

# The Cost of Language: Centroid Erasure Exposes and Exploits Modal Competition in Multimodal Language Models

Akshay Paruchuri<sup>1,4</sup>, Ishan Chatterjee<sup>2</sup>, Henry Fuchs<sup>3</sup>, Ehsan Adeli<sup>4</sup>, Piotr Didyk<sup>1</sup>

<sup>1</sup>Università della Svizzera italiana, <sup>2</sup>University of Washington

<sup>3</sup>UNC Chapel Hill, <sup>4</sup>Stanford University

akshaypa@stanford.edu

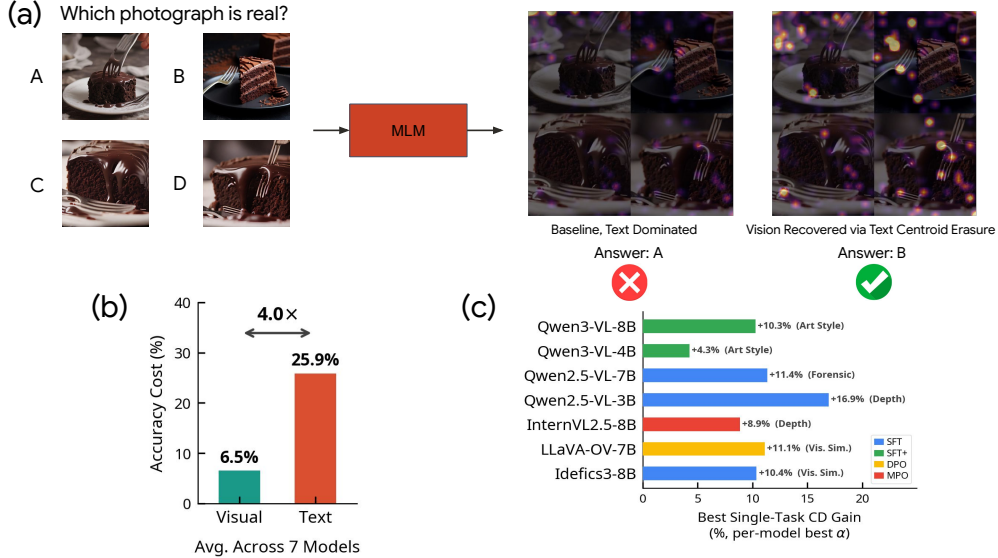


**Figure 1: Centroid erasure exposes and exploits modal competition in multimodal language models.** (a) In one BLINK example, text centroid replacement raises visual attention at L22 (3.6%  $\rightarrow$  11.1%), while text centroid contrastive decoding (TCCD) flips the prediction from incorrect to correct. The displayed heatmaps are qualitative; quantitative attention is computed separately after RoPE. (b) Across seven models and six BLINK tasks, post-image text replacement costs  $4\times$  more accuracy than visual replacement (25.9% vs. 6.5%). Since text replacement also changes the task and answer interface, the gap measures relative dependence. (c) At oracle best-per-task  $\alpha_{\text{interp}}$ , TCCD recovers up to +16.9% accuracy on an individual task, with every model showing positive gains on at least one task.

## Abstract

Multimodal language models systematically underperform on visual perception tasks, yet the structure underlying this failure remains poorly understood. We propose centroid replacement, mapping tokens to their nearest K-means centroid and removing within-cluster residual structure, as a controlled probe for modal dependence. Across seven models spanning four architecture families, post-image text replacement costs  $4\times$  more accuracy than visual replacement on six BLINK tasks; VPBench and MedBLINK preserve this aggregate ordering. Since text replacement also disrupts the task and answer interface, the gap measures relative dependence. Text centroid contrastive decoding (TCCD) exploits this asymmetry without retraining, recovering up to +16.9% accuracy on an individual task at oracle best-per-task  $\alpha_{\text{interp}}$ ; every model gains on at least one task, although gains vary across tasks and models. Together, centroid replacement provides a retraining-free audit of modal dependence on labeled data, while TCCD offers a reusable inference-time intervention. Code and artifacts are available at: <https://github.com/yahskapar/centroid-erasure>

## 1 Introduction

Multimodal language models (MLMs) extend the capabilities of large language models (LLMs) to visual inputs, yet they systematically underperform on tasks that require visual perception. On benchmarks such as BLINK (Fu et al., 2024) (depth estimation, visual similarity, forensic detection), even state-of-the-art MLMs fall short of human perceptual capabilities and of what their own vision encoders can achieve in isolation (Tong et al., 2024b; Fu et al., 2025). Recent work has traced this failure to the language model itself: visual representations encode the information needed to solve these tasks, but the LLM backbone fails to use it (Fu et al., 2025). Together, these findings point to a modal imbalance favoring text (Deng et al., 2025; Wu et al., 2025), further reflected in gradient and attention skew toward text.

These perceptual failures carry substantial consequences as MLMs are deployed in high-stakes domains. In healthcare (Tu et al., 2024), frontier MLMs can generate detailed clinical findings for images that were never provided while achieving top chest X-ray question-answering scores without visual input (Asadi et al., 2026). Such ungrounded perception errors can propagate into triage, diagnosis, and treatment planning. The problem is further compounded in agentic systems: personal-health architectures rely on each perception step being faithful to the input (Heydari et al., 2025), and a single instance of text-dominated perception can propagate through downstream reasoning in single-agent systems (e.g., Merill et al., 2026) and multi-agent pipelines (Kim et al., 2024). Measuring how predictions depend on language versus visual structure, and testing whether targeted inference-time interventions help without retraining, is therefore not merely an academic question but a practical requirement for trustworthy multimodal deployment.

Recently, NIST AI 800-4 (Rao et al., 2026) identified functionality monitoring as an open challenge for deployed AI systems, ensuring that a model’s outputs are grounded in the intended evidence channels rather than merely appearing correct on benchmark metrics. Existing approaches generally require labeled data, human review, or task-specific telemetry. A retraining-free forward-pass probe that quantifies relative dependence on fine-grained text versus visual structure on a labeled audit set can complement these approaches.

In this work, we propose *centroid replacement* as a controlled probe for modal dependence: collapsing each token to its nearest K-means centroid and measuring the resulting accuracy cost. This removes within-cluster residual structure while retaining the centroids themselves; throughout, we call this residual removal “centroid erasure”. Applying the same replacement rule separately to visual and post-image text tokens reveals a stark asymmetry. Across seven models spanning four architecture families, text replacement costs  $4\times$  more accuracy than visual replacement (25.9% vs. 6.5% on average; Figure 1b). We call the resulting dependence asymmetry *modal competition*. These tasks place answer-discriminating evidence in the image while text supplies the task, candidate semantics, and answer interface; because text replacement also disrupts that interface, the probe measures relative dependence rather than a causal effect of harmful language competition.

We further exploit this asymmetry through *text centroid contrastive decoding* (TCCD): running a second forward pass with post-image text residual structure removed and contrastively decoding the original output against this reference. At oracle best-per-task  $\alpha_{\text{interp}}$ , TCCD recovers up to +16.9% accuracy on an individual task, and every model gains on at least one task, although gains vary across tasks and models (Figure 1c). We test specificity with an equal-norm matched-random control and a directional sweep of contrastive strength  $\alpha_{\text{cd}}$  (Section 4.2). Beyond BLINK, separately scored VPBench and MedBLINK portfolios retain the aggregate text-visual ordering, while recovery is more variable (Section 3.4). An exploratory checkpoint comparison gives a larger oracle best-per-task mean gain for SFT/DPO-labeled models (+5.6%) than for SFT+/MPO models (+1.5%); fixed and cross-validated protocols preserve this ordering, but architecture, data, baseline accuracy, and task composition remain confounded.

Our contributions are:

1. **Revealing a consistent asymmetry in modal dependence.** Across seven models and four architecture families, removing post-image text residual structure costs  $4\times$  more accuracy than removing visual residual structure; separately scored VPBench and MedBLINK portfolios preserve the aggregate ordering (Section 3).
2. **Providing a retraining-free inference-time intervention.** In the seven-model, six-task BLINK core, TCCD recovers up to +16.9% accuracy on an individual task at oracle best-per-task  $\alpha_{\text{interp}}$ . The negative- $\alpha_{\text{cd}}$  side of a directional sweep is monotonic on five of six tasks, and an equal-norm matched-random control tests specificity (Section 4).
3. **Auditing modal dependence on labeled data.** Positional-span and layer ablations characterize dose- and depth-dependent sensitivity, while an exploratory stage probe tracks visual replacement cost from 22% at the encoder to 2% by L16 (Section 4.2). Together, replacement cost and TCCD response provide retraining-free audit measures, with TCCD response treated as secondary (Section 5).

Code, centroid artifacts, and aggregate results are available at: <https://github.com/yahskapar/centroid-erasure>

## 2 Related Work

**Modal imbalance in multimodal language models.** The observation that MLMs under-utilize visual information has emerged from multiple independent lines of evidence. Fu et al. (2025) showed that vision encoders encode the information needed for visual perception tasks, but the LLM backbone fails to use it; Deng et al. (2025) confirmed text dominance across VLM families, and Shahgir et al. (2026) showed that VLMs favor semantic anchors over visual detail. Li et al. (2025) found that language token features carry over  $10\times$  the average absolute gradient of vision tokens, while Wu et al. (2025) identified cross-modal attention imbalance and proposed a training-time fix. Earlier work traced modality gaps to contrastive pretraining dynamics (Liang et al., 2022) and under-optimized unimodal representations (Peng et al., 2022; Laurençon et al., 2024; Jiang et al., 2025). Our work complements these findings with a geometric measure of text-visual dependence and a task- and model-dependent training-free intervention.

**Contrastive decoding for multimodal language models.** Contrastive decoding methods typically mitigate hallucinations by subtracting a degraded reference distribution from the model’s output logits. VCD (Leng et al., 2024) contrasts original and noise-distorted visual inputs, while LCD (Manevich & Tsarfaty, 2024) subtracts the text-only LLM distribution. DoLa (Chuang et al., 2023) contrasts later- and earlier-layer logits, SDCCD (Xia et al., 2026) shuffles visual patches, and OPERA (Huang et al., 2024) penalizes over-trusted attention patterns during beam search. However, Yin et al. (2025) showed that gains on POPE (Li et al., 2023) are largely distributional artifacts rather than genuine hallucination suppression, questioning CD’s role as a universal intervention. TCCD differs in its intervention locus: it constructs the degraded reference by replacing text-side representations at an intermediate layer, then contrasts the resulting output logits. We treat variation in its gains as a secondary post-hoc measure of how strongly the measured dependence responds to this intervention, not as a universal correction or a training objective (Table 28, Section F.12).

**Representation geometry in language models.** Recent work has revealed rich geometric structure in LLM representations. Huh et al. (2024) argued that vision and language models converge toward a shared representation, Lee et al. (2025) found shared global and local geometry across LLM token embeddings, and Park et al. (2024) showed that categorical concepts form polytopes in representation space. In the multimodal setting, Papadimitriou et al. (2025) found a small stable dictionary of unimodal concepts that form cross-modal bridges in CLIP and SigLIP embedding spaces. Qi et al. (2025) found that randomly permuting vision token embeddings causes only 0.2–2.7% performance drops, suggesting that VLMs treat visual tokens as a semantic bag of tokens. Token compression methods

such as FastV (Chen et al., 2024a) and VisionZip (Yang et al., 2025) similarly exploit visual token redundancy. Our centroid analysis connects to this work by applying K-means decomposition to *both* visual and text tokens within the same model, directly measuring a cross-modal asymmetry in dependence on within-cluster structure.

### 3 Measuring Modal Competition

We propose *centroid replacement* as a controlled probe for modal dependence. For each model, we fit  $K$ -means centroids ( $K=256$ ; sensitivity to  $K$  and  $N$  in Section D.4) on hidden-state activations from one intermediate layer per modality (text L12, visual L16; Section A.2). The fitting pool contains  $N=2,000$  MS-COCO images and uses no BLINK labels or evaluation prompts. Centroid replacement interpolates each token toward its nearest centroid:  $\mathbf{x}' = \mu_k + \alpha_{\text{interp}} \cdot (\mathbf{x} - \mu_k)$ . At  $\alpha_{\text{interp}}=0$ , every token collapses to its cluster center and all within-cluster variation is discarded; at  $\alpha_{\text{interp}}=1$ , the original representation is preserved. We refer to the former as *centroid erasure*: it erases within-cluster variation while preserving cluster identity. Accuracy costs and gains are absolute changes in accuracy, reported in percent and macro-averaged across tasks unless stated otherwise. The text intervention covers the full post-image sequence, including the question, instructions, and answer options, but not pre-visual template tokens. We apply the same replacement rule separately to text and visual tokens, then measure the resulting accuracy cost on BLINK (Fu et al., 2024), a benchmark of visual perception tasks reformulated as multiple-choice questions.

We study six of BLINK’s fourteen tasks in depth: Forensic Detection, Visual Similarity, Art Style, Counting, Relative Depth, and Spatial Relation (Section A.3). They provide measurable baselines and span a broad range of perceptual demands. The CD results separate them into two descriptive groups, TEXT-COMPETES and TEXT-NEEDED, which we analyze in Section 4; because these labels are assigned from the observed outcomes, they do not define an independent task taxonomy. For Qwen2.5-VL-7B, several omitted tasks have low baselines under the fixed four-letter readout, while Multi-view Reasoning and Jigsaw are higher and were not included in our exploration (Section B). Fixed-protocol results for all fourteen tasks appear in Section F.8. Evaluations use this fixed readout and deterministic greedy decoding; Wilson intervals and two centroid-refit sensitivity analyses are reported in Section E, with model details in Section A.1.

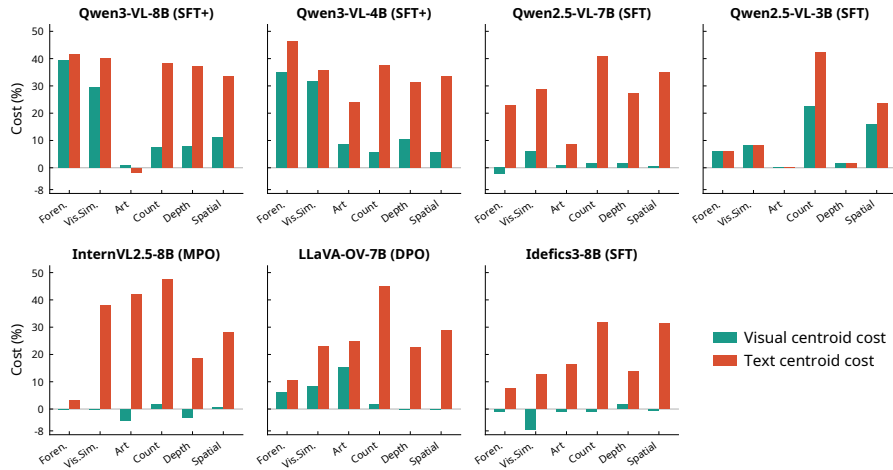


Figure 2: **The text–visual asymmetry is consistent across models.** Per-task centroid replacement cost (visual in teal, text in red) across seven models. Text replacement is more costly for every model when averaged across tasks and in 37 of 42 model–task comparisons, with four ties and one reversal.

### 3.1 Within-Cluster Visual Variation Is Largely Redundant at the Decision Layer

Replacing all visual tokens with their nearest centroids ( $\alpha_{\text{interp}}=0$ ) costs remarkably little at the decision layer. Across seven models, the mean accuracy drop is only 6.5% (Table 1). InternVL2.5-8B (Chen et al., 2024c) and Idefics3-8B (Laurençon et al., 2024) actually improve slightly ( $-0.9\%$  and  $-1.5\%$  cost), indicating that the removed within-cluster variation can sometimes be mildly harmful under this protocol. This is consistent with prior observations that visual token compression preserves performance (Chen et al., 2024a; Yang et al., 2025). The result is specific to the decision layer: an exploratory stage probe finds a 22% cost at the vision-encoder output (Section 4.2).

### 3.2 Text Centroids Are Costly: The $4.0\times$ Asymmetry

Applying the same centroid replacement to post-image text tokens yields a dramatically different result. Collapsing these tokens to their nearest centroids costs 25.9% accuracy on average,  $4.0\times$  the visual cost (Table 1). Text cost exceeds visual cost in 37 of 42 model–task comparisons, with four ties and one reversal (Figure 2). The two interventions share the fitting procedure, number of centroids, images, and evaluation, but use different layers (text L12, visual L16) and token counts. The asymmetry therefore measures dependence at these two intervention points rather than a matched-layer causal effect. Under this protocol, predictions depend more on fine-grained post-image text structure than on fine-grained visual structure. Because the replaced text includes the question, instructions, and answer options, the probe does not separate harmful language priors from necessary task-interface text; distinguishing the two would require matched-layer and matched-damage controls.

Table 1: **Centroid replacement cost across seven models.** Mean accuracy cost when replacing visual or post-image text tokens with their nearest centroids ( $\alpha_{\text{interp}}=0$ ). Text replacement costs  $4.0\times$  more than visual on average. Negative values indicate that replacement *improves* accuracy. † Ratio uses  $|\text{vis. cost}|$ ; per-task breakdown in Figure 2.

| Model          | Family                           | Training | Vis. Cost   | Text Cost    | Ratio                         |
|----------------|----------------------------------|----------|-------------|--------------|-------------------------------|
| Qwen3-VL-8B    | Qwen (Bai et al., 2025a)         | SFT+     | 16.1%       | 31.5%        | $2.0\times$                   |
| Qwen3-VL-4B    | Qwen (Bai et al., 2025a)         | SFT+     | 16.2%       | 34.7%        | $2.1\times$                   |
| Qwen2.5-VL-7B  | Qwen (Bai et al., 2025b)         | SFT      | 1.4%        | 27.2%        | $19.2\times$                  |
| Qwen2.5-VL-3B  | Qwen (Bai et al., 2025b)         | SFT      | 9.1%        | 13.7%        | $1.5\times$                   |
| InternVL2.5-8B | InternVL (Chen et al., 2024c)    | MPO      | $-0.9\%$    | 29.4%        | $34.4\times^\dagger$          |
| LLaVA-OV-7B    | LLaVA (Li et al., 2024)          | DPO      | 5.2%        | 25.8%        | $4.9\times$                   |
| Idefics3-8B    | Idefics (Laurençon et al., 2024) | SFT      | $-1.5\%$    | 18.9%        | $12.7\times^\dagger$          |
| <b>Mean</b>    |                                  |          | <b>6.5%</b> | <b>25.9%</b> | <b><math>4.0\times</math></b> |

### 3.3 Consistency Across Architectures

The text–visual asymmetry is consistent across all seven models spanning four architecture families (Figure 2, Table 1). The ratio of text-to-visual cost ranges from  $1.5\times$  (Qwen2.5-VL-3B) to  $34.4\times$  (InternVL2.5-8B); for every model, text replacement is more costly when averaged across the six tasks. The magnitude varies in informative ways. SFT/DPO-labeled models show visual costs from  $-2$  to  $9\%$  and text costs from  $14$  to  $27\%$ , while the two Qwen3 SFT+ checkpoints show elevated costs for both modalities ( $16\%$  visual,  $31\text{--}35\%$  text). InternVL2.5-8B and Idefics3-8B, the two models with negative visual cost, show the lowest decision-layer sensitivity to within-cluster visual variation. These patterns show that the asymmetry is not confined to one architecture or training recipe, although architecture, data, tokenization, baseline accuracy, and training vary together and prevent attributing the differences to any one factor.

**The signal is robust to norm-tail filtering.** To test whether attention-sink or near-zero-norm tokens distort the centroid fit (Kang et al., 2025; Luo et al., 2025), we refit Qwen2.5-VL-7B under three exclusion variants. Text centroid cost remains  $+27.2\%$  across all four conditions

(unfiltered plus three variants), while mean best- $\alpha_{\text{interp}}$  CD changes by at most 0.68% (Section D.5, Table 11). Thus, the Qwen result is not driven by either norm tail.

### 3.4 Cross-Benchmark Consistency

The asymmetry is not confined to BLINK. Across VPBench, MedBLINK, and MME, the dependence *measurement* generalizes more broadly than TCCD recovery.

**Measurement.** Text versus visual replacement costs are +11.4% versus +3.7% on VPBench ( $3.1\times$  across seven models) and +9.3% versus +4.6% on MedBLINK ( $2.0\times$  across nine models, including two MedGemma checkpoints [Sellersgren et al., 2025; 2026](#)). VPBench’s documented marker sensitivity remains an untested input-level confound ([Feng et al., 2025](#)). On balanced, option-letter-free MME, text cost is at least visual cost for all ten models with validated visual-token spans (greater on nine, with one tie). For Qwen2.5-VL-7B, text replacement lowers balanced accuracy from 85.7% to 50.0%, while visual replacement leaves it at 85.8%. This excludes an A/B/C/D answer-letter space and a fixed yes bias as sufficient explanations, although text replacement may still induce a default to “no” or disrupt task-critical predicate text. A broader 10-model  $\times$  8-benchmark MCQA grid uses a different answer-token readout and is therefore reported separately as a scoring sensitivity (Section F.4). Full portfolio and MME results appear in Section F.3.

**Recovery.** Recovery is narrower than the dependence measurement. Under the primary answer-token readout, non-oracle gains occur mainly on BLINK (5 of 7 models under cross-validated  $\alpha_{\text{interp}}$ ). Auxiliary forced-choice analyses are mixed on MMBench and null on MMMU and MathVista. On free-form tasks, first-token TCCD hurts OK-VQA, centroid-replaced captioning reduces object recall, and both DocVQA replacements collapse ANLS. Complete protocols and results appear in Sections F.3, F.5 and F.11.

## 4 Correcting Modal Competition at Inference Time

The  $4\times$  text–visual asymmetry from Section 3 suggests that a text-replaced pass may provide a useful contrastive reference. We test this idea with text centroid contrastive decoding (TCCD), which can recover visual perception accuracy on some tasks and models without retraining.

### 4.1 Text Centroid Contrastive Decoding

Our method proceeds in two steps. First, we erase within-cluster residual structure at layer L12 by replacing each post-image text token with its nearest centroid, interpolated by a strength parameter  $\alpha_{\text{interp}}$  (Section 3). Second, we run both the original and replaced forward passes and contrastively decode,  $\text{logits}_{\text{cd}} = \text{logits}_{\text{orig}} + \alpha_{\text{cd}} \cdot (\text{logits}_{\text{orig}} - \text{logits}_{\text{replaced}})$ , where  $\alpha_{\text{cd}}$  controls the contrastive strength. Throughout the primary protocol, we fix  $\alpha_{\text{cd}}=1$ , as in standard VCD, LCD, DoLa, and SDCD, and tune only  $\alpha_{\text{interp}}$ . Our fixed protocol uses  $\alpha_{\text{interp}}=0.4$  (Section E.3); Section C reports the full seven-model sweep and separates leave-one-task-out and oracle best-per-task settings from the exploratory sensitivity analyses.

Table 2: **Text centroid CD results on Qwen2.5-VL-7B.** Per-task accuracy and change under oracle best per-task  $\alpha_{\text{interp}}$ , with the fixed protocol ( $\alpha_{\text{interp}}=0.4$ ) shown separately. The best-per-task column is an oracle upper bound; all significance tests use the fixed protocol.

| Task        | $n$ | Base  | +CD   | $\Delta$ best | $\Delta$ 0.4 |
|-------------|-----|-------|-------|---------------|--------------|
| Forensic    | 132 | 47.7% | 59.1% | +11.4%        | +7.6%        |
| Vis. Sim.   | 135 | 76.3% | 86.7% | +10.4%        | +8.9%        |
| Art Style   | 117 | 55.6% | 62.4% | +6.8%         | +6.8%        |
| Counting    | 120 | 66.7% | 69.2% | +2.5%         | +0.8%        |
| Depth       | 124 | 79.0% | 80.6% | +1.6%         | −2.4%        |
| Spatial     | 143 | 88.8% | 88.8% | +0%           | −2.1%        |
| <b>Mean</b> |     | 69.0% | 74.5% | +5.4%         | +3.3%        |

Table 2 shows per-task results on Qwen2.5-VL-7B. At oracle best per-task  $\alpha_{\text{interp}}$ , TCCD improves or maintains accuracy on all six tasks (mean +5.4%; +3.3% at the fixed setting). The largest oracle gains occur on Forensic Detection (+11.4%), Visual Similarity (+10.4%), and Art Style (+6.8%); we call these TEXT-COMPETES tasks. Counting (+2.5%), Spatial Relation (+0.0%), and Relative Depth (+1.6%) form the TEXT-NEEDED group. These labels and the oracle gains are defined on the same outcomes; the corresponding per-task tests are therefore nominal, while fixed-setting inference is reported in Section E.3.

**Dose-response.** To probe directionality, we sweep  $\alpha_{\text{cd}}$  from  $-1.0$  (amplifying the replaced-text reference) to  $+1.0$  (contrasting against it and effectively suppressing it; Figure 3). Negative  $\alpha_{\text{cd}}$  hurts every task by 20.5–38.3% at  $-1.0$  and monotonically on five of six (Counting deviates by one item at  $-0.3$ ); positive  $\alpha_{\text{cd}}$  selectively benefits the TEXT-COMPETES group and leaves TEXT-NEEDED near baseline. This directional, task-selective pattern is consistent with a structured signal rather than random noise, although the exploratory sweep does not by itself establish a causal mechanism.

**Answer-logit resampling control.** Repeated answer-logit sampling does not recover the same gains. We also draw option letters repeatedly from one frozen forward-pass softmax under a seed-pinned protocol (Section F.6). Mixed voting (one greedy draw plus  $k - 1$  stochastic draws at  $T=0.7$ , ties toward greedy) stays near the greedy result ( $0.0 / -0.1 / -0.2\%$  at  $\text{Vote}@ \{2, 5, 10\}$ ); pure-stochastic voting changes the mean by  $-3.2 / -3.4 / -0.7\%$ , and repeated greedy draws are identical. Unlike independent-path self-consistency, this approximately  $1 \times$ -compute control isolates sampling variance without generating separate reasoning paths (Wang et al., 2022). TCCD uses two forward passes and gains +3.3% at the fixed protocol (+5.4% oracle).

**Comparison to contrastive decoding baselines.** Table 3 compares fixed-setting runs on the same six BLINK tasks and primary answer-token readout. TCCD (+3.27%) exceeds text-only LCD (+2.43%), VCD (+2.03%), DoLa-low (+1.61%), and SDCD (+0.87%), while DoLa-high is marginally higher (+3.42% versus +3.27%, a difference of 0.15 percentage points). All methods use the same task rows, scoring, and contrastive strength, but come from separate runs; TCCD is competitive rather than uniformly best. On a separate primary-scored blank-image LCD run, TCCD and LCD make complementary errors, with 36 and 37 uniquely correct items, respectively. A label-free majority vote improves the better method by +1.3%, compared with a +4.7% label-dependent oracle union. Further baseline, scoring, MMBench, visual-side, and external-judge analyses appear in Sections F.5, F.9 and F.10.

**Practical operating points.** At a shared  $\alpha_{\text{interp}} = 0.4$ , Qwen2.5-VL-7B gains +3.3% on average. Pooling paired predictions across the six tasks, TCCD corrects 75 baseline errors while changing 50 correct answers to errors (nominal two-sided McNemar  $p = 0.025$ ;

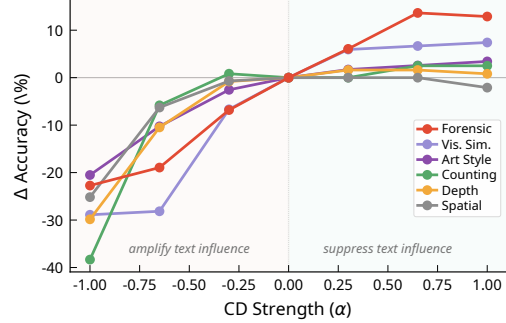


Figure 3: **Directional, task-selective dose-response** (Qwen2.5-VL-7B,  $\alpha_{\text{interp}}=0.4$ ; exploratory 11,218-image held-out COCO+ScienceQA cache,  $K=512$ ). Negative  $\alpha_{\text{cd}}$  hurts every task; positive  $\alpha_{\text{cd}}$  selectively benefits the post-hoc TEXT-COMPETES group. On the negative- $\alpha_{\text{cd}}$  side, Counting is the sole monotonicity exception.

Table 3: **Fixed-setting contrastive decoding baselines** (Qwen2.5-VL-7B, six BLINK tasks,  $n=771$ ). Mean accuracy change under the primary bare-continuation readout at  $\alpha_{\text{cd}}=1$ ; TCCD also fixes  $\alpha_{\text{interp}}=0.4$ . Methods use separate runs on the same task rows.

| Method                  | Protocol | Mean $\Delta$   |
|-------------------------|----------|-----------------|
| Text centroid CD (ours) | fixed    | +3.27%          |
| Text-only LCD           | fixed    | +2.43%          |
| VCD                     | fixed    | +2.03%          |
| DoLa low / high         | fixed    | +1.61% / +3.42% |
| SDCD                    | fixed    | +0.87%          |

asymptotic, without continuity correction). No individual task is significant after Bonferroni correction for six comparisons. In an exploratory in-sample policy sweep, applying the second forward pass to the 50% least-confident answers by top-2 logit margin retains +3.5% at  $1.5\times$  compute (blanket CD in the same experiment: +3.2% pooled at  $2\times$ ); the selection rule and coverage require held-out calibration. Across models, leave-one-task-out CV- $\alpha_{\text{interp}}$  yields a modest +0.9% mean, positive on 5 of 7 models and at or below zero on two (Qwen3-VL-4B  $-5.8\%$ , InternVL2.5-8B  $-1.8\%$ ; Table 4).

Table 4: **TCCD effect across reported training approaches.** Six-task mean baseline accuracy and TCCD change under fixed  $\alpha_{\text{interp}}=0.4$ , leave-one-task-out selection, and oracle best-per-task selection. Here  $g = (a_{\text{TCCD}} - a_{\text{base}})/(1 - a_{\text{base}})$ , averaged across the six tasks. Per-(model, task) Wilson CIs appear in Section C.

| Model                | Train. | Base | $\Delta \alpha_{\text{interp}}=0.4$ | $\Delta \text{CV-}\alpha_{\text{interp}}$ | $\Delta \text{best-}\alpha_{\text{interp}}$ | $g$   |
|----------------------|--------|------|-------------------------------------|-------------------------------------------|---------------------------------------------|-------|
| Qwen3-VL-8B          | SFT+   | 73%  | $-2.1\%$                            | $+0.9\%$                                  | $+1.6\%$                                    | +0.04 |
| Qwen3-VL-4B          | SFT+   | 76%  | $-6.8\%$                            | $-5.8\%$                                  | $-0.2\%$                                    | -0.01 |
| Qwen2.5-VL-7B        | SFT    | 69%  | $+3.3\%$                            | $+3.3\%$                                  | $+5.4\%$                                    | +0.16 |
| Qwen2.5-VL-3B        | SFT    | 55%  | $+3.9\%$                            | $+5.9\%$                                  | $+8.6\%$                                    | +0.19 |
| InternVL2.5-8B       | MPO    | 68%  | $-1.0\%$                            | $-1.8\%$                                  | $+3.1\%$                                    | +0.11 |
| LLaVA-OV-7B          | DPO    | 68%  | $+2.0\%$                            | $+1.5\%$                                  | $+4.7\%$                                    | +0.14 |
| Idefics3-8B          | SFT    | 59%  | $+0.5\%$                            | $+2.4\%$                                  | $+3.8\%$                                    | +0.09 |
| <i>SFT/DPO avg.</i>  |        |      | $+2.4\%$                            | $+3.3\%$                                  | $+5.6\%$                                    | +0.15 |
| <i>SFT+/MPO avg.</i> |        |      | $-3.3\%$                            | $-2.2\%$                                  | $+1.5\%$                                    | +0.05 |

## 4.2 Where Is the Intervention Effective?

**Positional-span ablation.** We selectively replace four positional ranges: all post-image text, its first 70% or last 30%, and the pre-visual prefix (Section D.3). These are not parsed question/option segments, and the pre-visual prefix includes template and control tokens. At  $\alpha_{\text{interp}} = 0.4$ , all-post-image replacement is strongest overall (Forensic +7.6%, Vis. Sim. +8.9%, Art Style +6.8%), while the pre-visual prefix gives Forensic +12.1%. Across doses, the pre-visual prefix is the only single range with a positive TEXT-COMPETES mean throughout, the post-image suffix helps only at reduced doses, and no single range matches the full post-image span’s +7.8% optimum. The ablation maps where the intervention is effective without assigning a semantic or causal locus.

**Layer sweep.** We apply text centroid CD at 16 individual layers (L0–L26) of Qwen2.5-VL-7B (Figure 6, Section D.1). The TEXT-COMPETES group benefits consistently from L4 through L22, while TEXT-NEEDED is hurt over much of the same range; effects diminish near the extremes. Thus, the pattern is not an artifact of choosing L12, although the groups themselves remain outcome-defined.

**Where visual representations become coarsely summarizable.** An exploratory Qwen2.5-VL-7B probe finds a visual cost of 22% at the vision-encoder output, 21% after the projector, and 2% at L16; a layer sweep places the drop between L12 and L16 ( $n=771$ ; Section F.7). This is consistent with visual states becoming centroid-summarizable by the middle of the LLM and provides visual-side support independent of the text task interface. Stage-specific fits use partly different samples, so the localization remains exploratory.

**Specificity control.** To test whether the learned centroid directions matter, we compare TCCD with exactly equal-norm random directions. At the fixed primary protocol ( $\alpha_{\text{interp}} = 0.4$ ), nearest-centroid TCCD gains +7.77% on the TEXT-COMPETES mean versus +0.24% for the random directions; across all six tasks, the means are +3.27% versus +0.40% (seed 42; Section D.2). This single-seed comparison supports sensitivity to the tested centroid direction beyond displacement magnitude alone, but one random-direction null does not establish that K-means is unique.

**Relation to distribution-shift explanations.** Any contrastive reference can shift the output distribution (Yin et al., 2025). We therefore compare TCCD with equal-norm random directions and sweep  $\alpha_{cd}$  in both directions. Their different behavior suggests that the gains are not explained by perturbation magnitude alone, but these controls do not rule out every distribution-shift explanation or identify a unique mechanism.

## 5 Quantifying Modal Dependence and Correctability

TCCD does not uniformly improve every model or task (Section 4). We therefore distinguish two complementary quantities. The primary signal is the text–visual accuracy–cost asymmetry, whose aggregate ordering also recurs on VPBench, MedBLINK, and MME (Section 3.4). TCCD recovery is a secondary measure of *correctability*: how much performance is recoverable under this particular intervention. It varies with  $\alpha_{interp}$ , checkpoint, baseline accuracy, and task composition. Together, the two quantities describe both modal dependence and how much of the measured dependence is actionable without retraining.

### 5.1 Correctability Varies Across Checkpoints

Table 4 separates recovery under a shared fixed setting, task-held-out calibration, and an oracle upper bound. TCCD improves 4 of 7 checkpoint means at fixed  $\alpha_{interp} = 0.4$ , 5 of 7 under leave-one-task-out selection, and 6 of 7 under oracle best-per-task selection. The oracle results show that the intervention exposes recoverable performance, while their gap from the non-oracle results shows that selecting its strength remains part of the problem. In this seven-checkpoint sample, the four SFT/DPO-labeled checkpoints average +5.6% under oracle selection, compared with +1.5% for the three SFT+/MPO-labeled checkpoints. Fixed (+2.4% vs. −3.3%) and leave-one-task-out (+3.3% vs. −2.2%) settings preserve this descriptive ordering, although recovery remains negative for some individual checkpoints.

Some of this variation tracks baseline headroom. Across the 42 oracle-selected model–task cells, a crossed restricted maximum-likelihood model with model and task random intercepts estimates a 0.213-percentage-point decrease in recovery for each one-percentage-point increase in baseline accuracy (SE 0.052; nominal Wald  $p = 4.57 \times 10^{-5}$ ; marginal  $R^2 = 0.379$ ). Yet after normalizing each gain by the remaining error, Hake gain is positive for 6 of 7 model averages (mean  $g = +0.103$ ; nominal two-sided one-sample  $t$ -test  $p = 0.0076$ ). Thus, under oracle tuning, TCCD recovers about 10.3% of the remaining errors on average in this sample. Recovery therefore remains positive after accounting for available headroom, so baseline performance and intervention-based correctability provide complementary summaries of a checkpoint.

These analyses do not isolate a training-recipe effect. The reported training labels covary with architecture, model family, data, and baseline performance; recovery already subtracts baseline accuracy; and both statistical summaries use oracle-selected outcomes. A controlled comparison would vary training while holding the base model and data fixed. Establishing whether TCCD correctability tracks independent gradient- or attention-based measures of modal balance also remains future work.

We therefore interpret TCCD recovery as correctability under the tested intervention, not as a direct measure of whether training has balanced the modalities. A large gain shows that some errors are recoverable without retraining under that protocol; a near-zero gain can instead reflect ceiling effects, task composition, or architecture-specific sensitivity. The primary audit signal remains the text–visual accuracy–cost asymmetry, while TCCD response adds a secondary view of how much measured dependence is actionable. Together, they provide a lightweight labeled audit that complements training-time diagnostics (Wu et al., 2025): the audit requires fitted centroid banks, labels, and one replacement pass per modality, but no retraining.

## 6 Discussion and Conclusion

**Broader implications.** Our findings suggest that the  $4\times$  text–visual asymmetry is not peculiar to one architecture: the aggregate ordering recurs across the evaluated model families and perception-focused benchmarks. This has practical consequences. In perception-critical settings such as medical imaging (Asadi et al., 2026) and agent pipelines (Heydari et al., 2025), centroid replacement offers a lightweight way to measure dependence on fine-grained text versus visual structure using a labeled audit set. For model developers, it complements ordinary accuracy at the aggregate evaluation level, although it does not certify that any individual answer is visually grounded. The decision-layer coarseness observed in Section 3.1 is also consistent with geometric explanations for visual token compression (Chen et al., 2024a; Yang et al., 2025).

**Application to monitoring.** Our probe maps onto the functionality-monitoring gap identified by NIST AI 800-4 (Rao et al., 2026). A concrete example is Qwen2.5-VL-7B on VPBench depth estimation: it reaches 64.7% accuracy while its measured visual cost is  $-0.7\%$  and text cost is  $+14.4\%$  (aggregate  $+0.6\%/+11.0\%$ , Table 20). Task-level accuracy does not surface this dissociation; centroid replacement does. Operational use would require a labeled sentinel set and periodic re-measurement.

**Limitations.** We evaluate the primary six-task analysis on BLINK validation because test-set answers are not public (Fu et al., 2024; 2025). The task grouping is exploratory, and tests after oracle  $\alpha_{\text{interp}}$  selection are nominal. Across five Qwen2.5-VL-7B centroid refits, the maximum per-task standard deviation at the primary fit’s frozen oracle  $\alpha_{\text{interp}}$  is 2.5 percentage points. On Qwen2.5-VL-7B, fixed-dose TCCD raises aggregate ECE by  $+0.028$ , reduced to  $+0.006$  under an exploratory selective-coverage policy. Demonstrated TCCD recovery is limited to discriminative, predominantly forced-choice visual QA; our generative and OCR experiments do not show a benefit. The text and visual probes use L12 and L16 with unequal token counts, without matching either the information removed or the damage induced. Post-image text also contains the task interface, so full replacement is a dependence stress test rather than evidence of a harmful language prior; the fixed TCCD protocol uses partial replacement. A fully protocol-matched same-layer comparison and a cross-modal control matched for task-relevant information loss remain future work.

**Conclusion.** Centroid replacement reveals a  $4\times$  text–visual cost asymmetry across seven multimodal language models, with the same aggregate ordering on VPBench, MedBLINK, and a balanced, option-letter-free MME control. TCCD exploits this structure to recover up to  $+16.9\%$  accuracy on an individual BLINK task without retraining, at oracle best-per-task  $\alpha_{\text{interp}}$ ; recovery varies across tasks and models under fixed and cross-validated settings. We frame the asymmetry as a labeled observational audit rather than a causal training signal or label-free monitor. An exploratory stage probe further traces visual cost from 22% at the encoder to 2% by L16. Together, centroid replacement exposes a recurring modal imbalance and turns part of it into a reusable, retraining-free intervention.

## 7 Acknowledgements

This work was supported primarily by the European Research Council under the European Union’s Horizon 2020 research and innovation program (Grant 804226, PERDY), with additional support from the U.S. National Institutes of Health (Grant AG08916) and a Stanford Institute for Human-Centered Artificial Intelligence (HAI) Hoffman-Yee Award.

## References

- Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaei, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. MIRAGE: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.
- Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3867–3876, 2025.
- Haiwen Feng, Long Lian, Lisa Dunlap, Jiahao Shu, XuDong Wang, Renhao Wang, Trevor Darrell, Alane Suhr, and Angjoo Kanazawa. Visually prompted benchmarks are surprisingly fragile. *arXiv preprint arXiv:2512.17875*, 2025.
- Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations, 2025. URL <https://arxiv.org/abs/2506.08008>.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.
- A Ali Heydari, Ken Gu, Vidya Srinivas, Hong Yu, Zhihan Zhang, Yuwei Zhang, Akshay Paruchuri, Qian He, Hamid Palangi, Nova Hammerquist, et al. The anatomy of a personal health agent. *arXiv preprint arXiv:2508.20148*, 2025.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.
- QingYuan Jiang, Longfei Huang, and Yang Yang. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. *arXiv preprint arXiv:2502.20120*, 2025.

- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 2025.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeon-hoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37: 87874–87907, 2024.
- Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: Insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. URL <https://arxiv.org/abs/2408.12637>.
- Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. *arXiv preprint arXiv:2503.21073*, 2025.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, et al. MBQ: Modality-balanced quantization for large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4167–4177, 2025.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.
- Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.
- Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models. *arXiv preprint arXiv:2510.08510*, 2025.
- Avshalom Manevich and Reut Tsarfaty. Mitigating hallucinations in large vision-language models (LVLMs) via language-contrastive decoding (LCD). In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6008–6022, 2024.
- Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, et al. Transforming wearable data into personal health insights using large language model agents. *Nature Communications*, 17:1143, 2026. doi: 10.1038/s41467-025-67922-y.
- Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces. *arXiv preprint arXiv:2504.11695*, 2025.
- Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*, 2024.
- Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8238–8247, 2022.

- Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- Anita Rao, Andrew Keller, Neha Kalra, Ryan Steed, Kweku Kwegyir-Aggrey, Kevin Klyman, Diane Staheli, and Amanda Bergman. Challenges to the monitoring of deployed AI systems: Center for AI standards and innovation, March 2026. URL <https://doi.org/10.6028/NIST.AI.800-4>.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.
- Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.
- Haz Sameen Shahgir, Xiaofu Chen, Yu Fu, Erfan Shayegani, Nael Abu-Ghazaleh, Yova Kementchedjheva, and Yue Dong. Vlms need words: Vision language models ignore visual detail in favor of semantic anchors. *arXiv preprint arXiv:2604.02486*, 2026.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024a.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.
- Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Chen Henry Wu, Neil Kale, and Aditi Raghunathan. Mitigating modal imbalance in multimodal reasoning. *arXiv preprint arXiv:2510.02608*, 2025.
- Yuxuan Xia, Siheng Wang, and Peng Li. Sdcd: Structure-disrupted contrastive decoding for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2601.03500*, 2026.
- Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025.
- Hao Yin, Guangzong Si, and Zilei Wang. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in mllms? *arXiv preprint arXiv:2504.10020*, 2025.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

## A Evaluation Details

### A.1 Models and Implementation

We evaluate seven instruction-tuned MLMs spanning four architecture families (Qwen (Bai et al., 2025a), InternVL (Chen et al., 2024c), LLaVA (Li et al., 2024), Idefics (Laurençon et al., 2024)) and four training-recipe labels reported for the checkpoints (SFT, DPO (Rafailov et al., 2023), SFT+, MPO (Zhu et al., 2025)). Table 5 lists all models with their HuggingFace identifiers and vision encoders. All models are loaded in bfloat16 precision on a single NVIDIA A6000 (48GB). Unless otherwise noted, model evaluations use greedy decoding, eliminating sampling variance, although low-level hardware and kernel nondeterminism may remain.

**Why Qwen2.5-VL-7B is the primary deep-dive model.** We center our detailed analyses (Figures 3, 8, 6 and Tables 2, 25) on Qwen2.5-VL-7B because it is a widely benchmarked open MLM at a practical 7B scale (Fu et al., 2024; 2025; Deng et al., 2025; Qi et al., 2025) and its response is sufficiently large for detailed geometric analysis. We evaluate the core dependence measurement and TCCD response across six additional checkpoints spanning three other architecture families. These cross-sectional comparisons establish breadth but do not isolate training-recipe effects.

Table 5: **Model registry.** Seven instruction-tuned MLMs spanning four architecture families, four reported training-recipe labels, and parameter scales from 3B to 8B. All use HuggingFace checkpoints.

| Model          | Family   | Train. | Size | HuggingFace ID                          | Vis. Enc. |
|----------------|----------|--------|------|-----------------------------------------|-----------|
| Qwen2.5-VL-3B  | Qwen     | SFT    | 3B   | Qwen/Qwen2.5-VL-3B-Instruct             | Qwen2-VL  |
| Qwen2.5-VL-7B  | Qwen     | SFT    | 7B   | Qwen/Qwen2.5-VL-7B-Instruct             | Qwen2-VL  |
| Qwen3-VL-4B    | Qwen     | SFT+   | 4B   | Qwen/Qwen3-VL-4B-Instruct               | Qwen3-VL  |
| Qwen3-VL-8B    | Qwen     | SFT+   | 8B   | Qwen/Qwen3-VL-8B-Instruct               | Qwen3-VL  |
| InternVL2.5-8B | InternVL | MPO    | 8B   | OpenGVLab/InternVL2.5-8B-MPO-hf         | InternViT |
| LLaVA-OV-7B    | LLaVA    | DPO    | 7B   | llava-hf/llava-onevision-qwen2-7b-ov-hf | SigLIP    |
| Idefics3-8B    | Idefics  | SFT    | 8B   | HuggingFaceM4/Idefics3-8B-Llama3        | SigLIP    |

### A.2 Centroid Fitting Procedure

We fit  $K$ -means centroids ( $K=256$ ) on hidden-state activations extracted from layer L12 (text) and L16 (visual, used for Section 3.1). For the cross-model comparison (Tables 1 and 4) we use 2,000 MS-COCO images streamed with data seed 1337,  $K$ -means seed 42, and  $\alpha_{cd}=1.0$ . Each image is rendered with the fixed harvest prompt “Describe what you see in this image.”, followed by a newline and “Answer:”; the text bank retains 16 post-image tokens per image (32,000 total). The fit uses no BLINK labels or evaluation prompts. Its generic prompt does not cover BLINK’s diverse semantics, and possible image-source overlap was not measured. This is the *primary protocol*; earlier exploratory analyses use an 11,218-image held-out COCO+ScienceQA cache with  $K=512$ . Visual and text tokens are separated by position and fitted independently. A 30-cell grid spans  $N \in [1,000, 50,000]$  and  $K \in [128, 2,048]$  (mean oracle-best  $\Delta = 5.2\% \pm 0.4\%$  across cells; Section D.4);  $N=2,000$ ,  $K=256$  is a low-compute setting near the lower end of that grid, not its minimum. The modality fits differ substantially in token count (707,414 visual vs. 32,000 text tokens before filtering; Table 11), so the comparison uses parallel rather than token-count-matched fits.

### A.3 Benchmark and Prompt Format

We evaluate on BLINK (Fu et al., 2024), a benchmark of 14 classic computer vision tasks reformulated as visual multiple-choice questions. Each sample consists of one or more input images, a text prompt with answer options, and a single correct answer. Figure 4 shows one representative example per task in the six-task analysis. We use a uniform four-letter answer-token readout across BLINK: at the final prompt position, we take the argmax over the bare continuation logits for A–D, the same position at which the contrastive-decoding logits are formed. Some BLINK prompts expose only two or three choices; on these items, selecting an undisplayed letter counts as an incorrect answer. The same scorer is applied to the baseline and every intervention pass, so all reported costs and gains are paired changes under a fixed readout (see the full template below).

**Full prompt for the Figure 1(a) sample.** We pass BLINK’s question and answer options verbatim through the model’s chat template, using the default Qwen2.5-VL system prompt. The resulting input for the Forensic Detection sample in Figure 1(a) (and identically for all Forensic Detection items) is:

**System:** You are a helpful assistant.

**User:**  You are a judge in a photography competition, and now you are given the four images. Please examine the details and tell which one of them is most likely to be a real photograph. Select from the following choices.  
 (A) the first image  
 (B) the second image  
 (C) the third image  
 (D) the fourth image

The four candidate images are concatenated horizontally into a single image before templating; the  placeholder expands to one pad token per visual token (1,314 for this sample). The answer is scored by the argmax over the option-letter logits at the final position, not by free-form generation.

### A.4 Figure 1(a) Exemplar and Attention Calculation

Figure 1a uses Qwen2.5-VL-7B on BLINK Forensic Detection item val\_Forensic.Detection\_18 (zero-based row 17 at the pinned revision). Its four candidate images expand to a 1,314-token visual span (half-open indices [15, 1329]); post-merge grid 18×73). Both the original pass and the pass with all post-image text replaced ( $\alpha_{\text{interp}}=0$ ) predict the incorrect answer A. TCCD contrasts their output logits at  $\alpha_{\text{cd}}=1$  and selects the correct answer B.

To measure attention, we reconstruct attention from the answer-scoring input position after Qwen’s multimodal rotary embedding and causal mask, using the Transformers 5.4.0 eager-attention formula while retaining SDPA for the model forward. At L22, mean-over-head attention mass on the visual span rises from 3.63% in the original pass to 11.09% in the reference. At L16/L20/L22, the respective values are 2.84/3.86/3.63% and 4.50/9.43/11.09% (means 3.44% and 8.34%). The mean rises at all three layers, although only 14/28, 15/28, and 16/28 heads increase; heads are not independent replicates.

The displayed heatmaps are softmax-normalized pre-RoPE QK scores, reshaped to 36×36 from the first 1,296 visual positions. They are qualitative views rather than the post-RoPE attention measurement reported above. The right heatmap comes from the text-replaced pass; TCCD predicts B by contrasting the logits of the two passes. This example is illustrative rather than an average result.



Figure 4: **Task examples from BLINK.** One representative sample per task showing the input image(s), prompt with answer choices, and the correct answer. Top row: TEXT-COMPETES tasks (Forensic Detection, Visual Similarity, Art Style), where TCCD yields larger gains. Bottom row: TEXT-NEEDED tasks (Counting, Relative Depth, Spatial Relation), where gains are smaller or absent.

## B Task Selection and Excluded Tasks

We focus on six of BLINK’s 14 tasks, chosen for baseline measurability and coverage of low-, mid-, and high-level perceptual demands. Several remaining tasks have low Qwen2.5-VL-7B baselines under the fixed four-letter readout, while Multi-view Reasoning and Jigsaw are higher; the six-task analysis is therefore not a complete benchmark estimate. We assign the TEXT-COMPETES/TEXT-NEEDED labels post hoc from the CD outcomes rather than use them for task selection. Table 6 reports every remaining baseline, and Section F.8 gives fixed-protocol results for all fourteen tasks on two models.

Table 6: **BLINK tasks outside the six-task analysis.** Qwen2.5-VL-7B baselines under the primary fixed four-letter readout. Several tasks have low baselines, while Multi-view Reasoning and Jigsaw are higher. Results on all fourteen tasks appear in Table 27.

| Task                      | Baseline (%) |
|---------------------------|--------------|
| Multi-view Reasoning      | 55.6         |
| Jigsaw Puzzle Solving     | 55.3         |
| Object Localization       | 45.1         |
| Visual Correspondence     | 32.6         |
| Semantic Correspondence   | 32.4         |
| Relative Reflectance      | 32.1         |
| Functional Correspondence | 17.7         |
| IQ Test                   | 19.3         |

## C Full Alpha Sweep Results

Figure 5 shows the complete centroid interpolation sweep ( $\alpha_{\text{interp}}$  from 0.0 to 0.8) for all seven models under the primary protocol ( $N=2,000$  COCO images,  $K=256$ ,  $\alpha_{\text{cd}}=1.0$ ). Each panel displays per-task CD delta as a function of erasure strength, with the fixed protocol ( $\alpha_{\text{interp}}=0.4$ ) marked by a dashed vertical line. SFT/DPO-labeled checkpoints show sharper task-selective peaks in this sample, while SFT+/MPO checkpoints are flatter. Because the checkpoints differ in other respects, this is a descriptive association rather than an isolated effect of post-training.

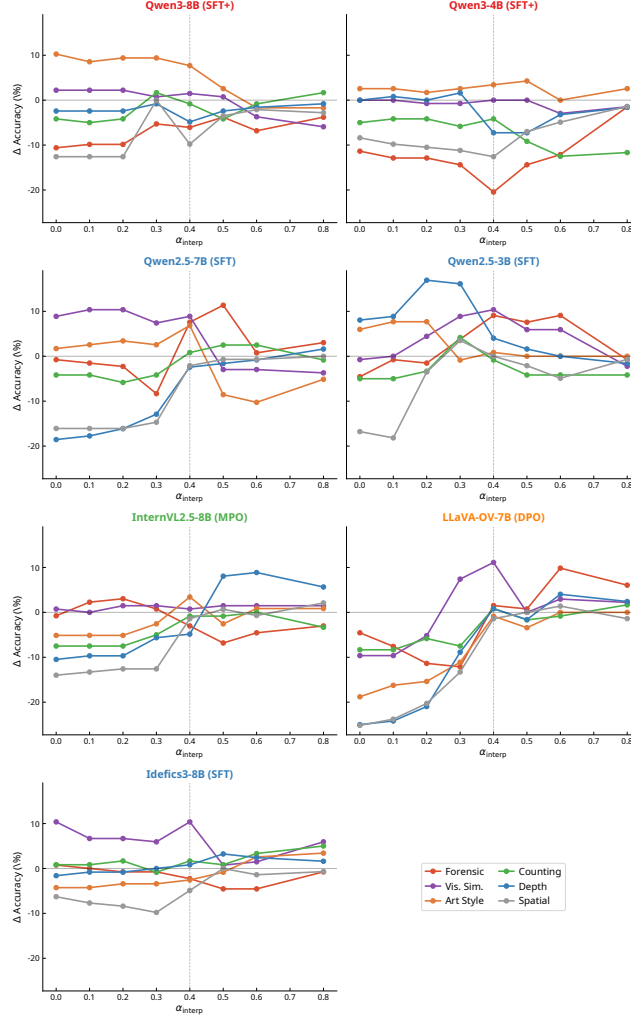


Figure 5: **Per-model alpha sweep under the primary protocol.** CD delta (%) versus centroid interpolation strength  $\alpha_{\text{interp}}$  for all seven models ( $N=2,000$ ,  $K=256$ ,  $\alpha_{\text{cd}}=1.0$ ). The dashed vertical line marks the fixed protocol,  $\alpha_{\text{interp}}=0.4$ . SFT/DPO-labeled checkpoints (Qwen2.5, LLaVA-OV, Idefics3) show sharper task-selective peaks than SFT+/MPO checkpoints in this descriptive cross-checkpoint comparison.

### C.1 Full Per-Model, Per-Task Breakdown with Wilson CIs

Table 7 reports the complete per-(model, task) breakdown at oracle best per-task  $\alpha_{\text{interp}}$ : baseline and TCCD accuracy with Wilson 95% CIs, the delta, and selected  $\alpha_{\text{interp}}$ . Since  $\alpha_{\text{interp}}$  is selected and evaluated on the same items, the table reports confidence intervals

but not inferential significance; fixed- $\alpha_{\text{interp}}$  and CV- $\alpha_{\text{interp}}$  means are in Table 4, and refit sensitivity is in Section E.4.

Table 7: **Per-(model, task) TCCD at oracle best per-task  $\alpha_{\text{interp}}$**  ( $N=2,000$  COCO images,  $K=256$ ,  $\alpha_{\text{cd}}=1.0$ ). Baseline and +CD accuracies and  $\Delta$  are in %, with Wilson 95% CIs. Same-set selection makes this an oracle analysis.

| Task                  | Base [95% CI]     | +CD [95% CI]      | $\Delta$ | $\alpha_{\text{interp}}^*$ |
|-----------------------|-------------------|-------------------|----------|----------------------------|
| <b>Qwen3-VL-8B</b>    |                   |                   |          |                            |
| Forensic              | 66.7 [58.3, 74.1] | 62.9 [54.4, 70.6] | −3.8     | 0.5                        |
| Vis. Sim.             | 87.4 [80.8, 92.0] | 89.6 [83.3, 93.7] | +2.2     | 0.0                        |
| Art Style             | 45.3 [36.6, 54.3] | 55.6 [46.5, 64.2] | +10.3    | 0.0                        |
| Counting              | 64.2 [55.3, 72.2] | 65.8 [57.0, 73.7] | +1.7     | 0.3                        |
| Depth                 | 88.7 [81.9, 93.2] | 87.9 [81.0, 92.5] | −0.8     | 0.3                        |
| Spatial               | 87.4 [81.0, 91.9] | 87.4 [81.0, 91.9] | +0.0     | 0.3                        |
| <b>Qwen3-VL-4B</b>    |                   |                   |          |                            |
| Forensic              | 71.2 [63.0, 78.2] | 69.7 [61.4, 76.9] | −1.5     | 0.8                        |
| Vis. Sim.             | 83.0 [75.7, 88.4] | 83.0 [75.7, 88.4] | +0.0     | 0.0                        |
| Art Style             | 70.9 [62.2, 78.4] | 75.2 [66.7, 82.2] | +4.3     | 0.5                        |
| Counting              | 63.3 [54.4, 71.4] | 59.2 [50.2, 67.5] | −4.2     | 0.1                        |
| Depth                 | 83.1 [75.5, 88.6] | 84.7 [77.3, 90.0] | +1.6     | 0.3                        |
| Spatial               | 87.4 [81.0, 91.9] | 86.0 [79.4, 90.8] | −1.4     | 0.8                        |
| <b>Qwen2.5-VL-7B</b>  |                   |                   |          |                            |
| Forensic              | 47.7 [39.4, 56.2] | 59.1 [50.6, 67.1] | +11.4    | 0.5                        |
| Vis. Sim.             | 76.3 [68.5, 82.7] | 86.7 [79.9, 91.4] | +10.4    | 0.1                        |
| Art Style             | 55.6 [46.5, 64.2] | 62.4 [53.4, 70.6] | +6.8     | 0.4                        |
| Counting              | 66.7 [57.8, 74.5] | 69.2 [60.4, 76.7] | +2.5     | 0.5                        |
| Depth                 | 79.0 [71.0, 85.3] | 80.6 [72.8, 86.6] | +1.6     | 0.8                        |
| Spatial               | 88.8 [82.6, 93.0] | 88.8 [82.6, 93.0] | +0.0     | 0.8                        |
| <b>Qwen2.5-VL-3B</b>  |                   |                   |          |                            |
| Forensic              | 31.1 [23.8, 39.4] | 40.2 [32.2, 48.7] | +9.1     | 0.4                        |
| Vis. Sim.             | 55.6 [47.1, 63.7] | 65.9 [57.6, 73.4] | +10.4    | 0.4                        |
| Art Style             | 47.0 [38.2, 56.0] | 54.7 [45.7, 63.4] | +7.7     | 0.1                        |
| Counting              | 68.3 [59.6, 76.0] | 72.5 [63.9, 79.7] | +4.2     | 0.3                        |
| Depth                 | 53.2 [44.5, 61.8] | 70.2 [61.6, 77.5] | +16.9    | 0.2                        |
| Spatial               | 77.6 [70.1, 83.7] | 81.1 [73.9, 86.7] | +3.5     | 0.3                        |
| <b>InternVL2.5-8B</b> |                   |                   |          |                            |
| Forensic              | 27.3 [20.4, 35.4] | 30.3 [23.1, 38.6] | +3.0     | 0.2                        |
| Vis. Sim.             | 83.7 [76.6, 89.0] | 85.2 [78.2, 90.2] | +1.5     | 0.2                        |
| Art Style             | 72.7 [63.9, 79.9] | 76.1 [67.6, 82.9] | +3.4     | 0.4                        |
| Counting              | 72.5 [63.9, 79.7] | 72.5 [63.9, 79.7] | +0.0     | 0.6                        |
| Depth                 | 69.3 [60.8, 76.8] | 78.2 [70.2, 84.6] | +8.9     | 0.6                        |
| Spatial               | 81.8 [74.7, 87.3] | 83.9 [77.0, 89.0] | +2.1     | 0.8                        |
| <b>LLaVA-OV-7B</b>    |                   |                   |          |                            |
| Forensic              | 35.6 [28.0, 44.1] | 45.5 [37.2, 54.0] | +9.8     | 0.6                        |
| Vis. Sim.             | 70.4 [62.2, 77.4] | 81.5 [74.1, 87.1] | +11.1    | 0.4                        |
| Art Style             | 71.8 [63.0, 79.2] | 71.8 [63.0, 79.2] | +0.0     | 0.6                        |
| Counting              | 70.8 [62.2, 78.2] | 72.5 [63.9, 79.7] | +1.7     | 0.8                        |
| Depth                 | 74.2 [65.8, 81.1] | 78.2 [70.2, 84.6] | +4.0     | 0.6                        |
| Spatial               | 82.5 [75.5, 87.9] | 83.9 [77.0, 89.0] | +1.4     | 0.6                        |
| <b>Idefics3-8B</b>    |                   |                   |          |                            |
| Forensic              | 30.3 [23.1, 38.6] | 31.1 [23.8, 39.4] | +0.8     | 0.0                        |
| Vis. Sim.             | 57.8 [49.3, 65.8] | 68.1 [59.9, 75.4] | +10.4    | 0.0                        |
| Art Style             | 62.4 [53.4, 70.6] | 65.8 [56.8, 73.8] | +3.4     | 0.8                        |
| Counting              | 57.5 [48.6, 66.0] | 62.5 [53.6, 70.6] | +5.0     | 0.8                        |
| Depth                 | 61.3 [52.5, 69.4] | 64.5 [55.8, 72.4] | +3.2     | 0.5                        |
| Spatial               | 85.3 [78.6, 90.2] | 85.3 [78.6, 90.2] | +0.0     | 0.5                        |

**Cross-model transfer of interpolation strength.** The sweeps above select settings separately for each model; the shared fixed protocol uses  $\alpha_{\text{interp}} = 0.4$ , while calibration is evaluated separately. To test cross-model transfer, we ran leave-one-model-out selection over the seven sweeps: for each held-out model, we selected the  $\alpha_{\text{interp}}$  that maximized mean CD delta on the other six and then applied it to the held-out model. The mean held-out delta is  $-1.4\%$ ; the two-sided Wilcoxon test gives  $p = 0.22$  (the one-sided test for positive recovery gives  $p = 0.92$ ), and no single global  $\alpha_{\text{interp}}$  is positive on average at any grid point. These results motivate prospective per-model calibration.

## D Ablation Studies

### D.1 Layer Sweep

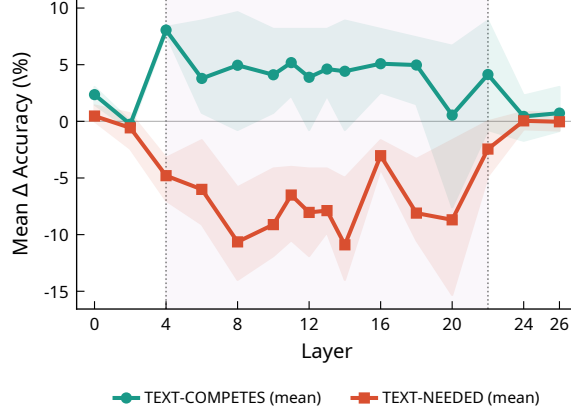


Figure 6: **Task-group separation across model depth.** Mean CD delta for TEXT-COMPETES and TEXT-NEEDED across 16 layers of Qwen2.5-VL-7B ( $\alpha_{\text{interp}}=0$ ,  $\alpha_{\text{cd}}=0.65$ ; exploratory 11,218-image held-out COCO+ScienceQA cache,  $K=512$ ). The L12 centroid bank is injected at every layer. Group means separate from L4 through L22, indicating that the pattern is not specific to one injection layer.

We apply text centroid CD ( $\alpha_{\text{interp}}=0$ ,  $\alpha_{\text{cd}}=0.65$ ) at 16 layers (L0–L26) of Qwen2.5-VL-7B to test whether the separation is localized or distributed across depth, using the same L12 centroid bank rather than refitting at each layer. Figure 7 shows per-task deltas. The TEXT-COMPETES group benefits from L4 through L22, while TEXT-NEEDED is hurt over much of the same range; effects shrink near the extremes. The separation is robust to injection layer across L4–L22. Because the groups were defined post hoc and the same L12 bank is used throughout, this sweep does not localize a mechanism.

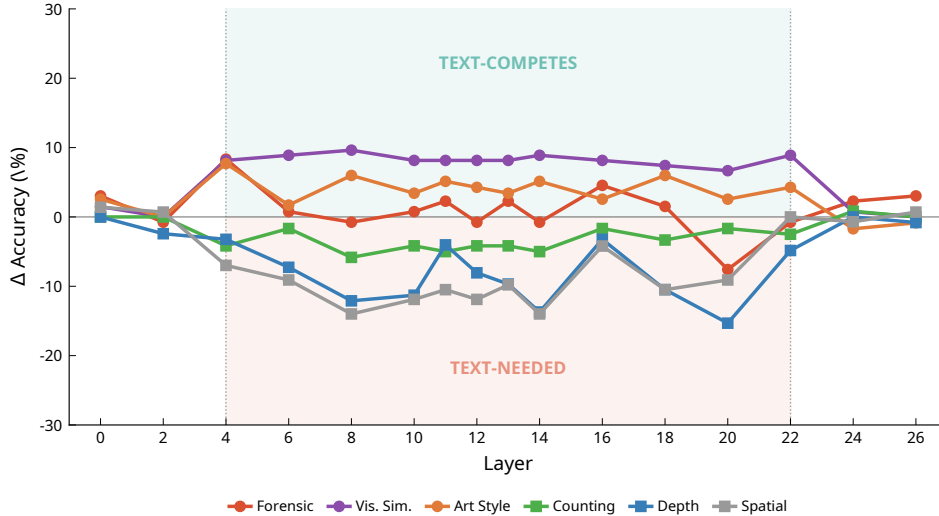


Figure 7: **Per-task layer sweep (Qwen2.5-VL-7B).** Individual task CD deltas across 16 injection layers at  $\alpha_{\text{interp}}=0$  and  $\alpha_{\text{cd}}=0.65$ , using the same L12 centroid bank and complementing the group means in Figure 6. The shaded L4–L22 interval marks the range where the group means separate; individual task trajectories vary and do not all preserve the group-level direction at every layer.

## D.2 Specificity Controls

At the fixed primary protocol (Qwen2.5-VL-7B, L12, full post-image tail,  $\alpha_{\text{interp}} = 0.4$ ,  $\alpha_{\text{cd}} = 1$ ), we compare TCCD’s nearest-centroid residual with an equal-norm random-direction reference. For each token, the random direction is scaled to exactly the residual norm, and both arms receive the same  $(1 - \alpha_{\text{interp}})$  scaling. With seed 42 set once for the deterministic run order, TCCD yields +7.77% across the three TEXT-COMPETES tasks versus +0.24% for matched random directions; across all six tasks, the means are +3.27% and +0.40% (Table 8). This equal-norm control shows that perturbation direction matters beyond displacement magnitude alone. With one seed and one null distribution, however, it does not distinguish the specific role of  $K$ -means geometry.

Table 8: **Equal-norm matched-random specificity control (Qwen2.5-VL-7B)**. Mean accuracy change (%) at the fixed primary protocol: full post-image tail, L12,  $\alpha_{\text{interp}}=0.4$ ,  $\alpha_{\text{cd}}=1.0$ , seed 42.

| Reference direction              | TEXT-COMPETES | All six tasks |
|----------------------------------|---------------|---------------|
| Nearest-centroid residual (TCCD) | +7.77         | +3.27         |
| Equal-norm random direction      | +0.24         | +0.40         |

## D.3 The Full Positional-Span-by-Dose Grid

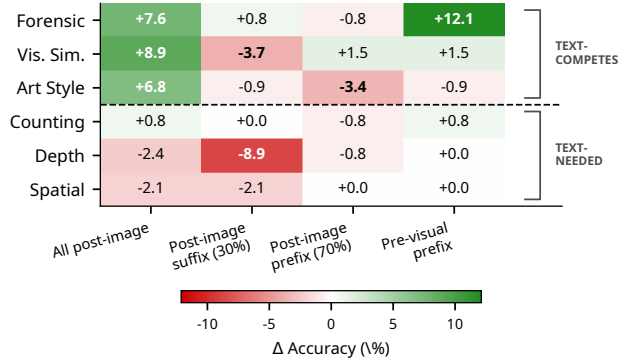


Figure 8: **Dose-dependent sensitivity across positional text spans**. CD effect when selectively applying centroid replacement to the full post-image tail, its first 70% or last 30%, or the pre-visual prefix (Qwen2.5-VL-7B,  $\alpha_{\text{interp}} = 0.4$ ,  $\alpha_{\text{cd}} = 1.0$ ). These ranges are positional heuristics rather than semantic parses. Full-post-image replacement produces the strongest gains on TEXT-COMPETES tasks (Forensic +7.6%, Vis. Sim. +8.9%, Art Style +6.8%). Pre-visual-prefix replacement yields a large gain on Forensic Detection (+12.1%). Suffix replacement at this dose is largely flat or mildly negative; at  $\alpha_{\text{interp}} = 0.3$  it recovers TEXT-COMPETES gains (Table 9).

Table 9 reports the complete positional-span  $\times$  dose grid at  $\alpha_{\text{interp}} \in \{0.2, 0.3, 0.4, 0.6\}$  under the primary protocol ( $\alpha_{\text{cd}}=1.0$ , L12,  $K=256$ ,  $n=771$ ). The implementation labels question, options, and system denote positional spans: the first 70% of the post-image tail, its last 30%, and the pre-visual prefix, respectively. The pre-visual range includes template and control tokens. All sources use the same pipeline and identical baselines; at  $\alpha_{\text{interp}}=0.4$ , post-image- and pre-visual-prefix cells reproduce the primary run to the fourth decimal. The effects vary with dose: pre-visual-prefix replacement is the only single range whose TEXT-COMPETES mean is positive at every dose (peaking at +4.3% at 0.4, with Forensic Detection from +6.8% to +12.1%), suffix replacement contributes only at reduced doses (+4.9% to +5.5% at 0.2–0.3), and post-image-prefix replacement is null to mildly negative at every dose. No single range at any dose matches full-post-image replacement at its own

optimum (+7.8% at 0.4). These results locate sensitivity by token position, not semantic role.

Table 9: **Positional-span-by-dose grid, summary (Qwen2.5-VL-7B)**. Mean CD  $\Delta$  (%) by dose for TEXT-COMPETES (top) and all six tasks (bottom). Values come from the primary full-tail sweep, suffix dose-response control, and prefix positional-grid evaluation, all using the same pipeline.

|                              | $\alpha_{\text{interp}}=0.2$ | $\alpha_{\text{interp}}=0.3$ | $\alpha_{\text{interp}}=0.4$ | $\alpha_{\text{interp}}=0.6$ |
|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| TEXT-COMPETES mean (3 tasks) |                              |                              |                              |                              |
| Full post-image tail         | +3.8                         | +0.5                         | <b>+7.8</b>                  | -4.2                         |
| Post-image suffix (30%)      | +4.9                         | <b>+5.5</b>                  | -1.3                         | -4.3                         |
| Post-image prefix (70%)      | -1.7                         | -1.5                         | -0.9                         | -0.6                         |
| Pre-visual prefix            | +1.4                         | +2.1                         | <b>+4.3</b>                  | +2.5                         |
| All-task mean (6 tasks)      |                              |                              |                              |                              |
| Full post-image tail         | -4.4                         | -5.0                         | +3.3                         | -1.9                         |
| Post-image suffix (30%)      | -0.6                         | -0.2                         | -2.5                         | -2.6                         |
| Post-image prefix (70%)      | -1.1                         | -1.3                         | -0.7                         | -0.3                         |
| Pre-visual prefix            | -0.1                         | +1.1                         | +2.3                         | +1.4                         |

Table 10: **Positional-span-by-dose grid, per-task detail** (post-image-prefix and pre-visual-prefix replacement;  $\Delta$  in %). \* marks nominal, unadjusted McNemar  $p < 0.05$  with no multiplicity correction. Post-image-suffix summaries are in Table 9; full-post-image cells are in the  $\alpha_{\text{interp}}$  sweep of Section C.

| Task                                 | $\alpha_{\text{interp}}=0.2$ | $\alpha_{\text{interp}}=0.3$ | $\alpha_{\text{interp}}=0.4$ | $\alpha_{\text{interp}}=0.6$ |
|--------------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| <b>Post-image prefix (first 70%)</b> |                              |                              |                              |                              |
| Forensic                             | +0.0                         | -0.8                         | -0.8                         | +1.5                         |
| Vis. Sim.                            | +0.0                         | +1.5                         | +1.5                         | +0.0                         |
| Art Style                            | -5.1*                        | -5.1*                        | -3.4*                        | -3.4                         |
| Counting                             | -0.8                         | -1.7                         | -0.8                         | +0.8                         |
| Depth                                | +0.0                         | +0.0                         | -0.8                         | +0.8                         |
| Spatial                              | -0.7                         | -1.4                         | +0.0                         | -1.4                         |
| <b>Pre-visual prefix</b>             |                              |                              |                              |                              |
| Forensic                             | +7.6                         | +9.1*                        | +12.1*                       | +6.8*                        |
| Vis. Sim.                            | +0.0                         | +0.7                         | +1.5                         | +0.7                         |
| Art Style                            | -3.4*                        | -3.4*                        | -0.9                         | +0.0                         |
| Counting                             | -0.8                         | +0.0                         | +0.8                         | +0.8                         |
| Depth                                | -2.4                         | +0.8                         | +0.0                         | -0.8                         |
| Spatial                              | -1.4                         | -0.7                         | +0.0                         | +0.7                         |

#### D.4 $N \times K$ Sensitivity

We ran a held-out 30-cell  $N \times K$  scaling grid (Qwen2.5-VL-7B,  $\alpha_{\text{cd}}=1.0$ , per-task best  $\alpha_{\text{interp}}$ ). The surface is essentially flat: mean best  $\Delta = 5.2\% \pm 0.4\%$  across all 30 cells with  $N \in \{1,000, 2,000, 5,000, 10,000, 20,000, 50,000\}$  and  $K \in \{128, 256, 512, 1,024, 2,048\}$ , with no monotone trend in either direction. The signal is robust to both hyperparameters in this oracle-selected grid; we use  $N=2,000$  and  $K=256$  for cost.

### D.5 Sensitivity to Sink and Dead Tokens

High-norm “sink” tokens (Kang et al., 2025; Luo et al., 2025) and near-zero-norm “dead” tokens (Kang et al., 2025) could add noise to centroid fitting and inflate the apparent sufficiency of text centroids. We test this possibility on the primary checkpoint by refitting both visual and text centroids on Qwen2.5-VL-7B under three exclusion variants applied to the 2,000-image COCO cache: (i) NO\_DEAD drops bottom-5% norm tokens; (ii) NO\_SINK drops top-1% norm tokens; (iii) NO\_EITHER applies both filters. Table 11 reports the resulting mean visual-centroid cost, mean text-centroid cost, and mean oracle best- $\alpha_{\text{interp}}$  CD delta on BLINK 6 tasks.

Table 11: **Sensitivity to sink/dead-token exclusions (Qwen2.5-VL-7B, BLINK 6 tasks).** Centroids are refit on 2,000 COCO activations under three exclusion variants. NO\_DEAD drops bottom-5% norm tokens; NO\_SINK drops top-1% norm tokens; NO\_EITHER applies both filters. Mean oracle best- $\alpha_{\text{interp}}$  CD delta changes by at most 0.68% from baseline, and the text centroid cost is identical across all four variants.

| Variant                          | Vis. Kept | Text Kept | Vis. Cost | Text Cost | Mean Best $\Delta$ |
|----------------------------------|-----------|-----------|-----------|-----------|--------------------|
| Baseline                         | 707,414   | 32,000    | +1.4      | +27.2     | +5.4               |
| NO_DEAD (bottom-5% norm)         | 672,043   | 30,400    | +0.3      | +27.2     | +4.8               |
| NO_SINK (top-1% norm)            | 700,339   | 31,680    | +1.7      | +27.2     | +5.0               |
| NO_EITHER (both)                 | 664,968   | 30,080    | +2.3      | +27.2     | +5.2               |
| $\Delta$ vs baseline (NO_DEAD)   | —         | —         | −1.08     | +0        | −0.68              |
| $\Delta$ vs baseline (NO_SINK)   | —         | —         | +0.26     | +0        | −0.40              |
| $\Delta$ vs baseline (NO_EITHER) | —         | —         | +0.89     | +0        | −0.25              |

Across the four variants, mean oracle CD delta shifts by at most 0.68%, while text cost remains +27.2%. Visual cost moves from +1.4% to +2.3% under NO\_EITHER. Thus, the Qwen result is insensitive to the tested norm-tail exclusions. This control does not test whether sink tokens explain the negative visual costs observed for InternVL2.5-8B or Idefics3-8B.

## E Statistical Analysis

### E.1 Confidence Intervals

Table 12 reports Wilson 95% confidence intervals for all six tasks on Qwen2.5-VL-7B at the fixed protocol ( $\alpha_{\text{interp}}=0.4$ ,  $\alpha_{\text{cd}}=1.0$ ,  $K=256$ ,  $N=2,000$  COCO images). Greedy decoding produces deterministic predictions, so intervals reflect finite-sample binomial uncertainty rather than run-to-run variance. Visual Similarity (+8.9%,  $h=0.227$ ) exceeds Cohen’s conventional “small” threshold of 0.2, and Forensic Detection (+7.6%,  $h=0.152$ ) and Art Style (+6.8%,  $h=0.139$ ) sit close to it. The remaining tasks show small or null effects at this fixed  $\alpha_{\text{interp}}$ ; oracle per-task selection recovers additional gains (Table 2).

Table 12: **Wilson 95% confidence intervals (Qwen2.5-VL-7B, fixed  $\alpha_{\text{interp}}=0.4$ ,  $\alpha_{\text{cd}}=1.0$ ,  $K=256$ ).** Baseline and +CD accuracies and  $\Delta$  are in %. Greedy decoding produces deterministic predictions; intervals reflect finite-sample binomial uncertainty, not run-to-run variance. Cohen’s  $h$  clears the conventional “small” threshold of 0.2 on Visual Similarity and sits near it on Forensic Detection and Art Style.

| Task      | $n$ | Base [95% CI]     | +CD [95% CI]      | $\Delta$ | $h$   |
|-----------|-----|-------------------|-------------------|----------|-------|
| Forensic  | 132 | 47.7 [39.4, 56.2] | 55.3 [46.8, 63.5] | +7.6     | .152  |
| Vis. Sim. | 135 | 76.3 [68.5, 82.7] | 85.2 [78.2, 90.2] | +8.9     | .227  |
| Art Style | 117 | 55.6 [46.5, 64.2] | 62.4 [53.4, 70.6] | +6.8     | .139  |
| Counting  | 120 | 66.7 [57.8, 74.5] | 67.5 [58.7, 75.2] | +0.8     | .018  |
| Depth     | 124 | 79.0 [71.0, 85.3] | 76.6 [68.4, 83.2] | −2.4     | −.058 |
| Spatial   | 143 | 88.8 [82.6, 93.0] | 86.7 [80.2, 91.3] | −2.1     | −.064 |

### E.2 Paired-Inference Scope

Cohen’s  $h$  in Table 12 is a descriptive difference between two proportions. Because baseline and intervention predictions are paired on the same items, independent-group power calculations do not apply; paired power depends on the discordant cells. Wilson intervals describe marginal accuracies, while McNemar tests use paired changes. Tests reported after per-task  $\alpha_{\text{interp}}$  selection are explicitly nominal and unadjusted. At fixed  $\alpha_{\text{interp}} = 0.4$ , the six-task aggregate gives  $p = 0.025$ , but no individual Qwen2.5-VL-7B task survives Bonferroni correction. Across the other checkpoints, five effects meet the two-sided per-model Bonferroni threshold: gains on Qwen2.5-VL-3B (Visual Similarity, +10.4%) and Qwen3-VL-8B (Art Style, +7.7%), and losses on Qwen3-VL-8B (Spatial Relation, −9.8%) and Qwen3-VL-4B (Forensic Detection, −20.5%; Spatial Relation, −12.6%). The fixed protocol therefore yields both gains and harms across models, consistent with Table 4.

### E.3 Fixed vs. Oracle Interpolation Strength

We report the fixed protocol ( $\alpha_{\text{interp}}=0.4$ ) separately from the same-set per-task optimum, which is an oracle analysis. Table 13 compares them on Qwen2.5-VL-7B. The fixed-protocol mean is +3.3%, the oracle mean is +5.4%, and their ratio is 0.60 using the unrounded means. Per-task selection increases the gain, while the fixed protocol preserves the same group-level pattern.

Table 13: **Fixed  $\alpha_{\text{interp}}=0.4$  vs. per-task oracle (Qwen2.5-VL-7B;  $\Delta$  in %).** Mean gains are +3.3% at fixed  $\alpha_{\text{interp}}=0.4$  and +5.4% at the same-set per-task optimum; their ratio is 0.60 using unrounded means. The task-level pattern (TEXT-COMPETES  $\gg$  TEXT-NEEDED) appears under both protocols.

| Task      | Fixed 0.4 | Oracle | $\alpha_{\text{interp}}^*$ |
|-----------|-----------|--------|----------------------------|
| Forensic  | +7.6      | +11.4  | 0.5                        |
| Vis. Sim. | +8.9      | +10.4  | 0.1                        |
| Art Style | +6.8      | +6.8   | 0.4                        |
| Counting  | +0.8      | +2.5   | 0.5                        |
| Depth     | -2.4      | +1.6   | 0.8                        |
| Spatial   | -2.1      | +0     | 0.8                        |
| Mean      | +3.3      | +5.4   | —                          |

#### E.4 Centroid Fitting Variance

We quantify centroid-fit sensitivity in two ways. First, we refit centroids five times with the primary implementation, varying only  $K$ -means seed on the fixed COCO subset (data seed 1337,  $N=2,000$ ,  $K=256$ ,  $\alpha_{\text{cd}}=1.0$ ). Table 14 evaluates each refit at the per-task  $\alpha_{\text{interp}}^*$  selected on the primary fit, and Table 15 uses fixed  $\alpha_{\text{interp}}=0.4$ . Second, a  $3 \times 5$  grid (three data subsets by five  $K$ -means seeds) was computed with the same generic harvest prompt but a separate  $K=512$  scikit-learn harvest/evaluation implementation. Table 16 summarizes within-seed and across-seed standard deviations; these are sensitivity summaries, not fitted factorial variance components. The second grid uses a different  $K$  and fitting backend from the primary  $K=256$  faiss implementation, and its absolute deltas differ by as much as  $\sim 4.5\%$ . We therefore use each implementation to assess its own fit-to-fit variation rather than compare their absolute effects.

Three observations hold for Qwen2.5-VL-7B. (1) At the fixed protocol, maximum unrounded per-task refit variation is  $\sigma = 1.83\%$ ; the separate  $3 \times 5$  grid has maximum reported  $\sigma_{\text{total}} = 1.53\%$ , and the group ordering holds in all five primary refits and all 15 separate-implementation fits. (2) Frozen oracle  $\alpha_{\text{interp}}^*$  is more sensitive (up to  $\sigma = 2.5\%$  on Forensic), with mean gains across refits from +3.8% to +5.4%. (3) Full text-replacement cost is +27.2% across the five primary refits, while visual cost ranges from +1.0% to +2.8%; the resulting ratio ranges from  $9.7 \times$  to  $26.7 \times$  without changing direction. This analysis does not refit the other six models and therefore does not establish cross-model variance.

Table 14: **Refit stability of the reported per-task deltas (Qwen2.5-VL-7B, primary implementation, data seed 1337).** CD delta (%) across five independent  $K$ -means fits (seeds 42, 800, 1337, 2024, 8320), evaluated at the frozen per-task  $\alpha_{\text{interp}}^*$  selected on the primary fit. The seed-42 column reproduces Table 2;  $\alpha_{\text{interp}}$  is not reselected for each fit. Forensic Detection has the largest sensitivity ( $\sigma=2.5\%$ ).

| Task      | km=42 | km=800 | km=1337 | km=2024 | km=8320 | $\sigma$ |
|-----------|-------|--------|---------|---------|---------|----------|
| Forensic  | +11.4 | +9.1   | +3.8    | +8.3    | +9.1    | 2.5      |
| Vis. Sim. | +10.4 | +10.4  | +9.6    | +9.6    | +10.4   | 0.4      |
| Art Style | +6.8  | +6.8   | +6.8    | +4.3    | +7.7    | 1.2      |
| Counting  | +2.5  | +2.5   | +3.3    | +2.5    | +3.3    | 0.4      |
| Depth     | +1.6  | +0.8   | -0.8    | -1.6    | +0.8    | 1.2      |
| Spatial   | +0    | +0     | +0      | +0      | +0      | 0.0      |
| Mean      | +5.4  | +4.9   | +3.8    | +3.9    | +5.2    | —        |

Table 15: **Refit stability at the fixed deployment protocol (Qwen2.5-VL-7B, primary implementation,  $\alpha_{\text{interp}}=0.4$ ).** Same five refits as Table 14; the seed-42 column reproduces Table 12. The maximum unrounded per-task  $\sigma$  is 1.83%, and the TEXT-COMPETES > TEXT-NEEDED group separation holds in all five refits.

| Task        | km=42 | km=800 | km=1337 | km=2024 | km=8320 | $\sigma$ |
|-------------|-------|--------|---------|---------|---------|----------|
| Forensic    | +7.6  | +11.4  | +12.1   | +11.4   | +8.3    | 1.8      |
| Vis. Sim.   | +8.9  | +7.4   | +8.2    | +8.2    | +8.2    | 0.5      |
| Art Style   | +6.8  | +6.8   | +6.8    | +4.3    | +7.7    | 1.2      |
| Counting    | +0.8  | +2.5   | +2.5    | +1.7    | +2.5    | 0.7      |
| Depth       | -2.4  | -3.2   | -3.2    | -2.4    | -3.2    | 0.4      |
| Spatial     | -2.1  | -0.7   | -0.7    | -1.4    | -2.1    | 0.6      |
| <b>Mean</b> | +3.3  | +4.0   | +4.3    | +3.6    | +3.6    | —        |

Table 16: **Sensitivity summary over the  $3 \times 5$  fit grid** (fixed  $\alpha_{\text{interp}}=0.4$ ,  $\alpha_{\text{cd}}=1.0$ ).  $\sigma_{\text{kmeans}}$  is the within-data-seed SD across five K-means seeds, averaged over three data seeds;  $\sigma_{\text{data}}$  is the across-data-seed SD averaged over K-means seeds;  $\sigma_{\text{total}}$  is the SD across all 15 fits. These descriptive SDs are not formal factorial variance components.

| Task       | $\sigma_{\text{kmeans}}$ | $\sigma_{\text{data}}$ | $\sigma_{\text{total}}$ |
|------------|--------------------------|------------------------|-------------------------|
| Forensic   | 1.4                      | 1.3                    | 1.5                     |
| Vis. Sim.  | 0.5                      | 0.4                    | 0.5                     |
| Art Style  | 0.9                      | 0.8                    | 1.0                     |
| Counting   | 0.6                      | 0.4                    | 0.7                     |
| Depth      | 0.5                      | 0.6                    | 0.8                     |
| Spatial    | 0.5                      | 0.5                    | 0.6                     |
| <b>Max</b> | 1.4                      | 1.3                    | <b>1.5</b>              |

## F Additional Analysis

### F.1 Calibration

Table 17 reports expected calibration error (ECE) before and after text centroid CD on Qwen2.5-VL-7B with an earlier exploratory protocol and 11,218-image held-out COCO+ScienceQA cache ( $K=512$ ,  $\alpha_{\text{interp}}=0.4$ ,  $\alpha_{\text{cd}}=0.65$ ; 10 equal-width bins). ECE increases on five of six tasks (mean  $\Delta\text{ECE} = +0.017$ ) as mean confidence rises from 0.785 to 0.841. Under the primary protocol on the full six-task set ( $n=771$ , 15 equal-width bins), TCCD at  $\alpha_{\text{interp}}=0.4$  and  $\alpha_{\text{cd}}=1$  raises aggregate ECE from 0.105 to 0.133 ( $\Delta\text{ECE} = +0.028$ ) while raising pooled accuracy by 3.2%. Applying TCCD only to the least-confident 25% of items by top-2 margin holds  $\Delta\text{ECE}$  to +0.006 while retaining +3.0% accuracy; at 50% coverage,  $\Delta\text{ECE}$  is +0.022 with +3.5%. The top-2 margin and these coverage levels were selected in the same sweep rather than on held-out data, and temperature scaling was not tested.

Table 17: **Calibration analysis under the earlier exploratory protocol** (Qwen2.5-VL-7B, 11,218-image held-out COCO+ScienceQA cache,  $K=512$ ,  $\alpha_{\text{interp}}=0.4$ ,  $\alpha_{\text{cd}}=0.65$ ; 10 equal-width bins). Text centroid CD increases ECE on five of six tasks (mean  $\Delta\text{ECE} = +0.017$ ).

| Task        | $\text{ECE}_{\text{B}}$ | $\text{ECE}_{\text{CD}}$ | $\Delta\text{ECE}$                       | $\text{Conf}_{\text{B}}$ | $\text{Conf}_{\text{CD}}$ |
|-------------|-------------------------|--------------------------|------------------------------------------|--------------------------|---------------------------|
| Forensic    | .103                    | .110                     | <span style="color: red;">+.007</span>   | .522                     | .620                      |
| Vis. Sim.   | .093                    | .085                     | <span style="color: green;">−.008</span> | .856                     | .907                      |
| Art Style   | .277                    | .299                     | <span style="color: red;">+.022</span>   | .833                     | .880                      |
| Counting    | .158                    | .185                     | <span style="color: red;">+.027</span>   | .820                     | .874                      |
| Depth       | .024                    | .061                     | <span style="color: red;">+.038</span>   | .785                     | .845                      |
| Spatial     | .037                    | .055                     | <span style="color: red;">+.018</span>   | .894                     | .919                      |
| <b>Mean</b> | .115                    | .133                     | <span style="color: red;">+.017</span>   | .785                     | .841                      |

## E.2 Cross-Task Centroid Transfer

Table 18 compares four centroid sources under two different protocols: COCO (primary,  $N=2,000$ ,  $K=256$ ,  $\alpha_{cd}=1$ , per-task best  $\alpha_{interp}$ ) and earlier exploratory  $K=512$  fits from COM-PETES-only, NEEDED-only, or all six BLINK tasks (fixed  $\alpha_{interp}=0.4$ ,  $\alpha_{cd}=0.65$ ). Task-derived banks use unlabeled input activations from the first 50 evaluation items in each contributing task and overlap the evaluation set, making the comparison transductive. COCO gives the largest mean (+5.4% vs. +1.1%/+1.3%/+0.4%), but the protocols also differ in  $K$ , fit size, evaluation pipeline, contrastive strength, interpolation selection, and fit-evaluation overlap. The table therefore does not isolate a general-geometry effect. COCO fitting uses no BLINK labels or evaluation prompts; possible image-level source overlap was not measured.

Table 18: **Cross-task centroid-source comparison (Qwen2.5-VL-7B;  $\Delta$  in %)**. COCO uses the primary protocol ( $N=2,000$ ,  $K=256$ ,  $\alpha_{cd}=1$ , per-task best  $\alpha_{interp}$ ); task-derived columns use earlier exploratory  $K=512$  fits from the first 50 evaluation inputs per task, at fixed  $\alpha_{interp}=0.4$  and  $\alpha_{cd}=0.65$ . Because the protocols and fit/evaluation overlap differ, the comparison is descriptive.

| Task        | COCO        | COMP.       | NEED.       | All 6       |
|-------------|-------------|-------------|-------------|-------------|
| Forensic    | +11.4       | +2.3        | +7.6        | +2.3        |
| Vis. Sim.   | +10.4       | +0.7        | +3.0        | +0.7        |
| Art Style   | +6.8        | +0          | +0.9        | +0          |
| Counting    | +2.5        | +0.8        | −0.8        | +1.7        |
| Depth       | +1.6        | +4.0        | −0.8        | +0          |
| Spatial     | +0          | −1.4        | −2.1        | −2.1        |
| <b>Mean</b> | <b>+5.4</b> | <b>+1.1</b> | <b>+1.3</b> | <b>+0.4</b> |

### F.3 Cross-Benchmark Consistency

The text–visual asymmetry extends beyond BLINK, while recovery is more mixed: some oracle best-per-task settings improve accuracy, but positive recovery under non-oracle selection concentrates in BLINK. We evaluate five further benchmarks spanning visually-prompted perception (VPBench; Feng et al., 2025), general perception (CV-Bench, Tong et al., 2024a; MMStar, Chen et al., 2024b; MMVP, Tong et al., 2024b), and clinical perception (MedBLINK; Asadi et al., 2026). Table 19 summarizes them; Table 20 and Table 21 provide per-model detail, including two clinically fine-tuned MedGemma models.

**VPBench.** VPBench’s visual markers can produce accuracy swings of up to 13% under design changes (Feng et al., 2025). Across 7 models  $\times$  2 tasks (depth:  $n = 1,730$ ; semantic correspondence:  $n = 1,023$ , per model), aggregate text cost is +11.4% and visual cost +3.7% ( $3.1\times$ ). Three per-task best- $\alpha_{\text{interp}}$  gains have nominal unadjusted McNemar  $p < 0.05$  after oracle selection. Because our hidden states are downstream of the visual markers, this experiment does not test robustness to marker design.

**CV-Bench, MMStar, MMVP.** This exploratory sweep uses canonical answer scoring and measures text cost only. Per-model means over the four task cells (CV-Bench depth/distance, MMStar, MMVP) span +16% to +29%, while MMVP considered alone is near zero across all seven models. Mean CD gain is +0.3% over 28 cells. Without visual costs, this sweep cannot establish a text–visual asymmetry.

**MedBLINK.** MedBLINK covers 8 clinical perception tasks on seven general checkpoints, MedGemma-1.5-4B, and MedGemma-27B (Sellersgren et al., 2026; 2025). Each model is evaluated on 1,283 items (task  $n$ : 200/141/134/200/200/88/176/144). Across 72 model–task points, aggregate text cost is +9.3% and visual cost +4.6% ( $2.0\times$ ); the direction holds in 8 of 9 model averages, with LLaVA at  $0.9\times$ . Thirteen positive best-per-task cells have nominal unadjusted McNemar  $p < 0.05$  after oracle selection; fixed  $\alpha_{\text{interp}} = 0.4$  is reported alongside. The largest auxiliary recovery is MedGemma-4B on image-enhancement detection: 50.00%  $\rightarrow$  76.87% (+26.87%) at oracle  $\alpha_{\text{interp}} = 0.5$  ( $n=134$ ; exact McNemar  $b=3$ ,  $c=39$ , nominal  $p = 5.63 \times 10^{-9}$ ). As a post-selected result, it is an upper bound rather than evidence of fixed-dose or clinical-wide benefit. Both MedGemma checkpoints show text-dominant aggregate costs, although this cross-sectional comparison cannot attribute the pattern to clinical fine-tuning.

**Additional forced-choice recovery checks.** These analyses use a space-prefixed answer-token scorer rather than the primary bare-token scorer, so we treat them as scoring sensitivities rather than primary-protocol accuracies. Qwen2.5-VL-7B gains +2.5% under strict MMBench CircularEval on the 938 base questions with all four dataset rotations (76.5%  $\rightarrow$  79.0%; exact McNemar  $b=6$ ,  $c=29$ ,  $p = 1.2 \times 10^{-4}$ ), while the five-model MMBench mean is  $-3.2\%$ , with three nominally significant negative models. Fixed-setting results are null on MMMU ( $-0.2\%$ ,  $p=0.93$ ,  $n=847$ ) and MathVista ( $+0.9\%$ ,  $p=0.60$ ,  $n=540$ ).

Table 19: **Cross-benchmark summary (costs and deltas in %).** Aggregate asymmetry is observed in every row where both modalities were measured. <sup>†</sup>The exploratory CV-Bench/MMStar/MMVP sweep measured text cost only and therefore does not establish an asymmetry. “Wins” are nominal  $p < 0.05$  counts after oracle best- $\alpha_{\text{interp}}$  selection and are not family-wise corrected.

| Benchmark                         | #Models | Vis. Cost | Text Cost | Asym        | #Sig. Wins | #Points |
|-----------------------------------|---------|-----------|-----------|-------------|------------|---------|
| BLINK (primary)                   | 7/7     | +6.5      | +25.9     | $4\times$   | 10         | 42      |
| CV-Bench+MMStar+MMVP <sup>†</sup> | 7/7     | —         | +23.2     | —           | —          | 28      |
| VPBench                           | 7/7     | +3.7      | +11.4     | $3.1\times$ | 3          | 14      |
| MedBLINK (general)                | 7/7     | +5.0      | +9.7      | $1.9\times$ | 11         | 56      |
| MedBLINK (medical)                | 2/2     | +3.1      | +8.0      | $2.5\times$ | 2          | 16      |

Table 20: **VPBench summary (7 models  $\times$  2 tasks; baseline accuracies, costs, and deltas in %).** Aggregate text cost is +11.4% versus +3.7% visual. Because marker variants were not tested, this table does not assess VPBench’s known marker sensitivity. Best- $\alpha_{\text{interp}}$  deltas and nominal win counts use oracle selection.

| Model                | Depth | Semantic | Vis. Cost | Text Cost | Asym            | Best $\Delta$ | #Sig. |
|----------------------|-------|----------|-----------|-----------|-----------------|---------------|-------|
| Qwen2.5-VL-7B (SFT)  | 64.7  | 32.9     | +0.6      | +11.0     | 17.3 $\times$   | −0.6          | 0/2   |
| Qwen2.5-VL-3B (SFT)  | 54.0  | 30.4     | +4.5      | +4.7      | 1.0 $\times$    | +3.5          | 1/2   |
| Qwen3-VL-8B (SFT+)   | 75.2  | 47.5     | +10.5     | +23.5     | 2.2 $\times$    | +0            | 0/2   |
| Qwen3-VL-4B (SFT+)   | 72.5  | 41.4     | +7.9      | +19.2     | 2.4 $\times$    | +2.3          | 1/2   |
| InternVL2.5-8B (MPO) | 59.8  | 30.8     | +0        | +7.3      | vis $\approx$ 0 | +1.0          | 1/2   |
| LLaVA-OV-7B (DPO)    | 67.6  | 30.6     | +2.0      | +11.3     | 5.6 $\times$    | +0.7          | 0/2   |
| Idefics3-8B (SFT)    | 55.4  | 25.7     | +0.2      | +2.8      | 12.1 $\times$   | +1.1          | 0/2   |

Table 21: **MedBLINK summary (9 models  $\times$  8 tasks; baseline accuracy, costs, and deltas in %).** Best- $\alpha_{\text{interp}}$  is an oracle upper bound; #Sig. counts nominal, unadjusted  $p < 0.05$  wins after selection. Aggregate text cost exceeds visual cost for 8 of 9 models (LLaVA is the exception). † The Idefics ratio uses |visual cost|. The MedGemma comparison is observational and does not isolate a fine-tuning effect.

| Model                           | Mean Base | Vis. Cost | Text Cost | Asym           | Best $\Delta$ | Fixed $\Delta$ | #Sig. |
|---------------------------------|-----------|-----------|-----------|----------------|---------------|----------------|-------|
| Qwen2.5-VL-7B (SFT)             | 54.3      | +0.8      | +7.9      | 10.4 $\times$  | +2.4          | −3.7           | 1/8   |
| Qwen2.5-VL-3B (SFT)             | 52.0      | +5.3      | +5.6      | 1.1 $\times$   | +3.4          | −0.1           | 2/8   |
| Qwen3-VL-8B (SFT+)              | 70.0      | +11.0     | +23.8     | 2.2 $\times$   | +3.2          | −1.4           | 1/8   |
| Qwen3-VL-4B (SFT+)              | 65.6      | +11.4     | +18.9     | 1.7 $\times$   | +1.4          | −2.2           | 2/8   |
| InternVL2.5-8B (MPO)            | 49.3      | +0.6      | +2.9      | 4.8 $\times$   | +3.2          | −2.3           | 1/8   |
| LLaVA-OV-7B (DPO)               | 52.8      | +7.0      | +6.5      | 0.9 $\times$   | +3.2          | −2.1           | 2/8   |
| Idefics3-8B (SFT)               | 48.5      | −0.7      | +1.9      | 2.6 $\times$ † | +3.8          | +0.9           | 2/8   |
| MedGemma-4B (clin. SFT)         | 53.8      | +0.7      | +7.5      | 11.4 $\times$  | +5.1          | −2.9           | 1/8   |
| MedGemma-27B (clin. SFT, 8-bit) | 54.8      | +5.6      | +8.5      | 1.5 $\times$   | +3.8          | −1.5           | 1/8   |

#### F.4 Auxiliary Answer-Token Scoring Sensitivity Across 10 Models and 8 Benchmarks

We also examine ten models for which the implementation has architecture-specific visual-token span finders (Qwen2.5-VL-7B/3B, Qwen3-VL-8B/4B, LLaVA-OV-7B, InternVL2.5-8B, Idefics3-8B, and Gemma-3-4B/12B/27B); the Qwen2-VL-7B fallback row is excluded. Seven MCQA benchmarks were scored with single space-prefixed letter tokens rather than the bare continuation tokens used in the primary evaluation. We therefore treat these as answer-scoring sensitivity results. Under this scorer, text cost exceeds visual cost in all 60 non-chance cells; MMVP’s ten cells remain near chance. Raw logits and predictions were not retained, so primary-protocol rescoring requires a rerun. MME instead uses the protocol-matched binary yes/no scorer and gives nine strict text-cost orderings and one tie. Measurement used full replacement ( $\alpha_{\text{interp}} = 0$ ); recovery used fixed  $\alpha_{\text{interp}} = 0.4$  and  $\alpha_{\text{cd}} = 1$ , with  $n=300$ – $1,000$  per cell.

Table 22: **Auxiliary answer-token-scoring sensitivity (10 models  $\times$  8 benchmarks).** Costs and recovery are reported in %. Seven MCQA rows use the alternative scorer defined above and therefore measure scoring sensitivity rather than primary-protocol accuracy. MME uses protocol-matched binary scoring. Significance counts are nominal and unadjusted; discordant counts were not retained.

| Benchmark     | $n$  | Text cost (%) | Vis. cost (%) | Recovery (%) | Nominal sig. +/- |
|---------------|------|---------------|---------------|--------------|------------------|
| A-OKVQA       | 1000 | +55 to +67    | 0 to +31      | −0.9         | 0 / 3            |
| SEED-Bench    | 1000 | +38 to +52    | 0 to +36      | −0.6         | 0 / 2            |
| ScienceQA-IMG | 1000 | +37 to +62    | 0 to +9       | −0.3         | 0 / 2            |
| RealWorldQA   | 436  | +17 to +39    | 0 to +19      | −0.9         | 0 / 0            |
| MMStar        | 1000 | +19 to +38    | 0 to +27      | −1.1         | 0 / 3            |
| CV-Bench      | 1000 | +5 to +45     | −1 to +22     | −5.1         | 1 / 8            |
| MME           | 1000 | +23 to +39    | −1 to +31     | −2.7         | 1 / 6            |
| MMVP          | 300  | −2 to +3      | −3 to +1      | +0.1         | 0 / 0            |

Table 23 gives the full RealWorldQA result under the alternative scorer and omits ratios when visual cost is within one item of zero. On option-letter-free MME with the protocol-matched binary scorer, Qwen2.5-VL-7B has baseline balanced accuracy 0.857, text-replaced 0.500, and visual-replaced 0.858 ( $n=1,000$ ; 500 yes/500 no). This pattern cannot be explained solely by A/B/C/D tokenization or a fixed yes bias. It does not, however, exclude intervention-induced response bias or disruption of task-critical text.

Table 23: **Auxiliary RealWorldQA sensitivity under the alternative answer-token scorer** across ten models with validated visual-token spans ( $n=436$ , full replacement,  $\alpha_{\text{interp}} = 0$ ; costs in %). Ratios with visual cost within one item of zero are reported as “vis  $\approx$  0.”

| Model          | Text cost (%) | Vis. cost (%) | Ratio           |
|----------------|---------------|---------------|-----------------|
| Qwen2.5-VL-7B  | +36.2         | +1.8          | 19.8 $\times$   |
| Qwen3-VL-8B    | +38.3         | +15.8         | 2.4 $\times$    |
| Qwen2.5-VL-3B  | +30.5         | +19.3         | 1.6 $\times$    |
| Qwen3-VL-4B    | +39.0         | +14.7         | 2.7 $\times$    |
| LLaVA-OV-7B    | +31.9         | +0.5          | 69.5 $\times$   |
| InternVL2.5-8B | +35.8         | +0.2          | vis $\approx$ 0 |
| Idefics3-8B    | +24.5         | −0.2          | vis $\approx$ 0 |
| Gemma-3-4B     | +16.5         | +0.7          | 24 $\times$     |
| Gemma-3-12B    | +26.6         | +16.7         | 1.6 $\times$    |
| Gemma-3-27B    | +27.1         | +18.1         | 1.5 $\times$    |

#### F.5 Contrastive Decoding Baselines and Auxiliary Sensitivity

Table 3 compares methods using the primary answer scorer at fixed settings: TCCD +3.27% ( $\alpha_{\text{interp}}=0.4$ ), text-only LCD +2.43%, VCD +2.03%, DoLa-low/high +1.61%/+3.42%, and

SDCD +0.87%, all at  $\alpha_{\text{cd}}=1$ . These values come from separate implementations on the same six BLINK task rows. Table 24 gives an additional sensitivity analysis. Its left block fixes  $\alpha_{\text{interp}}=0.4$  and selects  $\alpha_{\text{cd}}$  from  $\{0.2, 0.4, 0.6, 0.8, 1.0\}$  under the alternative answer-token scorer; the auxiliary LCD arm uses a black-image proxy. Raw logits and predictions were not retained, so bare-continuation rescoring requires rerunning. The right block reports primary-scored LCD/VCD on the other six models at fixed  $\alpha_{\text{cd}}=1$ . Methods are comparable within each block, not across them.

Under the alternative scorer, retaining one dataset row per MMBench base ID gives TCCD/LCD/VCD/DoLa flip-indicator correlations of  $r=0.46/0.41/0.61$  ( $n=1,176$ ); TCCD and LCD are uniquely correct on 27 and 30 items, respectively. Their oracle union reaches 86.4% versus 84.1% for the better method (baseline 82.9%), a +2.3% label-dependent upper bound. In a separate primary-scored BLINK portfolio using blank-image LCD, TCCD and LCD are uniquely correct on 36 and 37 items, the oracle union reaches 77.6% versus 72.9% (+4.7%), and an option-letter majority vote realizes +1.3%. These oracle unions are label-dependent upper bounds rather than deployable selection rules.

On alternative-scored MMBench, the same vote gains 1.0% over baseline but remains 0.2% below the better single method. The restricted CircularEval sensitivity uses the 938 bases with all four canonical dataset rotations: baseline 718/938, TCCD 741/938, exact McNemar  $b = 6, c = 29, p = 1.17 \times 10^{-4}$ . It excludes 53 three-rotation bases because a constructed fourth rotation would move label-referential fixed options; a custom four-way treatment of all 991 bases gives +2.8%. These MMBench results use canonical row groupings but not the primary answer-token scorer.

An auxiliary OPERA-inspired (Huang et al., 2024) per-letter attention-penalty screen uses the alternative scorer on five tasks ( $n=8$  each). Mean  $\Delta$  is zero for every  $\lambda \in \{0.5, 1, 5, 10, 50\}$ : all cells are unchanged through  $\lambda = 10$ , and two opposite one-item flips cancel at  $\lambda = 50$ . This small greedy single-token screen is not an implementation of OPERA, which applies its penalty during beam search; a generation-based comparison would be required.

Table 24: **Extended CD baselines.** Left: Qwen2.5-VL-7B under the alternative answer scorer at fixed  $\alpha_{\text{interp}}=0.4$ , selecting  $\alpha_{\text{cd}}$  at 0.4, by leave-one-task-out CV, or per-task oracle from  $\{0.2, \dots, 1.0\}$ . Right: primary-scored LCD/VCD on the other six models at fixed  $\alpha_{\text{cd}}=1$  (mean  $\Delta$  in %, six BLINK tasks,  $n=771$  each). Compare methods only within each block.

| Qwen2.5-VL-7B (alt.) | $\alpha_{\text{cd}}=0.4$ | CV- $\alpha_{\text{cd}}$ | Oracle | Model          | LCD   | VCD   |
|----------------------|--------------------------|--------------------------|--------|----------------|-------|-------|
| TCCD (ours)          | +1.65                    | +1.67                    | +3.22  | Qwen2.5-VL-3B  | +5.18 | +5.43 |
| Blank-image LCD      | +2.68                    | +2.55                    | +4.54  | Qwen3-VL-8B    | +3.35 | -0.97 |
| VCD                  | -0.52                    | -1.17                    | +0.27  | Qwen3-VL-4B    | +1.04 | -2.97 |
|                      |                          |                          |        | InternVL2.5-8B | -0.01 | +0.63 |
|                      |                          |                          |        | LLaVA-OV-7B    | +2.74 | -1.09 |
|                      |                          |                          |        | Idefics3-8B    | +0.45 | +1.77 |

## F.6 Answer-Logit Resampling: Per-Task Detail

Table 25 gives the mode-level summary and Table 26 per-task deltas. Seed 42 is used; all option-letter draws come from one frozen forward-pass softmax at  $T=0.7$  with no top- $p$  truncation, and ties favor the first inserted draw (greedy in mixed mode). This adds negligible model compute beyond one forward pass. Because all draws reuse one softmax, it is an answer-logit resampling control rather than conventional self-consistency across generation paths.

## F.7 Three-Stage Sufficiency Probe and Visual Coarsening Curve

This exploratory Qwen2.5-VL-7B probe uses full replacement,  $K=256$ , and  $n=771$  BLINK items. Raw-encoder and post-projector centroids were fit on 300 COCO images; the L16 point reuses the primary 2,000-image bank. Costs are 22% at the raw vision-encoder output,

Table 25: **TCCD vs. seed-pinned answer-logit resampling (Qwen2.5-VL-7B)**. Mean  $\Delta$  vs. greedy (%), six BLINK tasks ( $n=771$ ). All votes reuse one frozen softmax ( $\approx 1\times$  model compute); TCCD uses two forward passes.

| Method                                                                        | Vote@2 | Vote@5 | Vote@10 |
|-------------------------------------------------------------------------------|--------|--------|---------|
| Mixed draws ( $\approx 1\times$ )                                             | 0.0    | -0.1   | -0.2    |
| Stochastic draws ( $\approx 1\times$ )                                        | -3.2   | -3.4   | -0.7    |
| Greedy repeats ( $\approx 1\times$ )                                          | 0.0    | 0.0    | 0.0     |
| Text centroid CD, fixed $\alpha_{\text{interp}}=0.4$ ( $2\times$ )            |        | +3.3   |         |
| Text centroid CD, oracle best-per-task $\alpha_{\text{interp}}$ ( $2\times$ ) |        | +5.4   |         |

Table 26: **Per-task answer-logit-resampling deltas** (% vs. greedy, Qwen2.5-VL-7B). Mixed = one greedy plus  $k - 1$  stochastic draws; stochastic =  $k$  draws.

| Task      | Mixed |       |       | Pure-stochastic |       |       |
|-----------|-------|-------|-------|-----------------|-------|-------|
|           | @2    | @5    | @10   | @2              | @5    | @10   |
| Forensic  | 0.00  | -0.76 | +0.76 | -9.09           | -4.55 | 0.00  |
| Vis. Sim. | 0.00  | +0.74 | -1.48 | -0.74           | -2.96 | +0.74 |
| Art Style | 0.00  | +0.85 | +0.85 | 0.00            | -1.71 | +0.85 |
| Counting  | 0.00  | 0.00  | 0.00  | -4.17           | -0.83 | 0.00  |
| Depth     | 0.00  | -0.81 | 0.00  | -2.42           | -6.45 | -3.23 |
| Spatial   | 0.00  | -0.70 | -1.40 | -2.80           | -4.20 | -2.80 |

21% after the projector, and 2% at L16. A layer-fit sweep gives L0/L4/L8/L12/L16/L20/L24 costs of 24/23/22/21/2/0/-0%. The drop suggests visual states become centroid-summarizable between L12 and L16, but stage, fit sample, and prompt context are not fully matched, so this is not a controlled localization of modal competition.

## F.8 All Fourteen BLINK Tasks

Table 27 reports fixed-protocol TCCD on all fourteen BLINK tasks for Qwen2.5-VL-7B and Qwen3-VL-8B. Across the eight remaining tasks, the descriptive mean of sixteen model-task deltas is -0.9%; we do not pool a McNemar test across models because the same items recur. Multi-view Reasoning is not assigned to either task group and reaches an oracle +7.5% on Qwen2.5-VL-7B ( $\alpha_{\text{interp}}=0.8$ ; nominal unadjusted  $p=0.012$  after selection). Functional Correspondence’s fixed +12.3% sits on a low 17.7% baseline under the fixed four-letter readout and may therefore reflect a floor effect.

## F.9 Symmetric Visual-Side Contrastive Decoding (VIS-CCD)

The symmetric intervention contrasts a *visual*-centroid-erased reference under the same protocol. At oracle best-per-task  $\alpha_{\text{interp}}$ , the VIS-CCD/TCCD mean-delta ratio is 0.24 on InternVL2.5-8B (+0.75% vs. +3.15%), 0.76 on LLaVA-OV-7B (+3.54% vs. +4.68%), and 1.03 on Qwen2.5-VL-7B (+5.59% vs. +5.45%). These post-selected comparisons show that intervention-level asymmetry is weaker and more model-dependent than measurement-level asymmetry, without a consistent advantage for the text-side intervention.

## F.10 Sighted-Blind Judge Contrast on Selected Flips

For 79 successful TCCD flips selected under the auxiliary answer-token scorer (37 BLINK and 42 MMBench), two external judges answered each original MCQA item with and without the image. The May 2026 API run used gemini-3.1-pro-preview and gemini-3.1-flash-lite at temperature 1.0. Both prompts required A/B/C/D-only answers, and the embedded question and options were identical. Pro agreed with the corrected an-

Table 27: **All fourteen BLINK tasks at fixed  $\alpha_{\text{interp}}=0.4$**  under the fixed four-letter readout (baseline and  $\Delta$  in %; McNemar  $p$ ). Top: the six analyzed tasks; bottom: the remaining eight. Several Qwen3-VL-8B baselines in the lower block are substantially higher than the corresponding Qwen2.5-VL-7B baselines.

| Task                      | $n$ | Qwen2.5-VL-7B |              |       | $p$ | Qwen3-VL-8B |              |       |
|---------------------------|-----|---------------|--------------|-------|-----|-------------|--------------|-------|
|                           |     | Base (%)      | $\Delta$ (%) |       |     | Base (%)    | $\Delta$ (%) | $p$   |
| Forensic Detection        | 132 | 47.7          | +7.6         | .096  |     | 66.7        | −6.1         | .011  |
| Visual Similarity         | 135 | 76.3          | +8.9         | .014  |     | 87.4        | +1.5         | .317  |
| Art Style                 | 117 | 55.6          | +6.8         | .088  |     | 45.3        | +7.7         | .003  |
| Counting                  | 120 | 66.7          | +0.8         | .655  |     | 64.2        | −0.8         | .739  |
| Relative Depth            | 124 | 79.0          | −2.4         | .491  |     | 88.7        | −4.8         | .034  |
| Spatial Relation          | 143 | 88.8          | −2.1         | .491  |     | 87.4        | −9.8         | .006  |
| Multi-view Reasoning      | 133 | 55.6          | +0.0         | 1.000 |     | 55.6        | +0.0         | 1.000 |
| Jigsaw                    | 150 | 55.3          | −2.0         | .706  |     | 56.0        | −4.0         | .366  |
| Object Localization       | 122 | 45.1          | −1.6         | .739  |     | 63.9        | +0.0         | 1.000 |
| Visual Correspondence     | 172 | 32.6          | −2.9         | .446  |     | 77.3        | −2.9         | .025  |
| Semantic Correspondence   | 139 | 32.4          | −2.9         | .493  |     | 50.4        | +1.4         | .480  |
| Relative Reflectance      | 134 | 32.1          | +0.7         | .796  |     | 44.8        | −2.2         | .549  |
| Functional Correspondence | 130 | 17.7          | +12.3        | .011  |     | 42.3        | −4.6         | .221  |
| IQ Test                   | 150 | 19.3          | −2.0         | .655  |     | 26.7        | −4.0         | .157  |

swer on 93.7% of sighted calls versus 48.1% of blind calls (per-source gaps: BLINK +43.2%; MMBench +47.6%); Flash-Lite gave 81.0% versus 44.3%. Sighted raw inter-judge agreement was 65/79, with reported  $\kappa = 0.645$ ; the paired  $4 \times 4$  table needed to reconstruct  $\kappa$  was not retained. Sighted and blind calls also used different thinking levels and output-token caps. Because sighted and blind calls differ in both image availability and inference settings, and ground truth defines the selected flip set, this comparison supports the plausibility of the corrections but does not isolate image dependence or estimate net recovery. A complementary blank-image comparison is inconclusive: +2.4% with the image versus +0.6% without (interaction +1.8%; one-sided per-sample sign test, 92 positive versus 78 negative,  $p = 0.159$ ,  $n=771$ ).

### F.11 Open-Ended Generation and DocVQA

The generative checks use Qwen2.5-VL-7B and the primary centroid bank. On the first 2,000 OK-VQA validation items, first-token-only TCCD (text L12,  $\alpha_{\text{interp}} = 0.4$ ,  $\alpha_{\text{cd}} = 1$ , greedy continuation up to 8 tokens) lowers canonical VQA soft accuracy by 6.6% (paired  $t(1999) = -8.454$ ,  $p = 5.33 \times 10^{-17}$ ); exact match falls 7.2% (exact McNemar  $b=224$ ,  $c=80$ ,  $p = 6.06 \times 10^{-17}$ ). On 2,000 streamed COCO validation images, a second generate pass with the L12 replacement hook active throughout ( $\alpha_{\text{interp}} = 0.4$ , up to 48 tokens, prompt “Describe this image in a few sentences.”) lowers object recall by 2.9% (paired  $t(1999) = -5.799$ ,  $p = 7.72 \times 10^{-9}$ ) with  $\Delta\text{CHAIR}_i = -0.001$ . The captioning arm labeled tccd applies centroid-replaced generation rather than contrastive decoding and therefore does not use  $\alpha_{\text{cd}}$ . On DocVQA ( $n=300$ ), greedy prefill-only full replacement at text L12 or visual L16 collapses ANLS (baseline 0.945; text 0.00; visual 0.06). A gold-string containment diagnostic gives visual cost 0.90 versus text cost 0.71, reversing the perception-competition pattern where the image carries the answer text. Together, these checks show no benefit in the tested generative and OCR settings.

## E.12 Method Positioning

Table 28 compares centroid replacement with the prior work discussed in Section 2, organized by each method’s intervention locus and use. Centroid replacement operates on modality-specific hidden states at intermediate layers and serves first as a measurement instrument; TCCD uses the text-replaced pass as its contrastive reference.

Table 28: **Method positioning.** Centroid replacement operates on intermediate hidden-state centroids, used first as a measurement instrument for modal dependence; prior methods operate at other loci or are descriptive.

| Method                                                                | Operates on                                | What it does                                                                                                 |
|-----------------------------------------------------------------------|--------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| Centroid replacement / TCCD (ours)<br>LCD (Manevich & Tsarfaty, 2024) | hidden states (mid-stack)<br>output logits | measures text-vs-visual dependence; contrasts a text-erased pass<br>contrasts the text-only LLM distribution |
| VCD (Leng et al., 2024)<br>Fu et al. (2025)                           | input pixels<br>analysis                   | contrasts a noise-distorted image<br>visual information encoded but unused                                   |
| Wu et al. (2025)<br>Deng et al. (2025); Li et al. (2025)              | training time<br>behavior; quantization    | instance-level modality mixing<br>text-over-vision bias; balanced quantization                               |
| Huh et al. (2024); Qi et al. (2025);<br>Papadimitriou et al. (2025)   | representation<br>geometry                 | cross-modal geometry, descriptive                                                                            |

## **G LLM Disclosure**

LLMs assisted with prose editing, code review and implementation, and plotting author-generated data. The authors determined the scientific content and verified the resulting prose, figures, references, software changes, and experimental results. LLMs were also used as external judges only in the post-hoc Gemini analysis in Section [F.10](#); otherwise, they did not generate experimental data or determine the paper’s claims.