VISION-LANGUAGE MODELS CAN REPORT WHICH IMAGE'S VISUAL-TOKEN ACTIVATIONS WERE PERTURBED

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a vision-language model (VLM) report which of two images had its internal visual-token activations perturbed while the visible inputs themselves remain unchanged? We study this question in three instruction-tuned VLM checkpoints by perturbing decoder-side activations at candidate-associated visual-token positions and asking the model which candidate was affected. Because successful localization could simply reflect damaged visual information, we separately test whether measured visual functions remain intact and use controls to probe simpler explanations. At operating points satisfying both the category criterion and a secondary counting guard, localization rises from 0.499 to 0.688 in Qwen2.5-VL and from 0.505 to 0.656 in Qwen3-VL, with both effects surviving whole-window multiplicity correction. Idefics3 shows a smaller but statistically detectable effect. Norm-restored and crossover controls show that activation-norm inflation and fixed display-side preference are not sufficient explanations, while matched-quality and visible-corruption comparisons bound stronger perturbation-specific interpretations. Finally, in Qwen3-VL, perturbation identity is near-ceiling linearly decodable at both $p = .30$ and $p = .60$, yet the fitted probe directions have sharply different steering efficacy: little targeted steering at $p = .30$ but material bidirectional steering at $p = .60$, where the counting guard no longer holds. We call this phenomenon *perturbation reportability*: candidate-specific internal perturbation information can become behaviorally reportable, while linear availability and steering efficacy remain distinct under the tested intervention procedure.

1 INTRODUCTION

Two images are shown to a vision-language model (VLM). Their pixels remain unchanged, but we perturb the internal visual-token activations associated with one of them: can the model tell which image was affected? By asking this question, we test whether a VLM's own behavior can reveal information about an internal intervention. The distinction matters for interpretability and model monitoring because information may be present in internal representations without being behaviorally accessible or causally useful. We refer to this operational ability as *perturbation reportability*.

Recent work on language models shows that models can identify which of two text candidates had its activations perturbed with dropout or Gaussian noise (Fornasiere et al., 2026), building on differential tests designed to reduce simple output-side confounds (Hahami et al., 2025). We adopt the same candidate-specific setup, perturbation families and forced-choice readout. Extending this test to VLMs, however, is not simply a change of modality. Perturbing one candidate's visual-token activations may make its visual representation less usable, so a correct localization response could reflect degradation rather than access to perturbation-linked information. The visual setting therefore requires us to ask not only whether the model can localize the intervention, but whether it can do so while measured visual function remains intact and simpler cues are ruled out. To our knowledge, prior VLM work has not tested whether a model can behaviorally localize a candidate-specific internal visual perturbation while explicitly checking preservation of measured visual function.

Figure 1 summarizes how we test this question in Qwen2.5-VL, Qwen3-VL and Idefics3-8B-Llama3. Two candidate images are presented and one is selected for intervention. We perturb

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

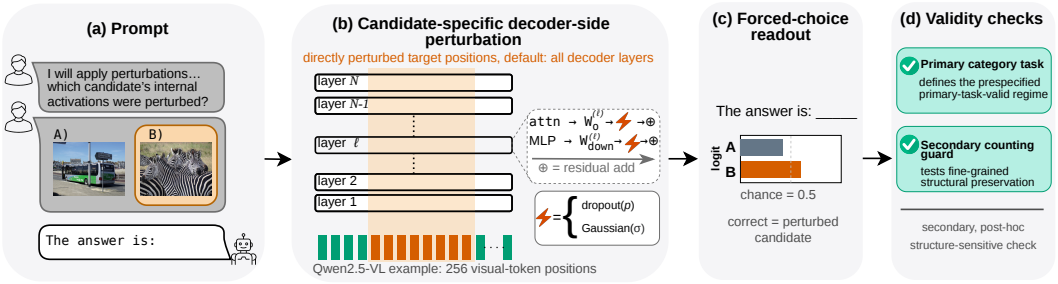

Figure 1: **Candidate-specific visual perturbation reportability protocol.** (a) Two candidate images are presented, with one selected for internal intervention. (b) Decoder-side activations at the selected candidate's visual-token positions are perturbed at the attention output projection and multilayer perceptron (MLP) down-projection before residual addition; input pixels remain unchanged in the intervention experiments. (c) The model reports which candidate was perturbed via a forced-choice A/B readout. (d) Interpretation is conditioned on the prespecified category criterion and, for the Qwen models, a secondary post-hoc counting guard for finer-grained visual preservation.

decoder-side activations only at visual-token positions associated with that candidate, while leaving both input images unchanged. The model then identifies the affected candidate through a forced-choice A/B readout. To interpret the localization result, we evaluate each intervention strength against a prespecified category criterion and, for the Qwen models, a secondary post-hoc counting guard for finer-grained visual preservation.

We next ask what could account for successful localization. Norm-restored interventions test whether activation magnitude provides a sufficient cue, while paired crossover trials test whether the readout follows the perturbed candidate rather than a fixed response side. Matched-quality and visible-corruption controls probe whether the effect can instead be explained by degradation-like changes in representation quality. Finally, we compare behavioral reportability with linear decoding and causal steering to test whether reportable, linearly available and steerable perturbation information coincide.

Contributions. We make three contributions. **(i) Visual perturbation reportability:** we adapt candidate-specific perturbation reporting to VLMs and evaluate it across three instruction-tuned checkpoints, with explicit checks of whether measured visual function remains intact. **(ii) Construct validity:** we test whether localization can be explained by activation-norm changes, fixed response-side preferences, or more general degradation cues using norm-restored, crossover, matched-quality and visible-corruption controls. **(iii) Reportability, availability and steering:** we compare behavioral localization with linear decodability and causal steering in Qwen3-VL, showing that these three forms of access to perturbation-linked information need not coincide.

2 Related Work

Internal-state reporting and controlled activation perturbations. Recent work asks whether language models can report privileged information about their own internal states (Binder et al., 2025; Lindsey, 2026). Simple disturbance-detection tasks, however, can be confounded by output-side effects, motivating differential tests that ask which of multiple candidates was altered rather than whether any disturbance occurred (Hahami et al., 2025). Most closely related, Fornasiere et al. (2026) perturb candidate-specific language-model activations with dropout and Gaussian noise and show that models can identify which of two text candidates was affected. Their text setting, however, leaves open a different interpretive problem in VLMs: localization may arise from reduced visual usability, degradation of finer-grained structure despite preserved coarse recognition or a generic degradation cue rather than perturbation-specific information. These gaps motivate our study of perturbation reportability in VLMs and its relation to linear availability and causal steering.

Decodability, faithfulness and causal control. Behavioral reports and probe accuracy do not by themselves establish faithful or causally relevant access to internal computation. Chain-of-

thought explanations can omit factors that influenced a model's prediction (Turpin et al., 2023), and activation-verbalization methods need not reveal privileged information from the target model's internal state (Li et al., 2025a). Likewise, linearly decodable information need not be used by the model's behavior (Ravichander et al., 2021; Elazar et al., 2021), and recent work directly contrasts linear decodability with causal control in a language-identity setting (Srivastav, 2026). We therefore distinguish behavioral reportability, linear availability and steering efficacy rather than treating successful decoding as evidence of knowledge, awareness or causal use.

Mechanistic analysis and intervention in VLMs. A growing literature analyzes and manipulates internal visual representations in VLMs. Causal-tracing and mechanistic studies localize visual information across token positions, layers and submodules (Palit et al., 2023; Neo et al., 2024; Kaduri et al., 2025; Li et al., 2026; Schaumlöffel et al., 2026), while other work edits or steers internal states to alter visual influence, reduce hallucination, causally probe visual representations or describe steered visual features (Liu et al., 2025; Li et al., 2025c;b; Deng et al., 2026; Ferrando et al., 2026). Related "vision-language introspection" work uses internal signals for hallucination diagnosis and self-correction (Liu et al., 2026). These studies establish external analysis and causal manipulation of visual representations, but not the same differential candidate-specific perturbation-reportability setting with explicit visual-function and construct-boundary controls.

3 METHOD

3.1 CANDIDATE-SPECIFIC PERTURBATION REPORTABILITY

Each trial presents two candidates, A and B, of the same modality, with exactly one selected for an internal activation intervention (Figure 1). In the core visual experiments, the input pixels remain unchanged: only decoder-side activations associated with the selected image are perturbed. The model is then asked which candidate was perturbed, and we score a forced-choice next-token A/B readout. Let $a(s)$ denote localization accuracy at perturbation strength s; chance performance is 0.5.

We use *candidate-specific perturbation reportability* purely as an operational term for this intervention-to-readout relationship. Throughout, "report" refers only to the forced-choice next-token readout, not to open-ended verbal description. It does not imply perturbation-specific semantics, awareness or a dedicated self-monitoring mechanism. The criteria used to evaluate visual-function preservation are defined in Section 3.3.

3.2 CANDIDATE-SPECIFIC ACTIVATION INTERVENTIONS

Let T_c denote the decoder positions associated with candidate $c \in \{A, B\}$, and let y be the candidate selected for intervention. For decoder layers $\mathcal{L}$, we intervene on the outputs of the self-attention output projection $W_O^{(\ell)}$ and MLP down-projection $W_{\text{down}}^{(\ell)}$, immediately before their residual additions (Figure 1). Let $h_{\ell,t}^{(u)} \in \mathbb{R}^d$ denote either projected sub-block output, with $u \in \{\text{attn}, \text{mlp}\}$. For $\ell \in \mathcal{L}$ and $t \in T_y$, we apply either inverted dropout or additive Gaussian noise, with $0 \le p < 1$, $\sigma \ge 0$, and $j = 1, \ldots, d$:

$$\widetilde{h}_{\ell,t}^{(u)} = \begin{cases} \dfrac{m_{\ell,t}^{(u)} \odot h_{\ell,t}^{(u)}}{1-p}, & m_{\ell,t,j}^{(u)} \sim \text{Bernoulli}(1-p), \\ h_{\ell,t}^{(u)} + \epsilon_{\ell,t}^{(u)}, & \epsilon_{\ell,t}^{(u)} \sim \mathcal{N}(0, \sigma^2 I_d). \end{cases} \tag{1}$$

Here $m_{\ell,t}^{(u)} \in \{0,1\}^d$. Unless otherwise stated, $\mathcal{L}$ contains all decoder layers. Only positions associated with the selected candidate are directly modified; the intervention does not directly modify the vision encoder, multimodal projector, non-target decoder positions or input pixels. Masks and noise are sampled independently across targeted positions, layers and sub-blocks and are re-seeded per trial. Runtime checks verify intervention confinement and intended target coverage; model-specific hook mappings and targeting verification details are reported in Appendices B and C respectively.

To test whether localization requires perturbation-induced norm changes, we additionally use norm-restored variants. Given an original vector h and its perturbed counterpart b, we rescale

$$\widetilde{h}_{\mathrm{NR}} = \|h\|_2 \frac{b}{\|b\|_2 + \varepsilon}. \tag{2}$$

Here $\varepsilon = 10^{-8}$; if $\|b\|_2 = 0$, the implementation returns h unchanged. Otherwise, this restores the original L_2 norm up to numerical tolerance while retaining the perturbed direction, allowing us to test whether activation-norm inflation is necessary for reportability. Verification and implementation details are provided in Appendix G.

3.3 Tasks and Visual-Function Criteria

Each trial contains two independent labels: which candidate is perturbed and which candidate is correct on a content-based task. For vision, we use COCO val2017 image pairs (Lin et al., 2014) with disjoint annotated category sets and sample perturbation side independently of the content label. We evaluate with 20 frozen paraphrased prompt templates and a single forward pass, aggregating next-token logits over tokenizer-valid realizations of A and B. For each question, the higher-scoring candidate defines the forced-choice prediction; exact score ties are assigned deterministically to A. The prediction is evaluated against the corresponding localization or content label. Additional prompt, dataset and text-anchor details are given in Appendix A.

Our prespecified primary criterion uses image-category accuracy. Let $\mathcal{G}$ denote the finite tested strength grid and let $q_{\mathrm{cat}}(s)$ denote category accuracy at strength s, with $q_{\mathrm{cat}}(0) > 0$. We define

$$\mathcal{S}_{\mathrm{primary}} = \left\{ s \in \mathcal{G} : \frac{q_{\mathrm{cat}}(s)}{q_{\mathrm{cat}}(0)} \geq 0.95 \right\}. \tag{3}$$

This point-ratio rule defines the primary-valid intervention set but does not imply preservation of all visual capabilities. For localization inference, $\mathcal{S}_{\mathrm{primary}}$ is held fixed across permutations: category correctness is defined by the independently sampled content label, whereas the randomization test relabels only the perturbed candidate. Thus, family membership is invariant to the localization test's perturbation-side sign flips. Uncertainty for category and counting ratios is reported in Appendix D.

Because category recognition is coarse, we additionally evaluate a secondary, post-hoc structure-sensitive guard using COCO counting. For the Qwen models, we call a regime *dual-criterion-valid* when both the primary category ratio and the analogous counting ratio are at least 0.95. Counting was not measured for Idefics3, so its regimes are described only with respect to the prespecified primary criterion.

Before interpreting any visual perturbation sweep, we also require a real/scrambled/blank prerequisite gate confirming reliance on genuine visual content: the real–blank accuracy gap must be at least 0.15 with exact McNemar $p < 0.01$, while blank performance must be equivalent to chance within a prespecified ± 0.05 margin. Gate statistics and counting-task construction appear in the appendix.

3.4 Construct-Validity Controls

We use four controls to test alternative explanations for candidate-specific reportability. First, the norm-restored intervention in Eq. 2 tests whether localization requires perturbation-induced activation-norm inflation.

Second, paired crossover trials hold candidate identities, display order, prompt, strength and perturbation randomness fixed while switching only which candidate is perturbed. We measure the corresponding paired shift in the A-versus-B localization margin and repeat the crossover after swapping candidate display order. A shift toward the newly perturbed candidate is inconsistent with a fixed display-side preference, while the order-swapped replication tests whether the effect depends on a particular content-to-slot assignment.

Third, we repeat matched trials with an alternative question asking which candidate's internal representation is less coherent. Candidate pair, order, perturbed side, intervention strength, hooks and perturbation realization are held fixed; only the question family changes. Because only the perturbation question explicitly announces that an intervention occurred, differences between the two questions are not interpreted as a clean test of perturbation-specific semantics.

Finally, we compare internal perturbations with visible input corruption using Gaussian blur, with patch scrambling included in the full appendix analysis. For the primary frontier comparison, both arms use the matched representation-quality question. We match observed candidate-identification accuracy using a frozen ± 2 percentage-point tolerance and compare questioned-image counting accuracy as the functional-cost measure. Targets without an observed match are treated as non-computable rather than interpolated or extrapolated. Visible corruption modifies the input image, whereas the internal intervention leaves its pixels unchanged. Equal-position ($K = 15$) and decoder-depth restrictions are reported as supplementary characterization in the appendix.

3.5 Linear Availability and Causal Interventions

We distinguish linear availability from steering efficacy of a fitted probe direction. We train an L_2-regularized logistic probe on the final answer-position residual to predict the perturbed candidate, using item-grouped cross-validation with train-fold-only standardization and nested selection of the regularization strength. Unperturbed ($p = 0$) trials provide a negative control; probe accuracy is interpreted only as linear decodability, not knowledge, awareness or causal use. Full fitting details are given in Appendix K.

For steering, we refit the same probe on the full training population and convert its standardized coefficient w to a raw-residual direction $d = w/\sigma$, where σ is the fitted feature scale, then normalize $\hat{d} = d/\|d\|_2$. At the final decoder answer position we intervene as

$$h' = h + s\,\alpha s_L \hat{d}, \qquad s_L = \sqrt{\mathbb{E}\Big[(\hat{d}^\top h)^2\Big]},$$

where $s \in \{-1, +1\}$ selects the target candidate and α controls intervention strength. Fresh item-disjoint trials are compared with per-trial random unit directions matched to the same injection norm. We report changes in the A/B margin, flips relative to each trial's own unsteered prediction, and normalized displacement $\|\Delta h\|_2/\|h_{\text{base}}\|_2$ for $\|h_{\text{base}}\|_2 > 0$. These experiments test sufficiency of the fitted direction under this intervention procedure, not necessity or use of all linearly available information; complete controls and definitions are in Appendix L.

3.6 Models and Statistical Analysis

Our primary visual experiments use Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a) and Idefics3-8B-Llama3 (Laurençon et al., 2024). Unless otherwise stated, inference uses bfloat16 and $n = 1000$ trials per perturbation strength. Primary sweeps use seed 42; the Qwen headline sweeps are independently repeated with seed 43 using resampled trials and perturbations. Exact checkpoints and key experimental configurations are reported in the appendix; complete strength grids, software versions and environment details are provided in the anonymous code artifact.

For binomial accuracies, we report 95% Wilson confidence intervals; paired or item-dependent contrasts use clustered bootstrap intervals at the corresponding experimental unit. For localization, let $\hat{a}(s)$ denote forced-choice accuracy. We define

$$\Delta(s) = \hat{a}(s) - \hat{a}(0), \qquad T = \max_{s \in \mathcal{S}_{\text{primary}},\, s>0} \Delta(s),$$

where T is tested using $B = 10{,}000$ candidate-pair-level perturbation-side sign-flip permutations. A single flip is applied consistently to all occurrences of a pair across available cells, including $s = 0$, and the one-sided add-one value is $p_{\max} = (1 + \#\{T^{(b)} \geq T_{\text{obs}}\})/(B+1)$. The 95% simultaneous threshold c^* is the corresponding 0.95 quantile of the null maxima. Full resampling and interval construction, including the differing Qwen and Idefics3 reuse structures, are given in Appendix N.

4 Results

4.1 Candidate-Specific Perturbations Are Reportable Under Measured Visual-Function Criteria

All three tested checkpoints pass the prerequisite visual-reliance gate before perturbation localization is interpreted (Appendix E). Figure 2 shows the visual-token dropout sweeps together with the prespecified category criterion and the secondary, post-hoc counting guard.

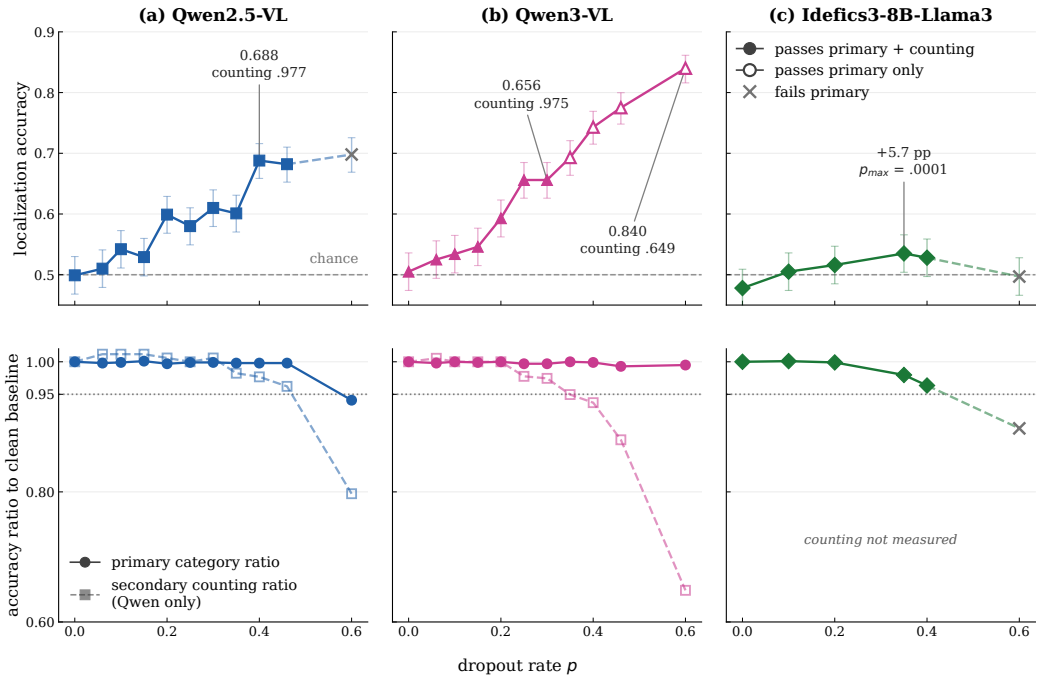

Figure 2: **Visual perturbation localization and visual-function criteria.** Top panels show localization accuracy under all-layer visual-token dropout; bottom panels show accuracy relative to the clean baseline for the prespecified category task and, for the Qwen models, the secondary questioned-image counting task. Chance localization is 0.5; the horizontal 0.95 line marks the criterion threshold. Filled Qwen markers pass both criteria, open markers pass the primary criterion only, and crosses fail the primary criterion. Idefics3 is assessed only with the primary criterion because counting was not measured for that model.

Table 1: **Main visual-dropout regimes used in subsequent analyses.** Counting entries show clean $\rightarrow$ perturbed questioned-image accuracy, with the corresponding ratio in brackets. The category criterion is prespecified; counting is secondary, post-hoc and was not measured for Idefics3.

Model	p	Localization	Category	Counting $q_0 \rightarrow q_p$ [ratio]	Regime
Qwen2.5-VL	.40	.688	Pass	.891→.870 [.977]	Both pass
Qwen3-VL	.30	.656	Pass	.964→.939 [.975]	Both pass
Qwen3-VL	.60	.840	Pass	.964→.625 [.649]	Counting fails
Idefics3-8B-Llama3	.35	.535	Pass	–	Primary only

Both Qwen models show candidate-specific reportability before the secondary counting guard breaks. Qwen2.5-VL rises from 0.499 localization accuracy at $p = 0$ to 0.688 at $p = .40$ (+18.9 pp), while Qwen3-VL rises from 0.505 to 0.656 at $p = .30$ (+15.1 pp). Both operating points satisfy the category and counting criteria; the corresponding absolute counting accuracies are given in Table 1. Across the primary category-valid strength families, both Qwen sweeps give $p_{\max} < 10^{-4}$ with a simultaneous threshold of $c^* = +5.5$ pp, which both headline changes exceed.

Reportability continues to increase in Qwen3-VL beyond the regime that preserves the secondary counting criterion. At $p = .60$, localization reaches 0.840, while questioned-image counting falls to a ratio of 0.649. The prespecified category criterion still passes, but this stronger operating point lies outside the counting-preserving regime.

The phenomenon is also detectable in Idefics3-8B-Llama3, but at substantially smaller magnitude: localization rises from 0.478 at baseline to 0.535 at $p = .35$ (+5.7 pp). Its whole-window test gives $p_{\max} = 0.0001$ and $c^* = +2.4$ pp, with all four primary-valid nonzero changes exceeding

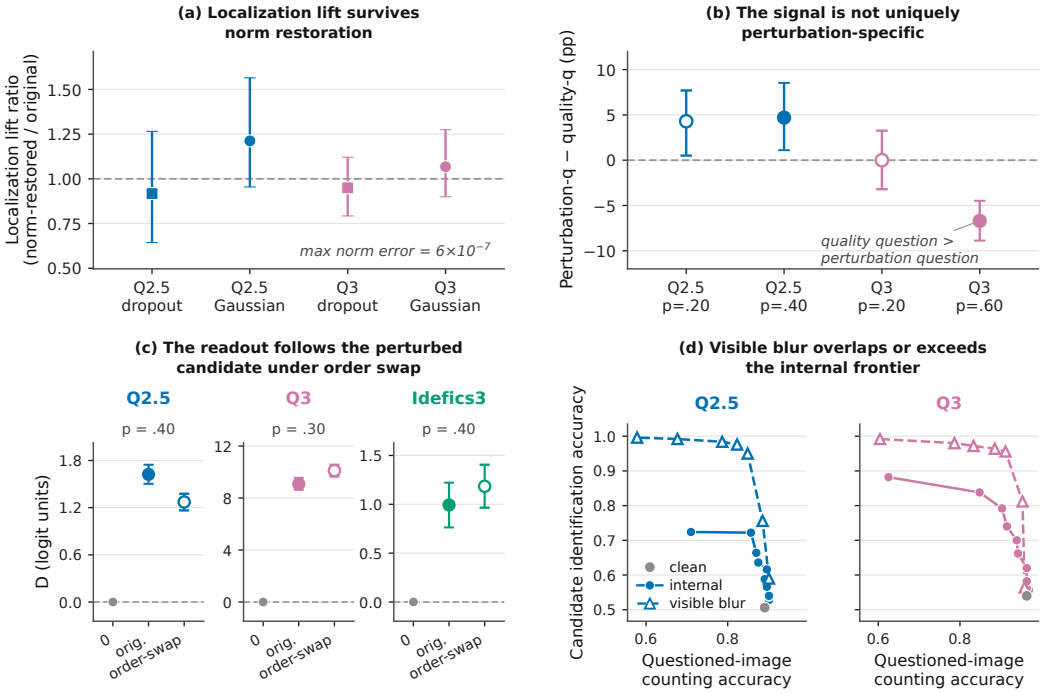

Figure 3: **Construct-validity controls and interpretation boundary.** (a) Ratio of the localization change from $p = 0$ under norm-restored versus original perturbations. (b) Paired accuracy difference between the perturbation question and a matched representation-quality question. (c) Paired crossover contrasts when the perturbed candidate is switched, before and after swapping display order; $D = 0$ exactly at $p = 0$. (d) Candidate identification versus questioned-image counting accuracy for internal perturbation and visible Gaussian blur. Crossover values are in model-specific logit units; full norm-restoration and visible-corruption results are reported in the appendix.

the simultaneous threshold. Thus, candidate-specific perturbation reportability is detectable across all three tested checkpoints; in both Qwen checkpoints, it is already present at operating points satisfying both measured visual-function criteria.

4.2 Controls Bound the Interpretation of the Candidate-Specific Signal

Having established perturbation reportability, we next ask what information supports it and which simpler explanations can be excluded (Figure 3).

Norm inflation and fixed-side preference are not sufficient. Restoring each perturbed vector to its original L_2 norm preserves comparable localization changes across the tested Qwen conditions, with norm-restored/original ratios ranging from 0.92 to 1.21 and maximum relative norm error 6×10^{-7}. Thus, activation-norm inflation is not required for the observed reportability. The paired crossover control further shows that the readout follows the perturbed candidate rather than a fixed display side: at Qwen3-VL $p = .30$, $D = 9.09\,[8.63, 9.56]$ with 98.2% of pairs positive, and remains positive after swapping display order ($D = 10.11\,[9.66, 10.56]$); the $p = 0$ control is exactly zero. Qwen2.5-VL and Idefics3 show the same qualitative pattern (Appendix F).

Perturbation-specificity is not established. On matched trials, Qwen2.5-VL favors the perturbation question over a representation-quality question by 4.3 pp at $p = .20$ and 4.7 pp at $p = .40$. Qwen3-VL shows no difference at $p = .20$, while at $p = .60$ the quality question performs better (0.887 vs. 0.820). Because a matched quality judgment can perform similarly or better, these results do not support interpreting the accessible information as uniquely encoding perturbation identity.

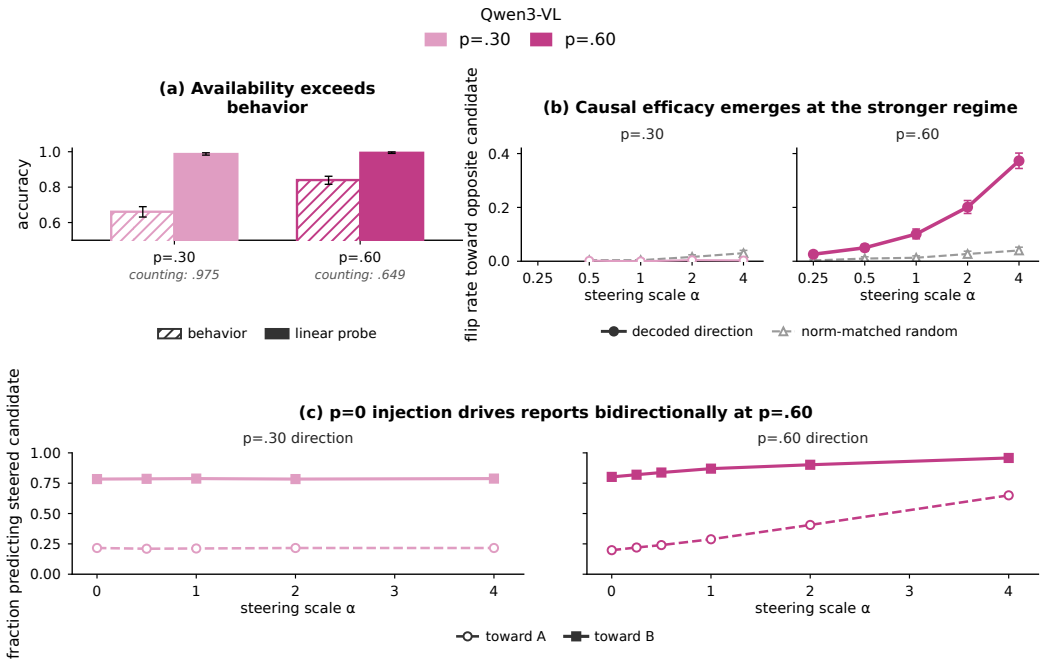

Figure 4: **Linear availability and steering efficacy differ across tested regimes.** (a) Perturbation identity is nearly perfectly linearly decodable at both $p = .30$ and $p = .60$, while behavioral localization remains lower; the counting guard passes at $p = .30$ but fails at $p = .60$. (b) The fitted $p = .30$ direction produces little targeted steering relative to norm-matched random directions at the tested scales, whereas the fitted $p = .60$ direction produces dose-dependent opposite-candidate flips. (c) With no perturbation applied ($p = 0$), injecting the fitted $p = .60$ direction steers the readout toward either candidate depending on sign, while the fitted $p = .30$ direction produces little change. These interventions establish sufficiency of the fitted $p = .60$ direction under the tested procedure, not necessity.

Visible corruption defines a further interpretation boundary. We compare internal perturbation with visible Gaussian blur at matched candidate-identification accuracy under the frozen ± 2 pp rule. Thirteen of sixteen prespecified contrasts have no measured match and are therefore left non-computable rather than interpolated or extrapolated (Appendix H). Among the computable blur contrasts, Qwen2.5-VL is consistent with overlapping frontiers (Δcounting $= -0.010$, 95% CI $[-0.037, 0.016]$), while for Qwen3-VL visible blur preserves more counting accuracy at comparable identification (Δcounting $= -0.052$, $[-0.089, -0.016]$).

Together, these controls rule out activation-norm inflation and fixed display-side preference as sufficient explanations, but do not support the stronger claim that the reportable signal is uniquely perturbation-specific or separable from degradation-like representational consequences.

4.3 Linear Availability and Steering Efficacy Differ Across Tested Regimes

Perturbation identity is nearly perfectly linearly decodable in Qwen3-VL at both $p = .30$ and $p = .60$, despite substantially lower behavioral localization. These regimes differ in visual-function preservation: both category and counting criteria pass at $p = .30$, whereas the counting guard fails at $p = .60$ (Table 1). On the probe evaluation population, decoding accuracy is 0.988 at $p = .30$, where behavioral localization is 0.661 and the counting ratio remains 0.975. We therefore ask whether similarly high linear availability yields comparable steering efficacy of the independently fitted probe directions (Figure 4).

Despite similar probe accuracy, the fitted directions have sharply different steering effects. At $p = .30$, little targeted steering is detected relative to norm-matched random directions at the tested

scales. At $p = .60$, steering along the fitted direction exceeds the random control by 8.8 pp at $\alpha = 1$ and 17.4 pp at $\alpha = 2$. The contrast cannot be reduced simply to lower probe accuracy or smaller injected displacement: $p = .30, \alpha = 4$ produces a larger normalized residual displacement than $p = .60, \alpha = 2$ (0.148 vs. 0.114), yet yields only 0.3% versus 20.1% targeted flips. The fitted directions also have little overlap with the direct A/B answer-logit direction ($|\cos| < 0.059$). This displacement comparison does not establish geometric equivalence between the independently fitted directions or rule out estimator- or scaling-dependent effects.

The fitted $p = .60$ direction is also sufficient to steer the readout without an underlying perturbation. At $p = 0$, steering toward B increases B reports from 0.802 to 0.958, while reversing the direction increases A reports from 0.198 to 0.650 at the largest tested scale; the fitted $p = .30$ direction produces little change.

Thus, near-ceiling linear decodability is accompanied by sharply different steering efficacy of the fitted probe directions in this Qwen3-VL comparison. Material bidirectional steering is observed for the fitted $p = .60$ direction, where the secondary counting guard has already failed (0.649). These results do not establish native use or necessity of the decoded direction, nor do they rule out other causally useful directions at $p = .30$; additional controls are reported in Appendix L.

5 Discussion and Limitations

Across the tested checkpoints, perturbation reportability varies in strength but remains after norm restoration and in the paired crossover tests. This makes activation-norm inflation and a fixed display-side preference unlikely to explain the effect on their own. The matched-quality and visible-corruption controls point to a more cautious interpretation: the model may be reporting a change in representational state that is also accessible through degradation-like cues. We therefore do not interpret the effect as evidence for a dedicated introspective mechanism.

Qwen3-VL also shows a clear difference between linear availability and steering efficacy. Perturbation identity is nearly perfectly decodable at both $p = .30$ and $p = .60$, but the fitted directions behave very differently when used for steering. The $p = .30$ direction has little targeted effect, whereas the $p = .60$ direction supports bidirectional control. Notably, this stronger effect appears only after the secondary counting guard has failed. The result shows that high linear decodability alone does not guarantee an equally effective steering direction. It does not establish native use or necessity, and other causally useful directions may still exist at $p = .30$.

There are several limits to what we can conclude. The counting guard is secondary and post-hoc and was evaluated only for the Qwen models. Idefics3 shows a smaller effect and stronger answer-side bias, the perturbations are artificial and we test only a small set of models, visual-function criteria and fitted directions. Future work should test broader architectures and visual abilities, study naturally occurring internal errors and ask when perturbation-linked information becomes steerable while measured visual function is still preserved.

6 Conclusion

We introduced *perturbation reportability* to test whether a VLM can behaviorally report which candidate's internal visual-token activations were perturbed while the visible inputs remain unchanged. Across three tested checkpoints, candidate-specific reportability is detectable, with both Qwen models showing substantial localization at operating points satisfying the measured category criterion and secondary counting guard. Construct-validity controls rule out activation-norm inflation and fixed display-side preference as sufficient explanations, while matched-quality and visible-corruption comparisons bound stronger perturbation-specific interpretations.

Our results also separate behavioral reportability, linear availability and steering efficacy. In Qwen3-VL, perturbation identity is nearly perfectly linearly decodable in two regimes, yet the fitted probe directions show sharply different steering efficacy; material bidirectional steering appears only in the stronger regime, where the counting guard has failed. This distinction matters for VLM interpretability: reporting internal-state information, decoding it from activations and steering behavior along a fitted direction are related but non-equivalent empirical questions.

AI USE STATEMENT

Generative AI tools were used in this work to assist with software development, experimental-design review, interpretation of experimental results and manuscript preparation. AI tools provided feedback on experimental design, suggested candidate controls and robustness checks and helped stress-test proposed interpretations and alternative explanations. The research question, final methodological design, experimental protocol, choice of controls and scientific conclusions were determined by the authors, who independently evaluated all AI-assisted suggestions and decided which experiments and analyses to conduct.

AI-assisted coding tools were used to help implement, edit and debug portions of the experimental pipeline. All AI-assisted code was subsequently reviewed, tested and executed by the authors, including checks of intervention targeting and experimental outputs; all reported results were obtained from author-run experiments.

Generative AI tools were additionally used to assist with drafting and editing portions of the manuscript, improving clarity and organization, identifying potentially relevant literature and checking or formatting references. The authors independently consulted the cited sources and verified the numerical results, statistical claims and scientific interpretations against the underlying experimental outputs.

The authors reviewed all AI-assisted material and take responsibility for the final content of this work, including its text, claims, code and other artifacts.

ETHICS STATEMENT

This work studies the internal behavior of vision-language models under controlled activation interventions and does not involve human-subject experiments or the collection of new personal data. The experiments use publicly available datasets, including COCO and WikiText-103, solely for model evaluation and controlled stimulus construction. We do not release new person-level annotations or sensitive information.

The methods developed here are intended to improve understanding of how models represent and report information about their internal states. While such techniques may contribute to future work on model monitoring, interpretability and reliability, they could also be misinterpreted as evidence of stronger forms of self-awareness or introspection than the experiments support. We therefore deliberately restrict our claims to *perturbation reportability* and distinguish behavioral reporting from internal availability, causal use, and dedicated introspective mechanisms. The authors take responsibility for the ethical presentation and interpretation of the results.

REPRODUCIBILITY STATEMENT

All experiments were run on NVIDIA A40 GPUs. The Method section specifies the intervention protocol, validity criteria and statistical procedures, while the appendix provides dataset and prompt construction, exact model checkpoints and hook mappings, sampling and resampling procedures, probe and steering specifications and supporting robustness analyses. We additionally submit an anonymous supplementary code archive containing the implementation, experiment configurations, evaluation scripts, frozen reference results and instructions for regenerating the numerical results reported in the main paper and the supporting analyses used to audit those claims. Random seeds and experimental settings are fixed where applicable, and the archive documents the required data organization and model revisions. Upon acceptance, we plan to make the code publicly available to facilitate independent replication and further research.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang

Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. QWEN3-VL Technical Report, 11 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. QWen2.5-VL Technical Report, 2 2025b. URL https://arxiv.org/abs/2502.13923.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: language models can learn about themselves by introspection, 5 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0a6059857ae5c82ea9726ee9282a7145-Abstract-Conference.html.

Zehao Deng, Tianjie Ju, Zheng Wu, Liangbo He, Jun Lan, Huijia Zhu, Weiqiang Wang, and Zhuosheng Zhang. Causal probing for internal visual representations in multimodal large language models, 5 2026. URL https://arxiv.org/abs/2605.05593.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Javier Ferrando, Enrique Lopez-Cuena, Pablo Agustin Martin-Torres, Daniel Hinjos, Anna Arias-Duart, and Dario Garcia-Gasulla. Language models can explain visual features via steering, 3 2026. URL https://arxiv.org/abs/2603.22593.

Damiano Fornasiere, Mirko Bronzi, Spencer Kitts, Alessandro Palmas, Yoshua Bengio, and Oliver Richardson. Language models recognize dropout and Gaussian noise applied to their activations, 4 2026. URL https://arxiv.org/abs/2604.17465.

Ely Hahami, Ishaan Sinha, Lavik Jain, Josh Kaplan, and Jon Hahami. Detecting the Disturbance: A Nuanced view of introspective abilities in LLMs, 12 2025. URL https://arxiv.org/abs/2512.12411.

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

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 8 2024. URL https://arxiv.org/abs/2408.12637.

Millicent Li, Alberto Mario Ceballos Arroyo, Giordano Rogers, Naomi Saphra, and Byron C. Wallace. Do activation verbalization methods convey privileged information?, 9 2025a. URL https://arxiv.org/abs/2509.13316.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: mechanistic interpretability and hallucination mitigation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37): 31645–31653, 3 2026. doi: 10.1609/aaai.v40i37.40431. URL https://ojs.aaai.org/index.php/AAAI/article/view/40431.

Qing Li, Jiahui Geng, Derui Zhu, Zongxiong Chen, Kun Song, Lei Ma, and Fakhri Karray. Internal Activation Revision: Safeguarding Vision Language models without Parameter update. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27428–27436, 4 2025b. doi: 10.1609/aaai.v39i26.34954. URL https://ojs.aaai.org/index.php/AAAI/article/view/34954.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 13–19 Jul 2025c. URL https://proceedings.mlr.press/v267/li25ca.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Jack Lindsey. Emergent introspective awareness in large language models, 1 2026. URL https://arxiv.org/abs/2601.01828.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b4008025c2182bfe16fcc8566ee14d64-Paper-Conference.pdf.

Shuliang Liu, Songbo Yang, Dong Fang, Sihang Jia, Yuqi Tang, Lingfeng Su, Ruoshui Peng, Yibo Yan, Xin Zou, and Xuming Hu. Vision-language introspection: Mitigating overconfident hallucinations in MLLMs via interpretable bi-causal steering. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38518–38543, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1784. URL https://aclanthology.org/2026.acl-long.1784/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in Vision-Language models, 10 2024. URL https://arxiv.org/abs/2410.07149.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards Vision-Language Mechanistic Interpretability: a Causal Tracing Tool for BLIP, 8 2023. URL https://arxiv.org/abs/2308.14179.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Timothy Schaumlöffel, Martina G. Vilas, and Gemma Roig. Mechanisms of object localization in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31356–31365, June 2026.

Arnav Srivastav. Steering the language axis: from linear decodability to causal control, 6 2026. URL https://arxiv.org/abs/2608.12334.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

A PROMPTS, TASKS AND DATASETS

Figure 5 illustrates the common candidate-specific protocol for the supplementary text anchor and the primary visual setting. Each example presents candidates A and B, exactly one of which is selected independently for an internal intervention. Evaluation uses a single forward pass and a forced-choice next-token readout rather than free-form generation. For each question, we aggregate logits over the frozen tokenizer-valid realizations of A and B and select the candidate with the larger score. We use 20 frozen paraphrased templates per task family; candidate formatting, answer instruction, and answer-token sets are fixed before evaluating perturbation strengths.

Visual data and task construction. Visual trials are constructed from COCO val2017 (Lin et al., 2014). Candidate images are paired only when their annotated category sets are disjoint, so the content-task label can be determined independently of the perturbed side. The intervention side is then sampled independently of the content label. The prespecified primary visual task is category discrimination. Because this task is intentionally coarse, selected Qwen conditions are additionally evaluated with the secondary, post-hoc COCO counting guard described in Section 3.3. Counting was not evaluated for Idefics3.

Sampling and reuse. The primary seed-42 Qwen dropout sweeps use 1000 trials at each of eleven strengths and share the same frozen sampling stream across the two Qwen checkpoints. Across the complete sweep, the saved records contain 1993 distinct candidate-image pairs comprising 3986 distinct image identifiers; individual strength cells contain 767–810 distinct pairs because some pairs are sampled more than once with different templates. The 20 localization templates are sampled within this trial construction. Strength cells are sampled independently rather than from a fully crossed item set. The independent seed-43 Qwen sweep contains 1994 distinct pairs (3988 image identifiers), with only one pair shared with the seed-42 pair union. Idefics3 instead uses a fully crossed set of 781 candidate pairs across its six tested strengths. Accordingly, statistical resampling and permutation procedures cluster at the candidate-pair level rather than treating 1000 trials as 1000 independent items.

Counting-guard construction. The secondary counting guard uses a frozen two-image COCO task asking whether a specified image contains more than k instances of a named object category. Ground-truth counts are derived from COCO instance annotations. Items are constructed so that positive cases contain at least $k + 2$ annotated instances of the queried category, providing a margin from the decision boundary, while the other candidate image contains no annotated instance of that category. The frozen item set is counterbalanced over yes/no answer, count threshold k, candidate display side and queried-image side. The model answers the binary judgment through the same fixed forced-choice readout protocol used elsewhere.

For intervention analyses, the *questioned image* is the candidate named in the counting question and the *distractor image* is the other candidate. We evaluate the two cases separately according to whether the intervention targets the questioned or distractor image. At strength s, the counting guard is the questioned-image counting accuracy relative to its corresponding clean baseline, $q_{\mathrm{count}}(s)/q_{\mathrm{count}}(0)$, and passes when this ratio is at least 0.95. This criterion is secondary and post-hoc and is evaluated only for the Qwen models.

Question families. The localization templates ask which candidate's internal activations were perturbed; a representative wording is *"I will apply perturbations to one candidate's internal activations. Which candidate's internal activations were perturbed?"* The matched construct-validity control instead asks which candidate's internal representation is less coherent. Candidate pair, display order, perturbed side, intervention settings, and perturbation realization are held fixed in the matched comparison; only the question family changes. The complete frozen paraphrase banks are included in the submitted code artifact.

Supplementary text anchor. The text arm is used only as an auxiliary anchor and does not support the primary VLM claims. Text candidates are drawn from WikiText-103 (Merity et al., 2017), using the frozen sentence pool and the same candidate-specific forced-choice intervention protocol shown in Figure 5. Its purpose is to verify that the implementation recovers the expected candidate-

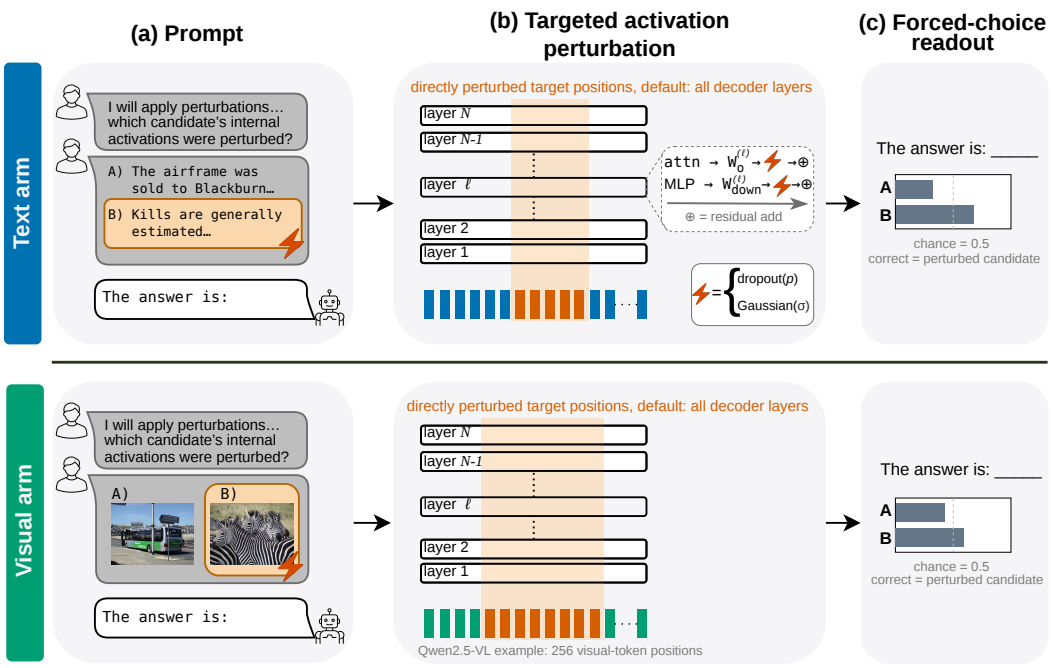

Figure 5: **Candidate-specific text and visual task examples.** The same forced-choice protocol is used for the supplementary text anchor (top) and primary visual experiments (bottom). Exactly one candidate's associated decoder-side activations are targeted with dropout or Gaussian noise, while the readout asks which candidate was affected. In the visual arm, the input pixels themselves are unchanged by the internal intervention.

Table 2: **Model checkpoints used in the primary visual experiments.** Exact Hugging Face snapshot revisions are reported to make the evaluated checkpoints immutable and reproducible.

Checkpoint	Frozen revision
`Qwen/Qwen2.5-VL-7B-Instruct`	`cc594898137f460bfe9f0759e9844b3ce807cfb5`
`Qwen/Qwen3-VL-8B-Instruct`	`0c351dd01ed87e9c1b53cbc748cba10e6187ff3b`
`HuggingFaceM4/Idefics3-8B-Llama3`	`fddb4ff79181e55a994674777e06cd5456ce3dc3`

localization behavior in a text setting closely related to prior work; visual and text results are compared further in Appendix J.

B Model Configurations and Hook Mapping

Model checkpoints. The primary visual experiments use the official instruction-tuned Hugging Face checkpoints listed in Table 2. Inference uses bfloat16 unless otherwise stated.

Decoder-side intervention mapping. Across all three models, interventions are applied inside the language decoder to the output of the self-attention output projection and the MLP down-projection, immediately before residual addition. For Qwen2.5-VL and Qwen3-VL, decoder layers are resolved from `model.language_model.layers`; for Idefics3, the corresponding Llama decoder stack is resolved from the loaded model wrapper at runtime. The targeted submodules are summarized in Table 3.

Candidate-specific visual positions. For each two-image prompt, visual-token positions are recovered from the processed input sequence and partitioned into the contiguous runs corresponding to candidates A and B. Only the span associated with the selected candidate is directly modified. Runtime checks require the recovered candidate spans to be nonempty and disjoint before an in-

Table 3: **Decoder-side hook mapping.** Both sub-block outputs are targeted before residual addition.

Component	Submodule	Intervention point
Self-attention	`o_proj`	pre-residual
MLP	`down_proj`	pre-residual

Table 4: **Uncertainty in the primary category-preservation ratio for the main Qwen regimes.** Ratios are $q_{\text{cat}}(p)/q_{\text{cat}}(0)$. Intervals are 95% candidate-pair-clustered bootstrap intervals from the frozen sweep records. The prespecified validity rule uses the point ratio threshold 0.95.

Model	p	Category ratio	95% CI
Qwen2.5-VL	.40	.998	[.994, 1.002]
Qwen3-VL	.30	.997	[.991, 1.004]
Qwen3-VL	.60	.995	[.987, 1.003]

tervention is applied. The vision encoder, multimodal projector or merger, non-target candidate positions, and input pixels are not directly modified in the core experiments. Targeting verification is reported in Appendix C.

C Targeting Verification

The intervention implementation includes runtime checks to verify that perturbations are applied to the intended candidate-specific decoder positions and requested submodules. For each trial, the selected candidate determines the target position set T_y, while the requested decoder-layer set determines where hooks are active. At every selected layer, interventions are applied at the self-attention output projection and MLP down-projection immediately before residual addition, as described in Section 3.2.

Target coverage and confinement. The runtime checks verify that the candidate-specific target positions are recovered before intervention and that the requested hooks are resolved for the selected decoder layers. Only positions in T_y are passed through the perturbation operation. Positions associated with the non-selected candidate are not directly modified, nor are the vision encoder, multimodal projector or merger and input pixels. Dropout masks and Gaussian-noise samples are generated independently across targeted positions, layers and sub-blocks and are re-seeded per trial.

Failure handling. Targeting checks are evaluated before accepting an experimental forward pass. A run is not treated as valid if the requested candidate positions or intervention modules cannot be resolved consistently with the configured experiment. This prevents silent changes in tokenization or model wrapping from converting a candidate-specific intervention into a broader decoder perturbation. The model-specific hook locations used by these checks are summarized in Appendix B.

D Full Perturbation Sweeps and Multiplicity-Aware Inference

Figure 6 reports the full visual-dropout sweeps underlying the operating points highlighted in Section 4.1. For the Qwen models, the figure shows localization together with the prespecified category ratio and the secondary, post-hoc counting ratios for both the questioned and distractor images. The two criteria agree at the headline Qwen2.5-VL $p = .40$ and Qwen3-VL $p = .30$ settings. At stronger Qwen3-VL perturbations, localization continues to increase while questioned-image counting deteriorates, whereas distractor-image counting remains substantially closer to its clean baseline.

Whole-window multiplicity control. For each model, the inferential family contains the tested nonzero strengths satisfying the prespecified primary category criterion. Let

$$\Delta(s) = \hat{a}(s) - \hat{a}(0)$$

denote the change in localization accuracy from the unperturbed baseline. We test the maximum change over the family with $10{,}000$ candidate-pair-level perturbation-side sign-flip permutations.

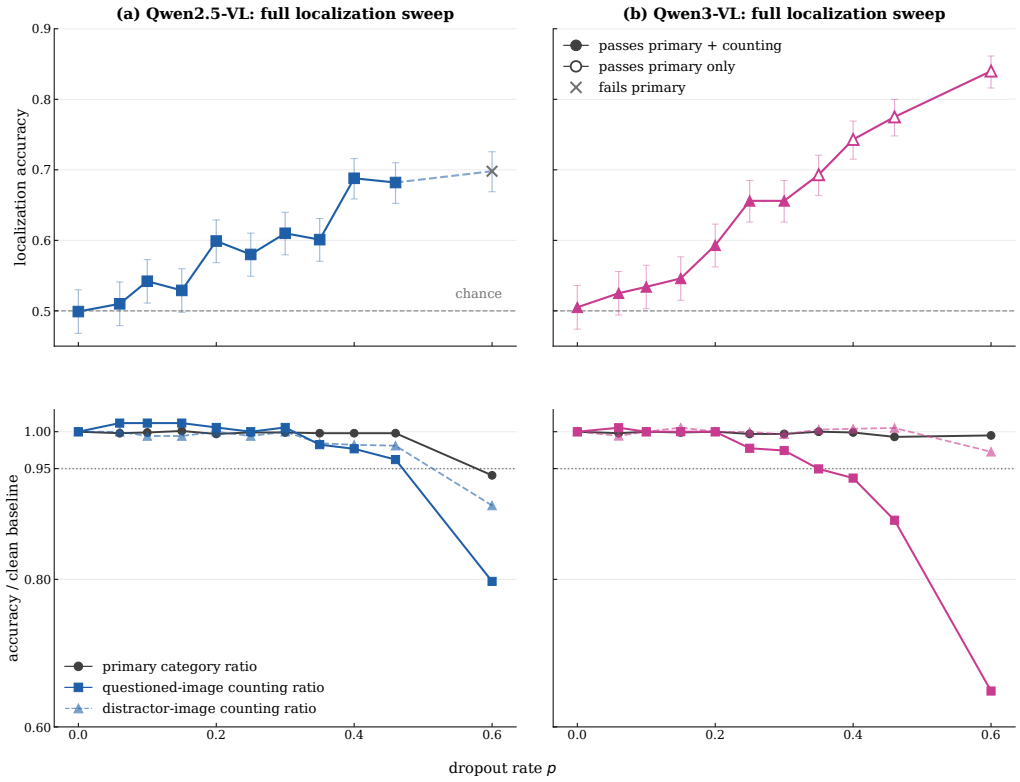

Figure 6: **Full perturbation sweeps and visual-function criteria.** Top panels show candidate-localization accuracy across the complete tested visual-dropout sweeps for Qwen2.5-VL and Qwen3-VL. Bottom panels show accuracy relative to the corresponding clean baseline for the prespecified category task and the secondary, post-hoc counting guard, with counting reported separately when the questioned or distractor image is perturbed. The horizontal 0.95 line marks the criterion threshold and the dashed 0.5 line marks chance localization. Filled markers denote settings where both category and counting criteria pass, open markers denote settings satisfying only the primary category criterion and crosses denote failure of the primary criterion. The distractor-image counting curves remain comparatively stable, whereas questioned-image counting degrades at stronger perturbation strengths.

Table 5: **Secondary counting guard at the main Qwen regimes.** Accuracies refer to the questioned image. Intervals on the absolute accuracies are 95% Wilson or item-clustered intervals according to the frozen analysis; ratio intervals are paired item-clustered bootstrap intervals from the saved counting trials. The pass/fail rule itself remains the prespecified point-ratio rule $q_{\text{count}}(s)/q_{\text{count}}(0) \geq 0.95$. The clean baseline cells contain $n = 192$ trials; each reported perturbed endpoint contains $n = 1000$ trials.

Model	p	Clean $q_{\text{count}}(0)$	Perturbed $q_{\text{count}}(p)$	Ratio [95% CI]
Qwen2.5-VL	.40	.891 [.839,.927]	.870 [.827,.910]	.977 [.949,1.004]
Qwen3-VL	.30	.964 [.927,.982]	.939 [.910,.968]	.975 [.952,.996]
Qwen3-VL	.60	.964 [.927,.982]	.625 [.562,.689]	.649 [.581,.716]

The resulting family-wise inference is represented by the corrected $p_{\max}$ and the one-sided simultaneous threshold c^*; the pointwise intervals reported below are not simultaneous intervals.

For Qwen2.5-VL, the primary family contains nine strengths. The $p = .40$ headline change is $+18.9$ pp with a pointwise item-clustered 95% interval of $[+14.7, +23.1]$ pp. For Qwen3-VL, the family

Table 6: **Whole-window inference for the headline localization effects.** Δ is the change from the model's own $p = 0$ localization accuracy. Intervals shown for the Qwen headline cells are pointwise item-clustered bootstrap intervals; c^* is the family-wise one-sided 95% threshold.

Model	Family	Headline	Δ (pp)	$p_{\max}$	c^* (pp)
Qwen2.5-VL	9	$p = .40$	$+18.9\,[14.7, 23.1]$	$< 10^{-4}$	$+5.5$
Qwen3-VL	10	$p = .30$	$+15.1\,[10.9, 19.4]$	$< 10^{-4}$	$+5.5$
Idefics3	4	$p = .35$	$+5.7$	0.0001	$+2.4$

Table 7: **Visual-reliance prerequisite gates.** Accuracies are measured before the perturbation sweeps. All three models satisfy the prespecified real–blank and blank-equivalence requirements.

Model	Real	Scrambled	Blank	Gate
Qwen2.5-VL	.997	.992	.485	Pass
Qwen3-VL	.993	.998	.506	Pass
Idefics3	.988	.976	.493	Pass

contains ten strengths and the $p = .30$ change is $+15.1$ pp with pointwise interval $[+10.9, +19.4]$ pp. Both whole-window tests give $p_{\max} < 10^{-4}$ and $c^* = +5.5$ pp.

For Idefics3, the primary family contains four nonzero strengths. The one-sided whole-window test gives $p_{\max} = 0.0001$ and $c^* = +2.4$ pp; all four primary-valid changes exceed this threshold. The effect is therefore detectable after accounting for the tested strength window, although its magnitude is substantially smaller than in the Qwen checkpoints and its response-side bias is analyzed separately in Appendix F.

Sampling and independent replication. The Qwen strength cells are independently sampled rather than fully crossed across strengths. In the seed-42 sweeps, individual cells contain 767–810 unique candidate pairs among 1000 trials, and 37.8%–41.7% of a nonbaseline cell's pairs also occur in the baseline cell; no candidate pair occurs in all eleven strength cells. The permutation procedure therefore synchronizes perturbation-side flips whenever the same pair occurs in multiple cells without assuming complete cross-strength pairing.

The headline sweeps were independently repeated with seed 43, resampling candidate pairs, candidate sides, prompt templates and perturbation realizations. For Qwen2.5-VL, localization changes from 0.512 at $p = 0$ to 0.653 at $p = .40$, a $+14.1$ pp change with item-clustered 95% interval $[+9.9, +18.4]$ pp. For Qwen3-VL, localization changes from 0.470 to 0.680 at $p = .30$, a $+21.0$ pp change with interval $[+16.9, +25.2]$ pp. Thus, the direction and substantial magnitude of both headline effects reproduce under an independently resampled trial population. Exact randomization and resampling rules are given in Appendix N.

E Visual-Reliance Prerequisite Gates

Before interpreting any visual perturbation sweep, we verify that the category task depends on genuine visual input rather than prompt-side cues alone. The prerequisite gate compares performance with the original images, spatially scrambled images and blank inputs. A configuration passes only if the real–blank accuracy gap is at least 0.15 with exact McNemar $p < 0.01$ and blank-input performance is equivalent to chance within the prespecified ± 0.05 margin.

All three models pass the prerequisite gate. The high real–blank gaps establish that the category judgments rely on visual content rather than the text prompt alone. At the same time, performance remains high after spatial scrambling, showing that this coarse category task does not require intact spatial structure. This limitation motivates the secondary, post-hoc counting guard used for the Qwen models in Section 3.3 and Appendix D.

Table 8: **Bias-robust localization and paired crossover.** AUC is computed from the continuous localization margin. Crossover values are in model-specific logit units and should not be compared in magnitude across models.

Model	p	Acc.	AUC	D	$D > 0$
Qwen2.5-VL	.40	.688	.734	1.62	92.4%
Qwen3-VL	.30	.671	.818	9.09	98.2%
Qwen3-VL	.60	.840	.965	–	–
Idefics3	.40	.528	.618	.99	66.0%

F Response Bias and Paired Crossover

Aggregate forced-choice accuracy can be distorted by a fixed preference for candidate A or B. We therefore supplement accuracy with the area under the receiver-operating-characteristic curve (AUC) of the continuous A-versus-B localization margin and with paired crossover trials that hold the prompt, candidates, display order, intervention strength and perturbation randomness fixed while switching only the perturbed candidate. A positive crossover contrast D indicates that the readout shifts toward the newly perturbed candidate.

The continuous-margin analysis remains informative despite the raw response-side preferences: AUC is $0.818\,[0.790, 0.844]$ for Qwen3-VL at $p = .30$ and $0.618\,[0.581, 0.652]$ for Idefics3 at $p = .40$, while the corresponding $p = 0$ controls are consistent with chance. The Qwen3-VL $p = .30$ accuracy of 0.671 is from this fresh bias-analysis holdout and therefore differs slightly from the 0.656 estimate in the main sweep.

The paired crossover gives the stronger control against a fixed side preference. At Qwen3-VL $p = .30$, $D = 9.09\,[8.63, 9.56]$ and remains positive after swapping candidate display order, $D = 10.11\,[9.66, 10.56]$; the $p = 0$ control is exactly zero. Qwen2.5-VL and Idefics3 likewise show positive crossover contrasts at $p = .40$. The Idefics3 crossover is evaluated at this nearby primary-valid setting rather than the $p = .35$ headline point. Together, these controls show that a fixed response-side preference is insufficient to explain the candidate-specific localization effect.

G Norm-Restored Perturbation Controls

Inverted dropout and additive Gaussian noise can change activation magnitude, creating a potential cue to the perturbed candidate. We therefore repeat selected Qwen conditions after restoring each perturbed target vector to its original L_2 norm. The maximum relative norm-restoration error across these runs is 6×10^{-7}, well below the implementation tolerance.

Norm restoration preserves substantial localization lift in all four tested conditions. For Qwen2.5-VL, the norm-restored/original lift ratios are $0.92\,[0.64, 1.27]$ for dropout at $p = .40$ and $1.21\,[0.95, 1.56]$ for Gaussian noise at $\sigma = .80$. For Qwen3-VL, the corresponding ratios are $0.95\,[0.79, 1.12]$ at $p = .60$ and $1.07\,[0.90, 1.27]$ at $\sigma = .32$. Thus, activation-norm inflation is not required for candidate-specific localization at these tested settings. This control addresses the norm cue specifically and does not by itself establish preservation of visual function.

H Visible-Corruption Construct Boundary

We compare the internal intervention with visible input corruption to test whether candidate identification is distinguishable from cues associated with degraded visual quality. The visible controls apply Gaussian blur or patch scrambling to one input image without decoder hooks, whereas the internal arm leaves the input pixels unchanged. For the primary frontier comparison, both arms use the matched representation-quality question so that visible corruption is not evaluated with a prompt that explicitly presupposes an internal perturbation. We match observed candidate-identification accuracy within a frozen ± 2 percentage-point tolerance and compare questioned-image counting accuracy at the matched operating points. For each prespecified identification target, each intervention arm contributes the measured cell with the smallest absolute identification mismatch; if two cells

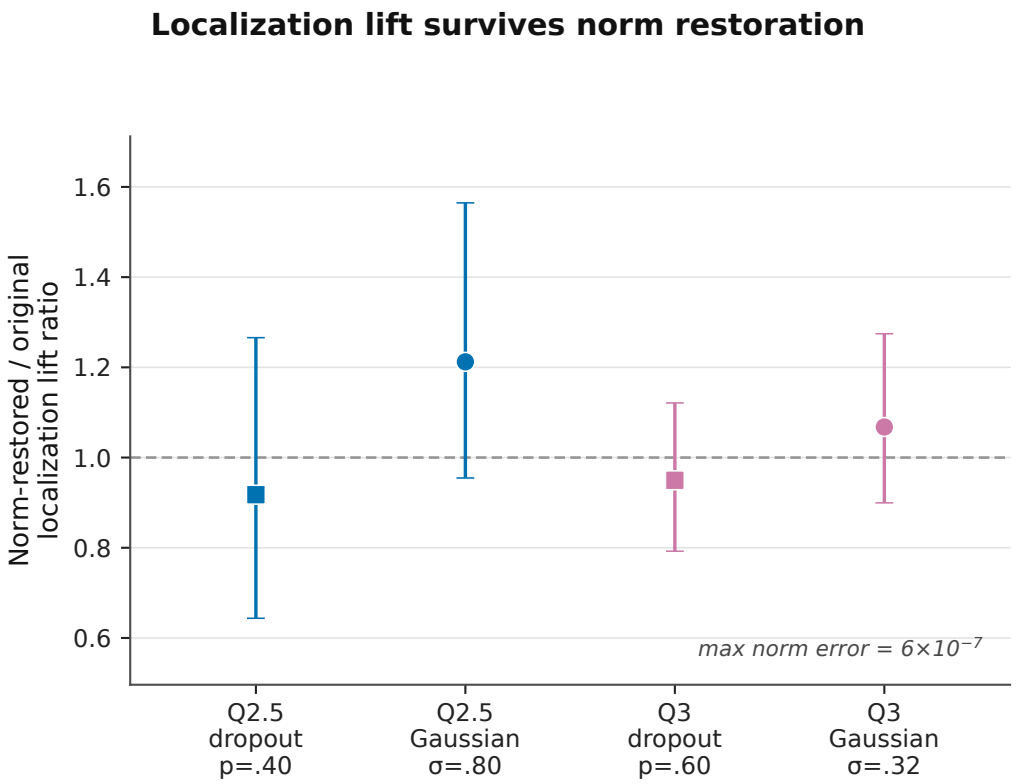

Figure 7: **Localization lift under norm-restored perturbations.** Points show the ratio between localization lift under the norm-restored intervention and the corresponding original intervention; a ratio of 1 denotes equal lift. Error bars are 95% cluster-bootstrap confidence intervals over candidate pairs. Results are shown for dropout and Gaussian noise at the tested Qwen endpoints.

are exactly tied, the first cell in the frozen strength-ordered grid is selected. A contrast is declared non-computable if either arm's nearest measured cell lies outside the tolerance; no interpolation or extrapolation is substituted for the primary comparison.

The measured corruption curves are too coarse for many exact matches: $13/16$ prespecified contrasts are non-computable under the frozen tolerance and are not replaced by interpolated estimates. Among the computable blur contrasts, Qwen2.5-VL is consistent with overlapping frontiers, with an internal-minus-blur counting difference of -0.010 (95% CI $[-0.037, 0.016]$). For Qwen3-VL the corresponding difference is -0.052 $[-0.089, -0.016]$, indicating that visible blur preserves more counting accuracy at comparable candidate identification. Thus, these controls provide no evidence that internal localization is irreducible to degradation-like consequences, while the unchanged input pixels remain an important distinction between the two intervention channels.

I Intervention Extent and Linear Availability

To test whether the full-span effect is driven simply by perturbing many visual positions, we restrict the intervention to exactly $K = 15$ positions while keeping the same decoder-side perturbation mechanism. Figure 9 compares these restricted interventions with the corresponding full-span conditions and evaluates both forced-choice localization and linear decodability at the final answer position.

For Qwen2.5-VL at $p = .40$, the $K = 15$ condition yields behavioral localization of 0.501 and probe accuracy of $0.567\,[0.538, 0.596]$, compared with 0.688 behavior and $0.993\,[0.987, 0.998]$ probe accuracy for the full-span condition. For Qwen3-VL, $K = 15$ gives 0.504 behavior and

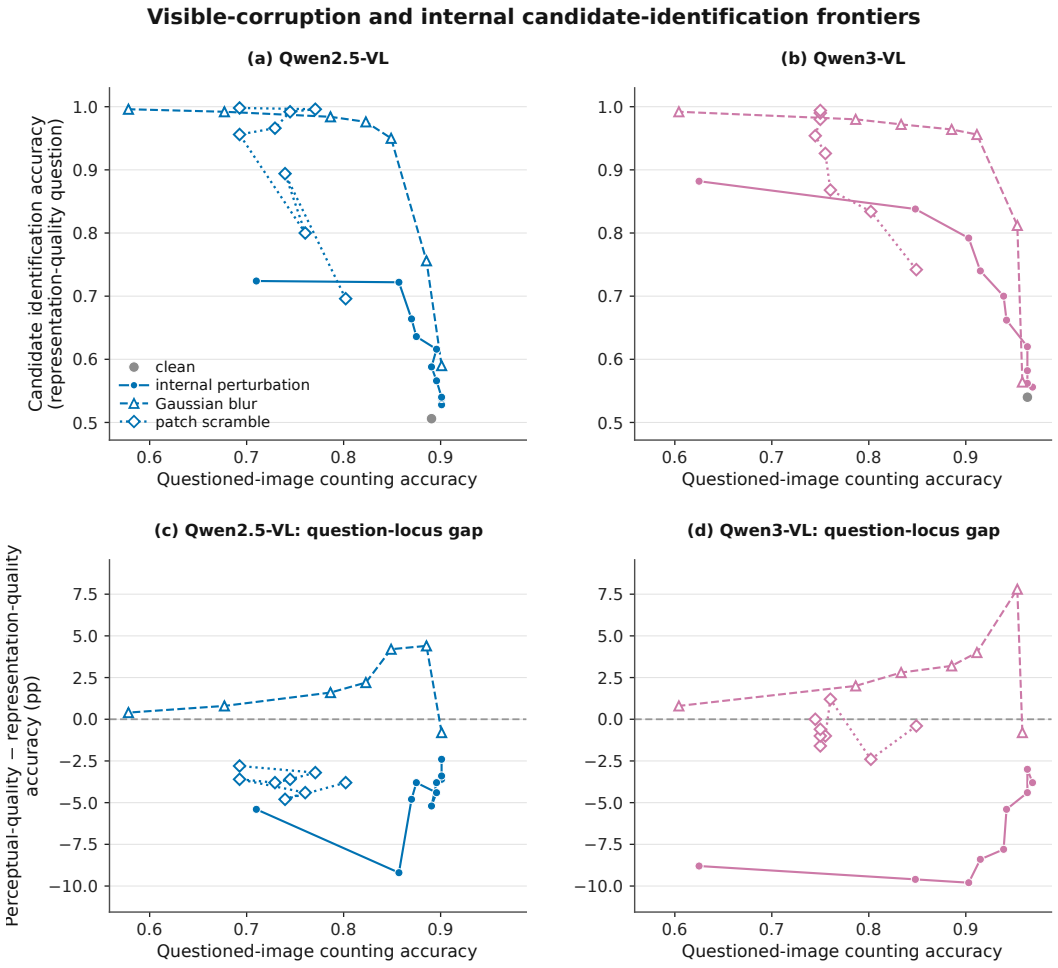

Figure 8: **Visible-corruption and internal candidate-identification frontiers.** (a,b) Candidate-identification accuracy under the matched representation-quality question as a function of questioned-image counting accuracy for internal visual-token perturbations, visible Gaussian blur and patch scrambling. Visible corruption modifies the input image whereas the internal intervention leaves its pixels unchanged. Primary matched comparisons use the frozen ± 2 percentage-point identification tolerance; $13/16$ prespecified contrasts have no observed match and are therefore treated as non-computable rather than interpolated or extrapolated. (c,d) Descriptive question-locus gaps between the perceptual-quality and representation-quality questions. These secondary curves characterize how the accessible quality signal depends on question framing and are not used as an additional validity criterion.

$0.512\,[0.480, 0.542]$ probe accuracy at $p = .30$, rising to 0.534 and $0.681\,[0.652, 0.709]$ at $p = .60$. By contrast, full-span probe accuracy is $0.988\,[0.981, 0.994]$ at $p = .30$ and $0.995\,[0.991, 0.999]$ at $p = .60$.

Thus, visual-token extent materially contributes not only to the forced-choice effect but also to the linear availability of perturbation identity. The near-ceiling full-span probes should therefore not be interpreted as evidence that perturbation identity is equally available under narrowly localized interventions.

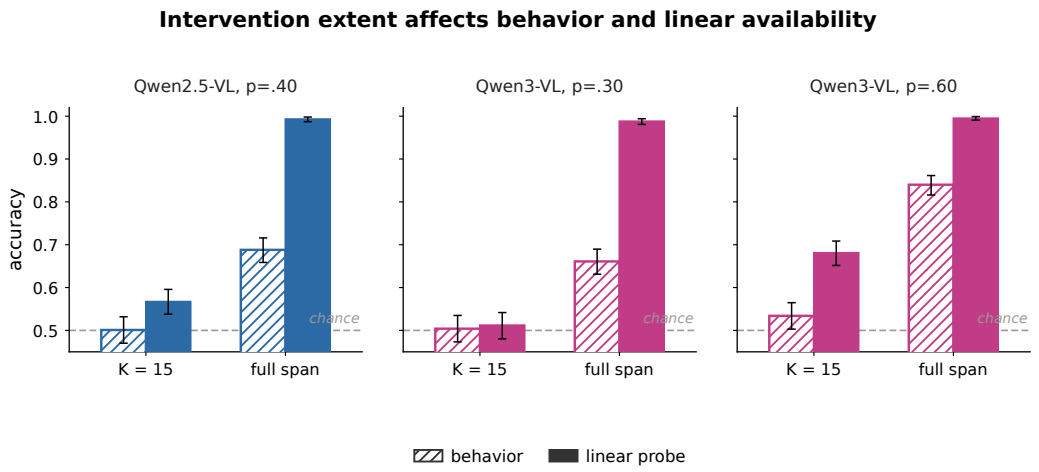

Figure 9: **Intervention extent affects behavior and linear availability.** Behavioral forced-choice localization and held-out linear-probe accuracy are compared between interventions restricted to exactly $K = 15$ visual-token positions and the corresponding full-span conditions. Restricting intervention extent substantially attenuates both measures, including a near-chance probe at Qwen3-VL $p = .30$. Error bars denote 95% confidence intervals and the dashed horizontal line marks chance accuracy. Behavioral values shown here are measured on the corresponding probe-evaluation populations and may therefore differ slightly from the independently sampled main-sweep estimates.

J Decoder Depth and Text-Anchor Analyses

We next restrict dropout to contiguous decoder-depth bands to characterize where candidate-specific localization can be induced. The supplementary text arm provides an anchor for comparison with visual-token interventions under the same forced-choice protocol. For the cross-arm comparison, we use $p = .30$ for Qwen2.5-VL and $p = .35$ for Qwen3-VL, the largest tested strengths satisfying the relevant primary-task criterion across the displayed depth conditions.

At the comparison strengths, the two models show the same ordering. For Qwen2.5-VL at $p = .30$, text localization is highest in Q1 (0.757), whereas visual localization is highest in Q2 (0.574). For Qwen3-VL at $p = .35$, text localization likewise peaks in Q1 (0.808) and visual localization in Q2 (0.673). The broader depth sweeps show the same qualitative separation, although it is not absolute.

These results indicate that candidate localization depends on the computational location of the intervention as well as its extent. They do not identify a necessary circuit or a dedicated modality-specific monitoring mechanism; rather, the restricted interventions identify decoder ranges that are sufficient to induce the observed forced-choice effect.

K Linear-Probe Details

We measure linear availability of perturbation identity from the final answer-position residual using an L_2-regularized logistic-regression probe. Features are converted to float32 and evaluated with five-fold GroupKFold cross-validation using the underlying item identifier as the group, so examples from the same item never cross the train–test boundary. Within each outer training fold, a StandardScaler is fit on that fold only and the inverse regularization strength C is selected by four-fold grouped cross-validation from

$$C \in \{10^{-4}, 10^{-3.5}, \ldots, 10^{2}\},$$

using classification accuracy. The logistic regression uses the `lbfgs` solver, an L_2 penalty, a fitted intercept, no class reweighting, and a maximum of 5000 iterations. Predictions reported in Table 9 are strictly out of fold.

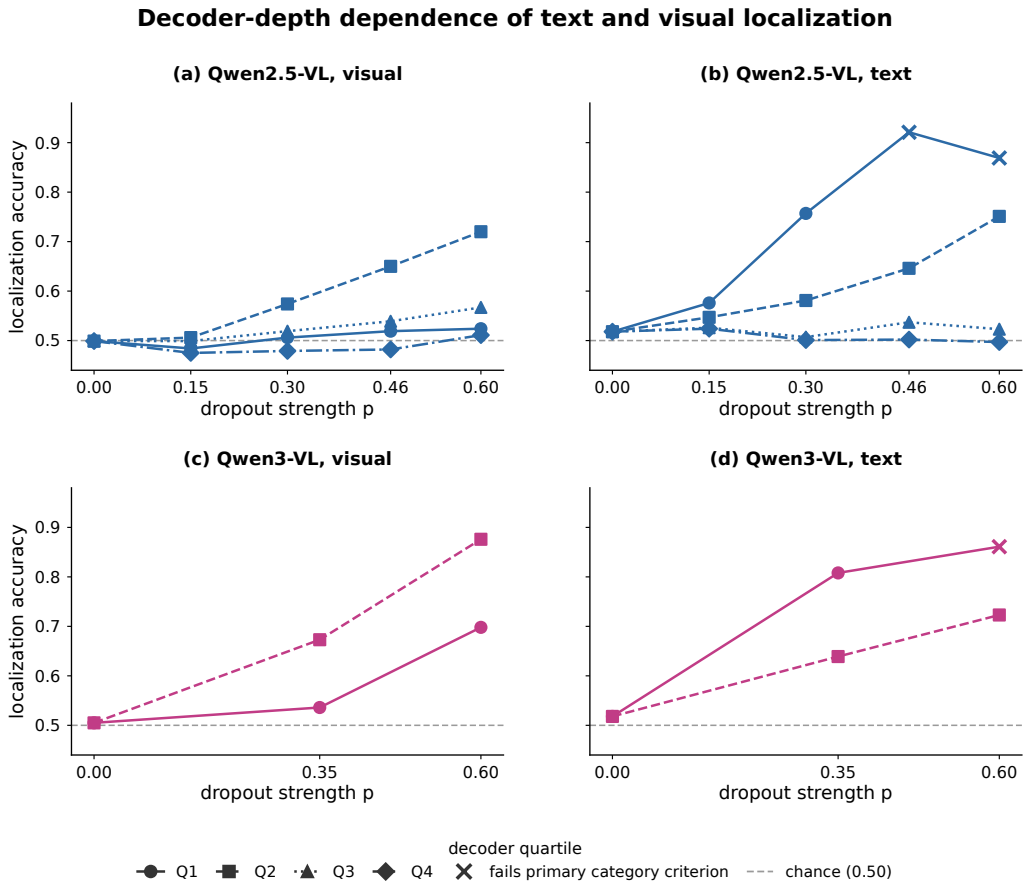

Figure 10: **Decoder-depth dependence of text and visual candidate localization.** Depth-restricted dropout is applied to contiguous decoder bands. For Qwen2.5-VL, Q1–Q4 denote successive decoder quartiles; the Qwen3-VL replication evaluates Q1 and Q2. At the common comparison strengths, text localization is strongest in Q1 whereas visual localization is strongest in Q2 in both models. Crosses mark conditions that fail the corresponding primary category criterion and are shown descriptively only. The secondary counting guard was not measured for these depth-specific conditions. Restricted-band effects identify layer ranges sufficient to induce localization and do not establish necessity.

Each newly extracted condition contains 1000 trials with perturbed-candidate labels counterbalanced $500/500$. Confidence intervals are percentile intervals from 2000 item-clustered bootstrap resamples using frozen seed 20260819. Behavioral accuracy is evaluated on the same trial population when the cell is newly extracted. For steering, the probe is subsequently refit on all 1000 training rows using the same scaler, solver and inner grouped C-selection procedure; the selected C values are 3.1623 for Qwen3-VL at $p = .30$, 0.1 for Qwen3-VL at $p = .60$, and 1.0 for Qwen2.5-VL at $p = .40$.

The full-span representations make perturbation identity nearly perfectly linearly decodable at all three reported conditions. The Qwen3-VL $p = .30$ result is particularly informative because both category and counting criteria pass there: probe accuracy is 0.988 while behavioral localization is 0.661. The latter is measured on the probe-evaluation population and therefore differs slightly from the independently sampled main-sweep estimate of 0.656. These probe results establish linear availability only; they do not establish knowledge, awareness or causal use by the model's readout.

As an independent negative-control check, the fresh extent-analysis populations give $p = 0$ probe accuracies of $0.527\,[0.494, 0.558]$ for Qwen2.5-VL and $0.507\,[0.477, 0.538]$ for Qwen3-VL. The

Table 9: **Full-span linear-probe results.** Probe accuracy measures linear decodability of the perturbed candidate from the final answer-position residual. Behavioral values correspond to the associated forced-choice evaluation population.

Model	Condition	Probe accuracy	Behavior
Qwen2.5-VL	$p = .40$	.993 [.987, .998]	.688
Qwen3-VL	$p = .30$	.988 [.981, .994]	.661
Qwen3-VL	$p = .60$	.995 [.991, .999]	.840

frozen stop-and-debug rule declares probe leakage only if the $p = 0$ bootstrap interval has a lower bound above chance (0.5); neither control triggers this criterion. The strong attenuation under $K = 15$ interventions is reported separately in Appendix I.

L Causal Steering and Intervention Controls

We test steering efficacy of the fitted probe direction on fresh evaluation sets whose items are disjoint from those used to fit the corresponding direction. Let w, b denote the logistic-regression parameters in standardized feature space, and let μ, σ be the fitted feature-wise scaler statistics. We convert the classifier exactly to raw residual space as

$$d = w/\sigma, \qquad b' = b - d^\top \mu, \qquad \hat{d} = d/\|d\|_2,$$

where division by σ is applied elementwise. We verify that $\mathrm{sign}(d^\top h + b')$ reproduces the fitted pipeline's predictions on the training features. The independently fitted Qwen3-VL $p = .30$ and $p = .60$ directions have cosine similarity 0.458.

The steering scale is

$$s_L = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{d}^\top h_i)^2},$$

computed on the corresponding 1000-row probe-training population. This gives $s_L = 29.0712$ at $p = .30$ and 50.0689 at $p = .60$ for Qwen3-VL (19.2455 for Qwen2.5-VL at $p = .40$). At the final decoder-layer output and answer position, before the model's final normalization, steering applies

$$h' = h + s\,\alpha s_L \hat{d},$$

with the addition computed in float32 and cast back to the model dtype. Because probe class 1 denotes candidate B, $s = +1$ targets B and $s = -1$ targets A. For opposite-candidate steering, the sign is chosen separately for each trial to target the candidate opposite that trial's own unsteered prediction.

For the random-direction control, each evaluation trial receives a deterministic independently sampled vector $r \sim \mathcal{N}(0, I)$, normalized to unit length and scaled to the same norm αs_L as the probe-direction intervention. The trial-specific random vector is reused across steering scales so that only its magnitude changes with α. Every steered trial has a cached unsteered reference prediction. A targeted flip occurs when the steered prediction differs from that reference in the specified steering direction; for opposite-candidate steering this is a switch to the other candidate. Endpoint flip rates use all 1000 trials in the corresponding arm, with item-clustered confidence intervals.

For a trial i, normalized displacement is

$$r_{\Delta,i} = \frac{\|\Delta h_i\|_2}{\|h_{\mathrm{base},i}\|_2} = \frac{\alpha s_L}{\|h_{\mathrm{base},i}\|_2},$$

because $\hat{d}$ is unit norm. The injection implementation additionally checks the realized $\|\Delta h\|_2$ against αs_L for each intervention cell.

The contrast cannot be reduced simply to lower probe accuracy or smaller injected displacement at $p = .30$. At $\alpha = 4$, the $p = .30$ direction produces normalized displacement $r_\Delta = 0.148$, exceeding the 0.114 produced by the $p = .60$ direction at $\alpha = 2$, yet the corresponding flip rates

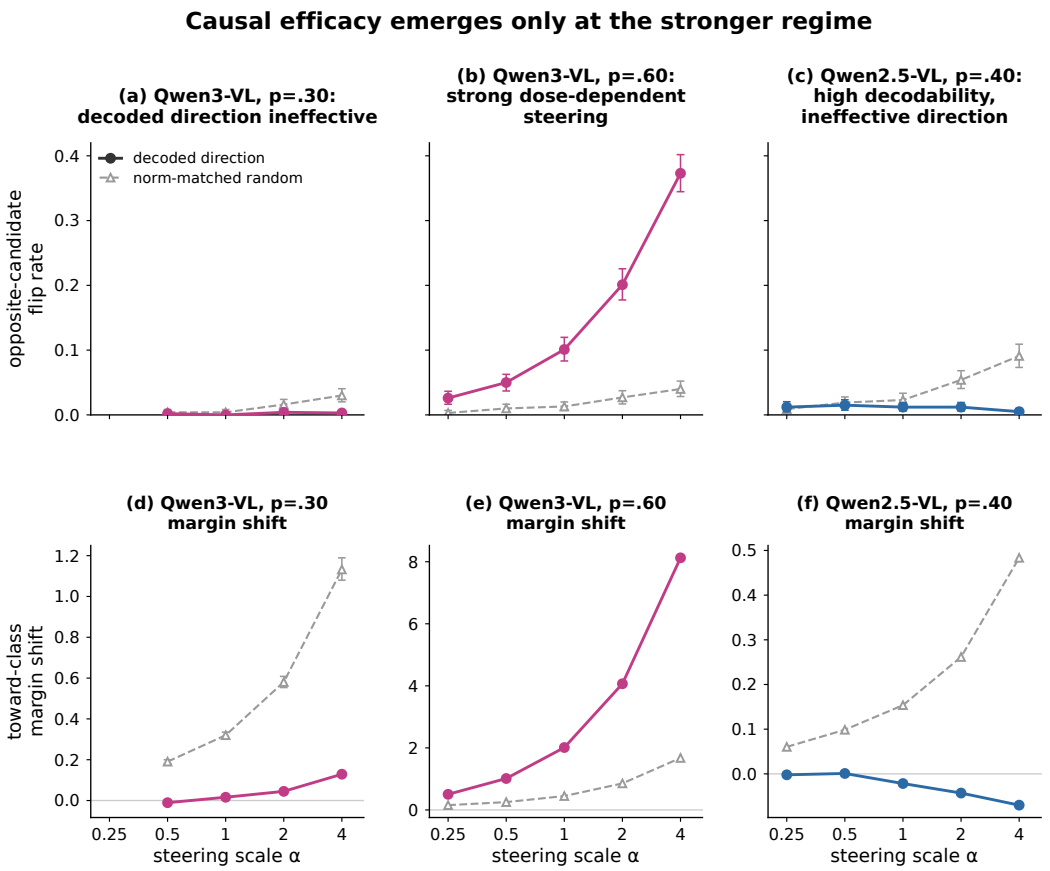

Figure 11: **Steering efficacy of decoded perturbation directions.** (a–c) Opposite-candidate flip rates under steering along the decoded direction and norm-matched random directions. The fitted probe direction produces little targeted steering for Qwen3-VL at $p = .30$ and Qwen2.5-VL at $p = .40$, despite high linear decodability, whereas Qwen3-VL at $p = .60$ exhibits a strong dose-dependent response. (d–f) Toward-class margin shifts show the same qualitative separation. Error bars denote 95% confidence intervals. The margin-shift panels use different vertical scales to expose the within-regime dose responses. These results establish sufficiency of the fitted Qwen3-VL $p = .60$ direction under the tested intervention, not necessity.

Table 10: **Geometry and displacement checks for Qwen3-VL steering.** The displacement comparison is deliberately conservative: the $p = .30$ intervention has the larger normalized residual displacement.

	$p = .30$	$p = .60$
Probe accuracy	.988	.995
Counting ratio	.975	.649
Max. answer-logit $\lvert\cos\rvert$	.0086	.0581
Compared steering scale	$\alpha = 4$	$\alpha = 2$
Normalized displacement r_Δ	.148	.114
Opposite-candidate flips	.003	.201

are only 0.3% and 20.1%, respectively. These results establish sufficiency of the fitted Qwen3-VL $p = .60$ direction under the tested intervention, not necessity. The fitted directions are also nearly orthogonal to the direct forced-choice readout. We construct the primary readout vector as the difference between the sums of the unembedding rows for the frozen A- and B-token variant sets, and register an uppercase-only variant as a second check. The corresponding cosines are $-0.0002/-$

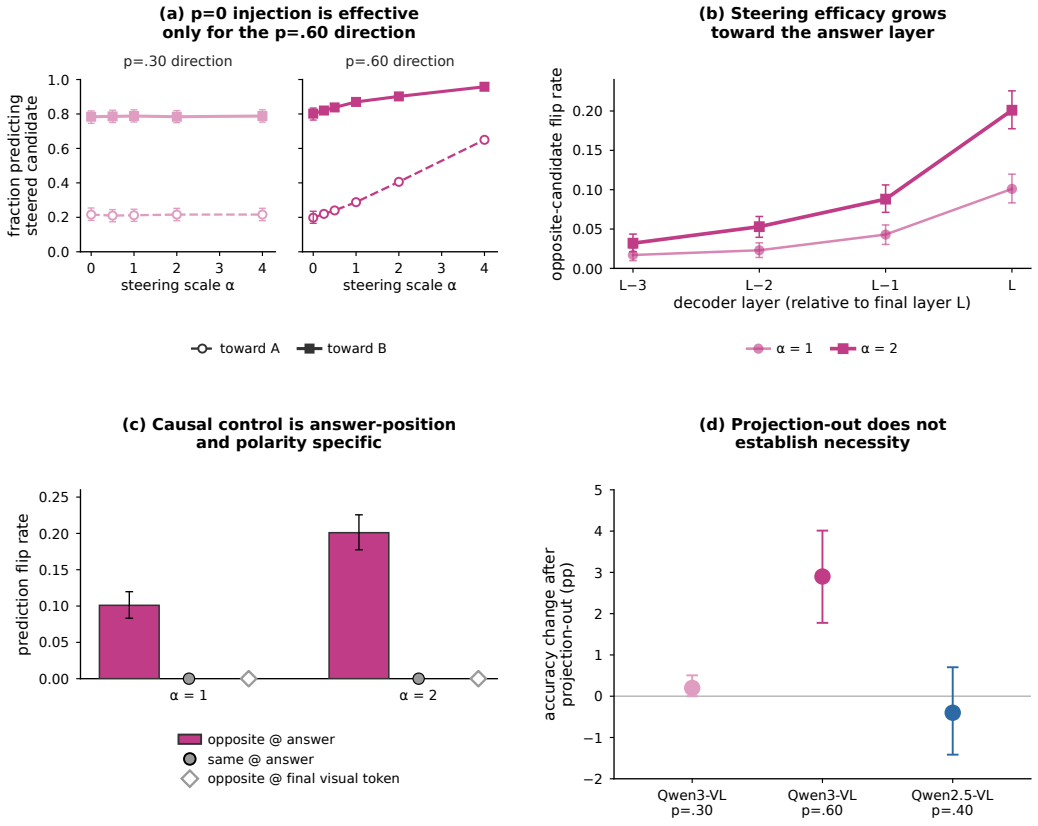

Figure 12: **Additional controls for causal steering.** (a) With no underlying perturbation ($p = 0$), injecting the fitted Qwen3-VL $p = .60$ direction steers the forced-choice readout bidirectionally according to its sign, whereas the fitted $p = .30$ direction produces little change. (b) Opposite-candidate steering becomes progressively stronger toward the final decoder layer. (c) At the tested scales, opposite-direction steering at the answer position produces flips, whereas same-direction steering and opposite-direction injection at the final visual-token position do not. (d) Projection-out interventions do not establish necessity: removing the decoded direction does not produce the accuracy impairment required for such a conclusion. Error bars denote 95% confidence intervals.

0.0086 at $p = .30$ and $-0.0581/-0.0487$ at $p = .60$, giving the maximum absolute overlaps reported in Table 10. Projection-out at the same final answer position uses

$$h' = h - (h^\top \hat{d})\hat{d}.$$

With no underlying perturbation, the $p = .60$ direction is sufficient to control the forced-choice readout: positive steering raises B reports from 0.802 to 0.958, while reversing the direction raises A reports from 0.198 to 0.650. The same vector has no measurable steering effect when injected at the final visual-token position, and the decoder-layer minisweep strengthens toward the answer-position readout. By contrast, the $p = .30$ direction leaves the corresponding $p = 0$ readout essentially unchanged ($0.786 \rightarrow 0.788$ toward B and $0.210 \rightarrow 0.216$ toward A).

These interventions establish causal sufficiency of the fitted $p = .60$ direction under the tested steering procedure, while showing sharply different steering efficacy of the fitted probe direction despite near-ceiling linear decodability at both tested regimes. They do not establish that perturbation-linked information is unused by the model's native computation, rule out other causally useful directions at $p = .30$, or establish necessity through the projection-out analyses.

M Additional Diagnostics and Robustness

Realized perturbation magnitude. We additionally test whether trial-to-trial variation in realized perturbation magnitude predicts localization within a fixed nominal intervention strength. For dropout, realized magnitude varies only weakly within each strength, making these within-cell associations intrinsically low-powered; the observed coefficients show no consistent dose direction. The headline-cell slope is -0.097 $[-0.418, 0.236]$. Gaussian perturbations provide greater within-strength variation, but the two model-level associations with detectable dose tracking have opposite signs. We therefore find no consistent evidence that residual trial-level variation in realized perturbation magnitude explains localization beyond the nominal intervention setting.

We also examine intervention extent as a continuous trial-level quantity where it varies naturally. None of the 18 tested extent–localization associations yields a confidence interval excluding zero. These analyses are diagnostic rather than mechanistic: the narrow within-strength magnitude distributions limit statistical power, so null associations should not be interpreted as evidence that perturbation magnitude is irrelevant. Realized dropout fraction itself was not logged in the committed runs and is therefore not analyzed post hoc.

N Statistics, Provenance and Reproducibility

Statistical procedures. Unless otherwise stated, production perturbation-sweep cells contain $n = 1000$ trials. Binomial accuracies are reported with 95% Wilson intervals. Paired crossover uncertainty is obtained by resampling candidate pairs, while probe, AUC and other item-dependent analyses use item-clustered resampling at the corresponding experimental unit. The visual-reliance prerequisite uses an exact McNemar test for paired real–blank outcomes together with two one-sided tests for equivalence of blank performance to chance within the prespecified ± 0.05 margin.

For whole-window localization inference, let C_s be the number of correct localization judgments among the $n = 1000$ trials at strength s, let

$$\hat{a}(s) = \frac{C_s}{n}, \qquad \Delta(s) = \hat{a}(s) - \hat{a}(0),$$

and let $\mathcal{S}_{\text{primary}}$ be the primary category-valid nonzero strength set defined in Equation 3. The family null is

$$H_{0,s} : \Delta(s) \leq 0 \qquad \text{for all } s \in \mathcal{S}_{\text{primary}},$$

and the observed max statistic is

$$T_{\text{obs}} = \max_{s \in \mathcal{S}_{\text{primary}}} \Delta(s).$$

For permutation replicate b, each unique candidate pair k receives one Rademacher perturbation-side flip. That flip is applied to every occurrence of the same pair in every available strength cell, including the $p = 0$ cell; flipping a pair exchanges which perturbation-side label is treated as correct. The resulting permuted changes $\Delta^{(b)}(s)$ define

$$T^{(b)} = \max_{s \in \mathcal{S}_{\text{primary}}} \Delta^{(b)}(s).$$

With $B = 10{,}000$ permutations, the one-sided whole-window value is

$$p_{\max} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\left[T^{(b)} \geq T_{\text{obs}}\right]}{B + 1}.$$

The family-wise one-sided 95% threshold is

$$c^* = Q_{0.95}\left(T^{(1)}, \ldots, T^{(B)}\right),$$

using the conservative "higher" empirical quantile; a strength clears the simultaneous threshold iff $\Delta(s) > c^*$.

The Qwen strength cells are independently sampled and only partially overlap in candidate pairs. A repeated pair receives the same flip wherever it occurs, but the procedure does not treat trial index

as a cross-strength pairing key. Across the seed-42 Qwen sweep, cells contain 767–810 unique pairs and their overlap with the baseline pair set is 37.8%–41.7%. By contrast, the Idefics3 sweep is fully crossed over the same 781 candidate pairs across all six tested cells, so the pair-level flips are synchronized across the complete strength grid. The Qwen and Idefics3 permutation streams use frozen seeds 20260909 and 20260821, respectively.

Pointwise confidence intervals and family-wise inference are kept distinct. For the Qwen localization changes, pointwise intervals are item-clustered two-sample percentile-bootstrap intervals, resampling the pair sets of the baseline and target cells independently. The simultaneous claim is represented by c^* and $p_{\max}$; we do not label the pointwise bootstrap intervals as simultaneous intervals. The secondary counting guard is not used to define this inferential family.

Randomness and independent replication. Primary sweeps use seed 42. The Qwen headline sweeps are independently repeated with seed 43, resampling trials, candidate sides, templates and perturbation realizations rather than replaying the original trial population. Random-number streams for perturbations are separated across decoder layers and sub-blocks. Additional analysis-specific bootstrap and permutation seeds are fixed in the corresponding configuration files and are distributed with the analysis code.

Provenance and reproducibility. Each reported experiment retains its configuration, trial-level outputs and analysis artifacts together with the information required to reconstruct the corresponding aggregate result. Analysis scripts regenerate reported statistics and figure-source data from these saved records rather than from manually transcribed values. Runtime checks verify candidate-specific targeting before accepting intervention runs and analyses include consistency checks against frozen reference quantities where applicable. Repeated executions of deterministic analysis paths and overlapping extraction cells were also used to verify reproducibility; in particular, independently executed overlapping Qwen3-VL $p = 0$ trials produced bit-identical predicted labels and A/B logits.

The anonymous code artifact provides the model and processor configurations, intervention mappings, strength grids, prompt resources, seed construction, analysis scripts and environment information needed to reproduce the reported experiments. Raw COCO images are not redistributed; experiments instead reference the publicly available dataset and the frozen sampling metadata required to reconstruct the evaluated candidate pairs.