymmetric configuration conservative evidence for our claim.

Ledger basis. Quantization compression ratios can be reported on a payload-only main ledger ($4\times$ and $8\times$, the basis of Figure 4). Metadata-inclusive accounting ($\approx 3.2\times$ and $\approx 5.3\times$) is required for comparing asymmetric and per-channel configurations. We label the basis at each quantization comparison. The reversal in Section 5 is not a byte-matched win because it involves different effective memory budgets.

Open problems. Additional model families, positional encodings, and information-density regimes may revise the map. Summaries may help when evicted mass is less concentrated or value outliers are controlled; learned query-agnostic policies remain outside this training-free study.

8 Conclusion

Prefix caching requires query-agnostic compression. In this paper, we investigate the query-agnostic visual KV compression scheme under matched memory budgets. Our research reveals that with bf16 retained cores, sampled tokens have non-negative point-estimate margins over summaries in all eight cells, although most intervals include zero. Meanwhile, selection and sampling are complementary: together they provide a practical deployment strategy. Under deeper eviction, sampling preserves individual token information while summary aggregates it. This decouples two commonly confused design decisions: query-agnostic selection determines which token information will survive after eviction, while representation determines whether the surviving bytes will retain individual tokens or aggregate them. The attention-mass bound and mechanism analysis explain when summaries have limited value. These results guide the selection of compression methods and elucidate their limitations.

AI USE STATEMENT

Language-model assistants were used for software engineering, numerical analysis support, and manuscript drafting. The authors designed the hypotheses and statistical protocol, verified the reported calculations, and take full responsibility for the scientific content and the paper's compliance with the conference policy.

REPRODUCIBILITY STATEMENT

The paper specifies the model configurations, evaluation metrics, byte-matched comparisons, decoding procedure, and paired-bootstrap protocol. The authors will provide the implementation and evaluation materials in accordance with the conference reproducibility policy.

REFERENCES

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b5b939436789f76f08b9d0da5e81af7c-Abstract-Conference.html.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Baobao Chang, Junjie Hu, and Wen Xiao. Pyramidkv: Dynamic KV cache compression based on pyramidal information funneling. *CoRR*, abs/2406.02069, 2024. doi: 10.48550/ARXIV.2406.02069. URL https://doi.org/10.48550/arXiv.2406.02069.

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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, volume 15139 of *Lecture Notes in Computer Science*, pp. 19–35. Springer, 2024. doi: 10.1007/978-3-031-73004-7_2. URL https://doi.org/10.1007/978-3-031-73004-7_2.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length LLM inference with KV cache quantization. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/028fcbcf85435d39a40c4d61b42c99a4-Abstract-Conference.html.

Soumyadeep Jana, Sagar Nishad, and Sanasam Ranbir Singh. Moment-kv: Momentum-based decode-time kv cache compression for long generation, 2026. URL https://arxiv.org/abs/2605.29873.

Jang-Hyun Kim, Jinuk Kim, Sangwoo Kwon, Jae W. Lee, Sangdoo Yun, and Hyun Oh Song. Kvzip: Query-agnostic KV cache compression with context reconstruction. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/f4eaa4b8f2d08edb3f0af990d56134ea-Abstract-Conference.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pp. 611–626. ACM, 2023a. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pp. 611–626. ACM, 2023b. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

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. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/28ab418242603e0f7323e54185d19bde-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen (Henry) Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32332–32344. PMLR / OpenReview.net, 2024b. URL https://proceedings.mlr.press/v235/liu24bz.html.

Bole Ma, Jan Eitzinger, Harald Köstler, and Gerhard Wellein. Kamera: Unified position-invariant multimodal KV cache for training-free reuse. *CoRR*, abs/2606.23581, 2026. doi: 10.48550/ARXIV.2606.23581. URL https://doi.org/10.48550/arXiv.2606.23581.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=HMrcv7Q4Ub.

Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. LOOK-M: look-once optimization in KV cache for efficient multimodal long-context inference. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pp. 4065–4078. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.235. URL https://doi.org/10.18653/v1/2024.findings-emnlp.235.

Zhongwei Wan, Xinjian Wu, Yu Zhang, Yi Xin, Chaofan Tao, Zhihong Zhu, Xin Wang, Siqi Luo, Jing Xiong, Longyue Wang, and Mi Zhang. D_{2}O: Dynamic discriminative operations for efficient long-context inference of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HzBfoUdjHt.

Ao Wang, Hui Chen, Jianchao Tan, Kefeng Zhang, Xunliang Cai, Zijia Lin, Jungong Han, and Guiguang Ding. Prefixkv: Adaptive prefix KV cache is what vision instruction-following models need for efficient generation. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/87b8bc2289d95bda4a8b620fc2232a9b-Abstract-Conference.html.

Lyuke Wang, Zhuo Li, and Guangxu Zhu. Viscache: Visual kv cache pruning for efficient vision large language model inference, 2026. URL https://arxiv.org/abs/2608.24063.

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. *CoRR*, abs/2508.15717, 2025. doi: 10.48550/ARXIV.2508.15717. URL https://doi.org/10.48550/arXiv.2508.15717.

Yuhang Zhan, Lisi Chen, and Shuo Shang. Reskv: Reconstructing omitted attention contributions for fixed-budget KV cache compression. *CoRR*, abs/2607.29591, 2026. doi: 10.48550/ARXIV.2607.29591. URL https://doi.org/10.48550/arXiv.2607.29591.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zhang25s.html.

Yuxin Zhang, Yuxuan Du, Gen Luo, Yunshan Zhong, Zhenyu Zhang, Shiwei Liu, and Rongrong Ji. Cam: Cache merging for memory-efficient llms inference. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58840–58850. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/zhang24n.html.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark W. Barrett, Zhangyang Wang, and Beidi Chen. H2O: heavy-hitter oracle for efficient generative inference of large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract-Conference.html.

A Verification Gates and Segmented Parity

Four verification gates were designed and passed before any production run. The adjudication of each gate is recorded in A.1–A.3. All four gates use shared per-sample parity records.

Table 4: Agreement between segmented and unsegmented inference.

Model	Samples	Adjudication	Note
LLaVA-1.5-7B	2/2	token-for-token identical	—
Qwen2.5-VL-7B	1/2	token-for-token identical	—
Qwen2.5-VL-7B	1/2	score-equivalent	degenerate sample; both paths collapse into a repeated-newline output and diverge at the seventh token on a content word, scoring identically

Table 5: Byte gaps in byte-matched comparisons ($\rho = 0.2$, $n = 150$).

Cell	Mean relative gap	Largest gap
LLaVA-1.5 TextVQA	0.58%	508 B
LLaVA-1.5 ChartQA	0.56%	508 B
LLaVA-1.5 DocVQA	0.53%	448 B
LLaVA-1.5 VRSBench	0.51%	448 B
Qwen2.5-VL TextVQA	0.43%	448 B
Qwen2.5-VL ChartQA	0.57%	448 B
Qwen2.5-VL DocVQA	0.43%	448 B
Qwen2.5-VL VRSBench	0.69%	448 B

A.1 The summary-injection gate

Injecting an empty summary (count $= 0$, zero-byte moments) reproduces the unpatched eager attention output bit for bit. The gate excludes the possibility that the summary-injection harness introduces numerical side effects of its own. The summary readout runs in log space. With an empty summary, the log-partition correction and the weighted-value contribution reduce to zero; the exact equality check verifies that the output logits are identical.

A.2 Segmented-inference agreement

The following table reports output agreement between the segmented path (image-prefix prefill $\rightarrow$ compression $\rightarrow$ question-segment prefill $\rightarrow$ decode) and unsegmented full inference. On Qwen2.5-VL, passing positions explicitly suppresses the model's internal rotary-position offset update. To address this mRoPE issue, the position indices and the accumulated rotary-position offset are passed back explicitly between the two prefill segments, so that the positional-encoding state of the second segment matches the unsegmented path.

A.3 Byte-accounting assertions

The following table reports the per-sample memory-accounting check for every byte-matched comparison (Section 3 and Appendix C give the formulas). The pre-registered thresholds (Section 3) are met at $\rho = 0.2$: all eight cells pass, with absolute gaps of 448–508 B against the 512 B ceiling.

At the deeper $\rho = 0.05$ budget, the relative ceiling is exceeded. Ten of twelve cell–configuration combinations have a mean relative gap above 1% (1.04–1.31%); the two below it are Qwen2.5-VL DocVQA (0.94%) and TextVQA (0.95%). This is a granularity effect: one bf16 token is 512 B, a larger fraction of the smaller persistent budget. At roughly 39 KB, it is a 1.3% floor, which motivates the 2% ceiling in the deep-regime composition round (Appendix G.6).

The substantive fairness claim is unchanged. The sampled method never uses fewer persistent bytes than the summary method, the single-sided violation count is zero, and no absolute gap exceeds 512 B. Thus a relative ceiling calibrated at one budget cannot be carried unchanged to a budget four times smaller; the main compression-regime results report the $\geq 10\times$ configurations at their actual ratios.

B Proofs and the Theory-to-Evidence Correspondence

B.1 Proofs and auditable limits

Setup. Notation is that of Section 4: retained core C, evicted set D, logits x_i and masses z_i, Z_S and m_S for any set S, mass ratio $r = Z_D/Z_C$, and the outputs o_{full}, o_{core} and $\tilde{o}$. The readout of the stored summary statistics (the moment sketch) uses log-normal moment matching with a first-order value term and is one instance of $(\hat{S}, \hat{Z})$. Retaining uniformly sampled evicted tokens is not, as explained by the escape argument in Section 4.

Proof of Proposition 1. *(a)* Both m_C and m_D are convex combinations of $\{v_i\}$, so $\|m_C - \bar{v}\| \le R$ and $\|m_D - \bar{v}\| \le R$, and the triangle inequality gives $\|m_D - m_C\| \le 2R$. Substituting the definitions yields the identity $o_{\text{full}} - o_{\text{core}} = \frac{Z_D}{Z_C+Z_D}(m_D - m_C) = \frac{r}{1+r}(m_D - m_C)$, which proves (a). *(b)* Substituting $\hat{S} = Z_D m_D + E_S$ and $\hat{Z} = Z_D + E_Z$ into $\tilde{o}$ and factoring $(Z_C + Z_D)o_{\text{full}}$ out of the numerator gives the exact identity

$$\tilde{o} - o_{\text{full}} = \frac{E_S - E_Z\, o_{\text{full}}}{Z_C + Z_D + E_Z} = \frac{E_S - E_Z\, o_{\text{full}}}{Z_C + \hat{Z}}, \tag{4}$$

so the triangle inequality gives $\|\tilde{o} - o_{\text{full}}\| \le \big(\|E_S\| + |E_Z|\,\|o_{\text{full}}\|\big)/(Z_C + \hat{Z})$. The hypothesis $\hat{Z} \ge Z_D$ makes $Z_C + \hat{Z} \ge Z_C + Z_D$ and yields the denominator in the statement. Under only $\hat{Z} \ge 0$ the same argument gives $Z_C + \hat{Z} \ge Z_C$, which is the weaker variant quoted there; the stronger denominator is not available from non-negativity alone, since $\hat{Z}$ may underestimate Z_D arbitrarily. Finally $\|\tilde{o} - o_{\text{core}}\| \le \|\tilde{o} - o_{\text{full}}\| + \|o_{\text{full}} - o_{\text{core}}\|$, and part (a) bounds the second term by $\frac{2r}{1+r}R$, which gives the stated bound on $\|\tilde{o} - o_{\text{core}}\|$.

Proof of Corollary 1. The condition $r \le \varepsilon_m$ is the stated retained-mass guarantee of the question-conditioned selector. Since $r \mapsto 2r/(1+r)$ is increasing on $r \ge 0$, substitution gives $\frac{2r}{1+r}R \le \frac{2\varepsilon_m}{1+\varepsilon_m}R \le 2\varepsilon_m R$, which vanishes as $\varepsilon_m \to 0$.

Proposition 2 (Outliers are not reconstructible from moments)**.** *Let $n \ge 2$ and model a tile's value side as $V \in \mathbb{R}^{n\times d}$. Suppose its stored representation is a deterministic permutation-invariant function only of the first moment $M_1 = \frac{1}{n}\sum_i v_i$ and a rank-ℓ factorization derived from $M_2 = \frac{1}{n}V^\top V$. Then:* (i) Support is unidentifiable. *For a constant base tile $V_{\text{base}} = \mathbf{1}v_{\text{base}}^\top$ and any outlier w with $\|w\| = R$, the entire family $\{V_j = V_{\text{base}} + e_j w^\top\}_{j=1}^n$ maps to the same (M_1, M_2). Every such reconstruction is constant on this family, and its worst-case Frobenius error is at least $R/\sqrt{2}$.* (ii) Mean reconstruction dilutes outliers. *Under a uniform prior over the outlier position, the Bayes-optimal permutation-invariant reconstruction is the mean $v_{\text{base}} + w/n$. If attention assigns share α within the tile to the outlier and the readout normalizer is $Z_C + \hat{Z}$, this reconstruction has output error $\frac{Z_D}{Z_C+\hat{Z}}R\,|\alpha - 1/n|$. When $\hat{Z} = Z_D$, this becomes $\frac{r}{1+r}R\,|\alpha - 1/n|$. A budget of K tiles resolves support only to tile granularity, giving mean-estimator dilution RK/n.* (iii) Rank adds no support information. *The unordered sum M_2 enumerates which value directions occurred, not which value occupied each row. Thus increasing ℓ can resolve value directions but not row support.*

Proof of Proposition 2. *(i)* Expand directly: $V_j^\top V_j = V_{\text{base}}^\top V_{\text{base}} + V_{\text{base}}^\top e_j w^\top + w e_j^\top V_{\text{base}} + w e_j^\top e_j w^\top$, in which $V_{\text{base}}^\top V_{\text{base}} = n\, v_{\text{base}} v_{\text{base}}^\top$, $V_{\text{base}}^\top e_j = v_{\text{base}}$ independently of j, and $e_j^\top e_j = 1$. Hence M_2 does not depend on j; the same expansion gives $M_1(V_j) = v_{\text{base}} + w/n$. Any readout is therefore constant across the family, and two members of it are $\|e_j w^\top - e_{j'} w^\top\|_F = \sqrt{2}R$ apart, so the worst-case Frobenius error is at least $\sqrt{2}R/2 = R/\sqrt{2}$. *(ii)* Under a uniform prior over the outlier position, a permutation-invariant estimator outputs the mean $\mu = v_{\text{base}} + w/n$ in every row, which is Bayes-optimal among such estimators. If attention gives the outlier share α, the true numerator is $Z_D(v_{\text{base}} + \alpha w)$ while the mean reconstruction supplies $Z_D(v_{\text{base}} + w/n)$. Their difference has norm $Z_D R\,|\alpha - 1/n|$; division by the readout normalizer gives $\frac{Z_D}{Z_C+\hat{Z}}R\,|\alpha - 1/n|$. Under exact mass recovery $\hat{Z} = Z_D$, this is $\frac{r}{1+r}R\,|\alpha - 1/n|$. Splitting the tile into K equal-support regions reduces the mean-estimator dilution to RK/n. *(iii)* By (i), the unidentifiability of

n distinct supports does not depend on ℓ, because a factorization derived from M_2 recovers no row information. *On a weaker energy-conservation convention.* Spreading the outlier energy R^2 over n rows and ℓ directions would give a per-row amplitude of $R/\sqrt{n\ell}$. The two bounds proved here are the worst-case bound $R/\sqrt{2}$ when the support is unknown and the dilution bound RK/n when the support is known at tile granularity. These bounds are stronger and simpler, and both are independent of ℓ. The measurements show this same pattern. $(16, 0)$ and $(4, 2)$ do not differ on the value-side rescue, and rank above zero contributes only the $\sigma^2/2$ term of Proposition 3. Nothing here relies on the weaker bound.

Proposition 3 (Dispersion correction diverges)**.** *For a tile with logits $x \in \mathbb{R}^n$, population mean μ, population variance σ^2, and $x_{\max} = \max_i x_i$, the log-normal moment-matching estimate $\hat{Z} = ne^{\mu+\sigma^2/2}$ of $Z = \sum_i e^{x_i}$ satisfies*

$$\hat{Z}/Z \;\geq\; e^{\sigma^2/2-(x_{\max}-\mu)}. \tag{5}$$

For $n > 1$ and logits $(s, 0, \ldots, 0)$ with $s \geq 0$, $\mu = s/n$, $\sigma^2 = s^2(n-1)/n^2$, and $x_{\max} - \mu = s(n-1)/n$, so

$$\hat{Z}/Z \;\geq\; \exp\Big(\frac{s^2(n-1)}{2n^2} \;-\; \frac{s(n-1)}{n}\Big). \tag{6}$$

This lower bound exceeds 1 for $s > 2n$ and diverges as $s \to \infty$.

Proof of Proposition 3. Since $Z = \sum_i e^{x_i} \leq n\, e^{x_{\max}}$,

$$\hat{Z}/Z \;\geq\; \frac{n\, e^{\mu+\sigma^2/2}}{n\, e^{x_{\max}}} \;=\; e^{\sigma^2/2-(x_{\max}-\mu)}. \tag{7}$$

For the one-hot tile of the statement, direct computation gives $\mu = \frac{1}{n}\big(s + 0 \cdot (n-1)\big) = s/n$ and $\sigma^2 = \frac{1}{n}\big[(s-\frac{s}{n})^2 + (n-1)(\frac{s}{n})^2\big] = \frac{s^2(n-1)}{n^2}$, while $x_{\max} - \mu = s - s/n = s(n-1)/n$. Substituting yields the stated exponent, which is positive exactly when $s > 2n$. This worst-case step is loose by up to a factor n, because the true mass of a one-hot tile is $Z = (n-1) + e^s$ rather than $n\, e^s$; the exact ratio is

$$\hat{Z}/Z \;=\; \frac{n \exp\big(\frac{s}{n} + \frac{s^2(n-1)}{2n^2}\big)}{(n-1) + e^s}, \tag{8}$$

which is larger than the bound above whenever $n > 1$. Note also that $\sigma^2 \leq s^2/4$ for every n, attained at $n = 2$, so no one-hot tile reaches the $\sigma^2 \approx s^2$ regime. The exponents measured in Section 4 ($\sigma^2/2$ peaking at 3831) therefore come from real tiles whose logit spread across many positions far exceeds this idealization; they sit deep in the amplifying regime, and the bound above is used only to exhibit the mechanism, not to size it.

Lemma 1 (Conditional per-block RoPE cancellation)**.** *Fix a two-dimensional block b. If all pre-RoPE keys in a tile equal a common vector $u^{(b)}$ and their positions are $p_j = p_0 + j\Delta$, then the post-RoPE keys $k_j^{(b)} = R(\omega_b p_j)u^{(b)}$ satisfy*

$$\frac{\|\bar{k}^{(b)}\|}{\|u^{(b)}\|} = \frac{|\sin(n\omega_b\Delta/2)|}{n\,|\sin(\omega_b\Delta/2)|} = O(1/n), \tag{9}$$

when $\omega_b\Delta$ is bounded away from $0 \pmod{2\pi}$. No such bound follows when pre-RoPE content varies within the tile or positions are not equally spaced. The result concerns key norms, not attention logits.

Lemma 1 (structural argument). Under its shared-content and equal-spacing assumptions, the lemma gives a geometric-series cancellation in each RoPE block. The centroid method merges LLaVA-1.5's 576 visual tokens over 16 equal-area tiles, so the relevant tile size is roughly 36 positions rather than 576. At that scale, rotating blocks can be attenuated while low-frequency blocks with $\omega_b\Delta n \ll 1$ are barely rotated. The near-zero merged-token score on all four LLaVA cells is consistent with this conditional per-block mechanism; it is not a proof that every attention logit vanishes. Under mRoPE (Qwen2.5-VL, with temporal, height, and width components split across dimensions), the positional structure differs, so cancellation need not be complete. This is a structural interpretation rather than a full spectral analysis of RoPE.

Auditable limits.

1. *Single-query model.* It does not cover multi-head or grouped-query coupling, nor error accumulation over multi-step decoding. The sketch readout is applied independently per query, so error does not propagate across steps. This is an implementation fact, not part of any theorem here.
2. *R assumed bounded.* No parametric assumption is made on the value distribution, so Proposition 2 is a distribution-free worst case; that is precisely what makes it complementary to the heavy-tail and outlier measurements.
3. *Oracle bound, not claimed tight.* Part (a) of Proposition 1 is an oracle ceiling; the attainable gain also carries the estimator-error term of part (b), and we make no tightness claim.
4. *Informal lemma.* Lemma 1 is structural and cited rather than proved spectrally.
5. *Mechanism, not optimality.* The coverage-precision argument for why sampling escapes Proposition 2 is a mechanism statement plus measurement. A strict expected-coverage calculation would need assumptions on the mass distribution, which we do not make, and we claim no optimality.
6. *Measurement protocol, amended once.* The sketch readout was originally active at decode steps only. Because the sampled-token methods keep their extra tokens physically present in the cache during question prefill, that placement was not matched across prefill and decoding: the two estimator families did not reach the first answer token through the same path. Every sketch method reported in this paper was re-measured with the readout also active at question prefill. The amendment, its scope, the evidence that only the sketch methods are affected, and the reproduction path are filed in the prefill-stage protocol amendment; it is recorded as a dated erratum rather than folded in silently.

B.2 The correspondence table

Section 4 gives three theoretical mechanisms plus a fourth correspondence from the rotation experiment. Table 6 pairs the principal theoretical predictions with their supporting measurements.

C The Per-Sample Byte Ledger

C.1 Byte Formulas

Every byte figure in this appendix is per layer and per KV head unless a total is explicitly labelled as one. This is the unit used for the ledger's comparisons, allowing the sketch, token and quantization quantities below to be compared directly. Whole-cache totals are obtained by multiplying by $L \times H$ (C.2).

Token bytes. Each retained token stores its K and V in bf16,

$$\text{token_bytes} \;=\; 2 \times d_{\text{head}} \times 2 \;=\; 2 \times 128 \times 2 \;=\; 512\,\text{B},$$

the leading factor 2 counting K and V and the trailing one counting bf16. Summed over heads this is $2 \times d_{\text{kv}} \times 2$ per layer, where d_{kv} is the total KV width (LLaVA-1.5: $4096 = 32$ heads $\times\ 128$; Qwen2.5-VL's grouped-query attention aggregates onto the same convention).

Summary bytes. Per tile,

$$\text{sketch_bytes}_{\text{tile}} \;=\; (2d + 3d\ell) \times 2 + 4,$$

where d is the head dimension (128), ℓ is the spectral rank, the factor 2 is bf16 and the 4 bytes are the count (int32). Total summary bytes are this quantity times K tiles times L times H.

Table 6: Theory-to-evidence correspondence. Each prediction is paired with its primary empirical evidence.

#	Theory result / prediction	Evidence location
1	Proposition 1 (unified ceiling): recoverable benefit is limited by evicted attention mass.	Section 4; Figure 2
2	Corollary 1 (query-aware limit): query-aware selection leaves little repair headroom.	Section 4; Figure 2
3	Proposition 2 (outliers): moment summaries cannot reconstruct isolated value outliers, regardless of rank.	Section 4; Figure 3
4	Proposition 3 (dispersion): large variation can destabilize higher-rank moment corrections.	Section 4; Table 9
5	Lemma 1 (phase cancellation): positional rotation can cancel merged keys.	Section 4; Figure 3
6	Rotation experiment: quality depends on error direction, not error magnitude.	Section 4; Figure 3
7	Continuum verdict: at equal bytes, sampled-token retention matches or exceeds moment summaries.	Section 6.3; Table 3
8	Selection and sampling compose: sampling remains beneficial across selection methods.	Appendix H; Figure 5
9	INT4 model dependence: similar error can yield different quality across model families.	Section 5; Figure 4
11	Asymmetric-scaling decoupling (extends Proposition 2): grouping, not scaling symmetry or bit width, determines quality; smaller error need not improve quality.	Appendix F; Section F.2
12	K carries the load (extends Proposition 2): key grouping, not value grouping, restores quality; smaller error need not improve quality.	Appendix F; Section F.7

Table 7: Equal-byte moment-summary ladder.

Configuration (K, ℓ)	Sketch bytes	Token equiv.	Per-tile arithmetic
$(1, 10)$	8196 B	16.01 tokens	$1 \times ((256 + 3840) \times 2 + 4)$
$(4, 2)$	8208 B	16.03 tokens	$4 \times ((256 + 768) \times 2 + 4)$
$(16, 0)$	8256 B	16.13 tokens	$16 \times (256 \times 2 + 4)$

Quantization bytes. For the whole visual cache, including both K and V, the metadata-inclusive total is

$$B_{\text{effective}} = 2Ln_{\text{tokens}}d_{\text{kv}} \text{ bits}/8 + n_{\text{groups}}m,$$

where $d_{\text{kv}} = Hd$ is the width of one K or V tensor per layer and m is metadata per group: 2 B for a symmetric absmax scale and 4 B when an affine zero-point is stored. The first term is the payload-only quantity; metadata is disclosed separately in Appendix F.3. Per-row grouping makes one group for every (layer, KV head, token, K/V) row, so $n_{\text{groups}} = 2LHn$. KIVI-style grouping uses 32-token groups for K and 32-channel groups for V, giving $n_{\text{groups}} = LH\,[d\lceil n/32\rceil + n\lceil d/32\rceil]$. On LLaVA-1.5 TextVQA ($n = 576$, $LH = 1024$), these counts are respectively 1,179,648 and 4,718,592. Relative to payload, symmetric per-row metadata is 3.1% at INT4 and 6.2% at INT2; affine per-row metadata is 6.2% and 12.5%; and affine KIVI-style metadata is 25.0% and 50.0%, respectively.

C.2 The Equal-Byte Ladder

The three configurations of Section 6.2 are all aligned to the same byte budget. The maximum absolute gap across the three configurations is 60 B (<1%), on the same per-(layer, KV head) basis, so budget confounding is removed from cross-K comparisons. Whole-cache totals multiply these figures by $L \times H$. The model configurations are derived from the frozen byte fields in Appendix F.7.3. LLaVA-1.5 (LLaMA-2 multi-head attention, no grouped-query attention) has $L = 32$, $H = 32$, $L \times H = 1024$, with a head dimension of 128. Qwen2.5-VL-7B (grouped-query attention) has $L = 28$, $H = 4$, $L \times H = 112$, with a head dimension of 128. *Note.* An earlier draft of this appendix listed LLaVA-1.5 as 28 layers (Qwen2.5-VL's depth) and described the quantization group

Table 8: Median compression ratios.

Method	LLaVA	Qwen
INT8 (2×)	2.0×	2.0×
INT4 (4×)	4.0×	4.0×
INT2 (8×)	8.0×	8.0×
Pruned (5.0×)	5.0×	5.0–5.1×
+ summaries / merged / sampled (4.4×)	4.4×	4.4–4.5× (VRS 4.0–4.1×)
+ sampled ($\geq$ 10×)	12.8×	12.6–13.8× (VRS 9.8×)

as 16 tokens. Both are corrected here against the frozen byte fields (Appendix F.7.3 and the group counts of Appendix F.3). The grouping axes used in the experiments are the two given in C.1. No measured quantity changes. The sketch-byte column and the 60 B ladder gap are per (layer, KV head) and therefore independent of L and H, and the metadata fractions of J.5 carry $L \times H$ in both numerator and denominator, where it cancels.

C.3 Sampled-Token Retention Bytes

Seventeen uniformly sampled tokens $\times$ 512 B $=$ 8704 B. Compared with the summary configurations' 8196–8256 B, this gives a gap of about 448–508 B ($\approx$5.5%). Under the per-cell constraint, the byte-matched claim still holds because the sampled method uses slightly more memory than the sketch, a conservative bias in favour of the summary side. The $\approx$5.5% gap is measured on the evicted-slot attachment alone. The pre-registered ceiling of Appendix A.3 is stated on the whole persistent ledger, which also includes the retained core; on that basis all eight cells are inside the 512 B ceiling, as Table 5 reports. The control for the mixture configuration (the third round, Section 6.2) is 34 sampled tokens $\times$ 512 B $=$ 17408 B; the mixture itself is 17 tokens plus a 16-tile moment summary, $8704 + 8256 = 16960$ B, a gap of 448 B ($\approx$2.6%).

C.4 Where the Compression-Ratio Labels Come From

The compression-ratio labels of the main compression-regime results and Figure 4 are per-cell medians of the frozen byte ledger, on the payload-only main-ledger basis. The metadata-inclusive effective ratios of the per-channel quantization configurations use a different basis and are reported in Appendix F. The Qwen-side intervals arise from dynamic resolution. Different images have different token counts, so at a fixed ρ the compression ratio varies with the token count. LLaVA-1.5 emits a fixed 576 visual tokens per image, so its column has no interval at all. VRSBench is carved out separately in the last two rows because it sits at the low end of that spread. Its images carry a median of 324 visual tokens against 580–744 on the other three Qwen tasks, so at $\rho = 0.2$ the retained core is a median of 64 columns and the payload — 17 sampled tokens, or the 16 columns of a moment summary — is 25% of it rather than the 11–14% elsewhere. The payload is additive and does not shrink with the image, which is what pulls the ratio down; the same arithmetic gives 9.8× instead of 12.6–13.8× at $\rho = 0.05$.

D The σ^2 Attribution (the Dispersion Correction in Measured Form)

D.1 Configurations and Statistics

The probe covers the single sample that drives the Qwen2.5-VL TextVQA monotone-repair failure (sample id 1196), measured over all 28 layers at KV head 0, with the last prompt token's key serving as a proxy for the decode-time query distribution. Head 0 is a convention, not a restriction, as the earlier single-layer probe showed that the phenomenon is head-wide. For each configuration, the probe visits $28 \times K$ tiles, giving the denominators 28, 112 and 448 below. A tile counts as "flipped" when its $\sigma^2/2$ term alone moves that tile from a marginal contribution ($<$1% of the attention weight) to a dominant one ($>$50% of the attention weight). The mean term points in the correct direction, but the variance term overrides it. The mean shift $qk - \mathrm{lse}_{\mathrm{core}}$, which a rank-0 readout uses, is negative throughout, with 90th-percentile values of -8.3 to -8.6. Mean matching alone would therefore imply that the deleted mass is negligible ($r \approx e^{-6}$). The rank-> 0 readout multiplies by $e^{\sigma^2/2}$,

Table 9: Dispersion correction by sketch setting.

Configuration (K, ℓ)	$\sigma^2/2$ p90	$\sigma^2/2$ max	mean shift p90 / max	σ-driven flips
$(1, 10)$	23.8	62.0	-8.6 / -6.6	4/28
$(4, 2)$	27.9	**3831.1**	-8.4 / -6.3	11/112
$(16, 0)$	0 (by construction)	0	-8.3 / -6.0	0/448

where $\sigma^2/2$ has a 90th percentile of 24–28 and a maximum of 3831 under $(4, 2)$. This factor flips 11 of 112 tiles from negligible to dominant. Because $\sigma^2/2$ is a log-space quantity, the multiplicative factor at the maximum is e^{3831}. The correction is thus not merely large but numerically extreme, with 2 of the 112 tiles reaching overflow level.

D.2 Mechanism reading

The maximum $\sigma^2/2$ under $(4, 2)$ (3831) is far larger than the maximum under $(1, 10)$ (62), for two reasons.

- Under $(4, 2)$ the $K = 4$ tiles are small ($\approx$144 tokens per tile), so the logit outlier of a single hotspot token drives that tile's variance very large; the rank-2 low-order moments capture the direction of the variance but place no constraint on its magnitude.
- Under $(1, 10)$ the $K = 1$ tile is global (576 tokens), so a hotspot is diluted in the global mean; the rank-10 higher-order moments give a better spectral approximation, but they too cannot recover the positional information that the mean has diluted away.

The probed sample is a text-dense TextVQA image whose answer appears in a high-attention text instance, and the flipped tiles concentrate on exactly such regions. As Proposition 3 predicts, on heavy-tailed tiles the divergence of the dispersion correction is structural, not an implementation defect. This probe covers one sample by design and identifies a mechanism. The population-level consequence is reported in the four-cell ladder of Section 6.2, where every configuration fails its pre-registered criterion.

E TTFT Amortization and Peak Memory

E.1 Measurement protocol

On an H100 we measure time-to-first-token (TTFT) and peak KV memory for both models under three prefix configurations.

- *full-prefix execution*: every question prefills the full visual KV (no prefix caching);
- *full cache*: the full visual KV cache is reused (prefix caching, uncompressed);
- *compressed cache*: the compressed visual KV cache is reused (prefix caching, $\rho = 0.2$ plus 17 sampled tokens).

There are $n = 48$ queries, not 48 images: the manifest carries one question per image, so eight question texts are reused across the first six unique images, and reusing one image prefix across its eight questions is exactly what the amortization measures. The construction is recorded in the source file's image–question mapping. This is a pure latency-and-memory microbenchmark; no answer is scored or consumed. TTFT is wall-clock time through the first-token logits, so each timed region runs the image and question prefill and performs no decode step. The TTFT figures are means over the 48 queries; the peak KV figures are the maximum over them.

E.2 Reading

With one image and many questions, a single compressed prefill is amortized over N questions, and each question's marginal TTFT is only the latency of prefilling the question segment. The KV sequence being scanned is shorter, so this latency falls even below the TTFT of the uncompressed

Table 10: First asymmetric-round configurations and isolated variables.

Configuration	Grouping convention	Variable isolated
per-channel INT4 quantization / *INT2 configuration*	*Exactly* the same groups as the symmetric configuration (rows of 128 dimensions), with only min/max asymmetry and a zero-point added (16 levels at INT4, 4 levels at INT2)	Symmetry alone
per-channel INT4 quantization / *INT2 configuration*	KIVI-faithful (Liu et al., 2024b): K asymmetric per channel (one group per 32 tokens along the visual-token order), V asymmetric per token (one group per 32 channels); no residual recent window, since a static prefix has no decode-order arrivals (deviation disclosed in Section F.5)	Grouping axis and symmetry change together (the quantizer a reviewer has in mind)

Table 11: First asymmetric-round adjudication questions and verdicts.

Question	Verdict	Count (threshold ± 0.02)
Question 1 (per-channel quantization vs symmetric)	**NO_RESCUE_ROW**	INT4: 0/4 interval lower bounds above $+0.02$, and 4/4 interval upper bounds at or below $+0.02$. INT2: 1/4 $\rightarrow$ MIXED. The headline statement takes the more severe of the two.
Question 2 (asymmetric K/V quantization vs symmetric)	**RESCUE_KIVI**	INT4: 2/4 interval lower bounds above $+0.02$ (both Qwen cells entirely positive). INT2: 4/4.
Question 3 (does the $4\times$ row flip?)	**FLIP**	The best asymmetric configuration is per-channel INT4 quantization in every cell; 3/4 cells have an interval lower bound above 0.

cache. *Correction.* An earlier draft of Section 1 and of this paragraph quoted the memory range as 2.6–3.4×. The lower endpoint came from dividing integer-rounded MiB (21 against 8) in an internal report; the frozen aggregate carries the unrounded values 20.890625 and 7.6015625 MiB, whose ratio is 2.75×. Both endpoints quoted here are now the frozen ratios, $313.5/91.5 = 3.4\times$ and $20.890625/7.6015625 = 2.75\times$. No measurement changed. Qwen2.5-VL's absolute peak KV is far smaller than LLaVA-1.5's (20.9 against 313.5 MiB) because of grouped-query attention (4 KV heads against LLaVA-1.5's 32 MHA heads) and dynamic-resolution small images.

F Asymmetric Quantization and KIVI-Style Grouping

The quantization rows of the main compression-regime results and Figure 4 use per-row symmetric absmax, with one group per (layer, KV head, token) row of 128 dimensions. This appendix examines whether that convention is a straw man for the whole quantization family and how the family's performance changes under a different convention. This comparison was pre-registered before evaluation. Its adjudication and the prior comparison records were preserved unchanged; the symmetric and retained-token references are reused without rerunning them.

Design and scope. This round covers four cells (LLaVA-1.5-7B and Qwen2.5-VL-7B × TextVQA and ChartQA) at $n = 150$, using the same manifest and seed-42 sample order as the frozen baseline round, paired per sample. DocVQA and VRSBench were *not* run in this round (Section F.5).

Prespecified adjudication. The following frozen wordings were pre-registered under the registered analysis plan and are reproduced verbatim.

F.1 Per-Cell Numbers ($n = 150$, Absolute Scores)

The following paired differences cover the same four cells at $n = 150$ (paired bootstrap, 2000 resamples, seed 42). Each contrast occupies two rows: its paired mean difference and its interval. The row labels use the pre-registered question numbers of Section F. *retained tokens*(16,0) is the code identifier of the sampled-token retention reference at $\rho = 0.2$, the '$4.4\times$ sampled tokens' row

Table 12: First asymmetric-round configuration means.

		INT4 (nominal 4×)			INT2 (nominal 8×)			References	
Cell	Full	sym	per-ch.	asym. K/V	sym	per-ch.	asym. K/V	Pruning 5×	Sampled 4.4×
LLaVA-1.5-7B–TextVQA	0.633	0.626	0.626	**0.634**	0.256	0.525	**0.663**	0.436	0.455
LLaVA-1.5-7B–ChartQA	0.200	**0.207**	0.193	**0.207**	0.080	0.140	**0.193**	0.160	0.160
Qwen2.5-VL-7B–TextVQA	0.873	0.536	0.000	**0.886**	0.000	0.000	**0.867**	0.776	0.785
Qwen2.5-VL-7B–ChartQA	0.793	0.527	0.000	**0.793**	0.000	0.000	**0.807**	0.460	0.547

Table 13: Paired differences (mean; 95% CI).

Contrast	LLaVA–Text	LLaVA–Chart	Qwen–Text	Qwen–Chart
	$+0.0002$	-0.0133	$\mathbf{-0.5365}$	$\mathbf{-0.5267}$
Q1 PC INT4 − symmetric	$[-0.019, +0.019]$ $+0.2687$	$[-0.040, +0.013]$ $+0.0600$	$[\mathbf{-0.612}, \mathbf{-0.457}]$ 0.0000	$[\mathbf{-0.607}, \mathbf{-0.447}]$ 0.0000
Q1 PC INT2 − symmetric	$[+0.194, +0.345]$ $+0.0081$	$[+0.020, +0.107]$ 0.0000	$[0, 0]$ $\mathbf{+0.3494}$	$[0, 0]$ $\mathbf{+0.2667}$
Q2 asymmetric K/V INT4 − symmetric	$[-0.012, +0.029]$ $+0.4072$	$[-0.020, +0.020]$ $+0.1133$	$[\mathbf{+0.279}, \mathbf{+0.428}]$ $+0.8673$	$[\mathbf{+0.187}, \mathbf{+0.347}]$ $+0.8067$
Q2 asymmetric K/V INT2 − symmetric	$[+0.331, +0.479]$ $+0.1785$	$[+0.060, +0.167]$ $+0.0467$	$[+0.814, +0.915]$ $+0.1007$	$[+0.747, +0.873]$ $+0.2467$
Q3 asymmetric K/V INT4 − retained tokens (16,0)	$[+0.109, +0.251]$	$[-0.007, +0.100]$	$[+0.050, +0.154]$	$[+0.167, +0.327]$

Table 14: INT4 attribution: scaling versus grouping.

Contrast	Element-wise relative error (vs symmetric)	Qwen downstream Δ
per-channel quantization INT4 (grouping fixed, only symmetry switched)	↓: 0.169 → 0.136, 0.167 → 0.134 (LLaVA 0.197 → 0.144, 0.198 → 0.145)	**intervals entirely negative** (−0.537 / −0.527)
asymmetric K/V quantization INT4 (grouping axis switched too)	↓↓↓: 0.069 / 0.068 (LLaVA 0.068 / 0.068)	**intervals entirely positive** (+0.349 / +0.267)

of the main compression-regime results. Bold marks the four intervals carrying Q1 and Q2's pre-registered INT4 adjudications: the two entirely negative per-row asymmetric INT4 differences on Qwen and the two entirely positive per-channel INT4 differences. It is not a general significance marker: 13 of the 20 intervals clear zero, and the remaining rows are descriptive.

F.2 Mechanism: the Load-Bearing Variable Is the Grouping Axis

Four readings.

1. **The attribution goes to the grouping axis, not to symmetry.** Only the move to KIVI's per-channel K and per-token V grouping rescues Qwen's INT4. At the same grouping, switching to asymmetric scaling alone fails to rescue it. Instead, Qwen's score collapses from 0.536 to 0.000. The zero-point is therefore not the load-bearing structure; *the per-channel dynamic-range structure is*. Section F.7 further identifies *K's* per-channel dynamic range as the load-bearing structure and shows that re-grouping V contributes zero, per sample.

2. **Error norm and downstream quality decouple for the third time** (it extends the correspondence table of Proposition 2 by one row, in the frozen wording "error-norm/downstream decoupling holds under asymmetric scaling"). The first two were the same error norm with downstream scores differing by 0.27–0.34 across the two model families (Section 5), and rotated INT4 with a 33% lower error norm on Qwen, where quality falls $0.54 \rightarrow 0.01$ (Section 4, Figure 3d). Here per-channel quantization lowers the error by about 20% while Qwen's downstream score moves by -0.53. This follows the same pattern as the first two. The causal quantity is not how *large* the error is but *which coordinates* it lands on.

Table 15: Group-metadata overhead.

Config.	Metadata / payload	Nominal	Effective
Per-channel INT4	6.2%	4×	3.76×
Asymmetric K/V INT4	25.0–25.5%	4×	**3.19–3.20**×
Asymmetric K/V INT2	50.0–51.1%	8×	**5.30–5.33**×

3. **The INT2 cliff is partly an artifact of per-row grouping.** *INT2 configuration* pulls Qwen's INT2 back from 0.000 to 0.867 and 0.807 (both intervals entirely positive), with $+0.407$ and $+0.113$ on the LLaVA side. The $8\times$ statement of Section 5 that pruning beats INT2 is therefore pinned to the per-row symmetric INT2 convention (Section F.4).
4. **What this design does not separate.** The asymmetric K/V quantization configuration changes the grouping of K and of V together, so this design *cannot* say which half carries the rescue. The ablations that change K alone or V alone were not run in this round and were disclosed as an open item. Section F.7 closes this item. The 2×2 grouping factorial adjudicates the rescue as carried by K, with Δ_K recovering $+0.88$ and $+0.79$ on the two Qwen cells while Δ_V is zero, per sample.

F.3 Group-Metadata Overhead and the Two Ledgers (with an Erratum)

The payload-only accounting's byte-counting formula is word-for-word identical to the frozen baseline round's and counts payload only. Scale and zero-point bytes stay outside it and are disclosed separately. The following values are recomputed from the frozen fields (scale bytes + zero-point bytes against the quantized payload).

Erratum. The original record for the round behind Section F.1 reported $\approx 82\%$ metadata overhead for per-channel INT4 using the wrong payload denominator. That figure is void and appears nowhere else in the paper.

Recomputation from the frozen fields gives 25.0% metadata for asymmetric K/V INT4 and 50.0% for INT2; all results use these values. Metadata grows as bit width falls, explaining why effective quantization compression saturates near $\approx 5.3\times$. Across cells, asymmetric K/V INT4 is reported as 3.17–$3.20\times$ effective (derivation in Section F.7.3).

Reading the two ledgers. Figure 4 reports payload-only ratios ($4\times$ and $8\times$). The asymmetric configurations require metadata-inclusive accounting ($3.2\times$ and $5.3\times$), so the apparent $4\times$ reversal is not a byte-matched win. Per-channel INT4's effective ratio is below sampled-token retention's $\approx 4.4\times$, trading more memory for higher score.

F.4 Consequences for the Main Text

The main text incorporates this appendix's adjudications as follows.

1. *The* $4\times$ *row.* The claim is pinned to per-row symmetric INT4, and carries the scaling-scheme-dependence sentence (the frozen wording of Section F) together with a ledger-basis note.
2. *The* $8\times$ *row.* The claim is pinned to per-row symmetric INT2. The *INT2 configuration* reversal is reported as an *exploratory observation* across four cells, with no pre-registered adjudication rule and an explicit post-hoc label.
3. *The* $\geq 10\times$ *row.* This row is unaffected and gains a structural argument. The quantization family's effective ratio saturates near $\approx 5.3\times$ and cannot enter that bracket. Reaching $\geq 10\times$ therefore requires eviction, which forces the question of what the deleted bytes should buy.
4. *The central verdict.* Unaffected, in the frozen wording: quantization acts on retained bytes, and the comparison that allocates the evicted bytes is made on the same ledger. The tokens-versus-summaries counts of 8/8, 8/8 and 4/4 are unchanged.

Table 16: Quantization configurations on four additional cells ($n = 150$).

		Per-row sym*		Per-row per-channel		Per-channel asym. K/V		
Cell	Full	INT4	INT2	INT4	INT2	INT4	INT2	*retained tokens*(16,0)*
LLaVA-1.5-7B–DocVQA	0.172	0.185	0.021	**0.192**	0.101	0.169	0.158	0.075
Qwen2.5-VL-7B–DocVQA	0.925	0.691	0.000	0.004	0.000	**0.918**	0.906	0.427
LLaVA-1.5-7B–VRSBench	0.439	**0.446**	0.385	**0.446**	0.428	0.439	**0.446**	0.448
Qwen2.5-VL-7B–VRSBench	0.425	0.294	0.000	0.007	0.000	**0.419**	0.410	0.414

5. *Figure 4.* Not redrawn; the caption carries the scaling-scheme note.
6. *The correspondence table of Proposition 2.* Extended by an eleventh row (Appendix B).

F.5 Conservativeness and Disclosed Deviations

- Asymmetric INT2 uses four levels, while symmetric INT2 is ternary $\{-1, 0, 1\}$ with three. The asymmetric configuration is therefore *strictly stronger* at INT2. That per-channel quantization is still not enough (Qwen INT2 stays at 0.000) therefore runs in the direction that favours this paper's claim, and the asymmetric K/V quantization rescue is disclosed together with this conservative feature.
- The residual full-precision recent window of the KIVI paper (Liu et al., 2024b) does not apply in a prefix-only setting because the visual KV is a static prefix with no decode-order arrivals. It was not implemented, and the deviation is disclosed here. Had that window been available it could only have made the KIVI configuration stronger, in the same direction as above.
- This round covered only 4 of the 8 cells. Neither *docvqa* (the cell where quantization reverses under the symmetric convention) nor *vrsbench* ran an asymmetric configuration in this round, so the coverage–fidelity trade-off was untested under KIVI grouping and the main text's claim about that axis stays on the per-row symmetric convention. Section F.6 discharges this limitation by completing those four cells in the map-completion round. The map's quantization rows go from 4/8 to 8/8, and the axis is adjudicated as surviving, but with its shape changed to 'quantization of the coverage type wins, the precision type ties' (Sections J.8.2 and J.8.3).
- The smoke run at $n = 4$ passed all five implementation-integrity gates. The symmetric mode agrees value for value with the frozen baseline round's symmetric-quantization reference, and the hard floor is $0.7292 \geq 0.70$. The byte fields equal the frozen formula per sample, the asymmetric K/V quantization group count equals the ceil formula, and every dequantized value is finite.

F.6 The Map-Completion Round: the Quantization Family Fills In the Map ($4/8 \to 8/8$) and the Fate of the Information-Density Axis

This pre-registered comparison covers the four remaining model–benchmark cells at $n = 150$ and $\rho = 1.0$ (pure quantization, with no eviction). It contrasts the full cache, symmetric row-wise INT4 and INT2, and two asymmetric K/V grouping variants. The matched retained-token reference is reused unchanged from the main matrix.

F.6.1 The Quantization Rows Reach 8/8 (GENERALIZES)

Rule (pre-registered, executed mechanically): *RECOVERED* $\Leftrightarrow$ CI-lower(per-channel INT4 quantization $-$ *full cache*) ≥ -0.05; *GENERALIZES* $\Leftrightarrow \geq 3/4$ cells *RECOVERED*. 4/4 $\to$ **GENERALIZES**. Joined with the four cells of Section F.1, KIVI-style per-channel grouping recovers full-cache quality at INT4 in **8/8 cells**. With the same figure and the same group structure, the per-row asymmetric configuration still collapses outright on the two Qwen grouped-query cells (INT4 at 0.004 and 0.007, INT2 at 0.000 and 0.000), and the per-row symmetric configuration reaches 0.691

Table 17: Per-channel INT4 versus the full cache. All cells recovered; entries give Δ [95% CI] and identical-score count.

Cell	Result
LLaVA–DocVQA	$-0.0033\,[-0.0202, +0.0113]$; 147/150
Qwen–DocVQA	$-0.0066\,[-0.0257, +0.0062]$; 145/150
LLaVA–VRSBench	$+0.0000\,[+0.0000, +0.0000]$; **150/150**
Qwen–VRSBench	$-0.0054\,[-0.0308, +0.0192]$; 144/150

Table 18: Per-channel INT4 versus sampled-token retention at $\rho = 0.2$.

Cell	Δ_B [95% CI]	Interval straddles zero
LLaVA-1.5-7B–DocVQA	$+0.0935\,[+0.0396, +0.1493]$	no
Qwen2.5-VL-7B–DocVQA	$+0.4913\,[+0.4216, +0.5622]$	no
LLaVA-1.5-7B–VRSBench	$-0.0090\,[-0.0585, +0.0387]$	yes
Qwen2.5-VL-7B–VRSBench	$+0.0055\,[-0.0398, +0.0495]$	yes

against a full cache of 0.925 on Qwen2.5-VL-7B–DocVQA and 0.294 against 0.425 on Qwen2.5-VL-7B–VRSBench. Both quantization rows of the map are thereby confirmed as artifacts of per-row grouping, not properties of the quantization family itself. Frozen wording (the registered 8/8 conclusion, consumed verbatim, $\{k\} = 4$): **The ledger note that accompanies that sentence at every citation.** The 8/8 is *not* a matched-memory win. The per-channel INT4 quantization configuration carries 25.0–26.1% group metadata across the eight cells, giving an effective ratio of 3.17–3.20×, against $\approx$4.4× for the sampled method. The two-ledger reading of Section F.3 applies word for word to every citation in this section.

F.6.2 The Information-Density Axis Survives, but Changes Shape (AXIS_SURVIVES)

Statistic (pre-registered, verbatim): Δ_B = per-channel INT4 quantization − *retained tokens*(16,0) at $\rho = 0.2$, paired by each matched example, with the intersection of n disclosed (150/150 in all four cells). The interval lower bound is above 0 in 2/4 cells, and the interval straddles zero in 2/4. The pooled DocVQA mean of $+0.2924$ and the pooled VRSBench mean of -0.0017 carry opposite signs, which gives the mechanical verdict **AXIS_SURVIVES**. The pre-registered rule requires opposite signs in the two dataset pools, or $\geq 3/4$ cells with intervals straddling zero. **Correction.** In the original report record on which Section F.6 is based, the header of the Δ_B table read "kivi_int4 − asym_row_int4", which does not match the numbers beneath it. A per-sample recomputation confirms that those numbers are the pre-registered Δ_B (LLaVA-1.5-7B–DocVQA example: asymmetric K/V quantization 0.1686 − *retained tokens*(16,0) 0.0751 = $+0.0935$, which agrees; whereas asymmetric K/V quantization − per-channel quantization = $0.1686 - 0.1923 = -0.0237$, which does not). The computation was correct and only the label was wrong; this appendix and the main text consume the numbers as Δ_B, and nothing was re-run. **Correction.** The frozen wording of the AXIS_SURVIVES branch says that "per-channel quantization wins on the coverage cell and **sampled tokens win on the precision cell**". The data do not support the second half. The Δ_B intervals of both VRSBench cells *straddle zero* ($-0.0090\,[-0.0585, +0.0387]$ and $+0.0055\,[-0.0398, +0.0495]$), indicating a tie rather than a win for sampling. The verdict label and counts are unchanged. The main text gives the following accurate reading. The correction *tightens* a statement that had been favourable to this paper's central verdict by giving up one claim that sampling wins. The thresholds and counts stay untouched. This is a mechanical correction permitted by the prespecified decision-rule clause.

F.6.3 Consequences for the Main Text

1. The third limitation of Section F.5 was that DocVQA and VRSBench had not run an asymmetric configuration, leaving the coverage–fidelity trade-off untested under KIVI grouping. This limitation is now discharged, and Section 7 no longer carries it.

Table 19: The 2×2 K/V grouping factorial under asymmetric min–max scaling.

	V whole row (1 group/token)	V in groups of 32 channels (4 groups/token)
K whole row (1 group/token)	per-channel quantization = per-channel quantization (read-only)	**per-channel quantization** (new)
K per channel (ceil(n_q/32) groups per channel)	**per-channel quantization** (new)	per-channel quantization = asymmetric K/V quantization (read-only)

Table 20: Paired K/V grouping effects (mean [95% CI]); bold K entries rescue quality.

Cell / width	Δ_K	Δ_V
Qwen–TextVQA / INT4	$+\mathbf{0.8799}$ $[+\mathbf{0.8281}, +\mathbf{0.9294}]$	$+0.0000$ $[+0.0000, +0.0000]$
Qwen–ChartQA / INT4	$+\mathbf{0.7933}$ $[+\mathbf{0.7267}, +\mathbf{0.8600}]$	$+0.0000$ $[+0.0000, +0.0000]$
LLaVA–TextVQA (negative control) / INT4	$+0.0066$ $[-0.0023, +0.0189]$	$+0.0059$ $[-0.0051, +0.0220]$
Qwen–TextVQA / INT2	$+\mathbf{0.8819}$ $[+\mathbf{0.8297}, +\mathbf{0.9285}]$	$+0.0050$ $[+0.0000, +0.0150]$
Qwen–ChartQA / INT2	$+\mathbf{0.7933}$ $[+\mathbf{0.7333}, +\mathbf{0.8533}]$	$+0.0000$ $[+0.0000, +0.0000]$
LLaVA–TextVQA / INT2	$+\mathbf{0.1431}$ $[+\mathbf{0.0819}, +\mathbf{0.2126}]$	$+0.0059$ $[-0.0317, +0.0425]$

2. The coverage–fidelity trade-off now has two measured shapes, each stated under its own convention and never merged into a single summary claim. *Under per-row symmetric grouping*, as used in Figure 4 and the main compression-regime results, quantization and sampled retention divide labour by task type, as Section 5 originally reported. *Under per-channel grouping*, as tested in this section, quantization wins on the coverage task and ties on the precision task.
3. Figure 4 is not redrawn (item 5 of Section F.4). The caption's scaling-scheme note gains one sentence stating that the per-channel configurations also recover on DocVQA and VRS-Bench (8/8), at an effective ratio of $\approx 3.2\times$, which is not equal memory.

F.7 The 2×2 Grouping-Factorial Round Identifies K as the Load-Bearing Axis

This round closes open item 4 of Section F.2. The asymmetric K/V quantization configuration changes the grouping of K and of V together, so the earlier design cannot say which half carries the rescue. Here K and V each take two grouping levels, and all four cells of the factorial use *asymmetric min-max* scaling. The scaling rule is therefore held fixed, and only the grouping axes vary. This avoids reintroducing the symmetry variable already excluded in the first asymmetric round. The scope is 3 cells $\times$ $n = 150 \times 2$ new configurations $\times$ 2 bit widths, with 750 records per cell. The selected cells are Qwen2.5-VL-7B–TextVQA and Qwen2.5-VL-7B–ChartQA, where the effect occurs, plus LLaVA-1.5-7B–TextVQA as the *negative control* that carried no effect in the first asymmetric round.

F.7.1 Adjudication: K_DRIVEN

Rule: Δ_K = per-channel quantization $-$ per-channel quantization (the K axis, V fixed); Δ_V = per-channel quantization $-$ per-channel quantization (the V axis, K fixed); *rescues* $\Leftrightarrow$ CI-lower $> +0.02$; the K-driven conclusion $\Leftrightarrow$ Δ_K rescues in $\geq 2/3$ cells and Δ_V does not. INT4 is the primary adjudication, with INT2 reported alongside. INT4: Δ_K rescues 2/3 (both Qwen cells) and Δ_V rescues none $\rightarrow$ **K_DRIVEN** (per-channel INT4 quantization). Frozen wording (the registered K-driven conclusion, consumed verbatim, with the placeholders filled from Table 20):

Naming. The formal prose name of the per-channel quantization configuration is *K per-channel with V per-token-row*. It is neither a subset implementation nor an equivalent rewrite of KIVI. It carries no residual recent window and shares only K's per-channel grouping with KIVI. It is therefore an independent grouping form rather than a KIVI variant. The two zero-width intervals for Δ_V reflect *per-sample identity*, not degenerate records. On Qwen2.5-VL-7B–TextVQA and Qwen2.5-VL-7B–ChartQA the scores of per-channel INT4 quantization and per-channel INT4 quantization

Table 21: K/V grouping factorial at INT4 ($n = 150$).

Cell	Full	K row V row*	K row V groups	K/channel V row	K/channel V groups*
Qwen2.5-VL-7B–TextVQA	0.873	0.000	0.880	0.000	**0.886**
Qwen2.5-VL-7B–ChartQA	0.793	0.000	**0.793**	0.000	**0.793**
LLaVA-1.5-7B–TextVQA	0.633	0.626	0.632	0.632	**0.634**

Table 22: K/V grouping factorial at INT2 ($n = 150$).

Cell	Full	K row V row*	K row V groups	K/channel V row	K/channel V groups*
Qwen2.5-VL-7B–TextVQA	0.873	0.000	**0.882**	0.005	0.867
Qwen2.5-VL-7B–ChartQA	0.793	0.000	0.793	0.000	**0.807**
LLaVA-1.5-7B–TextVQA	0.633	0.525	**0.668**	0.531	0.663

Table 23: INT4 error and downstream difference versus full cache (mean relative error; Δ [95% CI]; PC = per-channel).

Cell	Config.	Error	Downstream Δ
Qwen–TextVQA	PC	0.0685	$+0.0068\,[-0.0073, +0.0240]$
Qwen–TextVQA	PC	0.0968	$\mathbf{-0.8731}\,[\mathbf{-0.9200}, \mathbf{-0.8227}]$
Qwen–TextVQA	PC (round 1)	0.1357	$-0.8731\,[-0.9200, -0.8227]$
Qwen–TextVQA	Asym. K/V (round 1)	0.0694	$+0.0128\,[-0.0010, +0.0328]$
Qwen–ChartQA	PC	0.0658	$+0.0000\,[+0.0000, +0.0000]$
Qwen–ChartQA	PC	0.0963	$\mathbf{-0.7933}\,[\mathbf{-0.8600}, \mathbf{-0.7267}]$
Qwen–ChartQA	PC (round 1)	0.1344	$-0.7933\,[-0.8600, -0.7267]$
Qwen–ChartQA	Asym. K/V (round 1)	0.0676	$+0.0000\,[+0.0000, +0.0000]$
LLaVA–TextVQA	PC	0.0658	$-0.0006\,[-0.0050, +0.0036]$
LLaVA–TextVQA	PC	0.1001	$-0.0013\,[-0.0179, +0.0164]$
LLaVA–TextVQA	PC (round 1)	0.1444	$-0.0072\,[-0.0195, +0.0020]$
LLaVA–TextVQA	Asym. K/V (round 1)	0.0676	$+0.0007\,[-0.0041, +0.0060]$

are exactly the same on 150/150 samples (both configurations collapse to 0.000). Re-grouping V yields exactly zero gain on Qwen. The reverse control is equally clear. The scores of per-channel INT4 quantization and per-channel INT4 quantization are per-sample identical on 150/150 samples of Qwen2.5-VL-7B–ChartQA and on 145/150 of LLaVA-1.5-7B–TextVQA. Changing K alone delivers the whole effect of changing K and V together.

F.7.2 Mechanism: a Fourth Decoupling of Error Norm from Downstream Quality (Row 12 of the Proposition 2 Table)

The error metric is mean element-wise relative error; the final column is the paired difference against *full cache*. Configurations marked "first round" are the read-only references of Section F.1. Two entries of the error-norm column coincide across cells at four decimals: per-channel quantization reads 0.0658 on both Qwen ChartQA and LLaVA TextVQA (the underlying means are 0.065775 and 0.065830), and asymmetric K/V quantization reads 0.0676 on both (0.067607 and 0.067612). Each is a mean over its own cell's 150 samples; the agreement is rounding, not a repeated row.

Within this 2×2 factorial, the error norm is *neither sufficient nor necessary*. Compressing the error from 0.1357 to 0.0968 (-29%, per-channel quantization $\rightarrow$ per-channel quantization) leaves downstream quality unchanged. Both Qwen cells stay at 0.000, per-sample identical. Meanwhile 0.0685 (per-channel quantization) and 0.0694 (asymmetric K/V quantization) are all but indistinguishable numerically, yet both recover downstream to the full-cache level. The causal quantity is not how *large* the error is but *which coordinates of which tensor* it lands on. Error that lands along K's per-channel direction is absorbed, while error that lands on V's channel grouping amounts to no repair at all. This is the fourth measured decoupling. The first three were the same error norm with different downstream quality across the two families, rotated INT4 with a 33% lower error norm on Qwen, where quality goes $0.54 \rightarrow 0.01$, and per-channel quantization with an error about 20% lower while

Table 24: Group-metadata fraction at INT4.

Cell (mean n_q)	K row V row	K row V groups	K/channel V row	K/channel V groups
Qwen2.5-VL-7B–TextVQA (742.5)	0.158 (formula 0.161)	0.156 (0.156)	0.062 (0.063)	0.251 (0.254)
Qwen2.5-VL-7B–ChartQA (516.5)	0.163 (0.163)	0.156 (0.156)	0.062 (0.063)	0.257 (0.257)
LLaVA-1.5-7B–TextVQA (576.0)	0.156 (0.156)	0.156 (0.156)	0.062 (0.062)	0.250 (0.250)

Qwen moves by -0.53. These are rows 9, 6 and 11 of Appendix B. The current result extends the correspondence table of Proposition 2 to a twelfth row.

F.7.3 Group Metadata: the Cheapest Sufficient Fix Is Not KIVI

The metadata fraction is $f = (\text{scale} + \text{zero})/\text{INT payload}$. Table 24 compares its per-sample recomputation with the frozen formula.

Every absolute discrepancy is at most 0.4 percentage points, and the verification exception list is empty.

At INT4, grouping K per channel while retaining per-token-row V has the smallest metadata fraction (6.2%) and an effective $3.76\times$ ratio. It recovers full-cache quality on both Qwen cells. Grouping both K and V per channel instead costs 25.0–25.7% metadata and yields only 3.17–$3.20\times$ effective compression, without a material quality advantage. Thus K's grouping, rather than V's, supplies the sufficient repair.

This conclusion is restricted to INT4. The leanest sufficient INT4 configuration remains near 3.4–$3.8\times$ effective compression and cannot reach the $\geq 10\times$ regime, where eviction is necessary. INT2 uses a separate ledger and reaches approximately $5.3\times$ under asymmetric K/V grouping; it does not alter the INT4 comparison.

F.7.4 Negative Control and Conservativeness

LLaVA-1.5-7B–TextVQA is the zero-effect cell of the first asymmetric round (per-channel quantization 0.626 against a full cache of 0.633). At INT4, $\Delta_K = +0.0066\,[-0.0023, +0.0189]$ and $\Delta_V = +0.0059\,[-0.0051, +0.0220]$, neither of them significant. The negative control stays negative by design, and the 2×2 does not manufacture an effect in a cell that has none. At INT2, Δ_K rises to $+0.1431$ (interval entirely positive), because *INT2 configuration* falls to 0.525 in this cell while *INT2 configuration* reaches 0.668, above *INT2 configuration*'s 0.663. Re-grouping V is therefore slightly harmful here. This is the K-axis effect becoming visible only once the bit width drops; it is reported as a secondary INT2 result and does not enter the INT4 primary adjudication. Conservativeness carries over from Section F.5: asymmetric INT2 uses four levels where symmetric INT2 is ternary with three, and the KIVI paper's residual recent window is not implemented in a prefix-only setting. All four cells of this factorial use asymmetric scaling, so these conservative features do not affect the *comparison* between Δ_K and Δ_V, whose configurations share the same scaling rule. They bear only on cross-convention comparisons with the symmetric configuration.

F.7.5 Dated Revision (Zero-Adjudication Clause)

Checker correction. The first smoke run of this round produced four "byte formula mismatch (asym_row)" FAILs per cell. The diagnosis was an arithmetic error in the *smoke checker itself*. A multiplication by 2 was duplicated because the byte-count calculation already returns bytes and the caller multiplied by 2 again. The data were never wrong. This round's per-channel quantization records agree field for field with the frozen artifacts of the first asymmetric round (LLaVA-1.5-7B–TextVQA, $n_q = 576$: the K/V payload is 1,179,648; scale = zero = 2,359,296). With the checker fixed, the smoke run was re-run and passed on 3/3 cells (degenerate-case tests at bits $= 4/2$, plus the group formula, the byte formula and the parity gate against the frozen first-round artifacts: 12/12 records). This is a checker fix, not a formula changed in order to pass a gate: the failing check contradicted the very first-round data it was meant to verify. The correction's sha256 digest is recorded in the verification materials.

F.7.6 Consequences for the Main Text

1. *The attribution subsection of Section 5, second layer.* It now reads that the dominant factor is *K's* grouping axis rather than the grouping axis as such: re-grouping V carries no part of the rescue (Δ_V is zero per sample on both Qwen cells). Reading 1 of Section F.2 is replaced accordingly, and open item 4 of Section F.2 is closed.
2. *The correspondence table of Proposition 2 (Appendix B).* A twelfth row records that per-channel quantization lowers the error by 29% with no downstream movement, giving the fourth decoupling of this section.
3. *The mechanism sentence is pinned to the K axis.* "KIVI-faithful" stays in the abstract and in the contribution bullets, because the configuration behind the 8/8 really is asymmetric K/V quantization; but wherever the paper explains *why* the rescue happens, it names K's per-channel dynamic range. Where the main text offers the cheapest sufficient fix, it names the configuration as K per-channel with V per-token-row and reports its $3.44\times$ effective ratio.
4. *Figure 4.* Not redrawn; the caption note gains one sentence on the K axis, merged with item 3 of Section F.6.3 into a single note.

G Composition in the Deep-Compression Regime: Quantized Core plus Sampled Tokens

This appendix uses the same source as Appendix F.6. The pre-registration was frozen before the first trace, and the aggregation's sha256 record was re-verified as identical at adjudication. The scope is 8 cells $\times$ $n = 48$ at $\rho = 0.05$, with five new configurations plus the frozen controls from the continuum round (retained-token method at $\rho = 0.05$ and *pruned retained core*, read out of the frozen control records, **read-only and not re-run**). There are 240 records per cell (5×48) and 1920 records in total. This budget is the $\geq 10\times$ row of the main compression-regime results. On the payload-only ledger, the frozen control condition's ratio is $294{,}912/23{,}552 = 12.5\times$ by the reference arithmetic of the round and $12.8\times$ as a per-cell median (Table 8). On the effective ledger, the basis on which this appendix adjudicates, the primary condition's per-cell medians are 9.97–$14.00\times$ (Table 31). These are the same budget on two bases, not two budgets, and the spread within each basis is the dynamic-resolution spread of Appendix C.4. **Ledger basis (mandatory).** The equal-byte constraint of this appendix and every adjudication statistic in it are computed on the **effective ledger**, group metadata included. The payload-only ledger is recorded per sample alongside it, using the byte formula of Appendix C.1 verbatim, as a secondary disclosure. Each statement below names the basis it uses.

G.1 Design and Configuration Definitions

The retained core is $\lfloor 0.05\, n_v \rfloor$ columns: 28 in every LLaVA-1.5 cell (576 visual tokens) and 16 in Qwen VRSBench (324), ranging from 4–38 in the other three Qwen cells as dynamic resolution varies. "Sampled tokens" are uniformly sampled evicted tokens appended to that core; retained-token method is the count solved per sample under the equal-byte constraint. The moment summary of moment-summary method is the *byte-matched moment summary* defined in the terminology note of L.3. It is not the round-three sampling-plus-summary mixture method or the "Mixture ($n = 48$)" row of the main continuum results.

Equal-byte solve. Taking the effective bytes of the frozen retained-token method at $\rho = 0.05$ as the budget, the affordable retained-token method is solved per sample, the quantum being one bf16 token = 512 B. The control condition holds **17 tokens**. The solved means and medians are reported below. On qwen_chartqa the median sits *above* the mean ($36 > 33.0$, and $116 > 106.8$ for retained-token method). The cause is image size, not the solve: 13 of 48 samples carry fewer than the modal 580 visual tokens (80–558), so their equal-byte budget is smaller and they solve 18–35 tokens, pulling the mean below the modal 36 that 27 samples hold. In the other seven cells, the mean and median agree to within one token. Both columns reproduce the frozen aggregation exactly (retained-token method, per cell and configuration).

Table 25: Five configurations of the deep-regime composition round.

Method	Retained core	Sampled tokens		Moment summary	Metadata scope	Role
pruned-cache method	per-channel INT4	none		none	core	attribution baseline (not gap-constrained)
retained-token method	per-channel INT4	bf16 retained-token method	×	none	core	**primary condition**
retained-token method	per-channel INT2	bf16 retained-token method	×	none	core	bit-width secondary condition
retained-token method	per-channel INT4	INT4 retained-token method	×	none	core ∪ sampled (single call)	exploratory condition (L.4)
moment-summary method	per-channel INT4	bf16 retained-token method	×	moment mixture	core	**central-verdict control condition**

Table 26: Retained-token counts (mean / median).

Cell	Core–INT4	Core–INT2	All–INT4
LLaVA: Text/VRS/Chart/Doc	36.0 / 36	39.0 / 39	114.0 / 114
Qwen–TextVQA	40.8 / 41	45.8 / 46	132.9 / 133
Qwen–DocVQA	40.9 / 41	45.9 / 46	132.8 / 133
Qwen–ChartQA	33.0 / 36	36.4 / 40	106.8 / 116
Qwen–VRSBench	27.0 / 27	29.0 / 29	87.0 / 87

Table 27: Δ_C: quantized core plus sampled tokens versus the 17-token control.

Cell	retained-token method	Control	full	Δ_C [95% CI]	CI lower > 0
llava_textvqa	0.283	0.221	0.670	$+0.0625\,[+0.0000, +0.1458]$	no (boundary)
qwen_textvqa	0.598	0.497	0.838	$+0.1017\,[-0.0052, +0.2125]$	no
llava_vrsbench	0.528	0.487	0.508	$+0.0417\,[-0.0417, +0.1250]$	no
qwen_vrsbench	0.324	0.312	0.428	$+0.0114\,[+0.0000, +0.0341]$	no (boundary)
llava_chartqa	0.188	0.125	0.229	$+0.0625\,[+0.0000, +0.1458]$	no (boundary)
qwen_chartqa	0.333	0.312	0.771	$+0.0208\,[-0.0417, +0.1042]$	no
llava_docvqa	0.069	0.061	0.216	$+0.0083\,[-0.0375, +0.0625]$	no
qwen_docvqa	0.143	0.086	0.952	$+0.0569\,[-0.0108, +0.1330]$	no

Replacing the bf16 retained core with per-channel INT4 therefore **buys** 1.6–$2.4\times$ **more sampled tokens** at the same effective budget ($17 \rightarrow 27$–41). Quantizing the sampled payload as well (retained-token method) buys 5.1–$7.8\times$ more ($17 \rightarrow 87$–133), because that method spends only about a quarter as much per token.

G.2 MIXED_C: A Quantized Core plus More Tokens against a bf16 Core with 17 Tokens

Δ_C = retained-token method $-$ the frozen retained-token method at $\rho = 0.05$, on the metadata-inclusive memory accounting, paired on identical examples, $n = 48$. The adjudication runs on the four main cells; the other four are disclosed in full. **MIXED_C**. Of the four adjudicating cells, **0/4** have a CI lower bound above zero at $n = 48$. Neither the DOMINATE nor the NO_DOMINATE branch of the frozen wording applies. The result is therefore reported as the pre-registered MIXED_C branch prescribes, with a per-cell table and one mechanism paragraph rather

Table 28: Sampled tokens versus byte-matched moment summaries. Scores are sampled / summary; Δ_D is paired [95% CI].

Cell	Scores	Δ_D
LLaVA–TextVQA	0.283 / 0.264	$+0.0194\,[-0.0764, +0.1209]$
Qwen–TextVQA	0.598 / 0.493	$+0.1049\,[-0.0527, +0.2582]$
LLaVA–VRSBench	0.528 / 0.434	$\mathbf{+0.0947}\,[\mathbf{+0.0208}, \mathbf{+0.1875}]$
Qwen–VRSBench	0.324 / 0.346	$-0.0225\,[-0.1345, +0.0911]$
LLaVA–ChartQA	0.188 / 0.146	$+0.0417\,[+0.0000, +0.1042]$
Qwen–ChartQA	0.333 / 0.250	$+0.0833\,[-0.0417, +0.2083]$
LLaVA–DocVQA	0.069 / 0.068	$+0.0014\,[-0.0667, +0.0764]$
Qwen–DocVQA	0.143 / 0.129	$+0.0136\,[-0.0703, +0.1023]$

than an umbrella claim. *Mechanism (no generalization).* All 8/8 point estimates are positive at $n = 48$, with magnitudes from $+0.008$ to $+0.102$, and no interval clears zero. The immediate reason is that at $n = 48$ the CI half-widths (± 0.017 to ± 0.109) are of the same order as the point estimates themselves. Thus **the design is under-powered to resolve the effect it observed**, rather than ambiguous about the effect's direction. The split of magnitude by task type (text-dense / semantic) is **not clean**. The six text-dense cells run from $+0.008$ to $+0.102$ (mean $+0.052$), and the two semantic cells are $+0.011$ and $+0.042$ (mean $+0.027$), with heavily overlapping intervals. Magnitude does not track the headroom left between the control and the full cache either (qwen_docvqa has the largest headroom, 0.866, and gains only $+0.057$). The one clean structure is **dose consistency**. On the two cells where 17 sampled tokens already buy nothing over blind eviction (sampled minus blind: -0.004 on qwen_vrsbench and 0.000 on llava_docvqa), raising the count to 27–41 tokens also buys almost nothing ($+0.011$ and $+0.008$); on the five cells where 17 tokens already pay (sampled minus blind from $+0.030$ to $+0.136$), more tokens keep paying ($+0.021$ to $+0.102$). llava_chartqa is the threshold case, where 17 tokens sit exactly at zero (sampled minus blind 0.000) and only 36 tokens cross it (within this cell, retained-token method $-$ pruned-cache method $= +0.042$). This subsection therefore supports only the statement that quantizing the retained core in exchange for more sampled tokens is positive in 8/8 cells and significant in 0/8, at $n = 48$.

G.3 VERDICT_CHALLENGED: Robustness of the Central Verdict under a Quantized Core

Δ_D = retained-token method $-$ moment-summary method. Both methods carry the same per-channel INT4 core and equal effective bytes; $n = 48$; the tie band is $|\Delta| < 0.005$. Rule: VERDICT_HOLDS $\iff$ at least 3/4 of the main cells satisfy CI-lower$(\Delta_D) > 0$ **or** both sides fall inside the tie band; otherwise VERDICT_CHALLENGED. Holding cells: 1/4 (llava_vrsbench only) $\rightarrow$ **VERDICT_CHALLENGED**.

Consequence clause (pre-registered, mandatory). Sampled tokens have higher mean scores than merged tokens in 8/8 cells and match or exceed moment summaries in 8/8 at $n = 150$, and beat the exploratory mixture in 4/4 (three cells at $n = 48$; Qwen ChartQA at its pre-registered $n = 150$ expansion). This comparison is scoped "with bf16 retained cores" wherever those counts appear in the abstract and in contribution bullet 2. This appendix carries every per-cell number. The quantized-core comparison is a pre-registered result, not one downgraded to exploratory.

Correction. The frozen wording of the VERDICT_CHALLENGED branch contains "we report the reversal verbatim", but **no reversal exists** in the data. Of the 8 cells, **0** have the moment summary significantly ahead. The only negative point estimate is qwen_vrsbench at -0.0225, with an interval of $[-0.1345, +0.0911]$ straddling zero. The direction is **positive in 7/8** cells (3 of the 4 main cells). The gate fired on the **absence of significance at** $n = 48$, not on a change of direction. The adjudication label and consequence clause are unchanged. The main text gives the following accurate statement. This correction likewise **tightens** the claim by giving up the stronger narrative of a reversal. Thresholds and counts are untouched.

Terminology note. The moment mixture summary above, and in the frozen wording, denotes the moment-summary method used in the deep-regime composition round. It uses the same per-channel INT4 core and reuses moment-summary method (the single-tile moment-summary structure, with

Table 29: Quantized versus bf16 sampled payload ($n = 48$; exploratory).

Cell	Δ_Q [95% CI]	Direction
qwen_docvqa	$+0.3668\,[+0.2496, +0.4904]$	positive, significant
qwen_chartqa	$+0.1875\,[+0.0417, +0.3333]$	positive, significant
llava_textvqa	$+0.1281\,[+0.0094, +0.2483]$	positive, significant
qwen_textvqa	$+0.0773\,[-0.0417, +0.2065]$	positive, non-significant
llava_docvqa	$+0.0247\,[-0.0465, +0.0910]$	positive, non-significant
qwen_vrsbench	$-0.0131\,[-0.0833, +0.0571]$	negative, non-significant
llava_chartqa	$-0.0625\,[-0.1458, +0.0000]$	negative, non-significant
llava_vrsbench	$-0.0739\,[-0.1572, -0.0114]$	**negative, significant**

Table 30: Metadata-inclusive attribution (mean [95% CI]): A = pruned cache − pruned core; B = retained tokens − pruned cache; C = QC2 − retained tokens.

Cell	A	B	C
LLaVA–TextVQA	$+0.0000\,[+0.0000, +0.0000]$	$+0.1132\,[+0.0389, +0.2056]$	$+0.0259\,[-0.0938, +0.1405]$
Qwen–TextVQA	$+0.0130\,[+0.0000, +0.0391]$	$+0.2244\,[+0.0955, +0.3511]$	$-0.0451\,[-0.1802, +0.0896]$
LLaVA–VRSBench	$+0.0208\,[+0.0000, +0.0625]$	$+0.0869\,[-0.0189, +0.2060]$	$-0.0417\,[-0.1250, +0.0417]$
Qwen–VRSBench	$-0.0417\,[-0.1042, +0.0000]$	$+0.0489\,[-0.0618, +0.1626]$	$+0.0339\,[-0.0417, +0.1146]$
LLaVA–ChartQA	$+0.0208\,[+0.0000, +0.0625]$	$+0.0417\,[+0.0000, +0.1042]$	$-0.0625\,[-0.1458, +0.0000]$
Qwen–ChartQA	$+0.0000\,[+0.0000, +0.0000]$	$+0.0625\,[-0.0625, +0.1875]$	$+0.0417\,[-0.0625, +0.1458]$
LLaVA–DocVQA	$+0.0000\,[+0.0000, +0.0000]$	$+0.0083\,[-0.0375, +0.0625]$	$-0.0208\,[-0.0625, +0.0000]$
Qwen–DocVQA	$+0.0000\,[+0.0000, +0.0000]$	$+0.0868\,[+0.0208, +0.1688]$	$+0.0594\,[-0.0302, +0.1594]$

$K = 1$ and rank 9). Its budget is re-solved on the metadata-inclusive memory accounting, and it occupies the byte slots the sampled tokens would otherwise have occupied. It is **not** the round-three falsification condition consisting of "17 sampled tokens plus per-tile moment summaries (8256 B)", and it is **not** the "Mixture ($n = 48$)" row of the main continuum results. To distinguish it from those two, the main text uses "the byte-matched moment summary" throughout the abstract, contribution bullet 2 and the superposition paragraph of Section 5. The definition of Δ_D (retained-token method − moment-summary method), the adjudication label, the per-cell numbers and every count are unchanged, word for word.

G.4 Quantizing the Sampled Payload as Well (Exploratory, No Gate)

Δ_Q = retained-token method − retained-token method. The former applies INT4 **once** over the union of core and sampled tokens, so the metadata falls on the union; at the same effective budget it therefore retains 5.1–7.8× as many sampled tokens (87–133 against 27–41), at the cost of each sampled token dropping from bf16 to INT4. 5/8 positive and 3/8 negative at $n = 48$; the pre-registered condition of "$\geq 3/4$ of 8 sharing a sign" is not met (so the global condition fails). The configuration is a natural extension of the KIVI setting, in that the metadata falls on the union; it is not the configuration of the KIVI paper (Liu et al., 2024b), and it carries no pre-registered adjudication rule, so every mention of it is labelled post-hoc. The strongest statement this comparison supports remains per-cell. In the $\geq 12\times$ regime the trade-off between "INT4 keeping more sampled tokens" and "bf16 keeping fewer" yields significant differences in both directions. More quantized tokens help on qwen_docvqa, qwen_chartqa and llava_textvqa, whereas llava_vrsbench shows a **win for fidelity**. At $n = 48$ and with no gate, these remain post-hoc, per-cell observations.

G.5 Attribution inside the Composition (Descriptive, No Gate)

The conditional rule. One candidate sentence about where the deep-regime gains come from was pre-registered as writable **only if** the interval for pruned-cache method − *pruned retained core* straddles zero **and** the interval for retained-token method − pruned-cache method is entirely positive. Both conditions hold in **0/4** main cells, so the condition is not satisfied and the sentence appears nowhere in the paper. *Why 0/4* (mechanical detail, against misreading). On llava_textvqa the second condition holds ($+0.1132\,[+0.0389, +0.2056]$, entirely positive), but the first condition's interval is a **degenerate zero-width interval** $[0.0000, 0.0000]$, with 48/48 samples scoring identically. This does not satisfy the pre-registered "straddles zero" predicate, since that predicate requires ci_lower

Table 31: Compression ratios (payload / effective); retained core and union differ by sampled tokens.

Cell	Retained core	Retained union	Pruned cache
LLaVA: Text/VRS/Chart/Doc	13.40 / **12.84**	16.23 / 12.82	82.29 / 64.90
Qwen–TextVQA	14.88 / **14.00**	17.55 / 13.85	80.84 / 60.21
Qwen–DocVQA	14.81 / **13.93**	17.51 / 13.83	80.43 / 59.97
Qwen–ChartQA	13.41 / **12.84**	16.00 / 12.67	80.00 / 63.34
Qwen–VRSBench	10.45 / **9.97**	12.58 / 9.83	81.00 / 58.91

$< 0 <$ ci_upper. Under the prespecified decision-rule clause no on-the-spot exemption is granted. This is a mechanical outcome in the conservative direction. The data point the same way the sentence would have, but the pre-registered predicate is not satisfied, so the sentence is not written. *Descriptive content that is writable* (no gate). Swapping the bf16 core for a per-channel INT4 core is **in itself** worth almost nothing. In 4 of the 8 cells, the difference is zero per sample. The largest difference is $+0.021$, and qwen_vrsbench is -0.042. The tokens purchased with the freed budget do provide value. The contrast retained-token method $-$ pruned-cache method is positive in 8/8 cells at $n = 48$, from $+0.008$ to $+0.224$. Dropping the bit width from INT4 to INT2 buys 3–5 more tokens ($36 \rightarrow 39$ / $41 \rightarrow 46$) with an indeterminate sign of gain, from $+0.059$ to -0.063 and none significant, which supports no claim.

G.6 Ledger Verification and Compression Ratios

The gap rule follows the continuum round's convention of an absolute gap ≤ 512 B and a per-cell mean relative gap $\leq 2\%$. The 1% threshold does not apply here, because the quantum is one bf16 token $= 512$ B, which against a $\approx$39 KB budget forms a granularity floor of 1.3%.

- The deep-regime composition round has **1920 records** in total (8 cells $\times$ 5 configurations $\times$ 48). Per-sample verification over the 40 cell–configuration combinations gives **0 single-sided violations** (no exceptions), which the pre-registration requires to be exactly zero.
- For the four constrained configurations (retained-token method, retained-token method, retained-token method, moment-summary method; 1536 records) **no gap exceeds 512 B**. The measured maximum absolute gaps are 16–512 B, with the upper end exactly the quantum of one bf16 token. The mean relative gaps are 0.04%–0.82%.
- pruned-cache method sits below the budget by design, being a pure attribution condition with no gap constraint. Its measured absolute gaps are 14,080–21,664 B and its mean relative gaps are 0.34–0.47 (that is, 34–47%).
- At smoke time, the sentinel checks show that retained-token method is 36.0 on llava_textvqa, inside $[30, 45]$, and 40.8 on qwen_textvqa, inside $[25, 45]$; both pass.

Under the frozen writing rule, the payload figures maintain continuity with the paper's main payload-only ledger (Appendix C.1). The effective figures are this round's basis of comparison. The configuration on which the two bases diverge most is retained-token method, because its metadata falls on the union of core and sampled tokens. Across the eight cells, its payload ratio spans 12.6–$17.6\times$. Its effective ratio spans 9.8–$13.9\times$, and this configuration is cited on the effective basis. Both low endpoints occur on qwen_vrsbench ($12.58 \rightarrow 9.83$, payload to effective), the cell with the fewest sampled tokens (median retained-token method 27, against 36–41 elsewhere). Over the other seven cells, the same configuration runs 16.0–$17.6\times$ on the payload basis. It runs 12.7–$13.9\times$ on the effective basis.

G.7 Consequences for the Main Text

1. **The $\geq 10\times$ row is unchanged.** The quantized-core-plus- more-tokens comparison is significant in 0/4 of its adjudicating cells at $n = 48$, so the last row of the map still identifies sampled-token retention as the only contender. Exactly one sentence is added to state that quantizing the retained core in exchange for more sampled tokens is positive in 8/8 cells and significant in 0/8, the design being under-powered at $n = 48$.

Table 32: Selection-quality ordering at a shared budget.

	TextVQA		ChartQA		DocVQA		VRSBench	
Selector	L	Q	L	Q	L	Q	L	Q
Grid-uniform	0.436	0.776	0.160	0.460	0.064	0.416	0.448	0.407
FastV-blind	0.548	0.842	0.193	0.493	0.128	0.512	0.393	0.377
Question-conditioned[†]	0.609	0.884	0.208	0.750	0.165	0.841	0.508	0.449
Strictly monotone	✓	✓	✓	✓	✓	✓	×	×

2. **The central verdict carries a scope restriction (mandatory).** The 8/8 and 4/4 counts in the abstract, in contribution bullet 2, in Section 6.3 and in the main continuum results all carry "with bf16 retained cores". The restriction applies **only** to the new condition in which the core is quantized; the counts and adjudications of the earlier rounds under the original bf16-core condition are unchanged word for word.
3. **The two-ledger rule of Appendix F.3 carries over.** Every ratio and every equal-byte constraint in this appendix is on the effective basis, and each sentence names one basis only.
4. **Statements the paper does not make.** The conditional sentence of L.5 appears nowhere, since 0/4 main cells satisfy its pre-registered predicate. The 5/8 of L.4 is reported per cell at $n = 48$ and is nowhere generalized. The positive direction of Δ_D in 7/8 cells is nowhere described as the central verdict replicating under a quantized core. Replication would require at least 3/4 of the main cells to be significant, and 1/4 was measured.

G.8 Independent Recomputation Check for the Three Quantization Rounds

Every quantity destined for the paper was recomputed per sample from the JSONL records. There were 87 in all, comprising point estimates and both ends of every interval. They cover the map-completion adjudication, the 2×2 grouping factorial (INT4 and INT2, Δ_K and Δ_V), the robustness comparison of the central verdict under a quantized core, the quantized-sampled-payload comparison, and all configuration means, metadata fields and ledger-violation counts of the map-completion round.

- **Every point estimate matches exactly**, the maximum deviation being 0.0000.
- The maximum deviation of the interval endpoints is **0.0208**, exactly the single-sample score quantum at $n = 48$, namely $1/48 = 0.02083$; the corresponding deviation at $n = 150$ is $\leq 0.0068 \approx 1/150$. Its source is the quantile granularity of the percentile bootstrap over discrete scores, with different resampling orders between the two runs. It is **not a difference in the data**.
- Every adjudication is identical under both sets of intervals. This includes the map-completion verdict, the fate of the information-density axis, the attribution of the rescue to K, the mixed verdict on the quantized core plus more tokens, the challenged verdict on the central verdict's robustness, and the 5/8 count of the quantized-sampled-payload comparison.
- Every number in the paper is taken from the registered result set, the frozen aggregation product. The recomputation in this subsection is a check only, and is not a second reporting basis.

H Selection Lineage and Composition with Selection Quality

Grid-uniform, FastV-blind, and the question-conditioned reference span the selection-quality spectrum (Table 32). Six of eight cells follow a strictly monotone ordering. The exceptions occur on VRSBench in both models, where FastV-blind falls below grid-uniform.