FIRA: FIXED-SUPPORT INTERVENTION AND READOUT AUDIT FOR VISUAL TOKEN PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A stable visual-token ranking can coexist with failures shared by the compared selectors. Fixed-Support Intervention and Readout Audit (FIRA) holds selected positions fixed while recording actual outputs and candidate-restricted decisions. On 800 A-OKVQA images, changing the readout reduces historical LLaVA delete–zero ranking interactions by 19.625 points in NF4 and 18.625 in FP16. Qwen2.5-VL with MMTok/DivPrune supports instead exhibits shared failures that largely cancel in selector comparisons. Native OCRBench-v2 text-image choices provide 536 questions and seven complete arms. On 429 confirmation questions, full-input accuracy is 68.30%, the shared excluded-correct shift is 19.35 [15.81,23.10] points and the differential shift is 3.73 [-1.18,8.51], with 97.5% paired-image intervals. The task meets its frozen difficulty criterion. Another 7,504 decisions confirm positive shared shifts at 25/75% retention. Saved traces from 194 receipts reveal different multi-token candidate decisions in 59/970 Qwen and 116/970 Monkey-OCR cells. These results motivate reporting absolute answer outcomes alongside relative rankings under explicit answer contracts.

1 INTRODUCTION

Retaining selected visual tokens and measuring answer accuracy is a common way to assess visual evidence. For text-image tasks, that measurement depends on both the intervention and the rule that turns model outputs into answers. *Fixed-Support Intervention and Readout Audit* (FIRA) studies the robustness of the resulting selector comparison.

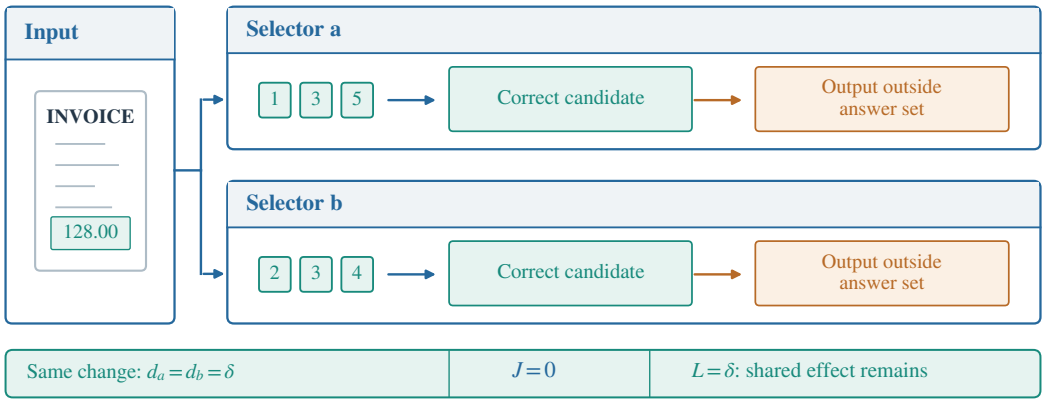


Figure 1: The measurement problem (schematic). A restricted candidate may be correct while the actual output falls outside the accepted answer set. Equal exclusion changes $d_a = d_b = \delta$ yield $J = 0$ but $L = \delta$: the ranking interaction hides a shared change. This symbolic example is not a measured prediction.

Figure 1 highlights the central blind spot: failures shared by selectors can cancel in a ranking interaction. FIRA keeps selected positions fixed within each selector, varies the declared intervention,

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.

and preserves both actual outputs and candidate decisions. It reports differential and shared effects with paired uncertainty and observed event counts.

Prior work establishes zeroing artifacts, calibration sensitivity and replacement dependence (Tan et al., 2026; Neo et al., 2025; Parodi et al., 2026). FIRA jointly tests intervention, selector and answer contract. Historical LLaVA and controlled Qwen comparisons distinguish unequal from shared failures. Published-selector audits cover 800 A-OKVQA and 3,565 TextVQA-derived confirmation questions; native OCRBench-v2 adds 429 confirmation questions and three budgets. Finally, 194 receipt traces connect these findings to sequence decisions. Together, these comparisons expose answer failures hidden by rankings.

2 Related work

Removal-based explanations. Covert et al. (2021) already organize explanations by removal procedure, explained behavior and summary statistic. Perturbation bias, optimized ablations, VLM mean replacement and faithfulness evaluation are established (Rong et al., 2022; Li and Janson, 2024; Neo et al., 2025; Zheng et al., 2025; Song et al., 2026). FIRA builds on this framework with a fixed-support empirical audit of selector differences and shared answer-contract failures.

Pruning and answer decisions. Tan et al. (2026) identify a FastV zeroing artifact and study pruning calibration; Wang et al. (2024) establish discrepancies between first-token candidates and generated answers. Answer-set mass is also established (Wiegreffe et al., 2023). FIRA jointly analyzes these issues through observed decision events: unequal exclusion effects change selector interactions, whereas shared effects can cancel in them. The native text-image evaluation demonstrates the latter below the full-input ceiling. Mean/difference coordinates and probability identities make the two effects separately measurable.

Text-rich tasks. Spatial-origin and integrity studies caution against equating selected locations with exclusively local evidence (Liu et al., 2026; Huang et al., 2026). Our complete-answer and receipt cohorts evaluate how answer contracts affect evidence measurement in text-rich tasks. Appendix S gives the extended comparison; Appendix Y records selected published external-selector results.

3 Fixed-Support Intervention and Readout Audit

FIRA fixes each selector's retained positions across interventions and audits how readout rules change reported outcomes (Figure 2). Paired estimands are defined below.

3.1 Interventions, selectors, and estimands

Let $z(x) \in \mathrm{R}^{N\times d}$ denote the real-valued visual states at an explicitly declared intervention site, and let selector s choose $S_s(x, q)$ with budget k. For each selector and example, we hold S_s fixed across operators. The four operators have different meanings: **Zero**: replace excluded rows with zero, retaining the dense graph. **Mean**: replace excluded rows with a frozen mean estimated from a separate image pool; retain the dense graph. **Mask**: prevent attention access to excluded keys in the declared layers while retaining rows. **Delete**: remove excluded rows and preserve the logical positions of retained rows. The operators have different computational meanings. Native FastV and the newer TextVQA/native-choice audits intervene at layer 2, after image–question mixing, so retained text states can contain visual information. The older 80-image TextVQA panel intervenes at decoder entry. Each site is analyzed under its own information-access conditions.

Let $U_i(s, o, r)$ be the score on example i under selector s, operator o, and readout r. For attention a and random b, define

$$D^r_o = \frac{1}{n}\sum_i \{U_i(a, o, r) - U_i(b, o, r)\}, \tag{1}$$

$$I^r_{o,o'} = D^r_o - D^r_{o'}, \qquad J^{r,r'}_{o,o'} = I^r_{o,o'} - I^{r'}_{o,o'}. \tag{2}$$

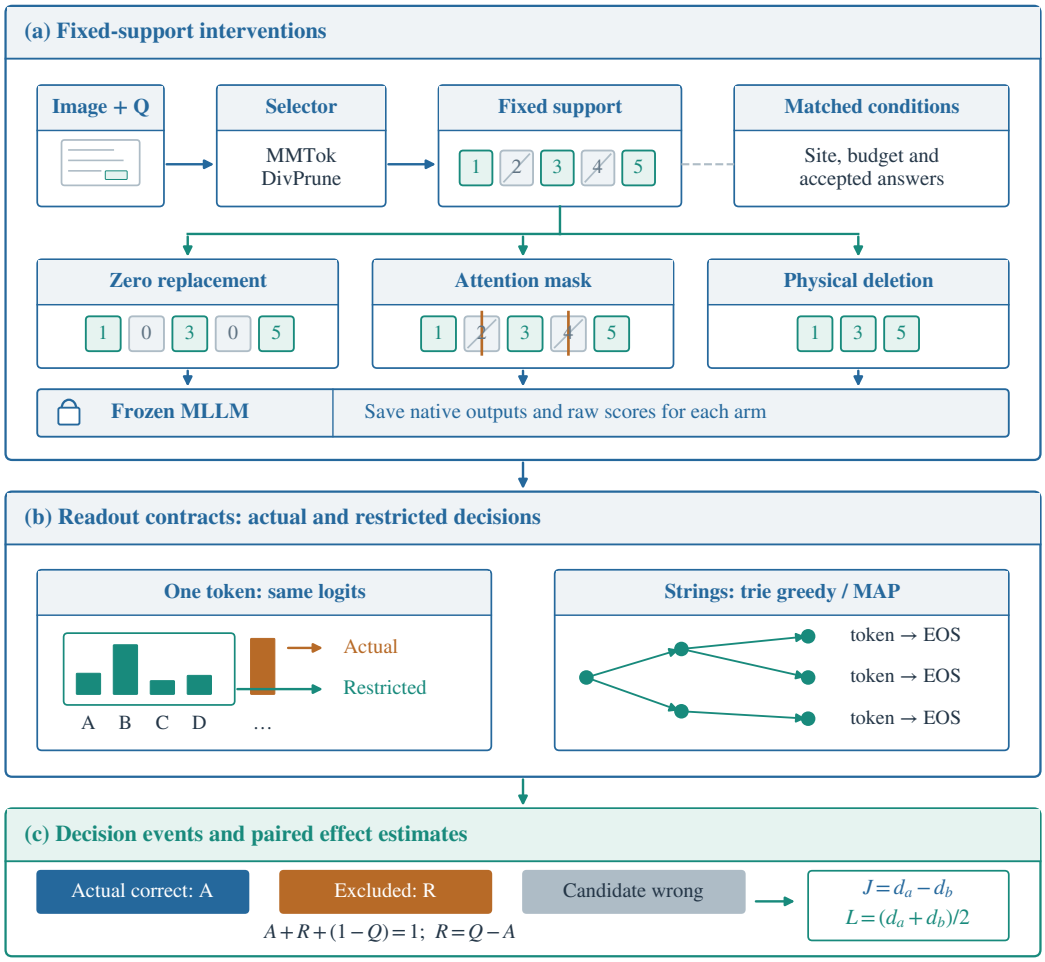


Figure 2: FIRA's fixed-support audit. (a) Hold each selector's support and intervention site fixed across zero, mask and delete; mean replacement appears in historical panels. Save native outputs and raw scores. (b) Contrast vocabulary-wide and accepted-set argmax on the same logits. For strings, reconstruct trie-greedy decisions from saved branches and compare them with whole-sequence MAP. (c) Partition actual-correct, excluded-correct and candidate-wrong events under the stated answer contract; paired J, L compare exclusion changes across operators. Tokens, bars and tree nodes are schematic.

I measures an operator–selector interaction; J describes its dependence on the readout. These paired differences preserve example identity. Repeated arms do not increase the number of independent images. Historical prospective contrasts and later descriptive readout changes retain their distinct statistical status. These contrasts depend on the score scale as well as the decision rule: under a common affine score change $U' = aU + b$, $I' = aI$. Consequently, comparing two bounded scores does not by itself isolate a change in the model's decision mechanism.

3.2 Evaluation procedure

Prospective comparisons fix the cohort, site, support, budget, score and contrasts before model calls; archived-output analyses retain their retrospective status. Every selector reuses its support across operators. Actual outputs, restricted decisions and paired event counts are retained throughout (Appendix N).

A structural control: masking versus deletion. In attention-only causal decoders, masking every excluded key in every downstream layer matches physical deletion in exact arithmetic when logical

positions, retained inputs and pointwise operations agree. Layer induction proves this structural control (Appendix N). Earlier text states can retain image information; zero replacement leaves accessible rows and generally differs.

3.3 Answer decision rules

Native single-token outputs. Historical LLaVA outputs use legacy substring, strict-uppercase and casefold letter scores. Modern Qwen pairs raw vocabulary argmax with argmax over accepted uppercase aliases, using a common tie order: A–D for four-choice tasks, or each native question's option letters. Restriction rescales no logits and generates no new answer; these are single-token decisions.

Decision-event accounting. For a single raw-logit argmax, use the same accepted uppercase aliases and tie order for actual and restricted decisions. Let A be strict actual correctness, Q candidate correctness and R the event that the candidate is correct but the actual output lies outside the accepted set. Then $A = Q - R$: restricting the decision can remove this exclusion, but does not repair the emitted answer. Applying the same paired contrast separates candidate, exclusion and legacy scoring-rule terms (Appendix L). This event partition describes decision outcomes under the stated single-step contract; visual-information attribution requires additional controls.

Shared failures can cancel in a ranking interaction. Write $\bar{R}_{s,o} = n^{-1} \sum_i R_i(s, o)$ and let $d_s = \bar{R}_{s,\text{zero}} - \bar{R}_{s,\text{delete}}$. Under the single-step partition,

$$J_{\text{strict,candidate}} = d_a - d_b, \qquad L = \tfrac{1}{2}(d_a + d_b). \tag{3}$$

If $d_a = d_b = \delta$, then $J = 0$ for any admissible δ, while $L = \delta$. Consequently, a small interaction cannot bound the shared exclusion change. L is an average across selectors, not the intersection of their per-image failure events. FIRA reports these mean/difference coordinates alongside actual accuracy. The original controlled-cohort shared analysis is exploratory. Later TextVQA-derived, published-selector and native-choice extensions freeze J, L jointly before their respective initial model calls; engineering amendments remain disclosed. The single-step assumptions remain essential.

Beyond single-token decisions. Receipt generation and whole-string candidate probability are different decisions. Trie-greedy restriction preserves the same local argmax at shared prefixes: if free decoding finishes in the fixed candidate token language, both completions agree and $A = EQ$. Appendix V states the EOS, alias and tie assumptions. We reconstruct trie-greedy decisions from archived native prefix probabilities and score whole-string MAP separately. Complete-answer TextVQA scores remain a separate measurement (Appendix U).

4 Experimental setup

Table 1 states the independent units of the central panels. Model and data changes are not isolated architectural interventions. We retain all designated examples, including incorrect original answers, invalid outputs, and capped generations.

The A-OKVQA cohort uses one identity-selected question per image, with recorded image-ID/RGB history exclusions. The two historical LLaVA pipelines retain 288 of 576 visual rows at zero-based layer 2, using layer-1 attention averaged across heads from the final input query. Random supports are fixed by question/index hashes; the mean uses 128 separate calibration images. Appendix M gives the exact prompt, layer convention, support rule, mean construction and generation settings.

Uncertainty resamples independent images, pairs or groups. Prespecified two-, four- and five-endpoint families use 97.5%, 98.75% and 99% intervals; later descriptive analyses use unadjusted 95% intervals. The original and added-budget families remain separate. Appendix A.1 records every family, replicate count and seed. These intervals do not correct the entire adaptive research history or establish equivalence when crossing zero.

Table 1: Central evidence. These are specified cohorts and tasks, not full benchmark leaderboard scores. The 40-pair TextVQA cohort is reused across controls.

Panel	Units	Model / path	Central measurement
A-OKVQA	800 images, one question each	LLaVA-1.5-7B NF4/FP16; Qwen2.5-VL-7B BF16	Historical: 9 arms/pipeline; Qwen: 7 arms per selector pair; differential/shared effects
TextVQA-derived	899 development + 3,565 confirmation questions; 2,830 images	Qwen2.5-VL-7B BF16	7 arms per selector pair; one-step four-choice audit; J, L frozen jointly
OCRBench-v2 choices	107 development + 429 confirmation questions; 506 images	Qwen2.5-VL-7B BF16	Native alternatives; 7 arms; fixed-layer J, L audit, not full-benchmark scoring
TextVQA	40 pairs / 80 images	Qwen2.5-VL-7B; InternVL3-8B	10% retention, complete-answer teacher probabilities
CORD	95 eligible test receipts	Qwen3.5-9B; MonkeyOCRv2-B-Und	50% retention, 3 selectors, total-value generation
WildReceipt	64 earlier and 194 prospective groups, separate cohorts	Qwen3.8-27B; MonkeyOCRv2-B-Und	Total extraction; original prospective interaction and later sequence reconstruction

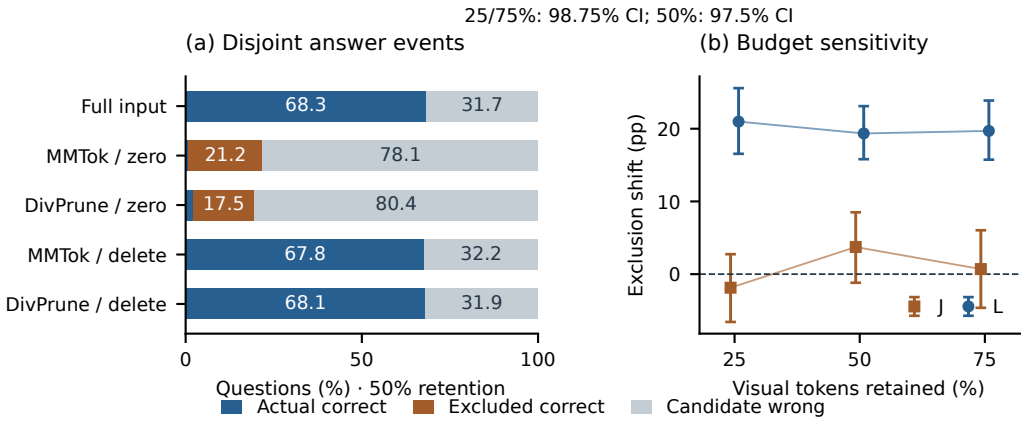


Figure 3: Native OCRBench-v2 choice confirmation, all 429 questions/405 images. (a) The same 50% audit partitioned into actual-correct (A), candidate-correct but excluded (R), and candidate-wrong ($1 - Q$) events. Bars include every question. (b) Differential J and average shared L shifts at all three tested budgets. New 25/75% endpoints form a four-test 98.75% image-bootstrap family; the original 50% two-test family retains 97.5% intervals. The joined points guide reading, not interpolation or a test of equal effects. All mask arms and development scores remain in Appendices AC and AH.

Table 2: Same-cohort delete–zero interactions in percentage points. Historical intervals for J are post-hoc unadjusted 95%; the new Qwen interval is prespecified 97.5% for a two-endpoint family with $I_{\text{candidate}}$. All 800 images remain included. Native preparation and numerical backends differ between model pipelines.

Model pipeline	I_{strict}	$I_{\text{candidate}}$	$J = I_{\text{strict}} - I_{\text{candidate}}$
LLaVA-1.5 NF4	24.750	5.125	19.625 [16.500, 22.750]
LLaVA-1.5 FP16	22.375	3.750	18.625 [15.500, 21.750]
Qwen2.5-VL BF16	1.500	0.750	0.750 [-1.625, 3.000]

5 RESULTS

Figure 3 shows the native-choice decision events and budget sensitivity. Appendix Figure 13 displays all seven arms at every budget and all receipt readouts. Appendix Figures 11 and 9 preserve the preceding audits and complete historical comparisons.

5.1 THE NATIVE SELECTOR COMPARISON CHANGES WITH THE READOUT

In NF4, under the original single-token scorer, attention's advantage over random selection is -22.625 points with zero replacement, 2.750 with mean replacement, and 2.500 with physical deletion or masking. All four originally fixed contrasts are shown in Table 23; mean and deletion cannot be called equivalent from a zero-crossing interval.

The historical NF4 and FP16 pipelines both show readout sensitivity. Exact event accounting gives $J = 19.625$ and 18.625 points, respectively (Table 2). Candidate restriction leaves a residual operator interaction; actual emitted answers are recorded separately. The complete original readout breakdown, its paired exploratory intervals, and the legacy-scoring decomposition remain in Appendix Z.

5.2 REPLICATION ON QWEN2.5-VL

We newly execute Qwen2.5-VL-7B (Bai et al., 2025) on the same 800 historical images and A–D questions. Layer-2 intervention, previous-layer last-query attention, question/index-hash random support, half retention and one-step raw-logit argmax are fixed. Actual and restricted decisions share the uppercase alias set and binary correctness score. Native preprocessing yields 99–529 visual rows (mean 351.63), versus 576 in LLaVA; chat rendering is also model-specific. This tests a controlled model pipeline, not a pure architecture intervention or independent dataset.

The controlled Qwen pipeline exhibits a distinct shared-failure regime: $J = 0.750$ [97.5% interval -1.625, 3.000] and the candidate interaction is 0.750 [-2.000, 3.500]. Full-input candidate accuracy is 88%, so this result is not explained by a full-input ceiling. Both zeroing selectors instead reach an actual-accuracy floor (Table 20). Their exclusions largely cancel in the ranking difference; the subsequent exploratory average exclusion change is 25.250 [95% interval 22.438, 28.188]. All 5,600 cells are complete (Appendix Q).

Contract failures are only part of the accuracy loss. New post-hoc accounting on the same 800 images gives an average actual deletion–zero loss of 86.875 points. This equals a 61.625-point candidate-decision loss plus 25.250 points of additional excluded-correct outputs (Table 21). The larger term is candidate-decision loss: changing the answer restriction alone does not restore deletion performance. The accounting separates observed score differences under the fixed decision contract. In particular, *shared* denotes averaging across selectors: 169 images have excluded-correct events in both zeroing arms, 36 only in attention and 30 only in random.

A separate probability diagnostic. Historical NF4 gold-letter log probability has a 3.600-nat delete–zero interaction: 3.493 nats come from answer-set mass and 0.107 from within-set probability (Appendix Table 9). This chain-rule accounting supports readout sensitivity in those saved scores; it does not quantify the fraction of accuracy loss or visual information.

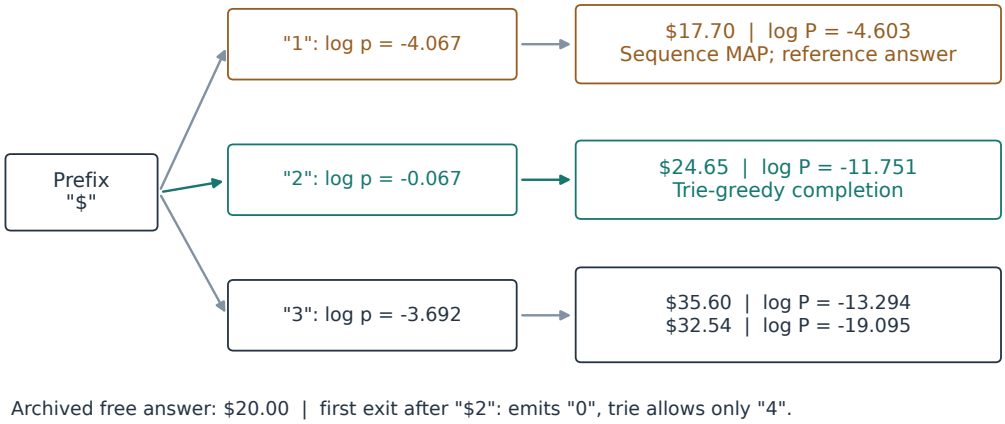


Figure 4: A real receipt from the 194-group archive. Whole-string MAP selects `$17.70`, local trie-greedy selects `$24.65`, and the native output is `$20.00`. Log P includes every token and EOS; arrows abbreviate remaining paths. This post-hoc illustration comes from the 15 full-input MonkeyOCR trie/MAP disagreements (Appendix AG); cohort results use all examples.

Table 3: Same 194 receipt images and fixed candidate strings, in percent. Actual responses are archived native generations; trie decisions are reconstructed from their paired teacher branches. The readouts solve different decision problems. All five arms appear in Appendix AG.

Full-input model	Actual EM	Trie greedy	Sequence MAP
Qwen3.8-27B	85.05	97.42	100.00
MonkeyOCRv2-B-Und	72.68	92.27	100.00

5.3 TextVQA-derived evaluation

A subsequent frozen panel retains 4,464 questions on 2,830 TextVQA validation images after excluding 336 previously used FIRA/reference image IDs. Image-level hashing assigns 899 questions to development and 3,565 to confirmation. All seven arms are complete: 31,248 decisions. Distractors come deterministically from official training answers; this is a derived four-choice diagnostic, not standard open-answer TextVQA. These images have prior SC-VTP evaluation exposure.

On confirmation, $J = 1.46$ [97.5% interval 0.36, 2.55] points and $L = 24.10$ [22.60, 25.63]. The mean actual deletion–zero gap is 93.39 points: 69.30 candidate-decision points plus 24.10 excluded-correct points. Thus shared failure is substantial even where the differential effect is small. Full-input accuracy is 96.49% (development: 96.66%), exceeding the frozen 95% difficulty ceiling; this panel therefore supports conditional failure accounting rather than the planned non-saturated confirmation. Appendix W retains every arm, protocol detail and scope restriction.

5.4 Sequence decisions in receipt extraction

Figure 4 illustrates the readout difference. On the same 194 WildReceipt groups, full-input whole-string MAP is 100% for both models. A later reconstruction uses every saved candidate-token conditional probability, including EOS, to traverse the original candidate trie greedily. Images, four candidates and five arms are unchanged. All 388 closed native records, 4,090 repeated prefix edges and 1,940 cell-wise inclusion identities pass verification; this is a saved-score analysis, not new constrained-generation execution.

Across all five arms, trie and MAP choose different strings in 59/970 Qwen and 116/970 Monkey cells. On zeroing arms, mean trie accuracy is below MAP by 5.67 [3.09,8.51] and 11.08 [7.22,15.21]

Table 4: MMTok/DivPrune-derived supports, half retention, Qwen2.5-VL BF16. J, L are percentage-point exclusion contrasts; full is actual/candidate accuracy under their shared answer contract. Four jointly specified endpoints use image-cluster bootstrap. The 899-question development partition is descriptive and retained in Appendix AA.

Cohort	Questions	Full (%)	J [98.75% interval]	L [98.75% interval]
A-OKVQA	800	88.00	-0.25 [-3.13, 2.63]	24.63 [21.19, 28.19]
TextVQA-derived	3,565	96.49	1.96 [0.53, 3.39]	24.60 [22.95, 26.28]

points (descriptive 95% group intervals). Thus a full-input MAP ceiling does not imply the same ceiling under a local decision rule. Canonical-token correctness equals normalized text EM in every cell, without added aliases. These decision-rule findings complement the original receipt interaction intervals, which continue to span zero. Earlier TextVQA replacement results and receipt details remain in Appendices AF and AE.

5.5 Shared failure extends to published selection rules

We replace attention/random with MMTok (Dong et al., 2026) and DivPrune (Alvar et al., 2025), obtaining 36,848 new one-step decisions over all 800 historical A-OKVQA and 4,464 existing TextVQA-derived questions. The two methods select original token positions from decoder-entry features; we reuse those positions for zero/delete/mask at zero-based layer 2. This tests selector coverage under a fixed audit, not their native entry-pruning accuracy or a new holdout. The frozen extension designates J, L on A-OKVQA and TextVQA confirmation jointly, before its model calls.

Both shared-effect intervals are positive. On A-OKVQA, full accuracy is 88%, the differential interval crosses zero, and the shared exclusion shift is 24.63 points. On TextVQA-derived confirmation it is 24.60, compared with a 1.96-point differential effect; the known 96.49% full-input ceiling remains. Thus the large shared term persists beyond attention/random supports. All 36,848 decisions satisfy the audited single-step event identity. Neither this identity nor candidate restriction restores actual generation. Full arm scores, implementation checks and the distinction from native pruning benchmarks are in Appendix AA.

5.6 Native OCRBench-v2 evaluation

To test the shared effect below the derived-task ceiling, we retain all 536 eligible native text-image choices from OCRBench v2 (Fu et al., 2025), with publisher-provided questions and alternatives. Image hashing assigns 107/429 questions to development/confirmation (101/405 images), without image overlap. These are text-rich reasoning and science QA, not pure transcription or the full benchmark. Appendix AC accounts for all 10,000 public questions and the exclusions.

The same Qwen model, half budget, MMTok/DivPrune supports and zero/delete/mask operators produce all 3,752 decisions. Full-input actual accuracy is 78.50% on development and 68.30% on confirmation; both meet the fixed 35–95% criterion. Confirmation gives $J = 3.73$ [-1.18, 8.51] and $L = 19.35$ [15.81, 23.10] points, using the frozen two-endpoint 97.5% image-cluster intervals. The mean actual deletion–zero gap is 66.55 points, comprising 47.20 candidate-decision points and 19.35 excluded-correct points. Thus shared failures persist on publisher-provided choices at non-ceiling full-input accuracy.

Candidate accuracy and response baselines. A secondary plan, fixed after launch but before inspecting correctness, compares candidate scores with uniform choice and always choosing the first option. Averaged over both zeroing selectors, candidate accuracy is 20.75%, versus 21.80% uniform expectation and 16.08% for the fixed first option. The paired difference from uniform is -1.05 [-4.28, 2.34] points (descriptive 95% interval). Table 29 shows how a substantial restricted score coexists with little probability on any valid answer. Full input assigns 99.70% mean mass to accepted answers. The response references are not neutral erasure controls or repairs; neither nonzero candidate accuracy nor L alone identifies preserved visual evidence. Appendix AC retains all arms, sources, option counts, exclusions and the disclosed interface recovery.

Shared exclusion persists across tested budgets. A follow-up retention-budget sensitivity panel repeats all 536 questions and seven arms at 25% and 75% retention: 7,504 new decisions. Its four confirmation endpoints were fixed before these calls; the reused cohort is not a new holdout. At 25/75%, L is 20.98 [16.55,25.58] / 19.70 [15.74,23.87] points, while J is -1.86 [-6.57,2.75] / 0.70 [-4.63,6.03] (98.75% image intervals). Both shared-shift lower bounds are positive, meeting the frozen sensitivity criterion (Figure 3b). This supports shared exclusion across the three observed budgets; it does not establish equal effects or a new task replication. Appendix AH retains every arm and control.

6 Discussion

FIRA connects relative selector comparisons to absolute answer outcomes. Fixed-support interventions test execution changes; explicit answer contracts determine how those changes enter the score.

Relative rankings and shared failures. LLaVA and Qwen reveal unequal and shared exclusion effects, respectively. Native OCRBench-v2 shows positive shared shifts at all three tested budgets. L averages selector-specific changes, rather than counting joint per-image failures; the J intervals include zero without establishing equivalence. Both coordinates distinguish absolute from differential effects.

Answer contracts in text-image tasks. Candidate accuracy evaluates a declared choice set; generation claims concern actual responses. Receipt reconstruction exposes decision-dependent ceilings by separating sequence MAP from trie-greedy choices. Exclusion need not indicate a format error. Actual responses, candidate decisions and answer-set mass should be reported together; visual attribution additionally requires image-blind or source-swap controls.

Evaluation scope. The controlled pipelines reuse 800 images with model-specific preprocessing, formats and supports. Older TextVQA controls reuse 80 images; the larger derived audit has prior exposure and exceeds its frozen difficulty ceiling. Native choices meet that criterion for text-rich reasoning/science QA with metadata options, rather than open-ended OCR or image-only options. Receipt primary intervals include zero. Sequence reconstruction reuses 194 groups and canonical candidates without new decoding. Merchant independence and backbone exposure are unverified. Other intervention sites and cross-attention architectures require separate native qualification. The results concern evaluation robustness; deployed OCR accuracy improvement remains untested.

7 Conclusion

FIRA combines fixed-support interventions, explicit answer decisions and paired event accounting. Differential and shared contrasts distinguish selector interactions from exclusion changes averaged across selectors. Native OCRBench-v2 yields a 19.35-point shared shift, positive at all three tested budgets. Receipt reconstruction separates actual responses, trie-greedy choices and sequence MAP. Together, these results support reporting actual answers alongside candidate scores and both absolute and relative intervention effects.

Reproducibility statement

Frozen protocols, model/data revisions, all designated outputs, exposure accounting and paired numerical replays are retained in the supplement. The appendices distinguish saved-score verification from native execution.

AI use statement

Generative AI assisted research design, experiments, analysis and writing. Appendix O details its use and verification scope.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *CVPR*, 2025. https://arxiv.org/abs/2503.02175.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, et al. Qwen2.5-VL Technical Report. *arXiv:2502.13923*, 2025. https://arxiv.org/abs/2502.13923.

Ritabrata Chakraborty, Rajatsubhra Chakraborty, Shivakumara Palaiahnakote, Angelo Cangelosi, and Umapada Pal. Vision-Language Models are Fragile Multilingual Associators. *arXiv preprint arXiv:2608.12333*, 2026. https://arxiv.org/abs/2608.12333.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by Removing: A Unified Framework for Model Explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. https://www.jmlr.org/papers/v22/20-1316.html.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive Entity Retrieval. *ICLR*, 2021. https://arxiv.org/abs/2010.00904.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. MMTok: Multimodal coverage maximization for efficient inference of VLMs. In *ICLR*, 2026. https://arxiv.org/abs/2508.18264v2.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, et al. OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning. *arXiv:2501.00321v2*, 2025. https://arxiv.org/abs/2501.00321v2.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. *Proceedings of ICML*, PMLR 70:1321–1330, 2017. https://proceedings.mlr.press/v70/guo17a.html.

Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary C. Lipton. Goodhart's Law Applies to NLP's Explanation Benchmarks. *Findings of EACL*, 2024. https://aclanthology.org/2024.findings-eacl.88/.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong Yu, Laizhong Cui, and Qi Tian. Nüwa: Mending the Spatial Integrity Torn by VLM Token Pruning. *ICLR*, 2026. https://arxiv.org/abs/2602.02951.

Jiwan Kim, Kibum Kim, Wonjoong Kim, Byung-Kwan Lee, and Chanyoung Park. Why and when visual token pruning fails? A study on relevant visual information shift in MLLMs decoding. *arXiv:2604.12358v1*, 2026. https://arxiv.org/abs/2604.12358v1.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. *NeurIPS*, 2024. https://arxiv.org/abs/2409.09951.

Feixiang Liu, Qiang Qiu, Hao Zhang, and Xinyue Wang. Beyond Accuracy: Auditing Spatial Provenance in Visual Token Pruning for OCR-Critical MLLM Inference. *arXiv:2608.00077v2*, 2026. https://arxiv.org/html/2608.00077v2.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards Interpreting Visual Information Processing in Vision-Language Models. *ICLR*, 2025. https://arxiv.org/abs/2410.07149.

Felipe Parodi, Jordan K. Matelsky, and Melanie Segado. Zero-Ablation Overstates Register Content Dependence in DINO Vision Transformers. *arXiv:2604.14433v1*, 2026. https://arxiv.org/html/2604.14433v1.

Yao Rong et al. A Consistent and Efficient Evaluation Strategy for Attribution Methods. *ICML*, 2022. https://proceedings.mlr.press/v162/rong22a.html.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models. *EMNLP*, pp.9895–9901,2021. https://aclanthology.org/2021.emnlp-main.779/.

Xurui Song, Weishi Wang, Zhongqi Yue, Kuluhan Binici, Tao Bai, Hongxin Shao, Daniel Dahlmeier, and Jun Luo. Visual Attention Faithfulness in Vision-Language Models is Heterogeneous. *arXiv:2609.00830v1*, 2026. https://arxiv.org/html/2609.00830v1.

Kaizhen Tan, Yang Feng, Heqing Du, Hanzhe Hong, Siru Tao, and Xin Xu. When Does Visual Token Pruning Improve Calibration? The Role of Evidence Coverage in MLLMs. *arXiv:2604.12035v3*, 2026. https://arxiv.org/html/2604.12035v3.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. *Proceedings of CVPR*, 2022. https://arxiv.org/abs/2204.03162.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. *Findings of ACL*, pp. 7407–7416, 2024. https://aclanthology.org/2024.findings-acl.441/.

Sarah Wiegreffe, Matthew Finlayson, Oyvind Tafjord, Peter Clark, and Ashish Sabharwal. Increasing Probability Mass on Answer Choices Does Not Always Improve Accuracy. *EMNLP*, pp. 8392–8417, 2023. https://aclanthology.org/2023.emnlp-main.522/.

Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI. *ICLR*, 2025. https://arxiv.org/abs/2410.02970.

A Complete NF4 factorial and reproducibility scope

These appendices document Fixed-Support Intervention and Readout Audit (FIRA).

A.1 Statistical families

Uncertainty resamples the declared independent images, pairs or groups. The original A-OKVQA four-interaction family uses 98.75% intervals; the original TextVQA two-endpoint family uses 97.5%; original receipt five-endpoint families use 99%. The prospective WildReceipt, controlled Qwen and new TextVQA-derived families each use 97.5% intervals for two endpoints; the last jointly designates J, L. The MMTok/DivPrune extension uses 98.75% intervals for four J, L endpoints across two cohorts. Native choices use 97.5% intervals for the original two endpoints; new 25/75% sensitivity uses a separate four-endpoint 98.75% family. Post-hoc readout, loss accounting, response-baseline and sequence-reconstruction analyses use unadjusted 95% intervals. Appendices retain the exact replicate counts, seeds and original intervals. None corrects the entire adaptive research history, and zero-crossing intervals are not equivalence certificates.

Table 5: NF4 factorial. All scientific cells remain in the denominator. Invalid means not exactly uppercase A–D. Candidate accuracy is a saved-score readout, not newly constrained generation.

Condition	Original (%)	Strict (%)	Casefold (%)	Candidate (%)	Invalid /800
Full	76.375	76.375	76.375	76.375	0
Attention delete	77.375	77.375	77.375	77.375	0
Attention mask	77.375	77.375	77.375	77.375	0
Attention zero	0.000	0.000	6.500	24.500	798
Attention mean	76.875	76.875	76.875	76.875	0
Random delete	74.875	74.875	74.875	74.875	0
Random mask	74.875	74.875	74.875	74.875	0
Random zero	22.625	22.250	22.250	27.125	155
Random mean	74.125	74.125	74.125	74.125	0

The NF4 panel retains 288 of 576 visual rows, with attention and random supports bound across all four operators. Physical removal and masking yield identical argmax outputs on all 800 images for

each selector, but no bit-identical full score vectors. For attention the maximum absolute differences range from 0.0087890625 to 0.0390625; for random they range from 0.0078125 to 0.0390625. Both medians are 0.015625. Task-output agreement therefore does not certify numerical equivalence.

Canonical letter IDs are 319, 350, 315 and 360. Exact-uppercase alias sets are A=(68,319,29909), B=(69,350,29933), C=(70,315,29907), and D=(71,360,29928). Ties use the smallest original vocabulary ID. The original readout audit checked every saved NPZ checksum, complete 32,000-score fingerprint, output ID and response. The existing NF4 CPU replay reconstructs its table means from 7,200 archived readout cells and independently regenerates the original four bootstrap intervals. It additionally verifies 14,400 mass-decomposition identities. This replay does not re-run the model or newly audit all hidden tensors.

A.2 Matched native FP16 pipeline

A completed nonquantized FP16 counterpart uses the same ordered 800 input records, including token IDs, questions and image identities, with the same single-token answer contract. It completed all nine arms: 7,200 scientific cells and 7,240 generation calls including engineering replays. The original FP16 technical gate and closed analysis passed; the subsequent saved-score analysis performed no new model calls. This completed A-OKVQA execution is distinct from the interrupted MMStar audit.

Table 6: Historical delete–zero operator–selector interaction in percentage points on the same 800 images. Each original-scorer interval belongs to its original four-contrast within-precision family. Strict-letter and candidate-letter values are saved-score readouts; Appendix M reports new exploratory intervals, separately from the original family. Candidate restriction does not regenerate answers. Argmax over all exact-uppercase aliases gives the same candidate result in both precisions.

Precision	Original scorer [98.75% interval]	Strict	Candidate
NF4	25.125 [20.875, 29.500]	24.750	5.125
Native FP16	22.500 [18.500, 26.625]	22.375	3.750

Within each precision, a selector's retained 288-of-576 support is fixed across operators. Across precisions, random supports agree on all 800 examples, whereas attention supports agree exactly on only 3; this counts equal sets, not their overlap magnitude. FP16 recomputes native attention rankings and uses its own calibrated mean. Precision therefore changes a matched pipeline, rather than being isolated with the NF4 supports and mean held fixed. The contraction under candidate readout occurs in both pipelines; it establishes neither precision equivalence nor invariance, and supplies no confirmation of the modern OCR endpoint. All values in Table 6 come from already closed historical analyses; these precision-table updates use saved scores. The Qwen2.5-VL experiment in Section 5.2 is a separate, newly completed model execution.

B Receipt population and outcome definitions

CORD-v2 uses revision `7f0115a4b758a71d6473b8d085751692da2fef98`, official test split. Eligibility is one nonempty `gt_parse.total.total_price`; 95 of 100 rows qualify. The original reference is retained, with only a leading case-insensitive Rp and optional period/whitespace removed according to the frozen currency-free prompt. Exact image hashes are checked against the 100-image CORD validation cohort. This does not certify merchant independence or absence from backbone pretraining.

WildReceipt's original test file has 472 records. The frozen field task requires one nonempty Total_value, a valid annotated total box, a nonempty normalized store string absent from training, an archive image, and no encoded-image match in the recorded history. It leaves 289 eligible records and 258 groups before deterministic selection of 64 groups. The archived dataset SHA256 is `f47646e16033d95c85a0b3a180a39d90cc354a729e51a280211b2fe7cae532bd`. Every exclusion and selected filename is retained. Ground-truth boxes and store labels never en-

ter token selection or the prompt. Original printed currency remains in the WildReceipt reference and requested answer, unlike the explicitly currency-free CORD prompt.

Qwen3.8-27B uses revision `1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0`; MonkeyOCRv2-B-Und uses `42a66faaf073bf8c67ec4b279d96dc798dd9599d`. The latter is the understanding checkpoint, not the parsing checkpoint. The nominal full budget is 1,280, with actual maxima of 1,280 Qwen rows and 5,120 Monkey rows. All interventions retain $\max(1, \lfloor N/2 \rfloor)$ rows. Decoder positions are preserved and physical cache slots are rebased. These are qualified research implementations, not claims about the latest author serving stack.

Generation is greedy with the native EOS and a 64-token cap. No capped output is discarded. NCS uses lowercasing and whitespace collapse, followed by $\max(0, 1 - d/\max(|a|, |p|, 1))$ for Levenshtein distance d. ANLS is NCS when NCS exceeds 0.5 and zero otherwise. Exact match uses the same text normalization. A current independent dynamic-programming implementation recomputes NCS, ANLS and exact match for all 2,544 archived receipt generations, checks paired pixels/features/supports, and reproduces all four central 99% intervals. These are string-level and paired-record checks, not fresh GPU experiments.

C Additional controls and sensitivity analyses

The additional controls examine source discrimination, replacement direction, restoration mechanisms and task transfer. Each is reported with its original evaluation family and decision criterion, separating mechanism-specific findings from broader attribution or deployment claims.

- On the reused 40 TextVQA pairs, fixed-support mean replacement exceeds a 50% image-blind normalized QA bound in the original two-endpoint analysis. The six-endpoint sensitivity gives excess intervals [9.4375,29.9375] points for Qwen and $[-1.4375, 20.6875]$ for InternVL; the joint claim fails under that broader family.
- On 64 synthetic pixel-identical-region pairs, both original models read all six-digit codes correctly. After retaining the common blank-region rows, the candidate source contrast is 13.360921 nats for Qwen and 0.108032 for InternVL; the latter fails the prespecified $\log 2$ lower-bound criterion. Exact free generation is 39.0625% and 0%, respectively. Candidate discrimination is not arbitrary-string recovery.
- Equal-coordinate-energy reference controls reject magnitude alone as an explanation, but the pool mean is not uniquely best: Qwen's prespecified alternate sign direction exceeds it by 5.138 QA points [0.826,10.000], with an exploratory 95% interval.
- First-layer normalization and softmax-dilution compensation fail their joint recovery criteria. These failures reject the specific tested restoration accounts; they do not prove an alternative universal mechanism.
- Selective-QA and state-projection tests do not meet their qualified transfer criteria from TextVQA to VQAv2. No deployable confidence method is claimed.
- The Monkey high-norm versus DivPrune rank-reversal criteria fail on both CORD and WildReceipt. Sign changes of sample means do not replace the joint interval criterion.
- Small whole-page, source-edit and oracle-selection panels remain development diagnostics. They do not supply a new OCR algorithm or an independent benchmark result here.

D Complete prospective OCR extension

The prospective extension uses every one of the 194 remaining eligible WildReceipt normalized store-name groups after excluding the 64 recorded earlier groups. One image per group is selected by a frozen filename hash, independently of model output. The original official test contains 472 records; the single-total task and declared exclusions leave 289 records in 258 groups before this separation. This is a specified total-extraction population, and a normalized store string does not certify an independent merchant or template.

Qwen3.8-27B and MonkeyOCRv2-B-Und each receive five arms: full input, and high-norm and DivPrune supports paired under zeroing and deletion at half retention. For each image–arm cell, the

protocol obtains one greedy output with the original 64-token cap and the complete probabilities of four frozen candidate strings. Candidates consist of the original total and three distinct distractors from training annotations, selected by a fixed nonnumeric pattern, length and hash rule. They do not enter the question prompt. Each candidate branch starts from fresh cache state, uses the same pixels, features and retained support, and scores all original string tokens followed by the tokenizer's primary EOS. The candidate set is a diagnostic construction that includes the reference; its mass is not the probability of every semantically valid answer. Each model completed all 970 free generations and 3,880 complete candidate branches. Qualification on two previously evaluated images is excluded from these counts.

The primary quantity is $J = I_{\mathrm{NCS}} - I_{\mathrm{candidate\ accuracy}}$, where I compares the high-norm minus DivPrune difference under deletion with the same difference under zeroing. Two model-specific endpoints use 20,000 paired group-bootstrap draws with seed 2026091902 and two-sided 97.5% intervals. NCS and candidate accuracy are different task scores, so this J includes both a decision-rule change and a scoring-scale change. The separately frozen descriptive supplement reports

$$J = \big(I_{\mathrm{EM}} - I_{\mathrm{candidate\ accuracy}}\big) + \big(I_{\mathrm{NCS}} - I_{\mathrm{EM}}\big). \tag{4}$$

The identity holds per group and after averaging. The first term compares two binary correctness measures; the second records the contribution of partial credit for generated strings. Both are signed interactions that can cancel. They do not identify error fractions or visual grounding, and receive no new significance test or inherited primary interval. Every group and both models remain in the analysis.

The frozen primary analysis also reports each I, all five arm means, complete strings, exact match, ANLS, caps and actual-versus-candidate correctness tables. A zero-crossing primary interval is not an equivalence result; if neither model establishes a direction, this extension does not confirm the proposed modern OCR readout dependence. An operator main effect, a sign change in sample ranks or a later descriptive decomposition cannot substitute for that result. Both completed endpoints include zero (Table 31). The integrated completion audit verifies all 194 closed image chunks per model, frozen supports and source identities, every generation/candidate branch, and agreement with the original analyses. Qwen has no capped outputs; Monkey has four, all retained. No new bootstrap, threshold, candidate set or endpoint was chosen after these results.

Table 7: All prospective receipt arms, each with 194 groups. Scores are on a 0–100 scale; caps are counts. These secondary summaries have no new confidence intervals.

Model	Arm	NCS	ANLS	EM	Candidate	Caps
Qwen3.8	Full	95.72	94.15	85.05	100.00	0
Qwen3.8	High-norm zero	90.27	86.64	76.29	98.97	0
Qwen3.8	High-norm delete	89.03	84.55	72.68	98.45	0
Qwen3.8	DivPrune zero	88.58	85.79	68.04	97.42	0
Qwen3.8	DivPrune delete	87.92	85.16	65.98	96.39	0
MonkeyOCRv2	Full	89.02	84.88	72.68	100.00	0
MonkeyOCRv2	High-norm zero	84.47	78.07	67.01	98.45	3
MonkeyOCRv2	High-norm delete	85.09	79.34	67.53	99.48	0
MonkeyOCRv2	DivPrune zero	81.92	76.79	55.67	98.97	1
MonkeyOCRv2	DivPrune delete	82.33	77.28	57.73	98.97	0

Table 8: The prespecified descriptive identity, in signed score points. Terms partially cancel; they have no new inferential status and cannot rescue the primary confirmation.

Model	$I_{\mathrm{EM}} - I_{\mathrm{candidate}}$	$I_{\mathrm{NCS}} - I_{\mathrm{EM}}$	J
Qwen3.8-27B	-2.062	0.968	-1.094
MonkeyOCRv2-B-Und	-2.577	1.759	-0.818

E HISTORICAL DESCRIPTIVE TABLES

Table 9: Post-hoc NF4 descriptive decomposition, 7,200 full score vectors and 14,400 candidate-set cells. Values are paired interactions, not unconditional arm means. The probability identity is numerical to 8.9×10^{-16} at the contrast level in the original audit. No new intervals or causal interpretation are assigned.

Delete–zero interaction (nats)	$\log P(G)$	$\log P(C)$	$\log P(G \mid C)$
Canonical letter tokens	3.599760	3.492578	0.107182
All exact-uppercase aliases	3.586683	3.479124	0.107559

Table 10: Within-receipt deletion-minus-zero differences averaged across three selectors. Different models have different actual visual-token counts, so this is not a controlled architecture comparison. Both prespecified Monkey selector rank-reversal criteria fail.

Dataset	Model	Delete − zero NCS points	Original 99% interval
CORD (95)	MonkeyOCRv2-B-Und	20.48	[13.24, 28.24]
CORD (95)	Qwen3.5-9B	0.01	[−1.73, 1.97]
WildReceipt (64)	MonkeyOCRv2-B-Und	4.54	[0.42, 8.90]
WildReceipt (64)	Qwen3.8-27B	−0.13	[−3.01, 2.63]

F Evidence provenance and analysis families

The evidence ledger links central analyses to source hashes. Statistical replays verify archived scores and strings; native execution checks are recorded separately, together with failed and interrupted historical runs. The prospective receipt panel completed both model-specific readout-interaction tests, whose intervals include zero. Later native-choice and sequence analyses address distinct tasks and decision rules and retain their own statistical status; they do not replace those primary receipt endpoints. The evidence here concerns fixed-support evaluation rather than learned-selector training.

G A completed paired-source follow-up on the same receipt population

The four-candidate full-input task is at ceiling in both recent models. A subsequently frozen follow-up asks whether two complete amount strings are assigned to their respective receipt images. Pairwise image–text matching is an established diagnostic design (Thrush et al., 2022); our matching arithmetic is not a new principle, and this field task is not a reproduction of Winoground's compositional task.

All 194 previously evaluated images enter 97 disjoint pairs. Pairing uses only annotation strings, fixed shape/edit-cost matching and deterministic rules; each pair has distinct normalized answers. It uses no predictions for choosing partners. For a pair (i, j) with answers (y_i, y_j), let $L(i, y)$ be the complete teacher-forced log probability of the canonical string followed by EOS, using the original image and question. The assignment margin is

$$M_{ij} = L(i, y_i) + L(j, y_j) - L(i, y_j) - L(j, y_i). \tag{5}$$

Assignment accuracy A_{ij} is one for a positive margin, zero for a negative margin, and one-half for an exact tie. An additive image-only or answer-only term in these scores cancels. This elementary cancellation neither removes image–answer interactions unrelated to field recognition nor proves spatial grounding. Reference answers never enter the question prompt, but the two-answer decision set is constructed with annotation access.

The backbone revisions, pixels, prompts, half-retention supports and five arms are identical to the earlier modern OCR experiment. Each model performs 1,940 new complete-answer scoring branches: own and partner answers for every image–arm cell. Every own-answer branch reproduces the earlier target IDs, token log probabilities and initial raw-logit hash exactly. Two previously exposed qualification images are separate from the scientific counts. No new free answers are generated.

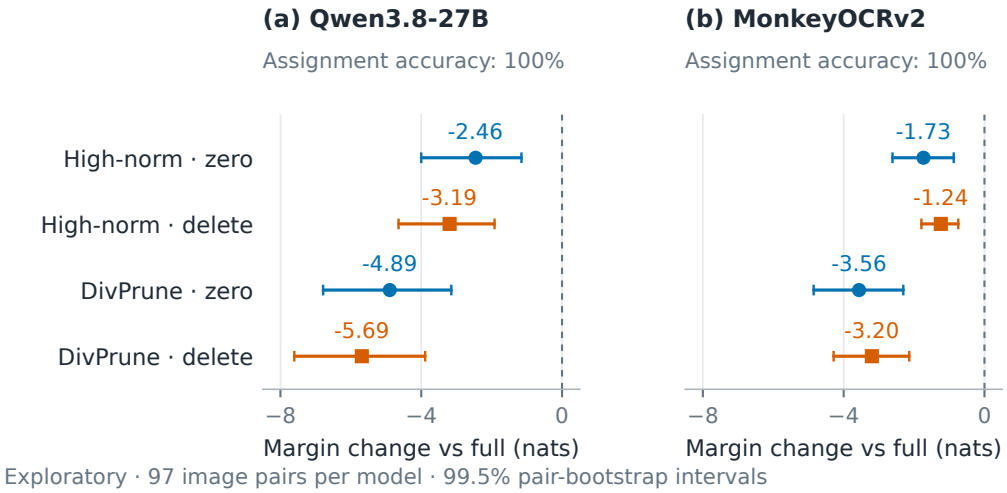


Figure 5: All eight exploratory pruning-minus-full assignment-margin contrasts on the same 97 frozen pairs per model. Whiskers are 99.5% pair-bootstrap intervals from the nominal ten-contrast family; the two operator–selector interactions remain unconfirmed and are retained in the complete table below. Every arm retains 100% assignment accuracy. Lower margins alone do not establish information loss: score calibration is an alternative explanation.

Define $K = I_A - I_{C_4}$, where I uses the same high-norm versus DivPrune, deletion versus zeroing contrast as the original modern experiment. C_4 is the previously measured four-candidate accuracy, averaged over each pair's two images. The fixed two-model family uses 20,000 resamples of the 97 intact pairs, seed 2026092101, and two-sided 97.5% percentile intervals.

Table 11: Complete follow-up, using all 194 reused images and 97 fixed pairs in each model. Both endpoints include zero. Original OCR endpoints remain unchanged.

Model	K (points)	97.5% interval
Qwen3.8-27B	−0.5155	[−2.5773, 1.5464]
MonkeyOCRv2-B-Und	−1.0309	[−3.0928, 0.0000]

Both models achieve 100% assignment accuracy in every arm, without ties; neither nonzero K is confirmed. These ceilings establish neither readout equivalence nor repaired free-generation errors. The reused population is not an independent replication. A closed-run audit verifies all 194 chunks per model, common pairing, own-branch replay, five arm summaries, both intervals and process cleanup. No pairing, candidates or endpoint is replaced after observing these results.

H Exploratory continuous margins on all frozen pairs

Figure 5 visualizes every pruning-minus-full contrast.

We reuse all 97 pairs, 194 images, five arms and two models from Appendix G, with no new model calls. Accuracies and mean margins were already known. Closed-chunk hashes and token-level score sums are verified; candidates, pairing, EOS and unnormalized sequence scoring are unchanged.

Each model has four pruning-minus-full contrasts and one high-norm/DivPrune by delete/zero interaction. We resample intact pairs 40,000 times, seed 2026092122. The 99.5% percentile intervals provide a nominal Bonferroni 95% family adjustment across these ten contrasts. Coverage is approximate; the adjustment does not correct post-hoc selection across the research history.

Table 12: All ten exploratory contrasts. Pruning margins decrease despite 100% assignment accuracy; both interactions remain unconfirmed. Calibration changes remain an alternative to information loss: positive score rescaling preserves accuracy but changes margins.

Model	Contrast	Mean (nats)	99.5% interval
Qwen	High-norm zero − full	-2.457	[-3.999, -1.153]
Qwen	High-norm delete − full	-3.194	[-4.645, -1.918]
Qwen	DivPrune zero − full	-4.894	[-6.787, -3.146]
Qwen	DivPrune delete − full	-5.689	[-7.607, -3.890]
Qwen	Operator–selector interaction	0.057	[-0.818, 0.987]
Monkey	High-norm zero − full	-1.733	[-2.612, -0.874]
Monkey	High-norm delete − full	-1.240	[-1.791, -0.744]
Monkey	DivPrune zero − full	-3.563	[-4.853, -2.306]
Monkey	DivPrune delete − full	-3.199	[-4.286, -2.137]
Monkey	Operator–selector interaction	0.130	[-0.824, 1.036]

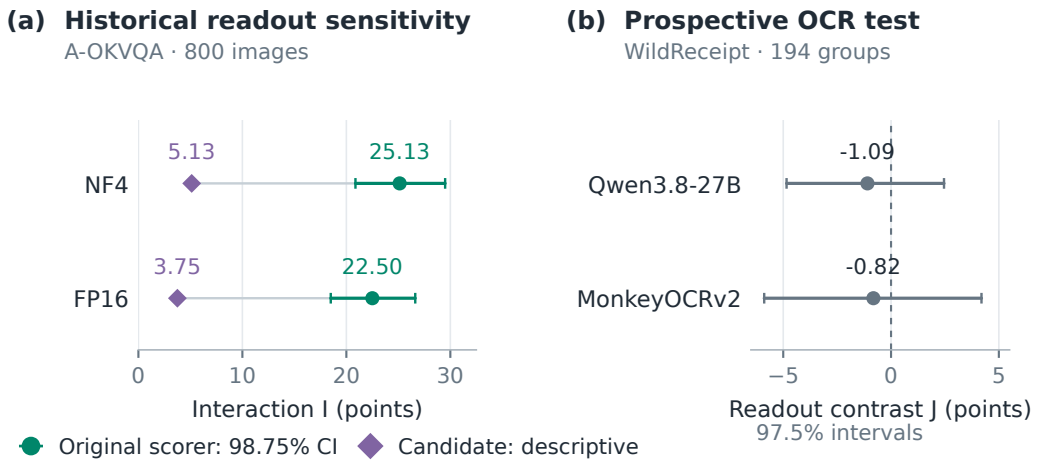


Figure 6: Historical readout sensitivity and its prospective transfer test. Left: the delete–zero selector interaction under original scoring and post-hoc candidate-letter restriction on the same 800 A-OKVQA images. Whiskers show original-family 98.75% intervals only; candidate points are descriptive and have no inherited interval. Gray lines connect the two readouts within each pipeline. FP16 recomputes attention supports and the reference mean. Right: the separately defined prospective receipt interaction J in both recent models, with 97.5% group-bootstrap intervals. These cross-task panels have different estimands and are not an effect-size comparison. Both modern intervals include zero.

I Historical intervals and precision check

The four-panel main figure expands the measurement mechanism. Figure 6 preserves the original interaction intervals and matched FP16 result. FP16 recomputes its supports and reference mean; it is not an intervention isolating precision alone.

J An archived receipt with different answer readouts

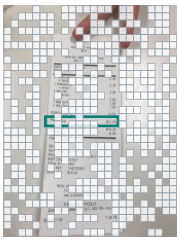
We select one illustration from the 194 prospective Qwen3.8 groups by a fixed post-hoc rule: full generation equals the reference, at least one pruning arm generates a different string while its candidate decision is correct, and every displayed response has at most 45 characters. Of 54 eligible groups, the minimum SHA256 of a fixed figure-specific prefix and sample ID selects the example below. It is not a new test set or a claim about a typical error.

(a) One receipt, fixed candidates, five recorded intervention arms



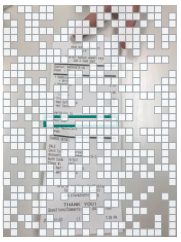
Reference: $14.00

High-norm support



384 / 768 tokens

DivPrune support



384 / 768 tokens

Original amount-due line



Currency symbol is part of the transcription target

Same retained positions for zero and delete

(b) Free output and complete-candidate scores disagree

Arm	Free output	Complete-string log probability (nats)			
		$68.50	**$14.00**	$12.18	$44.05
Full	**$14.00**	-23.601	-0.001	-20.585	-23.559
High-norm / zero	**$14.00**	-23.794	-0.002	-20.691	-25.720
High-norm / delete	**$14.00**	-22.015	-0.002	-19.570	-23.313
DivPrune / zero	**14.00**	-20.116	-0.902	-18.608	-23.658
DivPrune / delete	**14.00**	-20.527	-1.227	-18.532	-22.637

Qwen3.8-27B · selected illustration · candidates include the reference; generation is not repaired

Figure 7: Actual image supports, generated strings and complete-string log probabilities for one Qwen3.8 receipt. Both selectors retain 384 of 768 visual tokens and reuse their support across zeroing and deletion; all text is retained. The full prompt requests the printed currency symbol. DivPrune outputs "14.00" while the reference is "$14.00"; candidate scoring still selects that reference from four frozen strings. Shading marks relative candidate likelihood; printed values are unnormalized sequence log probabilities including EOS. The selected illustration does not confirm the prospective interaction, repair generation, or localize the information used by the model.

The original total-value annotation identifies the amount-due region; its enlarged crop is display-only. Every sequence score is checked against its saved token log-probability sum. The two operator arms have identical retained indices within each selector. Candidate strings are unchanged from the frozen experiment: “$68.50”, “$14.00”, “$12.18”, and “$44.05”. The reference’s score is -0.902 under DivPrune zeroing and -1.227 under deletion, versus approximately -0.001 at full input; this is a selected example, not a population estimate or evidence of causal information loss.

K A FINITE-CANDIDATE TEMPERATURE COUNTEREXAMPLE

This diagnostic concerns post-processing a fixed vector of complete-string scores, using the same candidate set for both images. It does not run a new model or estimate the cause of any experimental margin change. Positive temperature scaling is an established calibration construction (Guo et al., 2017). For scores $\ell_i(c)$ and $T > 0$, define

$$q_{i,T}(c) = \frac{\exp(\ell_i(c)/T)}{\sum_{u \in C} \exp(\ell_i(u)/T)}. \tag{6}$$

The strict monotonicity of the exponential preserves candidate ordering and ties. For a pair with distinct correct answers a and b, both in C, its crossed log margin is

$$M_T = \log q_{i,T}(a) + \log q_{j,T}(b) - \log q_{i,T}(b) - \log q_{j,T}(a) \tag{7}$$

$$= \{\ell_i(a) + \ell_j(b) - \ell_i(b) - \ell_j(a)\}/T = M/T. \tag{8}$$

The two normalizers cancel separately for each image. Hence sign-based assignment accuracy is unchanged, including the tie convention. For $M > 0$, any $T > 1$ decreases the margin without altering an archived neural representation or either decision. This gives a counterexample to interpreting margin magnitude alone as retained information. It does not establish calibrated probabilities, explain the eight observed losses, or invalidate margins as descriptive scores. Applying temperature inside autoregressive token softmaxes is a different intervention: candidate-dependent prefix normalizers need not cancel, so the result does not assert invariant sequence rankings in that setting. A common positive rescaling of all archived sequence scores is invertible and retains their rank information; actual pruning may change both score geometry and decisions.

L DECISION-SET EXCLUSION: COMPLETE ACCOUNTING AND SEQUENCE BOUNDARY

L.1 COMPACT ACCOUNTING IDENTITIES

For the common accepted set, E denotes actual-output eligibility and Q candidate correctness:

$$A = EQ, \qquad Q - A = (1 - E)Q = R. \tag{9}$$

For legacy score U, set $B = U - A$. Linearity of the paired contrast gives

$$I_U = I_Q - I_R + I_B. \tag{10}$$

For gold aliases $G \subset C$, the probability chain rule is

$$\log P(G) = \log P(C) + \log P(G \mid C). \tag{11}$$

These standard decision and probability identities organize the measured outcomes under declared answer rules and scores; they do not decompose visual information.

L.2 PROOF AND DEFINITIONS

With a shared total tie order, $t \in C$ implies $t = c$. Therefore the strict accepted-set score $A = 1[t \in C, f(t) = y]$ equals EQ, where $Q = 1[f(c) = y]$. Subtraction proves $Q - A = (1 - E)Q$. Linearity of the fixed four-cell contrast proves Equation 10. These are elementary properties of restricted argmax. They do not make candidate restriction actual constrained generation, equate probability mass with accuracy, or identify retained visual content.

Strict accepts all token IDs that decode to an exact uppercase A–D after the established stripping rule. The full alias candidate set uses those same IDs: A=(68,319,29909), B=(69,350,29933), C=(70,315,29907), D=(71,360,29928); ties prefer the smallest vocabulary ID. The separate canonical set has only (319,350,315,360). All four readouts remain in the released data. The previously reported case-insensitive readout accepts a different set and can change answers in both directions; it does not contradict the shared-set property.

Every cell of both complete 800-image, nine-arm analyses is included. We check both strict-correct/candidate-wrong and candidate-correct/strict-wrong directions. The former has zero cells; the latter equals candidate-correct excluded output in every cell. Under NF4 zeroing the attention/random counts are 196/39, giving $I_R = -(196 - 39)/800 \times 100 = -19.625$. FP16 counts are 189/40, giving -18.625. Full input, mean, deletion and masking have zero such cases. The complete operator/readout selector-order map is shown in Figure 9; no operator or image subset is selected for a favorable contrast.

The scoring-rule interaction $I_B = I_U - I_A$ is 0.375 in NF4 and 0.125 in FP16. Restricted interactions are 5.125 and 3.750. These exactly sum to the original interactions. The later review-driven analysis adds exploratory paired intervals (Appendix M) without changing the original success criterion or primary endpoints. Previous protocols, uncertainty families and known exposure remain unchanged; this is a descriptive refinement after the original outcomes were known. It is not an independent replication or a causal partition of neural computation.

L.3 GREEDY GENERATION CAN BREAK THE SHARED-SET INCLUSION ARGUMENT

Consider a normalized finite tree: $P(a) = 3/5$, $P(b) = 2/5$; $P(x \mid a) = 51/100$, $P(y \mid a) = 49/100$; and EOS follows b, ax and ay with probability one. The three full strings have probabilities 0.400, 0.306 and 0.294. Greedy emits ax. If $C = \{ax, b\}$ and the reference is ax, greedy is correct and belongs to the accepted set, but the candidate joint-likelihood decision chooses b and is wrong. Thus $A = EQ$ need not hold for greedy sequences even when accepted strings match. This basic greedy/MAP distinction marks a domain boundary of the single-step accounting; it does not explain either modern OCR interval or establish a new decoding result.

L.4 OPERATIONAL EVIDENCE RULES AND ORIGINAL SUMMARY

The original summary is retained to show probability-mass diagnostics and the unchanged prospective modern endpoints alongside the new decision accounting. The log-probability and accuracy decompositions concern different outcomes and must not be interchanged.

Table 13: Operational reporting rules grounded in the experiments. The protocol organizes evidence; its effect on deployed OCR accuracy is untested.

Observed pattern	Supported conclusion	Companion evidence
Selector comparison changes across operators	Performance depends on the intervention	Pair supports; disclose site, positions and computation
Generation and candidate decisions disagree	The answer rule changes the outcome	Keep actual strings; specify candidates; do not claim repaired generation
Accuracy saturates while margins fall	Continuous scores remain sensitive	Report all arms; check source dependence and score scale before information claims

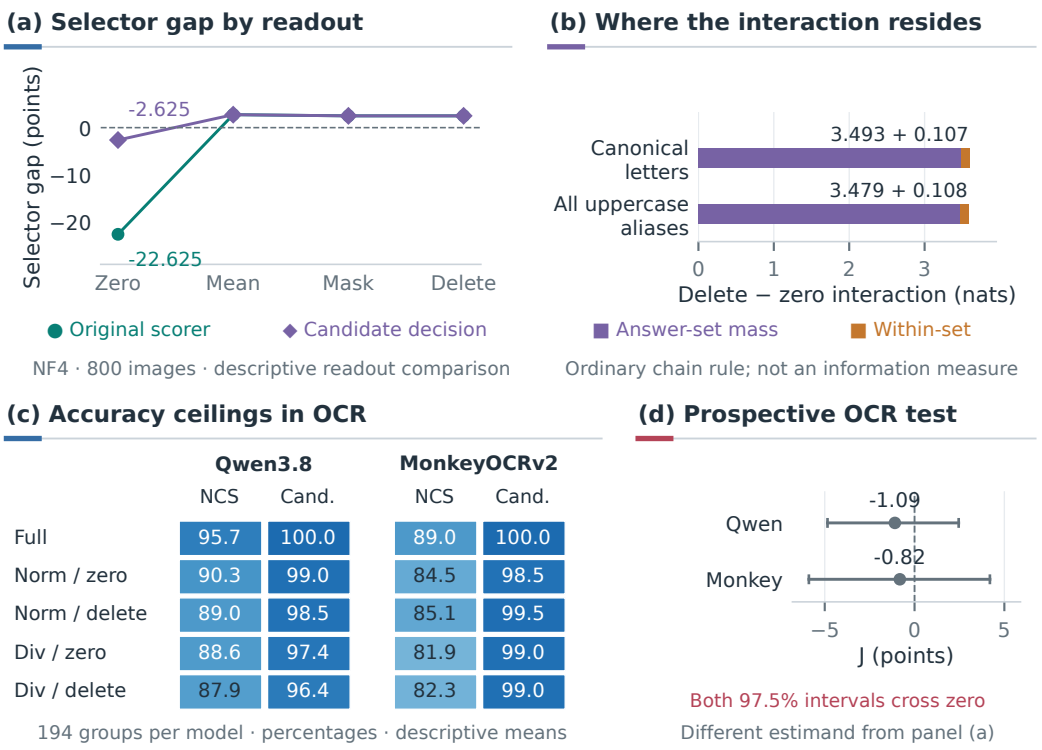


Figure 8: Evidence from historical measurement to prospective OCR. (a) Attention-minus-random accuracy on all 800 NF4 images under four operators and two readouts; the candidate view is post hoc. (b) Delete–zero log-score interactions split into answer-set mass and within-set discrimination for two candidate definitions; no new confidence intervals or information-loss interpretation. (c) All five arms in both modern models, 194 groups each; NCS and candidate accuracy are distinct scores. (d) The prespecified J intervals both include zero. Panels (a)–(c) are descriptive; Appendix I preserves the historical interaction intervals and FP16 check.

M EXACT HISTORICAL OPERATIONS AND PAIRED READOUT UNCERTAINTY

M.1 PROMPT, INPUT AND INTERVENTION SPECIFICATION

We use FastV commit `d1659729b5bf1be225e99ee15783deeea80f63b1`, LLaVA-1.5-7B, the author's Transformers 4.31.0 fork and tokenizer 0.13.3. NF4 uses bitsandbytes 0.43.1 double quantization with FP16 computation; FP16 recomputes its own attention supports and calibration mean. The original prompt template begins with a newline and is:

```
Analyse the image and choose the best answer for the
following question:
{question}
Options: (A) {choice_A} (B) {choice_B} (C) {choice_C} (D) {choice_D}
Just output the letter of the correct answer.
```

The first two displayed lines are a typesetting wrap: the exact template has one space between "the" and "following". Options are joined by a single space with no added punctuation. The supplement includes the literal template and all rendered conversations. The `llava_v1` conversation has user content `<image>` followed by a newline and the template, and an empty assistant slot. The author pad-aspect image processor produces a 336×336 image and 576 visual rows. Generation is greedy, one new token, no KV cache, with returned attentions and scores enabled; original stopping and output postprocessing are retained.

The intervention enters **zero-based decoder layer 2**, i.e., the third decoder layer, after layers 0 and 1 have processed the full image and question. Attention ranking averages all 32 heads from layer 1, takes the last input query's attention to the 576 image rows and applies native GPU `topk(288)`. We preserve the original top-k tie behavior, then restore selected positions to sequence order. This is separate from the smallest-vocabulary-ID rule for answer-token ties.

For random support, rank relative visual indices $i \in \{0, \ldots, 575\}$ by SHA256 of the literal string `202609277:random:` followed by question ID, a colon and decimal i. Take the first 288 and sort them numerically. Each selector's exact support is held fixed across its four operators. Delete gathers retained visual rows plus every text row, preserving original logical positions. Mask blocks excluded visual keys at layer 2 and every subsequent layer. Zero and mean overwrite excluded visual rows only once at the input to layer 2; all dense rows subsequently evolve normally. Earlier text states can therefore already contain visual information.

M.2 INDEPENDENT MEAN-POOL CONSTRUCTION

The calibration pool consists of 128 previously registered images from the original A-OKVQA Arrow, excluding all 800 evaluation images and the two qualification images. Sort eligible images by SHA256 of `202609276:image:` plus decimal image ID and take the first 128. Within each image, choose the question minimizing SHA256 of `202609276:question:` plus question ID. No model score enters this selection.

With FastV disabled, extract the 576 visual rows at layer-1 output (layer-2 input). Average rows within each image in FP64, then average the 128 image means in fixed order and cast once to FP16. NF4 and FP16 construct their means from their own native hidden states. The supplement binds all evaluation/calibration identities, prompts, token IDs and fingerprints to source files whose restored bytes match the historical hashes. It specifies the scientific operation; running a fresh full-model experiment still requires the pinned dependencies, public images and model weights.

M.3 EXPLORATORY PAIRED UNCERTAINTY FOR THE CENTRAL READOUT EFFECTS

For each of the same 800 images, first form delete–zero interaction vectors under original, strict and full-uppercase-alias candidate scoring. Form the three readout differences *within image*, then resample images jointly across all six vectors. We use 100,000 draws, NumPy default RNG seed 2026092204, percentile endpoints 0.025/0.975 and identical draws within each pipeline. These unadjusted 95% intervals were designed after the aggregate effects were known; they neither replace the original 98.75% family nor correct adaptive research history. The two precision pipelines share the same images.

Table 14: Complete exploratory readout contrast family with paired 95% intervals. The large readout change and positive residual concern different terms of the same comparison; neither implies repaired generation.

Delete–zero contrast (points)	NF4	FP16
Original interaction	25.125 [21.750, 28.500]	22.500 [19.250, 25.750]
Strict interaction	24.750 [21.375, 28.125]	22.375 [19.125, 25.625]
Candidate residual	5.125 [1.500, 8.750]	3.750 [0.250, 7.250]
Original − candidate	20.000 [17.000, 23.125]	18.750 [15.625, 21.875]
Strict − candidate	19.625 [16.500, 22.750]	18.625 [15.500, 21.750]
Original − strict	0.375 [0.000, 0.875]	0.125 [0.000, 0.375]

These intervals answer whether the central readout difference and residual are stable under paired resampling of the historical cohort. They do not supply independent modern-model validation. To identify a transfer boundary, a future controlled test must hold task, selector pair, support rule and answer contract fixed when changing the model, then vary one factor at a time. Its candidate difficulty must be fixed on separate development data before testing. The existing heterogeneous receipt panels and their ceiling-limited endpoints do not perform this test.

N AUDIT CONTRACT, INTERPRETATION AND CONTROLLED EXTENSION

Evaluation and reporting conditions. Prospective evaluations fix examples, support, intervention site, budget, accepted answers, score, resampling unit and contrasts before model calls. Retrospective analyses are identified separately. Native/all-keep replay and intervention-semantic checks precede execution, and all actual outputs are preserved alongside restricted decisions. An advantage for one declared operator/readout requires a positive paired lower bound; an advantage across a finite family requires positive simultaneous lower bounds in every designated cell. A zero-accuracy tie cannot discriminate between selectors, and a candidate ceiling cannot certify unchanged evidence. Harder candidates or sensitivity thresholds must be fixed on separate development data before a new confirmation. Candidate scoring remains valid for its declared task contract; it does not repair the previously emitted output. These conditions concern inference within the audited pipelines and leave the original MMTok/DivPrune benchmark results as separate evaluations.

Two different robustness claims. An operational pruning claim concerns actual generated performance under a particular deployed deletion operator. A token-importance claim additionally requires specifying what a counterfactual removal means. FIRA measures the latter claim's sensitivity without replacing the former's utility metric. Restricting the answer vocabulary can reveal a latent candidate decision but cannot retroactively change the generated response. The experiment therefore records both errors instead of relabeling forced-choice correctness as robust generation.

Worked historical decision audit. For NF4, strict actual-output delete–zero interaction is 24.750 points and restricted-candidate interaction is 5.125. Their 19.625-point difference comes from candidate-correct but contract-excluded actual outputs; the original literal scoring rule adds another 0.375, yielding its 25.125-point interaction. The paired interval for the strict-minus-candidate change is [16.500, 22.750], while the candidate residual is [1.500, 8.750]. Thus the numerical conclusion changes from a 25.125-point pipeline interaction to a 5.125-point restricted-decision interaction under the explicit alternative estimand. Both are retained; neither is presented as a recovered amount of visual information. This is a corrected interpretation of the existing comparison, not independent data validation.

Why masking is an execution control. Let retained hidden states agree at layer l. For a retained query at $l + 1$, discarded keys contribute zero attention weight, and retained keys/values and positional transforms agree with compact execution. Its attention output therefore agrees; position-wise normalization, feed-forward layers and residual addition preserve agreement. Induction proves equality at every later retained row. The statement requires masking every downstream attention layer, identical causal order and original logical coordinates. It excludes convolutions, recurrent memory, sequence-global normalization, position renumbering, dropout and caches whose state has not been compacted consistently. A hybrid model needs additional state control. Even under exact mathematical equivalence, fused low-precision kernels may differ numerically.

Matched scales before comparing readouts. For a modern multiple-choice replication, both actual and restricted readouts should use binary exact correctness with one accepted alias set and tie order. Comparing NCS with candidate accuracy additionally changes partial-credit scoring. The previous receipt endpoint is retained as originally specified, but cannot by itself isolate decision-rule sensitivity. Difficulty and distractors must be fixed on separate development data; choosing a new test subset because its candidate accuracy is lower would invalidate a prospective claim.

Controlled modern-model extension (completed). The separately frozen and now completed extension uses Qwen2.5-VL-7B on the same 800 historical A-OKVQA questions/images and letter prompt. It fixes the intervention at zero-based layer 2, derives attention scores from the previous layer's last input query, retains half of the visual rows, and uses the same question/index-hash random rule. Native image preparation caps the merged token count at 576; native chat rendering and image grids may differ from LLaVA, so the comparison is a controlled model-pipeline replication rather than a pure architecture intervention. Full input plus two selectors and zero/delete/mask give seven cells per image. One accepted uppercase-token alias set and a common vocabulary-ID tie order define both actual strict and restricted correctness. Every image remains in the denominator. The primary endpoint is strict-minus-candidate J; the candidate interaction I_Q is a second endpoint,

with paired image resampling and 97.5% intervals for the two-endpoint family. Native-hook replay, all-keep checks and data identity checks must pass before full execution.

This model execution uses an already studied cohort. All 5,600 cells and both frozen endpoints are reported in Appendix Q. The separate native-choice evaluation in Appendix AC extends the analysis to a text-rich setting below the full-input ceiling; broader open-ended document tasks require their own evaluation.

O FULL REPRODUCIBILITY AND ASSISTANCE DISCLOSURE

Reproducibility. The measurement section defines the interventions, fixed-support comparison and answer readouts. The evidence section and appendices identify dataset versions, eligible populations, model revisions, token budgets, scoring rules and statistical units. Frozen manifests and archived outputs bind images, supports and model executions; separate metric implementations check string scores and probability identities. The review supplement supplies exact historical operation definitions and all per-image interaction vectors; executable replay reproduces the new paired intervals. The depth supplement retains complete decision accounting and the sequence counterexample. Historical prospective tests and later descriptive analyses retain their distinct status, and incomplete native execution is recorded as incomplete.

AI assistance. Generative AI assistance was used in literature discovery and synthesis, hypothesis and method development, experimental design, implementation, formulation and checking of mathematical claims and proof sketches, data preparation, generation of synthetic control stimuli in the historical studies, result interpretation, scientific figures, manuscript drafting and translation. The recorded workflow includes source checks against primary literature, separate implementation checks, checks of data identity and metric calculations, and preservation of unsuccessful and interrupted experiments. Separate AI agent contexts assisted some implementation and audit passes; independence in these checks refers to code paths or contexts. The experimental sections distinguish completed model executions, CPU verification and planned work, and identify the scope of each validation.

P CONTINUOUS SOURCE-DISCRIMINATION DEFINITION

P.1 CONTINUOUS SOURCE DISCRIMINATION WHEN ACCURACY SATURATES

For images i, j with distinct answers a, b, retain all four complete-string scores and define

$$M_{ij} = \ell_i(a) + \ell_j(b) - \ell_i(b) - \ell_j(a), \qquad \ell_i(y) = \log P(y, \mathrm{EOS} \mid i, q). \tag{12}$$

Assignment accuracy records only the sign of M_{ij}. Any additive image-only term and answer-only term cancel in this crossed contrast. Nonlinear calibration and image–answer interactions do not cancel, so M is not an information measure or a correctness probability. Continuous margins under saturated accuracy are established (Chakraborty et al., 2026); here they supplement the fixed-support operator audit. Both sign and magnitude, together with actual generated answers, remain necessary. Their different estimands must not replace one another after observing results.

Q COMPLETE CONTROLLED QWEN2.5-VL REPLICATION

The same 800 historical image encodings, decoded RGB arrays, questions, option order and gold labels are recovered byte-for-byte from the pinned FastV public Arrow. A different public mirror and original COCO encodings fail the historical identity checks and are excluded before model execution; no approximate image substitution is used. The complete public validation Arrow has 1,145 questions; this study retains the same pre-existing identity-selected 800-image cohort, with one question per image. It does not become a pristine test set through model replacement.

Native Transformers 5.3.0 loads Qwen2.5-VL-7B-Instruct in BF16. Image preparation uses $\text{min_pixels} = 4 \cdot 28^2$ and $\text{max_pixels} = 576 \cdot 28^2$, and records actual visual counts: minimum 99, maximum 529 and mean 351.6275. This matches the retention fraction, not the absolute token count of LLaVA. The same literal letter question is rendered with the model's native image/chat template. Attention ranking recomputes the previous layer's last-query probabilities from its native Q/K projections and rotary transforms, then averages heads. Top-$\lfloor N/2 \rfloor$ positions are restored to original order. Random ranking uses the unchanged historical question/index hash rule on the native N rows. Supports are fixed before any intervention.

All seven arms use a single cache-free raw-logit argmax: greedy decoding, repetition penalty one and no logits processors. This explicitly defined decision differs from inheriting the checkpoint's generation preset. The candidate set enumerates every tokenizer ID decoding, after whitespace stripping, to exactly A, B, C or D; ties use ascending original vocabulary ID. Every raw argmax ID, decoded response, accepted-set event, candidate decision, candidate logit and full-logit fingerprint is archived. Full input plus attention/random crossed with zero, delete and mask gives 5,600 cells. Zero acts once at layer-2 input; mask blocks excluded keys throughout remaining layers; delete gathers rows once and carries original rotary coordinates through all remaining layers.

The first qualification stops at an all-keep replay discrepancy under mixed native SDPA kernel choices. Its failed status is retained. A distinct qualification uses the SDPA math backend consistently for every arm. Twenty same-arm native-forward/replay and all-keep checks on two historical rows give zero full-logit error and identical argmax. These rows remain in the full 800-image scientific denominator. Qualification is independent of answer correctness. Compact and dense BF16 paths can still differ numerically, so masking/deletion agreement is reported empirically rather than assumed from an exact-arithmetic proof.

Table 15: Complete new-model grid. Strict and candidate scores use the same accepted alias set. Candidate-correct excluded events explain their difference under the single-step decision partition; they do not repair the generated token.

Condition	Strict (%)	Candidate (%)	Eligible (%)	Excluded-correct (%)
Full	88.000	88.000	100.000	0.000
Attention zero	0.000	25.625	0.875	25.625
Attention delete	87.625	87.625	100.000	0.000
Attention mask	87.625	87.625	100.000	0.000
Random zero	0.000	24.875	0.750	24.875
Random delete	86.125	86.125	100.000	0.000
Random mask	86.000	86.000	100.000	0.000

The frozen primary family consists of $J_{\text{strict,candidate}}$ and $I_{\text{candidate}}$. Twenty thousand common image-bootstrap draws (NumPy default RNG, seed 2026092302) yield 97.5% intervals for these two endpoints. All source identities and arm counts are checked before analysis; no accuracy-based exclusions, checkpoint selection, candidate changes or early scientific stopping occur. Additional within-operator differences are descriptive and do not enlarge the primary claim.

Across 1,600 paired selector/image comparisons, mask versus delete gives 8 actual-token disagreements and 8 candidate-label disagreements. The maximum candidate-logit difference is 0.500. These BF16 observations are distinct from the zero-error same-arm replay checks.

The strict-minus-candidate interval includes zero: this controlled model pipeline does not establish the historical readout-interaction effect. This is a bounded replication result, not an equivalence

claim. Candidate-level uncertainty also includes zero. Full-input candidate accuracy is below ceiling. The models share task and identities, but native pixel preparation, token grids, templates and numerical pipelines differ. This experiment therefore narrows a concrete transfer question without identifying a single architectural cause. The full run shares a GPU with an existing job under a memory cap; no latency or throughput comparison is inferred.

R SHARED AND DIFFERENTIAL FAILURES IN THE CONTROLLED REPLICATION

This analysis was designed after the complete prespecified Qwen endpoints were observed. It retains every designated image and does not change the registered J or candidate-interaction results. For each image compute $R = (1 - E)Q$, then $d_a = R_{a,\mathrm{zero}} - R_{a,\mathrm{delete}}$, $d_b = R_{b,\mathrm{zero}} - R_{b,\mathrm{delete}}$, their mean and their difference. Twenty thousand paired image resamples with seed 2026092304 produce exploratory, unadjusted 95% intervals. All four quantities are reported:

Table 16: Complete exploratory family on the same 800 images. The main table preserves the wider prespecified 97.5% interval for J; this separate 95% reanalysis does not change its conclusion.

Exclusion-change statistic	Points	Exploratory 95% interval
Attention d_a	25.625	[22.625, 28.750]
Random d_b	24.875	[21.875, 27.875]
Shared mean L	25.250	[22.438, 28.188]
Difference J	0.750	$[-1.250, 2.750]$

These quantities describe candidate-correct outputs excluded by the declared actual answer contract, not a recovered amount of visual information. Candidate accuracy under zeroing is itself far below full input, so fixing answer format alone would not recover the original task accuracy. The common change locates one observed failure component rather than accounting for all damage caused by zeroing. The new lesson is methodological: ranking contrasts subtract away shared effects. Reporting both coordinates avoids interpreting their cancellation as an absence of answer-contract sensitivity. This is an elementary statistical fact supported by a newly executed controlled counterexample; the formulas themselves are not claimed as novel theory.

In the historical NF4 data, the corresponding attention and random exclusion changes are 24.500 and 4.875 points, giving $L = 14.6875$ and $J = 19.625$. The figure uses these descriptive historical coordinates and labels the new Qwen uncertainty separately. The pipelines share the same cohort; their different preprocessing, templates and numerical execution prevent identification of a uniquely architectural cause.

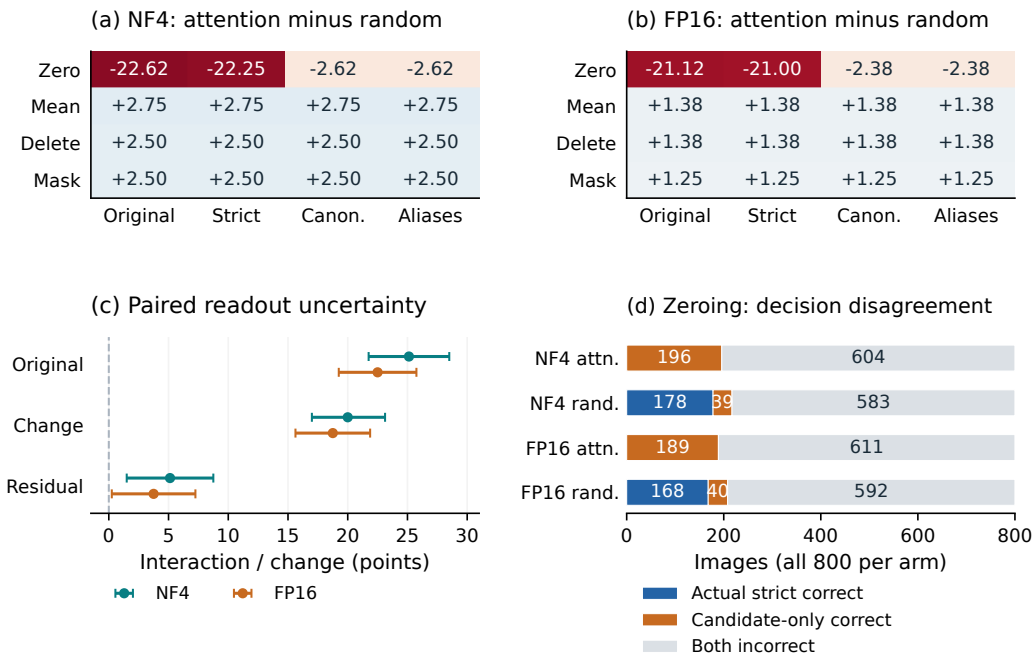


Figure 9: Observable decision dependence on the complete historical cohort. (a)–(b) Attention-minus-random accuracy for every operator and four readouts, using all 800 images in each pipeline; blue favors attention and red favors random. Canonical and full-uppercase-alias candidate readouts are distinct definitions even where scores agree. (c) Original interaction, original-minus-candidate change, and candidate residual with exploratory paired 95% intervals. NF4 and FP16 use the same 800 images; they are not independent cohorts. (d) Every zeroing outcome falls into actual strict-correct, candidate-only correct, or both-incorrect. Other operators have no candidate-only-correct cells. Panels (a), (b), (d) are descriptive; panel (c) adds post-hoc unadjusted intervals. Candidate decisions are saved-score readouts. FP16 recomputes its supports and mean.

S EXTENDED RELATED-WORK COMPARISON

Ablation and explanation evaluation. Covert et al. (2021) organize removal-based explanations by the removal procedure, the model behavior being explained, and the influence summary. Thus treating these choices jointly is an established framework. ROAD studies bias introduced by feature deletion (Rong et al., 2022); optimal ablation explicitly optimizes replacement values (Li and Janson, 2024). Explanation benchmarks can reward changes that do not improve explanations (Hsia et al., 2024). VLM mean ablation (Neo et al., 2025), register-reference comparisons (Parodi et al., 2026), and visual-attention faithfulness measurements (Song et al., 2026) establish the importance of the replacement reference. F-Fidelity tests evaluation validity through controlled degradation and addresses perturbation-distribution shift (Zheng et al., 2025). FIRA builds on these findings by holding selector support fixed, retaining complete outputs and separating intervention effects from answer decisions. Its paired empirical estimands complement training-based faithfulness evaluations. The processing-mode taxonomy of Song et al. (2026) concerns different models, data and query conventions and is outside this comparison.

Pruning and output reliability. Answer-set probability mass and its relation to surface-form competition are already formalized by Wiegreffe et al. (2023); increasing that mass need not improve accuracy. Wang et al. (2024) show that first-token candidate decisions can disagree with generated answers, including under restrictive prompts. FIRA uses these established distinctions to measure intervention-dependent answer failures. Tan et al. (2026) already separate pruning accuracy from confidence, identify zeroing versus removal in FastV, and compare normalized, unnormalized, and verbalized confidence. Their recorded answer-set masses are nearly one, so candidate normalization has little effect in those conditions. We examine a different empirical regime and estimand: does changing the answer decision rule change the operator-dependent comparison between selectors on the same examples? FIRA measures paired visual-intervention interactions, accounts for the associated decision events and tests their sensitivity in text-image tasks. The probability chain rule separates accepted-answer mass from the conditional decision within that set. These quantities describe the observed predictions; actual generation accuracy is reported separately from candidate accuracy.

Documents and spatial evidence. Spatial-origin audits distinguish token position from its effective receptive field (Liu et al., 2026). We likewise avoid equating retained locations with exclusively local evidence. Nüwa studies spatial-integrity damage under token pruning (Huang et al., 2026). Our receipt experiments use a specified field-extraction task to examine answer decisions under fixed token interventions. This connects spatial-evidence evaluation to sequence-level OCR outputs. Full document parsing and learned-selector training are separate evaluation settings.

T PAIRED ACCURACY-LOSS ACCOUNTING IN THE CONTROLLED COHORT

This additional analysis was specified after external review of the completed results. It introduces no model calls, changes no cohorts/supports/outputs, and leaves the two frozen primary endpoints and earlier shared-effect analysis intact. From each image's exact $A = Q - R$ relation,

$$A_{\text{delete}} - A_{\text{zero}} = (Q_{\text{delete}} - Q_{\text{zero}}) + (R_{\text{zero}} - R_{\text{delete}}).$$

All 800 images and seven arms are joined by question and image identity. The equality is verified for every image and selector before averaging. The two selectors are paired within each image, not treated as independent observations. Twenty thousand image-bootstrap draws with seed 2026092305 yield the new, unadjusted exploratory 95% intervals in Table 21. Its exclusion interval differs slightly from the earlier seed-2026092304 calculation because resampling draws differ; the older calculation remains unchanged.

The mean actual loss is 86.875 points, candidate loss 61.625 and exclusion increase 25.250. Excluded-correct events occur in both zeroing arms on 169 images, attention only on 36, random only on 30, and neither on 565; their union covers 235 images. Thus the mean L is not the proportion of images with failure in both arms. Eligibility itself is not correctness. An accepted but wrong letter remains an error, and a restricted correct label is not a successful actual response. This accounting does not establish that changing formatting alone would restore any generation or identify

how much visual evidence was lost. Portable per-image rows and a numerical replay accompany the analysis.

U COMPLETE-ANSWER AND RECEIPT MEASUREMENT DEFINITIONS

Complete-answer sufficiency. For an original answer $y = (y_1, \ldots, y_T, \mathrm{EOS})$, the TextVQA panel uses the arithmetic mean of non-EOS teacher-forced token probabilities,

$$A(y \mid z, q) = T^{-1} \sum_{t=1}^{T} p(y_t \mid z, q, y_{<t}), \quad \mathrm{Sgap}_o(S) = 1 - \frac{A(y \mid R_{S,o}(z), q)}{A(y \mid z, q)}. \tag{13}$$

This is not the joint probability of the string. Negative gaps remain in the analysis. Full-answer attention uses the complete answer and is a post-answer explanatory selector, not an online pruning method. Receipt experiments instead use complete greedy generation scored by normalized character similarity (NCS), thresholded ANLS, and exact match. We do not equate teacher probabilities, forced-choice accuracy and generated OCR quality.

Saturated candidate decisions. Receipt studies additionally retain crossed complete-string source margins when assignment accuracy saturates. These continuous scores are diagnostics, not information measures or replacements for the frozen accuracy endpoint; their definition and calibration limits appear in Appendix P.

V A SEQUENCE-ALIGNED AUDIT FOR COMPLETE CANDIDATE ANSWERS

Why change the comparison? Whole-string probability ranking and local greedy decoding solve different decision problems. The counterexample in Appendix L therefore prevents simply extending the single-step identity to the receipt panels. We instead specify a constructive, finite-candidate extension with the same local decision rule on both sides. Prefix-constrained decoding is established in autoregressive entity retrieval and incremental parsing (De Cao et al., 2021; Scholak et al., 2021); we apply these established tools to preserve an interpretable exclusion event across visual interventions.

Fixed contract. Let C be a finite set of valid *token sequences*, including EOS and all declared aliases, with an unambiguous answer mapping; let $C_\star \subseteq C$ be correct sequences. Every candidate fits the generation cap, EOS is a terminal edge, and every trie prefix admits a completion. At a prefix p, let $V_C(p)$ contain tokens extending p toward a member of C. For the same input, intervention, prompt, initial state, pre-mask scores $\ell(p)$ and token-ID tie order, compare

$$y_{\text{free}}: \quad \arg\max_{v \in V} \ell_v(p), \qquad y_{\text{trie}}: \quad \arg\max_{v \in V_C(p)} \ell_v(p)$$

at each successive decoding step. This is greedy traversal within the trie, not global sequence MAP or beam search. Define

$$E = 1\{y_{\text{free}} \in C\}, \quad A = 1\{y_{\text{free}} \in C_\star\}, \quad Q = 1\{y_{\text{trie}} \in C_\star\}.$$

Sequence inclusion. Under this contract, $E = 1$ implies $y_{\text{free}} = y_{\text{trie}}$, and hence $A = EQ$ and $R = Q - A = (1 - E)Q$.

Proof. If free decoding finishes in C, its first token is admissible. A vocabulary-wide maximum in the admissible subset is also that subset's maximum under the same tie order. The prefixes therefore agree. Repeat this argument until the common EOS edge. If $E = 0$, then $A = 0$ since $C_\star \subseteq C$. The two cases prove the identity. This deterministic coupling argument specifies sufficient execution conditions for the decision identity; it does not establish an information guarantee.

Locate the first exit. Record the first free-decoding prefix whose vocabulary-wide argmax is inadmissible. A valid free completion has no such exit. An invalid completion must exit by the maximum candidate depth. Its location distinguishes rejection before an answer token, after a partial word or number, and at termination. The original J/L accounting then applies to these sequence events with support and intervention fixed. Correct constrained decisions may still be guesses: neither inclusion nor a small R establishes retention of the visual evidence.

Table 17: Choose a readout comparison that supports the intended interpretation. Exact accounting is not a correctness or OCR-information certificate.

Comparison	Common decision mechanism	Status of $A = EQ$
One-step vocabulary / accepted tokens	Same logits, argmax, aliases and ties	Exact under the original contract
Free greedy / whole-string MAP	Local decisions versus global sequence probability	May fail even for a valid free answer
Free greedy / trie greedy	Same scores and tie order at every shared prefix	Exact under the stated sequence contract

Executable check and evidence boundary. A CPU reference implementation supplies trie construction, deterministic constrained/unconstrained traversal and first-exit logging. Synthetic checks cover 300 candidate/score configurations and every candidate gold label within each configuration (1,797 checks), including discrete-score ties and shared prefixes. An additional MAP counterexample is checked: free/trie greedy choose a EOS with probability0.306, whereas whole-string MAP chooses b EOS with probability0.3861. These remain reference-decoder checks. Appendix AG separately qualifies the recorded candidate tokens, EOS, raw-score processors and shared prefixes of the completed receipt runs, then reconstructs trie decisions from those saved probabilities. It adds

no native constrained-generation calls and does not change the original MAP experiment. Different prompts, beam search, length-normalized scoring or a different answer parser fall outside the proof. The finite-candidate audit does not establish open-vocabulary OCR robustness.

W COMPLETE TEXTVQA-DERIVED CONFIRMATION: POPULATION, CONTRACT AND OUTCOMES

Construction and exposure. We start with official TextVQA 0.5.1 validation (5,000 questions on 3,166 images). Excluding all 336 image IDs in two identified prior FIRA/reference cohorts removes 536 questions, leaving 4,464 on 2,830 images with no additional filtering. Exact decoded RGB checks find no duplicates within this cohort or overlaps with the 800-image A-OKVQA panel. This does not rule out resized duplicates or pretraining exposure. All 2,830 images have prior SC-VTP evaluation exposure; the cohort is new to this FIRA protocol, not a pristine repository-wide holdout.

The correct option is the modal human answer after Unicode NFKC, case-folding and whitespace normalization; ties use lexical order. Articles and punctuation are preserved. Three distinct distractors are modal answers from official *training* questions, ranked by question-word Jaccard similarity with a fixed SHA256 tie order; all annotated validation aliases are excluded. Correct option position is hashed from question ID. No model score or observed accuracy selects candidates. Unannotated valid alternatives remain possible. An initial normalization issue affecting articles was corrected before any model call, with its record preserved and no sample exclusions. This changes the task into one-step four-choice answering; none of its accuracies are official TextVQA scores.

Hash-sorting image IDs with seed 2026092306 places 566 images / 899 questions in development and 2,264 images / 3,565 questions in confirmation, with no image overlap. Both partitions, choices and endpoints were frozen before these model calls. The prespecified non-saturation gate requires full-input candidate accuracy in $[40\%, 95\%]$; failure does not permit resampling. Both partitions are retained despite exceeding 95%.

Execution. Qwen2.5-VL-7B-Instruct runs in BF16 with math SDPA, maximum 576 visual rows, half retention and the same layer-2 attention/random support protocol as the controlled A-OKVQA replication. All text is retained. Prompt options are A–D; native raw-logit argmax uses no generation logit processors. Candidate aliases comprise all tokenizer IDs whose stripped single-token decoding is an uppercase A–D; actual/restricted decisions share those IDs and token-ID tie order. This is one token, not free-form or trie-constrained multi-token generation. Frozen native/manual checks precede the full population; all 4,464 records publish input hashes and the seven arms. The completed panel contains 31,248 scientific decision cells. GPU sharing prevents interpreting elapsed time as isolated model efficiency.

Table 18: Every arm in both frozen partitions, percentages. Development has 899 questions; confirmation has 3,565. Actual A, candidate Q, eligibility E and excluded-correct $R = Q - A$ retain their one-step definitions. No arm or question is selected by outcome.

Partition	Arm	Actual	Candidate	Accepted	Excluded-correct
Dev.	Full	96.663	96.663	100.000	0.000
Dev.	Attention zero	0.667	25.806	4.672	25.139
Dev.	Random zero	2.781	25.362	12.013	22.581
Dev.	Attention delete	96.885	96.885	100.000	0.000
Dev.	Attention mask	96.663	96.663	100.000	0.000
Dev.	Random delete	94.661	94.661	100.000	0.000
Dev.	Random mask	94.438	94.438	100.000	0.000
Confirm.	Full	96.494	96.494	100.000	0.000
Confirm.	Attention zero	1.234	26.059	4.937	24.825
Confirm.	Random zero	3.198	26.564	13.436	23.366
Confirm.	Attention delete	96.101	96.101	100.000	0.000
Confirm.	Attention mask	96.045	96.045	100.000	0.000
Confirm.	Random delete	95.119	95.119	100.000	0.000
Confirm.	Random mask	95.091	95.091	100.000	0.000

Execution agreement and inference limits. Mask versus deletion yields 3/2 generated-token mismatches for attention/random in development and 2/8 in confirmation; aggregate similarity is not exact BF16 equivalence. The accounting identity $A = EQ$ holds in all 31,248 cells. Confirmation has

Table 19: The two jointly frozen endpoints. Intervals use 20,000 paired image-cluster bootstrap draws, seed 2026092307; primary inference uses confirmation. Development is reported for task diagnosis.

Partition	Endpoint	Estimate (pp)	97.5% interval
Dev.	J	2.558	[0.442, 4.730]
Dev.	L	23.860	[20.787, 26.974]
Confirm.	J	1.459	[0.364, 2.548]
Confirm.	L	24.095	[22.598, 25.628]

Table 20: Complete controlled replication: all 800 images in all seven arms. Values are percentages. Accepted output means membership in the declared alias set; it does not imply correctness. Attention delete/mask have equal aggregate scores, not necessarily identical decisions. Excluded-correct is R, a subset of candidate-correct events.

Qwen2.5-VL arm	Actual	Candidate	Accepted output	Excluded-correct
Full	88.000	88.000	100.000	0.000
Attention zero	0.000	25.625	0.875	25.625
Random zero	0.000	24.875	0.750	24.875
Attention delete / mask	87.625	87.625	100.000	0.000
Random delete	86.125	86.125	100.000	0.000
Random mask	86.000	86.000	100.000	0.000

mean actual deletion–zero gap 93.394 points [descriptive 95% interval 92.643,94.130], candidate gap 69.299 [67.758,70.804], and shared exclusion 24.095 [22.790,25.415]. These later descriptive gap intervals do not replace the primary 97.5% family. Full-input 96.494% accuracy [descriptive 95% interval 95.864,97.100] exceeds the frozen gate. A large L demonstrates observed exclusion under this task contract; it is neither generation repair nor a certificate of preserved OCR evidence. The new cohort expands sample scale and text-in-image coverage but leaves difficult, non-saturated open-answer OCR confirmation unresolved.

X Complete tables for the controlled A-OKVQA pipeline

The tables below retain the full earlier 800-image accounting, moved from the main text to avoid repeating numbers already shown graphically. They are unchanged in content and statistical scope.

Table 21: Paired loss accounting in the controlled cohort: the first row is the sum of the next two for every image and selector. New exploratory, unadjusted 95% intervals use 20,000 paired image resamples; this later analysis does not replace either prespecified endpoint or the earlier shared-effect interval.

Delete–zero loss (points)	Attention	Random	Mean [95% interval]
Actual accuracy	87.625	86.125	86.875 [84.625, 89.000]
Candidate accuracy	62.000	61.250	61.625 [58.062, 65.188]
Exclusion increase	25.625	24.875	25.250 [22.438, 28.125]

Y PUBLISHED CALIBRATION COMPARISONS AND FIRA'S DISTINCT ENDPOINT

The closest pruning-reliability comparison is Tan et al. (2026). The selected values below are reported in their Table 3 on LLaVA-1.5-7B / POPE (9,000 questions), not reproduced by us. All pruned rows use $K = 128$; full input uses 576. This source already identifies and corrects a zeroing artifact. FIRA therefore claims neither the first such diagnosis nor superior accuracy from that diagnosis.

Table 22: Selected paper-reported values from Tan et al., Table 3. Accuracy is higher-is-better; ECE/Brier are lower-is-better. This published POPE setting is distinct from our A-OKVQA and TextVQA-derived audits. FIRA uses separate cohorts and endpoints for its controlled comparisons.

Method	Accuracy (%)	ECE	Brier
Full input	86.9	.041	.099
SCOPE, coverage ($\alpha = 0$)	87.1	.016	.094
SCOPE, default ($\alpha = 1$)	86.9	.023	.095
VisionZip	84.9	.044	.111
Random (3 seeds)	83.5	.048	.119
FastV, corrected	77.8	.147	.180
FastV, zeroing artifact	50.1	.326	.360

Different empirical questions. Calibration metrics test whether confidence agrees with correctness. Our fixed-support actual/candidate events test whether excluding invalid emitted answers changes a selector comparison, including exclusions shared by both selectors. Neither metric family replaces the other. Published aggregate accuracy, ECE and Brier cannot recover per-example joint events or the paired uncertainty for J, L. A direct comparison requires the same examples, supports, operators and answer contract. Table 28 compares summaries on our completed outputs and is explicitly not a claim of superiority over SCOPE or corrected FastV.

Z COMPLETE HISTORICAL READOUT BREAKDOWN

Attention–zero produces only two exact uppercase-letter responses out of 800. Its remaining outputs include 230 lowercase letters, 18 empty postprocessed strings and 550 other responses. Exact case-insensitive scoring recovers 52 correct lowercase responses; attention–zero accuracy becomes 6.500%. Candidate restriction instead gives 24.500%, compared with 27.125% for random–zero, reducing the delete–zero interaction to 5.125 points. A new exploratory paired interval for this residual is [1.500, 8.750] points.

If accepted sets differ, candidate restriction need not preserve generated-answer correctness: attention–zero has 156 actual-wrong/candidate-correct cases but also 12 actual-correct/candidate-wrong cases under case-insensitive scoring. Nor does roughly chance candidate accuracy imply that the visual signal vanished; question priors and the forced-choice contract remain. The defensible conclusion is that the reported selector interaction strongly depends on what counts as an answer.

The completed native FP16 counterpart shows the same qualitative readout sensitivity on the same 800 examples: its original-scorer interaction is 22.500 points [98.75% interval: 18.500, 26.625], versus 3.750 under post-hoc candidate restriction (Appendix A.2). The parallel NF4 values are 25.125 and 5.125. The original-minus-candidate changes are 20.000 [17.000, 23.125] and 18.750 [15.625, 21.875] points for NF4/FP16; FP16's residual is 3.750 [0.250, 7.250]. These new 95% intervals are paired, exploratory and unadjusted (Appendix M). FP16 recomputes attention supports and its calibrated mean, so this is a matched-pipeline comparison, not fixed-support precision isolation or a modern OCR confirmation.

Exact accounting identifies the dominant observed term. The complete paired records satisfy Equation 9 under the declared common accepted set. Under NF4 zeroing, 196 attention and 39

Table 23: Complete original NF4 interaction family and descriptive readout sensitivities, on the same 800 images. Candidate means canonical-token restricted argmax; the complete exact-uppercase alias set gives identical correctness throughout. Only the original scorer has the displayed intervals. A degenerate empirical interval is not a population guarantee.

Interaction (points)	Original scorer [98.75% interval]	Strict	Casefold	Candidate
Mean − zero	25.375 [21.125, 29.750]	25.000	18.500	5.375
Delete − zero	25.125 [20.875, 29.500]	24.750	18.250	5.125
Delete − mean	−0.250 [−1.750, 1.250]	−0.250	−0.250	−0.250
Mask − delete	0.000 [0.000, 0.000]	0.000	0.000	0.000

random outputs are outside the accepted set while their restricted decisions are correct; no other operator has such cells. Thus

$$I_U^{\mathrm{NF4}} = 5.125 + 19.625 + 0.375 = 25.125. \tag{14}$$

FP16 gives $3.750 + 18.625 + 0.125 = 22.500$, with 189 and 40 corresponding zeroing cases. The large original interaction is therefore mostly an excluded-output contrast in this exact accounting. A residual interaction remains: candidate restriction does not erase operator dependence. The accompanying paired intervals quantify uncertainty in these empirical effects; the identities themselves do not identify information loss or repaired generation (Appendix L).

AA COMPLETE AUDIT WITH MMTOK- AND DIVPRUNE-DERIVED SUPPORTS

Design and scope. The extension tests whether shared excluded-correct changes persist under two published coverage/diversity rules. It reuses all 800 historical A-OKVQA questions and the existing 4,464-question TextVQA-derived diagnostic, including its 899/3,565 development/confirmation partition and known prior exposure. The latter comprises 2,830 images (566 development, 2,264 confirmation). No question or distractor is selected after observing this extension. This adds selector coverage and fresh model executions, not an independent held-out dataset or native benchmark accuracy.

Selection and intervention are distinct sites. The frozen model is Qwen2.5-VL-7B-Instruct in BF16 with Transformers 5.3.0 and math SDPA, maximum 576 visual tokens, half retention, and the existing uppercase A–D answer contract. MMTok uses the author's unmodified selection core at commit `8e2686836ed5`, with $\alpha = 0.5$, text/vision temperature 0.01, vision/vision temperature 0.2, full user text, and no keyword cleaning. Vision coverage uses means of four pre-merger patches in restored window order; merged vectors are verified against native decoder-entry visual rows. DivPrune uses the qualified reference selection port on the same merged vectors. Their original token indices are reused at zero-based layer 2 for zero replacement, physical deletion, or key masking. The first two layers retain full input, and can already mix image information into text states. These scores evaluate fixed layer-2 interventions on the selected supports. Earlier text-state mixing and the intervention site distinguish this setting from native decoder-entry pruning.

Qualification and complete accounting. The first two fixed questions from each task verify native-hook versus manual execution in all seven cells, plus all-keep checks and feature ordering. All 44 logged checks pass; the 28 designated native-replay cells have zero maximum logit error. Engineering examples remain in the complete scientific cohorts. All 36,848 decisions retain the top vocabulary token, decoded response, all 12 uppercase-alias candidate logits, answer-set log mass, support indices and input hashes. A separate replay verifies all immutable chunks and reconstructs candidate argmax with the same token-ID tie order, eligibility and correctness. The identity $A = EQ$ holds in every cell. This is a stored-decision verification, not a second native experiment.

Statistics. The frozen family comprises J, L on A-OKVQA and on TextVQA confirmation: 20,000 paired image-cluster bootstrap draws, seed 2026092402, with 98.75% percentile intervals. Both endpoints are fixed before new model calls, unlike the original A-OKVQA attention/random shared-effect follow-up. The cohorts and earlier outcomes were already known. Development is descriptive; its intervals do not supply an independent confirmation. Full-input difficulty remains 88.00% on A-OKVQA and 96.49% on TextVQA confirmation; the existing latter ceiling failure is unchanged.

Table 24: A-OKVQA: all 800 questions on 800 images. Scores are percentages; all seven cells remain included. This is the fixed-layer audit, not native method benchmark accuracy.

Support rule	Operator	Actual	Candidate	Eligible	Excluded correct
Full input	full	88.00	88.00	100.00	0.00
MMTok-derived	zero	0.13	24.63	1.25	24.50
MMTok-derived	delete	86.63	86.63	100.00	0.00
MMTok-derived	mask	86.88	86.88	100.00	0.00
DivPrune-derived	zero	0.00	24.75	0.63	24.75
DivPrune-derived	delete	87.00	87.00	100.00	0.00
DivPrune-derived	mask	86.50	86.50	100.00	0.00

Shared does not mean identical per-image failures. A-OKVQA: 155 questions have excluded-correct events in both zeroing arms, 41 only in MMTok, and 43 only in DivPrune. TextVQA-derived development: 154 questions have excluded-correct events in both zeroing arms, 67 only in MMTok, and 60 only in DivPrune. TextVQA-derived confirmation: 662 questions have excluded-correct events in both zeroing arms, 250 only in MMTok, and 180 only in DivPrune. Here L averages the two selector-specific exclusion changes; it is not an intersection count. Deletion and key masking

Table 25: TextVQA-derived development: all 899 questions on 566 images. Scores are percentages; all seven cells remain included. This is the fixed-layer audit, not native method benchmark accuracy.

Support rule	Operator	Actual	Candidate	Eligible	Excluded correct
Full input	full	96.66	96.66	100.00	0.00
MMTok-derived	zero	0.44	25.03	1.45	24.58
MMTok-derived	delete	95.66	95.66	100.00	0.00
MMTok-derived	mask	95.66	95.66	100.00	0.00
DivPrune-derived	zero	2.45	26.25	9.68	23.80
DivPrune-derived	delete	95.88	95.88	100.00	0.00
DivPrune-derived	mask	95.88	95.88	100.00	0.00

Table 26: TextVQA-derived confirmation: all 3,565 questions on 2,264 images. Scores are percentages; all seven cells remain included. This is the fixed-layer audit, not native method benchmark accuracy.

Support rule	Operator	Actual	Candidate	Eligible	Excluded correct
Full input	full	96.49	96.49	100.00	0.00
MMTok-derived	zero	0.39	25.97	1.88	25.58
MMTok-derived	delete	95.88	95.88	100.00	0.00
MMTok-derived	mask	95.88	95.88	100.00	0.00
DivPrune-derived	zero	2.66	26.28	10.60	23.62
DivPrune-derived	delete	95.99	95.99	100.00	0.00
DivPrune-derived	mask	95.96	95.96	100.00	0.00

disagree on the top vocabulary token for 5/7 A-OKVQA questions and 6/3 confirmation questions for MMTok/DivPrune respectively (development: 0/1). All disagreements remain included; native-replay qualification does not assert equality between distinct operators.

Coverage across selection rules. The common effect survives replacement of the original attention/random controls with published multimodal coverage and diversity rules. A-OKVQA gives a jointly specified positive shared effect without establishing a nonzero differential effect. TextVQA confirms both conditional contrasts on the existing derived task. This is evidence for joint absolute and relative reporting under the stated audit, not a new erasure principle, selector repair, calibration improvement, or independent non-ceiling OCR validation.

AB Comparison of reporting summaries on the original outputs

Recent pruning studies compare named selectors and several budgets before ablations. Calibration work additionally compares accuracy, ECE and Brier score (Tan et al., 2026); DSTP compares decoding-time evidence shifts across existing pruning methods (Kim et al., 2026). FIRA addresses a different question: what does a reported selector comparison reveal about actual answers? Table 28 applies alternative summaries to the same completed TextVQA-derived confirmation outputs. It requires no new model calls and is not a new pruning-performance benchmark.

Actual accuracy alone already detects severe pipeline failure. Joint reporting adds its observed relation to candidate decisions and shows why the differential effect alone misses much of the exclusion change: $L = 24.10$ [97.5% interval 22.60,25.63] versus $J = 1.46$ [0.36,2.55]. This result does not establish that FIRA detects more failures than calibration metrics; paired logits and a shared protocol would be needed for that comparison. Appendix Y places the closest published external-selector results alongside this distinction. Section 5.5 supplies new paired events for coverage selectors; published aggregate accuracy alone cannot supply them.

Table 27: All partition estimates in percentage points. A-OKVQA and confirmation belong to the four-endpoint family. Development is descriptive with the same nominal interval level. Intervals condition on the fixed model, supports and cohorts, and do not correct the broader adaptive research history.

Partition	J [98.75% interval]	L [98.75% interval]
A-OKVQA	-0.25 [-3.13, 2.63]	24.63 [21.19, 28.19]
TextVQA-derived development	0.78 [-2.37, 3.90]	24.19 [20.96, 27.62]
TextVQA-derived confirmation	1.96 [0.53, 3.39]	24.60 [22.95, 26.28]

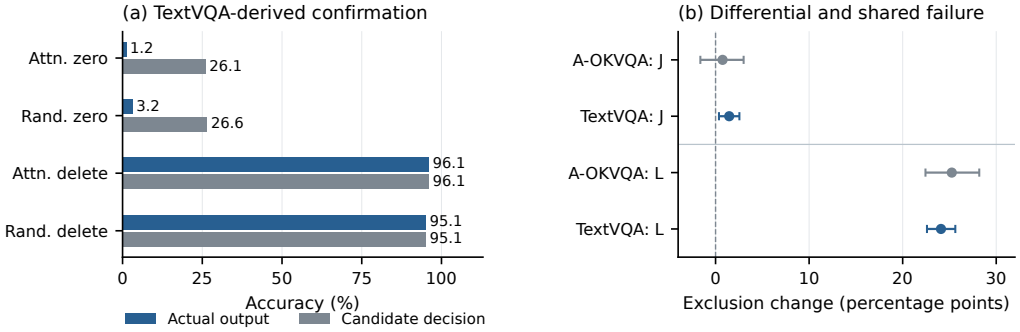


Figure 10: Preserved original attention/random result figure. Direct evidence of shared answer failures on Qwen2.5-VL. (a) All 3,565 TextVQA-derived confirmation questions: zeroing retains about 26% candidate accuracy but yields only 1–3% actual accuracy. These are derived four-choice scores, not standard TextVQA scores. (b) Separate horizontal intervals for differential J and shared L effects; the TextVQA estimates and A-OKVQA J use prespecified 97.5% intervals, while A-OKVQA L uses its exploratory 95% interval. Cohorts are distinct and use different prompts; these are conditional effects, not a controlled cross-task comparison. Both metrics are reported even when their magnitudes differ.

Table 28: Reporting comparison on identical saved decisions: 3,565 confirmation questions. Accuracy columns average attention/random zeroing arms; J, L compare zeroing with deletion, in percentage points. Dashes mean that the chosen summary omits that measurement, not that a method cannot compute it. This is an information-reporting comparison, not a detection-accuracy leaderboard or independent experiment.

Reported summary	Actual	Candidate	J	L
Actual accuracy only	2.22	—	—	—
Candidate accuracy only	—	26.31	—	—
Readout interaction only	—	—	1.46	—
Joint FIRA report	2.22	26.31	1.46	24.10

Table 29: Native-choice confirmation: descriptive means across MMTok/DivPrune on the same 429 questions (405 images), in percent. $A = Q - R$ counts decisions; answer-set mass is the mean summed vocabulary probability of accepted aliases, not accuracy. The exclusion gap is $L = 19.35$ points. No new model calls or primary tests enter this table.

Operator	Actual A	Candidate Q	Excluded R	Answer-set mass
Zero	1.40	20.75	19.35	1.86
Delete	67.95	67.95	0.00	99.69

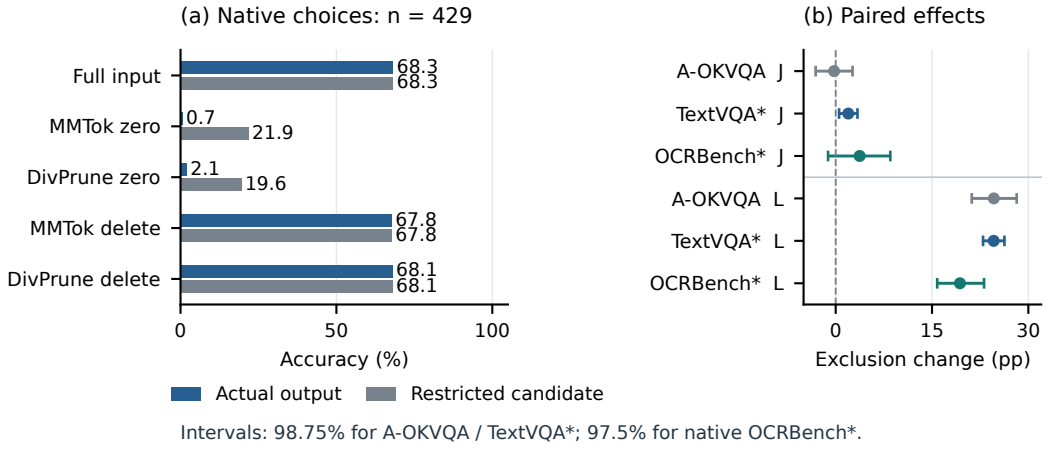


Figure 11: Fixed-layer MMTok/DivPrune support audits. (a) Actual and candidate accuracy on all 429 native OCRBench-v2 confirmation questions. (b) Differential J and shared L effects. A-OKVQA and TextVQA-derived intervals are the original 98.75% four-endpoint family; native OCRBench-v2 intervals are its 97.5% two-endpoint family. Asterisks mark specified derived/subset tasks, not benchmark leaderboard scores. All operators, development scores and prior results remain in the appendices.

AC NATIVE OCRBENCH-V2 CHOICE AUDIT

Data contract and history. The public `ling99/OCRBench_v2` revision is `c7e7cdf23bdb6774661e9b0caf0d9935a42feb8b`. Three original parquet files match their published SHA-256; all 10,000 IDs are accounted for. There are 9,301 non-multiple-choice exclusions, 40 non-single-uppercase annotations, and 123 cases whose options occur only inside the image. We do not infer their alternative set from the gold label. The remaining 536 questions contain an explicit contiguous A-starting option list in their original question text. A fixed anchored parser extracts the list and checks gold membership. Appending only an instruction to emit one listed capital letter preserves all original question/options. Candidate argmax and actual-output eligibility use exactly that question's native alternatives, including accepted tokenizer whitespace aliases and the same token-ID tie order.

The retained sources are CMMU (361), MMMU-Pro (137) and MMSI (38). Option counts are four (403 questions), ten (105), nine (15), three (6), five (4), eight (2) and six (1). This subset contains Chinese/English reasoning and science QA; no full-benchmark leaderboard comparison is made. Exact compressed-image and decoded-RGB comparisons find no overlap with the declared DocVQA, TextVQA and prior FIRA histories. This does not certify unknown histories or backbone exposure. RGB-image hashing with seed 2026092406 yields 107 questions/101 images for development and 429/405 for confirmation. No image crosses partitions; questions on the same image form one bootstrap cluster. These internal audit partitions are not official new benchmark splits.

Native execution and disclosed recovery. Qwen2.5-VL-7B-Instruct BF16 uses math SDPA, cap 576, half retention and layer-2 intervention. Decoder-entry MMTok/DivPrune supports remain fixed across operators, with window-permutation reversal checked against native decoder inputs. MMTok uses its pinned author core and default objective; DivPrune matches its qualified reference. This is an audit of derived supports at layer 2, not original entry-pruning accuracy.

The initial execution qualified two development questions but stopped on development ID 6609 after seven complete questions (49 cells): the author dynamic budget argument updates the inner selector, while its small-image early-return check reads an outer default of 64. The recovery synchronizes both budget fields to the same required k. Author objective/kernel parameters are unchanged. Seven CPU sizes (4,16,45,64,65,100,576) match the author coverage routine, retaining previous behavior above 64 rows. Original outputs/failure remain archived. All 536 questions are rerun separately; partial outcomes are not combined or selected. The original two development questions plus ID 6609 supply 33 checks: feature order, reference support, native/manual and all-keep replay. Every logit comparison has zero maximum error and identical argmax. No confirmation question supplies qualification. Cohort, arms, hypotheses and statistics stay fixed; the seven historical predictions are disclosed exposure.

Table 30: All native-choice arm scores in percent. Invalid outputs remain in every denominator. Development has 107 questions/101 images; confirmation has 429/405. All 3,752 cells satisfy the observed single-step event identity. The full model is the same fixed checkpoint throughout.

Partition	Arm	Actual	Candidate	Eligible	Excluded-correct
Development	Full	78.50	78.50	100.00	0.00
Development	MMTok zero	0.00	26.17	0.93	26.17
Development	MMTok delete	72.90	72.90	100.00	0.00
Development	MMTok mask	73.83	73.83	100.00	0.00
Development	DivPrune zero	0.00	25.23	2.80	25.23
Development	DivPrune delete	74.77	74.77	100.00	0.00
Development	DivPrune mask	74.77	74.77	100.00	0.00
Confirmation	Full	68.30	68.30	100.00	0.00
Confirmation	MMTok zero	0.70	21.91	2.56	21.21
Confirmation	MMTok delete	67.83	67.83	100.00	0.00
Confirmation	MMTok mask	67.83	67.83	100.00	0.00
Confirmation	DivPrune zero	2.10	19.58	5.59	17.48
Confirmation	DivPrune delete	68.07	68.07	100.00	0.00
Confirmation	DivPrune mask	67.60	67.60	100.00	0.00

Endpoints and response references. The primary confirmation family is J, L, using 20,000 paired-image bootstrap draws (seed 2026092407) and 97.5% percentile intervals. The 35–95% full-input actual-accuracy gate was fixed before original model calls and never used to exclude questions. All arms complete regardless of the gate. Development estimates are retained in the numeric supplement. Mask/delete raw-token disagreements are 1/0 for MMTok/DivPrune on development and 5/6 on confirmation. Both finite-precision arms are reported separately; the exact-arithmetic structural control does not imply bitwise identity for every BF16 execution. The secondary response-reference plan was recorded after recovery launch but before inspecting correctness. It uses 20,000 image-bootstrap draws (seed 2026092410), unadjusted descriptive 95% intervals, and reports every source/option-count stratum without selecting significant subgroups. Uniform expectation is $n^{-1}\sum_i 1/K_i$; always choosing A is a fixed label-blind rule. Option count and source are confounded with task content. A zero-crossing interval cannot establish chance equivalence, and neither reference is a neural erasure control. Token/letter frequency counts and answer-set probability are archived alongside every arm; candidate accuracy remains a separate decision from generation.

AD Decision outcomes with published selectors

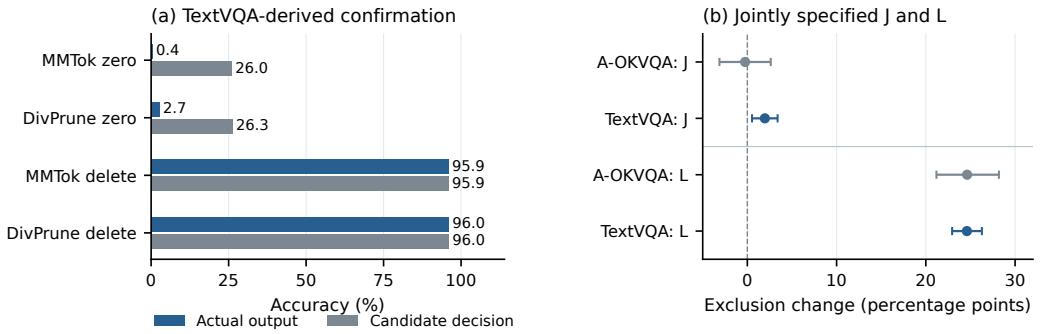


Figure 12: Published-selector support audit on Qwen2.5-VL. (a) Actual and candidate accuracy on all 3,565 TextVQA-derived confirmation questions, using MMTok/DivPrune supports at layer 2. These are four-choice diagnostic scores, not native method accuracies. (b) All four jointly prespecified J, L endpoints with 98.75% image-bootstrap intervals. The A-OKVQA J interval crosses zero while L remains large; TextVQA retains its known ceiling limitation. Original attention/random results remain in Appendix AA.

Table 31: Complete prospective receipt endpoints, in score points. The two readouts measure different tasks. Both original intervals include zero; the table retains the frozen endpoints and score definitions. All groups, including capped generations, remain in the analysis.

Model (194 groups each)	I_{NCS}	$I_{\mathrm{candidate}}$	J [97.5% interval]
Qwen3.8-27B	-0.579	0.515	-1.094 [-4.846, 2.457]
MonkeyOCRv2-B-Und	0.212	1.031	-0.818 [-5.883, 4.195]

AE RECEIPT READOUTS AND SENSITIVITY LIMITS

Single-step accounting does not extend automatically to greedy sequences: greedy generation and maximum whole-string likelihood can prefer different accepted answers. Appendix L retains a worked example. Receipt evaluation therefore preserves actual generation and treats candidate-string scoring as a distinct task.

The CORD test retains all 95 receipts with the frozen single-total annotation eligibility. WildReceipt uses 64 previously unevaluated normalized store-name groups excluded from the training store strings and the recorded image history. Each panel has full and resized-input controls plus random, high-norm and DivPrune selections paired under zero and deletion. The outcome is total-value text extraction at 50% retention, not full receipt parsing.

Deletion-minus-zero NCS averaged across three selectors is 20.48 points on CORD and 4.54 on the earlier WildReceipt cohort for MonkeyOCRv2, with original 99% intervals excluding zero. The Monkey result establishes a model-specific operator effect in these panels. Qwen intervals include zero, and the prespecified Monkey selector-rank reversal criteria are not met (Appendix Table 10).

The prospective study evaluates all 194 remaining eligible WildReceipt groups separately from those earlier 64. Each model runs 970 free generations and 3,880 fresh-cache candidate branches. The primary J contrasts high-norm versus DivPrune's delete-minus-zero interaction across NCS and four-candidate accuracy. Candidates are complete EOS-terminated strings, constructed from the reference and three frozen training-annotation distractors; they never enter the prompt. Table 31 reports both frozen endpoints. Both intervals include zero and allow several-point effects, leaving the prospective interaction unresolved. Because NCS additionally awards partial credit, J combines decision and score-rule changes: the descriptive EM–candidate terms are $-2.062/-2.577$, and the partial-credit terms are +0.968/+1.759 for Qwen/Monkey. These components retain their secondary status.

Both full-input models achieve 100% accuracy among the four candidates, while free-generation exact match is 85.05% for Qwen and 72.68% for Monkey. The 14.95/27.32-point separation exposes a ceiling in the candidate task despite substantial generated-answer errors. Appendix J shows an archived currency-symbol omission that candidate restriction does not repair. Appendix D gives all arms and the signed decomposition of J. A later paired-source follow-up on all 194 reused images also remains at ceiling, with both interaction intervals crossing zero (Appendix G).

AF COMPLETE-ANSWER REPLACEMENT RESULTS

The same 10% full-answer-attention supports on TextVQA yield mean sufficiency gaps of 0.992636 versus 0.059994 in Qwen2.5, and 0.986859 versus 0.024811 in InternVL3, under zero versus independent-pool mean replacement. The paired zero-minus-mean differences are 0.932641 [0.894262, 0.968103] and 0.962048 [0.932710, 0.985636], respectively, using the original 97.5% intervals.

This does not identify the mean as faithful erasure. In the corresponding comprehensiveness test, which replaces the selected 10% rather than the excluded 90%, the differences are much smaller: 0.069593 [0.025792, 0.113981] for Qwen and 0.001027 [-0.022009, 0.023214] for InternVL, with exploratory 95% intervals. The dose and replaced identities differ, so this asymmetry does not establish a general necessity–sufficiency law.

Coordinate-energy, reference-direction, image-blind and pixel-identical controls delimit this interpretation: the full family does not support a joint two-model grounding claim. Appendix C retains all outcomes. Mean replacement is a specified reference, not uniquely neutral erasure.

Receipt transfer remains a separate question. CORD and WildReceipt evaluate generated totals and complete-string candidate decisions. Their prospective interaction intervals remain unresolved, and full-input candidate accuracy reaches 100%. These different output regimes do not inherit the single-step identity. Appendix AE preserves all original outcomes, including the 194-group prospective study and model-specific operator effects.

AG COMPLETE SAVED-TRACE RECONSTRUCTION FOR RECEIPT SEQUENCES

Scope and provenance. This is a post-hoc analysis of the completed 194-group WildReceipt panel, after its original MAP/EM results were known. The plan was recorded before aggregating reconstructed trie outcomes, with no new model calls or outcome-based exclusions. It includes all 194 source groups, both Qwen3.8-27B and MonkeyOCRv2-B-Und, and every original full/high-norm/DivPrune zero/delete arm. Four candidate strings, ordering, prompts, images and half-retention supports are unchanged. These groups have prior research exposure; normalized store strings do not prove merchant independence. All intervals here are secondary, not replacements for the prospective receipt endpoints.

Why the archive is sufficient for this readout. Every original candidate branch used a fresh cache and recorded its target token's raw conditional log probability before forced-token processing, including the original terminal EOS. Their union contains every edge of the fixed candidate trie. At shared prefix p,

$$\log p(v \mid p) = \ell_v(p) - \log \sum_{u \in V} \exp \ell_u(p).$$

The common normalizer preserves the ordering of admissible next tokens. Thus these saved edge values suffice to traverse the trie under ascending token-ID ties; summing a candidate's edge values instead produces its sequence log probability. This reconstruction applies the existing decoding contract to the archived prefix scores. Raw logits outside those prefixes are unavailable.

Qualification and independent replay. All 388 immutable native-record hashes, completion statuses, model/source manifests and 1,940 cell identities are verified. Prepared pixels, visual features, grid and retained-index hashes agree between free and teacher branches and within each fixed-support operator comparison. Every sequence sum reproduces its stored score. All 4,090 repeated prefix-edge values agree exactly. Each cell's four teacher branches share its original first-logit hash; free first-logit fingerprints use another numeric representation and are not equated bytewise. Instead, recorded top-token decisions and all available pairs of root-edge log probabilities agree with the archived top-logit differences (2,834 distinct-token pairs across 1,124 cells). Archived generation uses greedy decoding without an active rank-changing score processor; the original EOS and 64-token cap remain fixed.

A second implementation eliminates candidate branches step by step without constructing an edge dictionary and reproduces all 1,940 trie completions. Every actual completion inside the finite candidate language equals its reconstructed trie completion; $A = EQ$ holds in all cells. Exact conditional-edge ties occur once for Qwen and 12 times for Monkey; the declared token-ID tie rule is retained. This is numerical qualification of saved traces, not a new native-decoder run or replay of hidden activations.

Table 32: All reconstructed receipt conditions; every row contains 194 images. Actual is original normalized text EM, equal here to canonical-token correctness A. Trie and MAP use the same four original EOS-terminated strings. $R = Q - A$; Disagree is the percentage of trie/MAP string disagreements. Cap counts actual generations reaching 64 tokens; all remain included. Eligibility E equals A in this particular panel, but not by definition.

Model	Arm	Actual	Trie	MAP	R	Disagree	Cap
Qwen27B	Full	85.05	97.42	100.00	12.37	2.58	0
Qwen27B	Norm zero	76.29	93.81	98.97	17.53	5.67	0
Qwen27B	Norm delete	72.68	93.30	98.45	20.62	6.70	0
Qwen27B	Div zero	68.04	91.24	97.42	23.20	7.22	0
Qwen27B	Div delete	65.98	92.27	96.39	26.29	8.25	0
Monkey	Full	72.68	92.27	100.00	19.59	7.73	0
Monkey	Norm zero	67.01	87.63	98.45	20.62	12.89	3
Monkey	Norm delete	67.53	86.60	99.48	19.07	13.40	0
Monkey	Div zero	55.67	87.63	98.97	31.96	12.37	1
Monkey	Div delete	57.73	87.11	98.97	29.38	13.40	0

Observed readout differences. Table 32 includes every arm. Qwen has 59 trie/MAP disagreements across 970 cells and Monkey has 116; these are repeated-image conditions, not 970 independent images. Full input has 5/194 and 15/194 disagreements, respectively. Under zeroing, the mean disagreement rates across selectors are 6.44% [3.87,9.28] and 12.63% [8.51,17.01]. Trie-minus-MAP accuracy is -5.67 [-8.51,-3.09] and -11.08 [-15.21,-7.22] points. Every contrast uses 20,000 paired source-group draws, seed 2026092501, and descriptive 95% intervals. These do not correct for the adaptive research history. A different decision rule does not retroactively make the old confirmation non-saturated.

Exclusion is not automatically a format failure. Here C is the original finite candidate token language, not the set of all syntactically valid amounts. A freely emitted amount can be outside C without violating the output format. We therefore compare canonical A with original normalized text EM: they agree in all 1,940 cells, with zero correct-text/noncanonical cases in this archive. This observed agreement is not a general alias guarantee. There are 59 first-token and 197 later-prefix exits for Qwen, versus 126 and 222 for Monkey; 714/622 completions remain inside C. First-exit location records a decision event, not a causal location of visual-information loss.

The shared zero-minus-delete exclusion shifts are $L = -3.09$ [-5.93,-0.52] for Qwen and $L = 2.06$ [-1.03,5.15] for Monkey. Differential shifts are $J = 0.00$ [-4.64,4.64] and $J = -1.03$ [-7.22,5.67]. Thus the sequence reconstruction does not establish the same positive shared zeroing shift in both receipt models; its supported increment is readout-dependent sequence decisions and an executable finite-language audit. The earlier prospective interaction results remain unchanged.

Figure 4 in the main text illustrates a full-input MonkeyOCR disagreement. The case was selected post hoc by the smallest fixed-prefix SHA256 among all 15 full-input trie/MAP disagreements; it does not change cohort estimates.

Portable artifacts. The supplement contains compact per-cell candidate-token traces, actual output IDs/text, the original chunk hashes, the complete table, a prefix-tree replay and independent candidate-elimination check. They require no model or original filesystem. This reconstructs only the declared finite-candidate decisions; arbitrary new answers, aliases, score processors or prompts require new native measurements.

AH COMPLETE RETENTION-SENSITIVITY AUDIT

Motivation and pre-call specification. The main native-choice panel fixes half retention. A later review raised whether shared exclusion depends on that single setting. We therefore specified 25% and 75% retention before any new model call, reusing the exact original 536-question cohort, question prompts, alternatives, image split and seven conditions. This is an adaptive sensitivity extension after observing the original results, not independent confirmation or a revised primary protocol. The new family jointly specifies $J_{25}, L_{25}, J_{75}, L_{75}$ on the 429-question confirmation partition, using 20,000 paired RGB-image bootstrap draws (405 groups), seed 2026092502, with 98.75% percentile intervals. Every group draw is shared by all four endpoints; the statistic remains question-weighted. The decision criterion requires both new L lower bounds to exceed zero. Original 50% endpoints and their 97.5% family remain unchanged.

Native implementation and complete accounting. The pinned Qwen2.5-VL-7B model, BF16 math-SDPA, layer 2, 576-row cap, author-core MMTok and qualified DivPrune ports are unchanged. Only $K = \max(1, \lfloor rN \rfloor)$ changes, with $r \in \{0.25, 0.75\}$. Text and logical positions are preserved. Each selector computes its support for each budget and reuses exactly that support across zeroing, deletion and masking. Support identities change across budgets, so the experiment does not isolate the number of deleted rows from their identity. The first two native layers keep the full visual input, as in the original panel.

Each budget first passes the same 33 engineering checks on the three previously designated development questions: native/manual logits and argmax, all-keep replay, feature order and DivPrune parity. The full runs then retain all 536 questions (107 development/429 confirmation; 101/405 RGB images), giving 3,752 decisions each and 7,504 new scientific decisions in total. Qualification calls are excluded from that count. The unmodified full-input condition is reexecuted at each budget: all 536 vocabulary-logit hashes and output IDs match the original 50% run exactly in both new panels. All 11,256 archived decisions across the three budgets satisfy $A = EQ$. There are no new question exclusions, model fitting or outcome-selected supports. Wall time is not used as an efficiency result.

Table 33: Development partition, all seven conditions at each budget; scores in percent. Mass is mean probability on accepted answer aliases. Full-input rows are reexecuted controls and retain all tokens. The 50% rows preserve the original results.

Retention	Arm	Actual A	Candidate Q	Excluded R	Eligible	Mass
25%	Full	78.50	78.50	0.00	100.00	99.74
25%	MMTok zero	0.00	17.76	17.76	1.87	0.45
25%	MMTok delete	69.16	69.16	0.00	100.00	99.74
25%	MMTok mask	70.09	70.09	0.00	100.00	99.74
25%	DivPrune zero	0.93	18.69	17.76	2.80	0.77
25%	DivPrune delete	71.96	71.96	0.00	100.00	99.72
25%	DivPrune mask	71.96	71.96	0.00	100.00	99.72
50%	Full	78.50	78.50	0.00	100.00	99.74
50%	MMTok zero	0.00	26.17	26.17	0.93	1.34
50%	MMTok delete	72.90	72.90	0.00	100.00	99.73
50%	MMTok mask	73.83	73.83	0.00	100.00	99.74
50%	DivPrune zero	0.00	25.23	25.23	2.80	2.24
50%	DivPrune delete	74.77	74.77	0.00	100.00	99.73
50%	DivPrune mask	74.77	74.77	0.00	100.00	99.73
75%	Full	78.50	78.50	0.00	100.00	99.74
75%	MMTok zero	6.54	24.30	17.76	14.02	12.43
75%	MMTok delete	76.64	76.64	0.00	100.00	99.71
75%	MMTok mask	76.64	76.64	0.00	100.00	99.72
75%	DivPrune zero	3.74	25.23	21.50	9.35	9.93
75%	DivPrune delete	76.64	76.64	0.00	100.00	99.75
75%	DivPrune mask	75.70	75.70	0.00	100.00	99.73

Table 34: Confirmation partition, all seven conditions at each budget; scores in percent. Mass is mean probability on accepted answer aliases. Full-input rows are reexecuted controls and retain all tokens. The 50% rows preserve the original results.

Retention	Arm	Actual A	Candidate Q	Excluded R	Eligible	Mass
25%	Full	68.30	68.30	0.00	100.00	99.70
25%	MMTok zero	0.47	20.51	20.05	1.63	0.88
25%	MMTok delete	64.80	64.80	0.00	100.00	99.67
25%	MMTok mask	64.57	64.57	0.00	100.00	99.67
25%	DivPrune zero	0.93	22.84	21.91	2.10	0.73
25%	DivPrune delete	65.50	65.50	0.00	100.00	99.68
25%	DivPrune mask	66.20	66.20	0.00	100.00	99.68
50%	Full	68.30	68.30	0.00	100.00	99.70
50%	MMTok zero	0.70	21.91	21.21	2.56	1.87
50%	MMTok delete	67.83	67.83	0.00	100.00	99.68
50%	MMTok mask	67.83	67.83	0.00	100.00	99.68
50%	DivPrune zero	2.10	19.58	17.48	5.59	1.85
50%	DivPrune delete	68.07	68.07	0.00	100.00	99.70
50%	DivPrune mask	67.60	67.60	0.00	100.00	99.69
75%	Full	68.30	68.30	0.00	100.00	99.70
75%	MMTok zero	7.93	27.97	20.05	13.29	12.52
75%	MMTok delete	68.30	68.30	0.00	100.00	99.70
75%	MMTok mask	68.07	68.07	0.00	100.00	99.71
75%	DivPrune zero	7.69	27.04	19.35	14.22	11.98
75%	DivPrune delete	68.07	68.07	0.00	100.00	99.70
75%	DivPrune mask	68.07	68.07	0.00	100.00	99.70

Table 35: Complete new endpoint family. Development is descriptive; no decision is selected from its observed effects. Confidence intervals condition on this reused population and do not correct the full adaptive research history.

Endpoint	Development (pp)	Confirmation (pp)	Confirmation 98.75% interval
J_{25}	0.00	-1.86	[-6.57, 2.75]
L_{25}	17.76	20.98	[16.55, 25.58]
J_{75}	-3.74	0.70	[-4.63, 6.03]
L_{75}	19.63	19.70	[15.74, 23.87]

Interpretation and replay. Both new L intervals exclude zero positively. The observed shared exclusion therefore persists at 25%, 50% and 75% retention in this fixed audit. Neither a constant effect across budgets nor a negligible differential effect is established. In particular, the single-token responses cannot diagnose verbosity or multi-token failure. The label L denotes the average of two exclusion changes, not their per-question intersection. Masking and deletion have respectively 6/7 and 10/7 confirmation output-token disagreements for MMTok/DivPrune at 25% and 75%; these are distinct native operators and all disagreements are retained. Actual accuracy, candidate accuracy and answer-set mass describe different quantities.

The supplement includes the pre-call protocol and binding, complete immutable chunk hashes, qualification results, native status snapshots and all 11,256 compact decision records. A separate portable implementation recomputes every arm mean, all four new image-bootstrap intervals and the joint criterion from the exported cells. It verifies saved decisions and does not claim a second native execution. Existing receipt outcomes and historical failures remain unchanged.

AI COMPLETE INTERVENTION AND READOUT OVERVIEW

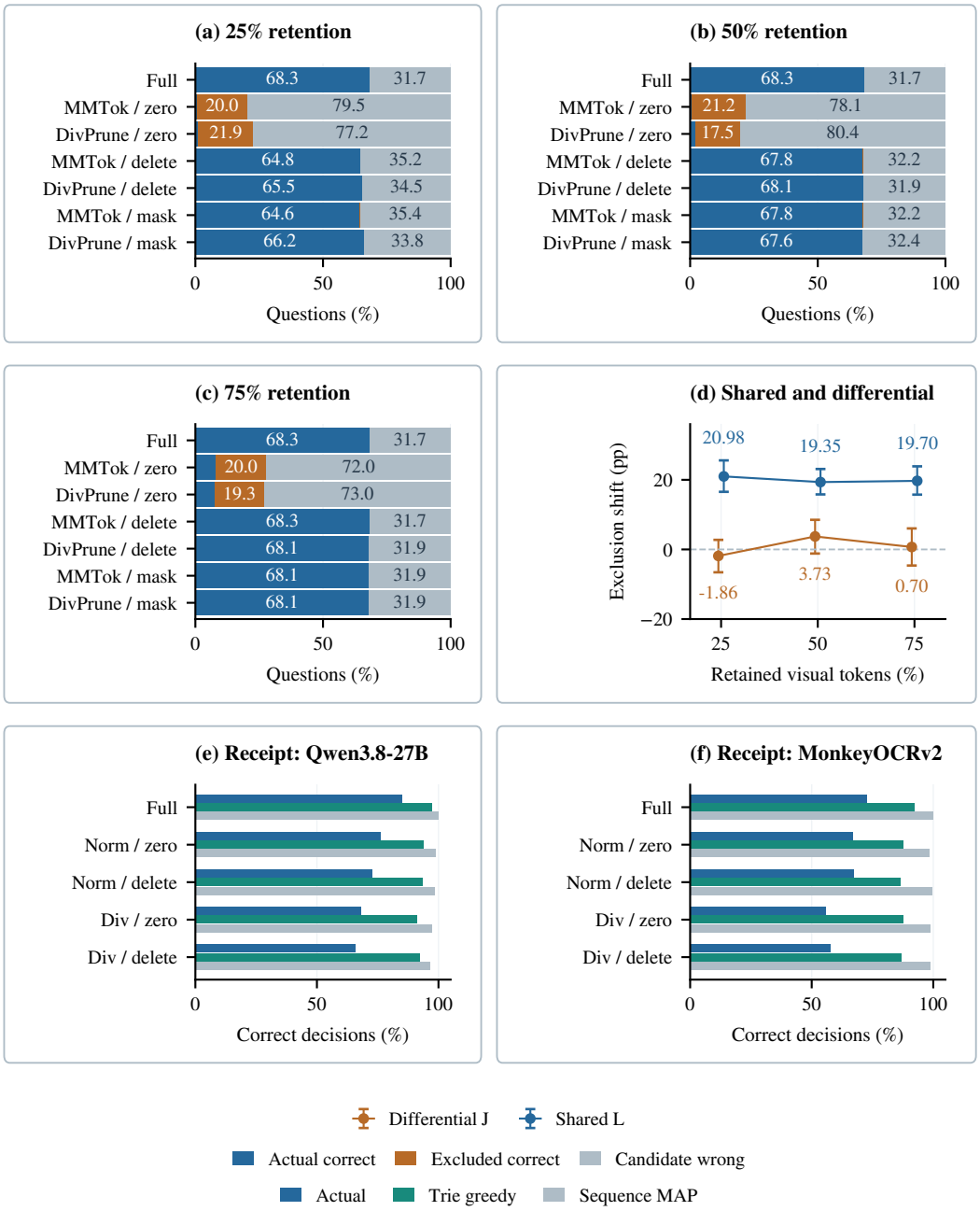


Figure 13: Complete decision accounting. (a–c) All seven native-choice arms on 429 confirmation questions, at every evaluated retention budget; disjoint actual-correct, excluded-correct and candidate-wrong events sum to 100%. (d) Both J and L, with the original 50% two-endpoint 97.5% family and the added 25/75% four-endpoint 98.75% family retained. (e–f) All five receipt arms on 194 groups per model, comparing archived actual correctness, reconstructed trie-greedy decisions and whole-string MAP. The latter two answer rules use the same stored branches but solve different decision problems. All panels preserve the original populations and denominators.