MEASURING, PREDICTING, AND DIAGNOSING COMPRESSION-INDUCED FAILURES IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual-token compression reduces the inference cost of vision-language models (VLMs). It is usually evaluated by average accuracy, which does not show which correct answers become incorrect. We characterize these *compression-induced failures* (CIFs) through paired reference and compressed runs, defining a CIF as a sample answered correctly by the reference but incorrectly after compression. We focus on post-prefill KV-cache eviction across ten benchmarks, three model families, four controlled retention policies, and adaptations of six published strategies. Methods with nearly equal accuracy often break different requests. On DocVQA, two adapted attention-based eviction methods differ by 0.19 points, yet 17.5% of the dataset fails under exactly one of them. The failures concentrate on benchmarks whose answers rely on small visual details such as text and layout. The paired runs also make these failures predictable and repairable. Output-confidence features rank CIFs at AUROC up to 0.89 and support selective fallback in paired simulation. Restoring pruned tokens repairs most failures, and at tight budgets score-guided selection repairs more than random or spatial selection. These results support reporting paired failure counts alongside average accuracy when evaluating compressed VLMs.

1 INTRODUCTION

Vision-language models (VLMs) encode each image into hundreds to thousands of visual tokens, and the KV cache built over these tokens takes most of the serving memory (Kwon et al., 2023). To reduce this cost, a large body of work discards most of the visual state, either by evicting KV-cache entries (Zhang et al., 2023; Liu et al., 2023; Li et al., 2024; Cai et al., 2024; Liu et al., 2024c) or by pruning visual tokens (Chen et al., 2024a; Zhang et al., 2025; Endo et al., 2025; Yang et al., 2025). These methods are usually evaluated by average benchmark accuracy at a matched budget. A method is regarded as safe to deploy if its accuracy stays close to that of the uncompressed model.

However, a deployed model serves individual requests, and average accuracy only counts how many of them break. We observe that the identity of the broken requests varies widely at the same accuracy. Figure 1 compares 116 pairs of compression conditions by their accuracy gap and by the fraction of samples that fail under exactly one condition of the pair. Adaptations of ElasticCache (Liu et al., 2024d) and H2O (Zhang et al., 2023), two attention-based eviction methods, differ by 0.19 points on DocVQA at 10% retention. Each of them breaks 25–26% of the dataset, and 17.5% of the dataset fails under exactly one of the two. Each of these requests is answered correctly under the other method, so the damage is method-specific rather than a shared set of hard samples. Moreover, the accuracy drop is a net quantity. New failures and reverse flips, i.e., samples that compression turns from wrong to right, cancel each other in the average (Proposition 1).

Sample-level damage has been measured before. Weight pruning concentrates its cost on identifiable examples (Hooker et al., 2019). Model updates are evaluated with negative flips and prediction churn (Yan et al., 2021; Milani Fard et al., 2016), and flip rates have been proposed for compressed LLMs (Dutta et al., 2024). Inference-time compression of VLMs differs from these settings in one respect. The compression decision is made per request and can be undone at serving time. Every failure thus has a counterfactual that we can run, namely the same request without compression. This lets us

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.

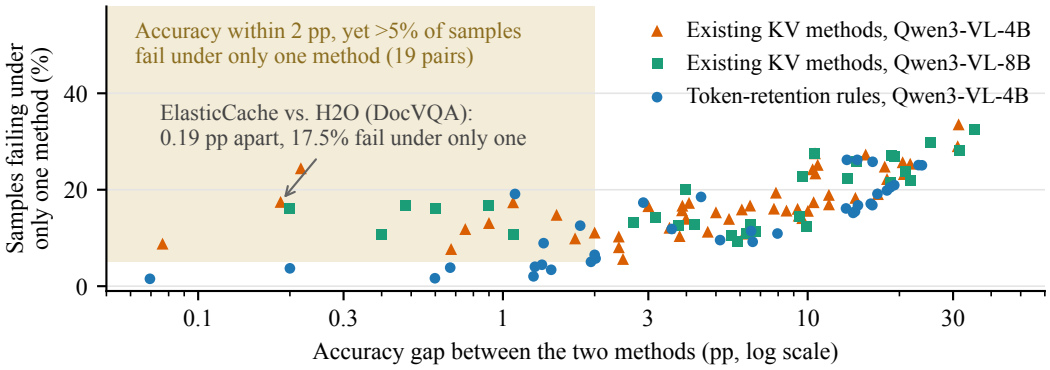


Figure 1: **Nearly equal accuracy, different broken requests.** Each point compares two compression methods on one dataset (116 pairs). Existing KV methods are adapted published KV-cache eviction strategies, and token-retention rules are the four controlled policies of §4.1. The y-axis is the fraction of the dataset that is a CIF under exactly one of the two methods (the exclusive-failure mass). Shaded: 19 pairs within 2 pp of accuracy that diverge on more than 5% of the dataset. Margin floors and matched-margin expectations are in Appendix B.2.

go beyond counting flips. We ask when a benchmark can observe the damage at all, whether the compressed model can indicate which of its own answers are broken, and whether a broken answer can be repaired, and by which evidence.

To answer these questions, we build a paired measurement protocol. Every sample runs under a reference configuration and a stress configuration that keeps 10% of the prompt's KV entries. A CIF is a sample that the reference answers correctly and the stressed run does not. The paired design removes generic sample difficulty and isolates the damage caused by compression itself. Figure 2 summarizes the protocol. We establish the results on Qwen3-VL-4B (Bai et al., 2025), replicate them on Qwen3-VL-8B, and test transfer on LLaVA-NeXT-Mistral-7B and InternVL3.5 (Liu et al., 2024a; Wang et al., 2025b). We first measure which requests break and what determines how many (§5.2). We then examine when a benchmark can register the damage (§5.3). Finally, we use the counterfactual to test prediction with fallback (§5.4) and token restoration (§5.5). In summary, our main contributions include:

- **Measuring.** We introduce a paired protocol for inference-time VLM compression, with CIF, reverse flips, and margin-calibrated failure-set turnover. On the affected datasets, 19 of 27 condition pairs within 2 points of accuracy diverge on more than 5% of the samples. At a matched token count, the identity of the kept tokens changes the CIF rate by up to an order of magnitude. A benchmark registers this damage only when its scored answer depends on the compressed cache (§5.2–5.3).
- **Predicting.** We show that the compressed pass signals its own failures. A logistic model over output-confidence features ranks CIF at AUROC up to 0.89 (six seeds), on par with hybrid triggers that add the compression trace. It transfers across datasets, compression backends, and model families. A fallback driven by it recovers 36–80% of the Qwen3-VL failures at 20% rerun coverage in paired simulation (§5.4).
- **Diagnosing.** We restore pruned tokens at equal budgets and compare selection policies. Restoring pruned tokens at a budget of 5% of the original visual-token count repairs 38–94% of the failures on Qwen3-VL and LLaVA-NeXT. Randomly restored tokens measure the effect of added capacity. At this budget, routing-score selection repairs 3–26 more points of the failure set than the better unguided selection, so the damage traces to specific lost tokens (§5.5).

2 Related Work

Efficient LVLM inference and token compression. KV-cache management is central to LLM serving (Kwon et al., 2023). Eviction and selection methods retain heavy-hitter or attention-selected entries (Zhang et al., 2023; Liu et al., 2023; Li et al., 2024), allocate budgets across layers (Cai et al., 2024), or quantize the cache (Liu et al., 2024c). In multimodal models, visual tokens make up most

of the sequence, so visual-token pruning and text-guided visual KV eviction have become common choices (Chen et al., 2024a; Zhang et al., 2025; Endo et al., 2025; Yang et al., 2025; Liu et al., 2026). These methods are evaluated by average accuracy at matched budgets. Recent analyses argue that this practice obscures what pruning removes (Wen et al., 2025). We measure a sample-level failure mode shared by these compressors and evaluate six published strategies under one protocol (§4.1).

Flips, churn, and compression disparity. Sample-level metrics for model changes are well established. Hooker et al. (2019) showed that weight pruning keeps top-line accuracy while concentrating damage on identifiable examples. For model updates, Milani Fard et al. (2016) defined prediction churn and Yan et al. (2021) defined negative flips. Dutta et al. (2024) advocate flip rates for compressed LLMs. In these terms, the CIF rate is the negative-flip rate of an inference-time compression intervention. On top of this rate, we measure failure-set turnover across compressors at matched budget and test when a benchmark registers the damage. Because per-request compression is reversible, we also predict flips and undo them by fallback. A restoration intervention tests which restored tokens repair them.

Hallucination in LVLMs. POPE (Li et al., 2023), HallusionBench (Guan et al., 2024), and AMBER (Wang et al., 2023) quantify hallucination at the object, illusion, and multi-dimensional levels. Mitigation methods intervene at decoding time (Huang et al., 2024; Leng et al., 2024) or on attention heads (Sarkar et al., 2025). These works treat hallucination as a property of the model and the input. We study a different cause, the removal of visual evidence at inference time. This cause is controllable, and object-level benchmarks such as POPE show little of its damage (§5.3).

Uncertainty, selective prediction, and fallback. Selective classification trades coverage for accuracy via a confidence score (Geifman & El-Yaniv, 2017; Hendrycks & Gimpel, 2017). Sampling-based estimators detect unreliable generations (Kuhn et al., 2023; Farquhar et al., 2024), and black-box consistency has been applied to VLM selective prediction (Khan & Fu, 2024). Our fallback is selective prediction whose recovery action is a rerun at higher fidelity. The trigger is output confidence (Hendrycks & Gimpel, 2017). It needs only token log-probabilities, and it ranks failures better than sampling-based scores at a fifth of their cost (§5.4). Unlike these lines of work, we pair every compressed run with an uncompressed counterfactual. This turns compression damage into per-request labels, which we measure, predict, and diagnose.

3 The CIF Measurement Protocol

Paired compressed inference. Let f denote an LVLM and $x = (v, q)$ an image–question pair. An inference configuration $c = (\pi, \rho)$ specifies a token-retention policy π and a prune ratio $\rho \in [0, 1)$ that keeps a fraction $1-\rho$ of the prompt's KV entries, and $f_c(x)$ denotes the answer generated under c. We fix a *reference* configuration c_{ref} (no compression, or low compression $\rho \leq 0.05$ where noted) and a *stress* configuration $c_{\text{str}} = (\pi, 0.9)$. With a task-specific correctness function $\text{corr}(\cdot)$, each sample falls into one of four paired outcomes. We define the target event

$$\text{CIF}(x) \;=\; \mathbb{1}[\,\text{corr}(f_{c_{\text{ref}}}(x)) \;\wedge\; \neg\,\text{corr}(f_{c_{\text{str}}}(x))\,], \tag{1}$$

and symmetrically a *reverse flip* $\text{rev}(x) = \mathbb{1}[\neg\,\text{corr}(f_{c_{\text{ref}}}(x)) \wedge \text{corr}(f_{c_{\text{str}}}(x))]$. The CIF rate over a dataset counts damage attributable to compression. Samples the model fails regardless are excluded. The reverse-flip rate bounds the harm of falling back to the reference. For two stress configurations c_1, c_2 with CIF sets A_1, A_2, the *failure-set turnover* $\tau(c_1, c_2) = 1 - |A_1 \cap A_2|/|A_1 \cup A_2|$ measures how much the failing set changes between policies. Average accuracy leaves this set undetermined.

Let ρ_{CIF} and ρ_{rev} denote the CIF and reverse-flip rates and $\text{acc}(c)$ the average accuracy. The two rates relate to the accuracy drop as follows.

Proposition 1 (What the average hides)**.** *For any reference–stress pair,* $\text{acc}(c_{\text{ref}}) - \text{acc}(c_{\text{str}}) = \rho_{\text{CIF}} - \rho_{\text{rev}}$ *exactly; hence the observed accuracy drop* Δ *only* lower-bounds *the induced damage,* $\rho_{\text{CIF}} \geq \max(0, \Delta)$*, while* ρ_{CIF} *can be as large as* $\min\big(\text{acc}(c_{\text{ref}}),\, 1 - \text{acc}(c_{\text{str}})\big)$.

Corollary 1 (Non-identifiability of failure identity)**.** *On a dataset of* n *samples, for any integer* k *with* $1 \leq k$ *and* $2k \leq n\,\text{acc}(c_{\text{ref}})$ *there exist two stress outcome patterns with* identical *average accuracy whose* CIF *sets are disjoint* ($\tau = 1$)*, together covering a* $2k/n$ *fraction of the dataset.*

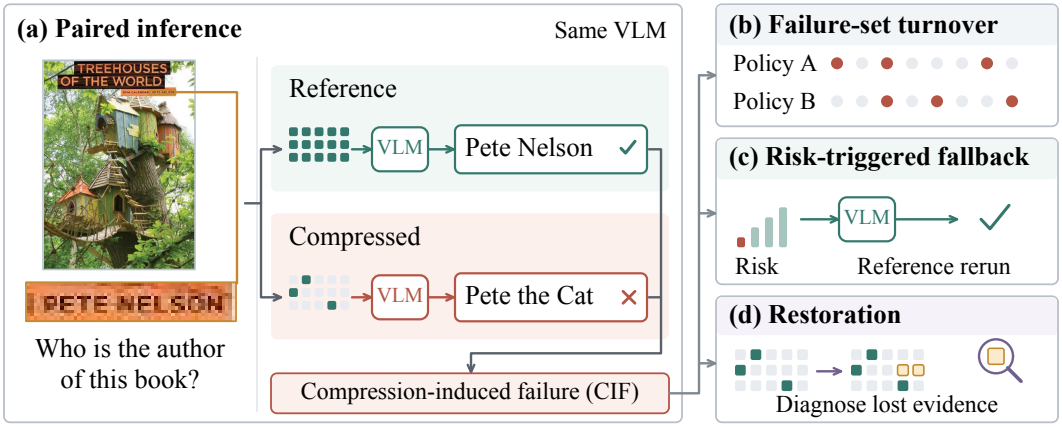


Figure 2: **The CIF measurement pipeline.** (a) In paired inference, the reference answers "Pete Nelson" correctly, while compression changes the answer to "Pete the Cat." (b) Paired outcomes reveal which samples fail under different compression policies. (c) Output confidence guides reference reruns to recover risky answers. (d) A separate restoration intervention tests which discarded visual evidence repairs failures. Token patterns and downstream panels are schematic, and the author line is enlarged.

Average accuracy thus bounds the size of the damage only through Proposition 1. For damage up to half the reference-correct mass, it places no constraint on which samples fail.

Proofs, the τ calibration bounds (margin floor, exclusive-failure mass, matched-margin expectation), and the fallback accounting identity are in Appendix A.1.

Risk prediction. During stress inference we log a feature vector $\Phi(x) \in \mathbb{R}^d$ (§4.2) and learn a risk score $s(x) = g(\Phi(x))$ that ranks samples by $\Pr[\text{CIF}(x) = 1]$. The features come only from the compressed pass, so the score is available at deployment without a reference run. Only the training labels need paired runs. The estimator g is a logistic regression, or a random forest for the high-dimensional families. Scores are trained and evaluated *out-of-fold* (5-fold, multiple seeds) within each dataset. We report AUROC and AUPRC for every feature family.

Risk-triggered fallback. Given a rerun budget (coverage) $b \in (0, 1)$, *fallback* replaces the stress outputs of the top-b risk fraction with the paired reference outputs. This simulates a deployment that reruns suspicious samples at high fidelity. The action is fixed. Only the *trigger*, the score that ranks samples for rerun, varies. The *simulated error* at b is the resulting error rate, including the harm from reverse flips. The *oracle* ranks by the true CIF label and tracks the analytic lower bound of Appendix A.1. Cost is reported as the rerun coverage b (§5.4).

Restoration intervention. To test whether the damage is tied to specific lost tokens, we define a post-pruning restoration intervention. After the stress run prunes its visual tokens, we restore $\lceil \gamma N_v \rceil$ tokens per active layer from the pruned set. Here N_v is the original visual-token count and γ the restoration budget. A selection policy chooses which tokens are restored. If a policy selects the tokens that the failed answers depend on, its restoration should repair more CIF cases than matched alternatives at equal γ (§5.5).

4 Experimental Design and Instrumentation

4.1 Instrumented compression and controlled retention policies

We instrument a token-level KV compression stack for Qwen3-VL. Pruning decisions and their inputs are logged per layer during prefill and generation. The stack scores visual tokens with an attention-derived routing score and keeps a subset within the budget set by the prune ratio ρ. Pruned tokens are evicted from the KV cache and are never attended to again. As in published KV-cache eviction methods, eviction acts on the cache after the prefill forward. The first generated token

is thus computed on the full cache, and pruning affects every later token. Within this fixed stack we implement four retention policies that differ only in the retention rule. Routing-score retention (RouteConf) keeps the tokens ranked highest by a text-weighted routing score (Appendix A.2). High-attention retention keeps the tokens with the highest unweighted text-to-visual attention. Random retention keeps a random subset. Spatially uniform retention takes a strided subset that covers the image grid. The four variants share the model, prompts, decoding, and harness, so differences in damage are attributable to the retention rule alone. The budget covers all prompt entries, and prompt-text entries are always kept. The retained visual ratio is therefore below 10% and depends on the prompt length (≈2.6% on OCRBench). All four policies keep the same number of visual tokens per sample. We report the measured retained visual ratios next to the damage.

Published strategies. We also evaluate five published KV-cache eviction strategies under the same protocol at matched 10% retention. They are StreamingLLM (Xiao et al., 2024), SnapKV (Li et al., 2024), H2O (Zhang et al., 2023), ElasticCache (Liu et al., 2024d), and PrefixKV (Wang et al., 2025a). We adapt each to Qwen3-VL from its released source and check retained ratio, cache reduction, output sanity, and backend identity. An unmodified run of the official SnapKV code lands within 0.7 pp of our version. Appendix A.3 lists deviations from published accuracies. A sixth subject, a FastV-style visual-token pruning criterion (Chen et al., 2024a), is evaluated in the same framework (Appendix A.3). All six pair against the same references and harness. They emit no routing traces, so their evaluation uses output-side signals only (§5.4).

4.2 Stress-run risk features

Each stress run logs a per-sample trace, which we summarize into four feature families. Retention features (5–6) cover the selected visual-token count and ratio and per-layer keep statistics. Attention features (51) summarize text→visual attention across layers and heads. Routing-extended features (90–99) describe the pre-selection routing-score distribution, including z-scores, selection margins, entropy, top-k concentration, and budget pressure. Generation-confidence features (33) are the mean and minimum token log-probability, token-entropy statistics, and length-normalized variants. A combined model uses the routing and generation families together. Since the estimators are simple, the comparisons measure the information in each family.

4.3 Fallback triggers

Beyond single-family scores, we evaluate three hybrid triggers at fixed coverage. The combined model is trained jointly on both families. Residual triggers spend a fraction of the budget by confidence and the remainder by routing rank. Two-gate rank fusion takes the mean or max of the confidence- and routing-rank percentiles. Random selection and the oracle bound every table, and all scores remain out-of-fold.

5 Experiments

5.1 Setup

Models. The main grid is Qwen3-VL-4B. It covers 10 datasets × 4 controlled policies, 6 published strategies on 7 of the datasets, the ρ sweep, risk, fallback, and restoration. Qwen3-VL-8B replicates the published-strategy evaluation, six-seed risk and fallback, restoration, and cost at a larger scale. Family transfer uses LLaVA-NeXT-Mistral-7B and InternVL3.5 (Liu et al., 2024a; Wang et al., 2025b) on 5 stress conditions × 4 datasets. LLaVA-NeXT covers damage, turnover, six-seed risk and fallback, and verified restoration. InternVL3.5 covers damage and turnover.

Datasets. OCRBench (Liu et al., 2024b), ChartQA (Masry et al., 2022), VizWiz (Gurari et al., 2018), AMBER (Wang et al., 2023), DocVQA (Mathew et al., 2021), TextVQA (Singh et al., 2019), GQA (Hudson & Manning, 2019), MMStar (Chen et al., 2024b), POPE (Li et al., 2023), and HallusionBench (Guan et al., 2024), all at full evaluation size (951–12,578 samples; Table 1a).

Configurations. Stress runs use $\rho = 0.9$ (retained KV ≈10%), and the dose–response sweep adds $\rho \in \{0.5, 0.7, 0.8\}$; published strategies run at matched 10%. References are strict no-prune except

Table 1: **Full-dataset compression damage under the paired-inference protocol.** (a) Controlled retention policies (Qwen3-VL-4B, $\rho = 0.9$): CIF rate, reference error, and mean retained visual-token ratio. The budget is 10% of all prompt entries, and prompt text is always kept, so the visual ratio is the same for all four policies and depends on the prompt length. Random retention causes the most damage on five of the six affected tasks. GQA, MMStar, POPE, and HallusionBench are the unaffected boundary. Their prompts leave most of the budget to text entries. (b) Four models from three families on their shared datasets. Routing is RouteConf on Qwen, the routing-score policy on LLaVA-NeXT, and a port of the same score on InternVL3.5; high-attention on InternVL3.5 is its attention-guided policy. SnapKV uses the window adaptation of Appendix A.3; Table 3 gives the full published-strategy evaluation.

(a) Controlled retention policies (Qwen3-VL-4B, $\rho = 0.9$; HallusionB. = HallusionBench).

Dataset	n	Ref. err.	CIF rate under retention policy				Retained ratio
			RouteConf	High-attn	Random	Spatial	
OCRBench	1000	.222	.362	**.357**	.513	.492	.026
ChartQA	2500	.162	**.072**	.085	.150	.138	.022
VizWiz	4319	.230	**.058**	.060	.085	.076	.037
AMBER	1004	.074	.144	**.119**	.148	.277	.035
DocVQA	5349	.056	**.023**	.029	.265	.219	.051
TextVQA	5000	.147	**.040**	.052	.224	.190	.052
GQA	12578	.379	.004	.004	.004	.004	.008
MMStar	1500	.403	.000	.000	.000	.000	.002
POPE	3000	.086	.000	.000	.000	.000	.009
HallusionB.	951	.304	**.004**	.006	.004	.007	.029

(b) Model families on the shared datasets (10% retention, strict references), CIF rate.

		Rout.	H-attn	Rand.	Spat.	SnapKV
ChartQA	Qwen-4B	**.072**	.085	.150	.138	.089
	Qwen-8B	**.060**	.072	.135	.123	.054
	LLaVA-7B	**.039**	.041	.189	.172	.095
	InternVL-8B	.008	**.006**	.081	.083	.021
DocVQA	Qwen-4B	**.023**	.029	.265	.219	.090
	Qwen-8B	**.032**	.042	.256	.218	.069
	LLaVA-7B	**.023**	.026	.234	.191	.076
	InternVL-8B	**.004**	**.004**	.211	.198	.021

		Rout.	H-attn	Rand.	Spat.	SnapKV
TextVQA	Qwen-4B	**.040**	.052	.224	.190	.123
	Qwen-8B	**.063**	.075	.225	.206	.091
	LLaVA-7B	**.010**	.012	.156	.121	.067
	InternVL-8B	.011	**.010**	.196	.181	.033
OCRBench	Qwen-4B	.362	**.357**	.513	.492	.345
	Qwen-8B	.362	**.357**	.479	.478	.317
	LLaVA-7B	.055	**.053**	.216	.177	.158
	InternVL-8B	.058	**.044**	.334	.299	.145

$\rho = 0.05$ for VizWiz, AMBER, POPE, and HallusionBench in the 4B matrix (all 8B and LLaVA-NeXT runs are strict). Decoding is greedy with matched lengths. Risk models are 5-fold out-of-fold, six seeds at both scales (Appendix A.2).

5.2 Nearly equal accuracy, different broken requests

Which requests break. Figure 1 covers all 116 condition pairs of the calibrated turnover analysis. These are pairs of published strategies at both scales and pairs of controlled policies on the 4B grid. Among the 27 pairs whose average accuracy differs by less than 2 points (gaps rounded to 0.01), 19 have an exclusive-failure mass above 5% of the dataset. On 4B DocVQA, ElasticCache and H2O are 0.19 points apart and break 1363 and 1390 samples. Only 909 of these are shared, so 17.5% of the dataset fails under exactly one of them. The largest case, PrefixKV vs. StreamingLLM on the same dataset, reaches 24.5% at a 0.22-point gap. The 8B replication shows the same pattern, e.g., 16.7% for ElasticCache vs. H2O on DocVQA and 16.2% for H2O vs. StreamingLLM on OCRBench at a 0.2-point gap.

Raw turnover τ has to be read against the margin between the two failure rates, since a large margin alone pushes τ upward (Proposition 2). For the 19 diverging pairs the margin floor is at most 0.30, and their τ of 0.19–0.79 lies above what the margin alone would force. Observed τ also stays below the matched-margin expectation for every pair (0.51 vs. 0.85 for ElasticCache–H2O on DocVQA). The failure sets thus share a set of hard samples, and the strategies differ on the samples around it. Pairs that share a scoring principle overlap the most. Routing and high-attention retention, whose scores are both derived from attention, stay at or below 3.7% on five of the six affected datasets.

AMBER is the exception, with 12.5% at a 1.8-point gap. The other three near-equal pairs below 5% are the spatial pairs on VizWiz (3.8–4.4%). Strategies built on different rules differ on many more samples at the same budget and accuracy.

How many requests break. Table 1a reports the controlled 4B matrix at $\rho = 0.9$ (bar-chart view in Figure 4). The damage concentrates on OCRBench, DocVQA, TextVQA, and ChartQA, whose answers rely on small visual details. MMStar, POPE, and HallusionBench show little or none (§5.3). With a fixed budget, the identity of the kept tokens decides how much breaks. Random retention keeps as many visual tokens as the guided policies. It breaks $11\times$ more samples on DocVQA (.265 vs. .023), $5.6\times$ more on TextVQA, and $2.1\times$ more on ChartQA. The published strategies at the same measured 10% budget span .090 (SnapKV) to .389 (PrefixKV) on DocVQA (Table 3). A FastV-style single-layer criterion (Chen et al., 2024a), transplanted into the same eviction framework, reaches .343 on DocVQA and .546 on OCRBench. Spatially uniform coverage stays close to the guided policies on scene-centric VizWiz, but it falls behind where evidence is spatially concentrated (DocVQA, OCRBench). The ratio matters as well. When ρ sweeps over $\{0.5, 0.7, 0.8, 0.9\}$ on ChartQA, DocVQA, and OCRBench (Figure 5), CIF rises monotonically under every policy, and the guided policies stay below random retention at every ratio. A nominal ratio alone therefore does not determine reliability, and the retention policy should be reported with it. Finally, Table 1b reproduces the concentration on Qwen3-VL-8B, LLaVA-NeXT-Mistral-7B, and InternVL3.5, with calibrated turnover above its margin floor (Appendix B).

5.3 When benchmarks register compression-induced failures

A benchmark can detect CIF only if its scored answer depends on computation after compression takes effect. Eviction follows prefill (§4.1), so the first generated token is always computed on the full cache. Every POPE answer is a single token, 95% of MMStar answers are one letter, and 83% of HallusionBench answers open with their yes/no. Their scores are fixed before any pruned entry is read, and the paired runs confirm this. POPE and MMStar yield exactly zero CIF under all four controlled policies and all six published strategies. and HallusionBench stays within 0.4–0.7% (Appendix C).

Answer placement is a necessary condition, but it does not fully explain the flat POPE results. We add a fixed answer prefix so that the yes/no becomes the fourth generated token, decoded after eviction. POPE accuracy then stays at the reference level under all four retention policies (.906–.911 vs. .909, with 8–14 CIF out of 3000). The kept prompt entries were computed on the full image during prefill, so object-level information may survive eviction through them. Text and layout details do not survive in the same way. Restricting Table 1a to reference answers of two or more tokens raises every text, chart, and document cell. OCRBench rises to .40–.57, while GQA stays at or below .02 under every policy (Table 6). Validating a compressor therefore requires tasks whose answers are decoded from the compressed cache and depend on small visual details.

5.4 The compressed pass signals its own failures

Every failure carries a paired label, so we can test whether the compressed pass itself indicates which of its answers are broken.

Prediction. On the main grid, every feature family ranks CIF above chance. Generation confidence is the strongest single family on all six affected datasets, with AUROC .72–.89 (Table 8; six seeds, std $\leq$ .014). The 8B replication (six seeds) gives the same ordering. Confidence is again the strongest single family on every dataset, with AUROC up to .88. On LLaVA-NeXT-Mistral-7B, a non-Qwen decoder, confidence ranks CIF at AUROC .73–.85 under routing retention and .68–.82 under SnapKV (Table 2). The signal is specific to CIF. Among stress-wrong samples, where every sample is an error, the focus families still rank CIF above always-wrong at AUROC .61–.83 in every cell. The signal also transfers across datasets (mean .75 over 600 leave-one-dataset-out cells) and across backends (mean .76 on the published strategies; Appendix D.2).

The confidence predictor is a logistic regression over the 33 log-probability and entropy features of the compressed output, trained out-of-fold.

Fallback. At a fixed rerun budget of 20% (Table 10), the best trigger per dataset cuts simulated error from .570 to .429 on OCRBench, from .163 to .114 on AMBER, and from .075 to .059 on

Table 2: **Output-signal control across model families (six seeds, out-of-fold).** CIF count, confidence AUROC, and fallback at 20% coverage with the confidence trigger: error against the stress and oracle errors, CIF cases recovered, and reverse flips undone. Qwen rows use RouteConf retention (8B DocVQA: exact-match view); LLaVA-7B rows use its routing policy and the SnapKV adaptation. All triggers and feature families: Tables 10 and 8.

Dataset	Model (condition)	CIF	AUROC	Error at 20% coverage Stress	Conf.	Oracle	Recovered / undone CIF	Reverse
ChartQA	Qwen3-VL-4B	180	.783	.222	.189	.153	96 (53%)	15 / 29
	Qwen3-VL-8B	151	.798	.201	.169	.143	88 (58%)	7 / 21
	LLaVA-7B (routing)	97	.734	.488	.478	.452	42 (44%)	18 / 52
	LLaVA-7B (SnapKV)	238	.679	.535	.513	.442	74 (31%)	18 / 74
DocVQA	Qwen3-VL-4B	122	.882	.075	.059	.053	96 (79%)	12 / 22
	Qwen3-VL-8B	596	.881	.226	.147	.116	442 (74%)	23 / 54
	LLaVA-7B (routing)	123	.748	.315	.309	.295	61 (50%)	32 / 61
	LLaVA-7B (SnapKV)	405	.717	.355	.335	.284	163 (40%)	53 / 126
TextVQA	Qwen3-VL-4B	202	.875	.175	.151	.136	155 (77%)	33 / 62
	LLaVA-7B (routing)	52	.850	.305	.302	.295	38 (72%)	22 / 33
	LLaVA-7B (SnapKV)	335	.823	.360	.323	.294	202 (60%)	19 / 44

DocVQA. Confidence alone is within .01 of the best trigger everywhere. The confidence trigger also works on the published strategies. Per-backend out-of-fold triggers recover 27–68% of each strategy's induced excess error at equal coverage, and RouteConf-trained scores transfer unchanged at .76 AUROC (Appendices D.2, D.3). The 8B replication matches across the full 5–50% coverage grid (.578 → .422 on OCRBench, six seeds; Figure 6, Appendix D.3). Reverse-flip harm is near zero where damage is large. On LLaVA-NeXT under SnapKV, the trigger recovers 31–60% of the CIF pool at 20% coverage and lowers error by .02–.04 (Table 2). Under the lightly damaging routing policy (CIF 1–4%) it recovers 44–72%. Reverse flips bound the net gain there. On DocVQA, 61 of 123 failures are repaired and 32 of 61 reverse flips are undone. A gap to the oracle remains (.429 vs. .370 on OCRBench).

Output signals suffice for control. Semantic entropy and self-consistency (K=5), two training-free sampling scores, trail confidence by 0.12 mean AUROC and recover 13.5 points less of the CIF pool at 20% coverage, at five times the cost (Table 9; Appendix D.2). Adding the compression trace shifts AUPRC by -0.043 to $+0.043$ across 11 dataset–model pairs (Table 8; mean $+0.005$), and hybrid triggers improve error by at most .01 over confidence alone. Output signals are thus enough for control, and every compressor produces them, including the trace-free strategies of Appendix A.3. A confidence predictor fit on Qwen also transfers unchanged to LLaVA-NeXT at AUROC .67–.79, against .73–.85 when refit (Appendix D.1). The routing score plays a different role in restoration, where it selects which pruned tokens to give back (§5.5).

Cost. Fallback is a per-request cascade (Chen et al., 2024c) to the reference configuration. We report its cost as the rerun coverage b. At $b = 20\%$, one request in five is rerun, and 36–80% of the Qwen3-VL failures are recovered. Peak memory is unchanged. Wall-clock cost depends on the serving stack. Our instrumented stress pass runs slower than the reference, so measured wall times are in Appendix D.4.

5.5 Failures are repairable, and the restored tokens matter

If the damage comes from specific lost evidence, giving that evidence back should repair it, and at an equal budget it should matter which tokens are given back. We test this on the frozen CIF sets of the routing-score stress runs in eight settings (Figure 3). The settings are Qwen3-VL-4B and 8B on four datasets and LLaVA-NeXT-Mistral-7B on ChartQA and DocVQA. After pruning, we restore $\lceil \gamma N_v \rceil$ tokens per active layer from the pruned set, where N_v is the original visual-token count and $\gamma \in \{5, 10, 20\}\%$. The restored tokens are selected by the routing score, by high attention, at random, or uniformly in space (spatial on Qwen only). Matched budgets and distinct restored tokens are verified per intervention (Appendix A.2).

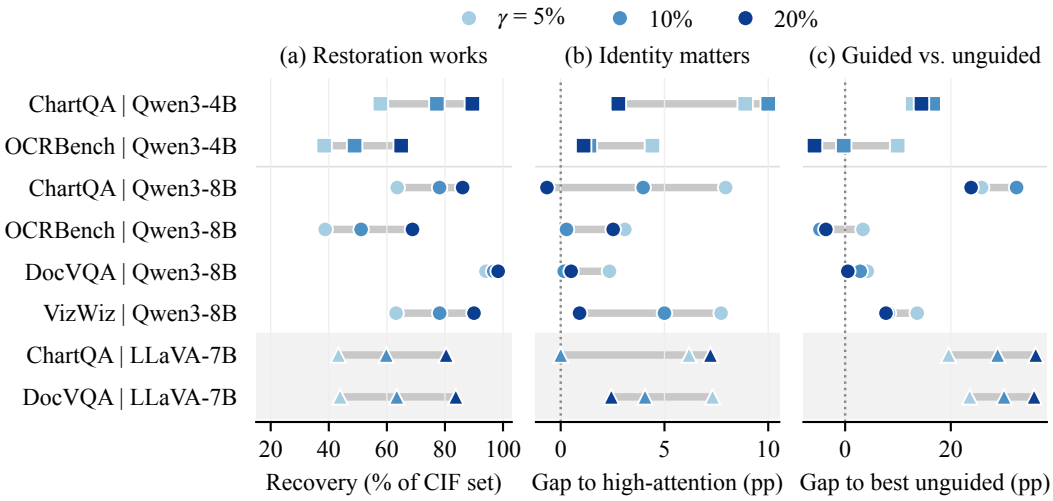


Figure 3: **Restoration repairs most failures, and the restored tokens matter at tight budgets.** (a) RouteConf recovery of the frozen CIF set at $\gamma = 5/10/20\%$ in eight settings: Qwen3-VL-4B/8B (squares/circles) and LLaVA-NeXT-Mistral-7B (triangles, shaded). (b) RouteConf minus high-attention (points of the CIF set): positive everywhere at 5%; by 20% near zero on Qwen, positive on LLaVA-NeXT. (c) RouteConf minus the better of random and spatial (random only on LLaVA-NeXT): by 20% it reverses on OCRBench, narrows on DocVQA and VizWiz, and stays at 14–36 pp on Qwen ChartQA and LLaVA-NeXT. Counts and tests: Appendix E.

Selection (Figure 3b,c). Random restoration adds the same number of tokens without any selection signal, so it measures the effect of added capacity. At $\gamma = 5\%$ it repairs 20–87% of the failures. Routing-score selection repairs 3–26 more points than the better of random and spatial selection, and 2–9 more points than high-attention selection. The high-attention contrast is significant in five of six Qwen settings (sign test $p \leq .013$, and $p = .052$ on 8B OCRBench). On LLaVA-NeXT, routing leads high-attention by 6–7 points at 5% and random by 20–36 points across budgets. The selection effect shrinks as the budget grows. By $\gamma = 20\%$ the guided selections converge on Qwen, and on OCRBench spatial selection overtakes routing at both scales.

Recovery (Figure 3a). Routing-score restoration at $\gamma = 5\%$ repairs 38–94% of the failures, rising to 65–98% at 20% (LLaVA-NeXT: 43–44% and 80–84%). Most failures are thus reversible with a small amount of restored evidence, and the equal-budget gaps show that the damage is tied to specific lost tokens.

6 Conclusion

This paper studies visual-token compression at the level of individual requests. With paired inference, we showed that compression changes which requests a VLM answers correctly, and that average accuracy leaves this change unspecified. Methods with similar accuracy break different requests, and with a fixed number of kept tokens, which tokens are kept decides how many requests break. The damage becomes visible on benchmarks whose answers are decoded from the compressed cache and rely on small visual details such as text and layout. The paired counterfactual also makes the failures predictable and repairable. The output confidence of the compressed pass ranks broken requests at AUROC up to .89 and recovers 36–80% of them at 20% rerun coverage. Restoring pruned tokens at 5–20% of the visual-token count repairs 38–98% of the failures, and routing-score selection repairs the most in all eight settings at the 5% budget.

Two reporting practices follow. Compressed VLMs should be reported with paired failure counts and the retention policy next to average accuracy, and they should be validated on tasks such as OCR and document understanding. We will release the protocol, instrumentation, and per-sample logs to support such reporting. The fallback gains are simulated on paired outputs, so their serving benefit still needs confirmation on an optimized stack. All eviction in this study follows prefill, and methods that prune during prefill remain to be evaluated (Appendix A.4).

AI USE STATEMENT

In this work, we used generative AI tools (large language models) to assist in formulating and proving the mathematical claims of Appendix A.1 (all statements and proofs were checked by the authors), to identify and summarize related literature, and to aid and polish the writing of the manuscript. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All benchmark evaluations run on full public evaluation sets with deterministic decoding and paired references. Restoration and cost profiling use the CIF subsets stated in Appendices E and D.4. Risk models are 5-fold out-of-fold over six seeds at both scales. Appendices A.2–D.4 document the configuration, verification, and measured costs. Code, instrumentation, per-sample paired logs, per-cell configurations, and analysis scripts will be released.

REFERENCES

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

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision – ECCV 2024*, pp. 19–35. Springer Nature Switzerland, 2024a. doi: 10.1007/978-3-031-73004-7_2.

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 37*, pp. 27056–27087. Neural Information Processing Systems Foundation, 2024b. doi: 10.52202/079017-0850.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024c.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems 37*, pp. 124347–124390. Neural Information Processing Systems Foundation, 2024. doi: 10.52202/079017-3950.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22826–22835. IEEE, 2025. doi: 10.1109/iccv51701.2025.02119.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*, pp. 4878–4887, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024. doi: 10.1109/cvpr52733.2024.01363.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617. IEEE, 2018. doi: 10.1109/cvpr.2018.00380.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024. doi: 10.1109/cvpr52733.2024.01274.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019. doi: 10.1109/cvpr.2019.00686.

Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10854–10863. IEEE, 2024. doi: 10.1109/cvpr52733.2024.01032.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024. doi: 10.1109/cvpr52733.2024.01316.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems 37*, pp. 22947–22970, 2024. doi: 10.52202/079017-0722.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024a. Official project blog.

Jizhihui Liu, Ruizi Han, Miao Zhang, Rui Shao, Xuebo Liu, Weili Guan, and Yaowei Wang. TGV-KV: Text-grounded KV eviction for vision-language models. In *International Conference on Machine Learning*, 2026. arXiv:2606.03075.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b. doi: 10.1007/s11432-024-4235-6.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems 36*, pp. 52342–52364, 2023. doi: 10.52202/075280-2279.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32332–32344. PMLR, 2024c.

Zuyan Liu, Benlin Liu, Jiahui Wang, Yuhao Dong, Guangyi Chen, Yongming Rao, Ranjay Krishna, and Jiwen Lu. Efficient inference of vision instruction-following models with elastic cache. In *Computer Vision – ECCV 2024*, pp. 54–69. Springer Nature Switzerland, 2024d. doi: 10.1007/978-3-031-72643-9_4.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208. IEEE, 2021. doi: 10.1109/wacv48630.2021.00225.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and iterate: Reducing prediction churn. In *Advances in Neural Information Processing Systems 29*, pp. 3179–3187, 2016.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter A. Beerel, and Souvik Kundu. Mitigating hallucinations in vision-language models through image-guided head suppression. *arXiv preprint arXiv:2505.16411*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019. doi: 10.1109/cvpr.2019.00851.

Ao Wang, Hui Chen, Jianchao Tan, Kefeng Zhang, Xunliang Cai, Zijia Lin, Jungong Han, and Guiguang Ding. PrefixKV: Adaptive prefix KV cache is what vision instruction-following models need for efficient generation. In *Advances in Neural Information Processing Systems 38*, pp. 104755–104781. Neural Information Processing Systems Foundation, 2025a. doi: 10.52202/085713-3159.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15537–15549. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.802.

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

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14294–14303. IEEE, 2021. doi: 10.1109/cvpr46437.2021.01407.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025. doi: 10.1109/cvpr52734.2025.01843.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74840–74857. PMLR, 2025.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems 36*, pp. 34661–34710, 2023. doi: 10.52202/075280-1506.

A Protocol, Implementation, and Scope

The appendix follows the order of Section 5. Appendix A gives the formal results, the implementation and its verification, the published-strategy adaptations, and the scope of the study. Appendix B supports §5.2 with the full damage and turnover analysis. Appendix C supports §5.3 with the answer-placement evidence. Appendix D supports §5.4 with all risk-ranking, fallback, and cost results. Appendix E supports §5.5 with the restoration counts and tests. The most important supporting items are the published-strategy adaptations (Appendix A.3), the margin-calibrated turnover (Appendix B.2), the CIF-specificity analysis (Appendix D.2), and the restoration verification (Appendix E).

A.1 Theoretical analysis

Throughout, fix a dataset of n samples and deterministic decoding, so each configuration c induces a fixed correct set $S_+(c) \subseteq \{1, \dots, n\}$ with $\mathrm{acc}(c) = |S_+(c)|/n$. For a reference–stress pair, the CIF set is $A = S_+(c_{\mathrm{ref}}) \setminus S_+(c_{\mathrm{str}})$ and the reverse-flip set is $R = S_+(c_{\mathrm{str}}) \setminus S_+(c_{\mathrm{ref}})$, with rates $\rho_{\mathrm{CIF}} = |A|/n$ and $\rho_{\mathrm{rev}} = |R|/n$, matching Eq. 1.

Proof of Proposition 1. Writing $I = S_+(c_{\mathrm{ref}}) \cap S_+(c_{\mathrm{str}})$, we have $|S_+(c_{\mathrm{ref}})| = |I| + |A|$ and $|S_+(c_{\mathrm{str}})| = |I| + |R|$; subtracting and dividing by n gives $\mathrm{acc}(c_{\mathrm{ref}}) - \mathrm{acc}(c_{\mathrm{str}}) = \rho_{\mathrm{CIF}} - \rho_{\mathrm{rev}}$. Since $\rho_{\mathrm{rev}} \geq 0$, $\rho_{\mathrm{CIF}} = \Delta + \rho_{\mathrm{rev}} \geq \Delta$, and $\rho_{\mathrm{CIF}} \geq 0$ always, giving the lower bound $\max(0, \Delta)$. For the upper bound, $A \subseteq S_+(c_{\mathrm{ref}})$ and $A \cap S_+(c_{\mathrm{str}}) = \emptyset$ give $\rho_{\mathrm{CIF}} \leq \min(\mathrm{acc}(c_{\mathrm{ref}}),\, 1 - \mathrm{acc}(c_{\mathrm{str}}))$. The bound is tight: with $a = |S_+(c_{\mathrm{ref}})|$ and $s = |S_+(c_{\mathrm{str}})|$ fixed, the overlap $|I|$ can be as small as $\max(0, a + s - n)$, so $|A| = a - |I|$ can reach $a - \max(0, a + s - n) = \min(a,\, n - s)$. □

Proof of Corollary 1. Let $k \geq 1$ be an integer with $2k \leq n\,\mathrm{acc}(c_{\mathrm{ref}}) = |S_+(c_{\mathrm{ref}})|$ and pick disjoint $A_1, A_2 \subseteq S_+(c_{\mathrm{ref}})$ with $|A_1| = |A_2| = k$ (possible since $2k \leq |S_+(c_{\mathrm{ref}})|$). Define stress outcome patterns $S_+(c_i) = S_+(c_{\mathrm{ref}}) \setminus A_i$ for $i = 1, 2$. Both have average accuracy $\mathrm{acc}(c_{\mathrm{ref}}) - k/n$, their CIF sets are exactly A_1 and A_2 with $A_1 \cap A_2 = \emptyset$, so $\tau(c_1, c_2) = 1$ and the exclusive-failure mass is $2k/n$. □

Proposition 2 (Turnover calibration)**.** *Let $A, B \subseteq \{1, \dots, n\}$ be the* CIF *sets of two stress configurations, both nonempty, and $\tau = 1 - |A \cap B|/|A \cup B|$. Then:* (i) $\tau \geq 1 - \min(|A|, |B|)/\max(|A|, |B|)$ *(the* margin floor*), with equality iff one set contains the other;* (ii) *the exclusive-failure mass satisfies $\big(|A| + |B| - 2|A \cap B|\big)/n \geq \big||A| - |B|\big|/n$;* (iii) *if B is a uniformly random subset of size $|B|$ placed independently of A, then $\mathbb{E}|A \cap B| = |A||B|/n$; the* matched-margin expectation *reported in Appendix B.2 is the plug-in value $\tau_{\mathrm{ind}} = 1 - \frac{|A||B|/n}{|A|+|B|-|A||B|/n}$ (a ratio of expectations, used as a reference point rather than as $\mathbb{E}\tau$).*

Proof. (i) $|A \cap B| \leq \min(|A|, |B|)$ and $|A \cup B| \geq \max(|A|, |B|)$, so $|A \cap B|/|A \cup B| \leq \min / \max$. Equality requires $|A \cap B| = \min(|A|, |B|)$ and $|A \cup B| = \max(|A|, |B|)$ simultaneously, i.e.,

the smaller set is contained in the larger; conversely, nesting attains both. (ii) From $|A \cap B| \leq \min(|A|, |B|)$, $|A|+|B|-2|A \cap B| \geq |A|+|B|-2\min(|A|, |B|) = \big||A|-|B|\big|$. (iii) $\mathbb{E}|A \cap B| = \sum_{x \in A} \Pr[x \in B] = |A| \cdot |B|/n$ since each sample lies in B with probability $|B|/n$; substituting this expectation into numerator and denominator of the Jaccard ratio (using $|A \cup B| = |A|+|B|-|A \cap B|$) gives τ_{ind}. □

Part (i) is why extreme-margin pairs are not evidence of disjoint failure sets (e.g., routing–random on DocVQA has $\tau = 0.93$ against a floor of 0.91). Part (ii) is why the exclusive-failure mass leads the reporting. Part (iii) supplies the independence reference, and every observed pair falls below it (§B.2).

Proposition 3 (Fallback accounting identity)**.** *Fix a stress configuration with error* err_{str} *and a trigger that, at coverage* b*, selects a set* F_b *(*$|F_b| = \lceil bn \rceil$*) whose stress outputs are replaced by the paired reference outputs. Then the simulated error is exactly* $\text{err}(b) = \text{err}_{\text{str}} - (G_b - H_b)/n$*, where* $G_b = |F_b \cap A|$ *(*CIF *samples recovered) and* $H_b = |F_b \cap R|$ *(reverse flips harmed). Every policy at coverage* b *obeys the analytic lower bound* $\text{err}(b) \geq \text{err}_{\text{str}} - \min(\lceil bn \rceil, |A|)/n$*, attained by an idealized selector that exhausts* A *first and fills any remainder outside* R *(feasible whenever* $\lceil bn \rceil \leq n - |R|$*). The tabulated oracle instead ranks by the true* CIF *label and fills the residual budget by risk rank, so its fill can select reverse flips (*$H_b > 0$*) and its error sits at or slightly above the bound.*

Proof. Replacing the stress output with the reference output changes correctness only on flip samples: a selected sample in A becomes correct (error count -1), a selected sample in R becomes incorrect ($+1$), and all other samples are unchanged. Hence the error count changes by $H_b - G_b$, giving the identity. For the lower bound, $G_b \leq \min(\lceil bn \rceil, |A|)$ and $H_b \geq 0$ for any policy, so $\text{err}(b) \geq \text{err}_{\text{str}} - \min(\lceil bn \rceil, |A|)/n$. For attainability, when $\lceil bn \rceil \leq n - |R|$ the idealized selector can exhaust A up to the budget and fill any remainder outside R, giving $H_b = 0$ and $G_b = \min(\lceil bn \rceil, |A|)$; the tabulated oracle's risk-rank fill can land in R, in which case it exceeds the bound by exactly H_b/n. □

A.2 Implementation and validation

Infrastructure. All compression experiments run through a single instrumented token-level KV compression stack integrated with the `lmms-eval` harness. The stack scores visual tokens per layer during prefill, retains a subset according to the active retention policy and prune ratio, evicts the remainder from the KV cache, and logs the full routing and retention trace per sample. Eviction is applied after the prefill forward returns, so the first output token is produced from the uncompressed cache. Generation-side features are computed from the stress run's token log-probabilities. Reference runs disable compression, or use $\rho = 0.05$ where noted.

Retention policies. All four policies share the per-layer budgets, always keep the prompt text, and differ only in how they rank visual tokens. Attention is averaged over heads in every layer. Inspired by Liu et al. (2026), the routing score (RouteConf) scores each visual token by text-to-visual attention weighted by text-to-text attention, and the budget is split across layers by text-to-visual attention mass. Each layer keeps its highest-scoring visual tokens, together with four leading sink tokens and the last token. High-attention retention ranks visual tokens by unweighted text-to-visual attention under the same budgets. Random retention ranks them by a fixed hash of token position, layer, and document. Spatially uniform retention keeps evenly spaced positions along the raster-ordered visual tokens. After prefill, every policy evicts the lowest-attention entry of each layer at each decoding step. Restoration adds $\lceil \gamma N_v \rceil$ tokens per layer from the pruned visual set, on top of the budget. Routing-score restoration takes the pruned tokens ranked highest by the routing score, high-attention restoration those with the highest unweighted attention, random restoration those ranked first by the hash, and spatial restoration evenly spaced pruned positions.

Correctness functions. $\text{corr}(\cdot)$ binarizes each benchmark's standard per-sample metric as computed by the evaluation harness (e.g., the ANLS-threshold criterion for DocVQA, the binary OCR-Bench accuracy score), applied identically to the paired reference and stress outputs of a cell. One exception applies. The Qwen3-VL-8B RouteConf transitions (Table 4) and the restoration analysis

use the stricter exact-match view of DocVQA, and the ANLS-threshold view of the same run yields 171 CIF cases (rate .032).

Verification. Restoration runs are checked for matched interventions. Pruned-token hashes agree across policies, restored counts agree, restored identities differ between policies, and per-sample cache records show a positive cache-length reduction. Every condition in Table 13 passes.

Risk models and seeds. Logistic regression (linear families) and random forests (extended families) are trained with 5-fold cross-validation, and all reported scores are out-of-fold. Both the Qwen3-VL-8B study and the Qwen3-VL-4B main grid aggregate the same six seeds, and reported values are means over seeds (per-seed std $\leq .014$ in AUROC and $\leq .010$ in simulated error).

A.3 PUBLISHED-STRATEGY ADAPTATIONS

Table 3: Published compression strategies as adaptations on the main grid (Qwen3-VL-4B, matched 10% retention), CIF rate. †: the subject prunes visual tokens with a single-layer criterion. SnapKV uses the window adaptation described in the text.

Strategy	ChartQA	DocVQA	OCRBench	TextVQA	VizWiz	POPE	MMStar
StreamingLLM (2024)	.187	.360	.512	.275	.106	.000	.000
SnapKV (2024)	**.089**	**.090**	**.345**	**.123**	**.075**	.000	.000
H2O (2023)	.146	.260	.519	.190	.096	.000	.000
ElasticCache (2024d)	.128	.255	.496	.237	.107	.000	.000
PrefixKV (2025a)	.174	.389	.557	.314	.211	.000	.000
FastV-style pruning† (2024a)	.164	.343	.546	.263	.103	.000	.000

Adaptation and validity. The five strategies are source-derived adaptations to Qwen3-VL. Each cell is checked for measured retained ratio at the 10% budget (all cells land in 0.100–0.101), KV-cache reduction, non-degenerate outputs, and backend identity. Average accuracy deviates from the published numbers by -8.8 pp (H2O, DocVQA) to $+52.5$ pp (SnapKV, ChartQA) on 4B and by -9.5 to $+38.3$ pp on 8B. The retention checks rule out under-compression as the cause of the positive deviations. SnapKV requires one further adaptation. The released recipe requires its 64-token observation window to be smaller than the prompt budget, and the strict 10% budget violates this on 100% of POPE, 94% of MMStar, 93% of ChartQA, and 71% of OCRBench prompts. Clamping the window to the budget evicts the entire visual prefix on such prompts, and the continuation degenerates after a correct first token. Every SnapKV cell in this paper therefore uses a window of $\min(64, \text{budget}/2)$, so pooled selection always receives at least half of the budget. The same rule is applied on LLaVA-NeXT and InternVL3.5. The clamped window scores 18.5% on ChartQA, close to the published 23.5%, and the adapted window scores 76.0%. This regime difference accounts for the large positive SnapKV deviations above.

Author-exact bound (official SnapKV). To bound the adaptation gap directly, the official SnapKV implementation was run unmodified on its supported LLaVA-NeXT-Mistral backbone at the same per-sample 10% budget against a shared reference. It lands within 0.52/0.69 pp of our adaptation on ChartQA/DocVQA, with CIF 257 vs. 238 and 444 vs. 405, and with *stress-failure-set* τ (computed over all stress errors, CIF plus both-wrong, rather than over CIF sets alone) of 0.16/0.18, against same-backbone cross-method stress-failure-set τ up to 0.51. The protocol-level quantities therefore agree.

FastV-style visual pruner. The sixth adaptation transplants FastV's single-layer (layer-2) attention selection criterion into the same eviction framework, matched to the main grid's measured budget. It differs from the released FastV in three respects. It evicts KV entries in every pruned layer instead of skipping computation from layer 2 onward, uses the framework's per-sample retention budget instead of FastV's global attention-rank budget, and retains decode-time eviction. Because it evicts after prefill, unlike the released FastV, which drops visual tokens from layer

2 onward during prefill, its opening token is unaffected. Its zero CIF on POPE and MMStar is consistent with this placement (§5.3). On the five affected datasets, stress-failure-set τ against the five KV strategies spans 0.09–0.64, and against the controlled policies it reaches 0.82 on DocVQA under high-attention retention.

A.4 Scope and Limitations

Published strategies are evaluated as source-derived adaptations. An author-exact run of official SnapKV agrees within 0.7 pp with stress-failure-set $\tau \leq 0.18$, and broader author-exact coverage is limited by official backbone support (Appendix A.3). Four 4B-matrix datasets use $\rho = 0.05$ references. Strict no-prune reruns of these datasets shift the CIF count by at most four cases and leave every conclusion unchanged. The cost model measures rerun budget rather than serving latency and presumes a warm reference configuration, and the instrumented research stack itself is tuned for logging rather than speed. Restoration is a diagnostic intervention conditioned on paired knowledge of the failure and evaluated on the RouteConf stress runs' failure sets; failure sets induced by other policies, and pre-pruning force-keep of routing-selected tokens, are the deployment-side tests to run next. All eviction in this study, controlled and adapted, follows prefill. Methods that prune during prefill, such as the released FastV, also expose the first output token, and that placement is outside our measurements. Most depth is on Qwen3-VL. LLaVA-NeXT replicates prediction and fallback on three datasets and restoration on two. InternVL3.5 contributes damage and turnover, and Qwen3.5 marks an applicability boundary.

B Which Requests Break: Damage and Turnover

This section supports §5.2. Appendix B.1 gives the damage under every policy, ratio, scale, and model family, and Appendix B.2 gives the margin-calibrated turnover behind Figure 1.

B.1 Damage across policies, ratios, and models

Figure 4 gives the bar-chart view of Table 1a, and Figure 5 plots the dose–response sweep behind §5.2.

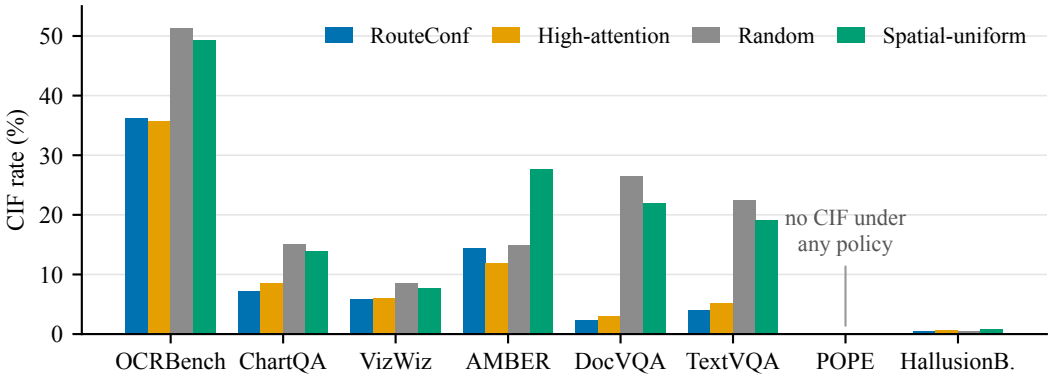


Figure 4: Bar-chart view of Table 1a: CIF rate at prune ratio 0.9 across eight of the ten evaluated benchmarks and four retention policies (Qwen3-VL-4B).

Qwen3-VL-8B paired transitions. Table 4 gives the full paired-outcome decomposition underlying §5.4. Reverse flips are an order of magnitude rarer than CIF on the OCR and document tasks, which bounds the harm of falling back to the reference. The 8B controlled cells on four datasets are listed in Table 1b.

Qwen3-VL-8B replication. The five strategies rerun on Qwen3-VL-8B for ChartQA, DocVQA, and OCRBench at 10% retention: CIF rates span .054–.152 (ChartQA), .069–.372 (DocVQA), and

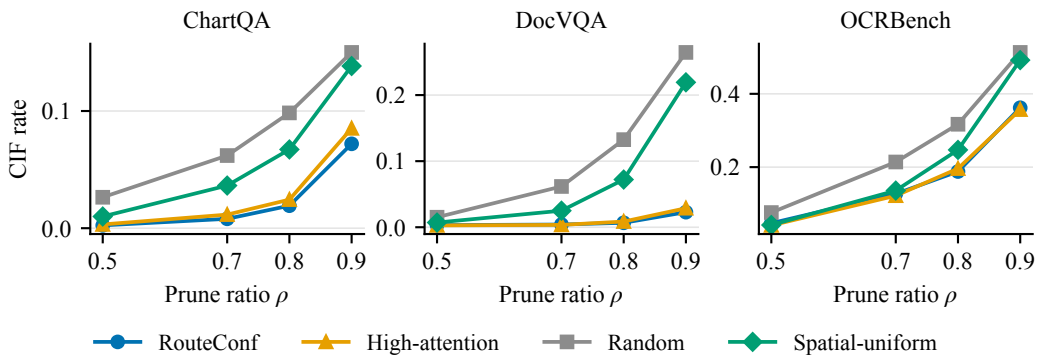


Figure 5: **Dose–response across retention policies.** CIF rate vs. prune ratio $\rho \in \{0.5, 0.7, 0.8, 0.9\}$ for all four retention policies on ChartQA, DocVQA, and OCRBench (Qwen3-VL-4B). Damage rises monotonically with ρ for every policy, and guided retention stays below random retention at every ratio.

Table 4: Paired transitions between the no-prune reference and the $\rho = 0.9$ stress run (Qwen3-VL-8B scale replication, RouteConf retention). DocVQA is scored under exact-match correctness in this table and in the restoration analysis; the ANLS-threshold view of the same run yields 171 CIF (rate .032), and the 4B grid (Table 1a) and the published-strategy evaluation use the ANLS-threshold view.

Dataset	n	Both+	CIF	Reverse	Both−	CIF rate	Rev. rate
AMBER	1004	832	110	39	23	.110	.039
VizWiz	4319	3210	220	56	833	.051	.013
ChartQA	2500	1976	151	21	352	.060	.008
DocVQA	5349	4088	596	54	611	.111	.010
OCRBench	1000	414	362	8	216	.362	.008

.317–.558 (OCRBench), and turnover reaches 0.88 (DocVQA, H2O vs. SnapKV). The wide damage spread at matched budget and the failure-set disagreement both match the 4B structure.

Family transfer: LLaVA-NeXT-Mistral-7B and InternVL3.5. Reference accuracy: LLaVA-NeXT ChartQA 53.0 (paper 53.0), DocVQA ANLS 63.3 (paper 63.7), TextVQA 64.7, OCRBench 49.4; InternVL3.5-8B ChartQA 87.0, DocVQA 92.3, TextVQA 77.9, OCRBench 78.4. CIF rates at 10% retention on ChartQA, DocVQA, TextVQA, and OCRBench are in Table 1b. The InternVL3.5 routing port applies the RouteConf score with a single global 10% visual budget, since that runner prunes the same positions in every layer. Pairwise CIF-set turnover on LLaVA-NeXT reaches 0.75/0.79 for routing–SnapKV on ChartQA/DocVQA against margin floors of 0.59/0.70, and the extreme routing–random pair on TextVQA ($\tau = 0.95$, floor 0.93) is margin-forced. InternVL3.5 shows stress-failure-set turnover 0.80 against a floor of 0.77 on DocVQA (attention–random).

B.2 Margin-calibrated turnover

Pairwise turnover, margin-calibrated. Failure-set turnover τ for every pair of published-strategy adaptations per dataset (both scales), together with the controlled-policy pairs of the main grid, feeds Figure 1. For every pair we report τ alongside three calibration quantities: the *margin floor* $1 - \min(|A|, |B|)/\max(|A|, |B|)$ (the smallest τ that the failure-rate gap permits), the *matched-margin expectation* (the τ obtained if the two sets were placed independently given their sizes), and the *exclusive-failure mass* $(|A| + |B| - 2|A \cap B|)/n$ (the fraction of the dataset failing under exactly one condition, robust to the small-set pathology of τ). Three regularities stand out. (i) 19 of the 27 pairs within 2 pp of average accuracy have exclusive-failure mass above 5% of the dataset, up to 24.5%. Gaps are rounded to 0.01 pp before the strict < 2 test. Fifteen of the 19 are published-strategy pairs. The other four are controlled pairs, i.e., random vs. spatial (ChartQA),

high-attention vs. random (VizWiz), and RouteConf vs. high-attention and vs. random on AMBER. The remaining eight are controlled pairs that share most failures, i.e., RouteConf vs. high-attention on five datasets and three spatial pairs on VizWiz. (ii) For the 19 diverging pairs the floor stays small (at most 0.30), so their turnover exceeds what the margin alone forces. (iii) Observed τ sits below the matched-margin expectation for every pair with both CIF sets nonempty (e.g., 0.51 vs. 0.85 for ElasticCache–H2O on DocVQA). The failure sets thus overlap more than independent placement predicts. The pair universe is the 116 pairs of Figure 1. It contains the ten published-strategy pairs on each affected dataset (4B, five datasets; 8B, three) and the six controlled-policy pairs of the 4B grid on each of its six affected datasets.

Controlled-policy overlap. Table 5 reports the Jaccard overlap J between the CIF sets induced by different retention policies on Qwen3-VL-4B, where failure-set turnover is $\tau = 1 - J$. Score-based policies (RouteConf vs. high-attention) share most failures on OCR-heavy tasks (J up to 0.90 on OCRBench) but as little as $J = 0.35$ on AMBER despite nearly matched average error. Random retention shares fewer failures with RouteConf than high-attention retention does on every dataset (J from 0.07 on DocVQA to 0.67 on OCRBench). On DocVQA and TextVQA these overlaps sit close to their margin floors, so they mostly reflect the size difference between the two failure sets.

Table 5: Jaccard overlap J of CIF sets between retention policies (Qwen3-VL-4B); turnover is the complement, $\tau = 1 - J$.

Dataset	RC ∩ High-attn	RC ∩ Random	RC ∩ Spatial
OCRBench	0.902	0.673	0.698
ChartQA	0.644	0.341	0.392
VizWiz	0.769	0.427	0.500
AMBER	0.354	0.210	0.233
DocVQA	0.517	0.068	0.072
TextVQA	0.637	0.142	0.159

C When Benchmarks Register Failures

This section supports §5.3. Appendix C.1 gives the answer-placement evidence and the answer-prefix variant, and Appendix C.2 documents two preconditions for applying the protocol.

C.1 Answer placement and the answer-prefix variant

Eviction follows the prefill forward (§4.1), so the opening token of every stress run precedes eviction. Its probability is bitwise identical to the reference on all 1220 TextVQA (8B) and 1500 MMStar (4B) samples compared. A score that is fixed by the first token can therefore change only through a failure to stop. Every POPE reference answer is one token, 95.2% of MMStar answers are one letter followed by the end-of-sequence token, and 83% of HallusionBench answers open with their yes/no. All ten MMStar stress conditions (four controlled policies, six adaptations) register zero flips in either direction. Pruning is active, since the output text changes on 71–379 samples per condition. The normalized answer letter still changes on no reference-correct sample, and 59–92% of the text changes keep the same normalized answer. The MMStar and GQA units run the same effective backends at the same total keep ratio (0.100) as the main grid, and the same pipeline registers 79–92 flips per policy on GQA. The noise-image control replaces every image with a fixed-seed random-RGB image under the strict no-prune protocol. MMStar then drops from 59.7% to 25.7% and POPE from 91.4% to a uniform "no" at 50.0%. Both benchmarks therefore require vision at prefill. Their zero CIF is consistent with the answer being decided before eviction. The answer-prefix variant moves the scored token behind eviction. It reuses the POPE random split with a post-prompt that asks for the sentence "The answer is yes." or "The answer is no.", so the yes/no is the fourth generated token, and the parser takes the last yes/no after "answer is". The strict no-prune reference scores .909 on all 3000 samples (standard prompt: .914). At $\rho = 0.9$, routing-score, high-attention, spatially uniform, and random retention score .911, .907, .906, and .906, with 14, 8, 10, and 10 CIF cases. The prefix tokens and the kept prompt entries were computed with the full image in context,

so this variant bounds the effect of answer placement and does not remove every path by which visual information survives eviction. Table 6 recomputes Table 1a on reference answers of two or more tokens, the samples whose scored answer depends on at least one token decoded from the compressed cache. The ordering of Table 1a is unchanged, and every text, chart, and document cell rises.

Table 6: CIF rate restricted to reference answers of two or more tokens (Qwen3-VL-4B, $\rho = 0.9$, the runs of Table 1a). Share: fraction of reference answers with two or more tokens. POPE has none.

Dataset	Share	RouteConf	High-attn	Random	Spatial
OCRBench	.90	.401	.396	.569	.545
ChartQA	.85	.084	.099	.175	.162
VizWiz	.80	.063	.066	.099	.089
AMBER	1.00	.144	.119	.148	.277
DocVQA	.91	.025	.031	.291	.242
TextVQA	.76	.051	.066	.293	.249
GQA	.21	.018	.014	.018	.018
MMStar	.05	.000	.000	.000	.000
HallusionB.	.82	.005	.008	.005	.009

C.2 Applicability preconditions: hybrid-attention Qwen3.5

Porting the instrumentation to Qwen3.5 (4B/9B), whose hybrid architecture confines full attention, and thus token-level KV pruning, to 8 of 32 layers, illustrates two preconditions of the protocol. **(1) The reference must have dynamic range.** Because CIF is upper-bounded by the number of reference-correct samples, a model with low reference accuracy on a task cannot register damage there. Under our matched short-generation protocol, Qwen3.5's reference is correct on only 10–15/1000 OCRBench samples, which makes paired analysis uninformative on the OCR and document tasks. **(2) Proxy scores must be read jointly with reverse flips.** On AMBER (Table 7), pruning shifts Qwen3.5's output mode so strongly that *random* retention *improves* the proxy score (854/1004 correct against a 472/1004 reference, reverse-flip rate 0.44), while spatially uniform retention drops it to 395/1004 (CIF rate 0.25). A proxy improvement under compression is not evidence that visual evidence was preserved.

Table 7: Qwen3.5-9B AMBER backend diagnostics under $\rho = 0.9$ full-attention KV pruning (reference correct: 472/1004). Random pruning *raises* the proxy score while flipping nearly half the answers.

Retention policy	Stress correct	CIF	Reverse flips
Routing-score	577	116	221
Random	854	63	445
Spatially uniform	395	254	177

D Prediction and Fallback

This section supports §5.4. Appendix D.1 reports the risk ranking on all three models, Appendix D.2 tests whether the signal is specific to CIF and compares sampling-based scores, Appendix D.3 reports fallback for every trigger and coverage, and Appendix D.4 gives the measured cost.

D.1 Risk ranking on Qwen3-VL-4B, Qwen3-VL-8B, and LLaVA-NeXT

Table 8 reports the six-seed out-of-fold ranking for every feature family on all three models (5-fold, mean over six seeds, per-seed std $\leq .014$ AUROC). Generation confidence is the strongest single family in every row. On Qwen3-VL-4B, HallusionBench (4 positives) and POPE (0 positives) are omitted as unstable or undefined. The same 4B tables for the spatially uniform and high-attention

backends reproduce this ordering in every cell with a defined confidence score. On LLaVA-NeXT, the SnapKV adaptation emits no routing trace, so only confidence is defined. It reaches AUROC .679/.717/.823 on ChartQA/DocVQA/TextVQA. AUPRC is low under LLaVA-NeXT routing retention because the CIF base rates are 1–4%.

Table 8: CIF risk ranking by feature family (six seeds, out-of-fold), AUROC / AUPRC. Best per row in **bold**. Generation confidence is the strongest single family in every row. Fusing routing features (Combined) shifts AUPRC by -0.043 to $+0.043$ on the Qwen rows, with no consistent gain.

Dataset	Random	Retention	Attention	Routing-ext	Confidence	Combined
Qwen3-VL-4B, RouteConf retention						
AMBER	.508 / .156	.667 / .225	.628 / .197	.630 / .189	**.724 / .281**	.713 / .264
ChartQA	.498 / .074	.571 / .084	.570 / .083	.605 / .116	**.783 / .211**	.776 / .169
DocVQA	.504 / .024	.580 / .062	.643 / .064	.591 / .057	**.882 / .219**	.877 / .197
OCRBench	.500 / .362	.675 / .542	.666 / .497	.703 / .555	.759 / .605	**.793 / .648**
TextVQA	.496 / .042	.542 / .055	.601 / .068	.595 / .064	**.875** / .234	.873 / **.266**
VizWiz	.493 / .057	.577 / .086	.663 / .121	.633 / .106	.885 / .294	**.893 / .327**
Qwen3-VL-8B, RouteConf retention						
AMBER	.522 / .122	.658 / .165	.669 / .189	.655 / .198	.715 / **.259**	**.739** / .257
ChartQA	.501 / .064	.615 / .095	.636 / .117	.655 / .106	.798 / .188	**.825 / .225**
DocVQA	.496 / .112	.530 / .141	.620 / .167	.613 / .181	.881 / **.542**	**.881** / .519
OCRBench	.494 / .357	.658 / .488	.694 / .543	.719 / .587	.821 / **.706**	**.824** / .703
VizWiz	.489 / .052	.581 / .071	.694 / .102	.676 / .108	.878 / .326	**.889 / .342**
LLaVA-NeXT-Mistral-7B, routing retention						
ChartQA	.487 / .040	.519 / .044	.492 / .042	.552 / .048	**.734 / .089**	.696 / .077
DocVQA	.491 / .023	.492 / .024	.583 / .029	.545 / .027	**.748 / .053**	.724 / .048
TextVQA	.505 / .011	.641 / .032	.598 / .016	.614 / .016	**.850** / .040	.838 / **.069**

Frozen cross-family transfer. Two questions are kept apart: whether a trained detector transfers unchanged, and whether the same recipe holds when refit on the new model. For each dataset and seed, the LLaVA-NeXT rows are split by stratified group 5-fold over image or document identity (ChartQA image hash, DocVQA document id, TextVQA image id), and the source model's training rows exclude every image and document in the target test fold. A confidence predictor fit on Qwen3-VL-4B rows of the same dataset and applied unchanged reaches AUROC .668/.679/.788 (ChartQA/DocVQA/TextVQA); fit on the other datasets of Qwen3-VL-4B, .672/.724/.758; from Qwen3-VL-8B, .689/.669 on ChartQA/DocVQA. The same recipe refit on LLaVA-NeXT reaches .734/.747/.846 on identical folds. At 20% coverage the frozen 4B predictor recovers 31–55% of the target CIF pool against 44–73% for the refit predictor. The signal therefore carries across decoder families. Target calibration recovers the remaining .05–.07 AUROC.

D.2 CIF-SPECIFICITY OF THE RISK SIGNAL, AND SAMPLING-BASED BASELINES

Specificity (three-target analysis). For each of the 11 dataset–scale pairs we train the same feature families against three targets: (a) CIF over all samples, (b) generic stress error, and (c) CIF versus both-wrong restricted to stress-wrong samples. In contrast (c) every sample is an error, so the error label itself provides no ranking signal, and any separation tracks whether the failure was compression-induced rather than the presence of an error. On (c), generation confidence and the combined family reach AUROC .61–.83 (mean .72; every cell more than two seed-stds above chance). The signal transfers: leave-one-dataset-out training averages AUROC .75 over 600 cells (99% above 0.5), and RouteConf-trained scores applied unchanged to the five published KV adaptations average .76 over 480 cells. The detector is also a competent generic stress-error ranker (target b), but (c) and both transfer matrices show that it carries CIF-specific information beyond that.

Sampling-based baselines. Five stochastic replicas per dataset (T=0.7, top-p 1.0, top-k 20) yield semantic-entropy and self-consistency scores against the same CIF target (Table 9). Confidence leads semantic entropy by 0.12 mean AUROC (per-dataset gaps 0.08–0.22) and recovers on average 13.5 points more of the CIF pool at 20% coverage. With five samples the training-free scores also

quantize risk into few levels, which makes their fallback selection coarse. The sampled replicas cost five stress passes per request against one for confidence.

Table 9: Sampling-based UQ baselines vs. generation confidence (Qwen3-VL-4B, RouteConf backend, CIF target). Semantic entropy (SE) and 1-consistency use $K{=}5$ samples at $T{=}0.7$; confidence is the trained six-seed out-of-fold score. Recovery is the fraction of the CIF pool recovered by fallback at 20% coverage.

Dataset	SE AUROC	1-cons. AUROC	Conf. AUROC	SE recov.	Conf. recov.
ChartQA	.698	.697	**.783**	.428	**.535**
DocVQA	.803	.802	**.882**	.721	**.788**
OCRBench	.664	.664	**.759**	.329	**.364**
VizWiz	.751	.747	**.885**	.530	**.803**
AMBER	.507	.507	**.724**	.234	**.429**

D.3 Fallback at every trigger and coverage

Figure 6 plots the 8B error-vs-coverage grid.

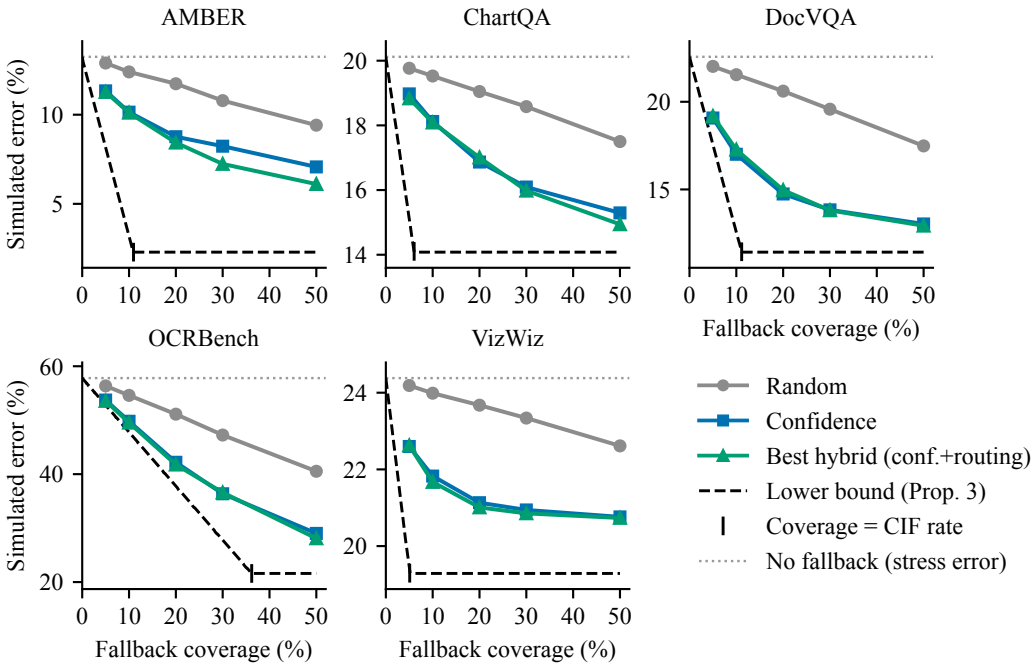


Figure 6: Simulated error vs. fallback coverage (Qwen3-VL-8B, six seeds). The best hybrid confidence+routing trigger (green) differs from confidence-only (blue) by at most .01 at any coverage. The dashed line is the analytic lower bound of Proposition 3. It falls with slope one until coverage reaches the CIF rate (tick) and is flat beyond it. The gap to this bound is the open headroom. Random selection (gray) shows the value of any learned trigger. The tabulated oracle of Table 10 fills its residual budget by risk rank and sits at or slightly above the bound.

Fallback at 20% coverage. Table 10 reports simulated error after fallback at 20% coverage for every trigger on both scales (six seeds, std $\leq$ .010). Confidence is within .01 of the best non-oracle trigger on every dataset at both scales. Table 9 compares confidence with the sampling-based scores.

Best policy by dataset and coverage (Qwen3-VL-8B). Table 11 summarizes the full 5×5 coverage grid behind the hybrid-vs-confidence comparison in §5.4. For each dataset–coverage cell it

Table 10: Simulated error after fallback at 20% coverage (six seeds), by trigger. Routing: routing-extended features alone. Resid.: residual 70/30 (confidence, then routing). 2-gate: rank fusion of confidence and routing percentiles. Best non-oracle trigger in **bold**. Hybrid triggers were run on Qwen3-VL-8B only (–).

Dataset	Stress err.	Random	Routing	Conf.	Comb.	Resid. 70/30	2-gate mean	2-gate max	Oracle
Qwen3-VL-4B									
AMBER	.163	.142	.132	**.114**	.118	–	–	–	.023
ChartQA	.222	.210	.204	**.189**	.194	–	–	–	.153
DocVQA	.075	.071	.068	**.059**	**.059**	–	–	–	.053
OCRBench	.570	.501	.456	.438	**.429**	–	–	–	.370
TextVQA	.175	.170	.164	.151	**.150**	–	–	–	.136
VizWiz	.273	.264	.256	**.233**	**.233**	–	–	–	.218
Qwen3-VL-8B									
AMBER	.133	.117	.093	.088	.090	.085	.093	**.084**	.028
ChartQA	.201	.191	.179	**.169**	.170	.170	.171	.172	.143
DocVQA	.226	.206	.192	**.147**	.150	.154	.168	.159	.116
OCRBench	.578	.511	.444	.422	.429	.429	**.418**	.432	.378
VizWiz	.244	.237	.228	.211	**.210**	**.210**	.219	.212	.196

reports the best non-oracle policy, its simulated error, and its delta against confidence-only (hybrids win 19 of 25 cells, by $< .01$ error on average). Negative deltas improve on confidence, and zero deltas occur exactly when confidence itself is the best policy.

Table 11: Best non-oracle trigger per dataset and coverage (Qwen3-VL-8B, six seeds). Δ is simulated error minus confidence-only simulated error at equal coverage.

	5%			10%			20%			30%			50%		
Dataset	Best	Err	Δ	Best	Err	Δ	Best	Err	Δ	Best	Err	Δ	Best	Err	Δ
AMBER	2g-m	.113	−.001	res	.101	−.000	2g-x	.084	−.003	2g-x	.072	−.010	comb	.061	−.010
ChartQA	comb	.188	−.001	comb	.181	−.000	conf	.169	.000	comb	.160	−.001	comb	.149	−.004
DocVQA	conf	.191	.000	conf	.170	.000	conf	.147	.000	comb	.138	−.000	comb	.129	−.001
OCRBench	2g-m	.536	−.002	2g-m	.495	−.003	2g-m	.418	−.004	conf	.364	.000	res-r	.281	−.009
VizWiz	conf	.226	.000	comb	.217	−.002	comb	.210	−.001	comb	.209	−.001	2g-m	.207	−.000

Triggers: conf = confidence-only; comb = jointly trained combined model; res = residual 70/30 (confidence then routing); res-r = residual 50/50 routing-first; 2g-m / 2g-x = two-gate rank fusion (mean / max).

Fallback on published strategies. Fallback with the confidence trigger (5-fold out-of-fold generation features per backend) at 20% coverage against the shared 8B references cuts simulated error by 9–42% relative and recovers 27–68% of each strategy's induced excess error. Reverse-flip harm is bounded by the strategies' reverse rates of 0.1–1.2%.

D.4 Cost profiling

These profiles measure the rerun budget behind the fallback costs of §5.4. The instrumented stress pass carries the routing and logging overhead of the protocol (§A.2), so it runs slower than the uncompressed reference on these subsets. Profiles use representative CIF subsets under matched decoding (greedy, 64 new tokens) on a single GPU.

With $T_{\text{ref}}/T_{\text{str}}$ of 0.861 (AMBER) and 0.674 (VizWiz), fallback costs $1.09\times$/$1.07\times$ at 10% coverage, $1.17\times$/$1.13\times$ at 20%, and $1.43\times$/$1.34\times$ at 50%. The 4B main grid profiles identically on its CIF subsets (ChartQA: stress 128 s vs. reference 102 s; OCRBench: 427 vs. 305 s; peak memory $\approx$10.8 GiB in both modes). This gives $T_{\text{ref}}/T_{\text{str}}$ of 0.71–0.80 and a cost of 1.14–$1.16\times$ at 20% coverage. These multipliers are relative to the stress pass. On a stack where compression genuinely saves time ($T_{\text{ref}}/T_{\text{str}} > 1$), the same coverage costs more relative to the compressed path—e.g., $T_{\text{str}} = 0.5\,T_{\text{ref}}$ gives $1.4\times$ at 20% coverage—though the total remains below a full-fidelity run

Table 12: Command-level wall time and memory on the instrumented stack (Qwen3-VL-8B, single GPU).

Dataset	Mode	Requests	Wall (s)	s/request	Peak mem (MiB)
AMBER	stress $\rho = 0.9$	110	267	2.43	18 892
AMBER	reference	110	230	2.09	18 892
VizWiz	stress $\rho = 0.9$	92	144	1.57	19 106
VizWiz	reference	92	97	1.05	19 106

($0.5 + 0.2 = 0.7\times$ the reference), so fallback still recovers a large share of the damage below uncompressed cost. Peak memory is weight-dominated at single-request scale and unchanged across modes.

E Restoration Details

Table 13: Recovered CIF cases at restoration budgets $\gamma = 5/10/20\%$, by selection policy, under verified matched interventions. Best per row in **bold**. Rows 4B and 8B are Qwen3-VL. Rows LLaVA-7B are LLaVA-NeXT-Mistral-7B on the frozen CIF sets of its routing-score stress runs, where no spatial policy was run. The 8B OCRBench set (356 cases) is frozen from a separate strict-reference run.

Model	Dataset (CIF n)	RouteConf	High-attn	Random	Spatial
4B	ChartQA (180)	**104 / 139 / 161**	88 / 121 / 156	81 / 105 / 126	73 / 109 / 135
4B	OCRBench (362)	**139** / 177 / 235	123 / 172 / 231	97 / 148 / 233	103 / **178** / **256**
8B	ChartQA (151)	**96 / 118** / 130	84 / 112 / **131**	57 / 65 / 89	47 / 69 / 94
8B	OCRBench (356)	**138** / 182 / 245	127 / 181 / 236	118 / 179 / **258**	126 / **199** / 257
8B	DocVQA (596)	**561 / 577 / 586**	547 / 576 / 583	519 / 537 / 568	536 / 560 / 583
8B	VizWiz (220)	**139 / 172 / 198**	122 / 161 / 196	99 / 136 / 176	109 / 154 / 181
LLaVA-7B	ChartQA (97)	**42 / 58 / 78**	36 / **58** / 71	23 / 30 / 43	–
LLaVA-7B	DocVQA (123)	**54 / 78 / 103**	45 / 73 / 100	25 / 41 / 59	–

Intervention semantics. Restoration is applied *after* pruning within the same forward pass. With budget γ, the stack restores $\lceil \gamma N_v \rceil$ visual tokens per active layer, selected from the tokens that the $\rho = 0.9$ run would otherwise prune, with identity determined by the restoration policy. It is a post-pruning intervention into the compressed cache, distinct from an independent run at a lower prune ratio. Measured final retained visual ratios (Qwen3-VL-8B CIF subsets) are approximately 0.097, 0.146, and 0.245 on ChartQA at $\gamma = 5/10/20\%$, and 0.088, 0.137, and 0.237 on OCRBench. On LLaVA-NeXT ChartQA, the mean per-layer cache grows from 225 tokens at $\gamma = 0$ to 336, 446, and 671 tokens at $\gamma = 5/10/20\%$. These are not the dataset-mean retained ratios of Table 1a plus γ: budgets are applied per active layer, and the stress run's baseline keep on the CIF subset differs from the dataset mean in Table 1a.

Pairwise significance. All $\binom{4}{2} = 6$ policy pairs per dataset–budget cell (three on LLaVA-NeXT) were tested with paired sign tests on the discordant recoveries. RouteConf vs. high-attention is significant at $\gamma = 5\%$ in five of six Qwen settings (4B ChartQA $p = .0009$, 4B OCRBench $p = .005$, 8B ChartQA $p = .012$, 8B DocVQA $p = .013$, 8B VizWiz $p = .005$) and marginal on 8B OCRBench ($p = .052$). RouteConf vs. random and vs. spatial are significant at 5% everywhere except 8B OCRBench. At $\gamma = 20\%$ the RouteConf–high-attention contrast is significant nowhere, because high-budget restoration saturates and the two guided selections stop being distinguishable. OCRBench shows the weakest separation at both scales, consistent with its uniformly high damage (Table 1a), where the lost evidence is broad enough that many selections eventually cover it. On LLaVA-NeXT, RouteConf vs. random is significant at every budget on both datasets ($p \leq .0013$). RouteConf vs. high-attention has discordant counts of 11:5, 9:9, and 8:1 on ChartQA and 20:11,

10:5, and 7:4 on DocVQA at $\gamma = 5/10/20\%$. Only the ChartQA 20% cell reaches significance ($p = .039$), since the failure sets are small. All p-values are reported unadjusted for multiple comparisons. At $\gamma = 5\%$, the primary RouteConf-vs-high-attention contrast points the same way in all eight settings. A tighter budget of $\gamma = 2\%$, run on 4B ChartQA only, shows the same ordering (RouteConf 76 vs. high-attention 60 of 180).

Interpretation. A small, policy-selected subset of pruned KV entries repairs most failures post hoc. This is consistent with localized evidence loss. At tight budgets, routing-score selection repairs more failures than attention magnitude or coverage heuristics. At generous budgets the selections converge, and on OCRBench spatial selection leads.