AUDITING CAPABILITY RETENTION UNDER CROSS-LAYER KV SHARING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-layer key–value cache sharing diminishes the memory footprint of vision-language models; however, similarity-based acceptance criteria might fail to detect impairments in specific visual capabilities. We conduct an audit to determine whether the statistics employed in the approval process for sharing accurately predict the preservation of object–attribute binding. Utilizing reconstructed and adapted sharing mechanisms, we differentiate between static inter-layer similarity and representation similarity measured on the cumulative modified model. Additionally, we compare acceptance rules under controlled sharing operations and matched cache-reduction budgets. The evaluated gates can admit configurations with significant binding degradation, whereas aggregate visual question-answering and object-presence scores offer an incomplete assessment of the loss. Moreover, evaluating substitutions individually does not establish the retained capability when they are combined. We introduce *RecallGate*, a lightweight acceptance rule that temporarily incorporates each proposed substitution and evaluates the cumulative configuration on a small visual-recall battery. Matched-budget comparisons demonstrate that capability-based acceptance can preserve substantially more binding capability than proxy-based alternatives. Across various model families and selection strategies evaluated, its effectiveness is contingent upon probe sensitivity and baseline competence. These findings advocate for capability-specific evaluation of compression gates and independent validation of the final shared configuration, while positioning RecallGate as empirical screening rather than a general safety guarantee.

1 INTRODUCTION

Visual tokens increase the KV-cache storage required by vision-language models alongside text, constraining concurrent serving capacity (Kwon et al., 2023). A 336×336 image produces 576 visual tokens in the standard LLaVA-1.5 setting, while Qwen2-VL supports variable visual-token counts through dynamic resolution (Liu et al., 2024b; Wang et al., 2024). Cross-layer sharing can reduce this storage by allowing selected decoder layers to reuse tensors computed elsewhere.

Post-hoc sharing modifies a pretrained model without retraining it, creating a practical question: what evidence is sufficient to accept a proposed substitution? Existing methods use different signals and operations, including cache merging in MiniCache (Liu et al., 2024a), dissimilar-pair selection in KVSharer (Yang et al., 2024), and attention-pattern divergence in Q Cache (Zhuang et al., 2026). These approaches motivate examining whether the scores used to approve sharing are informative about the capabilities retained after installation. We study object–attribute binding: associating attributes with the correct visual objects, a distinction that vision-language models can struggle to capture (Yuksekgonul et al., 2022). Prior compression evaluations motivate capability-specific assessment beyond aggregate scores (Li et al., 2025; Yuan et al., 2024; Li et al., 2024b). Moreover, evaluating a substitution in isolation need not reveal its effect when combined with previously accepted substitutions. Our central question is whether proxy-based approval adequately tracks binding retention after cross-layer sharing, and whether directly evaluating the target capability on the cumulative modified configuration can improve sharing decisions. We further ask whether passing such a capability-aware gate is sufficient to ensure retention on independently evaluated layouts.

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.

Two distinctions guide the investigation: whether approval evaluates an individual substitution or the cumulative configuration, and whether it measures representation fidelity or the target capability. KVSharer already evaluates cumulative candidate additions using final-hidden-state similarity and reverts failures (Yang et al., 2024). Our proposed contribution is therefore the use of a targeted capability statistic for acceptance, rather than cumulative evaluation itself. We introduce RECALLGATE, an empirical capability-aware acceptance rule. It temporarily installs each proposed addition to the current sharing configuration and evaluates the resulting model on a targeted capability mini-battery (Figure 1). The addition is retained only if the measured score meets a fixed floor relative to the unmodified model; otherwise, the previous configuration is restored. This directly tests the capability of interest under the proposed cumulative intervention, although passing a finite battery does not guarantee held-out retention. We evaluate a Q Cache GLA-style reconstruction and adapted wholesale KV replacement rules across four VLM families. These are not complete reproductions of the original methods. Targeted screening improves several matched-savings comparisons, but selection variability and layout-dependent failures limit capability preservation.

Contributions. We (i) audit proxy–binding relationships across 104 singleton interventions from four VLM families; (ii) evaluate capability-aware acceptance against proxy approval under matched-savings controls, including five new SmolVLM2 selections at a fixed sharing budget; and (iii) show that capability-aware approval can substantially improve binding retention without guaranteeing healthy-model retention, identifying configuration-dependent transfer and a negative singleton–joint interaction on independently evaluated layouts.

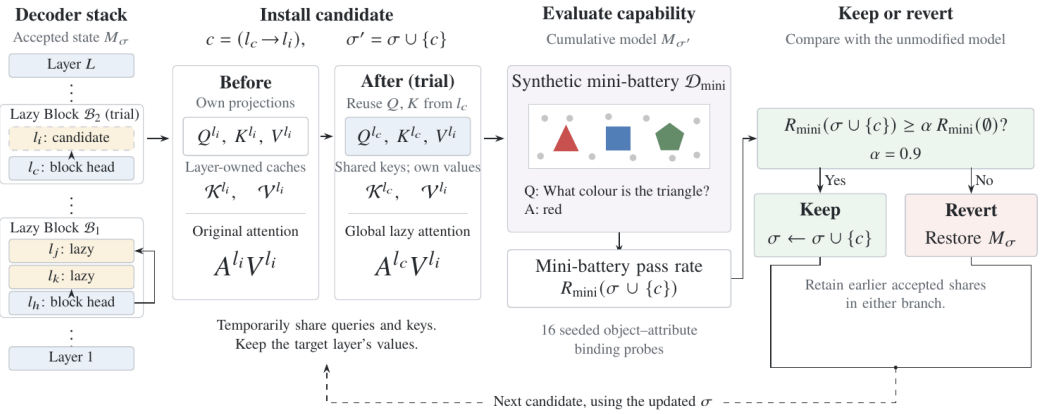


Figure 1: RecallGate temporarily installs each proposed share and retains it only if binding performance meets a fixed floor relative to the unmodified model. The diagram illustrates Q/K reuse with target-layer values retained; the same acceptance procedure applies to wholesale KV replacement.

2 BACKGROUND: CROSS-LAYER SHARING AND ACCEPTANCE

For a decoder that caches keys and values at every layer, nominal KV-cache storage is $2LBSn_\text{kv}d_\text{head}p$ bytes, where L is the number of layers, B the batch size, S the cached sequence length, n_kv the number of KV heads, d_head the head dimension, and p the bytes per scalar. This expression assumes uniform cache dimensions and precision across layers and excludes allocation overhead and other model memory. Cross-layer sharing reduces distinct cache storage when layers reuse tensors instead of storing separate copies. CLA (Brandon et al., 2024) and YOCO (Sun et al., 2024) incorporate cache sharing into model architectures; post-hoc methods instead modify pretrained models. Their candidate selection, acceptance criteria and tensor operations differ. Q Cache uses attention divergence between adjacent layers in its block-formation procedure (Zhuang et al., 2026); KVSharer evaluates additions to a cumulative sharing configuration using final-hidden-state similarity (Yang et al., 2024); and MiniCache merges cache representations while retaining magnitude information and selected distinct tokens (Liu et al., 2024a). An audit must therefore distinguish the approval signal from the operation it approves.

Object–attribute binding. Vision-language models can recognize objects and attributes without correctly associating them (Yuksekgonul et al., 2022). Binding probes test this correspondence, but benchmark biases can inflate apparent compositional ability (Hsieh et al., 2023). Prior evaluations of long-context methods and KV-cache compression similarly motivate capability-specific assessment (Li et al., 2025; Yuan et al., 2024; Li et al., 2024b).

Relation to visual-memory audits. CVMA (Chen et al., 2026) examines the relationship between current attention and the usefulness of visual regions in subsequent dialogue turns. Our study instead examines whether approval scores predict binding retention after cross-layer substitutions. The distinction is the intervention and its approval criterion, rather than the general observation that aggregate evaluation can obscure capability loss.

Evaluated adaptations. Our Q Cache GLA-style reconstruction (Zhuang et al., 2026) reuses source-layer queries and keys while retaining target-layer values. MC-R and KS-R adapt MiniCache-inspired similarity ranking (Liu et al., 2024a) and KVSharer-inspired dissimilarity ranking (Yang et al., 2024) to wholesale KV replacement. Recorded MC-R configurations include nonadjacent pairs and do not implement MiniCache's merge. Section 3.1 describes the inspected implementation; correspondence to every historical run remains unverified. These adaptations are not complete reproductions.

3 Problem Setup: Empirical Capability Gates

3.1 Sharing configurations and the audited gate

Let M have a frozen vision encoder and connector feeding an L-layer decoder. Candidate $c = (m, \ell)$, $m < \ell$, supplies source-layer tensors to target ℓ. Configuration σ contains accepted pairs, $n = n_{\mathrm{sh}}(\sigma)$ counts distinct shared targets, and M_σ denotes the modified model ($M_\varnothing$ is unmodified). MC-R/KS-R replace target K/V tensors, giving nominal savings n/L under uniform cache dimensions and precision, retained source tensors, and no separate target copies. The Q Cache GLA-style reconstruction (Zhuang et al., 2026) reuses block-head queries and keys but retains target values:

$$Q^{(\ell_i)} = Q^{(\ell_c)}, \qquad K^{(\ell_i)} = K^{(\ell_c)}, \qquad V^{(\ell_i)} = X^{(\ell_i)} W_V^{(\ell_i)}. \tag{1}$$

Here, $X^{(\ell_i)}$ is the target input under the installed configuration. Removing target key caches saves $n/(2L)$; query reuse adds no KV-storage saving. Neither fraction measures peak process memory or latency. The inspected MC-R/KS-R implementation ranks pairs by increasing/decreasing Euclidean distance, respectively, between fingerprints concatenating token-averaged K/V projections from the final calibration example. Fingerprints are normalized by default; ranking is fixed, with adjacency enforced only when configured. Acceptance measures cumulative hidden-state change:

$$D_{\mathrm{hid}}(\sigma') = 1 - \frac{1}{N} \sum_{i=1}^{N} \cos\big(\bar{h}_i(\sigma'), \bar{h}_i(\varnothing)\big)\,, \tag{2}$$

where $\bar{h}_i$ averages final-layer states over tokens for calibration example i. An eligible addition $\sigma' = \sigma \cup \{c\}$ passes when $D_{\mathrm{hid}}(\sigma') < \varepsilon$. RecallGate also requires the capability test below; rejected additions are reverted. Q Cache-style formation sequentially scans adjacent layers. For calibration set $\mathcal{D}_{\mathrm{cal}} = \{x_i\}_{i=1}^{N}$ and unmodified-model last-row attention distributions $a_i^{(\ell)}$, the score is

$$S(\ell - 1, \ell) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{JS}\Big(a_i^{(\ell-1)} \,\|\, a_i^{(\ell)}\Big)\,. \tag{3}$$

Adding ℓ requires $S(\ell - 1, \ell) < \varepsilon$ and satisfaction of formation constraints. This static adjacent-layer comparison can approve substitutions from an earlier block head. For capability performance C on a fixed evaluation distribution and required floor τ_C, we seek

$$g(\sigma) = 1 \quad \Longrightarrow \quad C(M_\sigma) \geq \tau_C. \tag{4}$$

Neither the audited proxy nor RecallGate establishes this guarantee.

3.2 RECALLGATE

RecallGate temporarily installs $\sigma' = \sigma \cup \{c\}$ and evaluates a seeded battery $\mathcal{D}_{\text{mini}}$ of $n_{\text{mini}} = 16$ V-BIND object–attribute probes. The POPE ablation substitutes object-presence evaluation. For battery pass rate R_{mini}, acceptance requires

$$R_{\text{mini}}(\sigma') \geq r_{\min}, \qquad r_{\min} = \alpha R_{\text{mini}}(\varnothing), \quad \alpha = 0.90. \tag{5}$$

The unmodified-model reference uses the same battery and remains fixed. Passing updates $\sigma \leftarrow \sigma'$; failure restores σ. The primary implementation first requires proxy approval. Removing this filter in the proxy-free ablation changes candidate exposure and cost, so it does not isolate additional computation. Acceptance on the reused battery guarantees neither held-out retention nor other capabilities. With equally weighted binary scores, an unmodified pass count K_0 requires $\lceil \alpha K_0 \rceil$ candidate passes: eight for $K_0 = 8$, fifteen for $K_0 = 16$, and none for $K_0 = 0$. Quantization, weak baselines and adaptive reuse (Dwork et al., 2015) limit interpretation.

3.3 STATISTICAL REPORTING

Fixed-configuration SDs describe evaluation-draw variation, not gain uncertainty. Tables 4 and 5 count three distinct evaluation-seed offsets once each. Selection summaries instead report sample SDs across five gating seeds on fixed evaluation items. Layout confidence intervals use paired item resampling within layouts; they do not quantify selection variability. Singleton associations use within-family Spearman correlations. Shared-layer dependence and heterogeneous associations limit inference. Appendix F provides statistical details.

4 EXPERIMENTAL SETUP

Models. We evaluate five checkpoints from four families (Table 1). Qwen2-VL-2B (Wang et al., 2024) provides binding headroom; LLaVA-1.5-7B (Liu et al., 2024b) is the Q Cache reconstruction target (Zhuang et al., 2026). OneVision-0.5B (Li et al., 2024a) and SmolVLM2-2.2B (Marafioti et al., 2025) broaden architectural coverage; LLaVA-1.5-13B tests scale within one family.

Table 1: Checkpoints and nominal fp16 cache storage. Savings assume a 12.5% reduction for one 4096-token sequence; KB/MB are decimal.

Model	KV/heads	Layers	Cache/token	Saved
LLaVA-1.5-13B	40/40	40	819 KB	419 MB
LLaVA-1.5-7B	32/32	32	524 KB	268 MB
Qwen2-VL-2B	2/12	28	28.7 KB	15 MB
LLaVA-OneVision-0.5B	2/14	24	12.3 KB	6 MB
SmolVLM2-2.2B	32/32	24	196.6 KB	100.7 MB

Interventions and acceptance rules. Q Cache comparisons use our all-token GLA-style Q/K-sharing reconstruction, retaining target-layer values. The unsuccessful visual-only VLA reconstruction is excluded from performance comparisons. MC-R and KS-R use adapted similarity-ranked and dissimilarity-ranked selection rules, respectively, under wholesale KV replacement. Candidate space, operation and maximum audition count are fixed within each comparison; a shared cap does not imply equal compute. Section 3.1 specifies the inspected implementation. Headline records establish thresholds, selected pairs, savings and evaluation seeds; these records alone do not establish code-version correspondence for every historical run. The adaptations are not complete reproductions.

Evaluation. V-BIND-NIAH uses 100 binding items per headline evaluation draw; X-FRAME-ID tests cross-frame identity. GQA (Hudson & Manning, 2019) and POPE (Li et al., 2023) provide descriptive scores on 100 and 30 sampled examples, respectively, using custom scoring. Synthetic gating and original evaluation use different item seeds but shared template families. The controlled SmolVLM2 KS-R experiment fixes $\varepsilon = 0.03$ and a two-share budget, repeating selection with gating seeds 7100–7104.

Appendix A specifies generation and scoring.

Comparison protocol. All conditions use the same evaluation harness. Thresholds are interpreted within selection principles because score definitions and directions differ. Matched-threshold comparisons fix the proxy threshold but may yield different savings. Matched-savings comparisons fix shared-target counts within a model and operation, controlling nominal compression dose. They do not necessarily match selected pairs, compute or proxy thresholds. Qwen2-VL uses different thresholds across acceptance rules (Table 4), confounding threshold and acceptance effects. SmolVLM2 and OneVision provide same-threshold, matched-savings comparisons.

5 RESULTS

5.1 RECONSTRUCTION CHECKS

Component checks cover scoring, block formation, prefill/decode projection reuse, and calibration, without establishing implementation equivalence. We did not recover the LLaVA-1.5-7B operating point reported by Zhuang et al. (2026): at 29.7% nominal savings, our 19-layer configuration reduces GQA from 68.0 to 16.0 and POPE from 76.7 to 56.7, with zero binding. Subsequent comparisons therefore concern acceptance rules within our evaluated reconstruction and should not be interpreted as evidence of failure of the original Q Cache method.

5.2 PROXY SCORES AND RETAINED BINDING

We evaluate 104 adjacent singleton interventions using 25 binding items per configuration and audit-specific baselines. Within-family Spearman correlations are -0.001 for Qwen2-VL, -0.089 for LLaVA-1.5, $+0.463$ for OneVision, and -0.324 for SmolVLM2. Negative correlations agree with the proxy's lower-is-safer ordering, but global rank association does not directly measure unsafe acceptance at an operational threshold. At $\varepsilon = 0.03$, accepted Qwen2-VL pairs average 72.0% binding versus 76.5% for rejected pairs. SmolVLM2's $L_9 \leftarrow L_8$ passes with score 0.02807 despite reducing binding from 80% to 16%; however, the headline thresholds 0.008 and 0.015 reject it. These results identify threshold-specific approval failures, without establishing that rejected candidates are generally safer or that nonlinear prediction is impossible. They also do not directly assess cumulative hidden-state gates. The random-selection control remains exploratory because savings differ and repetitions are limited.

Aggregate benchmark retention. Near-baseline aggregate scores can coexist with severe binding loss. At 8.3% nominal savings, SmolVLM2 under KS-R retains 100% of its unmodified POPE score and 102% of its GQA score, but only 16% of its binding accuracy (Table 2). Qwen2-VL under MC-R likewise exceeds its baseline POPE score while binding falls to zero. These normalized scores illustrate unequal task sensitivity, not statistical equivalence on aggregate benchmarks. Under more aggressive sharing, aggregate scores also deteriorate.

Table 2: Task-score retention relative to unmodified baselines, expressed as normalized percentages. SmolVLM2 MC-R and KS-R use single evaluation draws, not the three-draw means in Table 5. MC-R and KS-R replace keys and values; the Q Cache reconstruction retains target-layer values.

Family	Setting, layers shared	POPE	VQAv2	GQA	TextVQA	V-BIND
Qwen2-VL-2B	MC-R, 4/28	107%	83%	80%	39%	0%
SmolVLM2-2.2B	MC-R, 3/24	100%	91%	90%	78%	59%
SmolVLM2-2.2B	KS-R, 2/24	100%	94%	102%	84%	16%
SmolVLM2-2.2B	Q Cache (recon.), ε=0.03, 16/24	54%	15%	34%	1%	0%

Adjacent approval and installed substitutions. An adjacent approval score can authorize a substitution from an earlier block head. The Qwen2-VL comparison uses different source maps and evaluation baselines, so it does not isolate composition effects (Appendix D.2).

5.3 RECALLGATE AT MATCHED PROXY THRESHOLDS

Table 3 and Figure 2 compare acceptance rules at identical proxy thresholds under the GLA-style reconstruction. Savings differ, so these are operating-point comparisons, not fixed-dose effects. Scores are absolute percentages; ± denotes sample SD across three fixed-configuration evaluation draws, while plain entries are single draws. Binding uses 100 items per draw. Qwen2-VL's healthy mean is 87.0%, versus 84.0% in its separate singleton audit. The zero-share OneVision row repeats the healthy summary.

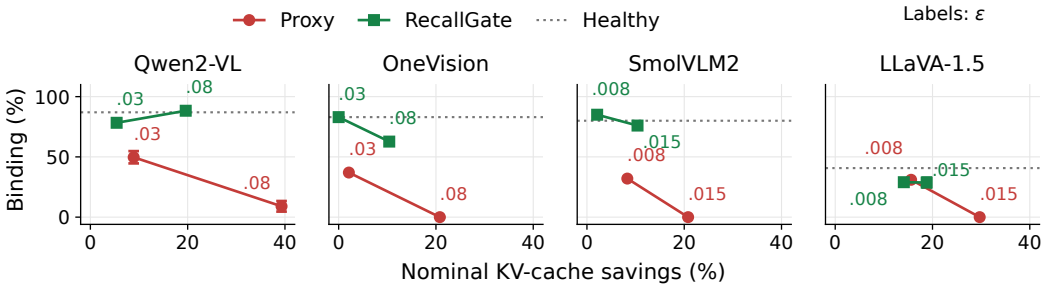


Figure 2: Binding versus nominal cache savings ($n/(2L)$). Points correspond to Table 3; labels denote proxy thresholds, bars show available SDs, and dotted lines mark healthy means. Connecting lines do not imply a frontier.

On Qwen2-VL at $\varepsilon = 0.08$, RecallGate selects 11 shared-key layers, achieving 19.6% nominal savings and $88.3 \pm 2.1\%$ binding, whereas proxy approval selects 22 layers, achieving 39.3% savings and $9.0 \pm 4.4\%$ binding. The improvement accompanies a substantial reduction in the number of shared layers and cannot be attributed solely to the acceptance statistic. At the higher-damage proxy settings, RecallGate yields $62.7 \pm 2.1\%$ binding for OneVision and $76.0 \pm 2.0\%$ for SmolVLM2, each at 10.4% nominal savings, whereas the corresponding proxy-approved configurations have zero binding at higher compression doses. The trade-off varies across models. On LLaVA-1.5 at $\varepsilon = 0.008$, RecallGate achieves 29.0% binding versus 31.0% under proxy approval despite sharing fewer layers. On OneVision at $\varepsilon = 0.03$, it rejects the only proposed share and therefore saves no cache. These cases limit a uniform-benefit interpretation.

5.4 RECALLGATE AT MATCHED NOMINAL CACHE SAVINGS

Table 4 compares proxy approval and RecallGate under wholesale KV replacement at matched nominal savings. Within each model and arm, the comparisons use the reported common candidate space, operation and maximum audition count, with shared-target counts matched. Savings are n/L. MC-R denotes the adapted similarity-ranked arm, whose recorded configurations include nonadjacent pairs; it is not a complete MiniCache reproduction. RecallGate has higher observed binding in four of five matched-savings comparisons of fixed configurations. Selection-level variability is reported separately in Table 7. SmolVLM2 and OneVision provide the most tightly controlled comparisons: both acceptance rules use the same proxy threshold, candidate principle, operation, and shared-target count. On SmolVLM2 KS-R, binding rises from $13.7 \pm 2.1\%$ to $84.0 \pm 1.7\%$ at 8.3% nominal savings, a 70.3 percentage-point gain. On OneVision KS-R, it rises from $20.3 \pm 6.1\%$ to $40.7 \pm 3.2\%$ at 25.0% savings. Because the compared configurations have the same shared-target count, these observed binding differences cannot be attributed to fewer shared targets. Qwen2-VL also shows higher observed binding at matched savings, but its differing proxy thresholds confound attribution to capability approval. LLaVA-1.5 selects identical recorded pairs and shows no observed difference. These comparisons concern fixed selected configurations. A SmolVLM2 replay reproduces all six historical compressed-condition aggregate scores across three evaluation offsets. Substring and unique-color scoring agree on all 900 replay responses, including healthy evaluations. This verifies these aggregate results, not every historical implementation.

Table 5 reports additional MC-R and KS-R configurations, including proxy thresholds, achieved cache savings, binding accuracy, and GQA and POPE scores. These results show how the observed trade-offs vary across models and acceptance rules; not all rows are matched on cache savings.

Table 3: Proxy approval versus RecallGate under the Q Cache GLA-style reconstruction at matched thresholds. Gray shading and bold text identify RecallGate, not statistical significance or superiority.

Method	ε	Lazy layers	Cache saved↑	Binding↑	GQA↑	POPE↑
Qwen2-VL-2B	–	0	0.0%	87.0 ± 3.5	69.0	90.0
Q Cache (recon.)	0.03	5	8.9%	49.7 ± 5.1	64.0	80.0
RecallGate (Ours)	**0.03**	**3**	**5.4%**	$\mathbf{78.3 \pm 3.2}$	**66.0**	**90.0**
Q Cache (recon.)	0.08	22	39.3%	9.0 ± 4.4	11.0	40.0
RecallGate (Ours)	**0.08**	**11**	**19.6%**	$\mathbf{88.3 \pm 2.1}$	**62.0**	**90.0**
LLaVA-OneVision-0.5B	–	0	0.0%	83.0 ± 2.0	72.0	86.7
Q Cache (recon.)	0.03	1	2.1%	37.0	64.0	86.7
RecallGate (Ours)	**0.03**	**0**	**0.0%**	$\mathbf{83.0 \pm 2.0}$	**72.0**	**86.7**
Q Cache (recon.)	0.08	10	20.8%	0.0 ± 0.0	20.0	60.0
RecallGate (Ours)	**0.08**	**5**	**10.4%**	$\mathbf{62.7 \pm 2.1}$	**46.0**	**76.7**
SmolVLM2-2.2B	–	0	0.0%	80.0 ± 1.0	59.0	80.0
Q Cache (recon.)	0.008	4	8.3%	32.0	46.0	83.3
RecallGate (Ours)	**0.008**	**1**	**2.1%**	**85.0**	**55.0**	**83.3**
Q Cache (recon.)	0.015	10	20.8%	0.0 ± 0.0	10.0	36.7
RecallGate (Ours)	**0.015**	**5**	**10.4%**	$\mathbf{76.0 \pm 2.0}$	**49.0**	**83.3**
LLaVA-1.5-7B	–	0	0.0%	40.7 ± 1.2	68.0	76.7
Q Cache (recon.)	0.008	10	15.6%	31.0	59.0	73.3
RecallGate (Ours)	**0.008**	**9**	**14.1%**	**29.0**	**61.0**	**73.3**
Q Cache (recon.)	0.015	19	29.7%	0.0 ± 0.0	16.0	56.7
RecallGate (Ours)	**0.015**	**12**	**18.8%**	$\mathbf{28.7 \pm 4.1}$	**50.0**	**76.7**

Table 4: Binding accuracy (%) at matched nominal cache savings under wholesale KV replacement. Means ± sample SD use three distinct evaluation-seed offsets for fixed configurations. Qwen2-VL uses proxy/RecallGate thresholds 0.015/0.03, so its observed differences do not isolate the effect of capability-based acceptance; all other comparisons use 0.03/0.03. Differences are computed from unrounded means and reported in percentage points.

Family	Arm	Layers	Saved	Healthy	Proxy	RecallGate	Δ Binding
Qwen2-VL	MC-R	3/28	10.7%	87.0 ± 3.5	0.0 ± 0.0	$\mathbf{75.3 \pm 4.0}$	**+75.3**
Qwen2-VL	KS-R	3/28	10.7%	87.0 ± 3.5	22.3 ± 5.1	$\mathbf{58.7 \pm 2.5}$	**+36.3**
SmolVLM2	KS-R	2/24	8.3%	80.0	13.7 ± 2.1	$\mathbf{84.0 \pm 1.7}$	**+70.3**
OneVision	KS-R	6/24	25.0%	83.0	20.3 ± 6.1	$\mathbf{40.7 \pm 3.2}$	**+20.3**
LLaVA-1.5	KS-R	4/32	12.5%	40.7	37.7 ± 2.9	$\mathbf{37.7 \pm 2.9}$	0.0

5.5 Controlled selection and layout transfer

At the common threshold $\varepsilon = 0.03$, proxy approval and all five new SmolVLM2 KS-R selections reach two shares, saving 8.3%. Evaluation uses identical items across conditions: 100 original-probe examples and 300 diagnostic-layout examples, with 100 per layout. Layout decoding is greedy with at most 24 output tokens.

All five RecallGate selections achieve higher observed accuracy than proxy approval on both evaluations, but only three exceed healthy layout accuracy. Across five gating-seed runs, binding is $86.6 \pm 2.2\%$ on original probes and $77.1 \pm 17.2\%$ on layouts (mean ± sample SD; repeated configurations retain their seed multiplicity). These previously evaluated layouts test new selections; they are not a fresh final test set.

Table 6: Controlled SmolVLM2 binding (%): 100 original-probe and 300 layout items. All compressed configurations save 8.3%. A/B occur in 3/2 of five gating-seed runs.

Condition	Original	Layouts
Healthy	80.0	81.7
Proxy	13.0	20.0
RecallGate A	**85.0**	**89.7**
RecallGate B	**89.0**	**58.3**

Table 5: Adapted selection rules under wholesale KV replacement. Scores are absolute percentages. Intervention binding summaries report mean ± sample SD across three distinct evaluation-seed offsets for fixed configurations, counting each offset once. Gray shading and bold identify RecallGate, not superiority or statistical significance.

Family	Selection principle, gate	ε	Saved	Binding↑	GQA↑	POPE↑
Qwen2-VL-2B	*unmodified*	–	–	87.0 ± 3.5	69.0	90.0
	MC-R, proxy rule	0.03	14.3%	0.0 ± 0.0	55.0	96.7
	MC-R, proxy rule	0.015	10.7%	0.0 ± 0.0	60.0	80.0
	MC-R, RecallGate	**0.03**	**10.7%**	$\mathbf{75.3 \pm 4.0}$	**71.0**	**83.3**
	KS-R, proxy rule	0.03	21.4%	32.3 ± 3.5	50.0	66.7
	KS-R, proxy rule	0.015	10.7%	22.3 ± 5.1	58.0	70.0
	KS-R, RecallGate	**0.03**	**10.7%**	$\mathbf{58.7 \pm 2.5}$	**66.0**	**96.7**
LLaVA-OneVision-0.5B	*unmodified*	–	–	83.0	72.0	86.7
	MC-R, proxy rule	0.03	25.0%	0.0 ± 0.0	23.0	46.7
	MC-R, RecallGate	**0.03**	**16.7%**	$\mathbf{69.0 \pm 2.6}$	**57.0**	**80.0**
	KS-R, proxy rule	0.03	25.0%	20.3 ± 6.1	34.0	46.7
	KS-R, RecallGate	**0.03**	**25.0%**	$\mathbf{40.7 \pm 3.2}$	**43.0**	**63.3**
SmolVLM2-2.2B	*unmodified*	–	–	80.0	59.0	80.0
	MC-R, proxy rule	0.03	12.5%	51.0 ± 5.3	53.0	80.0
	MC-R, RecallGate	**0.03**	**8.3%**	$\mathbf{91.3 \pm 2.1}$	**54.0**	**70.0**
	KS-R, proxy rule	0.03	8.3%	13.7 ± 2.1	60.0	80.0
	KS-R, RecallGate	**0.03**	**8.3%**	$\mathbf{84.0 \pm 1.7}$	**59.0**	**73.3**
LLaVA-1.5-7B	*unmodified*	–	–	40.7	68.0	76.7
	MC-R, proxy rule	0.03	9.4%	35.7 ± 2.3	63.0	83.3
	MC-R, RecallGate	**0.03**	**6.2%**	$\mathbf{34.3 \pm 1.5}$	**66.0**	**83.3**
	KS-R, proxy rule	0.03	12.5%	37.7 ± 2.9	70.0	80.0
	KS-R, RecallGate	**0.03**	**12.5%**	$\mathbf{37.7 \pm 2.9}$	**70.0**	**80.0**

Configuration A uses pairs $(0, 5), (0, 20)$ and is selected by seeds 7100, 7101 and 7104. Configuration B uses $(3, 14), (0, 4)$ and is selected by seeds 7102 and 7103. Despite higher original-probe accuracy, B transfers worse at identical savings. Improvement over proxy approval therefore does not ensure healthy-model capability retention. Proxy approval uses 29 auditions; A/B use 9/32 auditions and 4/3 capability checks, respectively (Table 12).

Fresh-layout validation. On 300 fresh examples across three additional synthetic layouts, frozen healthy, proxy, RecallGate A and B configurations achieve 80.3%, 16.3%, 88.0% and 63.3% accuracy, respectively. Observed accuracy follows A > healthy > B > proxy on each of the three fresh layouts. Thus, the observed configuration-dependent transfer pattern persists on this fresh synthetic evaluation (Appendix A.1).

Scoring sensitivity. Substring and unique-color scoring give identical correctness decisions on all 2,700 responses from nine earlier layout-evaluation files and 900 responses from the SmolVLM2 headline replay. This agreement does not validate scoring for other historical evaluations.

Singleton diagnostics. For configuration B, substitutions $(3, 14)$ and $(0, 4)$ achieve 91.7% and 86.7% layout binding individually, but 58.3% jointly, versus 81.7% healthy. The accuracy-scale interaction $I = C_{AB} - C_A - C_B + C_\emptyset$ is -38.3 percentage points (exploratory paired bootstrap 95% CI $[-45.7, -31.3]$). Although RecallGate tests cumulative configurations, its selection battery accepts B despite its layout-level loss. This identifies a negative interaction on these items, not its internal mechanism or prevalence.

5.6 Sensitivity to battery choice and gating samples

Battery choice and acceptance floor. Battery-size and acceptance-floor sweeps show sensitivity, without establishing an optimum (Appendix E.1). Replacing binding with POPE admits 14 Qwen2-VL shares at $\varepsilon = 0.08$, yielding 13% binding versus 12% for that ablation's proxy condition, with

76.7% POPE. This small observed binding difference highlights probe sensitivity; it does not establish equivalence.

Variation across gating samples. We repeat RecallGate selection using five gating seeds (7000–7004), with 100 binding evaluation items per selected configuration. Table 7 distinguishes Q Cache-style Q/K reuse from MC-R and KS-R wholesale KV replacement. Earlier 25-item evaluations are not pooled.

Table 7: Binding accuracy (mean ± sample SD) across five gating-seed selection runs. Ranges span these runs; cache savings vary across runs.

Model	Sharing strategy	ε	Cache saved (%)	Binding (%)	Binding range (%)
Q/K reuse: nominal savings $n/(2L)$					
Qwen2-VL	Q Cache-style	0.08	**14.3–19.6**	$\mathbf{79.6 \pm 11.9}$	**65–92**
SmolVLM2	Q Cache-style	0.015	**8.3–10.4**	$\mathbf{55.6 \pm 25.2}$	**28–74**
Wholesale KV replacement: nominal savings n/L					
Qwen2-VL	MC-R	0.03	**7.1–10.7**	$\mathbf{42.2 \pm 26.8}$	**18–80**
Qwen2-VL	KS-R	0.03	**7.1–14.3**	$\mathbf{64.0 \pm 15.6}$	**43–84**
SmolVLM2	MC-R	0.03	**8.3–12.5**	$\mathbf{83.0 \pm 20.1}$	**47–92**
SmolVLM2	KS-R	0.03	**8.3–16.7**	$\mathbf{78.8 \pm 18.4}$	**46–89**

These earlier runs vary both selected pairs and savings: Qwen2-VL MC-R spans 18–80% binding, while SmolVLM2 KS-R spans 46–89%. Unlike fixed-configuration evaluation draws in Tables 3 and 4, they describe selection outcomes but do not isolate a fixed-budget effect. Section 5.5 fixes the sharing budget for five new SmolVLM2 selections and still finds configuration-dependent transfer.

Limitations. Results concern reconstructed or adapted implementations and synthetic binding probes. The SmolVLM2 replay reproduces specific headline scores; code-version correspondence remains unresolved for other historical runs. Five controlled selection runs yield only two distinct configurations on one model, limiting inference about selection-level variability. Fresh-layout validation remains synthetic and does not establish broad reliability. Reported savings concern nominal KV-cache storage rather than measured peak GPU memory, end-to-end latency, or serving throughput.

6 Conclusion

The audited proxies inconsistently rank binding retention, while sampled aggregate scores can obscure binding loss. RecallGate has higher observed binding in several matched-savings comparisons, and all five controlled SmolVLM2 selections achieve higher observed accuracy than proxy approval. However, two selection runs, corresponding to the same distinct configuration, fall below healthy layout accuracy at identical savings; singleton diagnostics reveal a negative joint interaction for that configuration on the evaluated layouts. Qwen2-VL threshold confounds and the non-improving LLaVA comparison further limit generalization. These findings support targeted capability screening followed by independent validation, not guaranteed capability preservation.

7 AI Use Disclosure

Generative AI tools were used to assist with software implementation, literature search, reference formatting, figure preparation, and manuscript drafting and editing. They were also used during development to help implement synthetic probe generators, experimental utilities, and acceptance-rule code, and to support discussion and refinement of experimental hypotheses and analysis.

The research questions, experimental decisions, capability definitions, evaluation protocols, and interpretation of the reported findings were determined and reviewed by the authors. All reported

measurements originate from executed experiments, with summary statistics computed from saved experimental records; AI-generated estimates are not presented as experimental results. The authors reviewed and verified AI-assisted material and take responsibility for the methods, results, claims, and final manuscript.

REFERENCES

William Brandon, Mayank Mishra, Aniruddha Nrusimha, Rameswar Panda, and Jonathan Ragan-Kelley. Reducing transformer key-value cache size with cross-layer attention. *Advances in Neural Information Processing Systems*, 37:86927–86957, 2024.

Hong Chen, Kang Chen, Yuxuan Fan, Bo Wang, Yubo Gao, Yuanlin Chu, and Xuming Hu. Seen, said, or forgotten? a causal audit of visual kv memory across dialog turns. *arXiv preprint arXiv:2607.25467*, 2026.

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*, pp. 19–35. Springer, 2024.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *Advances in neural information processing systems*, 28, 2015.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A Roberts. The unreasonable ineffectiveness of the deeper layers. In *International Conference on Learning Representations*, volume 2025, pp. 81906–81920, 2025.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

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

Haoyang Li, Yiming Li, Anxin Tian, Tianhao Tang, Zhanchao Xu, Xuejia Chen, Nicole Hu, Wei Dong, Qing Li, and Lei Chen. A survey on large language model acceleration based on kv cache management. *arXiv preprint arXiv:2412.19442*, 2024b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, 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.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, et al. Scbench: A kv cache-centric analysis of long-context methods. In *International Conference on Learning Representations*, volume 2025, pp. 66063–66093, 2025.

Akide Liu, Jing Liu, Zizheng Pan, Yefei He, Gholamreza Haffari, and Bohan Zhuang. Minicache: Kv cache compression in depth dimension for large language models. *Advances in Neural Information Processing Systems*, 37:139997–140031, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024b.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *arXiv preprint arXiv:1603.05766*, 2016.

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.

Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang, Wenhui Wang, Shuming Ma, Quanlu Zhang, Jianyong Wang, and Furu Wei. You only cache once: Decoder-decoder architectures for language models. *Advances in Neural Information Processing Systems*, 37:7339–7361, 2024.

Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4065–4078, 2024.

Zhongwei Wan, Hui Shen, Xin Wang, Che Liu, Zheda Mai, and Mi Zhang. Meda: Dynamic kv cache allocation for efficient multimodal long-context inference. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2485–2497, 2025.

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

Junbin Xiao, Jiajun Chen, Tianxiang Sun, Xun Yang, and Angela Yao. Mukv: Multi-grained kv cache compression for long streaming video question-answering. *arXiv preprint arXiv:2605.22269*, 2026.

Yifei Yang, Zouying Cao, Qiguang Chen, Libo Qin, Dongjie Yang, Hai Zhao, and Zhi Chen. Kvsharer: Efficient inference via layer-wise dissimilar kv cache sharing. *arXiv preprint arXiv:2410.18517*, 2024.

Jiayi Yuan, Hongyi Liu, Shaochen Zhong, Yu-Neng Chuang, Songchen Li, Guanchu Wang, Duy Le, Hongye Jin, Vipin Chaudhary, Zhaozhuo Xu, et al. Kv cache compression, but what must we give in return? a comprehensive benchmark of long context capable approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4623–4648, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

Jiedong Zhuang, Lu Lu, Ming Dai, Rui Hu, Jian Chen, Qiang Liu, and Haoji Hu. Q cache: Visual attention is valuable in less than half of decode layers for multimodal large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14031–14039, 2026.

SUPPLEMENTARY MATERIAL

A EXPERIMENTAL DESIGN AND IMPLEMENTATION

A.1 MODELS, EVALUATION SETS AND CONFIGURATION

The study evaluates five checkpoints from four model families. Sharing changes the language tower; the vision encoder and connector remain frozen. GPU experiments used V100 hardware. The LLaVA-1.5-13B capability evaluations used CPU-only fp32 execution on high-memory nodes, with reported runtimes of 4.5–10.7 hours per configuration. The 13B binding evaluation uses the same nominal context-length setting of 3000 as LLaVA-1.5-7B; this setting does not specify the measured total input-token count. These evaluations are not serving-system throughput measurements. Tables 8 and 9 describe the evaluation suite and configuration dimensions. Available records identify settings, seeds and selected pairs, but do not establish code-version correspondence for every historical result.

Table 8: Evaluation suite and the capability targeted by each benchmark. V-BIND-NIAH is the primary probe for object–attribute binding after sharing.

Benchmark	Setting	Capability evaluated
V-BIND-NIAH	Synthetic	**Object–attribute binding** in long visual contexts
X-FRAME-ID	Synthetic	Cross-frame object identity and correspondence
GQA / VQAv2 / TextVQA	Real	Aggregate VQA, including reasoning and localized text recognition
POPE	Real	Object presence and hallucination

GQA (Hudson & Manning, 2019); VQAv2 (Goyal et al., 2017); TextVQA (Singh et al., 2019); POPE (Li et al., 2023).

Synthetic image generation. Each image is a 448×448 RGB canvas containing one triangle, one square and one pentagon. Three distinct colors are sampled without replacement from {red, blue, green, yellow, purple, orange} and assigned to the shapes. Shapes occupy three randomly selected positions on a 3×3 grid; the remaining six positions contain gray distractor dots. The queried shape is sampled uniformly. Examples use a seeded pseudorandom generator; finite samples are not explicitly balanced by color or shape.

Original binding-probe protocol. V-BIND-NIAH combines synthetic shape–color images with repeated neutral weather text and a question about a target object's color. All images precede the filler and question in the processor's chat template. For m images, the queried image has zero-based index $\mathrm{round}(d(m-1))$, where d is the configured depth; depth therefore controls image ordinal position, not token position. The filler uses approximately four characters per nominal token. Reported context settings are generation parameters, not measured total input-token counts.

Headline operating-point evaluations use 100 examples per draw, overriding the 25-example configuration default. The standard probe geometry uses six images for Qwen2-VL, two for OneVision and SmolVLM2, and one for LLaVA-1.5. Nominal context lengths are capped at 3000 for OneVision, SmolVLM2 and LLaVA-1.5. Headline depth-only evaluations use $d = 0.1$ at the shortest configured context length.

The original evaluator uses greedy decoding with at most six output tokens. A response is counted correct when its lowercased text contains the target color as a substring; additional color words are not rejected. This permissive rule can accept ambiguous responses containing multiple colors. The new-layout evaluation instead requires exactly one recognized color matching the target.

Aggregate evaluation in the inspected harness. The controlled KS-R configuration samples 100 questions from GQA's balanced validation split using seed 4000. Responses use greedy decoding with

Table 9: Reported experiment groups. This summary is not a run manifest: checkpoint revisions, exact proxy implementations, item IDs and selected pairs require the execution archive.

Experiment	Reported configuration
Singleton audit	Adjacent pairs: $27, 31, 23, 23$ across four families.
Q Cache comparisons	Thresholds and selected counts in Table 3; key-only saving $n/(2L)$.
Matched replacement	Layer counts and thresholds in Table 4; saving n/L.
RecallGate sensitivity	$\alpha \in \{0.80, 0.85, 0.90, 0.95\}$; battery sizes $8, 16, 32, 64$.
Random control	Two evaluation summaries over random seeds; doses differ.
13B scale check	Single evaluation draw per cell; operations distinguished.

at most six output tokens and are scored by lowercased target-answer substring matching. A loader audit of the cached 100-item GQA sample (seed 4000) resolves all images without decoding errors; the evaluator's denominator is therefore 100 for this sample. This check does not independently establish dataset identity across historical runs. POPE evaluation samples 30 rows without replacement from the `lmms-lab/POPE` test split using seed 3000, without explicit stratification by question category or answer. Generation is greedy with at most three output tokens. The parser assigns "yes" when that substring appears and "no" otherwise. Calibration images are sampled through the same dataset loader using a separate configured seed; different seeds do not guarantee disjoint images.

Verification of saved SmolVLM2 records. Recorded proxy pairs are $(5, 9), (0, 4)$; RecallGate pairs are $(0, 5), (0, 20)$. Both save 8.3% nominal cache storage. Replaying evaluation offsets 0, 1000 and 2000 exactly reproduces all six historical compressed-condition aggregate binding scores. Healthy scores are 80%, 81% and 79%; proxy scores are 13%, 16% and 12%; RecallGate scores are 85%, 85% and 82%. Their means and sample SDs are $80.0 \pm 1.0\%$, $13.7 \pm 2.1\%$ and $84.0 \pm 1.7\%$, respectively. Each offset contains 100 items and is counted once. Duplicate records of an offset are not independent evaluation draws. Substring and unique-color scoring agree on all 900 replay responses. All six historical compressed-condition result files match the source hashes recorded in the replay manifests. The replay verifies their aggregate scores, not every historical code revision or historical per-item agreement.

Anonymized archive and verification scope. The accompanying anonymized archive includes source snapshots, configuration records, saved predictions where available, dataset images and integrity manifests for the documented SmolVLM2 experiments. Packaged core imports and portable replay listing and transfer preparation checks passed; end-to-end GPU reruns of the portable package remain untested. These records do not establish provenance for all historical experiments or recover missing original-probe per-item predictions.

Layout decoding and scoring. Evaluation uses greedy decoding. A response is correct only when its recognized color words contain exactly one distinct color matching the target. Raw predictions are retained. For the earlier healthy, proxy and RecallGate A configurations, the six-token budget yields 81.7%, 17.0% and 89.7% accuracy, respectively. After observing truncation, reevaluation on identical items with a 24-token output budget raises proxy accuracy to 20.0%; the other accuracies remain unchanged. Proxy responses without a unique recognized color decrease from 34 to six; healthy and RecallGate A have zero at either budget. All new controlled-selection layout evaluations use the 24-token budget.

Paired uncertainty. For the earlier healthy, proxy and RecallGate A configurations, percentile intervals use 20,000 paired bootstrap resamples (seed `20260920`), sampling items within layouts and preserving 100 items per layout. The observed difference relative to proxy approval is 69.67 percentage points (95% CI [64.33, 75.00]): 212 items are correct only under RecallGate A and three only under proxy approval. The observed difference relative to healthy accuracy is 8.00 points ($95\% CI [5.00, 11.33]$): 25 items are correct only under RecallGate A and one only under the healthy model. These intervals describe item-level uncertainty for fixed configurations, not variation across gating seeds, uncertainty in the five-selection mean, or general transfer beyond the evaluated layouts.

Earlier selections with variable savings. Gating seeds 7000–7004 produce three distinct SmolVLM2 KS-R configurations, evaluated on identical 300-item layout sets with a 24-token output budget. Seeds 7000/7004 select configuration A, achieving 89.7% binding at 8.3% savings. Seeds 7001/7003 select configuration B, achieving 58.3% at the same savings. Seed 7002 achieves 39.3% at 16.7% savings; its greater compression prevents attributing the difference solely to selection variability. Identical configurations share one transfer evaluation. These earlier runs are separate from the controlled two-share experiment using seeds 7100–7104.

Table 10: Earlier SmolVLM2 layout accuracy (%): 100 items per layout and a 24-token output budget. Proxy and RecallGate A both save 8.3%.

Layout	Healthy	Proxy	RecallGate A
Horizontal	83.0	16.0	**90.0**
Vertical	84.0	26.0	**87.0**
Circular	78.0	18.0	**92.0**
Overall	81.7	20.0	**89.7**

Fresh-layout validation. We evaluate frozen healthy, proxy and RecallGate A/B configurations on 300 fresh examples: 100 each with diagonal-down, diagonal-up and asymmetric layouts. All conditions receive identical items, with greedy decoding, a 24-token output budget and unique-color scoring. All compressed configurations retain 8.3% nominal cache savings. Each condition produces 300 responses, with zero unparsed outputs. RecallGate A exceeds proxy accuracy by 71.7 percentage points $(95\%CI[66.3, 77.0])$; B exceeds proxy by 47.0 points $(95\%CI[40.7, 53.3])$. Relative to healthy accuracy, A differs by $+7.7$ points (95% CI [4.3, 11.3]), whereas B differs by -17.0 points (95% CI [-23.7, -10.7]). A exceeds B by 24.7 points $(95\%CI[18.7, 31.0])$. Intervals use 20,000 paired percentile bootstrap resamples within layouts (seed `20260925`). They are unadjusted for multiple comparisons and describe item-level uncertainty for frozen configurations, not variability across selection runs or layout families.

Table 11: Fresh-layout binding accuracy (%): 100 items per layout, using frozen configurations. Shading highlights the overall results; bold marks the highest accuracy in each row.

Layout	Healthy	Proxy	RecallGate A	RecallGate B
Diagonal-down	84.0	15.0	**96.0**	61.0
Diagonal-up	78.0	20.0	**85.0**	56.0
Asymmetric	79.0	14.0	**83.0**	73.0
Overall	80.3	16.3	**88.0**	63.3

A.2 Candidate proposal, approval and comparison controls

Candidate proposal and approval are separate choices. The Q Cache GLA-style reconstruction uses sequential adjacent-layer formation with a divergence threshold. The MC-R and KS-R arms port selection principles into the common harness; they do not reproduce the published methods' full implementations. RecallGate evaluates a proposed addition on the cumulative accepted configuration. The proxy-free ablation omits the preliminary divergence test. Table 12 reports the controlled SmolVLM2 KS-R comparison at $\varepsilon = 0.03$. The five RecallGate gating seeds produce two distinct configurations, all reaching the two-share budget. Original-probe accuracy uses 100 items; layout accuracy uses 300 identical items across conditions. Audition and capability-check counts do not imply matched computational cost.

For gating seeds 7100–7104, recorded runner hashes match the inspected implementation. Each entry enables capability gating without overriding its default relative floor, $\alpha = 0.90$. The legacy configuration value `r_min=0.875` is unused by this acceptance branch. For all seven controlled conditions, original source snapshots matched recorded hashes before anonymization. Native results agree with summaries on selected pairs and shared-layer counts; compressed runs retain 120 candidate-decision records. Archive manifests distinguish original hashes from anonymized-copy hashes. These

checks establish record consistency, not recovery of original-probe per-item answers or verification of other historical runs.

Table 12: Controlled SmolVLM2 KS-R results. All compressed configurations reach two shares, saving 8.3%. Accuracy is reported in percent. Checks count capability-battery evaluations. Seeds 7100, 7101, and 7104 select configuration A; seeds 7102 and 7103 select configuration B.

Condition	Source–target pairs	Original	Layouts	Auditions	Checks
Healthy	–	80.0	81.7	0	0
Proxy	$(5, 9), (0, 4)$	13.0	20.0	29	0
Seed 7100 (A)	$(0, 5), (0, 20)$	85.0	**89.7**	9	4
Seed 7101 (A)	$(0, 5), (0, 20)$	85.0	**89.7**	9	4
Seed 7102 (B)	$(3, 14), (0, 4)$	**89.0**	58.3	32	3
Seed 7103 (B)	$(3, 14), (0, 4)$	**89.0**	58.3	32	3
Seed 7104 (A)	$(0, 5), (0, 20)$	85.0	**89.7**	9	4

The comparison controls are: **I1**, a matched audition cap within ranked comparisons; **I2**, matched dose where explicitly reported; **I3**, common calibration data and modality ratio except when calibration is the manipulated variable; **I4**, first-eligible-pair truncation for no-filter controls; **I5**, deterministic evaluation and seeded generation; and **I6**, disjoint item seeds for calibration, the mini-battery and headline evaluation. Seed disjointness does not make the synthetic template families independent. A matched cap is a common maximum, not evidence that every run evaluates the same number of candidates.

A.3 Cumulative audition and installed operation

Algorithm 1 describes the ranked greedy audition with the source–target exclusion rule. Q Cache's sequential Lazy-Block formation is a distinct proposal procedure; it is not represented by those exclusions. The algorithm below is a reference description of score modes, not a verified trace of every historical run. The unconditional branch accepts eligible pairs without either proxy or capability approval. Proxy-free RecallGate is distinct: it disables only the proxy pre-filter while retaining capability approval. That implementation provenance must be checked before claiming a faithful KVSharer comparison.

For original KVSharer's strict similarity test $c > \mathcal{T}$, the divergence form is $1 - c < 1 - \mathcal{T}$; thus $\varepsilon = 1 - \mathcal{T}$. This algebra does not establish that the ported table thresholds used that transformation. Calibration and candidate logs are needed to verify it.

The implementation uses projection hooks during prefill and subsequent decoding. At each step, the designated source supplies the projection used by the target. Q Cache's audited operation retains target-layer values; the wholesale KV comparisons replace keys and values. Computing a target projection and subsequently overwriting it reproduces an intervention but does not establish a reduction in inference FLOPs or peak memory. The unsuccessful visual-only splicing reconstruction is excluded from performance comparisons and recorded in the reconstruction ledger.

B Cache accounting and offline cost

B.1 Architecture determines absolute memory value

For L decoder layers, n_{kv} KV heads, head dimension d, precision p bytes per scalar and context length S, the unmodified cache occupies $2Ln_{\mathrm{kv}}dpS$ bytes per sequence. Under the audited operations, n shared target layers give

$$\mathrm{KV_bytes}(\sigma) = n_{\mathrm{kv}}dpS \begin{cases} 2L - n, & \text{key-only sharing,} \\ 2(L - n), & \text{wholesale KV sharing.} \end{cases} \tag{6}$$

The corresponding saved fractions are $n/(2L)$ and n/L. These formulas assume the relevant source tensors remain stored and each shared target removes the indicated distinct storage. Head count and precision determine the absolute bytes saved; the fraction alone does not.

Algorithm 1 Ranked greedy audition with optional cumulative capability approval

Require: Ranked candidates R; layer budget B; audition cap A; proxy score mode and threshold ε; approval mode; mini-battery $\mathcal{D}_{\text{mini}}$ and floor $r_{\min}$ when required
Ensure: Accepted configuration σ, decision log and counters

$\sigma \leftarrow \varnothing$; target set $T \leftarrow \varnothing$; source set $S \leftarrow \varnothing$; $a \leftarrow 0$; $a_{\text{rec}} \leftarrow 0$
for (m, ℓ) in R **do**
  if $|\sigma| = B$ **then**
    break
  end if
  if $\ell \in T \cup S$ **or** $m \in T$ **then**
    continue
  end if
  if approval mode is unconditional **then**
    accept (m, ℓ) and update σ, T, S; **continue**
  end if
  if $a = A$ **then**
    break
  end if
  $a \leftarrow a + 1$
  $pass \leftarrow \text{true}$
  if proxy pre-filter is enabled **then**
    if score mode is static **then**
      $s_{\text{cand}} \leftarrow S(m, \ell; \mathcal{D}_{\text{cal}})$
    else
      temporarily install $\sigma' = \sigma \cup \{(m, \ell)\}$
      $s_{\text{cand}} \leftarrow 1 - \text{CosSim}(M_{\sigma'}, M_{\varnothing}; \mathcal{D}_{\text{cal}})$
      restore σ
    end if
    $pass \leftarrow (s_{\text{cand}} < \varepsilon)$
  end if
  if $pass$ **and** capability approval is enabled **then**
    temporarily install $\sigma' = \sigma \cup \{(m, \ell)\}$
    $a_{\text{rec}} \leftarrow a_{\text{rec}} + 1$
    $pass \leftarrow [R_{\text{mini}}(\sigma') \geq r_{\min}]$
    restore σ
  end if
  log candidate, available scores and decision
  if $pass$ **then**
    $\sigma \leftarrow \sigma \cup \{(m, \ell)\}$; $T \leftarrow T \cup \{\ell\}$; $S \leftarrow S \cup \{m\}$
  end if
end for
install the accepted configuration σ
evaluate M_σ; record selected pairs and achieved savings
return σ, decision log, a, a_{rec}

Table 13: Architecture and cache accounting at fp16 with a 4096-token context. KB and MB are decimal units; "n/r" means not reported. Bold emphasizes memory values. The final row is an architectural reference, not an evaluated checkpoint.

Model	Attention	KV/heads	L	d	KB/token	12.5% saves (MB)
LLaVA-1.5-7B	MHA	32/32	32	n/r	**524.3**	**268.4**
Qwen2-VL-2B	GQA	2/12	28	128	**28.7**	**14.7**
OneVision-0.5B	GQA	2/14	24	n/r	**12.3**	**6.3**
SmolVLM2-2.2B	MHA	32/32	24	64	**196.6**	**100.7**
LLaVA-1.5-13B	MHA	40/40	40	n/r	**819.2**	**419**
LLaVA-1.6-Mistral-7B	GQA	8/32	32	n/r	**131**	**67**

For example, 12.5% saves 268.4 MB on LLaVA-1.5-7B and 6.3 MB on OneVision at this context length, approximately a 43-fold difference using the rounded entries. This comparison concerns

KV storage, not total process memory. Different context lengths, attention designs and allocation strategies can change the practical benefit.

B.2 Measured storage reductions

After prefill, target tensors are aliased to their designated sources and distinct underlying storages are counted once. Borrow chains resolve to the block head. Table 14 checks the architectural saving formula at specified shared-layer counts; the wholesale-KV columns are byte-accounting checks at those counts, not a separate claim about the accepted MiniCache or KVSharer configurations.

Table 14: Measured and predicted fractions of KV-cache storage saved at fixed shared-target counts. Here, L is the number of decoder layers and n is the number of shared targets. Bold identifies measurements, which agree with predictions to the displayed precision. These are storage-accounting checks, not selected operating-point comparisons.

			Key-only sharing		Wholesale KV sharing	
Family	L	n	Measured	$n/(2L)$	Measured	n/L
Qwen2-VL-2B	28	11	**0.1964**	0.1964	**0.3929**	0.3929
SmolVLM2-2.2B	24	5	**0.1042**	0.1042	**0.2083**	0.2083
LLaVA-OneVision-0.5B	24	5	**0.1042**	0.1042	**0.2083**	0.2083
LLaVA-1.5-7B	32	12	**0.1875**	0.1875	**0.3750**	0.3750

A 385-token forward pass allocates 22,077,440 bytes on Qwen2-VL-2B and 151,388,160 on SmolVLM2-2.2B in fp32. These equal 57,344 and 393,216 bytes per token, respectively. The fp16 equivalents are 28.7 and 196.6 KB per token. Repeating the accounting by distinct tensor object instead of distinct storage gives the same totals in the four reported families, a check against misleading results from fused allocations. These cache-object measurements do not establish serving throughput or peak-memory savings.

B.3 Offline audition cost

RecallGate evaluates a mini-battery after a candidate passes the proxy pre-filter. Selection occurs offline; no approval test runs during serving, although selected configurations may have different serving costs. Table 15 reports proxy-selected layers (n_ε), RecallGate-selected layers (n_{RG}), and binding checks (a_{rec}). The observed equality $a_{\mathrm{rec}} = n_\varepsilon$ is specific to those runs and does not imply identical candidates. Table 16 separately reports the inspected Qwen MC-R traces. Without the pre-filter, adjacent-pair runs evaluate 27 candidates on Qwen2-VL, 23 on SmolVLM2 and OneVision, and 31 on LLaVA-1.5-7B. Ranked comparisons use a 150-audition cap. Each binding check evaluates multiple examples; counts alone do not measure runtime or token cost. Token-equivalent cost fields were not populated.

Table 15: Offline selection counts under the Q Cache-style reconstruction.

Model	ε	n_ε	n_{RG}	a_{rec}
Qwen2-VL	0.03	5	3	5
	0.08	22	11	22
SmolVLM2	0.008	4	1	4
	0.015	10	5	10
OneVision	0.03	1	0	1
	0.08	10	5	10
LLaVA-7B	0.008	10	9	10
	0.015	19	12	19
LLaVA-13B	0.008	10	8	10
	0.015	30	16	30

Table 16: Qwen MC-R trace counts. Both runs reached 150 auditions under a seven-layer maximum, accepting three layers (10.7% nominal savings).

Rule	ε	Proxy passes	Binding checks	Accepted layers
Proxy	0.015	3	0	3
RecallGate	0.030	17	17	3

C RECONSTRUCTION AND THRESHOLD PORTABILITY

C.1 COMPONENT VALIDATION AND UNRECOVERED OPERATING POINT

Table 17 records component-level reconstruction checks supporting the scope stated in the main paper. Passing these checks does not establish equivalence to the original implementation or its evaluation protocol.

Table 17: Reconstruction ledger. Checks validate aspects of the evaluated reconstruction, not full equivalence to the original method.

Component	Validation evidence	Verdict
Criterion	Static inter-layer last-row JS divergence; acceptance counts checked within the reconstruction	Check passed
Formation	Sequential scan checked at $\varepsilon \in [0.008, 0.03]$, with a maximum borrowing distance of six layers	Check passed
Operation	GLA-style reuse is applied during prefill and decoding. One share preserves aggregate scores on 12 synthetic shape-listing questions and 12 POPE items (0.583 and 0.833, respectively), although generated outputs differ	Check passed
Calibration	COCO-photo and painted-image calibration scores compared: $L_1 \leftarrow L_0$ gives 0.0326 versus 0.0320; the remaining compared scores span 0.33–0.55 versus 0.30–0.54	Compared
Cache consistency	For one image and $L_4 \leftarrow L_3$, cached and uncached processing agree on the top-ranked next token at all seven tested prefixes in each condition. Maximum absolute logit differences are 0.0156, 0.0156, and 0.0234 for unmodified, GLA-style, and visual-only processing, respectively, in fp16. All logits are finite	Agreement on tested prefixes
Projection reference	For $L_4 \leftarrow L_3$, visual-only hooks match a separate reference that recomputes source Q/K projections. Across three images and three prefixes per image, maximum next-token logit differences and all recorded projection-check differences are zero. The reference uses the same model attention and positional machinery	Check passed on nine inputs
Operating point	The reported high-savings, low-loss operating point was not recovered in the evaluated reconstruction sweep	Not recovered
Visual-only behavior	Inspected masking and identity controls pass. On 12 images, direct shape–color accuracy is $4/12$, versus $9/12$ unmodified and $7/12$ with GLA-style reuse ($L_4 \leftarrow L_3$). All 12 visual-only shape-listing answers are "Triangle" followed by EOS despite a 48-token allowance. Both cached and uncached processing select EOS after "Triangle" in the cache-consistency example	Diagnostic completed

Projection-level visual-only splicing reduces synthetic shape-listing accuracy from $7/12$ (0.583) to $2/12$ (0.167) after one share. Masking and identity controls pass, but equivalence to the original VLA method remains unverified; this variant is excluded from comparative performance tables.

C.2 THRESHOLD TRANSFER ACROSS FAMILIES

At $\varepsilon = 0.015$, the formation procedure selects $19/32$ lazy layers on LLaVA-1.5, $10/24$ on SmolVLM2, and none on Qwen2-VL or OneVision. The same numerical threshold therefore does not specify the same compression dose across checkpoints. Reported threshold settings and accepted layer sets are necessary to interpret comparisons.

C.3 ADJACENT COMPARISONS AND BLOCK-HEAD SUBSTITUTION

A sequential formation rule may compare each layer with its immediate predecessor while the installed operation borrows from the block head. Thus the pair tested for block membership need not be the source–target pair ultimately substituted. This procedural distinction motivates evaluating the installed configuration. It does not by itself establish a numerical bound on accumulated damage.

D ADDITIONAL EMPIRICAL RESULTS

D.1 SINGLETON AUDIT ACROSS FAMILIES

Each adjacent pair is installed independently. A negative correlation is consistent with the intended lower-is-safer ordering; a threshold rule need not exhibit a strong global rank correlation. Table 18 reports inconsistent signs across families. OneVision's positive association is adverse to that ordering; its exploratory $p_{\text{perm}} = 0.027$ becomes 0.108 after Bonferroni correction across four tests.

Table 18: Singleton audit. Baseline binding is absolute accuracy on the audit set. Spearman correlations use average ranks for ties. Unadjusted, exploratory p_{perm} values use 20,000 within-family permutations (seed 20260918), a two-sided absolute-statistic test and a plus-one correction. Shared layers may violate exchangeability. The combined summary uses Fisher-z weights $n_f - 3$.

Family	n	Baseline binding	$\rho(S, \text{binding})$	p_{perm}
Qwen2-VL-2B	27	0.84	-0.001	0.997
LLaVA-1.5-7B	31	0.28	-0.089	0.636
LLaVA-OneVision-0.5B	23	0.80	$+0.463$	0.027
SmolVLM2-2.2B	23	0.80	-0.324	0.130
Fisher-z summary	104	–	$+0.009$	0.930

The Fisher-z summary of within-family Spearman correlations is $+0.009$ ($p_{\text{perm}} = 0.930$), using weights $n_f - 3$. This summarizes within-family associations rather than pooling raw scores across models; opposing associations can cancel. No family test survives Bonferroni correction across four tests (OneVision: adjusted $p = 0.108$). These exploratory tests assume exchangeability and do not establish absence of association. LLaVA-1.5's low baseline binding limits the measurable range of deterioration.

D.2 SINGLETON AND BLOCK-SHARING COMPARISON

For the Qwen2-VL comparison in Section 5.2 ($\varepsilon = 0.03$), singleton substitutions $L_{12} \leftarrow L_{11}$, $L_{13} \leftarrow L_{12}$, $L_{16} \leftarrow L_{15}$, $L_{24} \leftarrow L_{23}$ and $L_{26} \leftarrow L_{25}$ score 0.68, 0.68, 0.68, 0.84 and 0.72 against a 0.84 baseline. The block configuration changes the second substitution to $L_{13} \leftarrow L_{11}$ and scores 0.44 against a separate 0.96 baseline. These unmatched source maps and evaluation baselines prevent attribution to composition.

D.3 RANDOM SELECTION CONTROL

Table 19 compares random proposal with similarity-based selection under Q Cache's operation. On Qwen2-VL at $\varepsilon = 0.03$, random selection saves more cache (12.5% versus 8.9%) and retains more binding (0.590 versus 0.497 ± 0.051). The random summaries use two seeds; the remaining rows also differ in dose. These observations motivate auditing the proxy but do not establish that random selection is generally superior.

Table 19: Random versus similarity-based selection under Q Cache's operation. Binding is absolute recall. Random values average two seeds; the configurations are not matched on savings.

		similarity selection		random selection	
Family	ε	saved	binding	saved	binding
Qwen2-VL-2B	0.03	8.9%	0.497 ± 0.051	12.5%	**0.590**
Qwen2-VL-2B	0.08	39.3%	0.090 ± 0.044	12.5%	**0.560**
SmolVLM2-2.2B	0.008	8.3%	0.320	4.2%	**0.860**
SmolVLM2-2.2B	0.015	20.8%	0.000	5.2%	**0.585**

D.4 Q CACHE AT THE 13B SCALE POINT

LLaVA-1.5-13B is an additional checkpoint in the same family, not a fifth family. Table 20 reports single-draw results. At $\varepsilon = 0.008$, binding is 0.31 under the proxy rule and 0.36 under RecallGate, against a healthy score of 0.38. We report this descriptive difference without an equivalence or significance claim.

Table 20: Q Cache-style reconstruction on LLaVA-1.5-13B. Nominal savings follow $n/(2L)$; task scores are fractions. Results use single evaluation draws, with 100 items for binding; no across-draw uncertainty is reported. Gray shading and bold identify RecallGate, not superiority.

ε	Acceptance rule	Layers	Saved	Binding↑	GQA↑	POPE↑
–	*unmodified*	0/40	0.0%	0.38	0.70	0.867
0.008	Proxy rule (adapted)	10/40	12.5%	0.31	0.69	0.800
0.008	**RecallGate**	**8/40**	**10.0%**	**0.36**	**0.68**	**0.867**
0.015	Proxy rule (adapted)	30/40	37.5%	0.00	0.00	0.433
0.015	**RecallGate**	**16/40**	**20.0%**	**0.45**	**0.60**	**0.767**

At $\varepsilon = 0.015$, the proxy accepts 30 layers and saves 37.5%, with binding 0.00, GQA 0.00 and POPE 0.433. RecallGate accepts 16, saves 20.0%, and scores 0.45, 0.60 and 0.767, respectively. Aggregate VQA is 0.833 unmodified, 0.000 under the proxy rule and 0.467 under RecallGate. The compression doses differ, so this is an operating-point comparison rather than a matched-savings result. X-FRAME-ID changes only from 0.52 unmodified to 0.48 in the configuration with zero binding, GQA and VQA. It is therefore a weak damage indicator in this example. The decision trace also contains a candidate $(36, 37)$ with proxy score $0.00521 < 0.008$ but mini-battery score 0.500 below a floor of 0.619. This establishes disagreement between approval criteria; it does not establish that this candidate alone causes the model's full performance loss.

D.5 RANKED SELECTION PRINCIPLES AT 13B

The MC-R and KS-R arms use wholesale KV replacement at 13B in Table 21. All four runs reach the 150-audition cap before the ten-layer budget. A space allowing one earlier-source/later-target pair per layer pair has $L(L-1)/2$ candidates: 780 at $L = 40$ and 378 at $L = 28$; a cap of 150 corresponds to approximately 19% and 40% of those spaces. These ratios compare the cap with the earlier-source/later-target candidate-space size; they do not establish actual unique-candidate coverage after eligibility exclusions.

Table 21: Adapted MC-R and KS-R comparisons on LLaVA-1.5-13B at $\varepsilon = 0.03$ under wholesale KV replacement (savings n/L). Scores are fractions from one evaluation draw; savings differ between approval rules. Gray shading and bold identify RecallGate, not superiority.

Principle	Acceptance rule	Layers	Saved	Binding↑	GQA↑	POPE↑
–	*unmodified*	0/40	–	0.38	0.70	0.867
MC-R	Proxy rule (adapted)	4/40	10.0%	0.40	0.66	0.800
MC-R	**RecallGate**	**5/40**	**12.5%**	**0.42**	**0.68**	**0.833**
KS-R	Proxy rule (adapted)	5/40	12.5%	0.41	0.64	0.767
KS-R	**RecallGate**	**3/40**	**7.5%**	**0.45**	**0.70**	**0.867**

Observed binding ranges from 0.40 to 0.45 versus 0.38 unmodified. MC-R with RecallGate selects five rather than four layers and has higher observed task scores; KS-R with RecallGate selects three rather than five. With single draws and different doses, these results do not establish statistical superiority or equivalence. They document a low-dose regime in which the measured proxy configurations do not show the severe binding loss seen at the looser Q Cache setting.

D.6 Calibration dependence and the unsupported dilution prediction

The reported calibration sweep does not support the predicted dilution response. Changing the modality mix makes the criterion stricter in the reported sweep: mean scores rise and fewer pairs pass. This is evidence against the proposed empirical response, not confirmation of a dilution mechanism. Four calibration draws produce $5, 5, 5, 6$ lazy layers, with the six-layer draw retaining less recall. Calibration composition and seed can therefore change the selected configuration and should be reported. This does not establish universal irreproducibility across laboratories.

D.7 Depth and benchmark sensitivity

The reported Qwen2-VL depth profile is non-monotone: substitutions targeting $\ell = 1$ and $\ell = 14$ are particularly damaging, with more tolerant targets between them. This observation concerns the audited substitutions and does not establish depth as a general predictor of binding loss.

E RecallGate sensitivity and empirical scope

E.1 Battery size and acceptance floor

Table 22: RecallGate sensitivity on Qwen2-VL at $\varepsilon = 0.08$ under key-only sharing. Binding is absolute recall. Only the default-floor cell reports three-draw mean $\pm$ SD; other cells are single draws. Gray shading and bold identify default settings, not optimality.

Sweep	Setting	Lazy layers	Saved	Binding
Floor α	0.80	13/28	23.2%	0.660
	0.85	13/28	23.2%	0.660
	0.90	**11/28**	**19.6%**	$\mathbf{0.883 \pm 0.021}$
	0.95	9/28	16.1%	0.920
Battery n_{mini}	8	9/28	16.1%	0.79
	16	**11/28**	**19.6%**	**0.90**
	32	11/28	19.6%	0.89
	64	7/28	12.5%	0.72

Table 22 provides the full cells behind the main sensitivity summary. Configurations need not be nested across battery sizes because both measured performance and the model-relative floor depend on the sampled battery. The reported results do not establish a monotonic benefit from increasing battery size or a statistically optimal acceptance floor.

Recorded offline audition times are 19.5, 27.7, 26.3 and 45.8 minutes for 8, 16, 32 and 64 items, respectively. These are individual observed timings, not a fitted scaling law. The default battery-size cell uses a single evaluation draw (0.90), whereas the default-floor cell reports three evaluation draws (0.883 ± 0.021).

E.2 What the battery actually measures

RecallGate accepts when the measured mini-battery score of the cumulative configuration meets or exceeds its relative floor. This is an empirical decision on a finite battery, not a guarantee of 90% retention on an independent evaluation distribution. The floor does not separately guarantee every capability represented in the battery. A weak baseline, an insensitive probe or adaptation to a reused battery can yield a configuration that passes the gate but loses evaluation capability. The POPE-battery ablation motivates checking probe sensitivity. The non-improving LLaVA comparisons alone do not establish whether baseline competence, battery sensitivity or candidate selection explains the outcome. Disjoint seeds reduce direct item reuse, but the mini-battery and evaluation probes share template families. The study does not establish transfer to arbitrary recall distributions, trained cross-layer sharing, latent-KV architectures, or frontier-scale VLMs.

F STATISTICAL REPORTING AND CONDITIONAL THEORY

F.1 VERIFICATION OF REPORTED STATISTICS

Singleton Spearman correlations use average ranks for ties. The combined summary averages Fisher-transformed within-family correlations with weights $n_f - 3$ before inverse transformation, where n_f counts family candidates. Recomputation from saved records reproduces all four correlations and the combined value 0.008533, rounded to 0.009. This is not a pooled Spearman correlation over all 104 interventions. Exploratory two-sided permutation tests use $20{,}000$ within-family shuffles (seed `20260918`) and a plus-one correction (Phipson & Smyth, 2016). Shared-layer dependence may violate exchangeability, and opposing associations can cancel in the summary. Unverified singleton confidence intervals are omitted. Overlapping intervals or identical rounded summaries do not establish equivalence.

F.2 AUXILIARY DEFINITIONS AND SCOPE OF THE THEORETICAL ANALYSIS

The empirical conclusions of this work are based on intervention measurements rather than theoretical guarantees. We record here two auxiliary observations that clarify the interpretation of the audited gates.

Cumulative versus static approval scores. For a deterministic cumulative score, rejecting a candidate and restoring the previous configuration leaves the score equal to that of the last accepted cumulative configuration, provided that no subsequent truncation changes the configuration. This property does not hold for a static pairwise score, because the score used to approve an individual pair need not describe the cumulative installed intervention. This distinction motivates evaluating the modified configuration directly, but does not itself guarantee capability retention.

Calibration dilution. For a token-averaged proxy statistic, suppose that the change relative to a reference is bounded by $\Delta d_{\max}$ on a subset occupying fraction ρ_S of positions and by δ_{bg} elsewhere. Splitting the average over these subsets gives

$$|\Delta\phi| \le \rho_S \Delta d_{\max} + (1 - \rho_S)\delta_{\mathrm{bg}}.$$

This bound only describes how a token-averaged statistic may depend on the fraction of affected positions. It does not imply that increasing the number of background tokens must dilute the proxy score, nor does it establish a threshold at which capability damage becomes undetectable. We initially considered such a dilution hypothesis as a possible explanation for proxy–capability mismatch. However, the calibration sweep in Appendix D.6 does not support the predicted response: changing the modality mixture increased the reported proxy scores and made the criterion more restrictive. We therefore treat dilution as an unsupported hypothesis rather than an explanation of the empirical results. The evidence for proxy–capability mismatch in this paper comes instead from the measured interventions reported in Sections 5.2–5.6.

G ADDITIONAL RELATED WORK

G.1 TOKEN ALLOCATION, EVICTION AND SHARING

MEDA (Wan et al., 2025), LOOK-M (Wan et al., 2024), MuKV (Xiao et al., 2026), FastV (Chen et al., 2024) and CVMA (Chen et al., 2026) address related allocation, merging, pruning or eviction questions. These operations differ from replacing representations across decoder layers. The present audit asks whether the approval signal predicts retained capability after a cross-layer intervention; it does not claim that thresholding is the only source of compression failure or that the same empirical result holds for all compression methods.

G.2 BACKBONES AND TARGETED EVALUATION

The model families are described by LLaVA (Liu et al., 2023), Qwen2-VL (Wang et al., 2024), LLaVA-OneVision (Li et al., 2024a) and SmolVLM (Marafioti et al., 2025). V-NIAH (Zhang et al., 2024) motivates needle-style long-context evaluation, while deep-layer pruning (Gromov et al., 2025)

studies layer redundancy. Our synthetic binding and cross-frame probes test specific consequences of sharing; they are not a complete assessment of visual competence.