WHICH COMPUTATIONS DRIVE VISUAL-TO-TEXT TRANSFER? A CAUSAL LEDGER FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models can incorporate visual evidence into text states, but how much each decoder block contributes to the answer-relevant visual effect remains difficult to quantify. We introduce reciprocal nested interventions that decompose a common image-conditioned answer effect into ordered block contributions. By replacing visual rows at successive decoder boundaries while retaining text states, we measure how much of the visual effect has entered text states by each depth. These contributions sum exactly to the total effect, yielding a layerwise ledger of cumulative transfer and residual visual dependence. On Qwen3.5-0.8B and Qwen3.5-4B, just two dominant writer blocks carry 83% and 71% of the total visual effect, respectively. This rapid accumulation establishes early causal closures at $K12$ (0.8B) and $K16$ (4B). The ledger directly guides compression: closure determines when visual tokens can be evicted, while writer share prioritizes which blocks' K/V inputs to protect in BF16 under quantization. These predictions depend on task and readout: transfer schedules shift across tasks, and generation can still require late visual access beyond an immediate-answer closure.

1 INTRODUCTION

Understanding how visual information flows through vision-language models (VLMs) is central to interpreting their reasoning. Prior mechanistic studies have identified specific pathways for this flow, showing that visual evidence transfers to text states—such as question-token representations (Zhang et al., 2025c) or prediction positions (Neo et al., 2025)—during generation. Building on this, recent inference techniques like VTW have demonstrated that visual tokens can be entirely dropped at certain depths with little performance degradation in evaluated settings (Lin et al., 2025). However, a critical gap remains: knowing *when* visual tokens become dispensable does not explain *how* this visual-to-text transfer accumulated across preceding layers. While it is natural to hypothesize that earlier computations established this effect, current withdrawal criteria fail to quantify how much each specific block contributes to the final answer, or how this transfer dictates precision sensitivity.

Existing interpretability tools struggle to fully answer this question. While attention analysis characterizes visual access (Chen et al., 2024a; Zhang et al., 2025b; Xing et al., 2025), it does not in general reflect true causal contributions (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Conversely, causal tracing and activation patching (Basu et al., 2024; Salazar et al., 2026), along with pathway interventions (Zhang et al., 2025c; Neo et al., 2025), identify important components and connections. Yet, their intervention effects at different locations rarely sum up to the total visual effect due to internal model interactions and compensation. What is missing is a unified accounting framework—one where cumulative transfer, per-layer contribution, and residual visual dependence are all measured against the same answer-level effect. To address this, we ask: *which specific blocks establish the answer-relevant visual effect in text states, how much does each contribute, and how does this transfer relate to precision sensitivity and the necessity of subsequent visual access?*

To achieve this unified accounting, this paper proposes *reciprocal nested intervention* (see Figure 1 for an illustration of our setup and notation). For two matched images under the same question, we run the forward pass of one image up to block k. We then replace its visual-token representations with those of the other image while retaining its text-token representations, and read out the final

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

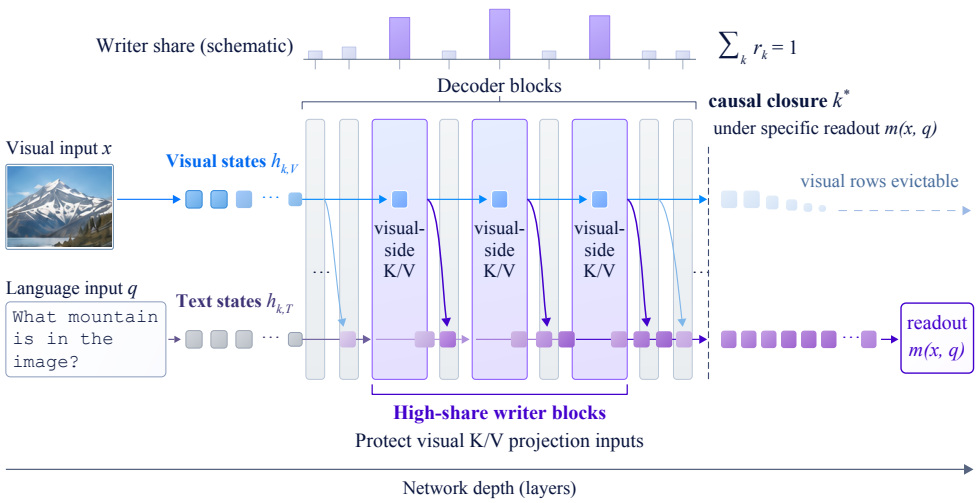

Figure 1: Schematic of visual-to-text transfer. Text queries read visual rows (blue) via the K/V projections of specific writer blocks (shaded), transferring answer-relevant evidence into text states (arrows). *Writer shares* r_k quantify each block's fractional contribution to the total answer effect (summing to 1). *Causal closure* k^* marks the boundary where residual visual dependence becomes persistently negligible.

prediction. Averaging the effects across both directions yields the *cumulative transfer* C_k: the exact amount of the total visual effect that has entered the text states by block k. By calculating adjacent differences $C_k - C_{k-1}$, we assign block-wise contributions that sum perfectly to the total visual effect. We term this layer-wise accounting the *transfer ledger*. As illustrated in Figure 1, this ledger yields two critical quantities for model compression: *writer share* (identifying specific blocks whose visual-row precision must be protected) and *causal closure* (predicting the earliest boundary where visual tokens can be safely evicted).

Empirically, applying this approach to Qwen3.5-0.8B and -4B (Qwen Team, 2026) reveals a highly concentrated transfer: the two largest writer blocks alone carry 83% and 71% of the total visual effect, establishing early causal closures at block 12 (0.8B) and block 16 (4B) where visual tokens can be safely dropped while still retaining 95% of the clean answer gap. This localized transfer also guides precision allocation. On Qwen3-VL (Qwen Team, 2025), the ledger-selected writer pair outperforms blocks selected by the attention-mass signals used by FastV and VTW. For the dominant writer $L7$ on Qwen3.5-0.8B, preserving only its K/V projection inputs in BF16 recovers 94% of the whole-block benefit. Despite this concentration, the ledger remains dynamically task- and readout-dependent. Reasoning-heavy MMStar items (Chen et al., 2024b) shift writer shares toward earlier blocks, while blocking late-layer visual attention during generation drops mathematics accuracy by up to 19 pp, proving that visual access remains critical after the prefill phase.

Our contributions are:

- We propose a quantitative accounting method to measure visual-to-text transfer, defining block-wise contributions and cumulative transfer against a unified answer-level effect.
- We reveal highly concentrated transfer dynamics in Qwen3.5 models, where just two blocks carry up to 83% of the visual effect, establishing early causal closures.
- We demonstrate that our transfer ledger effectively guides model compression, achieving strong performance retention in both visual token pruning and targeted quantization.

2 Related Work

Cross-modal information transfer in VLMs. Causal tracing places visual retrieval in early attention and feed-forward blocks and attributes it to a few image tokens (Basu et al., 2024). Attention knockouts trace image-to-question transfer through shallow and middle layers (Zhang et al., 2025c;

Yin et al., 2025; Kaduri et al., 2025), and later layers also read object information directly from visual tokens (Neo et al., 2025). Closest to ours, Salazar et al. (2026) show with counterfactual patching that the balance between direct and text-mediated routes depends on task, data, and prompt, and Wu et al. (2026) decompose the predictive information at each depth into modality-unique, shared, and synergistic components. Neither pathway effects nor information components sum to one answer-level total or say how much of it the visual rows still carry at a given depth. Our contribution is a fixed-order decomposition of a common answer-level effect, linking cumulative transfer, blockwise contributions, and residual visual dependence.

Visual-token withdrawal and precision sensitivity. FastV, SparseVLM, and PyramidDrop select visual tokens from attention or similarity signals inside the decoder (Chen et al., 2024a; Zhang et al., 2025b; Xing et al., 2025). Closest to the question of when visual rows are no longer needed, VTW withdraws all visual tokens at a depth chosen from output-distribution changes (Lin et al., 2025), DyVTE learns an input-dependent exit (Wu et al., 2025), and Information Horizon measures the prediction effect of deleting tokens at each depth (Wang et al., 2026); these criteria locate an endpoint, whereas the ledger also accounts for how the answer effect was assembled before it. MBQ, Q-VLM, and QIG calibrate uniform-precision quantization by modality balance, cross-layer dependence, or gradient-based sensitivity (Li et al., 2025b; Wang et al., 2024; Xiang et al., 2026), and failure analyses find that needed visual evidence shifts during reasoning (Kim et al., 2026) and that pruning and quantization alter circuits differently (Elluru et al., 2026). We use precision perturbations and token removal to test predictions of the transfer ledger, with port-level interventions further localizing the sensitive computations.

3 METHOD

We measure visual-to-text transfer by replacing visual states with those from a matched reference image while retaining the text states, then measuring the answer contrast that remains (Figure 2). We formalize this measurement as a transfer ledger with a closure criterion, validate the measurement, and derive testable predictions about visual-token eviction (physical removal) and precision perturbations.

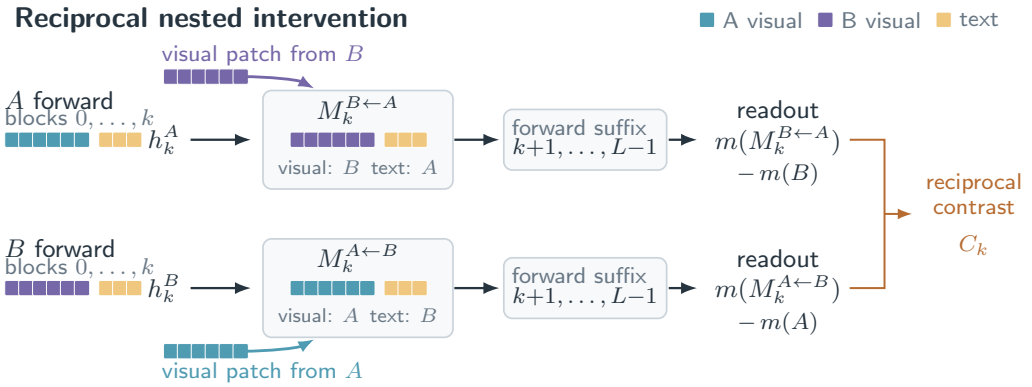

Figure 2: Pipeline of the reciprocal nested intervention.

3.1 SETUP AND THE TRANSFER ESTIMAND

Notation. An image x is encoded into visual tokens, concatenated with the text tokens of a question q, and passed through a decoder of L blocks. Let h_k be the hidden states at the output of block k, $k \in \{-1,\dots,L-1\}$, with h_{-1} the input embeddings. Blocks are zero-indexed; $L7$ denotes block 7, and $K12$ denotes the input to block 12. The rows are partitioned into *visual rows* V, contiguous and fixed across depth, and text rows; $h_{k,V}$ denotes the restriction to V, and the row class is the unit

of analysis throughout. Let $m(x, q)$ be a continuous answer score — in the certified instance below, the Yes/No log-likelihood margin on NaturalBench (Li et al., 2024).

Reciprocal nested intervention. To instantiate this comparison, let (A, B) be a semantically matched natural image pair sharing a question q, with directed total visual effect $\mathrm{TE} = s\,[\,m(A, q) - m(B, q)\,]$, where $s \in \{\pm 1\}$ aligns the semantic direction. Their visual rows match in count and position. The nested intervention $M_k^{B \leftarrow A}$ runs donor A to h_k^A, replaces *only* $h_{k,V}^A$ with clean recipient states $h_{k,V}^B$, retains all text rows, and runs the suffix to the readout. The reverse $M_k^{A \leftarrow B}$ is symmetric. We define *cumulative transfer* as the reciprocal, baseline-corrected contrast:

$$C_k \;=\; \frac{s}{2}\Big(\big[\,m(M_k^{B\leftarrow A}) - m(B)\,\big] - \big[\,m(M_k^{A\leftarrow B}) - m(A)\,\big]\Big). \tag{1}$$

C_k is the part of the A/B answer difference that has entered text states by block k, and can therefore still be read out once the visual rows are reverted. Two conventions anchor the ledger: the donor has not yet participated at $k = -1$, so $C_{-1} = 0$; and a restoration after the last block cannot affect the output, so $C_{L-1} = \mathrm{TE}$.

Writer contributions and exact accounting. Comparing adjacent intervention boundaries assigns the change in cumulative transfer to block k: $\Delta_k = C_k - C_{k-1}$. Under this fixed order, interactions with earlier states are attributed to the later block. Over a cohort of groups g we report ratios of means,

$$r_k \;=\; \frac{\mathbb{E}_g[\Delta_{k,g}]}{\mathbb{E}_g[\mathrm{TE}_g]}, \qquad F_k \;=\; \frac{\mathbb{E}_g[C_{k,g}]}{\mathbb{E}_g[\mathrm{TE}_g]}, \tag{2}$$

rather than averaging per-item ratios, which are unstable when the clean gap is near zero. We call r_k the *writer share* of block k and F_k the *completion* at k. Both are signed and are not truncated to $[0, 1]$, since under a fixed network order a block may apply a reverse correction or settle an earlier overshoot. Telescoping the ordered marginals between the two anchors yields an exact accounting identity,

$$\sum_{k=0}^{L-1} \Delta_k \;=\; \sum_{k=0}^{L-1} (C_k - C_{k-1}) \;=\; C_{L-1} - C_{-1} \;=\; \mathrm{TE}, \qquad \text{so} \qquad \sum_{k=0}^{L-1} r_k = 1. \tag{3}$$

Thus, the ledger accounts for the full visual effect under this ordered decomposition.

Causal closure. We next ask where dependence on visual rows becomes persistently negligible. The cohort-level remaining visual dependence is $\rho_k = 1 - F_k$. We freeze a practical-equivalence band $\epsilon = 0.05$ before inspecting any layer scan, take a 90% group-bootstrap interval for ρ_k, and define the *causal closure*

$$k^* \;=\; \min\Big\{\, k \;:\; \mathrm{CI}_{90}(\rho_j) \subseteq [-\epsilon,\, \epsilon] \;\text{ for all } j \geq k \,\Big\}, \tag{4}$$

with predicted eviction boundary $K^* = k^* + 1$, where eviction at K removes visual tokens before block K. The criterion asserts *equivalence* rather than absence of significance: the entire interval must lie inside the frozen band. The suffix condition $\forall j \geq k$ prevents a non-monotone curve from certifying a depth that later blocks contradict, and ϵ is a fraction of the clean visual effect rather than of accuracy. The ledger and closure are conditional on the model, stimulus distribution, and readout protocol.

Design choices. We use matched clean states rather than zero substitution, which can conflate information removal with the model's response to an anomalous intervention. Across four VLMs, zeroing the visual rows once and letting later blocks recompute them gives up to 35 pp lower accuracy than re-zeroing them after every block, and this gap is smaller for every nonzero replacement we tried (Appendix B). The matched donor also gives the intervention a semantic direction and a corresponding total effect TE.

3.2 Instrument validity

The accounting identity holds by construction, so it cannot by itself show that Δ_k measures transfer. We therefore check (a) that the implementation carries out the intended intervention exactly, (b) that an independent intervention finds the same writers, and (c) that the predicted closure holds when visual tokens are physically removed.

(a) Implementation fidelity. We audit eight quantities at every group and boundary on Qwen3.5-0.8B and Qwen3.5-4B, with zero tolerance: four algebraic endpoint checks and four implementation checks covering suffix equivalence, batch-path drift, visual-row restoration, and final recipient visual-state agreement. All eight discrepancies are exactly zero on both models (Appendices C and A).

(b) Convergent validity. We next test whether Δ_k identifies the same writer locations as an independent intervention. The *edge-pulse* intervention inserts donor visual rows for one block and restores recipient rows at its output. Unlike the donor-prefix marginal Δ_k, it isolates a single-block pulse, so we compare layer schedules rather than absolute magnitudes. Layerwise means correlate with Δ_k at Pearson 0.999 (0.8B) and 0.991 (4B), with exact recovery of the pre-registered top-k writers on both models (Appendix D).

(c) Predictive validity under visual-token eviction. The ledger predicts closure from state replacement alone, without deleting visual tokens. We test this prediction by eviction, measuring clean-gap retention as the mean image-conditioned answer gap after eviction divided by its clean reference. On the same NaturalBench cohort, the predicted $K12$ (0.8B) and $K16$ (4B) match the earliest boundaries certified to retain at least 95% of the clean gap; all deeper tested boundaries also pass (Figure 3). Accuracy changes there are $+0.59$ and $+0.39$ pp, respectively. Full scans and certification criteria are in Appendix E; cross-cohort transfer is tested in Section 4.2.

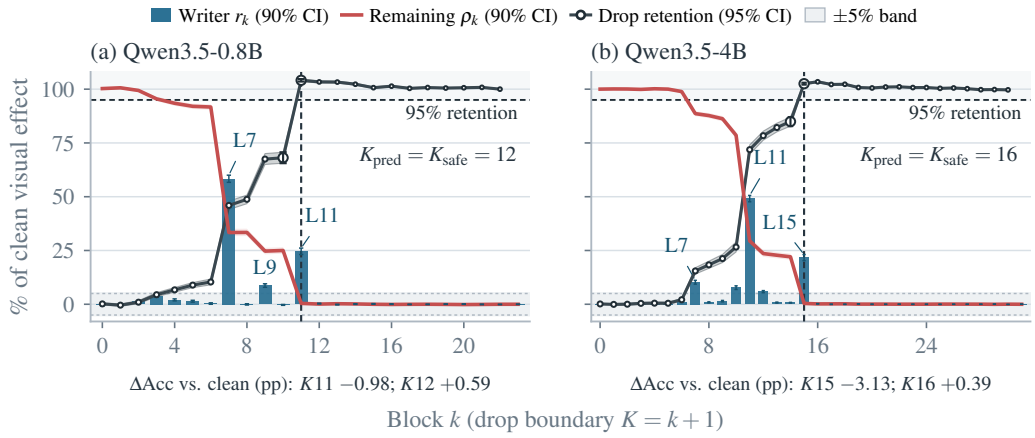

Figure 3: Transfer ledgers and physical-drop validation on NaturalBench.

3.3 From the ledger to compression decisions

We derive three deployment predictions from the transfer ledger before running the experiments in Section 4. Throughout, quantization applies only to visual-row activations.

P1: eviction at the closure. Under the same readout, eviction at K^* should preserve the clean visual effect, whereas earlier eviction should incur loss. Section 3.2 tests this prediction under the likelihood readout; Section 4.1 evaluates it at a deployment scale.

P2: protection ordered by writer share. If Δ_k captures the transfer performed by block k, quantizing a large writer should be more damaging than quantizing a small one. More strongly, the benefit of retaining high precision at a writer should scale with its writer share. We test this proportional prediction against non-writer blocks matched for quantization-error energy.

P3: the K/V port. Text queries read visual information through keys and values. With quantization confined to visual rows, protecting the K/V projection inputs should therefore recover most of the whole-block protection benefit, whereas protecting Q or O should recover little.

The resulting phase rule. Together, P1–P3 yield a simple phase rule: use reduced precision before closure, evict visual tokens afterward, and prioritize high-share writers' K/V inputs for precision protection.

4 EXPERIMENTS

Setup for Sections 4.1 and 4.2. Both sections use two 4B decoders: Qwen3-VL-4B-Instruct (Qwen Team, 2025), with 36 full-attention blocks, and Qwen3.5-4B (Qwen Team, 2026), with 32 blocks that interleave full and gated linear attention. We evaluate on EmbSpatial-Bench (Du et al., 2024), POPE (Li et al., 2023), and CV-Bench (Tong et al., 2024), reporting accuracy (F1 for POPE); later subsections state their own models and tasks. Each ledger comes from a matched-pair cohort as in Section 3 (NaturalBench for Qwen3.5-4B, a 216-group synthetic cohort for Qwen3-VL-4B; Appendix F), and no benchmark result is used to choose the boundary or the writers. Ledger construction caches boundary states and replays only each suffix, requiring approximately $L{+}1$ forward-equivalents per item; the NaturalBench runs take 24 and 22 minutes for Qwen3-VL-4B and Qwen3.5-4B, respectively, on a single A100. We evict visual tokens at the predicted boundaries: the input of block 20 on Qwen3-VL-4B and of block 16 on Qwen3.5-4B. The protected writers are each model's two largest-share blocks: 15 and 13 on Qwen3-VL-4B, 11 and 15 on Qwen3.5-4B. A4, A3, and A2 quantize visual-row activations to 4, 3, and 2 bits (asymmetric, per row) from block 3 through closure, and *protected* conditions keep the key and value projection inputs of the two writers in BF16.

Tables 1 and 2 score the complete test sets. All runs use lmms-eval (Zhang et al., 2025a) on NVIDIA A100 GPUs with greedy decoding and identical prompts within each model–benchmark pair; decoding is deterministic, so each condition is evaluated once and quoted differences are point estimates.

4.1 CAUSAL TRANSFER PREDICTS COMPRESSION-SENSITIVE DEPTHS

Eviction boundary. To test the likelihood-calibrated closure on generative benchmarks, we sweep visual-token eviction at block input K with all other settings fixed. Each curve uses its deepest eviction boundary as the reference (Appendix F).

Both checkpoints show stepwise recovery aligned with the ledger writers (Figure 4). On Qwen3-VL, the largest gains occur when blocks 13 and 15 first retain visual access; on Qwen3.5, the boundary that first admits the dominant writer $L11$ lifts all three benchmarks. The ledger-predicted boundaries are the first where all benchmarks come within 1.1 points of their reference and remain there at deeper boundaries, while one block earlier POPE and CV-Bench are still more than a point below.

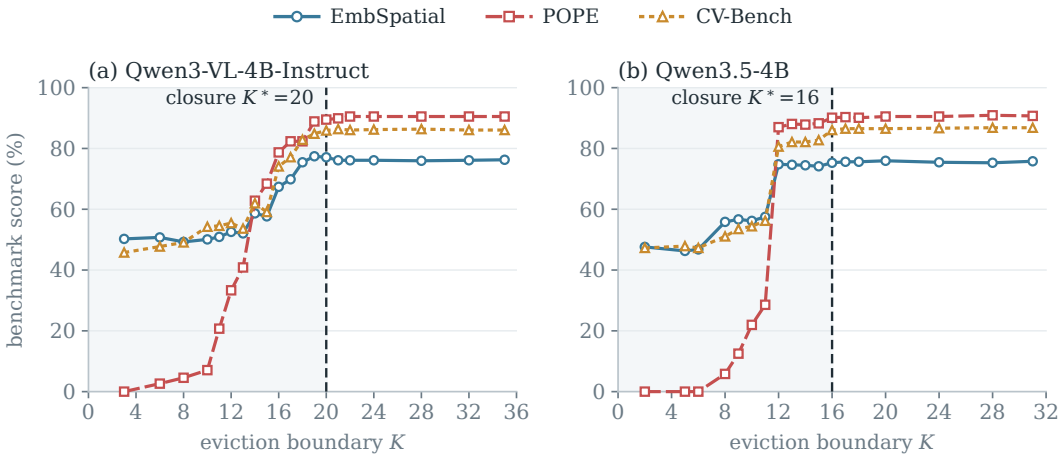

Figure 4: Eviction-boundary scans. Each point removes the visual rows at the input of block K and scores one benchmark; the dashed line marks the certified boundary. POPE is scored by F1.

Table 1: Writer selection at a fixed compression budget. Each column protects two zero-indexed decoder blocks, that is four key and value projection inputs.

	Qwen3-VL-4B			Qwen3.5-4B	
Selector	Causal	FastV	VTW	Causal	Control
Protected blocks	15+13	8+15	18+17	11+15	7+3
Combined writer share	49.4%	28.6%	16.6%	70.8%	7.6%
A4, EmbSpatial	75.59	74.45	75.11	72.42	70.66
A4, POPE	83.91	81.81	83.68	85.54	85.21
A4, CV-Bench	83.57	82.07	83.02	81.45	80.84
A3, EmbSpatial	65.58	63.87	64.84	67.42	56.18
A3, POPE	53.19	46.71	48.83	78.21	68.43
A3, CV-Bench	66.41	63.12	60.84	76.15	72.55
A2, EmbSpatial	62.69	58.43	54.37	66.21	45.08
A2, POPE	38.66	19.61	27.92	77.11	10.01
A2, CV-Bench	56.03	50.72	49.36	73.99	47.35

EmbSpatial on Qwen3.5 is already flat from $K{=}12$, making the predicted eviction boundary conservative for earlier-recovering benchmarks. Neither ledger saw these items or generative scoring, so the agreement provides out-of-sample support rather than a new equivalence certificate.

Writer selection. To test P2, we fix the boundary, bit width, and a two-block protection budget, varying only the protected blocks: the two largest-share writers, or the top two full-attention blocks by visual-column attention at the final prompt position (FastV) or averaged over text rows (VTW; Appendix F). On Qwen3.5 both attention signals select the writer pair, so the control instead protects the next two full-attention blocks by writer share, L7 and L3, testing whether writer identity rather than merely protection budget explains the gain.

The causal pair leads in every cell of Table 1. Margins are within 2.1 points under A4 and widen as precision falls, reaching tens of points on POPE at A2. Attention recovers the writers on one checkpoint but not the other, whereas the ledger ranks the blocks worth protecting on both.

4.2 Transfer-aware compression across decoder variants

We apply the ledger's eviction boundary and writer protection together on both decoders, without benchmark-specific tuning (Table 2). Eviction alone costs at most 1.3 points. Quantization costs substantially more, with the loss growing steeply as precision falls; at A2, POPE F1 drops to near zero on both models. Keeping the two writers' K/V inputs in BF16 recovers part of the loss in every cell, most visibly on POPE at A2 ($0.00 \rightarrow 38.66$ on Qwen3-VL-4B, $0.66 \rightarrow 77.11$ on Qwen3.5-4B). At A3 and A2 the recovered fraction is larger on Qwen3.5-4B, whose protected pair carries 70.8% of the transfer, than on Qwen3-VL-4B, where it carries 49.4%, consistent with P2. Thus, both ledger-selected decisions remain predictive when combined across the two tested decoders.

Table 2: The transfer-aware phase rule across checkpoints and activation bit widths. EmbSp. and CV-B. abbreviate EmbSpatial-Bench and CV-Bench.

	Qwen3-VL-4B			Qwen3.5-4B		
Condition	EmbSp.	POPE	CV-B.	EmbSp.	POPE	CV-B.
BF16, uncompressed	76.90	89.28	85.25	75.33	89.01	85.37
Eviction only	77.14	88.01	85.10	75.38	88.15	85.67
+ A4	73.13	79.95	81.46	69.09	81.58	78.96
+ A4, protected	75.59	83.91	83.57	72.42	85.54	81.45
+ A3	57.31	30.40	53.94	46.46	38.33	52.23
+ A3, protected	65.58	53.19	66.41	67.42	78.21	76.15
+ A2	42.47	0.00	42.00	39.64	0.66	39.76
+ A2, protected	62.69	38.66	56.03	66.21	77.11	73.99

4.3 WHAT EXPLAINS WRITER PROTECTION

All ablations use Qwen3.5-0.8B, whose predicted eviction boundary at $K12$ was validated in Section 3.2. Visual-row activations are quantized over $L3$–$L11$, and protected inputs remain in BF16. We first re-run the ledger on the NaturalBench cohort to measure how quantization changes transfer. Protection is then evaluated on 2,944 A-OKVQA (Schwenk et al., 2022) questions (four-choice likelihood, eviction at $K12$; Table 3), testing whether its benefit tracks writer identity, which input port carries it, and whether weak precision sensitivity implies dispensability.

Quantization degrades transfer before it degrades accuracy. Under A4 quantization, the total visual effect has already fallen by a fifth, while accuracy shows no detectable change; task accuracy can therefore conceal mechanistic degradation. At A2 only an eighth of the effect survives and the schedule is rerouted: its layerwise correlation with BF16 falls to 0.18, and the top writers move from $L7/L11/L9$ to $L9/L4/L5$. Dose curves for accuracy and writer marginals are in Appendix G.

Table 3: Recovery when only the protection target varies. A-OKVQA ($n = 2944$), visual-row A2 over $L3$–$L11$ with eviction at $K12$; BF16 59.48%, unprotected 42.19%, gap 17.29 pp. Energy is the target's share of the window's quantization error energy. Point estimates; intervals and max-T tests in Appendix H. †not significant under family-wise max-T.

Protected input activations	**Energy**	**Recovery (pp)**	**Share of gap**
$L7$ block (writer, full attention)	17.6%	+7.64	44%
$L3$ block (non-writer, same module type)	21.0%	+2.68	16%
$L4{+}L5$ blocks (early GDN)	13.2%	+1.56	9%
$L8{+}L9$ blocks (late GDN, energy-matched to the early pair)	14.0%	+0.44†	3%
$L7$ attention ($Q/K/V/O$)	10.3%	+7.24	42%
$L7$ MLP	7.3%	+0.41†	2%
$L7$ K/V only	6.8%	+7.17	41%
$L7$ Q only	3.4%	−0.07†	0%
$L11$ K/V only (second writer)	8.7%	+3.64	21%

Writer identity survives matched controls (P2). Table 3 fixes the quantization setting and varies only which inputs stay in BF16. Protecting writer $L7$ recovers 4.96 pp more than $L3$, despite the same module type and greater quantization-error energy at $L3$. For depth, we interpolate the recovery trend between two error-energy-matched Gated DeltaNet (GDN) pairs bracketing $L7$ ($L4{+}L5$ and $L8{+}L9$); $L7$ exceeds it by 6.78 pp. Both contrasts are max-T significant at $p < 10^{-5}$ (Appendix H), so the benefit tracks writer identity beyond module type, error energy, and depth. The protection benefit also roughly tracks writer share: $L7$'s K/V recovers 2.0× as much as $L11$'s, versus a 2.35× writer-share ratio (Appendices H and G).

The benefit is carried by the K/V port (P3). Protecting only $L7$'s K/V inputs recovers 94% of the whole-block benefit, although these activations account for just 6.8% of the window's quantization-error energy. The MLP and Q together hold most of the block's error energy, yet neither yields significant recovery after max-T correction (Table 3); neither does O (Appendix H). Error energy alone therefore does not explain the protection pattern across blocks or within $L7$. A separate 256-item W4 pilot finds no reliable writer-specific benefit from keeping the writers' attention weights in BF16 relative to a matched control pair (Appendix I).

Protection depends on continued visual access. Whole-block protection of $L7$ yields an additional 1.46 pp of recovery under eviction (Appendix J), consistent with a compensation account: while visual rows remain available, later blocks can partly re-read evidence that a corrupted writer failed to transfer.

Presence matters more than precision. The small benefit from protecting MLP precision does not mean the visual MLP is dispensable: disabling it entirely costs 10.94 pp where running it at A4 costs 2.34 pp in the same condition. Before closure, the prescription is reduced precision, not skipped computation.

4.4 TASK-DEPENDENT TRANSFER AND VISUAL ACCESS DURING DECODING

We examine how task distribution changes the transfer pattern and whether generated-token queries still depend on visual access in late blocks.

Tasks change writer allocation. On Qwen3.5-4B, we build capability-specific ledgers on MMStar (Chen et al., 2024b), sampling donors within each capability under a pre-registered rule and scoring transfer with a multiple-choice margin (Appendix K). In pre-registered contrasts, reasoning-heavy items assign about 14 percentage points more writer share to $L7$ and 8 points less to $L15$ than perception-and-instance items. They also show 8.22 points less residual visual dependence at $K12$, but by $K16$ the residual difference is no longer reliable. Thus, different task groups can reach similar late-stage residual dependence through different layerwise transfer schedules.

The same endpoint is not equally safe. Under generative evaluation, physical eviction at the likelihood-calibrated boundary $K16$ reduces overall MMStar accuracy by 5.53 pp, with losses concentrated in logical reasoning and mathematics (Appendix K.1). This motivates a direct test of visual access during decoding, with prefill held fixed.

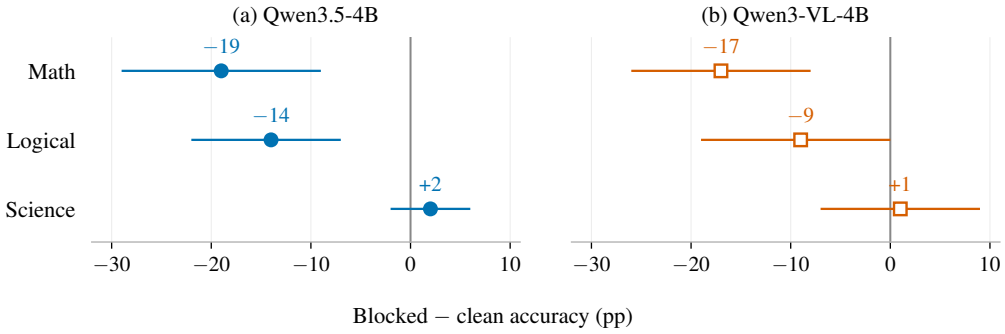

Figure 5: Decode-only visual blocking on MMStar for Qwen3.5-4B and Qwen3-VL-4B ($n = 100$ per task and model). Whiskers: paired-bootstrap 95% CIs. Model-specific protocols are given in Appendix L.

Late visual access exposes a readout mismatch. We block generated queries' visual-key attention in late blocks, with both arms sharing clean prefill and the first output token before generating freely. Sustained blockade reduces mathematics and logical-reasoning accuracy by 19 and 14 pp on Qwen3.5-4B. The dependence also appears in full-attention Qwen3-VL-4B, where mathematics accuracy falls by 17 pp (Figure 5; Appendix L). A plausible explanation is that tokens generated partway through a reasoning chain introduce visual retrieval demands not fully specified by the original question, and these details may be easier to read from refined visual representations in late blocks than from what prefill transferred into text states. Fixed-depth visual-token eviction is therefore vulnerable to this failure mode: a boundary calibrated at the answer position need not remain safe throughout generation.

5 CONCLUSION

In this work, we introduce an answer-level accounting framework for visual-to-text transfer, turning layerwise interventions into an exact decomposition of how visual evidence is established in text states. Across models and tasks, this accounting links the organization of transfer to downstream compression behavior while exposing an important limitation: visual-token dispensability under one readout does not imply independence from late visual access in reasoning-intensive generation. More broadly, this view turns visual-token eviction from a depth-selection problem into a question of how visual evidence is established, preserved, and subsequently accessed.

AI USE STATEMENT

We used AI-assisted tools primarily for language polishing and code development. All AI-assisted code was manually reviewed and verified by the authors to ensure correctness and consistency with the intended methodology.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun 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, 2025. arXiv:2502.13923.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 7400–7426, 2024. doi: 10.52202/079017-0237. arXiv:2406.04236.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. arXiv:2405.09818.

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)*, volume 15139 of *Lecture Notes in Computer Science*, pp. 19–35, 2024a. doi: 10.1007/978-3-031-73004-7_2. arXiv:2403.06764.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024. doi: 10.18653/v1/2024.acl-short.33.

Veeraraju Elluru, Arth Singh, Roberto Aguero, Ajay Agarwal, Debojyoti Das, and Hreetam Paul. Mechanistically interpreting compression in vision-language models, 2026. arXiv:2603.25035.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558, 2025. arXiv:2411.17491.

Jiwan Kim, Kibum Kim, Wonjoong Kim, Byung-Kwan Lee, and Chanyoung Park. Why and when visual token pruning fails? a study on relevant visual information shift in mllms decoding. In *European Conference on Computer Vision (ECCV)*, volume 17046 of *Lecture Notes in Computer Science*, pp. 244–263, 2026. doi: 10.1007/978-3-032-37281-9_15. arXiv:2604.12358.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, pp. 17044–17068, 2024. doi: 10.52202/079017-0542.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. URL https://openreview.net/forum?id=zKv8qULV6n. arXiv:2408.03326.

Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, Shengen Yan, Huazhong Yang, and Yu Wang. Mbq: Modality-balanced quantization for large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4167–4177, 2025b. doi: 10.1109/CVPR52734.2025.00394. arXiv:2412.19509.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5334–5342, 2025. doi: 10.1609/aaai.v39i5.32567. arXiv:2405.05803.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.07149.

Qwen Team. Qwen3-vl-4b-instruct. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen technical blog, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Israfel Salazar, Stella Frank, Dan Oneata, Desmond Elliott, and Constanza Fierro. Pathways of visual information flow in vision-language models, 2026. arXiv:2607.03358.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*, pp. 146–162, 2022. doi: 10.1007/978-3-031-20074-8_9.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356, 2024. doi: 10.52202/079017-2771.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Q-vlm: Post-training quantization for large vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 114553–114573, 2024. doi: 10.52202/079017-3638. arXiv:2410.08119.

Yahong Wang, Juncheng Wu, Zhangkai Ni, Longzhen Yang, Yihang Liu, Chengmei Yang, Ying Wen, Lianghua He, Xianfeng Tang, Hui Liu, and Yuyin Zhou. When token pruning is worse than random: Understanding visual token information in vllms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31910–31919, 2026. arXiv:2512.07580.

Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1002.

Hongxuan Wu, Yukun Zhang, and Xueqing Zhou. How vision becomes language: A layer-wise information-theoretic analysis of multimodal reasoning, 2026. arXiv:2602.15580.

Qiong Wu, Wenhao Lin, Yiyi Zhou, Weihao Ye, Zhanpeng Zeng, Xiaoshuai Sun, and Rongrong Ji. Accelerating multimodal large language models via dynamic visual-token exit and the empirical findings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 186578–186603, 2025. doi: 10.52202/085713-5612. arXiv:2411.19628.

Ziwei Xiang, Fanhu Zeng, Hongjian Fang, Rui-Qi Wang, Renxing Chen, Yanan Zhu, Yi Chen, Peipei Yang, and Xu-Yao Zhang. Fine-grained post-training quantization for large vision language models with quantization-aware integrated gradients. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3500–3510, 2026. arXiv:2603.17809.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Conical visual concentration for efficient large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14593–14603, 2025. arXiv:2410.17247.

Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in MLLMs: Unlocking pathways to faster inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9382–9391, 2025. arXiv:2503.13108.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025a. doi: 10.18653/v1/2025.findings-naacl.51.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74840–74857, 2025b. arXiv:2410.04417.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791, 2025c. arXiv:2411.18620.

A DETERMINISM AND STATISTICAL CONVENTIONS

Deterministic execution and matched condition batching keep repeated measurements comparable, with each item's paired conditions on the same device. Framework determinism flags alone did not stabilize margins or correctness; we therefore pin the FLA autotuner and require two full-grid runs per device to agree exactly.

Shared statistical protocol. The protection, composability and task-map experiments use $10{,}000$ paired bootstrap resamples stratified by device, keeping all conditions for an item together, and $200{,}000$ sign-flip max-T draws. The relevant sections specify each testing family; tests are two-sided for protection/composability and one-sided for positive task-map writers. Reported 95% intervals are pointwise, not simultaneous; max-T p-values test zero effects, not equivalence. Protection/composability equivalence requires the entire interval inside the frozen $\pm\epsilon_{\text{acc}}$ band, $\epsilon_{\text{acc}} = 1.953125$ pp. This accuracy tolerance is distinct from the normalized-effect closure band in Section 3.1. Edge-pulse, zero-substitution, and decode-only experiments use the protocols stated in their respective sections; task-map contrasts specify their own interval level.

B ZERO SUBSTITUTION

The advantage of persistent over one-shot zeroing shrinks under the tested nonzero replacements, motivating the matched-state intervention in Section 3.1.

Interventions and readout. At the output of zero-based block d, Z zeros visual rows once; M applies the same initial zeroing and masks visual keys with $-\infty$ in every subsequent *full-attention* block; Z_{pers} zeros visual rows after every block $\ell \geq d$. GDN/Mamba state pathways are not masked. These arms share the initial lesion, not their subsequent information content or computation. With $a(\cdot)$ denoting accuracy in percentage points, define

$$T = a(M) - a(Z), \qquad P = a(M) - a(Z_{\text{pers}}), \qquad R = a(Z_{\text{pers}}) - a(Z), \quad T = P + R.$$

Here P is the mask-versus-clamp contrast and R the zero-clamp advantage over release; the identity alone does not identify a pure softmax or semantic-regeneration mechanism. We evaluate 256 A-OKVQA items (indices 256–511) using BF16 eager inference and four-choice length-normalized likelihood. Pointwise 95% intervals use $10{,}000$ paired item-bootstrap resamples without multiplicity correction.

Mask, clamp, and release. Table 4 compares the three interventions on Chameleon-7B (Chameleon Team, 2024), Qwen2.5-VL-3B (Bai et al., 2025), LLaVA-OneVision-0.5B (Li et al., 2025a), and Qwen3.5-0.8B. These selected model–block pairs illustrate distinct contrast patterns rather than their prevalence across models.

Table 4: A-OKVQA accuracy contrasts ($n = 256$ per model): mask-minus-release (T), mask-minus-clamp (P), and clamp-minus-release (R). Entries are pp with pointwise 95% paired-bootstrap CIs; d is the zero-indexed intervention block.

Model/block	T [95% CI]	P [95% CI]	R [95% CI]
Chameleon $d15$	$+22.27$ $[+14.84, +29.69]$	-1.95 $[-6.25, +2.34]$	$+24.22$ $[+16.41, +31.64]$
Qwen2.5-VL $d19$	$+28.12$ $[+21.09, +35.16]$	-7.03 $[-12.89, -1.56]$	$+35.16$ $[+27.73, +42.58]$
LLaVA $d9$	$+12.89$ $[+7.81, +17.97]$	$+7.42$ $[+2.34, +12.50]$	$+5.47$ $[+0.78, +10.16]$
Qwen3.5 $d5$	$+1.95$ $[-1.56, +5.47]$	-2.73 $[-5.47, -0.39]$	$+4.69$ $[+1.17, +8.20]$

Clamp-versus-release effects do not exhaust the total discrepancy: P can reinforce or oppose R, even leaving a non-significant total despite significant components. In Qwen3.5, P also retains recurrent-state reachability. A zeroing curve alone therefore cannot separate semantic dependence from the response to the intervention.

Replacement controls. For operator o, one-shot replacement acts only after d; persistent replacement clamps after every $\ell \geq d$ to that layer's target r^o_ℓ. Define the response $S_o = a(\text{persistent}(o)) - a(\text{one}(o))$; only $S_{\text{zero}} = R$. Mean targets broadcast per-layer/channel visual-row means from 128 calibration images. Noise uses Gaussian draws standardized per channel before matching those calibration means and standard deviations. Cross-50% patches a fixed random half of positions with same-position clean donor rows, leaving the rest free; perm applies a fixed row permutation to the recipient's clean trajectory (exploratory). Thus these are distinct replacement controls, not uniformly distribution-preserving interventions. Chameleon uses 256 items; each Qwen model uses 252, excluding four singleton image grids under the pre-specified exact donor-grid matching rule. Table 5 reports the corresponding persistent-minus-one-shot accuracy differences.

Table 5: Persistent-minus-one-shot accuracy for zero and nonzero replacements on A-OKVQA: $n = 256$ for Chameleon and $n = 252$ for each Qwen model. Entries give S_o in pp with pointwise 95% paired-bootstrap CIs; positive values favor persistent replacement.

Model/block	Zero	Mean	Noise	Cross-50%	Perm
Qwen2.5-VL $d19$	$+36.11$ $[28.57, 43.65]$	-0.40 $[-3.17, 2.38]$	$+0.40$ $[-1.98, 2.78]$	0.00 $[-1.19, 1.19]$	$+0.40$ $[0.00, 1.19]$
Chameleon $d15$	$+24.22$ $[16.41, 31.64]$	-15.23 $[-23.44, -7.03]$	$+0.78$ $[-5.47, 7.42]$	$+0.39$ $[-0.78, 1.95]$	-0.78 $[-2.34, 0.78]$
Qwen3.5 $d5$	$+4.76$ $[1.59, 8.33]$	-1.59 $[-6.35, 3.17]$	-3.97 $[-7.14, -1.19]$	-0.40 $[-1.98, 1.19]$	$+0.40$ $[-1.59, 2.38]$

All twelve paired interactions $S_{o\neq\text{zero}} - S_{\text{zero}}$ have pointwise 95% intervals below zero, with estimates from -4.37 to -39.45 pp; this is not a family-wise certification. A zero-defined horizon therefore need not identify semantic transfer.

C AUDIT QUANTITIES

All eight audit discrepancies are exactly zero on Qwen3.5-0.8B and Qwen3.5-4B, supporting implementation fidelity in Section 3.2. The checks use zero tolerance. The grids contain 256 groups, two questions per group, and 25 or 33 boundaries: 12,800 and 16,896 rows, respectively; reciprocal directions are evaluated within each question. In the tensor checks, maxima also cover both questions and directions (indices suppressed). Here $\mathbf{z}$ denotes the selected answer-scoring logits and $\mathbf{u} = \text{Norm}(\mathbf{h}_{L-1})$ the final normalized hidden states returned by the suffix implementation.

$$E_{\text{start}} = \max_g |C_{-1,g}|, \qquad E_{\text{end}} = \max_g |C_{L-1,g} - \text{TE}_g|,$$

$$E_{\text{tel}} = \max_g \left| \textstyle\sum_k \Delta_{k,g} - (C_{L-1,g} - C_{-1,g}) \right|, \qquad E_{\text{ledger}} = \max_g \left| \text{TE}_g - \textstyle\sum_k \Delta_{k,g} \right|,$$

$$E_{\text{suffix}} = \max_{g,k} \|\mathbf{z}^{\text{clean-suffix}}_{g,k} - \mathbf{z}^{\text{full}}_g\|_\infty, \qquad E_{\text{dup}} = \max_{g,k} \|\mathbf{h}^{\text{dup}}_{g,k} - \mathbf{h}^{\text{ctrl}}_{g,k}\|_\infty,$$

$$E_{\text{restore}} = \max_{g,k} \|\widetilde{\mathbf{h}}_{g,k,V} - \mathbf{h}^{\text{recip}}_{g,k,V}\|_\infty, \qquad E_{\text{traj,final}} = \max_{g,k} \|\widetilde{\mathbf{u}}^{(k)}_{g,V} - \mathbf{u}^{\text{recip}}_{g,V}\|_\infty.$$

Audit scope. The first four quantities test algebraic consistency, not causal validity. The tensor checks cover clean-suffix equivalence, duplicate donor-prefix states, restored visual rows, and final recipient visual states. Text-row preservation follows from restricting replacement to visual rows; E_{restore} measures only those rows, and $E_{\text{traj,final}}$ checks the suffix endpoint rather than the intermediate trajectory.

The clean-recipient trajectory interpretation assumes that altered text states cannot feed back into visual rows, as guaranteed by causal ordering when text rows follow the visual segment and intervening operations preserve this dependency. Suffix states are recomputed without reusing donor caches; cached decoding requires a separate state audit.

D EDGE-PULSE CONVERGENCE

Edge-pulse and nested interventions recover the same leading writers on both Qwen3.5 models, supporting the convergent-validity test in Section 3.2.

Protocol. For every real block k: run clean recipient B to the input of block k; replace only visual rows with donor A rows; run exactly one block, so that A's visual information can reach text states; restore B's visual rows at the block output, retaining the updated text states; run the clean recipient suffix and form the reciprocal semantic effect. Both models use the same NaturalBench calibration cohort of 256 groups (512 reciprocal questions), with BF16 eager inference, thinking disabled, and 121 visual tokens.

Table 6: Edge-pulse and nested schedules on NaturalBench (256 groups per model). Pearson correlation compares layerwise means; the final column reports top-three overlap.

Model	Blocks	Pearson	Edge top-3	Nested top-3	Result
Qwen3.5-0.8B	24	0.99894	$L7/L11/L9$	$L7/L11/L9$	3/3, exact
Qwen3.5-4B	32	0.99120	$L11/L15/L7$	$L11/L15/L7$	3/3, exact

The 0.8B rule required $L7$ and $L11$ in the nested top-3, $L9$ in the top-5, and recovery of at least two of the three edge writers; the 4B rule was the same with $L11/L15$ and $L7$. The recovery criterion was frozen before inspecting the nested results.

Layerwise comparison. Figure 6 normalizes both schedules by the nested run's mean clean total effect; difference intervals use paired group bootstrapping.

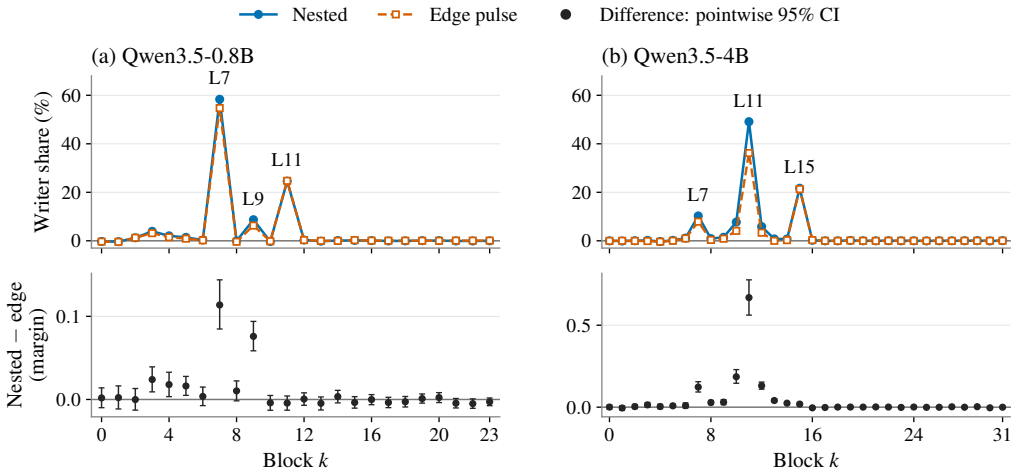

Figure 6: Full edge–nested schedules (top) and paired raw-margin differences with pointwise 95% intervals (bottom). All blocks are shown.

Nested shares exceed edge shares at early writers but nearly coincide at the final writer on both models, as expected when the ordered marginal assigns donor-prefix interactions to the current block.

E PHYSICAL-DROP VALIDATION

Physical-drop scans recover ledger-predicted boundaries on Qwen3.5-0.8B/4B, Qwen3-VL-4B, and LLaVA-OneVision-0.5B; Qwen3-VL-2B distinguishes retention safety from accuracy recovery (Section 3.2).

Shared criterion. Deletion occurs at the input of block K with surviving MRoPE positions preserved. Safety is defined on the retained clean-gap ratio: certified safe if the 95% lower bound is at least 95%, material loss if the upper bound is below 95%, and indeterminate otherwise. Intervals are pointwise, not simultaneous. Ledger and drop measurements are each normalized by their own clean reference, including in Figure 3; no common denominator is imposed across runs.

E.1 QWEN3.5 ON NATURALBENCH

Figure 7 shows all scanned boundaries; Table 7 retains exact outcomes near K^*. Accuracy changes are relative to clean and do not define safety. Every $K < K^*$ is materially unsafe and every $K \geq K^*$ is certified safe: $K^* = 12$ over $K1$–$K23$ on 0.8B and $K^* = 16$ over $K1$–$K31$ on 4B. On 0.8B, the largest upper bound before $K12$ is 70.68% and the smallest lower bound from $K12$ onward is 99.78%. On 4B, the largest upper bound before $K16$ is 87.02% and the smallest lower bound from $K16$ onward is 99.55%.

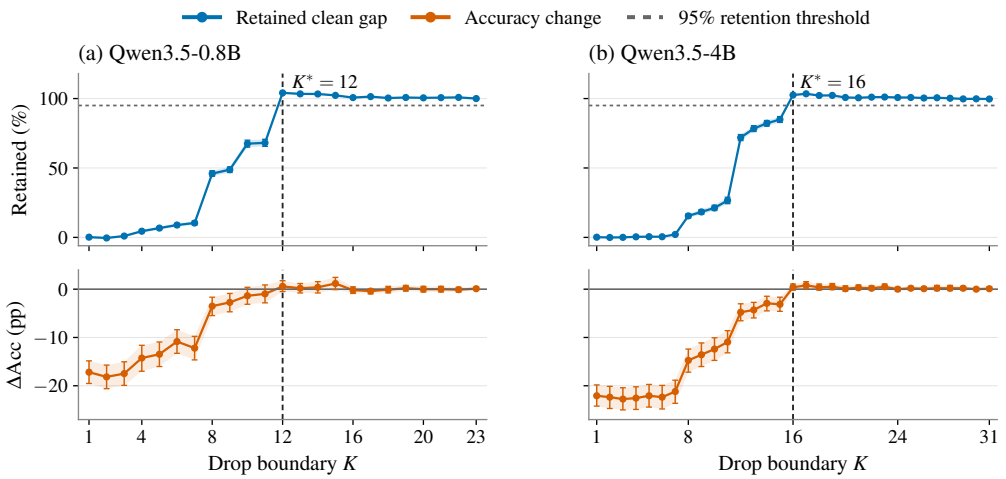

Figure 7: Full physical-drop scans. Whiskers and ribbons show pointwise 95% intervals; vertical lines mark K^*. Safety is determined by retention, not accuracy.

Table 7: Physical-drop outcomes adjacent to the predicted boundary. Intervals are 95%; bold marks K^*.

Model	K	Retained (%) [95% CI]	ΔAcc (pp) [95% CI]
Qwen3.5-0.8B	11	68.11 [65.58, 70.68]	−0.98 [−2.83, +0.88]
	12	104.12 [103.55, 104.71]	+0.59 [−0.49, +1.76]
	13	103.37 [102.86, 103.88]	+0.20 [−0.78, +1.17]
Qwen3.5-4B	15	84.93 [82.81, 87.02]	−3.13 [−4.59, −1.66]
	16	102.54 [102.10, 103.00]	+0.39 [−0.20, +1.07]
	17	103.43 [103.01, 103.87]	+0.78 [+0.10, +1.56]

E.2 QWEN3-VL AND LLAVA-ONEVISION

Figure 8 compares Qwen3-VL-2B/4B-Instruct with LLaVA-OneVision-0.5B (Li et al., 2025a). The Qwen3-VL runs use the 216-group synthetic cohort of Appendix F (864 predictions per condition), with the ledger's prompt and Yes/No log-likelihood margin. We scan every block input from $K3$ (after DeepStack injection) to the final block. The clean directed-gap means are 18.72 on 2B and 40.35 on 4B (ledger: 18.57 and 40.26, respectively); retention uses each scan's own reference and 95% group-bootstrap intervals. The LLaVA scan covers $K0$–$K24$, including embedding and terminal controls. All scans use the shared retention criterion.

On Qwen3-VL-4B, every $K \leq 18$ incurs material loss; $K19$ is indeterminate at 96.1% $[94.8, 97.3]$. Certification starts at $K20$ (101.0% $[100.7, 101.3]$), with accuracy change $+0.12$ pp $[0.00, 0.35]$, versus -1.97 pp $[-3.24, -0.81]$ at $K19$. All deeper boundaries pass and preserve all 864 predictions. Admitting blocks 13, 15, and 18 yields the three largest single-block gains.

On 2B, retention first passes at $K14$ (99.3% $[98.0, 100.5]$), but the accuracy change is -3.59 pp $[-5.21, -2.08]$. The ledger predicts $K15$, where accuracy recovers exactly, giving the more conservative boundary. Admitting writer block 11 lifts retention from 19.0% to 64.2%.

On LLaVA-OneVision-0.5B, $L9$ carries 51.2% of the total visual effect; $L9/L11/L12$ together account for 78.1%. Closure after $L12$ predicts $K13$, the earliest retention-certified suffix boundary. $K12$ is inconclusive; $K13$ retains 99.02% $[98.33, 99.74]$. Their accuracy changes are -4.00 and -0.39 pp, respectively.

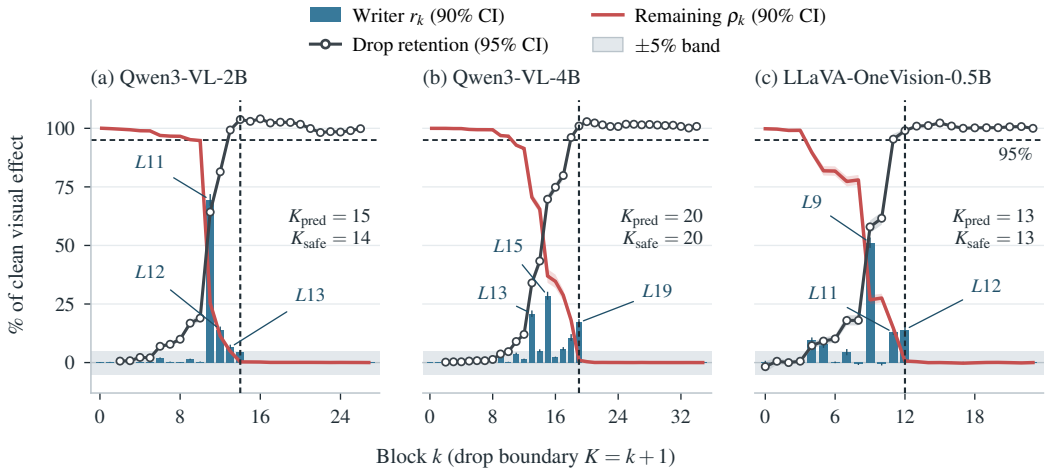

Figure 8: Transfer ledgers and physical-drop validation for Qwen3-VL and LLaVA-OneVision. Vertical lines mark K_{pred}; K_{safe} is the earliest boundary whose entire tested suffix satisfies the retention criterion.

F Deployment-scale ledger, boundary scans, and selectors

The Qwen3-VL-4B ledger selects writers $L15/L13$ and predicts boundary $K20$ for the deployment tests in Sections 4.1 and 4.2. We detail its calibration, the generative scan, and the attention-based controls here.

Ledger calibration. The Qwen3-VL-4B-Instruct ledger uses the estimand, nested intervention, and closure criterion of Section 3.1 on a cohort of 216 groups of synthetic matched image pairs, each pair differing in one attribute and sharing Yes/No questions scored by the log-likelihood margin. All strict audits of Appendix C are exactly zero. Cumulative transfer reaches 29.5% after block 13, 63.0% after block 15, 81.9% after block 18, and 99.0% after block 19, where the 90% interval of the remaining fraction first enters the $\pm 5\%$ band and stays inside it at every later block, giving closure $k^* = 19$ and boundary $K^* = 20$. The largest writer shares are block 15 at 28.6% $[26.8, 30.2]$ and block 13 at 20.9% $[19.5, 22.3]$, followed by blocks 19 (17.1%) and 18 (10.7%). Every block of this decoder is full attention, so all are protectable; the top two are protected in the main text. Physical drop on the same cohort confirms the boundary (Appendix E.2).

Scan grid. The scan of Figure 4 serves the deployed checkpoint with visual rows removed at the input of block K and scores the three benchmarks on fixed stride subsamples (POPE every 10th item, EmbSpatial every 6th, CV-Bench every 4th). Qwen3-VL-4B is scanned at $K \in \{3, 6, 8, 10, \ldots, 22, 24, 28, 32, 35\}$ and Qwen3.5-4B at $K \in \{2, 5, 6, 8, \ldots, 18, 20, 24, 28, 31\}$, consecutively across the whole recovery, so every step in the curves is resolved to a single block.

Qwen3-VL cannot be scanned below $K = 3$ because DeepStack injects multi-level visual features at blocks 0–2. The deepest boundary of each grid retains the visual rows through every block but the last and serves as the reference for its curve. Decoding is greedy.

Attention-based selectors. The FastV and VTW columns of Table 1 are layer selectors built from the attention signals those methods read, not reimplementations of the methods. Per-block attention mass on the visual columns is computed with the model's own eager attention on an unmodified BF16 forward, summed over heads, and averaged over 20 prompts from each of the three benchmarks; the FastV reading takes the final prompt position and the VTW reading averages over all text rows. The two full-attention blocks with the largest mass inside the quantization window are protected. On Qwen3-VL these are blocks $8{+}15$ and $18{+}17$; on Qwen3.5 both readings return the certified pair $11{+}15$, so the control protects the next two full-attention blocks by writer share, $7{+}3$.

G Dose curves under quantization

Visual-row quantization reduces the measured total visual effect before a detectable accuracy loss, supporting the dose-response result in Section 4.3.

We evaluate Qwen3.5-0.8B on the 256-group NaturalBench calibration cohort, quantizing visual-row activations over $L3$–$L11$. Quantization covers the inputs to all linear modules except linear-attention projections. Each condition measures all 24 block boundaries under the shared protocol (Appendices C and A). BF16 accuracy is 67.29% and $\mathrm{TE} = 3.134$.

At A4, TE retains 80.0% of BF16 while the accuracy-change interval includes zero. At A2, only 12.5% remains, schedule Pearson falls to 0.180, and the top-three writers change from $L7/L11/L9$ to $L9/L4/L5$ (Figure 9). $L7$ remains positive but tiny ($\Delta_7 = 0.026$, max-T $p = 0.0010$); $L11$ is no longer significant ($\Delta_{11} = 0.017$, $p = 0.182$). All conditions still close at $L11$, conditional on the surviving effect: A2's remaining fraction is 6.13% $[2.65, 9.38]$ after $L9$ and 1.84% $[-0.24, 3.82]$ after $L11$ (90% CIs).

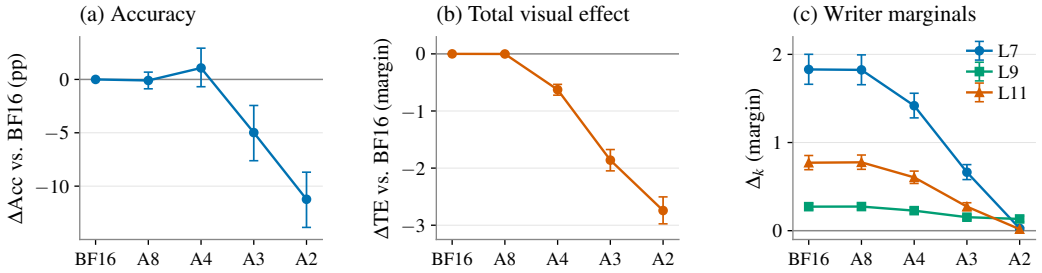

Figure 9: Dose curves under $L3$–$L11$ quantization. Whiskers are pointwise 95% group-bootstrap CIs; conditions are equally spaced.

Quantization-window control. Paired comparison with same-bit $L3$–$L9$ quantization isolates the addition of $L10/L11$: at A4 the TE change is -0.153 $[-0.185, -0.120]$ (95% CI); at A2, $L11$'s marginal falls from 0.393 to 0.017. The shorter window leaves a BF16 downstream recovery path, so its results cannot be pooled with the extended window.

H Protection attribution

Writer protection remains beneficial after controlling for depth, module type, and error energy, with most recovery concentrated in the K/V inputs (Section 4.3).

We compare protection targets on Qwen3.5-0.8B using $2{,}944$ A-OKVQA questions (indices 256–3199), scored by four-choice normalized likelihood. Visual-row activations use asymmetric per-

token A2 over $L3$–$L11$, followed by physical removal at $K12$; linear-attention projections remain in BF16. Inference follows Appendix A, with same-family max-T contrasts. Table 8 reports the targets and paired contrasts. BF16 is 59.48% and unprotected A2 is 42.19%, a gap of 17.29 pp.

Table 8: Protection targets and paired contrasts on Qwen3.5-0.8B, A-OKVQA ($n = 2944$), with visual-row A2 over $L3$–$L11$ and eviction at $K12$. Recovery is relative to unprotected A2 (42.19%); intervals are pointwise 95% CIs and p-values are family-wise max-T.

Target / contrast	Accuracy	Recovery (pp) [95% CI]	max-T p
$L7$	49.830%	$+7.643$ $[+6.352, +8.933]$	5×10^{-6}
$L3$	44.871%	$+2.683$ $[+1.529, +3.838]$	4.5×10^{-5}
$L4{+}L5$	43.750%	$+1.563$ $[+0.645, +2.480]$	0.00695
$L8{+}L9$	42.629%	$+0.442$ $[-0.170, +1.053]$	0.702
Same-family contrasts			
$L7 - L3$		$+4.959$ $[+3.635, +6.352]$	5×10^{-6}
$L7-$ position interpolation		$+6.781$ $[+5.626, +7.978]$	5×10^{-6}
$(L8{+}L9) - (L4{+}L5)$		-1.121 $[-2.072, -0.204]$	0.0847 (margin 5×10^{-6})
Operator and port within $L7$			
attention only	49.423%	$+7.235$ $[+5.944, +8.526]$	$< 5 \times 10^{-6}$
MLP only	42.595%	$+0.408$ $[-0.272, +1.053]$	0.675
K/V only	49.355%	$+7.167$ $[+5.944, +8.390]$	$< 5 \times 10^{-6}$
Q only	42.120%	-0.068 $[-0.815, +0.645]$	1.000
O only	41.576%	-0.611 $[-1.291, +0.068]$	0.309
Dual K/V ports			
$L7$ K/V only	49.35%	$+7.167$ $[+5.910, +8.424]$	$< 5 \times 10^{-6}$
$L11$ K/V only	45.82%	$+3.635$ $[+2.615, +4.688]$	$< 5 \times 10^{-6}$
both	52.00%	$+9.82$ $[+8.32, +11.31]$	$< 5 \times 10^{-6}$

Attention$-$MLP recovery is $+6.827$ $[+5.571, +8.118]$ (max-T $p < 5 \times 10^{-6}$). K/V also exceeds Q and O (both max-T $p < 5 \times 10^{-6}$), retaining 93.8% of full-block recovery. The attention/MLP split is additive-equivalent, with residual 0.000 $[-0.849, +0.883]$; the port split is not certified additive, with residual $+0.747$ $[-0.679, +2.208]$ extending beyond the frozen band.

Both conditional dual-port contributions remain positive (max-T $p < 10^{-5}$), but their interaction, -0.985 $[-2.310, +0.374]$ ($p = 0.535$), is inconclusive rather than additive-equivalent. The single-port recovery ratio is 1.97, compared against the writer-share ratio in Section 4.3. Protection-by-eviction contrasts are collected in Appendix J.

I Weight-only protection of the writer blocks

This pilot detects no reliable writer-specific benefit from protecting attention weights, complementing the activation-protection results in Section 4.3.

We evaluate Qwen3.5-0.8B on 256 A-OKVQA items (dataset indices 256–511), scored by four-choice likelihood. Under four-bit round-to-nearest weight quantization (W4 RTN) and visual-token eviction at $K12$, selected self-attention $Q/K/V/O$ weights remain in BF16. The writer pair is $L7{+}L11$; $L3{+}L15$ is a parameter- and shape-matched control (Table 9).

Writer specificity is $R(L7{+}L11) - R(L3{+}L15) = -3.125$ $[-6.250, 0.000]$, max-T $p = 0.160$; the margin specificity is -0.023 $[-0.128, +0.083]$, $p = 0.966$. These results do not establish equivalence between the protection targets.

J Interactions between quantization, protection, and eviction

Eviction increases the recovery from protecting $L7$, supporting the continued-access comparison in Section 4.3. Small quantization–eviction interactions, however, do not by themselves establish safe eviction.

Table 9: Weight-only protection on Qwen3.5-0.8B, A-OKVQA ($n = 256$), under W4 RTN and eviction at $K12$. Selected attention weights remain in BF16. Recovery is the accuracy gain over W4, in pp with 95% CIs.

Condition	Accuracy	Recovery (pp) [95% CI]
W4 baseline	52.734%	—
$L7$ attention BF16	53.516%	$+0.781$ $[-1.953, +3.516]$
$L11$ attention BF16	54.297%	$+1.562$ $[-0.781, +3.906]$
$L7{+}L11$ attention BF16	52.734%	0.000 $[-2.734, +2.734]$
$L3{+}L15$ attention BF16	55.859%	$+3.125$ $[-0.391, +6.641]$

Quantization–eviction interactions. Using the model and A-OKVQA cohort of Appendix H, we cross BF16/A4/A2/W4 with no eviction (E0) and eviction at $K8/K12/K16/K22$. A4/A2 quantize visual rows over $L3$–$L9$; W4 uses asymmetric group-128 weights excluding linear-attention projections. We use Appendix A's paired bootstrap protocol and max-T over each 12-interaction family.

Let $a(Q, K)$ denote accuracy in percentage points. We distinguish the eviction effect within a quantization setting from its change relative to BF16:

$$D_Q(K) = a(Q, K) - a(Q, \mathrm{E0}), \qquad I_{\mathrm{acc}}(Q, K) = D_Q(K) - D_{\mathrm{BF16}}(K).$$

For each quantity, equivalence requires the entire 95% interval inside the frozen $\pm\epsilon_{\mathrm{acc}}$ band (Appendix A). Thus, safety is assessed relative to the same-Q baseline, while interaction measures how quantization changes the effect of eviction (Figure 10).

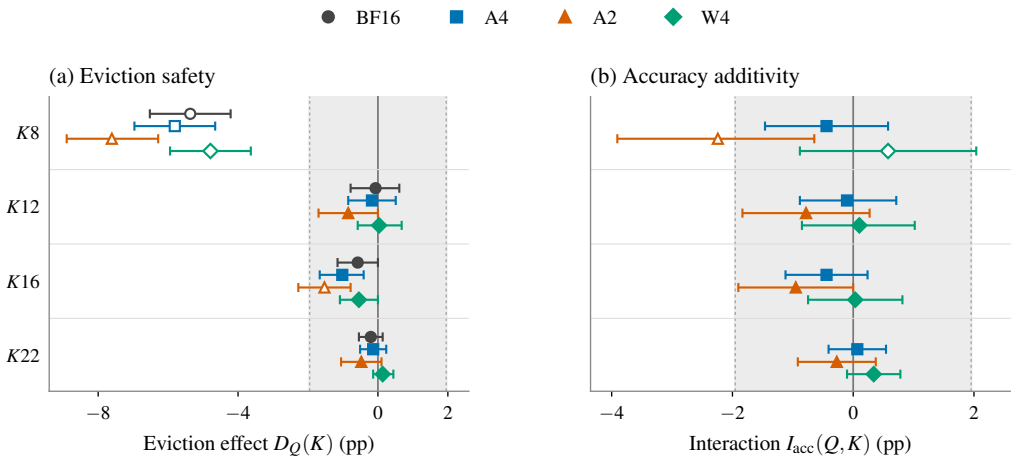

Figure 10: Eviction effects and accuracy interactions, with pointwise 95% CIs. Gray bands mark ± 1.953125 pp; filled markers have their entire CI inside the band. Panel axes use different ranges.

Ten of twelve accuracy interactions are equivalent to zero, including all nine at $K12/K16/K22$; A2/$K8$ and W4/$K8$ are inconclusive. The earliest tested boundary whose eviction-effect interval lies inside the band is $K12$ for all four settings. Yet A2/$K16$ fails safety equivalence ($D_Q = -1.53$ $[-2.28, -0.78]$ pp), despite an equivalent interaction. Approximate additivity therefore does not imply safe eviction at every deeper boundary. Additivity also depends on the outcome measure: margin interactions are not uniformly additive.

Writer protection under eviction. For the $L3$–$L11$ A2 configuration in Appendix H, let $R_T(K)$ be the accuracy recovery from protecting target T, relative to unprotected A2 at the same eviction setting. The interaction $R_T(K12) - R_T(\mathrm{E0})$ tests whether protection has the same benefit with and without eviction. The intervals for the non-writer block $L3$, early pair $L4{+}L5$, and late pair $L8{+}L9$

all fall inside the equivalence band. For the dominant writer $L7$, however, whole-block protection recovers an additional 1.46 pp under eviction (95% CI $[0.37, 2.55]$; max-T $p = 0.0356$), and the interval extends beyond the band.

The larger recovery under eviction is consistent with continued visual access supporting downstream compensation when rows remain available; this comparison does not directly isolate the compensating pathway.

K TASK-SPECIFIC TRANSFER AND READOUT MISMATCH

Reasoning-heavy MMStar capabilities allocate more transfer to $L7$ and less to $L15$ than perception-and-instance capabilities, while their late residual differences are not reliably resolved (Section 4.4).

Cohort and intervention. We evaluate Qwen3.5-4B on MMStar. Each capability is split into 16 pilot, 112 confirmation, and 122 holdout items. The descriptive map uses pilot and confirmation items; transfer-map inference uses only the 669 eligible confirmation items (of 672). Holdout items are excluded from donor selection, screening, and map estimation. The full-benchmark generative evaluation in Appendix K.1 is separate, not a holdout-only confirmation. Each item receives two same-capability natural donors drawn from the map cohort whose correct-option margin on the recipient question is ≤ 0; the readout is a four-choice length-normalized likelihood margin; images are aspect-preserving bicubic padded to 512×512 so that donor and recipient visual rows correspond one to one. Every donor runs all 33 nested boundaries. We use Appendix A's item-level protocol, keeping both donors together, with one-sided max-T over the 32 writer deltas.

Table 10: Capability-specific transfer on Qwen3.5-4B, MMStar: n counts the eligible confirmation items (669 total). Writer shares and remaining dependence are percentages of the clean visual effect, not accuracy; the NaturalBench row is a calibration reference.

		Writer share				Remaining dependence	
Capability	n	$L7$	$L10$	$L11$	$L15$	$K12$	$K16$
coarse perception	112	9.93%	15.51%	38.02%	17.97%	25.73%	1.71%
fine-grained perception	112	3.93%	12.81%	47.76%	15.45%	20.37%	0.67%
instance reasoning	111	13.68%	12.06%	42.39%	17.05%	25.86%	0.71%
logical reasoning	112	27.11%	8.89%	39.83%	10.24%	17.45%	1.80%
math	111	23.70%	10.02%	34.46%	7.11%	14.10%	1.33%
science & technology	111	19.83%	13.57%	32.49%	8.82%	16.70%	2.34%
confirmation overall	669	15.28%	12.44%	39.78%	14.01%	21.47%	1.36%
NaturalBench reference	—	10.25%	7.70%	49.13%	21.68%	—	—

Capability contrasts. Table 11 reports the pre-registered reasoning-heavy minus perception-and-instance comparison.

Table 11: Pre-registered task contrasts on Qwen3.5-4B, MMStar: logical reasoning, mathematics, and science ($n = 334$) minus perception and instance reasoning ($n = 335$). Differences are percentage points of writer share or remaining dependence, with 90% intervals.

Estimand	Difference [90% CI]	Decision
$L7$ writer share	$+13.94$ $[+8.68, +19.23]$	confirmed
$L15$ writer share	-8.05 $[-10.92, -5.17]$	confirmed
$L10$ writer share	-2.91 $[-5.98, +0.26]$	not confirmed
$L11$ writer share	-5.80 $[-11.81, +0.17]$	not confirmed
remaining at $K12$	-8.22 $[-12.56, -4.03]$	confirmed
remaining at $K16$	$+0.76$ $[-0.78, +2.29]$	no reliable difference detected
remaining at $K20$	$+0.72$ $[-0.28, +1.73]$	no reliable difference detected

The lower residual at $K12$ accompanies the shift toward earlier writers. The $K16/K20$ contrasts do not certify equal residual effects or identical capability-specific closures. The confirmation map

reproduces the pilot's aggregate $K16$ closure and top-writer identity, with a smaller $L7$ share. Two independent donors give an aggregate key-writer Pearson of 0.9990, the same top-3 and aggregate closure, with per-capability Pearson between 0.9778 and 0.9946.

Under the strict continuous-suffix criterion, five capabilities close at $K16$ and math at $K20$: although math is within the band at $K16$–$K18$, its $K19$ interval $[0.04, 5.02]\%$ exceeds the upper threshold by 0.02 pp. This small crossing still delays certified closure; fixed-endpoint comparisons cannot override the suffix criterion or establish a universal $K16$ closure.

K.1 Generative readout under physical eviction

On all $1{,}500$ MMStar items (250 per capability), eviction at $K16$ reduces Qwen3.5-4B's accuracy chiefly in logical reasoning and mathematics (Section 4.4). Table 12 reports micro accuracy and paired eviction-minus-BF16 differences, with pointwise 95% CIs from $10{,}000$ item-bootstrap resamples. The other four capability intervals include zero, which does not certify equivalence.

Table 12: MMStar accuracy under standard generation, with visual eviction at $K16$. Differences and pointwise 95% intervals are in percentage points.

Capability	BF16 (%)	Drop (%)	Drop−BF16 [95% CI]
coarse perception	67.60	67.20	-0.40 $[-2.00, +1.20]$
fine-grained perception	55.20	54.80	-0.40 $[-3.20, +2.40]$
instance reasoning	70.40	69.20	-1.20 $[-4.00, +1.20]$
logical reasoning	67.60	54.00	-13.60 $[-18.40, -9.20]$
math	80.00	62.00	-18.00 $[-24.40, -11.60]$
science & technology	56.00	56.40	$+0.40$ $[-3.60, +4.40]$
Overall (micro)	66.13	60.60	-5.53 $[-7.20, -3.87]$

L Late-layer visual access during decoding

To isolate decode-time dependence, we hold clean prefill and the first output token fixed; visual blocking still reduces mathematics accuracy on both tested decoders (Figure 5; Section 4.4).

Protocol. Both models use the same 300 MMStar items: 100 each from mathematics, logical reasoning, and science & technology, sampled from the non-holdout transfer-map cohort without accuracy or length filtering. Both arms share clean prefill and the first output token, then generate independently. The blocked arm masks each generated query's visual-key edges; visual K/V and all block computations are retained. BF16 greedy decoding uses a $1{,}024$-token budget and the original lmms-eval parser. Qwen3.5-4B uses the original prompt in non-thinking mode and blocks $L19/L23/L27/L31$. Qwen3-VL-4B-Instruct blocks $L20$–$L35$ at the preselected $K20$ boundary and removes only the trailing instruction "Answer with the option letter only."

Estimation and supplementary results. Pointwise 95% CIs use $10{,}000$ paired item-bootstrap resamples stratified by capability and device; exact McNemar tests are unadjusted. Pooled blocked-minus-clean accuracy changes are -10.33 $[-14.67, -6.00]$ pp for Qwen3.5 and -8.33 $[-13.33, -3.00]$ pp for Qwen3-VL. Retaining Qwen3-VL's letter-only instruction instead yields -1.00 pp and 291 clean single-token answers, versus 26 without it. This prompt comparison is descriptive; no prompt-by-blocking interaction was tested. The shared first token already supplies an option in $85/100$ Qwen3.5 science items. Unparsed responses count as incorrect, and token-capped items are retained. Correct-to-incorrect flips also occur without either arm reaching the token limit.