Pixel Decodability Is Not a Compression Signal: Causally Evaluating Importance Proxies for Visual KV-Cache Eviction

Anonymous authors
Paper under double-blind review

Abstract

Vision–language models (VLMs) retain, in the visual key–value (KV) cache of their language model, a substantial amount of pixel-decodable visual content. Yet we show, in our setting, that this retention is *task-inert*: across our preregistered tests, how much decodable content a unit retains never positively tracks whether the computation that answers the question causally relies on that unit — where retention relates to task structure at all, the association is weak and points the wrong way. We quantify how much decodable content each visual unit *retains* with a learned pixel-inversion decoder, and how much the model *causally uses* it with single-super-patch KV ablation (the teacher-forced drop in gold-answer log-probability), and relate the two within images under a preregistered, sign-calibrated, held-out design. Retention is decoupled from attention ($\rho \approx -0.07$ to -0.09) and, in a preregistered, well-powered null (held-out N=72, power ≈ 0.9; upper CIs exclude $\rho \gtrsim 0.06/0.01$), from causal utilization ($\rho \approx -0.01$ to -0.05). Yet utilization is not inert to every proxy: attention weakly but significantly tracks it ($\rho \approx +0.11$ to $+0.12$, held-out intervals excluding zero) — the only signal we find that does, serving as the design's positive control. We characterize pixel-decodable retention as an *informational axis* of the visual KV cache, orthogonal to the functional (attention–utilization) one. How much of this task-inert content a cache holds differs by architecture in our model pair: the encoder-free VLM retains 2.68× more than the encoder-based one. The engineering consequence is a controlled negative result: at super-patch granularity, deconfounded pixel-decodable retention ranks KV eviction no better than random; at token granularity it acquires only a weak inverse-importance signal at the larger budgets — dominated at every budget by attention magnitude, the weak-but-real proxy. In our setting, pixel-decodable reconstructability is not a competitive KV-compression signal at any granularity we test.

1 Introduction

Vision–language models (VLMs) carry a large visual key–value (KV) cache (hundreds of tokens per image, across every layer), and a growing body of work compresses it by *evicting* the visual units deemed least important. The importance signal is almost always a proxy: the magnitude of attention a unit receives, or the reconstructability of its representation. Both encode an implicit assumption: that the proxy tracks what the model actually *uses*. We ask whether that assumption holds by taking the assumption apart into three distinct questions about a visual unit's KV. How much decodable pixel content does it *retain*? How much attention does it *receive*? And how much does the model *causally rely* on it to produce the correct answer? These are routinely conflated; we measure them separately.

We instrument two of these axes directly. *Retention* is read out by a learned decoder that inverts a unit's KV back to pixels, a content-level probe, blind to the attention pattern (§3.1). *Utilization* is read out causally: we ablate a single super-patch's KV and measure the teacher-forced drop in the gold answer's log-probability, a per-unit causal readout of "does removing exactly this unit hurt the answer" (§3.2). Relating the two within each image, under a preregistered partial-correlation design with sign calibration on a held-out split (§3.5), we find a clean dissociation. Figure 1 lays out the three readouts and the relationships we measure between them.

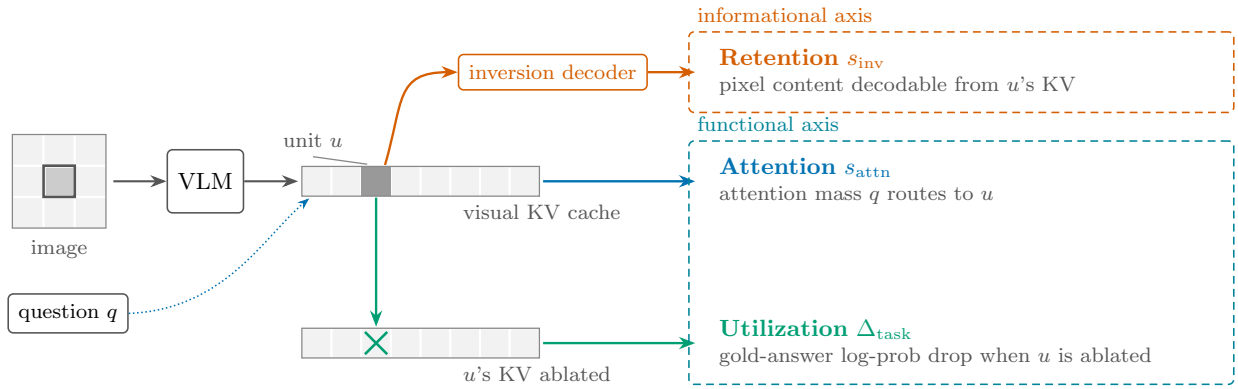

Figure 1: Three readouts of one visual KV unit u. *Retention* (s_{inv}): a learned pixel-inversion decoder measures how much of u's image content its KV still decodes to pixels — the informational axis. *Attention* (s_{attn}): the attention mass the question routes to u. *Utilization* (Δ_{task}): the drop in gold-answer log-probability when u's KV is ablated — together with attention, the functional axis. The paper measures the relationships between the two axes within images; estimates are in Figure 2.

Our findings, on TextVQA with an encoder-free (Gemma-4-12B) and an encoder-based (InternVL3.5-8B) VLM, are three. (i) Pixel-decodable retention predicts neither attention ($\rho \approx -0.07$ to -0.09) nor causal utilization: the covariate-adjusted partial correlation is $-0.01/-0.05$ across the two models with CIs including zero, under a design powered at ≈ 0.9 for the preregistered effect size of $\rho = 0.15$, and the hypothesized positive coupling never appears. (ii) Attention, by contrast, does carry a weak but genuine utilization signal ($+0.11$ to $+0.12$, held-out CIs excluding zero), and because the very same estimator detects it, the retention null reflects retention itself rather than an insensitive instrument. Retention is orthogonal to this functional signal; we cast pixel-decodable retention as an **informational axis** of the visual KV cache: content the cache carries and that is extractable, but that the task computation does not route through. (iii) How much of this task-inert content a cache holds differs by architecture *in our model pair*: the encoder-free model retains $2.68\times$ more than the encoder-based one. In the eviction experiment, the ordering predicted by (ii) is recovered on the primary causal-cost metric: attention-guided eviction beats random; deconfounded retention does not, in either direction of the ranking, on the super-patch grid — and at token granularity it helps only through a weak *inverse* signal (§4.5).

Contributions. (1) We define a *per-unit causal readout* of utilization (single-super-patch KV ablation) and use it to evaluate importance proxies directly, rather than through downstream accuracy. (2) We jointly quantify retention, attention, and utilization in the VLM visual cache, turning a two-way intuition into a three-way analysis that repositions retention as an informational axis orthogonal to the functional one rather than merely "decoupled." (3) We give a quantitative model-pair contrast ($2.68\times$) associating the encoder-free/encoder-based distinction with how much task-inert pixel content the cache retains. (4) We contribute a controlled negative result: deconfounded pixel-decodable retention ranks KV eviction no better than random at super-patch granularity, and acquires only a weak inverse-importance signal at token granularity, providing causal evidence, in our setting, against the assumption that pixel-level reconstructability implies importance.

The pixel-inversion instrument, and the finding that encoder-free VLMs reconstruct more from deep KV than encoder-based ones, originate in earlier privacy-oriented work (Zhou et al., 2026); the retention/utilization *dissociation*, the causal-utilization readout, the three-axis analysis, and the compression falsification are new here; the prior work quantified *what can be reconstructed*, not whether it is *used* (§2).

2 Related Work

Decodability is not causal use. A recurring finding across representation analysis is that information *recoverable* from a hidden state by a trained probe need not be information the network *uses* to produce

its output. Probing studies in vision transformers recover object counts and spatial attributes that the model does not causally rely on for its predictions (Huang & Chang, 2025), and analogous dissociations appear in language models, where structural information (e.g., bracket-matching state) is linearly decodable yet causally inert for the behavior it superficially predicts (Sharma et al., 2026). Methodologically, our utilization readout is in the ablate-and-measure lineage of amnesic probing (Elazar et al., 2021) — intervene on a representation and measure the behavioral change — applied at the granularity of a single visual KV unit. We inherit this decodable-≠-causal lens and instantiate it for a concrete engineering object, the visual key–value (KV) cache of a vision–language model (VLM), where the gap between what is retained and what is used has direct consequences for cache compression.

Retention, accessibility, and utilization in text-only caches. Closest in vocabulary is recent work formalizing *retention*, *accessibility*, and *utilization* of information in the KV cache of text-only LLMs (Ananthanarayanan et al., 2026). We deliberately keep our terminology aligned to make the comparison legible, but the instrumentation and the object of study differ on every axis. Their setting is text tokens; ours is visual tokens carrying pixel content. Their retention is read out with linear concept-vector probes and their accessibility with attention-derived graph reachability; ours is read out with an *independent, learned pixel-inversion decoder* — a content-level probe that never sees the attention pattern. Their utilization is inferred from downstream task accuracy; ours is a *per-unit causal readout* obtained by ablating a single visual super-patch's KV and measuring the teacher-forced drop in gold-answer log-probability. The shared vocabulary is a feature: it lets us state precisely that the retention/utilization dissociation we document is a *visual, pixel-level, causally-grounded* instance, not a re-run of the text-only result.

Reconstruction as a compression signal for VLM caches. A parallel line of work compresses VLM KV/token caches and, in several cases, uses *reconstructability* of the compressed representation as the pruning signal: retaining the units whose features can be recovered (Lee et al., 2026), or regressing value vectors to decide what is redundant (Lin et al., 2026); in text-only KV caches, reconstruction of the full context has likewise been used as a query-agnostic eviction score (Kim et al., 2025). These methods reconstruct in *feature* space (learned subspaces, value regressions, context re-generation) and report genuine compression payoffs there. Our probe operates one level lower (at the level of *pixel-decodable content*, using a decoder trained only to invert KV back to the image), and we use it *diagnostically* (to characterize an axis of the representation) rather than as a scoring function. The distinction matters for our negative result: it is specific to pixel-decodable content, while feature-space reconstruction targets a different, and evidently more task-aligned, quantity. Attention-magnitude pruning — FastV-style eviction of low-attention visual tokens (Chen et al., 2024) and its cache-eviction variants (Tu et al., 2025; Liu et al., 2026), where the latter's "retention" denotes a cache-persistence policy distinct from the pixel-decodable retention studied here — is the standard baseline our compression analysis compares against.

Prior work: KV→pixel inversion. The pixel-inversion decoder we use as a measurement instrument was introduced by Zhou et al. (2026), which framed KV-cache reconstructability as a *privacy* boundary and characterized *what* is reconstructable (spatial grid resolution, character readability, structural fidelity), showing that encoder-free VLMs reconstruct substantially more from deep KV than encoder-based ones. That work is entirely about reconstruction quality; it does not ask, and does not address, whether the reconstructable content is *causally used* by the model. While the inversion instrument and the architecture-differential reconstruction observation originate in Zhou et al. (2026), what is new here is the question that work did not pose: (i) a *causal readout* of utilization (single-unit KV ablation) and the finding that reconstructable retention does not predict it — a preregistered, well-powered dissociation, not a qualitative aside; (ii) the *attention* axis, turning a two-way measurement into a three-way analysis that positions retention as an informational axis orthogonal to the functional one (§4.3); and (iii) the reframe from a privacy attack into a cache-compression diagnostic, with a controlled eviction falsification (§4.5). The instrument is prior art; the dissociation, its causal grounding, and its compression consequence are not.

Limits of attention-based visual pruning evaluations. Finally, recent evaluation work shows that attention-guided visual-token pruning carries spatial biases and can degrade on tasks requiring peripheral evidence (Endo et al., 2025), complementing our finding that attention is only a *weak* predictor of causal utilization (§4.3): it is the better of the two proxies we study, but it is not a strong one.

3 Method

We measure three axes of a VLM's visual KV cache: how much pixel content it *retains*, how much attention it *receives*, and how much the model *causally uses* it — and relate them within a preregistered partial-correlation design. All quantities are defined per *super-patch* (a spatial unit of the visual grid) and estimated within-image to remove image-level confounds.

3.1 Retention: learned pixel inversion

For a fixed decoder layer ℓ, we read the visual KV of an image-only prefill (the image with an empty/neutral prompt, so that KV reflects encoded visual content rather than question-conditioned state) and train an inversion decoder to reconstruct the input image from that KV. The retention score s_{inv} of a super-patch is the reconstruction quality attributable to that patch's KV, measured as a *shuffle-adjusted excess*: the reconstruction score with the patch's KV in place minus the score under a within-image shuffle of KV positions, so that s_{inv} isolates position-specific decodable content from what any KV would afford on average. We report the main analysis at the primary layer $\ell = 0$ (the layer at which the instrument is trained and validated); deeper layers are secondary. To quantify measurement noise, we train the decoder under *two independent seeds* and take the two resulting s_{inv} maps as parallel measurements of the same latent quantity (used for disattenuation, §3.5). Decoder architecture, training protocol, and the shuffle-adjusted score are specified in Appendix B.

3.2 Utilization: single-unit causal ablation

Retention asks what can be *read out* of the KV; utilization asks what the model *would lose* without it. For each super-patch we ablate its visual KV (masking those key/value positions so downstream attention cannot attend to them) and measure the change in the teacher-forced gold-answer log-probability:

$$\Delta_{\text{task}}(u) \;=\; \log p(\text{gold} \mid \text{full KV}) \;-\; \log p(\text{gold} \mid \text{KV} \setminus u),$$

summed over the gold answer's tokens. A large positive Δ_{task} means the patch's KV was causally load-bearing for the correct answer; $\Delta_{\text{task}} \approx 0$ means it was inert. This is a *per-unit causal readout* of utilization, as opposed to accuracy-based proxies: it directly interrogates the counterfactual "does removing exactly this unit hurt the gold answer." That Δ_{task} carries genuine signal rather than noise is confirmed downstream (§4.3): attention weakly but significantly predicts it in both models (held-out cluster-bootstrap CIs exclude zero). This serves as the design's positive control (§4.2); an external criterion validation — Δ_{task} localizes OCR-grounded answer regions on both models — is in Appendix C. The readout is a *single-unit* intervention, so it can understate importance that is redundantly distributed across units, particularly after the ViT's bidirectional mixing in the encoder-based model (whose per-unit Δ_{task} is therefore a lower bound). We pair it with cumulative eviction (§4.5) and with a whole-image ablation arm, which removes *all* visual KV and yields mean gold-logprob drops of 8.1/9.9 for Gemma/InternVL (positive on 94%/90% of images): the small per-patch effects reflect distributed reliance.

3.3 Attention readout

The third axis, attention $s_{\text{attn}}(u)$, is the mean attention mass a super-patch's KV positions receive from the question's text tokens in the question-conditioned forward pass: for each of the model's full-attention layers (5, 11, and 17 in both models), we average the attention weights over heads and over the question's token positions to obtain a per-visual-token mass, take the equal-weight mean over the three layers, and sum over the tokens assigned to the super-patch. This is the FastV-style text→vision importance signal (Chen et al., 2024) that attention-based eviction methods rank by, computed at the same granularity as the other two readouts.

3.4 Unit of analysis and models

We study two VLMs chosen to span the architectural axis of interest: **Gemma-4-12B**, an *encoder-free* VLM (pixels enter the language model directly), whose visual grid varies with aspect ratio (e.g., 13×20;

252–275 tokens per image over our images); and **InternVL3.5-8B**, an *encoder-based* VLM (a ViT encoder mediates the visual input), whose single tile yields $16 \times 16 = 256$ tokens. Both grids are aggregated to a $4 \times 4 = 16$ super-patch grid per image; because Gemma's grid does not divide evenly into 4×4, its token-to-super-patch map is *non-uniform* (block sizes such as $[20, 20, 20, 20, 15, \dots]$ for a 13×20 image), and we use the exact block assignment throughout rather than assuming equal-sized patches. Super-patch aggregation trades spatial resolution for statistical power: a per-token unit would leave too few independent units per image for a within-image correlation, whereas 16 super-patches give a usable within-image sample while still localizing to regions; the coarser cumulative-eviction analysis (§4.5) provides a complementary check at the group level. Evaluation is on the official TextVQA validation set, a text-reading VQA task where the answer is localized to specific image regions. This localization is load-bearing for the instrument: it is what makes single-unit ablation informative.

3.5 Estimand: within-image rank partial correlation

Our primary quantity is the *within-image rank partial correlation* between retention and utilization. Within each image we rank super-patches by s_{inv} and by Δ_{task}, and we residualize both on a fixed covariate set (attention mass s_{attn}, two pixel-energy covariates ($s_{\text{pixel-edge}}$, $s_{\text{pixel-entropy}}$), and spatial position (super-patch row and column, center distance), so that the reported association is *not* explained by "this patch has more texture," "this patch is attended to," or "this patch is central." Rank residualization is done per image (16 units) and never across images. Uncertainty is quantified by a *cluster bootstrap over images* ($n_{\text{boot}} = 10{,}000$), which respects the fact that the 16 super-patches of an image are not independent.

Sign calibration and held-out confirmation. To avoid a post-hoc sign selection, we fix the *sign* of the hypothesized effect on a calibration split and *test* it on a disjoint validation split. Image IDs are partitioned three ways — decoder-training / calibration ($N = 48$) / validation ($N = 72$) — pairwise disjoint, so no image contributes to both the instrument and the test. The sign sign_{cal} is frozen from calibration; the verdict is read only on validation; the frozen sign is the sign of the Gemma calibration estimate, shared across both models (§4.2 reports both models' calibration values).

Disattenuation. Because s_{inv} is measured with noise, an observed partial correlation understates the true one. Using the two decoder seeds (§3.1) we estimate the split-half reliability of s_{inv} and apply a *conservative* disattenuation — dividing only by $\sqrt{\text{rel}_{\text{inv}}}$ (correcting for retention-side noise only), so that the reported effect is, if anything, an over- rather than under-estimate. This makes a null result robust: a corrected estimate near zero cannot be dismissed as measurement attenuation.

Power and effect-size bounds. The design was preregistered with a smallest effect size of interest of $\rho = 0.15$ for the retention–utilization test and a decoupling threshold of 0.5 for C1. The confirmatory test reads the held-out validation split ($N = 72$; the sign is fixed on the disjoint 48-image calibration split); at $N = 72$ the cluster-bootstrap design achieves power ≈ 0.9 against that SESOI. Beyond power, the interval itself bounds the effect: the two-sided upper limits exclude $\rho \gtrsim 0.06$ (Gemma) and $\rho \gtrsim 0.01$ (InternVL), so we can rule out even modest positive associations, not merely the SESOI. A null under these conditions is a *well-powered, bounded* null, not an underpowered absence of evidence. Two qualifications are stated once here and hold throughout. Power was computed by plasmode simulation: synthetic effects of the preregistered size were injected into the empirical per-image Δ_{task} structure and recovered with the identical rank-partial, cluster-bootstrap pipeline. And all correlation bounds are on the *measured* scale: the per-unit Δ_{task} readout has finite reliability (its cross-model consistency, +0.12, indicates the ceiling any per-unit correlate can reach), so true-scale bounds are weaker; the cumulative-eviction analysis (§4.5), which aggregates over units, is the leg of the argument that does not depend on per-unit reliability.

4 Results

All correlations below are within-image rank partial correlations on the held-out validation split ($N = 72$), estimated by cluster bootstrap over images ($n = 10{,}000$); the sign of each hypothesized effect was fixed on the disjoint calibration split ($N = 48$). We report both models (Gemma-4-12B, encoder-free; InternVL3.5-8B,

encoder-based). Figure 2 summarizes the four within-image relationships and the architecture contrast at a glance; Table 1 gives the full intervals.

4.1 Retention is decoupled from attention (C1)

Pixel-decodable retention and attention mass are, if anything, weakly *negatively* correlated within an image — far below the preregistered decoupling threshold of 0.5, and in the wrong direction to serve as a proxy. The rank correlation $\rho(s_{\text{inv}}, s_{\text{attn}})$ is -0.068 for Gemma (one-sided 95% bootstrap upper bound -0.017; two-sided CI $[-0.130, -0.007]$) and -0.088 for InternVL, holding across both the calibration and validation splits. The KV positions that retain the most decodable pixel content are not the positions the model attends to: retention is not a proxy for attention (throughout, "decoupled" means far below the preregistered proxy threshold, not a literally zero correlation).

4.2 Retention is decoupled from causal utilization (C2, well-powered null)

Result. Retention is likewise uninformative about which units the model *causally uses*. On validation, the covariate-adjusted partial correlation $\rho(s_{\text{inv}}, \Delta_{\text{task}} \mid s_{\text{attn}}, s_{\text{pixel}}, \text{pos})$ is -0.009 for Gemma (95% cluster-bootstrap CI $[-0.076, +0.057]$, tightly straddling zero) and -0.052 for InternVL (CI $[-0.113, +0.010]$, a weak negative whose interval still includes zero). The direction that matters is the one hypothesized in advance and fixed on calibration ($\text{sign}_{\text{cal}} = +1$: retention would *positively* predict utilization if reconstructable content were preferentially used). Neither model shows it: both point estimates are at or below zero, and the sign-oriented one-sided 95% lower bound is negative for both (Gemma -0.070, InternVL -0.105). This is a *well-powered, bounded* null (§3.5: power ≈ 0.9 at N=72 against $\rho = 0.15$; upper CIs exclude $\rho \gtrsim 0.06/0.01$), not an absence of evidence. Critically, the null is not an artifact of an insensitive instrument: the same within-image rank-partial estimation pipeline, on the same held-out split, detects a positive attention–utilization association ($\rho \approx +0.11$, CI excluding zero; §4.3). We therefore state a positive claim: within our setting, pixel-decodable retention does not predict causal utilization (the InternVL estimate leans weakly negative; Gemma's is essentially zero).

Robustness of the null. Three families of checks make the null robust rather than fragile. *Design-internal checks.* The sign fixed on calibration was weakly *positive* for Gemma ($+0.052$; the InternVL calibration estimate was -0.005, essentially zero and uninformative about direction) yet did not replicate on validation — consistent with an effect smaller than the between-split variation. The retention instrument is reliable (split-half 0.84/0.89 across decoder seeds; 0.82/0.88 for the exact within-image-ranked, covariate-residualized quantity entering the estimator), so the null cannot be explained away as measurement or residualization noise, and our conservative disattenuation (§3.5) only pushes the estimate further from — not toward — zero. *Estimator robustness.* A mixed-effects model with per-image random intercepts yields retention coefficients of $+0.005/-0.055$, and a per-image Wilcoxon test fails to reject zero for both models ($p = 0.80/0.41$): three estimators, one conclusion. *Measurement-choice robustness.* The null holds across decoder depth — recomputing s_{inv} for Gemma from layer-0/6/12 KV, each with its own decoder, yields $\rho = -0.009$, -0.020, -0.039 (all CIs include zero; never the hypothesized positive coupling), even as decodable retention itself weakens with depth (mean s_{inv} $0.108 \rightarrow 0.088 \rightarrow 0.066$, two-seed ensemble; the architecture contrast in Figure 2b uses the seed-wise estimate, 0.109), and C1 likewise holds at all three depths — so the task-inertness is not a primary-layer artifact (Gemma; the encoder-based decoder was trained only at layer 0). Nor is it manufactured by deconfounding: a progressive covariate ladder shows the *marginal* correlation is already ≈ 0 for Gemma (-0.004 $[-0.066, +0.059]$) and weakly *negative* for InternVL (-0.068 $[-0.131, -0.010]$) — the covariates absorb a negative texture confound, not a positive signal. Finally, Δ_{task} itself carries genuine cross-model structure: the two models' within-image utilization maps correlate at $+0.12$ $[+0.05, +0.20]$ on identical images and questions, bounding the readout's reliability away from zero.

4.3 Attention weakly but genuinely tracks utilization

Crucially, the utilization axis is *not* inert to every proxy: attention weakly predicts it. The partial correlation $\rho(s_{\text{attn}}, \Delta_{\text{task}} \mid s_{\text{pixel}}, \text{pos})$ is $+0.113$ for Gemma (CI $[+0.044, +0.182]$) and $+0.120$ for InternVL (CI

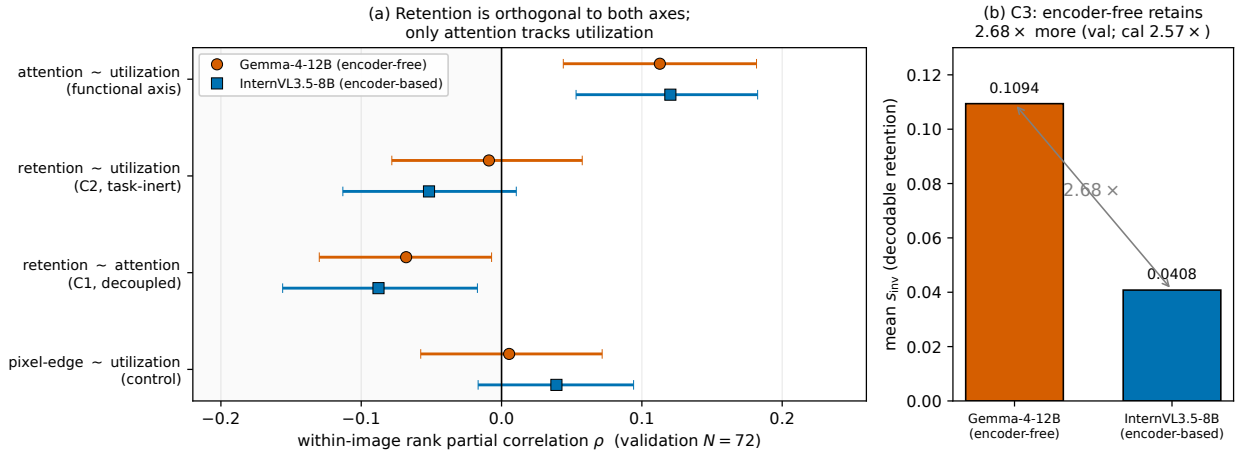

Figure 2: The dissociation results at a glance: (a) within-image partial-ρ estimates for attention–utilization, C2, C1, and the pixel-edge control in both models; and (b) the C3 retention contrast (validation 2.68×, 95% CI $[2.37, 3.06]$; calibration 2.57×).

Table 1: Within-image rank partial correlations, held-out validation ($N = 72$ images, 1152 units), 95% cluster-bootstrap CIs over images ($n = 10{,}000$).

Relationship	Gemma-4-12B (encoder-free)	InternVL3.5-8B (encoder-based)
C1 retention⊥attention $\rho(s_{\text{inv}}, s_{\text{attn}})$	-0.068 $[-0.130, -0.007]$	-0.088 $[-0.156, -0.017]$
C2 retention⊥utilization $\rho(s_{\text{inv}}, \Delta_{\text{task}} \mid \text{cov})$	-0.009 $[-0.076, +0.057]$	-0.052 $[-0.113, +0.010]$
attention∼utilization $\rho(s_{\text{attn}}, \Delta_{\text{task}} \mid \text{pix,pos})$	$+0.113$ $[+0.044, +0.182]$	$+0.120$ $[+0.053, +0.183]$
pixel-edge∼utilization $\rho(s_{\text{pixel-edge}}, \Delta_{\text{task}})$	$+0.005$ $[-0.058, +0.072]$	$+0.039$ $[-0.017, +0.094]$

C2 robustness (all agree on null): mixed random-intercept $\beta_{\text{inv}} = +0.005$ / -0.055; per-image Wilcoxon $p = 0.80$ / 0.41. Reliability (split-half) 0.84 / 0.89. C2 power ≈ 0.9 at held-out $N = 72$ vs SESOI $\rho = 0.15$; upper CIs bound the effect below $\rho \approx 0.06$ (Gemma) / 0.01 (InternVL).

$[+0.053, +0.183]$), both positive with held-out intervals that exclude zero. By contrast, pixel-energy covariates do not predict utilization ($\rho(s_{\text{pixel-edge}}, \Delta_{\text{task}}) = +0.005/+0.039$, CIs contain zero), ruling out "textured patches matter" as the explanation. This result does double duty. It confirms the Δ_{task} instrument carries real signal (something — attention — predicts it), which in turn hardens the C2 null (§4.2). And it establishes an asymmetry: of the axes we measure, attention is the *only* one with a statistically detectable relationship to utilization, and pixel-decodable retention is orthogonal to it. At $\rho \approx 0.11$ attention explains only about one percent of the within-image variance in Δ_{task}: it is the better proxy, not a good one. We use "functional axis" as shorthand for this signal — the only utilization-tracking signal we detect; the substantive claim is comparative: retention tracks utilization even less, indistinguishably from zero. In that comparative sense pixel-decodable retention behaves as an **informational axis** of the cache (content carried and extractable) that is orthogonal to the weak functional signal (content the computation acts on). One structural asymmetry deserves note when reading this comparison: retention is by construction a query-agnostic readout (image-only prefill), while attention and utilization are both measured under the question-conditioned pass. An eviction score must be computable before the answer in any case, so the comparison reflects deployment reality, but "the only proxy that tracks utilization" is also the only query-conditioned one (Appendix A examines how much query-conditioning attention actually uses).

4.4 Retention differs across the architecture pair (C3)

How much task-inert pixel content a cache retains differs sharply between our two models. At the primary layer, the encoder-free Gemma retains 2.68× more decodable pixel information per unit than the encoder-

based InternVL (cluster-bootstrap 95% CI $[2.37, 3.06]$; calibration $2.57\times$, consistent across splits and decoder seeds). A natural mechanistic reading is that, without a ViT bottleneck to re-encode the image first, pixel content enters the language model directly and persists there in decodable form — even though, per §4.2, that persistence buys no additional causal utility. This mechanistic reading remains a model-pair hypothesis: with one model per architecture class, the $2.68\times$ is confounded with every other difference between Gemma-4-12B and InternVL3.5-8B (parameter count, training data, tokenizer, visual token budget), so we report it as a *model-pair* contrast and defer an architecture-*class* claim to the multi-model replication discussed in §5.

4.5 Compression implication: retention is not a pruning signal

The dissociations above predict a concrete engineering consequence. Reconstruction-based importance admits two rival readings, and we test both. Under the *removable-redundancy* reading, a unit whose pixel content is highly reconstructable is redundantly stored and therefore safe to drop: evict the *highest*-retention super-patches. Under the *high-value* reading, high retention marks content worth keeping: evict the *lowest*-retention super-patches instead. If retention is task-inert, *neither* rule should beat evicting at random. We test it directly with a controlled KV-eviction Pareto on validation ($N = 72$). For each strategy we evict $k \in \{4, 8, 12\}$ of the 16 super-patches' visual KV in a single fresh forward pass and measure the teacher-forced gold-answer log-probability drop (lower is safer) and greedy-decoded VQA accuracy; random is averaged over 10 seeds; all differences use an image-level cluster bootstrap ($n = 2{,}000$). Strategies: attention-guided **FastV** (evict lowest-attention), **raw retention** (evict highest s_{inv}), **deconfounded retention** (evict highest within-image-residualized s_{inv}), **combined** (the sum of the within-image retention rank and the reversed attention rank, so that patches that are simultaneously high-retention and low-attention are evicted first), and **random**. Table 2 gives the full grid; the paired contrasts tell the story, and Figure 3 shows them.

Table 2: Compression Pareto — KV eviction on validation ($N = 72$). Each cell: mean gold-logprob drop (lower = safer) / VQA accuracy, per strategy $\times$ budget k; random averaged over 10 seeds; paired differences via image-cluster bootstrap (n=2,000). Bold marks the attention baseline (FastV).

Gemma-4-12B (encoder-free; baseline VQA 0.66)			
strategy	k=4	k=8	k=12
FastV (attention)	**+0.25 / 0.60**	**+0.68 / 0.56**	**+2.84 / 0.36**
Raw retention	+0.77 / 0.57	+2.40 / 0.44	+4.56 / 0.22
Deconfounded retention	+1.16 / 0.54	+3.20 / 0.40	+5.49 / 0.22
Combined	+0.53 / 0.54	+0.93 / 0.51	+3.20 / 0.34
Random	+1.46 / 0.45	+3.04 / 0.33	+5.17 / 0.19
InternVL3.5-8B (encoder-based; baseline VQA 0.73)			
strategy	k=4	k=8	k=12
FastV (attention)	**+0.30 / 0.67**	**+0.83 / 0.61**	**+2.49 / 0.56**
Raw retention	+0.63 / 0.62	+2.33 / 0.50	+5.46 / 0.31
Deconfounded retention	+1.54 / 0.57	+2.95 / 0.44	+5.35 / 0.28
Combined	−0.04 / 0.69	+0.60 / 0.61	+3.18 / 0.50
Random	+1.79 / 0.57	+3.53 / 0.46	+6.35 / 0.29

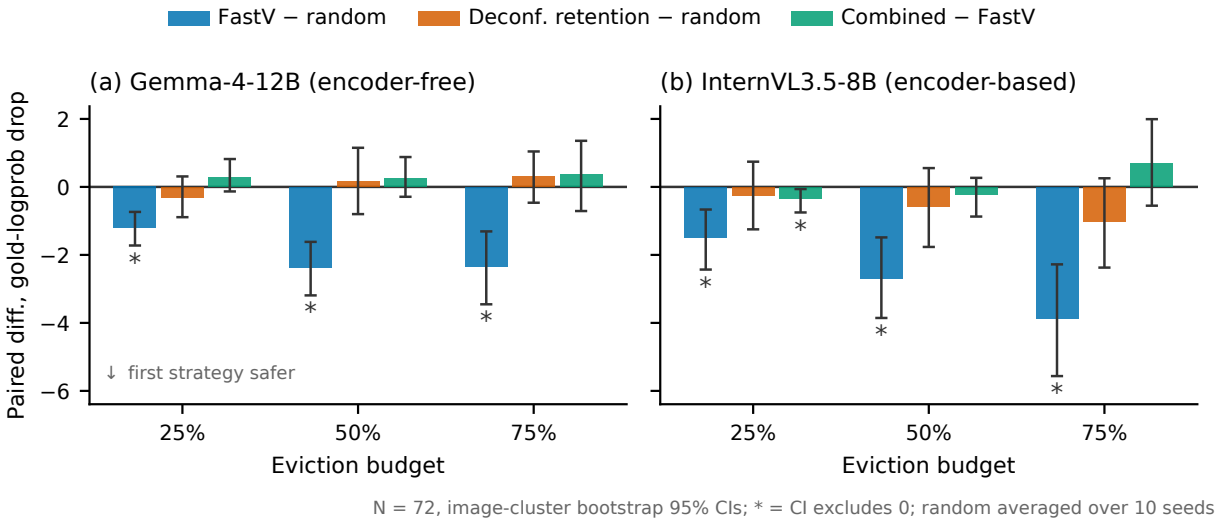

Figure 3: Eviction contrasts on the super-patch grid (validation, N=72): paired differences in gold-answer logprob drop at each budget (negative = first strategy safer; error bars = image-cluster bootstrap 95% CIs; * = CI excludes zero, as in Tables 2 and 3; random averaged over 10 seeds). FastV − random excludes zero at every budget (attention helps); deconfounded retention − random contains zero at every budget (no super-patch compression signal); combined − FastV shows no consistent gain, with the one exception of InternVL at the 25% budget. Full grid: Table 2; token-granularity contrasts: Table 3.

Paired contrasts on the super-patch grid (mean diff [95% CI]; $*$ = CI excludes 0; differences computed from unrounded per-image values):

- **Deconfounded retention − random** (main; expect ≈ 0): Gemma $-0.30/+0.16/+0.32$, InternVL $-0.25/-0.57/-1.01$ — **all include 0** (no super-patch compression signal).
- **FastV − random**: Gemma $-1.21^*/-2.36^*/-2.33^*$, InternVL $-1.49^*/-2.70^*/-3.86^*$ — **all exclude 0** (attention helps).
- **Combined − FastV**: Gemma $+0.28/+0.25/+0.35$, InternVL $-0.34^*/-0.22/+0.69$ — no consistent gain from adding retention. The one significant exception is InternVL at k=4, where the combined rule beats FastV on both metrics (-0.04 vs. $+0.30$ cost; 0.69 vs. 0.67 accuracy) — retention can complement attention at small budgets on one model, but the gain does not persist at larger budgets or on Gemma.

Deconfounded retention ranks eviction no better than random at super-patch granularity — on both models. The paired difference deconfounded retention − random has a 95% CI containing zero at every budget for both VLMs; at the larger Gemma budgets the point estimate is even mildly *positive* (deconfounded retention slightly worse than random). By contrast, *raw* retention does appear to beat random at the smallest budget (raw retention − random excludes zero: Gemma -0.69, InternVL -1.16), but that apparent advantage evaporates the moment retention is residualized within-image against texture, attention, and position. This is the compression-side confirmation, on a real cumulative-eviction forward pass, of the §4.2 null: retention's surface-level compression value has no detectable unique advantage once within-image confounds are removed. The paired super-patch intervals also *bound* what remains undetected: the maximum safety advantage of deconfounded retention over random compatible with the 95% CIs (0.88/0.77/0.45 gold-logprob for Gemma at k=4/8/12; 1.28/1.71/2.40 for InternVL) is smaller at every budget than the advantage FastV actually realizes (1.21–2.36 and 1.49–3.86 respectively) — so, at this granularity, "no better than random" is not merely a failure to reject, but an interval statement placing retention's best case below attention's observed case. One bookkeeping caveat for Gemma: because its token-to-super-patch blocks

are non-uniform (§3.4), the *realized* evicted-token fraction at a nominal budget varies slightly by strategy (per-strategy means 22.8%–25.2% at k=4; FastV in fact evicts marginally *fewer* tokens than the retention rules), a ≤ 2.4 percentage-point wobble that slightly favors the retention rules' safety and therefore cannot explain their failure to beat random. The secondary greedy-VQA-accuracy metric tells the same story with one wrinkle: paired deconfounded retention − random accuracy differences include zero at every budget for InternVL and at Gemma's two larger budgets, with a single marginal exception at Gemma's smallest budget (+0.09, CI $[+0.02, +0.16]$) that would not survive Bonferroni correction across the twelve strategy×budget contrasts and in any case points the wrong way for retention-as-importance: its sign favors *evicting* high-retention patches; on the super-patch grid the removable-redundancy reading that would support is rejected on the primary metric at every budget (at token granularity a weak effect in that direction does emerge; see below). The complementary policy of treating retention as a *keep* signal and evicting the *lowest*-retention units instead fares no better: on an additive per-unit proxy (summing single-unit Δ_{task} over the evicted set) on the same validation images, evicting the k lowest-retention super-patches is indistinguishable from random for Gemma and significantly *worse* than random for InternVL (e.g., +1.78 at k=8, CI excluding zero; raw and deconfounded scores behave alike). Neither direction of the retention ranking carries eviction value, so the null is not an artifact of having tested the "wrong end" of the score.

Token granularity. All results above operate on the 16-super-patch grid. Re-running the cumulative-eviction protocol at native token granularity (252–275 visual tokens per image for Gemma, 256 for InternVL; budgets of 25/50/75% of each image's actual token count, so at a given budget every strategy deletes exactly the same number of tokens) sharpens the comparative picture and triggers one preregistered narrowing (Table 3). The attention positive control holds: FastV − random excludes zero at every budget on both models. Deconfounded retention, in contrast with the super-patch null, acquires a weak *inverse*-importance signal: evicting the highest-retention tokens is safer than random at the 50% and 75% budgets on both models (Gemma -1.29 $[-2.02, -0.55]$ and -0.97 $[-1.79, -0.18]$; InternVL -1.13 $[-2.34, -0.10]$ and -1.31 $[-2.42, -0.30]$ gold-logprob), while the 25% intervals include zero and the effect is not monotone in budget (Gemma's 50% effect exceeds its 75%); a per-seed check confirms no single random seed drives it, and raw retention shows the same directional pattern (its 75% intervals marginally include zero). The preregistered narrowing therefore applies: "no better than random" is a super-patch-level statement, and at token granularity high pixel-decodable retention weakly marks *safe-to-evict* content — the direction opposite to what retention-as-importance would need. Unchanged, and now strengthened from an interval bound to a direct paired comparison, is the ordering that matters for practice: FastV − deconfounded retention excludes zero at every budget on both models, so attention dominates retention wherever the two disagree. The secondary VQA metric agrees in direction with the retention−random inversion where it is significant (Gemma's two smaller budgets; no InternVL budget reaches significance). The spatial counterpart of this inversion — retention anti-localizing OCR-grounded answer regions — is validated independently in Appendix C.

Table 3: Token-granularity eviction, paired gold-logprob contrasts (mean difference; negative = first strategy safer; image-cluster bootstrap, n=2,000; random averaged over 5 seeds within image; budgets are 25/50/75% of each image's actual visual-token count and identical across strategies). * = 95% CI excludes 0; key intervals in the text.

Model	Contrast	25%	50%	75%
Gemma	Deconf. retention − random	−0.17	-1.29^*	-0.97^*
Gemma	FastV − random	-1.53^*	-3.61^*	-4.52^*
Gemma	FastV − deconf. retention	-1.36^*	-2.33^*	-3.56^*
InternVL	Deconf. retention − random	−0.25	-1.13^*	-1.31^*
InternVL	FastV − random	-1.08^*	-2.25^*	-4.01^*
InternVL	FastV − deconf. retention	-0.83^*	-1.12^*	-2.70^*

Attention is the weak-but-real signal; retention adds no consistent gain. Attention-guided FastV is robustly the safest single-proxy strategy — FastV − random excludes zero at every budget for both models, and it also retains the most VQA accuracy (Table 2). Adding retention to attention buys no consistent improvement: combined − FastV contains zero at nearly every budget (and trends positive for Gemma).

On the primary causal-cost metric the super-patch ordering is unambiguous — attention > random ≈ deconfounded retention — and at token granularity attention dominates deconfounded retention at every budget (Table 3). It mirrors the correlational result exactly: attention weakly tracks utilization (§4.3) and helps compression; pixel-decodable retention positively tracks neither — at its strongest it weakly marks what is *safe to discard*. What the cache reconstructably retains is not what is worth keeping.

5 Discussion

Why is decodable content retained but not used? The dissociation is less paradoxical than it first appears once retention and utilization are seen as answering different questions about the same KV vector. Our inversion decoder asks whether pixel content is *recoverable* from a unit's KV by a learned (non-linear) decoder; causal ablation asks whether the model's forward computation *routes the answer through* that unit. A representation can be rich in recoverable content yet not be the locus the computation reads from. The architecture result (§4.4) suggests a mechanistic reading. In an encoder-free VLM, pixels enter the language model directly and their content remains decodable at the layer we probe, with no ViT bottleneck to re-encode them first. But being decodable at a spatial unit is not the same as being *used* per spatial unit: the information relevant to the answer is mixed across positions, so removing any single super-patch's KV rarely severs a load-bearing path (§4.2), even as the raw content sits there to be decoded. Attention, meanwhile, is trained toward task relevance rather than information content, which is why it — the only proxy we measure that does — weakly tracks utilization (§4.3). Retention measures what the cache *contains*; attention approximates what the computation *prefers*; and, among the proxies we measure, only the latter bears any relation to what the computation *needs*.

From probing principle to cache-compression evidence. The established decodable-≠-causal lens (§2) becomes operational here in three ways. First, the relevant hypothesis is not folklore but *practice*: KV- and token-compression methods do use reconstruction- and information-content signals, and one cannot know a priori whether pixel-decodable retention is among the signals that work — we show, causally, that it is not. Second, the effect is quantified, not asserted: a preregistered, well-powered, bounded null (not "we did not find an effect") with a working positive control on the same estimator, which distinguishes genuine inertness from an insensitive measurement. Third, the null is not global — attention *does* track utilization — so the finding is a specific claim about *which* axis is inert, not a blanket "nothing correlates." The contribution is a map and a causal instrument for drawing it, of which this dissociation provides one concrete entry.

Implications for KV-cache compression. The engineering reading of these results is a caution against reconstruction- or information-content-based importance scores for visual KV eviction. In our setting, deconfounded pixel-decodable retention ranks eviction no better than random at super-patch granularity (and helps at token granularity only through its weak inverse signal), whereas attention magnitude — the weak-but-real proxy — does at every granularity. Neither signal is strong (attention itself only reaches $\rho \approx 0.11$ with utilization), which suggests that, in our setting, single-proxy visual-cache pruning has a low ceiling and that causal or task-aware signals may be needed to push past it. A direct test reinforces the ceiling reading: a ridge combination of thirteen readily available per-unit signals (multi-layer attention variants, OCR density, position, pixel statistics), supervised on the calibration split by the causal Δ_{task} readout itself, fails to outperform plain attention on held-out validation ($\rho = 0.07/0.06$ vs. $0.09/0.13$; within-image rank correlations without covariate partialling, hence differing slightly from Table 1) — the extractable per-unit importance signal in these features is essentially exhausted by attention alone. Importantly, this is not a claim against *feature-space* reconstruction methods, which reconstruct a different quantity and report real gains; our negative result is specific to pixel-level decodability used as an eviction score.

What attention selects is nearly question-invariant. A post-hoc analysis (Appendix A) adds a structural dimension to the low single-proxy ceiling. On the validation images that carry two official questions, the retained sets an attention-based policy would keep are nearly identical across questions (token-level Jaccard 0.87–0.95 for Gemma and 0.74–0.94 for InternVL across budgets, against random-set expectations of 0.14–0.60), while the per-unit utilization readouts under the two questions correlate only weakly ($\rho \approx 0.09$, interval including zero — bounded by per-unit reliability, so we do not conclude query-specificity; Appendix A). Read together with §4.5 and Appendix C, the picture is coherent: the one weak functional proxy

operates as question-independent salience rather than question-conditioned routing, and the retention axis relates to task structure only with inverted sign — spatially (it anti-localizes answer regions) and behaviorally (evicting its highest-scoring tokens is safer than random at the larger budgets). Both observations mark out the same boundary: per-unit, query-agnostic signals — whether informational or attentional — appear to face a structural ceiling in our setting, which is precisely where query-conditioned or causal signals would have to enter.

Limitations. Our claims are deliberately scoped to the setting we measured: two VLMs (one encoder-free, one encoder-based), the TextVQA task, super-patch granularity, and a primary decoder layer. Several of these are load-bearing. The utilization readout is a single-unit ablation; importance that is redundantly distributed across units is not captured by the per-unit Δ_{task}, and for the encoder-based model the ViT's early mixing makes it a conservative lower bound (§3.2) — the cumulative-eviction analysis of §4.5 and the essentially-zero encoder-free estimate partly offset both concerns. TextVQA localizes answers to specific regions, which is what makes single-unit ablation informative — tasks with more diffuse visual dependence may show a different retention/utilization relationship. The architecture contrast rests on one model per class, so the 2.68× is a model-pair effect confounded with model scale and training; extending to additional encoder-free and encoder-based VLMs is what would license an architecture-*class* claim. The decoder-layer choice, by contrast, we tested directly: re-running the Gemma analysis at layers 0, 6, and 12 leaves the retention–utilization null intact at every depth (§4.2), so the task-inertness is not a primary-layer artifact; extending this depth sweep to the encoder-based model (whose secondary-layer decoders we did not train) remains future work. Finally, the cross-question invariance of attention-selected sets (§5, Appendix A) is itself a scoped, post-hoc observation on the official TextVQA question pairs, not a preregistered result.

Broader impact statement

This work evaluates measurement validity of importance signals for visual KV-cache compression. Its practical effect is to steer deployed-efficiency work away from a signal that does not withstand a causal test; we do not foresee direct negative societal impacts beyond those of VLM deployment generally.

Reproducibility statement

All quantities reported are computed from the official TextVQA validation set with preregistered splits and seeds; code, the per-unit measurement tables, the eviction harness, the frozen preregistration protocol (SESOI, thresholds, split procedure, sign-freezing rule), and the exact model checkpoints and revisions will be released.

References

Samhruth Ananthanarayanan, Ayan Sengupta, and Tanmoy Chakraborty. Understanding the physics of key-value cache compression for LLMs through attention dynamics, 2026. URL https://arxiv.org/abs/2603.01426.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2403.06764.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://arxiv.org/abs/2006.00995.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22826–22835, 2025. URL https://arxiv.org/abs/2412.13180.

Lianghuan Huang and Yingshan Chang. Causality ≠ decodability, and vice versa: Lessons from interpreting counting ViTs, 2025. URL https://arxiv.org/abs/2510.09794.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2503.03321.

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 *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.23416.

Jaeyeon Lee, Shunjie Wen, and Dong-Wan Choi. Moving beyond diversity: Visual token pruning as subspace reconstruction for efficient VLMs, 2026. URL https://arxiv.org/abs/2606.18681.

Yuping Lin, Jiayuan Ding, Yue Xing, Pengfei He, Jiliang Tang, and Subhabrata Mukherjee. A simple plug-in for improving eviction-based KV cache compression, 2026. URL https://arxiv.org/abs/2605.23258.

Sihao Liu, YuFan Xiong, Zhonghua Jiang, Zhaode Wang, Chengfei Lv, and Shengyu Zhang. RetentiveKV: State-space memory for uncertainty-aware multimodal KV cache eviction, 2026. URL https://arxiv.org/abs/2605.04075.

Aryan Sharma, Cutter Dawes, and Shivam Raval. Dissociating decodability and causal use in bracket-sequence transformers, 2026. URL https://arxiv.org/abs/2604.22128.

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*, 2025. URL https://arxiv.org/abs/2410.23317.

Chenyu Zhou, Qiliang Jiang, Shuning Wu, and Xu Zhou. The vision encoder as a privacy boundary: Visual-token side channels in encoder-free vision-language models, 2026. URL https://arxiv.org/abs/2606.14783.

A Post-hoc exploratory: cross-question stability of the attention axis

The analyses in this appendix were conducted after the preregistered study concluded, on the subset of validation images that carry two official TextVQA questions ($N = 34$); they are exploratory and were not preregistered.

A natural question raised by §4.3 and §4.5 is whether the weak-but-real attention signal is *query-conditioned* in what it selects. We measured this directly: for each of the 34 validation images with exactly two official questions, we computed the attention score under each question (identical estimand to §3.3) and compared the retained sets an attention-based eviction policy would keep.

Attention retained sets are nearly invariant across questions. At token granularity — the granularity at which eviction methods operate — the Jaccard overlap between the top-b sets under the two questions is 0.87/0.91/0.95 for Gemma and 0.74/0.86/0.94 for InternVL at $b = 25\%/50\%/75\%$ budgets (cluster-bootstrap CIs within ± 0.04; random-set expectation 0.14/0.33/0.60). Super-patch granularity gives the same picture, so the invariance is not an aggregation artifact. Nor is it an attention-*sink* artifact: stripping the most sink-like positions (up to 20% of positions, ranked by image-independent mean attention rank) leaves the overlap essentially unchanged (e.g., InternVL $0.745 \rightarrow 0.728$ at the 25% budget), and only $\sim$20% of the cross-question intersection consists of sink positions — the invariance is carried by image-driven, not position-driven, attention. Qualitatively, a query-independent component of visual attention has been reported before (Kang et al., 2025); the observation here is quantitative and selection-level: *the sets an eviction policy retains barely change with the question* (in our setting: TextVQA official pairs, two questions per image).

What this does and does not imply. The modal gold answers of the two questions differ for 88% of these images, and the two questions' causal-utilization maps correlate only weakly ($\rho = 0.086$ $[-0.028, +0.203]$). We explicitly do *not* conclude from this that utilization is query-specific: the per-unit Δ_{task} readout's own reliability bounds what any such cross-question correlation can reach — cross-*model* same-question consistency ($+0.117$ $[+0.004, +0.222]$ on this 34-image subset; $+0.12$ $[+0.05, +0.20]$ on the full validation set, §4.2) is of the same magnitude as cross-*question* same-model consistency, so unit-level measurement noise cannot be excluded as the driver of both. One diagnostic did resolve cleanly: the same-question control on the second questions ($\rho(s_{\text{attn}}, \Delta_{\text{task}}) = 0.029$, anomalously below the $\approx$0.11 of §4.3) is driven by low-annotator-agreement questions — in the high-agreement stratum the control recovers to 0.101, matching the first-question baseline (0.109) on the same images. The instrument behaves normally when the gold target is clean.

Reading. The selection-level stability of attention across queries is consistent with the picture of §4: attention-based eviction (FastV) is robust (§4.5) yet couples to causal utilization only weakly (§4.3) — what it selects is a largely query-independent notion of visual salience. We report this as a scoped, post-hoc observation on official question pairs of a text-reading task; whether it persists under deliberately region-divergent queries (e.g., multi-page document QA) is an open question we do not answer here.

B Instrument details: the pixel-inversion decoder and s_{inv}

Decoder. For each probed layer ℓ, the inversion decoder maps the layer-ℓ visual KV of an image-only prefill back to pixel space. Each 4096-dimensional KV token is projected linearly to a 48-channel spatial cell on the token grid; a convolutional stack (Conv→GELU, ConvTranspose upsampling ×2 twice, interleaved 3×3 convolutions) decodes the grid to a 3-channel image at 16 pixels per token cell ($\approx$3.37M parameters per layer). Decoders are trained with a pixel MSE objective (Adam, learning rate 3×10^{-4}, 40 epochs, batch size 4) on 200 images from the official TextVQA *training* split (50 held out for decoder validation), disjoint by official split from every image used in calibration or validation. Two decoders are trained per layer under independent seeds; the analysis uses their within-image z-scored ensemble, and the seed pair provides the split-half reliability used for disattenuation (§3.5).

Reconstruction score and s_{inv}. Reconstruction quality is measured by *gradient correlation*: the Pearson correlation between the concatenated horizontal and vertical spatial gradients of the decoded image and those of the ground-truth image, computed over the pixel region of a super-patch. For each super-patch u, the retention score is the shuffle-adjusted excess $s_{\text{inv}}(u) = g_u(\text{KV}) - \overline{g_u(\text{KV}_\pi)}$, where g_u is the regional gradient correlation with the KV intact and the second term averages three within-image random permutations π of KV positions. The shuffle baseline removes what any KV content would afford at that output location, so s_{inv} isolates position-specific decodable content; it is a dimensionless correlation difference, and the C3 contrast (§4.4) compares its per-unit means across models (ratio 2.68×, cluster-bootstrap 95% CI $[2.37, 3.06]$).

C Criterion validation: localizing OCR-grounded answer regions

The ablation readout Δ_{task} is our measure of causal utilization; a natural question is whether it has spatial validity of its own — whether high-Δ_{task} units coincide with where the task-relevant content actually sits. We validate this against an answer-localization criterion defined independently of any model-derived score: official Rosetta OCR boxes for the TextVQA validation images, matched to the answer string.

Setup. The preregistered protocol matches the normalized gold answer against OCR tokens with a three-tier rule (exact; multi-word subsequence; edit distance ≤ 2) and maps matched boxes to super-patches under a center-point criterion (IoU ≥ 0.1 as sensitivity). The implementation additionally admitted the ten normalized ground-truth answer variants as match candidates — a deviation from the preregistered gold-only wording that we report as the expanded surface, keeping the preregistration-faithful gold-only restriction as the primary reading; the two agree, and gold-only is stronger. Gold-only localizes the answer in N=52 of 72 held-out images (expanded surface N=59; positives per image 1:44, 2:11, 3:3, 4:1 on the expanded surface); unmatched images are spelling variants and reasoning-type answers, and they do not differ detectably from

matched images at baseline (gold log-probability and VQA-accuracy differences contain zero in the three model×metric cells with stored generative baselines; the Gemma VQA cell has none in the archived artifacts). For each image we rank the 16 super-patches by each score and compute answer-region AUC, aggregated with image-cluster bootstrap.

Δ_{task} localizes the answer region on both models. Gold-only AUC 0.791 $[0.703, 0.878]$ (Gemma) and 0.766 $[0.670, 0.849]$ (InternVL); expanded surface 0.723 $[0.626, 0.816]$ and 0.714 $[0.623, 0.808]$; all four exclude 0.5. The per-image paired contrast (answer vs. non-answer mean Δ_{task}, expanded surface) is 3.23 $[2.06, 4.53]$ and 4.16 $[2.09, 6.71]$. The ablation readout is thus anchored to an external criterion, not merely internally consistent. Attention localizes as well (0.731/0.800, expanded surface), consistent with §4.3. The pixel-edge control also localizes answer regions (0.707, both models) — text sits where texture is dense — which is precisely why the design residualizes against texture (§3.5); localizing the answer is not the same as predicting Δ_{task} (§4.3).

The retention score's answer-location association is inverted. We preregistered any departure of the retention AUC from 0.5 as an alarm on the C2 measurement, to be reported regardless of direction. The alarm fired, in the direction *away* from answer regions: AUC 0.393 $[0.322, 0.463]$ (Gemma) and 0.415 $[0.343, 0.485]$ (InternVL) on the expanded surface (0.408/0.371 gold-only). Two things follow. First, this is real, sign-invertible discriminative information: retention is not information-free about answer location. Second, it does not overturn the preregistered C2 result, which is a null of the *continuous* retention–utilization partial correlation; what the inverted AUC adds is that where retention relates to task-relevant location at all, the association points opposite to the direction any retention-as-importance proxy would need — consistent with the weak negative retention–attention correlation (C1) and with the token-granularity eviction result (§4.5).

Robustness. Dropping the edit-distance tier gives 0.734/0.713 on the expanded surface (N=44) and 0.846/0.759 gold-only (N=32); the IoU criterion agrees in direction (N=26: 0.770/0.633, the InternVL interval widening at this reduced N to include 0.5). The gold-only AUC cells were recomputed with an independent rank-sum implementation with exact agreement (8 summary and 416 per-image cells).