

Anchored by an Afterthought: Cross-Modal Numerical Anchoring in VLMs

Taehyeon Park¹ Jeonghun Park² Youngjae Yu²

¹Yonsei University ²Seoul National University
namam3gy@yonsei.ac.kr, {top321902, youngjaeyu}@snu.ac.kr

Abstract

When a vision-language model (VLM) is shown an irrelevant image containing a digit (the *anchor image*) alongside a numeric question, the model’s numeric answer shifts toward that digit. This matters because multi-image inputs (retrieval-augmented QA, multi-screenshot analysis, VLM-as-judge) routinely contain irrelevant numeric content. We call this phenomenon *cross-modal numerical anchoring*: across six VLMs and five numeric VQA datasets, the per-model mean SDF (signed direction-follow rate) ranges from 3 % to 24 %, with the effect growing stronger as the model’s confidence in the base answer decreases. Masking only the digit pixels of the anchor image, leaving the rest of the image intact, recovers the non-anchored answer, isolating the *digit pixels* as the causal channel, with the effect concentrated at *low confidence*. We propose a mitigation that uses this anchored-masked contrast as a calibration signal: at a single per-model integration site, we estimate a low-rank subspace spanning the anchoring directions and remove it via a projection at inference time. On LLaVA-OneVision-7B, the projection reduces anchoring ($\Delta\text{SDF}_a = -1.64$ pp, $\text{rel}\% -30.0$) without degrading accuracy on non-anchored inputs or average performance on six capability benchmarks (within ± 1 pp).

1 Introduction

Anchoring, the phenomenon in which an arbitrary cue presented before a decision pulls that decision toward itself, is a well-established cognitive bias in both human decision-making (Tversky and Kahneman, 1974; Mussweiler and Strack, 1999) and text-only large language models (LLMs) (Jones and Steinhardt, 2022; Echterhoff et al., 2024; Lou and Sun, 2026). A cue in any input modality could act as an anchor; visual anchoring in VLMs has not been evaluated systematically. We focus on the visual modality for two reasons: vision-language is

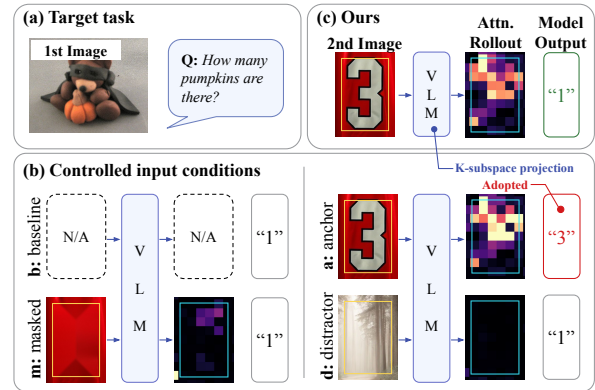


Figure 1: Illustrative overview of the four-condition stimulus setup and mitigation procedure. The same target image and question are evaluated under four second-image conditions: target-only baseline (b), anchor digit (a), digit-masked anchor (m), and neutral distractor (d). The visible digit pulls the output toward the anchor; our mitigation projects out the $(a - m)$ residual subspace at inference time.

the most common and highest-impact multimodal setting, and the visual cue is processed through a routing pathway (vision encoder + cross-modal fusion) that has no analogue in text-only anchoring. The gap matters in practice, since anchoring-like effects can routinely arise in multi-image deployment scenarios such as retrieval-augmented QA, multi-screenshot analysis, and VLM-as-judge. We refer to this phenomenon as *cross-modal numerical anchoring* and study it on six open-weight VLMs and five numeric VQA datasets, which we refer to as the main panel throughout the paper, providing behavioral measurement, causal isolation, and an inference-time mitigation.¹

Our stimulus extends numeric image-QA by adding a second image as the manipulated cue, giving a 4-condition design (b: target only; a: target with anchor; m: target with masked anchor; d: target with neutral distractor; Figure 1). The effect

¹Code: <https://github.com/SNU-PI/vlm-anchoring>

is graded: per-model mean SDF ranges from 3 to 24 % while per-model mean Adopt ranges from 2 to 19 %, with SDF exceeding Adopt on every model (ratio 1.2–1.9 $\times$), indicating a directional shift rather than literal copying. The pull also increases monotonically as the model’s confidence in the base answer decreases. To isolate the cause, we use the *anchored–masked contrast* ($a - m$): m reuses the same anchor image as a but masks the digit pixels. This contrast collapses the anchoring effect to the level of the neutral distractor d , narrowing the cause to the visible digit, whose effect is concentrated at low confidence. At the mechanism level, the anchoring signal is distributed redundantly across layers and is integrated at a per-model site.

We then propose a deployable mitigation that reduces the anchoring effect on five datasets without compromising accuracy on non-anchored inputs or average performance on six general-capability benchmarks. The mitigation reuses the anchored–masked contrast as the calibration signal: from the residual-stream difference between anchored and masked inputs, we estimate a low-rank subspace of those differences in a single calibration pass and remove it via a projection at inference time. Calibration is performed once in a separate phase and the resulting projection is then applied to every sample at inference, so neither anchor detection nor labeling is required at deployment.

Our contributions are threefold.

- **Phenomenon.** We isolate digit pixels as the causal channel through which an irrelevant image pulls a VLM’s numeric answer, via the anchored–masked contrast, with the effect concentrated at low confidence. Anchoring is positive on every cell of the main panel, and on the masked counterpart the effect collapses to the level of the neutral distractor.
- **Mechanism evidence.** We show that the anchoring representation is distributed redundantly across layers and multi-directional within layer at the integration site. These findings motivate the two design axes of the mitigation: the per-model integration site and the multi-direction subspace.
- **Mitigation.** We propose an anchor-label-free, deployable mitigation that uses the anchored–masked contrast ($a - m$) as the calibration signal. On LLaVA-OneVision-7B, the projection reduces anchoring on all five datasets while

preserving accuracy on non-anchored inputs and six general capability benchmarks. Cross-architecture verification on Qwen2.5-VL-7B replicates the mitigation chain at its own per-architecture cell.

2 Related Work

2.1 Anchoring in cognition and text LLMs

Anchoring is a canonical human bias whereby arbitrary or task-irrelevant values pull later numerical estimates toward themselves (Tversky and Kahneman, 1974; Mussweiler and Strack, 1999), with the effect modulated by uncertainty and persisting across non-identical cue formats such as physical magnitudes (Jacowitz and Kahneman, 1995; Oppenheimer et al., 2008). Recent work studies the same bias in text-only LLMs, documenting systematic cognitive-bias failures (Jones and Steinhardt, 2022; Hagendorff et al., 2023; Echterhoff et al., 2024), direct numeric anchoring from textual hints with incomplete prompt-level debiasing (Lou and Sun, 2026), and synthetic evidence on text-side mechanisms (Huang et al., 2026). We study the same uncertainty-modulated anchoring with the cue placed in an image rather than in the text prompt: the cue is an independently rendered digit, and the response is an open numerical VQA answer.

2.2 Cognitive bias and behavioral analysis of VLMs

Controlled-stimulus probes have exposed several VLM biases. VLMBias finds that familiar-subject priors override visual evidence in counting (Vo et al., 2026); AIpsych categorizes VLM hallucinations under cognitive-psychology labels such as sycophancy (Liu et al., 2025); Tinted Frames shows that question framing redirects visual attention (Fan et al., 2026); and longer reasoning does not automatically shield VLMs from visual distractors (Bae et al., 2025). We differ in cue source: our cue is a second image, sampled independently of the target, that contains only a rendered digit—not a target-subject property or a linguistic frame.

2.3 Visual cue manipulation in VLMs

Rendered text can function as a semantic visual signal: CLIP multimodal-neuron analyses revealed shared high-level representations across textual and visual depictions and introduced typographic attacks (Goh et al., 2021). Subsequent work extends this to multi-image typographic attacks (Wang et al.,

2025), image-rendered visual jailbreaks (Gong et al., 2025), encoder-side defenses against typographic circuits (Hufe et al., 2026), and causal mechanism-to-mitigation analyses of VLM bias (Weng et al., 2024). Our cue is also visual text-like content but carries a single numeric value rather than a class label, prompt, or harmful instruction; the outcome is an open-numeric, baseline-relative shift rather than a classification flip or attack success rate; and the paired masked condition isolates the digit pixels from second-image distraction.

2.4 Representation-level intervention: activation steering and concept erasure

Our mitigation belongs to the family of inference-time representation interventions, including activation addition from prompt-pair differences (Turner et al., 2024), Contrastive Activation Addition with single steering directions (Rimsky et al., 2024), reading-vector extraction via PCA on contrastive activation differences (Zou et al., 2023), attention-head interventions for truthfulness (Li et al., 2023a), and linear concept erasure (LEACE) (Belrose et al., 2023). We specialize representation-level intervention to cross-modal numerical anchoring along three axes: the calibration signal is the image-causal ($a - m$) contrast; the intervention site is the VLM language-model residual stream rather than a vision-encoder typographic circuit or selected attention heads; and we use a K -dimensional projection subspace because the anchoring representation is multi-direction within layer, which a rank-1 erasure or a single steering direction cannot fully capture. Because such interventions can affect unrelated behaviors (Chand et al., 2026), we evaluate capability preservation on six general-purpose VLM benchmarks.

3 Cross-Modal Anchoring: Stimulus and Measurement

3.1 4-condition stimulus (b / a / m / d)

For each question we construct four stimulus conditions (Table 1). Each condition isolates a distinct causal pathway: ($a - b$) is the total effect of adding the anchor, ($a - m$) is the effect of the digit pixels alone, and ($d - b$) controls for second-image distraction. We refer to the four conditions as *arms* (the b-, a-, m-, and d-arm), in the sense of the arms of a controlled experiment, and use *condition* and *arm* interchangeably. In the two-image arms (a, m, d) the target image is presented first and the

distractor image second; an image-order control with the reversed order shows that the anchoring effect is not a recency artifact of this ordering (Appendix E). The prompt template, the masked-image generation procedure, and stimulus examples are in Appendices A and O.

Symbol	Stimulus	Role
b	target only	baseline
a	target + anchor digit image	anchored stimulus
m	target + digit-masked anchor image	digit-pixel ablation of (a)
d	target + digit-free neutral image	distraction control

Table 1: The four stimulus conditions. In (**m**), the anchor image of (**a**) is reused with only the digit pixels masked out.

3.2 Anchored-masked contrast

The *anchored-masked contrast*—the per-sample difference between the a-arm and the m-arm, denoted ($a - m$)—serves two roles: it is the causal test of the phenomenon and the calibration signal of the mitigation. **a** and **m** are the same anchor image except for the digit-pixel region, so background, layout, and two-image input load are held fixed and ($a - m$) isolates the effect of the digit pixels alone. Second-image distraction is controlled separately by the d-arm, and $m \approx d$ is a falsifiable prediction: it is falsified if the panel-mean ($m - d$) on the primary metric SDF exceeds +0.3 pp; this is approximately one panel-mean standard error under our sample sizes (Appendix D.3), so a residual within this bound is empirically indistinguishable from sampling noise.

3.3 Metrics

We define anchoring as the directional shift of the prediction from the base answer (p^b) toward the anchor (z). The primary metric is Signed Direction-follow (SDF), the mean signed pull on $[-1, +1]$. The secondary metrics are Adopt, the literal-copy rate, Exact-match (EM, accuracy), and Pull, the magnitude-aware companion of SDF. Each metric is defined identically for every condition $c \in \{b, a, m, d\}$.

Notation. For each sample i , let gt_i be the ground truth, z_i the anchor digit value, and p_i^c the numeric prediction under condition $c \in \{b, a, m, d\}$.

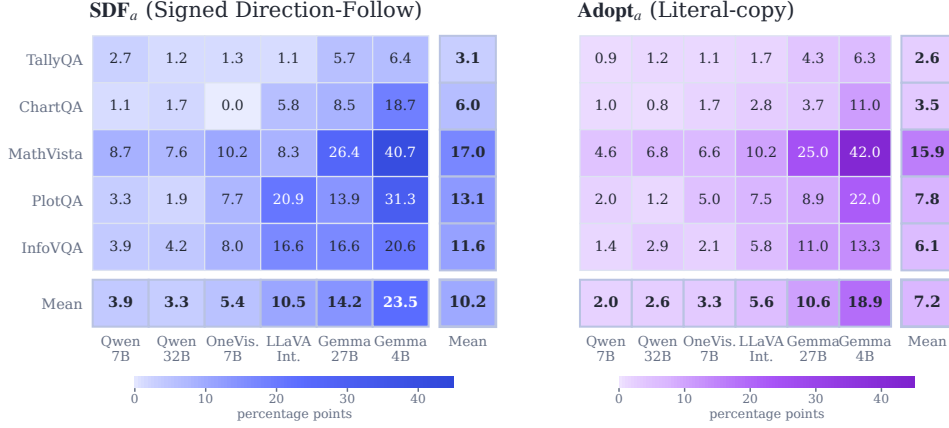


Figure 2: Cross-dataset anchoring across the main panel (per-cell values in Appendix D). The two panels show the same grid under the two metrics: the left panel reports the signed direction-follow rate SDF_a, the right panel the literal-copy rate Adopt_a. Both are positive in all cells, with SDF_a larger than Adopt_a on every model (per-model ratio 1.2–1.9×), indicating that anchoring is graded pull rather than exact copying.

For each c :

$$\begin{aligned}
 \text{SDF}_c &= \mathbb{E}_i[\text{sign}((p_i^c - p_i^b)(z_i - p_i^b)) \mid p_i^b \neq z_i], \\
 \text{Adopt}_c &= P[p_i^c = z_i \mid p_i^b \neq z_i], \\
 \text{EM}_c &= P[p_i^c = g_i^t], \\
 \text{Pull}_c &= \mathbb{E}_i\left[\text{clip}\left(\frac{p_i^c - p_i^b}{z_i - p_i^b}, -1, +1\right) \mid p_i^b \neq z_i\right].
 \end{aligned}$$

SDF, Adopt, and Pull exclude $p_i^b = z_i$ rows, which have no room for anchoring; EM is computed over all samples. The same definitions apply to all four arms. Only the a-arm shows the anchor digit, so on the b-, m-, and d-arms these values reflect chance and any incidental overlap between the model’s prediction and z rather than anchoring, providing the no-anchor baseline against which the a-arm effect is read. Pull extends SDF by also crediting how far the prediction moves toward the anchor, clipped at ± 1 (definition and rationale in Appendix G).

3.4 Models and datasets

Models. LLaVA-OneVision-7B (Li et al., 2025a), LLaVA-Interleave-7B (Li et al., 2025b), Qwen2.5-VL-7B / 32B (Bai et al., 2025), and Gemma3-4B / 27B (Gemma Team, 2025). **Datasets.** TallyQA (Acharya et al., 2019), ChartQA (Masry et al., 2022), PlotQA (Methani et al., 2020), InfoVQA (Mathew et al., 2022), and MathVista (Lu et al., 2024). Specs, filters, and the anchor inventory are in Appendices B and C.

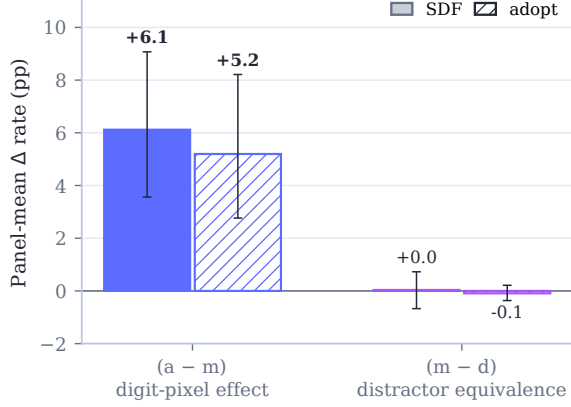
4 Phenomenon

4.1 Graded pull, not literal copy

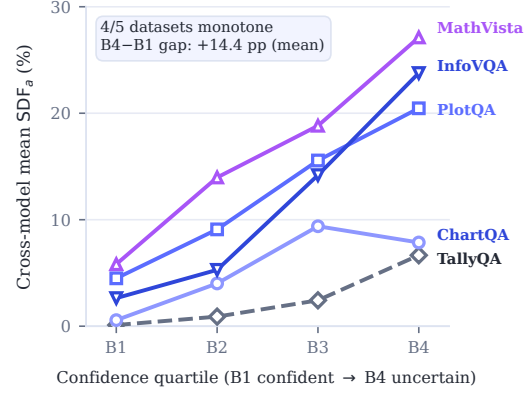
The anchoring effect is observed on all cells of the main panel (Figure 2). The per-model mean SDF ranges from **3.3 %** (Qwen2.5-VL-32B) to **23.5 %** (Gemma3-4B), and the Adopt range from **2.0 %** (Qwen2.5-VL-7B) to **18.9 %** (Gemma3-4B). SDF exceeds Adopt on every model in the panel (ratio **1.2–1.9×**, mean 1.5×), indicating that the anchoring effect lies in *graded pull* (SDF) rather than *literal copy* (Adopt). Within each family the smaller model is more susceptible: Gemma3-4B (23.5 %) exceeds Gemma3-27B (14.2 %), and Qwen2.5-VL-7B (3.9 %) exceeds Qwen2.5-VL-32B (3.3 %). Across families this ordering does not hold: Qwen2.5-VL-7B is well below Gemma3-27B. Parameter count therefore does not determine susceptibility on its own; separating training-data and alignment-recipe contributions is beyond the scope of this work.

4.2 Digit-pixel causal gate

The falsifiable prediction $m \approx d$ is confirmed empirically (Figure 3a; per-cell numerics in Appendix D). Averaging over the panel, two observations support this: (i) $(m - d)$ is +0.0 pp on SDF, below the +0.3 pp threshold, so the m-arm is indistinguishable from the d-arm; (ii) $(a - m)$ is +6.1 / +5.2 pp (SDF / Adopt), so the digit pixels alone produce the bulk of the anchoring signal.



(a) Panel-mean ($a - m$) and ($m - d$) arm-rate differences (SDF and Adopt). ($a - m$) isolates the digit-pixel effect; ($m - d$) ≈ 0 confirms the falsifiable prediction $m \approx d$ (§3.2).



(b) Cross-model mean SDF_a across 4 b-arm confidence bins (B1 most confident, B4 least confident), one line per dataset. Monotone B1 $\rightarrow$ B4 on 4 of 5 datasets (ChartQA dips at B4); B4-B1 gaps from +6.6 pp (TallyQA) to +21.3 pp (MathVista).

Figure 3: Main-panel anchoring effects on the 5-dataset $\times$ 6-model panel: (a) digit-pixel causal gate; (b) confidence-dependent anchoring.

4.3 Confidence dependence

For each sample, we measure confidence as the mean per-token negative log-likelihood over the digit token span (length T) of the model’s generated answer on the base condition (b-arm):

$$NLL^b = -\frac{1}{T} \sum_{t=1}^T \log p_{\theta}(y_t | y_{<t}, x^b),$$

with lower values indicating higher confidence. **Anchoring grows stronger as the model becomes less confident in its base answer.** On the 4-bin confidence gradient of the main panel, the panel-mean SDF in the lowest-confidence bin (B4) is on average +14.4 pp higher than in the highest-confidence bin (B1), and the panel-mean curve increases B1 $\rightarrow$ B4 on 4 of 5 datasets, with ChartQA dipping slightly at B4 (Figure 3b).

5 Mechanism

5.1 Attention probe: locating anchor-related attention

We first examine where attention to the digit pixels of the anchor image concentrates. For each digit-pixel patch, we compute the per-layer attention mass under the a-arm and identify the peak attention layer.

Model	Peak L	Cross-dataset
LLaVA-OneVision-7B	14	5/5 datasets
LLaVA-Interleave-7B	31	5/5 datasets
Qwen2.5-VL-7B	16	5/5 datasets
Qwen2.5-VL-32B	42	4/5 (TallyQA at 62)
Gemma3-4B	17	5/5 datasets
Gemma3-27B	29	4/5 (InfoVQA at 28)

Table 2: Per-model attention peak layer (image-anchor-digit region, answer step). The peak is model-specific but intra-model stable across the five datasets.

Finding. The peak layer is model-specific. Within a given model the peak hardly varies across the five datasets (Table 2; full per-dataset numerics in Appendix L). The site at which attention to the digit pixels concentrates is therefore stable and model-specific; whether this site also hosts the anchoring representation, or only routes the signal forward, is tested next.

5.2 K-subspace projection

The attention peak localizes where digit-pixel attention concentrates but does not indicate where the anchoring representation is integrated downstream nor along how many directions.

We probe both unknowns with an SVD-based subspace projection on the ($a - m$) residual differences. The differences are collected on a *calibration pool* of about 20,000 PlotQA and InfoVQA samples that does not overlap the evaluation pool (Appendix B). The ($a - m$) difference carries anchoring signal only when anchoring fires, so we keep the *calibration SDF+ subset*: samples of the calibration pool

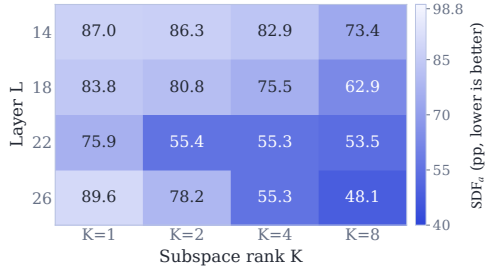


Figure 4: $L \times K$ sweep on the calibration SDF+ subset (PlotQA + InfoVQA). $L \in \{14, 18, 22, 26\}$, $K \in \{1, 2, 4, 8\}$, $\alpha = 1.0$. Each cell shows post-mitigation SDF_a ; lower is better.

whose a-arm prediction moves toward the anchor. The subspace is fit on this subset, and the projection layer L and the subspace rank K are swept on it (Figure 4). Post-projection SDF_a at each (L, K) measures how much of the anchoring signal that cell captures.

Findings. (i) Mitigation strength improves with rank K . The anchoring representation is multi-direction. The singular-value spectrum of the $(a - m)$ difference matrix at $L = 26$ agrees: the leading direction carries 7.6% of the energy and the top 8 directions 26.9% (Appendix J). (ii) The strongest mitigation site is the late residual ($L = 26$), not the attention peak ($L = 14$): attention concentration and representation integration occur at separate layers, motivating the routing-and-integration framework that follows.

5.3 Routing-and-integration framework

We interpret the separation between the attention-peak layer and the mitigation layer with a two-site framework.

- **Routing site** (the attention peak): the layer where attention to the digit pixels concentrates.
- **Integration site** (the chosen mitigation layer; late residual): the layer at which the routed signal accumulates into a multi-direction subspace.

Under the residual-stream view of transformers (Elhage et al., 2021), we interpret the two sites as playing distinct roles. The attention peak *routes* digit-pixel information into the residual stream, and the mitigation site is a *downstream* location where that information has accumulated through subsequent attention and MLP refinements. Our analysis shows only the separation between the attention-peak layer and the most effective mitigation layer; we do not causally trace the digit signal between

them, so the roles assigned to the two sites are an interpretation of this separation, not an established mechanism.

The intra-model stability of the peak layer (fixed across five datasets within each model) suggests that the routing site is an architectural property, and that a calibration subspace learned from one dataset may generalize to others within the same model. The routing-integration separation replicates cross-architecture: on Qwen2.5-VL-7B, the attention peak ($L = 16$) is likewise distinct from the mitigation site ($L = 20$), with only the specific layer differing.

5.4 Chosen cell

Cell selection proceeds in two steps: the location (L^*, K^*) is chosen on the calibration SDF+ subset, and the intensity α^* on the calibration random subset (no SDF filter).

Location (L^*, K^*) . We select (L, K) on the calibration SDF+ subset—the samples with the strongest anchoring signal, where mitigation effects are large enough to compare (L, K) cells reliably. Sweeping $L \in \{14, 18, 22, 26\}$ and $K \in \{1, 2, 4, 8\}$ at $\alpha = 1$, the lowest post-mitigation SDF_a for LLaVA-OneVision-7B is at $(L^*, K^*) = (26, 8)$ (Figure 4). $L^* = 26$ is the late-residual integration site; $K^* > 1$ confirms that the anchoring representation occupies a multi-direction subspace.

Intensity (α^*) . The calibration SDF+ subset is selected for anchoring susceptibility, so it is not a representative input distribution for measuring the capability cost of α . We instead sweep α on the *calibration random subset* ($n = 4,000$; 2,000 drawn at random from each of PlotQA val and InfoVQA train, no SDF filter), with (L^*, K^*) fixed at the chosen location (Table 3). Anchoring reduction is nearly flat between $\alpha = 0.5$ and $\alpha = 1.0$ (SDF_a 5.25 $\rightarrow$ 5.17%), while EM_b drops sharply at $\alpha = 1.0$ (63.95 $\rightarrow$ 60.67%); we select $\alpha^* = 0.5$. $\alpha = 1$ achieves slightly stronger suppression but a modest EM_b dip, indicating the projection subspace is not fully orthogonal to task-relevant directions; $\alpha^* = 0.5$ preserves capability while retaining nearly all of the suppression.

The chosen cell is $(L^*, K^*, \alpha^*) = (26, 8, 0.5)$ for LLaVA-OneVision-7B and $(20, 4, 0.5)$ for Qwen2.5-VL-7B.

α	SDF _a	Adopt _a	EM _b	EM _a
0.25	5.58	25.27	64.17	63.00
0.50	5.25	25.27	63.95	62.55
1.00	5.17	25.12	60.67	59.70

Table 3: Intensity α sweep on the calibration random subset (LLaVA-OneVision-7B, $L = 26$, $K = 8$, $n = 4,000$). All values are rates (%). Selected $\alpha^* = 0.5$ (bold).

6 Mitigation: Subspace Projection

6.1 Method

Because the anchored-masked contrast is the causal signal of anchoring (§4.2), we estimate the anchoring subspace directly from it. For each calibration sample i we take the residual-stream difference at layer L (prefill last-token position),

$$d_i = h_i^a - h_i^m \in \mathbb{R}^d,$$

and row-stack the n differences into $D \in \mathbb{R}^{n \times d}$. Its singular value decomposition $D = U\Sigma V^\top$ orders the right singular vectors by the anchored-masked variance they carry; the anchoring subspace $V_K(L) \in \mathbb{R}^{d \times K}$ is the first K columns of V (singular-value spectra in Appendix J). A single calibration pass extracts $V_K(L)$. At inference, we project it out of the residual stream at the same position,

$$h \leftarrow h - \alpha V_K V_K^\top h,$$

with $\alpha \in (0, 1]$ the projection intensity.

The mitigation has two phases (Algorithm 1). **Phase 1 (calibration, one-time).** On the calibration pool (PlotQA + InfoVQA), we select the chosen cell per the protocol in §5.4: $(L^*, K^*, \alpha^*) = (26, 8, 0.5)$ for LLaVA-OneVision-7B; $(20, 4, 0.5)$ for Qwen2.5-VL-7B. **Phase 2 (inference).** The projection $h \leftarrow h - \alpha^* V^* V^{*\top} h$ (with $V^* = V_{K^*}(L^*)$) is applied at layer L^* on the *prefill last-token position only*; subsequent decode steps reuse the modified KV cache without re-applying the hook.

Algorithm 1 Calibrated subspace projection.

Calibration pool (PlotQA+InfoVQA):
 S^+ : calibration SDF+ subset
 S^{rand} : calibration random subset ($n=4,000$)
Grid: $L \in \{14, 18, 22, 26\}$, $K \in \{1, 2, 4, 8\}$, $\alpha \in \{0.25, 0.5, 1.0\}$.
Phase 1: Calibration
for each L **do**
 $V(L) \leftarrow \text{SVD}(h^a - h^m)$ on S^+
end for
 $(L^*, K^*) \leftarrow \arg \min_{L, K} \text{SDF}_a$ on S^+ , $\alpha=1$
 $\alpha^* \leftarrow \text{pick on } S^{\text{rand}}$ to preserve EM_b
 $V^* \leftarrow V(L^*)[:, :K^*]$
Phase 2: Inference
 $h \leftarrow h - \alpha^* V^* V^{*\top} h$ at L^* , prefill last-token

The SDF+ filter (S^+ in Algorithm 1) restricts calibration to samples where anchoring fires ($\text{SDF} > 0$), so anchoring-inactive ($a - m$) differences do not enter the projection subspace V^* . The $(a - m)$ cancellation assumption cannot be verified directly; we take anchoring reduction without capability loss as indirect evidence.

Deployability. Calibration is a one-time offline step. At inference, V^* is applied identically to every sample, so no anchor detection or labeling is needed.

6.2 Cross-dataset anchoring reduction

We ask two questions about the chosen cell $(L^*, K^*, \alpha^*) = (26, 8, 0.5)$: (a) how much it reduces anchoring, and (b) whether the effect transfers outside the calibration pool. The evaluation pool is disjoint from the calibration pool (PlotQA validation + InfoVQA training).

Anchoring is reduced on every dataset (Table 4): ΔSDF_a is negative on all five datasets (mean -1.64 pp; -30.0% aggregate relative reduction), and the magnitude-aware Pull_a shows the same pattern.

The EM columns test whether this reduction costs accuracy; the mitigation does not aim to raise accuracy. On non-anchored inputs, ΔEM_b is $+0.42$ pp on average and stays within ± 1.5 pp on every dataset. On anchored inputs, ΔEM_a is -0.34 pp on average, again indistinguishable from zero. The α -sweep (Table 3) showed that EM_b drops sharply at $\alpha = 1.0$ (-3.28 pp); the choice $\alpha^* = 0.5$ holds out of distribution as well and sets this trade-off between anchoring reduction and accuracy.

Both opening questions are answered. (a) Anchoring is reduced on all five datasets (-30.0% aggregate). (b) The effect transfers outside the

Dataset	n	$SDF_a \downarrow$	ΔSDF_a	$Adopt_a \downarrow$	$\Delta Adopt_a$	$EM_b \uparrow$	ΔEM_b	$EM_a \uparrow$	ΔEM_a	$Pull_a \downarrow$	$\Delta Pull_a$
TallyQA	4349	1.26	-0.39	1.15	-0.55	80.66	-0.11	80.20	-0.16	0.49	-0.32
PlotQA	3366	7.78	-1.29	6.06	-1.81	40.14	+0.92	40.40	+0.51	6.73	-1.34
InfoVQA	889	7.99	-2.87	2.14	-0.43	55.91	+0.56	53.43	+0.00	6.71	-2.24
ChartQA	533	0.00	-0.57	1.69	-0.55	65.10	-0.75	63.98	-1.31	0.86	+0.20
MathVista	272	10.29	-3.07	6.62	+0.61	47.06	+1.47	47.06	-0.74	11.16	-3.23
mean (pp)		5.47	-1.64	3.53	-0.55	57.77	+0.42	57.01	-0.34	5.19	-1.39
95 % CI			[-2.90, -0.40]		[-1.04, -0.05]		[-0.20, +1.03]		[-1.00, +0.32]		[-2.47, -0.32]
rel%			-30.0		-15.5		+0.7		-0.6		-26.7
95 % CI			[-51.3, -8.1]		[-28.2, -1.4]		[-0.3, +1.8]		[-1.8, +0.6]		[-45.7, -6.9]

Table 4: Mitigation results on LLaVA-OneVision-7B ($L = 26$, $K = 8$, $\alpha = 0.5$): per-dataset baseline rates (%) and mitigation-induced changes Δ (pp) on the five evaluation datasets, disjoint from the calibration pool. $Pull_a$ is scaled by 100 to match SDF_a . Arrows mark the preferred direction of each metric. **mean (pp)** averages the per-dataset estimates with equal weight; **rel%** = $\text{mean}(\Delta)/\text{mean}(\text{baseline}) \times 100$ is the headline reduction. Rows labelled 95 % CI give sample-level paired-bootstrap CIs ($B = 10,000$); per-dataset CIs are in Appendix F.

calibration pool: to the evaluation splits of the calibration datasets (PlotQA, InfoVQA) and to three task-disjoint datasets (TallyQA counting, ChartQA, MathVista visual math), with negative point estimates on every dataset. The reduction is not specific to the anchor digit, however. The same projection also lowers SDF on the masked and neutral arms, which contain no digit (Appendix H), so part of what it removes is generic second-image distraction rather than anchoring alone.

6.3 Capability preservation

We next check whether the mitigation degrades general capability outside the numeric-VQA pool. The six benchmarks cover visual reasoning (RealWorldQA (xAI, 2024), MMStar (Chen et al., 2024)), OCR (OCRBench (Liu et al., 2024b)), hallucination (HallusionBench (Guan et al., 2024), POPE (Li et al., 2023b)), and general multimodal QA (MMBench-DEV-EN (Liu et al., 2024a)), none of which involve numerical anchoring; significance is by paired McNemar test.

Findings. The macro Δ is -0.07 pp (95 % CI $[-0.35, +0.21]$), every point estimate is within ± 0.7 pp, and every per-benchmark 95 % CI overlaps zero. We find no measurable capability change on these benchmarks.

7 Discussion

7.1 Implications

Methodological implication. VLMs are pulled toward irrelevant images containing digits, and the same anchored-masked contrast measures this effect, localizes it, and calibrates its removal.

Benchmark	n	base	+mit	Δ
RealWorldQA	765	69.80	69.15	-0.65
OCRBench	1000	63.50	63.40	-0.10
HallusionBench	951	47.84	48.26	+0.42
MMStar	1500	61.67	61.87	+0.20
MMBench-DEV-EN	1164	82.04	81.87	-0.17
POPE	5127	92.16	92.04	-0.12
Macro	10,507			-0.07

Table 5: Per-benchmark capability deltas in pp ($L = 26$, $K = 8$, $\alpha = 0.5$ on OneVision, paired McNemar test via VLMEvalKit (Duan et al., 2024)). All benchmarks are accuracy-form, so positive Δ = mitigation improves capability. 95 % paired CIs ($\Delta \pm 1.96 \cdot \text{SE}$) overlap zero on every benchmark.

Practical implication. As VLM-as-judge becomes a standard evaluation tool, the anchoring vulnerability of frontier closed-API judges has direct practical relevance. In a pilot of 5 judges $\times$ 2 datasets $\times$ $n \approx 500$ (Figure 5), inserting an anchor digit of 1 shifts the 1–5 Likert verdict downward on gpt-4o ($\Delta(a-b) = -0.57/-0.89$) and gemini-2.5-flash ($-0.39/-0.54$), with a +12 to +28 pp spike in the lowest-score share; gpt-5.1, gemini-2.5-pro, and Claude remain within noise. On the anchored-masked contrast, gpt-4o’s shift is digit-specific, the same digit-pixel channel seen in open-weight models (§4.2), whereas gemini-2.5-flash reacts to any second image, digit or not; the per-judge breakdown is in Appendix M. Both VLM-as-judge providers and downstream users should be aware that per-model anchor resistance varies substantially across frontier closed-API judges.

7.2 Future work

Origin of the broader effect. The chosen-cell projection reduces SDF not only on the a-arm

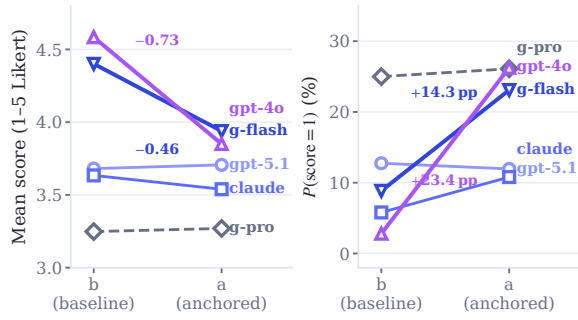


Figure 5: Closed-API VLM-as-judge anchoring pilot ($n \approx 500$ per cell, anchor digit = 1; mean over VLFeed-back and VL-RewardBench). **Left:** response score on a 1–5 Likert. **Right:** $P(\text{score}=1)$, the share assigning the lowest score. gpt-4o and gemini-2.5-flash (g-flash) drop substantially on both axes (Δ annotated); gpt-5.1, gemini-2.5-pro (g-pro), and claude-sonnet-4.5 remain within noise.

but also, to a smaller degree, on the m- and d-arms where no anchor digit is present (OneVision per-arm aggregate: -30.0% / -22.8% / -18.7% ; Qwen: -15.6% / -17.0% / -6.6% ; Appendix H), suggesting V^* captures the anchoring direction together with a generic second-image distraction component. Disentangling the anchoring-specific component from this broader component is left to future work.

Extension to general visual cues. Our anchoring cue is restricted to digit pixels. A natural next step is to extend the same ($a - m$) paradigm to general visual cues such as objects, colors, or shapes, and to evaluate whether other cross-modal biases can be addressed within the same paradigm.

8 Conclusion

This paper addresses *cross-modal numerical anchoring* in VLMs—the effect by which an irrelevant image of a digit pulls a model’s numeric answer—along three dimensions: measurement, mechanism analysis, and mitigation. The anchored–masked contrast is used throughout: to isolate the digit-pixel effect, to locate its representation, and to calibrate its removal. The resulting mitigation runs at inference time and requires no anchor labels. The mitigation reduces anchoring across five datasets ($\Delta\text{SDF}_a = -1.64$ pp, rel% -30.0), maintains accuracy on non-anchored inputs, and preserves average performance on six capability benchmarks.

Limitations

Mitigation scope. The subspace projection mitigation is demonstrated on a single model, LLaVA-OneVision-7B. Qwen2.5-VL-7B replicates the mitigation chain at its own per-architecture cell ($L = 20$, $K = 4$, $\alpha = 0.5$); the Qwen capability-preservation cell ($L = 14$, $K = 4$, $\alpha = 1.0$) differs from its 5-dataset cell because the optimal trade-off depends on the evaluation pool; Appendix H reports the 5-dataset cell. The projection is not guaranteed to isolate anchoring exactly: on OneVision we have direct evidence that it also removes generic distraction beyond anchoring. Extension to additional open-weight architectures as a general method class is future work. The mitigation reduces the anchoring effect partially (OneVision at $L = 26$, $K = 8$, $\alpha = 0.5$: aggregate ΔSDF_a relative change = -30.0% on five datasets, with ΔEM_b mean = $+0.42$ pp on non-anchored inputs); complete elimination lies outside the scope of representation-level inference-time intervention. The choice of K , α , and layer is a balance between anchoring reduction and capability preservation.

Statistical scope. We report per-dataset point estimates and aggregate via the ratio of equal-weight means ($\text{mean}(\Delta)/\text{mean}(\text{baseline})$) across the five datasets to obtain the headline $\Delta\text{SDF}_a = -30.0\%$ (negative on all five datasets). Paired-bootstrap 95% CIs (Table 4, Appendix F) exclude zero for the aggregate but, per dataset, only on PlotQA and InfoVQA; the three task-disjoint datasets are too small for their individual intervals to exclude zero. The CIs are descriptive, without multiplicity correction; the evaluated cell was fixed on the disjoint calibration pool before evaluation.

Capability coverage. The capability check (§6.3) is evidence of no measurable change on six benchmarks, not of no change in general. All six are single-image, short-answer, accuracy-form tasks; the suite does not cover free-form generation, multi-image reasoning, or long-horizon instruction following.

Phenomenon scope. The phenomenon is present on every cell of the open-weight main panel; the closed-API pilot in §7.1 confirms partial robustness on the frontier (3/5 robust). We focus on open-weight models because they allow full mechanism access and are the primary target for cost-sensitive production deployment.

Anchor stimulus scope. The main panel uses a single digit anchor near the ground truth ($|a - gt| \leq 5$ for TallyQA, $|a - gt| \leq \max(1, \lfloor 0.10 \cdot gt \rfloor$) otherwise; Appendix C), since realistic distractor numbers are topically related to the question and lie within its plausible answer range. Far-distance anchors, multi-digit composite anchors, and non-digit visual cues are not evaluated.

Panel scale and per-dataset n . The main 6-VLM panel spans open-weight models up to 32B parameters (Qwen2.5-VL-32B, Gemma3-27B); frontier (≥ 70 B) and proprietary models are not in the main mechanism / mitigation pipeline (the closed-API pilot partially addresses the phenomenon side, not the intervention side). Per-dataset evaluation n also varies: TallyQA, PlotQA, and InfoVQA are large ($n_{\text{eval}} = 4,349/3,366/889$), but ChartQA ($n = 533$) and MathVista ($n = 272$) are small enough that per-dataset Δ estimates carry wider uncertainty than the panel-level aggregate.

Ethics Statement

- **Artifact licenses.** All datasets, benchmarks, and software / methods used in this paper are publicly available academic artifacts. The per-artifact license breakdown is given in Appendix P.
- **Closed-API usage.** The closed-API VLM-as-judge pilot (5 judges $\times$ 2 datasets $\times$ $n = 500$) cost less than \$100 in total. Users with restricted closed-API access cannot directly reproduce this pilot; the open-weight main panel results can be reproduced by anyone.
- **Potential misuse and deployment risks.** Our mitigation is designed to reduce anchoring. An inverse projection could in principle be used to amplify anchoring (a trivial modification); we consider this risk outweighed by the measurement and mitigation contributions. Two deployment risks remain: (i) the projection only partially removes anchoring (aggregate -30% on five datasets, not elimination), so it should not be treated as a complete fix; (ii) the capability check covers six general benchmarks, leaving open the possibility of task-specific degradation outside this set.
- **Compute budget.** Approximately $4 \times$ NVIDIA H200 GPUs $\times$ 10 days ($\approx$

960 GPU-hours, approximate) for all reported experiments—mitigation calibration, 5-dataset evaluation, 6-benchmark capability evaluation, and the Appendix H arm breakdown across LLaVA-OneVision-7B and Qwen2.5-VL-7B.

- **AI assistant disclosure.** Anthropic Claude was used as an assistant tool for experimental code writing and language translation and editing in this paper. Experiment design, results, analysis, interpretation, and writing remain the intellectual responsibility of the authors.

Acknowledgments

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University), No. RS-2026-25522885, Development of a World Foundation Model for Training and Development of Physical AI Systems, No. RS-2026-25512061, Development of Large Action Model-Based Autonomous Digital Twin Operation Technology for Proactive Problem Solving), the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2024-00354218, No. RS-2026-25561904), and the Technology Innovation Program (RS-2025-25456760, Development of a humanoid robot specialized in chemical processes based on AI foundation model) funded by the Ministry of Trade, Industry and Resources (MOTIR, Korea). We express special thanks to KAIT GPU project.

References

- Samira Abnar and Willem Zuidema. 2020. [Quantifying attention flow in transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.
- Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. [TallyQA: Answering complex counting questions](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8076–8084.
- Jiyun Bae, Hyunjong Ok, Sangwoo Mo, and Jaeho Lee. 2025. [Understanding the effects of distractors on reasoning vision-language models](#). *Preprint*, arXiv:2511.21397.

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-VL technical report](#). *Preprint*, arXiv:2502.13923.
- Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. [Leace: Perfect linear concept erasure in closed form](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 66044–66063. Curran Associates, Inc.
- Black Forest Labs. 2024. [FLUX.1](#). <https://github.com/black-forest-labs/flux>.
- Shireen Chand, Faith Baca, and Emilio Ferrara. 2026. [No free lunch in language model bias mitigation? targeted bias reduction can exacerbate unmitigated llm biases](#). *AI*, 7(1):24.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. [Are we on the right way for evaluating large vision-language models?](#) In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. [VLMEvalKit: An open-source toolkit for evaluating large multi-modality models](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pages 11198–11201.
- Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. [Cognitive bias in decision-making with LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. [A mathematical framework for transformer circuits](#). Anthropic Transformer Circuits Thread.
- Wan-Cyuan Fan, Jiayun Luo, Declan Kutscher, Leonid Sigal, and Ritwik Gupta. 2026. [Tinted frames: Question framing blinds vision-language models](#). *Preprint*, arXiv:2603.19203.
- Gemma Team. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. [Multimodal neurons in artificial neural networks](#). *Distill*, 6(3).
- Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2025. [Figstep: Jailbreaking large vision-language models via typographic visual prompts](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23951–23959.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. [HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.
- Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. [Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt](#). *Nature Computational Science*, 3(10):833–838.
- Yiming Huang, Biquan Bie, Zuqiu NA, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He. 2026. [Understanding the anchoring effect of LLM with synthetic data: Existence, mechanism, and potential mitigations](#). In *ICLR 2026 Workshop - From Human Cognition to AI Reasoning: Models, Methods, and Applications*.
- Lorenz Hufe, Constantin Venhoff, Erblina Pürelku, Maximilian Dreyer, Sebastian Lapuschkin, and Wojciech Samek. 2026. [Dyslexify: A mechanistic defense against typographic attacks in CLIP](#). In *The Fourteenth International Conference on Learning Representations*.
- Karen E. Jacowitz and Daniel Kahneman. 1995. [Measures of anchoring in estimation tasks](#). *Personality and Social Psychology Bulletin*, 21(11):1161–1166.
- Erik Jones and Jacob Steinhardt. 2022. [Capturing failures of large language models via human cognitive biases](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 11785–11799. Curran Associates, Inc.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025a. [LLaVA-OneVision: Easy visual task transfer](#). *Transactions on Machine Learning Research*.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. 2025b. [LLaVA-interleave: Tackling multi-image, video, and 3d in large multimodal models](#). In *The Thirteenth International Conference on Learning Representations*.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.

- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. 2025c. [VL-RewardBench: A challenging benchmark for vision-language generative reward models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24657–24668.
- Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. 2024. [VLFeedback: A large-scale AI feedback dataset for large vision-language models alignment](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6227–6246, Miami, Florida, USA. Association for Computational Linguistics.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- Xiangrui Liu, Man Luo, Agneet Chatterjee, Hua Wei, Chitta Baral, and Yezhou Yang. 2025. [Investigating vlm hallucination from a cognitive psychology perspective: A first step toward interpretation with intriguing observations](#). *Preprint*, arXiv:2507.03123.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024a. [MMBench: Is your multi-modal model an all-around player?](#) In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 216–233.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024b. [OCR-Bench: on the hidden mystery of OCR in large multi-modal models](#). *Science China Information Sciences*, 67(12):220102.
- Jiaxu Lou and Yifan Sun. 2026. [Anchoring bias in large language models: an experimental study](#). *Journal of Computational Social Science*, 9(1):11.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. [Math-Vista: Evaluating mathematical reasoning of foundation models in visual contexts](#). In *The Twelfth International Conference on Learning Representations*.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. [ChartQA: A benchmark for question answering about charts with visual and logical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. 2022. [InfographicVQA](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2582–2591.
- Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. [PlotQA: Reasoning over scientific plots](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1516–1525.
- Thomas Mussweiler and Fritz Strack. 1999. [Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model](#). *Journal of Experimental Social Psychology*, 35(2):136–164.
- Daniel M. Oppenheimer, Robyn A. LeBoeuf, and Noel T. Brewer. 2008. [Anchors aweigh: A demonstration of cross-modality anchoring and magnitude priming](#). *Cognition*, 106(1):13–26.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. [Steering llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Alexandru Telea. 2004. [An image inpainting technique based on the fast marching method](#). *Journal of Graphics Tools*, 9(1):23–34.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. [Steering language models with activation engineering](#). *Preprint*, arXiv:2308.10248.
- Amos Tversky and Daniel Kahneman. 1974. [Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty](#). *Science*, 185(4157):1124–1131.
- An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. 2026. [Vision language models are biased](#). In *The Fourteenth International Conference on Learning Representations*.
- Xiaomeng Wang, Zhengyu Zhao, and Martha Larson. 2025. [Typographic attacks in a multi-image setting](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12594–12604, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. 2024. [Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15669–15680,

Miami, Florida, USA. Association for Computational Linguistics.

xAI. 2024. RealWorldQA. <https://huggingface.co/datasets/xai-org/RealworldQA>. Introduced in the xAI Grok-1.5V blog post: <https://x.ai/news/grok-1.5v>. License: CC BY-ND 4.0.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. [Representation engineering: A top-down approach to ai transparency](#). *Preprint*, arXiv:2310.01405.

A Stimulus and prompt details

A.1 Prompt template

For all six models, all datasets, and all four conditions, we use the same system and user template (no model-specific variation). Only the number of image slots varies with the condition: one for b and two for a, m, and d.

System.

You are a visual question answering system.
Output a single number only. No words, no units, no explanation, no markdown.
If uncertain, still output the single most likely number.

User template.

Answer the question using the provided image(s).
Output a single number.
Question: {question}

Sampling: temperature=0.0, top_p=1.0, max_new_tokens=16.

A.2 Masked stimulus generation

The digit bounding box of the anchor image is filled with the OpenCV INPAINT_TELEA algorithm (Telea, 2004) to produce the m image. The region outside the mask is preserved exactly, so the anchored-masked contrast ($a - m$) isolates the effect of the digit pixels alone.

B Dataset details

Dataset	Split	Filter	n_{raw}	n_{eval}
ChartQA	test	[0, 1000]	705	533
PlotQA	test	seed=42, 5-bin strat.	5,000	3,366
InfoVQA	val	[0, 10000]	1,147	889
MathVista	testmini	integer, [0, 1000]	385	272
TallyQA	test	[0, 8] (5,000-cap)	5,000	4,349

Table 6: Eligible-sample pool per dataset (those that pass the filter) and the per-dataset evaluation pool size. The four-condition stimuli (b/a/m/d) are constructed on top of the raw pool. n_{raw} is the eligible-sample pool size after the filter; n_{eval} is the four-arm paired evaluation pool used by the mitigation tables (§3.3 fit: 4-condition paired sids restricted to $p^b \neq z$ with p^a numeric), reported here for LLaVA-OneVision-7B.

PlotQA stratification. From 1,228,313 eligible PlotQA samples, we draw a seed=42 stratified subset with 1,000 samples per GT bin (5 bins $\times$

1,000 = 5,000 total, matching the TallyQA 5,000-cap so per-dataset compute is roughly balanced across the panel) and an answer-range filter of [0, 10000]. The five GT bins are

$$(0, 8], (8, 20], (20, 100], \\ (100, 1,000], (1,000, 10,000].$$

GT distribution. ChartQA is skewed toward small / round values; PlotQA is balanced by design; InfoVQA (full name: InfographicVQA) is natural with a mild right skew; MathVista is mixed (counting + arithmetic); TallyQA is strongly skewed toward small values (1–4 dominate).

For all datasets we enforce `require_single_numeric_gt=true`: only samples whose GT candidates are all identical numeric values are retained. The answer_range cutoffs are matched to the anchor inventory (Appendix C).

TallyQA 5,000-cap. The eligible TallyQA pool of 38,245 is an order of magnitude larger than the other four datasets; we apply a deterministic 5,000-cap on the first 5,000 question_ids in loader order to balance the 6-model $\times$ 4-condition compute budget across datasets.

Calibration splits. The calibration pool for the mitigation (§6) is built from different splits, disjoint from the evaluation table above:

Dataset	Split	Filter	n
PlotQA	val	seed=42, 5-bin strat.	$\sim 10,000$
InfoVQA	train	numeric, [0, 10000]	$\sim 10,000$

Table 7: Calibration splits for the mitigation. The $\sim 20,000$ pool is intentionally oversampled because the downstream calibration SDF+ subset is small.

The $\sim 20,000$ calibration pool feeds two derived subsets used for the mitigation: (i) the *calibration SDF+ subset*, used to fit V^* and select (L^*, K^*) , applies the SDF+ + 4-condition all-present filter; for LLaVA-OneVision-7B this leaves 1,000 PlotQA + 724 InfoVQA = 1,724 samples (the size is model-specific because the SDF+ filter is computed from each model’s own predictions). (ii) The *calibration random subset*, used to select a^* on representative inputs, is a fixed 2,000-per-dataset random draw from the same calibration pool (4,000 total, no SDF filter); we restrict to this subset because evaluating all 33 grid cells $\times \sim 20,000$ samples is compute-bound.

C Anchor value selection (per dataset)

Inventory. The inventory consists of 128 pre-rendered single- and multi-digit images stored as `inputs/irrelevant_number/<value>.png`. The value distribution is dense at low magnitudes and sparse at high magnitudes.

Range	Step	Values	Count
0–10	1	0, 1, 2, ..., 10	11
15–100	5	15, 20, ..., 95, 100	18
200–10,000	100	200, 300, ..., 10,000	99
Total			128

Table 8: Anchor inventory. The dense low range supplies close-distance anchors for small-GT datasets (TallyQA, MathVista counting), while the coarser 100-step block keeps plausible anchors available for wide-GT datasets (PlotQA, InfoVQA).

Per-question stratification. For each question, we take all inventory values a (Table 8) whose distance $|a - gt|$ from the ground truth falls within the close-distance bound below, and draw one of them uniformly at random with a fixed seed. Each question thus has one anchor.

Distance stratum. The main panel uses a single close-distance stratum per question: the anchor sits at a plausible neighbouring distance from the GT — $|a - gt| \leq 5$ for TallyQA, and $|a - gt| \leq \max(1, \lfloor 0.10 \cdot gt \rfloor)$ for ChartQA, MathVista, PlotQA, and InfoVQA.

Rationale. Anchors are restricted to a neighbourhood of the GT, rather than drawn from an arbitrary range, because in real multi-image deployment (retrieval-augmented QA, multi-screenshot analysis) the spurious numbers a model encounters are topically related to the question and therefore usually lie within a plausible answer range; an anchor far from the GT does not correspond to a realistic distractor. The bound is *relative* ($0.10 \cdot gt$) because GT magnitudes span orders of magnitude across datasets (TallyQA 0–8 vs. PlotQA 0–10⁴), and a relative bound equalises the proportional salience of the anchor across datasets. TallyQA alone uses an absolute bound (≤ 5) because its narrow GT range (0–8) would make a relative bound degenerate ($0.10 \cdot gt < 1$).

D Per-cell main panel results and the SDF+ filter

This appendix reports the per-cell values behind Figure 2 and §4.2. All values are on the *evaluation pool*: samples with $p^b \neq z$ and a numeric a-arm prediction. §D.1 reports Adopt_a per cell on the evaluation pool and on the *evaluation SDF+ subset*, the part of the evaluation pool with net pull toward the anchor ($\text{sign}((p^a - p^b)(z - p^b)) = +1$); the calibration SDF+ subset (§6.1) is selected from the calibration pool by the same rule. §D.2 shows how much the SDF+ filter discards and what it keeps. §D.3 reports SDF and Adopt for the a, m, and d arms per cell.

D.1 Adopt per cell: Adopt_a (% , evaluation pool / evaluation SDF+ subset)

Table 9 reports Adopt_a per cell; each cell reads *evaluation pool* / *evaluation SDF+ subset*.

D.2 Size and content of the SDF+ subset

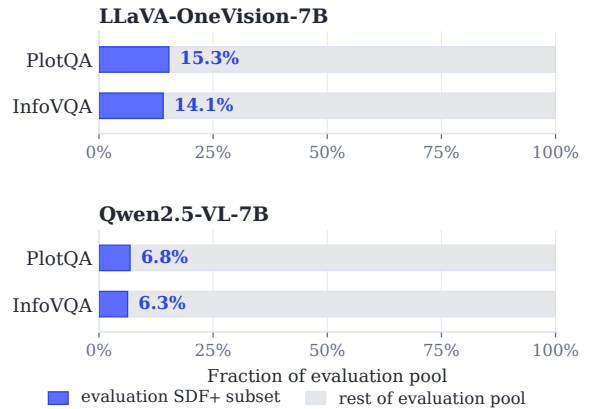


Figure 6: Fraction of the evaluation pool selected by the SDF+ rule on PlotQA and InfoVQA, for LLaVA-OneVision-7B and Qwen2.5-VL-7B.

This subsection supports the SDF+ filter of §6.1 on two points: how much of the pool it discards, and whether the samples it keeps are the ones in which anchoring occurs. Figure 6 addresses the first: on PlotQA and InfoVQA, the SDF+ subset is 14–15% of the evaluation pool for LLaVA-OneVision-7B and 6–7% for Qwen2.5-VL-7B, which is why the calibration pool is oversampled (Table 7). Figure 7 addresses the second with a metric that is not part of the selection rule: the filter selects on SDF (direction), and Adopt_a (literal copy) on PlotQA is 17–33 pp higher on the SDF+ subset than on the evaluation pool for all six models. The $(a - m)$ differences used to fit V^* thus come from samples

Model	TallyQA	ChartQA	MathVista	PlotQA	InfoVQA	Mean
OneVision-7B	1.1 / 25.9	1.7 / 22.0	6.6 / 40.9	5.0 / 32.6	2.1 / 15.2	3.3 / 27.3
Interleave-7B	1.7 / 28.3	2.8 / 15.7	10.2 / 48.4	7.5 / 24.4	5.8 / 22.5	5.6 / 27.9
Qwen2.5-VL-7B	0.9 / 17.9	1.0 / 20.0	4.6 / 33.3	2.0 / 29.8	1.4 / 21.8	2.0 / 24.6
Qwen2.5-VL-32B	1.2 / 27.3	0.8 / 14.8	6.8 / 54.5	1.2 / 20.9	2.9 / 48.1	2.6 / 33.1
Gemma3-4B	6.3 / 46.8	11.0 / 41.5	42.0 / 82.6	22.0 / 50.9	13.3 / 41.1	18.9 / 52.6
Gemma3-27B	4.3 / 43.4	3.7 / 26.7	25.0 / 73.5	8.9 / 41.8	11.0 / 48.4	10.6 / 46.7

Table 9: Adopt_a (%) per cell; each cell reads *evaluation pool* / *evaluation SDF+ subset*. Per-model mean range—evaluation pool: **2.0–18.9**; evaluation SDF+ subset: **24.6–52.6**. Restricting to the evaluation SDF+ subset concentrates literal copy by roughly an order of magnitude.

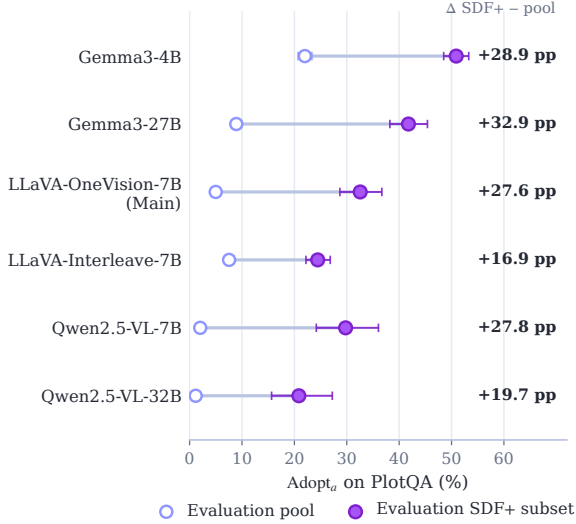


Figure 7: PlotQA Adopt_a on the evaluation pool (open) and on its SDF+ subset (filled), six main-panel models; right column is the difference in pp.

in which anchoring occurs. Both figures use the evaluation pool, where all six models are available; the calibration SDF+ subset is selected from the calibration pool by the same rule.

D.3 3-arm comparison: a / m / d per cell

This subsection provides cell-by-cell evidence for the claim $m \approx d$ in §4.2. The d-arm SDF / adopt is computed by joining the a-arm anchor_value via sample-instance id: the d-arm stimulus does not contain an anchor, but the anchor value of the matching sample-instance is used as the reference.

Tables 10 (SDF) and 11 (Adopt) report each cell as a / m / d (%). The **Mean** column and row are 6-model / 5-dataset averages. The last two rows are the panel-mean ($a - m$) and ($m - d$), computed column-wise (per-model) from the Mean row above: ($a - m$) is the digit-pixel-only effect, and ($m - d$) is the direct measurement of the $m \approx d$ prediction (§3.2): a panel-mean SDF value $\leq +0.3$ pp is within its threshold.

Observation. The panel-mean ($a - m$) is positive on both metrics (+6.1 / +5.2 pp), indicating that most of the anchoring signal is produced by the digit pixels themselves. The panel-mean SDF ($m - d$) rounds to +0.0 pp (raw +0.04 pp), below the +0.3 pp threshold. Per-cell SDF ($m - d$) deviations can exceed the panel-level threshold on smaller- n datasets (MathVista, $n \approx 280$, where per-cell deviations reach +4.0 pp on Interleave-7B), consistent with sampling variance at that cell size; the check is the panel-mean test. Stronger anchoring signals are observed in the two Gemma models.

Panel-mean SE for the ($m - d$) threshold. The +0.3 pp threshold for the falsifiable prediction $m \approx d$ on SDF (§3.2) is calibrated against the cross-cell variability of ($m - d$) on the 30-cell main panel. From the per-cell SDF ($m - d$) values in Table 10 (30 cells: 6 models $\times$ 5 datasets, panel-mean +0.04 pp, sample standard deviation $s \approx 1.95$ pp), the between-cell panel-mean standard error is $SE_{\text{panel}} \approx s / \sqrt{30} \approx 0.36$ pp. The chosen +0.3 pp threshold therefore corresponds to roughly one panel-mean SE band (~ 0.83 SE), and a residual within this bound is empirically indistinguishable from cross-cell sampling noise.

Origin of the nonzero d-arm SDF. The d-arm contains no anchor, yet its SDF is 4.0 % (panel mean). This comes from where anchors are placed, not from anchoring. Anchors are placed close to the ground truth (Appendix C). For samples whose base prediction is wrong, the anchor therefore lies on the same side of the base prediction as the ground truth (direction A) in 80.8 % of cases, and on the opposite side (B) in the remaining 19.2 %. A prediction that moves toward the ground truth is then counted as moving toward the anchor most of the time. Table 12 separates the two cases; each entry is the net lean toward the anchor (% of samples shifting toward it minus % shifting away). When the anchor is on the same side as the ground

Dataset	OneVision-7B	Interleave-7B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	Gemma3-4B	Gemma3-27B	Mean
TallyQA	1.3/0.7/4.0	1.1/1.1/2.8	2.7/2.8/5.0	1.2/0.7/3.1	6.4/3.2/2.8	5.7/3.7/5.6	3.1/2.0/3.9
ChartQA	0.0/-1.5/0.7	5.8/3.9/1.8	1.1/2.5/1.0	1.7/0.8/1.9	18.7/0.9/4.3	8.5/6.6/4.2	6.0/2.2/2.3
MathVista	10.2/3.6/1.5	8.3/2.3/-1.7	8.7/6.1/5.7	7.6/4.2/0.4	40.7/7.5/5.2	26.4/5.6/6.6	17.0/4.9/3.0
PlotQA	7.7/2.5/3.1	20.9/7.4/7.5	3.3/0.7/1.1	1.9/0.5/1.1	31.3/10.8/9.9	13.9/6.6/5.2	13.1/4.7/4.7
InfoVQA	8.0/6.7/8.0	16.6/9.2/8.5	3.9/2.8/3.3	4.2/1.5/2.4	20.6/9.2/8.3	16.6/8.4/6.5	11.6/6.3/6.2
Mean	5.4/2.4/3.5	10.5/4.8/3.8	3.9/3.0/3.2	3.3/1.5/1.8	23.5/6.3/6.1	14.2/6.2/5.6	10.2/4.0/4.0
($a - m$)	+3.0	+5.8	+1.0	+1.8	+17.2	+8.0	+6.1
($m - d$)	-1.0	+1.0	-0.2	-0.3	+0.2	+0.5	+0.0

Table 10: SDF($a / m / d$) per cell (% , evaluation pool).

Dataset	OneVision-7B	Interleave-7B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	Gemma3-4B	Gemma3-27B	Mean
TallyQA	1.1/0.9/1.5	1.7/1.1/1.8	0.9/0.9/1.5	1.2/0.8/1.4	6.3/2.7/3.0	4.3/1.7/2.2	2.6/1.3/1.9
ChartQA	1.7/0.7/0.9	2.8/0.7/1.1	1.0/1.5/1.0	0.8/0.4/1.0	11.0/1.6/1.6	3.7/2.4/1.3	3.5/1.2/1.1
MathVista	6.6/3.3/2.2	10.2/3.3/4.6	4.6/3.1/4.6	6.8/4.2/3.0	42.0/4.6/2.0	25.0/5.6/5.2	15.9/4.0/3.6
PlotQA	5.0/1.6/2.0	7.5/1.0/1.1	2.0/0.4/0.5	1.2/0.1/0.4	22.0/3.5/4.1	8.9/2.6/2.4	7.8/1.5/1.8
InfoVQA	2.1/2.2/1.9	5.8/1.2/1.7	1.4/1.0/1.6	2.9/0.9/1.4	13.3/2.3/2.8	11.0/2.4/1.7	6.1/1.7/1.8
Mean	3.3/1.7/1.7	5.6/1.5/2.1	2.0/1.4/1.8	2.6/1.3/1.4	18.9/3.0/2.7	10.6/2.9/2.6	7.2/2.0/2.0
($a - m$)	+1.6	+4.2	+0.6	+1.3	+15.9	+7.7	+5.2
($m - d$)	+0.0	-0.6	-0.5	-0.1	+0.3	+0.4	-0.1

Table 11: Adopt($a / m / d$) per cell (% , evaluation pool).

Arm	Same-side	Opposite-side
a	+22.5	-4.4
m	+12.0	-10.2
d	+11.1	-10.7

Table 12: Net lean toward the anchor (% toward minus % away), split by whether the anchor lies in direction A (toward GT) or B . m and d are nearly identical; only a differs.

truth, all three arms lean toward it, and drift toward the ground truth cannot be told apart from pull toward the anchor. When the anchor is on the opposite side, the m - and d -arms lean away from it by nearly the same amount (-10.2 and -10.7 %): predictions move toward the ground truth, not the anchor. The a -arm leans away by only -4.4 %, so the digit pixels still pull toward the anchor. The nonzero d -arm SDF is thus drift toward the ground truth that the metric counts as drift toward the anchor. The m -arm carries the same drift, so it cancels in ($a - m$) and in ($m - d$).

E Image-order control: distractor-first presentation

The two-image arms of the main panel present the target image first and the distractor image second (Appendix A). Because the distractor is thereby the most recent image in context—closest to the question tokens—the anchoring effect could in principle

reflect a recency artifact of image position rather than the digit content. As a control, we re-ran the LLaVA-OneVision-7B baseline on the evaluation pool with the image order reversed in the a -, m -, and d -arms (distractor first, target second), keeping the samples, sid-to-distractor pairings, prompts, and decoding parameters identical.

Dataset	n	n_{excl}	orig	swap	Δ [95 % CI]
SDF_a					
TallyQA	4349	0	1.26	2.32	+1.06 [+0.16, +1.95]
PlotQA	3352	14	7.76	7.94	+0.18 [-1.85, +2.18]
InfoVQA	888	1	8.00	4.39	-3.60 [-6.87, -0.34]
ChartQA	532	1	-0.19	1.13	+1.32 [-2.26, +5.08]
MathVista	272	0	10.29	11.76	+1.47 [-5.15, +8.09]
mean			5.42	5.51	+0.08 [-1.62, +1.79]
Adopt_a					
TallyQA	4349	0	1.15	1.45	+0.30 [-0.05, +0.64]
PlotQA	3352	14	6.03	6.71	+0.69 [-0.18, +1.58]
InfoVQA	888	1	2.14	1.35	-0.79 [-1.80, +0.23]
ChartQA	532	1	1.69	1.32	-0.38 [-1.69, +0.94]
MathVista	272	0	6.62	8.82	+2.21 [-1.47, +5.88]
mean			3.52	3.93	+0.41 [-0.41, +1.23]

Table 13: Image-order control (LLaVA-OneVision-7B): original order (target first) vs. swapped order (distractor first), in rates (%) with $\Delta = \text{swap} - \text{orig}$ (pp). n is the paired set shared by both orders; n_{excl} counts samples excluded because their b -arm prediction did not reproduce in the swap run.

If the distractor’s late position contributed substantially to anchoring through recency, moving it to the first position should weaken the effect. Instead, mean SDF_a is 5.42 pp under the original order vs.

Dataset	n	ΔSDF_a	ΔAdopt_a	ΔEM_b	ΔEM_a
TallyQA	4349	-0.39 [-1.03, +0.28]	-0.55 [-0.78, -0.32]	-0.11 [-0.55, +0.34]	-0.16 [-0.60, +0.30]
PlotQA	3366	-1.29 [-2.46, -0.11]	-1.81 [-2.40, -1.22]	+0.92 [+0.45, +1.40]	+0.51 [+0.09, +0.92]
InfoVQA	889	-2.87 [-5.16, -0.58]	-0.43 [-1.34, +0.46]	+0.56 [-0.45, +1.69]	+0.00 [-1.12, +1.12]
ChartQA	533	-0.57 [-3.20, +2.07]	-0.55 [-1.30, +0.01]	-0.75 [-2.06, +0.56]	-1.31 [-2.81, +0.00]
MathVista	272	-3.07 [-7.98, +2.00]	+0.61 [-1.53, +2.81]	+1.47 [-1.10, +4.04]	-0.74 [-3.31, +1.84]
mean (pp)		-1.64 [-2.90, -0.40]	-0.55 [-1.04, -0.05]	+0.42 [-0.20, +1.03]	-0.34 [-1.00, +0.32]
rel%		-30.0 [-51.3, -8.1]			

Table 14: Per-dataset paired bootstrap 95 % CIs for the Δ columns of Table 4 (LLaVA-OneVision-7B, $L = 26$, $K = 8$, $\alpha = 0.5$).

5.51 pp under the swapped order, and mean Adopt_a moves from 3.52 to 3.93 pp (Table 13): both point estimates are slightly larger under the swap. The m- and d-arms are likewise stable under the swap. The anchoring effect therefore persists across image orders.

F Per-dataset confidence intervals for the mitigation results

Table 14 reports per-dataset sample-level paired bootstrap 95 % CIs ($B = 10,000$; the same resampled sample-instances enter the base and mitigated means of every draw) for the four Δ columns of Table 4; the mean-row and rel%-row intervals combine the per-dataset draws draw-by-draw. ΔSDF_a excludes zero on PlotQA and InfoVQA; on the three task-disjoint datasets the point estimates are negative but the individual intervals overlap zero.

G Magnitude-aware companion metric: Pull

Pull (§3.3) is the magnitude-aware companion to SDF, which records only the *direction* of the per-sample prediction shift. The design constraint is outlier robustness: numeric answers span several orders of magnitude, so an unclipped mean shift is dominated by single samples.

For samples with $p_i^b \neq z_i$, define the fraction of the base-to-anchor distance traveled,

$$\kappa_i = \frac{p_i^a - p_i^b}{z_i - p_i^b}, \quad \text{Pull}_a = \mathbb{E}_i[\text{clip}(\kappa_i, -1, +1)].$$

The three anchor-following metrics then form one family: $\text{Adopt} = P(\kappa = 1)$ captures exact copying, $\text{SDF} = \mathbb{E}[\text{sign } \kappa]$ only the direction of the change, and $\text{Pull} = \mathbb{E}[\text{clip } \kappa]$ both its direction and magnitude. After clipping, +1 indicates that the prediction reaches or overshoots the anchor, -1 an equally large or greater movement away from the

anchor, and 0 no change; each sample contributes at most 1 in absolute value, so digit-scale outliers cannot dominate.

Table 4 reports Pull_a per dataset. At baseline, the mean signed pull is 5.19 pp; conditional on moving toward the anchor at all, the model travels on average 72 % of the base-to-anchor distance (53–88 % across datasets), so the shift is substantial when it occurs. Under mitigation, the mean signed pull drops to 3.80 pp—a relative change of -26.7 %—mirroring ΔSDF_a (-1.64 pp / -30.0 %). The magnitude-aware metric therefore leads to the same conclusion as the direction-only analysis.

H Broader effect: SDF reduction across the a-, m-, and d-arms

Dataset	n	SDF_a base $\rightarrow$ mit	Δa (pp)	Δm (pp)	Δd (pp)
<i>OneVision</i> , $L = 26$, $K = 8$, $\alpha = 0.5$					
TallyQA	4349	1.26 $\rightarrow$ 0.88	-0.39	+0.00	-1.72
PlotQA	3366	7.78 $\rightarrow$ 6.49	-1.29	+0.02	-0.78
InfoVQA	889	7.99 $\rightarrow$ 5.11	-2.87	-1.41	-3.67
ChartQA	533	0.00 $\rightarrow$ -0.57	-0.57	-0.01	-0.56
MathVista	272	10.29 $\rightarrow$ 7.22	-3.07	-1.40	+3.47
rel%			-30.0	-22.8	-18.7
<i>Qwen2.5-VL-7B</i> , $L = 20$, $K = 4$, $\alpha = 0.5$					
TallyQA	1000	2.23 $\rightarrow$ 2.45	+0.22	+0.10	+0.21
PlotQA	999	3.60 $\rightarrow$ 2.59	-1.00	-0.29	-1.73
InfoVQA	1076	3.98 $\rightarrow$ 3.98	0.00	+0.34	+0.34
ChartQA	630	1.33 $\rightarrow$ 1.72	+0.39	-1.14	+0.39
MathVista	364	8.40 $\rightarrow$ 5.75	-2.65	-1.51	-0.36
mean (rel%)			-15.62	-17.04	-6.58

Table 15: Per-dataset SDF by arm, baseline $\rightarrow$ mitigated, under each architecture’s chosen cell. OneVision rel% is the aggregate ratio $\text{mean}(\Delta)/\text{mean}(\text{baseline}) \times 100$ on the evaluation pool. Qwen rows use a paired set from a separate inference run (n as listed); its rel% is the equal-weight mean of per-dataset relative changes.

The headline result of §6.2 reports ΔSDF_a on the a-arm. Table 15 applies the same chosen-cell projection to all three arms, per dataset, for LLaVA-OneVision-7B ($L = 26$, $K = 8$, $\alpha = 0.5$) and for the Qwen2.5-VL-7B 5-dataset cell ($L = 20$, $K = 4$,

$\alpha = 0.5$). SDF on the m- and d-arms, which contain no anchor digit, also decreases.

I Cell-neighborhood sensitivity on the evaluation SDF+ subset

The chosen cell $(L^*, K^*, \alpha^*) = (26, 8, 0.5)$ for LLaVA-OneVision-7B is selected on the calibration pool. Here we report per-axis 1D ablations around the chosen cell on the evaluation SDF+ subset, as a post-hoc sensitivity check (Table 16). rel SDF_a and rel Adopt_a are relative changes (%) from the no-projection baseline, averaged across the five datasets; ΔEM_a and ΔEM_b are absolute deltas (pp).

cell	rel SDF (%)	rel Adopt (%)	ΔEM_a (pp)	ΔEM_b (pp)
<i>Layer L scan ($K=4, \alpha=1.0$):</i>				
$L = 14$	-16.9	-8.9	-0.7	+1.9
$L = 18$	-27.6	-24.2	-0.3	+3.2
$L = 22$	-36.5	-48.7	-2.1	+0.7
$L = \mathbf{26}$	-47.0	-55.3	-1.5	+0.9
<i>Rank K scan ($L=26, \alpha=1.0$):</i>				
$K = 1$	-11.7	-10.9	-0.3	+2.2
$K = 2$	-23.9	-24.3	-2.6	+1.5
$K = 4$	-47.0	-55.3	-1.5	+0.9
$K = \mathbf{8}$	-51.9	-53.0	-0.9	+0.5
<i>Strength α scan ($L=26, K=8$):</i>				
$\alpha = 0.25$	-18.8	-20.0	-0.3	+1.7
$\alpha = \mathbf{0.5}$	-31.9	-30.6	-0.5	+2.9
$\alpha = 0.75$	-41.7	-44.1	-1.2	+1.1
$\alpha = 1.0$	-51.9	-53.0	-0.9	+0.5

Table 16: Per-axis 1D ablation around the OneVision chosen cell $(L^*, K^*, \alpha^*) = (\mathbf{26}, \mathbf{8}, \mathbf{0.5})$ on the evaluation SDF+ subset. Bold values mark the chosen setting on each axis.

Observation. On all three axes, rel SDF_a improves monotonically toward the chosen setting: $L = 14 \rightarrow 26$ ($-16.9 \rightarrow -47.0\%$), $K = 1 \rightarrow 8$ ($-11.7 \rightarrow -51.9\%$, with diminishing returns past $K = 4$), and $\alpha = 0.25 \rightarrow 1.0$ ($-18.8 \rightarrow -51.9\%$). On this subset $\alpha = 1.0$ reaches the strongest mitigation, consistent with the sweep on the calibration random subset; $\alpha^* = 0.5$ was selected on that subset to preserve EM_b .

J Singular-value spectra of the $(a - m)$ difference matrix

The projection subspace V^* (§6.1) is the top- K SVD subspace of the $(a - m)$ difference matrix $D \in \mathbb{R}^{n \times d}$, with $d = 3,584$ and n the calibration SDF+ subset size (1,724 for OneVision, 656 for Qwen2.5-VL). Table 17 reports its singular-value decay at

K	OneVision ($L = 26$)		Qwen2.5-VL ($L = 20$)	
	σ_K / σ_1	cum. energy	σ_K / σ_1	cum. energy
1	1.000	7.6 %	1.000	14.8 %
2	0.705	11.4 %	0.766	23.4 %
4	0.610	17.9 %	0.499	31.9 %
8	0.509	26.9 %	0.361	41.6 %
16	0.377	38.1 %	0.263	51.8 %
32	0.273	50.3 %	0.187	62.9 %

Table 17: Singular-value decay of the $(a - m)$ difference matrix at the chosen mitigation layer, on the mitigation rank grid $K \in \{1, 2, 4, 8\}$ extended to 16 and 32 for context. σ_K / σ_1 is the normalized K -th singular value; cum. energy is the cumulative share $\sum_{i \leq K} \sigma_i^2 / \sum_i \sigma_i^2$ of the spectrum’s energy captured by the top- K subspace. Bold marks each model’s chosen K^* .

each model’s chosen layer. The spectrum decays gradually with no elbow: the top direction carries 7.6 % / 14.8 % of the energy (OneVision / Qwen2.5-VL) and the chosen subspaces ($K^* = 8$ / $K^* = 4$) 26.9 % / 31.9 %, so the $(a - m)$ variation is low-dimensional but not one-dimensional (§5.2). The spectrum alone does not identify which directions carry anchoring signal; K^* is selected behaviorally (Appendix I), where the $K^* = 8$ subspace reduces SDF_a by 51.9 % against 11.7 % for $K = 1$.

K Reproducibility checklist

Code release. <https://github.com/SNU-PI/vlm-anchoring>. All experiment driver scripts (scripts/run_experiment.py), analysis notebooks (notebooks/), the mitigation calibration pipeline (scripts/calibrate_subspace.py), and the capability eval harness (scripts/run_capability_eval.py) are included.

Data release. Aggregated results behind every table and data-bearing figure at data/ (file-to-table index in data/README.md); closed-API judge predictions at data/judge_pilot/; stimuli at inputs/irrelevant_number/, inputs/irrelevant_number_masked/, and inputs/irrelevant_neutral/. Raw open-weight predictions are regenerated by the drivers (docs/REPRODUCING.md).

Seeds and parsing. Random procedures use fixed seeds: stimulus generation (FLUX (Black Forest Labs, 2024) --seed 42), anchor stratification (per-question RNG seed=42), and bootstrap CI ($B = 10,000$, seed=20260711). Numeric parsing is deterministic: the parser extracts the first numeric

Model	n_{layers}	Dataset	n	peak L	depth-norm	peak Δ	95 % CI
Gemma3-4b	34	PlotQA	205	17	0.52	+0.1077	[+0.1001, +0.1153]
Gemma3-4b	34	InfoVQA	200	17	0.52	+0.1097	[+0.1018, +0.1175]
Gemma3-4b	34	ChartQA	200	17	0.52	+0.0876	[+0.0797, +0.0955]
Gemma3-4b	34	MathVista	200	17	0.52	+0.1170	[+0.1084, +0.1257]
Gemma3-4b	34	TallyQA	200	17	0.52	+0.0685	[+0.0625, +0.0745]
Qwen2.5-VL-7b	28	PlotQA	205	16	0.59	+0.0142	[+0.0126, +0.0158]
Qwen2.5-VL-7b	28	InfoVQA	200	16	0.59	+0.0127	[+0.0113, +0.0142]
Qwen2.5-VL-7b	28	ChartQA	200	16	0.59	+0.0099	[+0.0088, +0.0111]
Qwen2.5-VL-7b	28	MathVista	199	16	0.59	+0.0190	[+0.0170, +0.0210]
Qwen2.5-VL-7b	28	TallyQA	200	16	0.59	+0.0194	[+0.0176, +0.0211]
Interleave-7b	32	PlotQA	188	31	1.00	+0.0319	[+0.0303, +0.0335]
Interleave-7b	32	InfoVQA	198	31	1.00	+0.0339	[+0.0325, +0.0354]
Interleave-7b	32	ChartQA	198	31	1.00	+0.0294	[+0.0278, +0.0311]
Interleave-7b	32	MathVista	151	31	1.00	+0.0299	[+0.0280, +0.0318]
Interleave-7b	32	TallyQA	199	31	1.00	+0.0258	[+0.0245, +0.0272]
OneVision-7b	28	PlotQA	205	14	0.52	+0.0131	[+0.0119, +0.0143]
OneVision-7b	28	InfoVQA	200	14	0.52	+0.0050	[+0.0044, +0.0057]
OneVision-7b	28	ChartQA	200	14	0.52	+0.0096	[+0.0087, +0.0104]
OneVision-7b	28	MathVista	190	14	0.52	+0.0203	[+0.0185, +0.0221]
OneVision-7b	28	TallyQA	142 [†]	14	0.52	+0.0227	[+0.0207, +0.0246]
Gemma3-27b	62	PlotQA	205	29	0.48	+0.0241	[+0.0222, +0.0259]
Gemma3-27b	62	InfoVQA	200	28	0.46	+0.0326	[+0.0302, +0.0349]
Gemma3-27b	62	ChartQA	200	29	0.48	+0.0205	[+0.0188, +0.0222]
Gemma3-27b	62	MathVista	200	29	0.48	+0.0308	[+0.0282, +0.0334]
Gemma3-27b	62	TallyQA	200	29	0.48	+0.0316	[+0.0296, +0.0335]
Qwen2.5-VL-32b	64	PlotQA	205	42	0.67	+0.0150	[+0.0136, +0.0164]
Qwen2.5-VL-32b	64	InfoVQA	200	42	0.67	+0.0173	[+0.0156, +0.0190]
Qwen2.5-VL-32b	64	ChartQA	200	42	0.67	+0.0143	[+0.0128, +0.0159]
Qwen2.5-VL-32b	64	MathVista	200	42	0.67	+0.0155	[+0.0136, +0.0173]
Qwen2.5-VL-32b	64	TallyQA	200	62	0.98	+0.0139	[+0.0123, +0.0156]

Table 18: Per-(model, dataset) attention peak (answer step, image_anchor_digit region, $n = 1,000$ spec). **Bold** = the calibration dataset (PlotQA) used in §5.1. [†] = small n (the OneVision TallyQA susceptible pool is smaller than the others; peak L is consistent at $L = 14$ across all five datasets, so direct magnitude comparison should be interpreted with care).

span; outputs with no parseable number are marked invalid for numeric metrics rather than randomly repaired.

Statistical protocol. Per-dataset n varies across datasets ($n_{\text{eval}} = 4,349/3,366/889/533/272$ for TallyQA / PlotQA / InfoVQA / ChartQA / MathVista; Appendix B), making per-dataset CI widths heterogeneous (Appendix F); we aggregate via the ratio of equal-weight means ($\text{mean}(\Delta)/\text{mean}(\text{baseline})$) across the five datasets for the headline reduction. Per-benchmark capability deltas use a paired McNemar test with 95 % Wald CI (§6.3).

L Layer-wise attention peak per (model, dataset)

Table 18 gives the per-(model, dataset) values behind Table 2. For each cell we measure, at the answer step, the per-layer attention mass from the text tokens to the digit-pixel patches of the second image (image_anchor_digit) under the a- and d-

arms; peak $_{\Delta}$ is the a – d difference at its maximizing layer peak $_L$, and depth-norm = peak $_L/n_{\text{layers}}$. CIs are sample-instance bootstrap. The sample pool per dataset is the top and bottom susceptibility deciles of the OneVision panel ($n \leq 1,000$).

M VLM-as-judge anchoring pilot (closed-API)

We check whether cross-modal anchoring also appears on closed-API judges.

Setup. 5 closed-API judges $\times$ 2 standard judge benchmarks $\times$ 3 arms (**b**: image only / **a**: + anchor digit “1” / **m**: + Telea-masked anchor) $\times$ up to $n = 500$ paired samples per cell (effective per-cell $n = 437\text{--}498$ after parsing and provider filters). Judge task = *Visual Faithfulness 1–5 rating* (VLFeedback Silkie native rubric (Li et al., 2024)). The response selector is chosen (VLFeedback (Li et al., 2024): max average GPT-4V rating; VL-RewardBench (Li et al., 2025c): argmax human ranking), which yields a high-baseline sample set with room for the

Dataset	Judge	n	$b \rightarrow a \rightarrow m$ mean	$\Delta(a-b)$	$\Delta(m-b)$	$\Delta(a-m)$	$P(\text{score}=1) \ b \rightarrow a$
VLFeedback	gpt-4o	498	4.65 $\rightarrow$ 4.08 $\rightarrow$ 4.54	-0.57	-0.12	-0.46	3 % $\rightarrow$ 22 % (+18.9 pp)
VLFeedback	gpt-5.1	498	4.08 $\rightarrow$ 4.07 $\rightarrow$ 4.04	-0.01	-0.04	+0.03	9 % $\rightarrow$ 9 % (+0.0 pp)
VLFeedback	gemini-2.5-pro	484	3.89 $\rightarrow$ 3.95 $\rightarrow$ 4.02	+0.06	+0.10	-0.05	15 % $\rightarrow$ 15 % (+0.2 pp)
VLFeedback	gemini-2.5-flash	496	4.49 $\rightarrow$ 4.09 $\rightarrow$ 4.09	-0.39	-0.39	+0.01	9 % $\rightarrow$ 21 % (+11.5 pp)
VLFeedback	claude-sonnet-4-5	490	4.02 $\rightarrow$ 3.98 $\rightarrow$ 3.98	-0.04	-0.04	-0.00	5 % $\rightarrow$ 8 % (+3.1 pp)
VL-RewardBench	gpt-4o	498	4.52 $\rightarrow$ 3.62 $\rightarrow$ 4.13	-0.89	-0.39	-0.51	3 % $\rightarrow$ 31 % (+27.9 pp)
VL-RewardBench	gpt-5.1	497	3.27 $\rightarrow$ 3.34 $\rightarrow$ 3.20	+0.06	-0.08	+0.14	16 % $\rightarrow$ 15 % (-1.6 pp)
VL-RewardBench	gemini-2.5-pro	437 [†]	2.60 $\rightarrow$ 2.59 $\rightarrow$ 2.73	-0.01	+0.11	-0.14	35 % $\rightarrow$ 37 % (+1.8 pp)
VL-RewardBench	gemini-2.5-flash	485	4.32 $\rightarrow$ 3.77 $\rightarrow$ 3.71	-0.54	-0.60	+0.07	8 % $\rightarrow$ 26 % (+17.3 pp)
VL-RewardBench	claude-sonnet-4-5	491	3.25 $\rightarrow$ 3.10 $\rightarrow$ 3.09	-0.15	-0.15	+0.00	6 % $\rightarrow$ 13 % (+6.9 pp)

Table 19: Headline matrix (5 judge $\times$ 2 dataset $\times$ 3 arm). **Bold** = $|\Delta(a-b)| \geq 0.4$ (substantial effect).

Item	Value
Datasets	VLFeedback (MMInstruction/VLFeedback) and VL-RewardBench (MMInstruction/VL-RewardBench); target $n = 500$ each (effective per-cell $n = 437$ –498, see Table 19)
Anchor	digit “1” (inputs/irrelevant_number/1.png); m-arm uses the same image with digit pixels Telea-masked
Judges	gpt-4o, gpt-5.1, gemini-2.5-pro, gemini-2.5-flash, claude-sonnet-4-5-20250929
Reasoning	gpt-5.1, flash via reasoning_effort=minimal (non-reasoning); gemini-2.5-pro reasoning ON (provider policy)
Gateway	Staix OpenAI-compatible gateway (gateway.letsur.ai/v1); 2026-05-13/14 access

Table 20: Closed-API judge pilot configuration.

Pattern	Δ_{ab}	Δ_{mb}	Δ_{am}	Judges
Digit-specific	$\ll 0$	≈ 0	$\ll 0$	gpt-4o
Distract.-general	$\ll 0$	$\approx \Delta_{ab}$	≈ 0	gemini-2.5-flash
Null on both	≈ 0	≈ 0	≈ 0	gpt-5.1, gemini-2.5-pro, claude-sonnet-4-5

Table 21: Attack-pattern classification of judges. Subscripts: $\Delta_{ab} = \Delta(a-b)$ etc. gpt-4o detail: VLF / VLB $\Delta_{am} = -0.46/-0.51$. gemini-2.5-flash detail: $\Delta_{mb} \approx \Delta_{ab}$.

anchor to lower the score.

Two patterns. Splitting $\Delta(a-m)$ from $\Delta(m-b)$ separates two patterns (Table 21):

The digit-specific pattern (gpt-4o) is isolated by the anchored-masked contrast with the same sign as in §4.2; the $(a-m)$ contrast isolates this pattern on closed-API judges too. The distractor-general pattern (flash) is a distinct threat model (a multi-image instruction-following weakness against any second image).

Within-vendor ablations.

- **OpenAI generation (gpt-4o $\rightarrow$ gpt-5.1).** Within a vendor, the newer generation removes anchor susceptibility (mean $\Delta(a-b)$ drops from ~ 1 pt to noise). The cause (RLHF, scale, instruction-following) cannot be disentangled.
- **Google reasoning toggle (pro vs. flash).** Within the Gemini family, pro (reasoning on) shows no effect while flash (reasoning off) does. This is consistent with reasoning reducing anchoring. A pro-vs-flash capability gap acts as a confound (a clean ablation would require pro-with-reasoning-disabled, which provider policy disallows).

Caveats.

- Anchor = “1”—a single anchor digit. A run with anchor = “5” (pushing scores toward the ceiling) was not performed (an earlier 2-judge run was null due to a score-5 ceiling).
- The chosen-response selector yields a high-baseline sample set; the random-response selector attenuates the effect by approximately 10 $\times$ (the anchor effect is baseline-conditional).
- [†] gemini-2.5-pro VL-RewardBench: parse-failure rate 12 % (longer reasoning chain $\times$ max_output_tokens=2048 budget) $\rightarrow n_{\text{pair}} = 437$; sufficient for the null reading but interpret magnitude with care.
- Only gateway aliases are exposed; with the exception of claude-sonnet-4-5-20250929, models are cited by alias and access date.

Anchoring appears on two of the five closed-API judges (gpt-4o, gemini-2.5-flash).

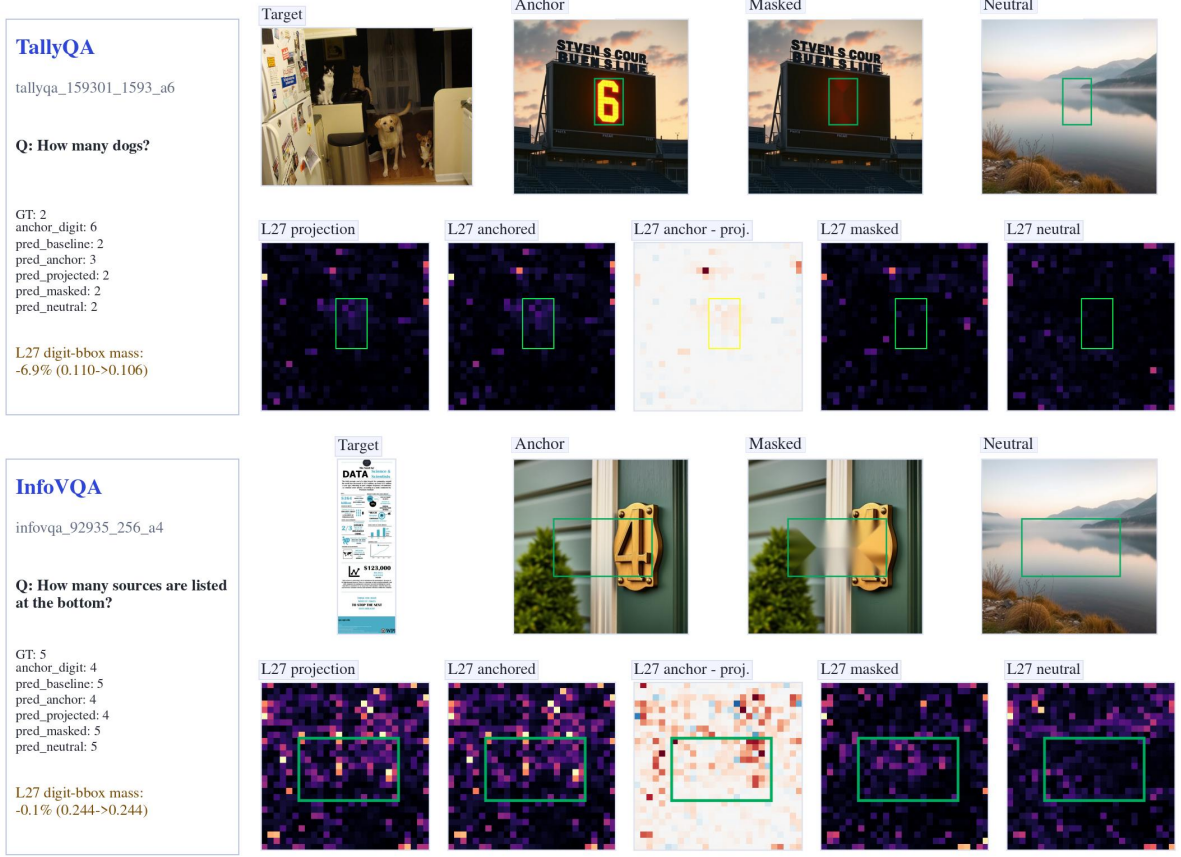


Figure 8: Exact residual rollout maps for LLaVA-OneVision-7B before and after calibrated subspace projection. Each dataset block shows one selected sample: the target image, anchor/masked/neutral control images with green digit-region overlays, and five $L = 27$ rollout-map views: projection, anchored, anchor-minus-projection, masked, and neutral. The left metadata panel reports the ground-truth answer, the anchor digit, model predictions for each condition, and the relative reduction in digit-bbox rollout mass computed from the raw rollout arrays. Green boxes mark the digit bounding boxes on the global 27×27 image-token grid of the second image.

N Rollout-map visualization details

This section expands the rollout visualization used in the paper figures. We use two LLaVA-OneVision-7B examples, one from TallyQA and one from InfoVQA, the model on which the mitigation is evaluated. The goal of the visualization is to inspect whether the final answer token’s decoder path places less mass on the digit-pixel region of the irrelevant anchor image after the calibrated subspace projection.

N.1 Visualization protocol

We use attention rollout (Abnar and Zuidema, 2020) to localize the answer-token path. For each selected sample, we first generate the answer under the relevant arm and then replay the prompt plus that answer teacher-forced. The query for the visualization is the final generated numeric token, because this is

the token whose representation directly produces the displayed answer. We compose head-averaged causal self-attention together with the residual path across decoder layers and retain only this final-token rollout row. Thus each map shows where the answer token routes probability mass over the preceding multimodal sequence.

N.2 Mapping sequence positions to the visual grid

LLaVA-OneVision represents each image with a global image-token grid plus additional AnyRes high-resolution tile tokens. For a clean spatial overlay we visualize only the global 27×27 image-token grid of the second image: the selected rollout row is sliced to this 729-token span and reshaped to a 27×27 grid. The digit bounding box from the anchor image is projected onto the same grid and drawn as the green overlay. All mass statistics in

this section are therefore digit-bbox mass on this displayed second-image grid, not mass over every AnyRes image token.

N.3 Arms and comparison policy

The panels run the four relevant arms independently: **a** (target plus anchor image), **a+projection** (the same prompt with the calibrated subspace projection at $L = 26$, $K = 8$, $\alpha = 0.5$), **m** (masked anchor), and **d** (neutral distractor). The projection-mitigated rollout is computed from the actual mitigated generation path. The difference maps show the anchored map minus the projection-mitigated map at the same layer. Positive red mass therefore marks spatial positions where the rollout mass decreased after mitigation. The masked and neutral maps are shown as controls: they help separate digit-pixel-specific mass from generic second-image distraction.

N.4 Observed pattern

In both examples, the projection reduces rollout mass on the digit bounding box at the intervention layer and in the following layer. On the TallyQA example, digit-bbox mass drops by 44.8 % at $L = 26$ and 50.8 % at $L = 27$; the digit fraction within the displayed second-image grid changes from $0.145 \rightarrow 0.104$ and $0.182 \rightarrow 0.121$, respectively. On the InfoVQA example, the corresponding reductions are 20.3 % and 23.7 %, with digit fractions $0.503 \rightarrow 0.457$ and $0.522 \rightarrow 0.464$. The masked and neutral controls provide the matching non-digit and ordinary-distraction views for the same samples. The maps are illustrative; the behavioral evidence is in the paired contrasts and the projection intervention (§6). The rollout maps show where the answer-token mass is concentrated in these selected examples.

O Examples

This appendix shows six four-condition items with the b/a/m/d answers of all six main-panel models; **b** uses the target image only, and **a**, **m**, **d** pair it with the anchor, masked, and neutral image.

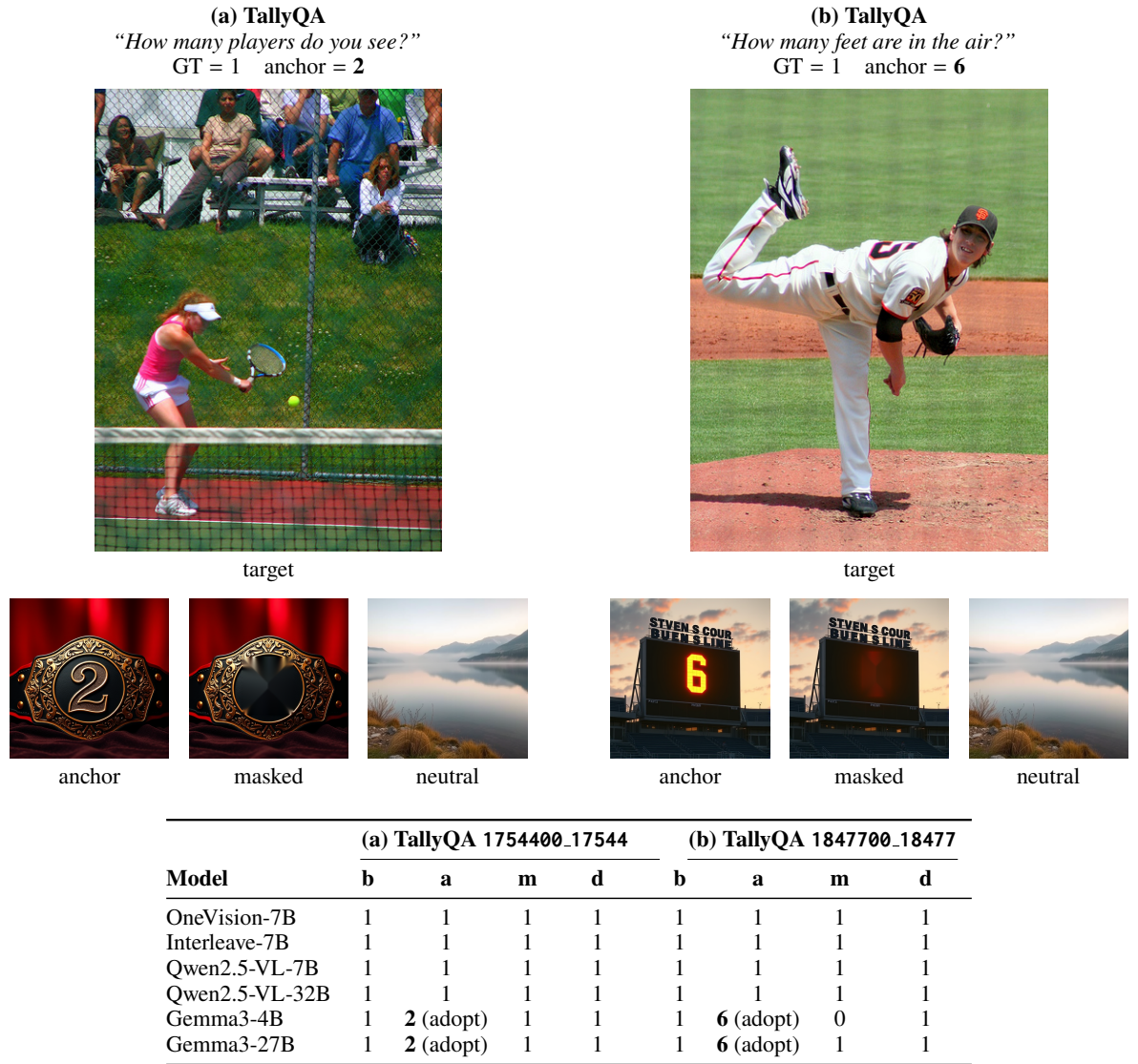
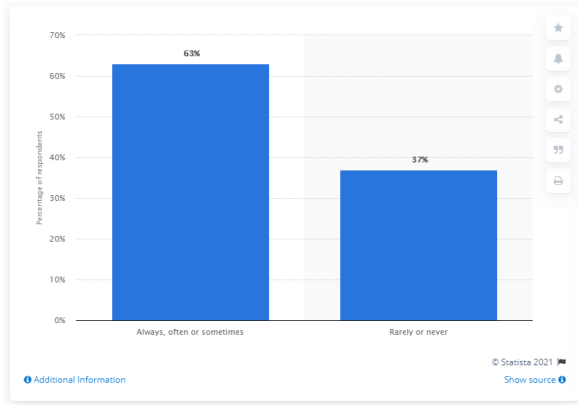


Figure 9: Four-condition examples: (a) TallyQA item 1754400_17544; (b) TallyQA item 1847700_18477. Of the six models, 2 adopt the anchor in (a) and 2 in (b).

(a) ChartQA
 “What’s the percentage of employees that never have stress?”
 GT = 37 anchor = 35



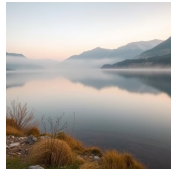
target



anchor

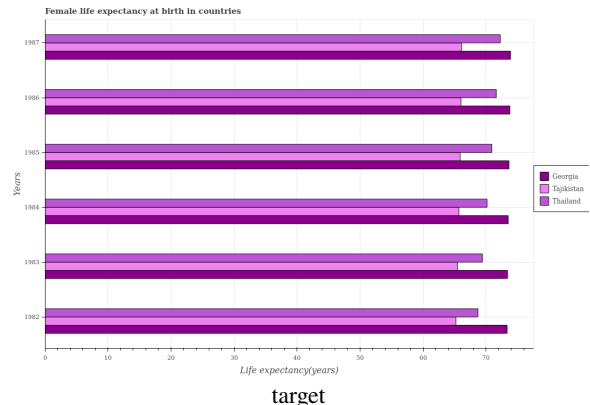


masked



neutral

(b) PlotQA
 “What is the label of the 2nd group of bars from the top?”
 GT = 1986 anchor = 1800



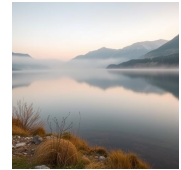
target



anchor



masked



neutral

Model	(a) ChartQA 1130_1130				(b) PlotQA 95263_11636			
	b	a	m	d	b	a	m	d
OneVision-7B	37	37	37	37	1986	1986	1986	1986
Interleave-7B	37	37	37	37	1980	1800 (adopt)	1988	1988
Qwen2.5-VL-7B	37	37	37	37	1986	1866 (SDF)	1986	1986
Qwen2.5-VL-32B	37	37	37	37	1986	1800 (adopt)	1986	1986
Gemma3-4B	37	35 (adopt)	37	37	67	70 (SDF)	70	70
Gemma3-27B	37	37	37	37	70	65	65	65

Figure 10: Four-condition examples: (a) ChartQA item 1130_1130; (b) PlotQA item 95263_11636. Of the six models, 1 adopts the anchor in (a); in (b) 2 adopt and 2 further move toward it (SDF).

(a) MathVista
 “A gymnast jotted down the number of cartwheels she did each day. What is the mode of the numbers?”
 GT = 10 anchor = 9

Doing cartwheels	
Day	Number of cartwheels
Sunday	10
Monday	10
Tuesday	9
Wednesday	4
Thursday	10
Friday	4
Saturday	9

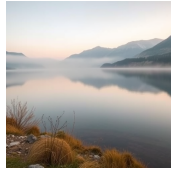
target



anchor

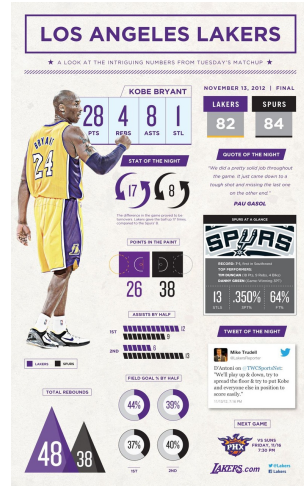


masked



neutral

(b) InfoVQA
 “What is Kobes jersey number?”
 GT = 24 anchor = 25



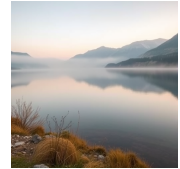
target



anchor



masked



neutral

Model	(a) MathVista 801_801				(b) InfoVQA 95488_699			
	b	a	m	d	b	a	m	d
OneVision-7B	10	10	10	10	24	24	24	24
Interleave-7B	10	10	10	10	24	25 (adopt)	24	24
Qwen2.5-VL-7B	10	10	10	10	24	24	24	24
Qwen2.5-VL-32B	10	10	10	10	24	24	24	24
Gemma3-4B	10	9 (adopt)	10	10	24	24	24	24
Gemma3-27B	10	10	10	10	24	24	24	24

Figure 11: Four-condition examples: (a) MathVista item 801_801; (b) InfoVQA item 95488_699. Of the six models, 1 adopts the anchor in (a) and 1 in (b).

P Artifact licenses

Licenses of all datasets, benchmarks, and software / methods used in this paper are listed in Table 22. Our use of each artifact is consistent with its license and intended use (academic research only; we do not redistribute modified versions where prohibited). The code release accompanying this paper is similarly intended for academic reproduction of our experiments, compatible with the research-purpose access conditions of the underlying datasets. Entries marked (not published) have no license explicitly stated on the artifact’s project page, GitHub, or HuggingFace card at the time of writing (2026-05); these artifacts are publicly accessible for academic research, and users should consult the original authors or the project page for definitive terms.

Artifact	License
<i>Numeric VQA evaluation datasets (§3.4)</i>	
TallyQA (Acharya et al., 2019)	Apache-2.0
ChartQA (Masry et al., 2022)	GPL-3.0
PlotQA (Methani et al., 2020)	CC-BY-4.0 (data); MIT (code)
InfoVQA (Mathew et al., 2022)	(not published)
MathVista (Lu et al., 2024)	CC BY-SA 4.0
<i>Capability benchmarks (§6.3)</i>	
HallusionBench (Guan et al., 2024)	BSD-3-Clause
RealWorldQA (xAI, 2024)	CC BY-ND 4.0
MMStar (Chen et al., 2024)	(not published)
POPE (Li et al., 2023b)	MIT
MMBench-DEV-EN (Liu et al., 2024a)	Apache-2.0
OCRBench (Liu et al., 2024b)	MIT
<i>Judge pilot datasets (Appendix M)</i>	
VLFeedback (Li et al., 2024)	(not published)
VL-RewardBench (Li et al., 2025c)	MIT
<i>Software / methods</i>	
VLMEvalKit (Duan et al., 2024)	Apache-2.0
FLUX.1-schnell (Black Forest Labs, 2024)	Apache-2.0
Telea (via OpenCV) (Telea, 2004)	Apache-2.0 (OpenCV)

Table 22: Licenses of artifacts used in this paper. (not published) = no license explicitly published by the original authors at the time of writing (2026-05).