# PERTURBED VISUAL KEYS AND VALUES AS A CONTRASTIVE REFERENCE FOR MULTI-IMAGE REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multi-image reasoning requires multimodal large language models (MLLMs) to integrate visual evidence across images. However, how cross-image interaction shapes the visual representations used for generation remains unclear. We examine this interaction through visual Keys and Values, which govern how text queries attend to and aggregate visual information. We observe that cross-image interaction changes both components: disabling its induced change in either one lowers accuracy. Describing later-image Keys and Values as coefficients along directions defined by earlier images, we find that cross-image interaction makes more of these coefficients negative. Removing or reversing the negative ones progressively weakens support for the correct answer. This degradation serves as an effective reference branch for Key-Perturbed Contrastive Decoding (K-PCD) and Value-Perturbed Contrastive Decoding (V-PCD), two training-free methods that contrast original predictions against internally perturbed visual branches at inference time. Across seven models from three families and four multi-image benchmarks, K-PCD and V-PCD achieve the two highest average accuracies and outperform state-of-the-art baselines in cross-model average accuracy on three of four benchmarks.

## 1 INTRODUCTION

Multimodal large language models (MLLMs) (Alayrac et al., 2022; Liu et al., 2023; 2024a; Bai et al., 2023; Chen et al., 2024c; Yin et al., 2024; Liu et al., 2025; Wang et al., 2024a) have made substantial progress in visual understanding, with recent architectures supporting multiple images within a shared context (Li et al., 2024a; Bai et al., 2025; Zhu et al., 2025; Li et al., 2024b; Laurençon et al., 2024). Multi-image reasoning requires models to compare, track, and combine evidence across image boundaries (Jiang et al., 2024; Zhao et al., 2024; Wang et al., 2025; Fu et al., 2024; Meng et al., 2025). Under causal attention, tokens from later images can attend directly to earlier ones, providing a pathway for cross-image information exchange. Understanding exactly how this interaction shapes the representations used for generation is crucial for improving multi-image reasoning.

Prior work examines positional effects, information leakage, and depth-dependent interaction patterns in multi-image models. For instance, unidirectional attention induces positional bias (Tian et al., 2025), while block-diagonal masking reduces this bias but impairs reasoning (Tian et al., 2025). Other studies show that joint encoding increases representational similarity (Park et al., 2025), delimiter tokens only partially suppress transfer (Lee et al., 2026), and cross-image mixing varies with depth (Das et al., 2026). Since text-token Queries in a transformer decoder access visual inputs strictly through visual Keys (which govern attention allocation) and visual Values (which supply the retrieved content), naturally, we narrow our focus to ask a more fundamental question: how does cross-image interaction shape the visual Keys and Values?

We first establish visual Keys and Values as functionally consequential sites of cross-image interaction. By comparing ordinary causal computation with a disabled condition (*i.e.*, where direct attention from later-image tokens to earlier-image tokens is masked), we isolate the cross-image-induced change in these representations. With the original causal mask restored, disabling this specific change in either visual Keys or Values alone lowers answer accuracy (§2.1). Unlike completely isolating images via attention masks, this links the effect directly to the representations that govern how visual information is attended to and aggregated. To understand the nature of this effect, we further ask: ① ***What does cross-image interaction do to visual Keys and Values?*** Since later images receive information from earlier ones, we measure the interaction-induced changes along directions defined by the earlier images. At both image and token levels, these shifts are consistently negative for Keys



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

in both analyzed models, with Values largely following the same pattern at smaller magnitude. When later-image Keys and Values are fitted in these directions, cross-image interaction makes more of their coefficients negative (§2.2).

② ***Does the correct answer depend on the negative components of later-image Keys and Values?*** This fit isolates directly editable components. We find that these negative components carry support for the correct answer, as removing or reversing them progressively reduces the correct candidate's margin over competing options (§2.3). This degradation yields a prediction in which the true answer specifically loses ground, providing an effective reference branch for contrastive decoding.

Based on these observations, we instantiate this mechanism as **Key-Perturbed Contrastive Decoding (K-PCD)** and **Value-Perturbed Contrastive Decoding (V-PCD)**. In both methods, the perturbed branch isolates and edits the negative components of later-image Keys or Values, leaving the images, causal attention mask, and model weights completely unchanged (§3). Across seven models from three families and four multi-image benchmarks, both methods achieve higher overall average accuracy than state-of-the-art inference-time baselines (§4.2).

Our contributions are as follows:

- We trace cross-image interaction directly to visual Keys and Values, demonstrating that disabling the cross-image-induced shift in either component impairs answer accuracy.
- We reveal that cross-image interaction shifts later-image Keys, and largely Values, away from earlier images, making negative components along earlier-image directions more prevalent. These negative components are functionally critical: removing or reversing them progressively degrades the correct answer's margin.
- We propose K-PCD and V-PCD, which exploit this internal degradation to form a targeted contrastive reference. Across seven models and four benchmarks, both achieve the two highest overall average accuracies among the evaluated training-free inference-time methods, and both lead in cross-model average accuracy on three of the four benchmarks.

## 2 Visual Keys and Values Under Cross-Image Interaction

We study how cross-image interaction relates to the visual Keys and Values that text-token Queries read, and refer to the Keys and Values as the two *sites* at which visual representations are measured and edited. We first test whether changes caused by cross-image interaction at these sites contribute to answer accuracy (§2.1). We then address ① by measuring how these changes are oriented relative to earlier-image directions, and by fitting later-image Keys and Values along these directions to isolate editable components (§2.2). Finally, we address ② by testing how editing later-image representations along these directions affects generation (§2.3).

**Preliminary: Multi-image attention.** Let $\mathcal{I}_1, \ldots, \mathcal{I}_J$ be the visual-token spans of $J$ images in prompt order. Under causal attention (Vaswani et al., 2017), a token in image $j$ can attend to visual tokens in images $i < j$. At a subsequent text position $\tau$, the visual contribution to one attention head $o_\tau^{\text{vis}}$ is:

$$o_\tau^{\text{vis}} = \sum_{j=1}^{J} \sum_{t \in \mathcal{I}_j} a_{\tau t} v_t, \qquad a_{\tau t} = \frac{\exp(q_\tau^\top k_t / \sqrt{d_h})}{\sum_{r \le \tau} \exp(q_\tau^\top k_r / \sqrt{d_h})}, \tag{1}$$

where $d_h$ is the head dimension and the denominator includes all causally visible tokens. Keys determine the attention weights $a_{\tau t}$, while Values determine the content combined with those weights. Cross-image updates to visual representations can affect both terms.

**Disabling cross-image interaction.** We compare two forward passes on the same input. The original condition uses standard causal attention, while the disabled condition removes only attention from later-image visual Queries to earlier-image visual Keys. Holding all other prompt inputs, token positions, and attention paths fixed, this comparison cleanly isolates direct visual-to-visual interaction. Throughout, cross-image interaction refers to this direct attention from later-image to earlier-image visual tokens; text tokens attend to all images in both conditions. We use superscripts $\mathrm{on}$ and $\mathrm{off}$ for quantities from the original and disabled conditions, respectively.

### 2.1 Cross-Image Changes in Both Keys and Values Affect Answer Accuracy

On 200 Mantis examples with LLaVA-OneVision-7B (Li et al., 2024a), we replace one visual component at a time with its counterpart obtained by disabling cross-image interaction. This follows

the activation-patching methodology developed for causal analysis of language models (Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024) and extended to multimodal models (Basu et al., 2024), restricted here to the visual Key and Value spans. Under the original causal mask, Key replacement supplies $(K^{\text{off}}, V^{\text{on}})$, and Value replacement supplies $(K^{\text{on}}, V^{\text{off}})$. We supply these visual tensors at every decoder layer during prefill and generation, while text-token Queries and non-visual states evolve under the intervention (Appendix S3).

Fig. 1 shows that disabling cross-image interaction lowers accuracy from $62.5\%$ to $50.5\%$. Replacing only Keys reduces accuracy to $54.5\%$, and replacing only Values reduces it to $50.5\%$, each with its own pattern of correctness transitions (Appendix S3). Thus, cross-image changes in each component contribute to answer accuracy.

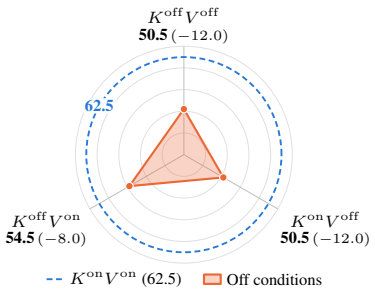



Figure 1: **Disabling cross-image interaction.** Superscripts indicate the source run of visual K/V; parentheses show the percentage-point change from $K^{\text{on}}V^{\text{on}}$.

**How Keys and Values affect the text readout.** To see how each component acts, we fix the original text Query at the answer position and decompose the readout change induced by each (Appendix S2). Key changes alter the total attention assigned to visual tokens, its distribution between images, and token selection within each image. Value changes alter the aggregated visual content mainly in direction, with the directional component approximately $7\times$ larger in norm than the magnitude component (Tab. S1). The two sites thus act through distinct channels, attention allocation and content direction, so we measure and edit them separately. Because these changes are largely directional, §2.2 asks which direction they take relative to earlier images.

### 2.2 Later-Image Keys and Values Shift Away from Earlier Images

While these results show that cross-image shifts in Keys and Values affect the final answer, the geometric nature of these internal changes remains unclear. We thus characterize them using the on and off conditions defined above: we define one direction per image, project the paired change onto earlier-image directions, express each later-image representation as coefficients along these directions, and compare these coefficients between conditions.

**Earlier-image directions.** Describing how a later image moves relative to those before it requires a frame, and the images themselves provide one. Let $x_t^s$ denote the Key or Value representation of visual token $t$ under condition $s \in \{\text{on}, \text{off}\}$, with Keys taken before rotary position encoding, Values at the Value projection output, and heads concatenated. Inspired by direction-based representation analysis (Bolukbasi et al., 2016; Kim et al., 2018; Park et al., 2024) and difference-of-means constructions (Marks & Tegmark, 2024), we define one direction per image by averaging within each image and centering across images:

$$\mu_i^s = \frac{1}{|\mathcal{I}_i|} \sum_{t \in \mathcal{I}_i} x_t^s, \qquad \bar{\mu}^s = \frac{1}{J} \sum_{k=1}^{J} \mu_k^s, \qquad u_i^s = \frac{\mu_i^s - \bar{\mu}^s}{\|\mu_i^s - \bar{\mu}^s\|_2}, \tag{2}$$

where $u_i^s = 0$ if the denominator vanishes. Each $u_i^s$ points from the shared center toward image $i$; those with $i < j$ are the earlier-image directions of image $j$. Throughout the paired analysis the directions are fixed to $u_i^{\text{off}}$, so the frame does not change with the pathway being measured. For $J = 2$, $u_1 = -u_2$, so the frame reduces to a single direction and the measurement becomes a single-direction projection; the edit in §2.3 remains well defined. Tab. 1 therefore reports the three- and four-image examples, and Appendix S4 reports the two-image and all-example splits.

**The direction of the induced changes.** For a token $t$ in a later image $j$, we project its change onto each earlier image's direction:

$$\Delta x_t = x_t^{\text{on}} - x_t^{\text{off}}, \qquad d_{t,i} = \langle \Delta x_t, u_i^{\text{off}} \rangle, \quad t \in \mathcal{I}_j,\ i < j. \tag{3}$$

A negative $d_{t,i}$ indicates a change directed away from earlier image $i$. We summarize this at two scales. Averaged over an image, with $\Delta\mu_j = \mu_j^{\text{on}} - \mu_j^{\text{off}}$:

$$D_{\text{earlier}} = \frac{\text{mean}_{j>1,\, i<j} \langle \Delta\mu_j, u_i^{\text{off}} \rangle}{\text{mean}_{j>1} \|\Delta\mu_j\|_2}, \tag{4}$$

Table 1: **Changes relative to earlier images** on the three- and four-image Mantis examples ($n = 71$). Early, middle and late are thirds of each model's decoder layers (0–8, 9–18 and 19–27 for LLaVA-OV-7B; 0–7, 8–15 and 16–23 for InternVL3-1B); layer 0 precedes the first cross-image update and is excluded. Panels follow the order of the text. Paired change $x^{\text{on}} - x^{\text{off}}$ projected onto earlier-image directions, summarized per image ($D_{\text{earlier}}$, Eq. 4) and per token ($R_{\text{neg}}$, Eq. 5). Negative-fraction change: original minus disabled share of negative coefficients (pp), each condition fit in its own directions (Tab. S3). Coefficients decreasing: share of later-image coefficients with $c^{\text{on}} < c^{\text{off}}$, both conditions refit in the disabled pass's directions. Appendix S4 reports mean coefficient changes, the remaining example splits, and confidence intervals.

| | | Image-level $D_{\text{earlier}}$ | | | Token-level $R_{\text{neg}}$ (%) | | | Negative-fraction change (pp) | | | Coefficients decreasing (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Site | Early | Middle | Late | Early | Middle | Late | Early | Middle | Late | Early | Middle | Late |
| LLaVA-OV-7B | K | −0.178 | −0.220 | −0.194 | 66.5 | 73.5 | 70.8 | +2.5 | +19.1 | +11.2 | 72.2 | 81.9 | 79.9 |
| | V | −0.069 | −0.109 | −0.062 | 58.8 | 64.6 | 59.1 | +3.4 | +21.9 | +13.2 | 61.4 | 70.1 | 62.4 |
| InternVL3-1B | K | −0.115 | −0.253 | −0.255 | 59.2 | 67.1 | 66.3 | +2.7 | +3.3 | +4.6 | 65.4 | 76.7 | 73.8 |
| | V | +0.058 | −0.140 | −0.078 | 44.5 | 62.6 | 55.4 | −1.7 | +3.7 | +2.5 | 39.2 | 67.5 | 57.8 |

and counted over individual tokens:

$$R_{\text{neg}} = \text{mean}_{j>1,\, i<j,\, t\in\mathcal{I}_j} \mathbf{1}[d_{t,i} < 0]. \tag{5}$$

The two are related by $\langle \Delta\mu_j, u_i^{\text{off}} \rangle = |\mathcal{I}_j|^{-1} \sum_{t\in\mathcal{I}_j} d_{t,i}$: the first measures how far the average token moves, the second how widely that sign is shared. Only together do they separate a shift driven by a few tokens from one distributed across the image.

**Reconstruction coefficients.** The projections above require both passes, which an inference-time method does not have; the same directions, however, can describe a single representation rather than a difference. Stacking them into $U_j = [u_1, \ldots, u_{j-1}]^\top$, we reconstruct each later-image token by a ridge fit (Hoerl & Kennard, 1970):

$$c_t = (U_j U_j^\top + \lambda I)^{-1} U_j x_t, \qquad \lambda = 0.01. \tag{6}$$

We call the entries of $c_t$ the *coefficients* of this fit and $U_j^\top c_t$ the corresponding *reconstruction terms* in representation space; §2.3 edits the latter. Because the earlier-image directions are not constrained to be orthogonal, the coefficients are joint reconstruction weights rather than independent projections (Appendices S4 and S8 discuss $\lambda$).

**How cross-image interaction changes the coefficients.** Negative coefficients select the reconstruction terms edited by the method in §3, so we examine how cross-image interaction changes these coefficients in two steps. First, we fit each condition in its own current directions, as the method does at inference time, and compare the fraction of negative coefficients. Since the frame is centered ($\sum_k (\mu_k - \bar{\mu}) = 0$), negative coefficients also encode image distinctness present without cross-image interaction, so the informative quantity is the change in this fraction between conditions. Second, because this comparison allows the frame itself to change, we refit both conditions in the disabled pass's directions $u_i^{\text{off}}$ and compare the coefficients one by one, so that only the representations differ.

**Cross-image interaction makes negative coefficients more prevalent.** In Tab. 1, Key changes point away from earlier images in both models: image means do so at every depth, and $59.2\%$–$73.5\%$ of individual tokens move in the same direction. When each condition is fit in its own directions, cross-image interaction raises the fraction of negative coefficients in 11 of 12 site–depth cells, by up to $21.9$ pp on these examples and $28.9$ pp over all examples (Tab. S3). With the frame held fixed, $65.4\%$–$81.9\%$ of the Key coefficients decrease, so the representations themselves shift in the negative direction even when the directions are unchanged. The effect is largest in middle layers in the token-level and fixed-direction panels for both models; at the image level it peaks in the middle band for LLaVA-OV-7B and in the late band for InternVL3-1B. Values follow the same direction in almost all settings, with smaller paired projections than Keys and a comparable rise in the negative fraction; the only exception is the early band of InternVL3-1B. The consistency across image-level shifts, token-level distributions, and coefficient trends shows that the observed directional movement is broadly distributed across representations rather than being driven by outliers.

Thus, negative coefficients encode how a later image is set apart from earlier images, and cross-image interaction sharpens this separation. §2.3 tests whether this separation supports the correct answer.

### 2.3 Removing or Reversing Negative Reconstruction Terms Weakens the Correct Answer's Margin

Unlike the paired change $x_t^{\text{on}} - x_t^{\text{off}}$ measured in §2.2, the reconstruction terms describe the current representation in earlier-image directions and need no paired run, which is what an inference-time method requires. We therefore test directly whether removing or reversing the negative terms weakens support for the correct candidate at the first answer position.

**Perturbation.** We apply Eqs. 2–6 to the branch's current representations to obtain $U_j$ and $c_t$ (superscripts are dropped, since each branch uses its own current directions). For a later-image token $x_t$, let $c_{t,-} = \min(c_t, 0)$ elementwise, so that $U_j^\top c_{t,-}$ collects its negative reconstruction terms (the negative components referred to in §1). We edit these terms with strength $\alpha$:

$$x_t(\alpha) = x_t - 2\alpha\, U_j^\top c_{t,-}, \tag{7}$$

which scales the negative reconstruction terms by $1 - 2\alpha$, so $\alpha = -\frac{1}{2}, 0, \frac{1}{2}, 1$ strengthen, retain, remove, and reverse them. This additive form is the operation used by activation steering (Subramani et al., 2022; Turner et al., 2023; Li et al., 2023a; Zou et al., 2023; Rimsky et al., 2024); unlike weight-level editing (Meng et al., 2022), it leaves parameters unchanged and affects only the current forward pass. Each operation is applied at all decoder layers during prefill, separately to Keys and Values, with directions and coefficients recomputed from the edited branch at each layer, so the perturbation includes the downstream effects of earlier edits.

**Measuring the effect on the answer.** We run this diagnostic on LLaVA-OV-7B and Qwen2.5-VL-7B, scoring single-token option labels at the first answer position on 64 Mantis multiple-choice examples. Alongside candidate accuracy, we measure the correct answer's margin against the strongest incorrect option, selected under the original branch and held fixed across edits:

$$y^- = \arg\max_{y \in \mathcal{Y} \setminus \{y^\star\}} z_0(y), \qquad m_\alpha = z_\alpha(y^\star) - z_\alpha(y^-), \qquad g_\alpha = m_0 - m_\alpha, \tag{8}$$

where $y^\star$ is the correct option, $\mathcal{Y}$ the option set, and $z_\alpha$ the next-token logits under strength $\alpha$, with $\alpha = 0$ giving the original branch. Positive $g_\alpha$ indicates reduced support for the correct answer relative to the same competitor.

Removal and reversal progressively reduce both accuracy and margins in both models (see Tab. 2). On LLaVA-OV-7B, reversal lowers Key- and Value-side accuracy from $57.81\%$ to $42.19\%$ and $51.56\%$, with mean margin losses of $0.729$ and $0.630$, i.e., $76\%$ and $66\%$ of the original mean margin. Qwen2.5-VL-7B shows the same ordering with larger effects: from the original $79.69\%$, removal lowers accuracy to $62.50\%$ and $70.31\%$, and reversal to $45.31\%$ and $31.25\%$. Across both models and sites, reversal yields positive losses on $68.8\%$–$78.1\%$ of examples. Appendix S5 reports full results and candidate transitions. Strengthening also lowers Key-side accuracy, but removal and reversal produce larger margin losses at both sites; on the Value side, strengthening leaves the margin nearly unchanged (mean $g_\alpha = -0.06$ and $0.06$ vs. $0.63$ and $2.21$ for reversal on LLaVA-OV-7B and Qwen2.5-VL-7B).

Table 2: **Prediction effects of all-layer negative-coefficient editing** on 64 Mantis examples per model. Accuracy (%) is the argmax over candidate labels at the first answer position. Margin loss $g_\alpha$ uses the incorrect competitor selected under the original branch; original mean margins are $0.955$ (LLaVA-OV-7B) and $2.229$ (Qwen2.5-VL-7B). Remove and Reverse act on the negative reconstruction terms. The original row is shared by both sites.

| | | | LLaVA-OV-7B | | Qwen2.5-VL-7B | |
|---|---|---|---|---|---|---|
| Site | Operation | $\alpha$ | Acc. | Mean $g_\alpha$ | Acc. | Mean $g_\alpha$ |
| — | Original | 0 | 57.81 | 0.000 | 79.69 | 0.000 |
| K | Strengthen | $-0.5$ | 50.00 | 0.077 | 64.06 | 0.696 |
| K | Remove | 0.5 | 54.69 | 0.189 | 62.50 | 0.861 |
| K | Reverse | 1 | 42.19 | 0.729 | 45.31 | 2.072 |
| V | Strengthen | $-0.5$ | 57.81 | $-0.062$ | 76.56 | 0.063 |
| V | Remove | 0.5 | 54.69 | 0.327 | 70.31 | 0.553 |
| V | Reverse | 1 | 51.56 | 0.630 | 31.25 | 2.210 |

Together, these results show that the **negative reconstruction terms carry support for the correct answer**: removing them lowers its mean margin, and reversing them lowers it most. As the directions are non-orthogonal, we read the edit as weakening the separation that cross-image interaction strengthens (§2.2), not as guaranteed movement toward each earlier image. In decoding, negative-coefficient editing also outperforms positive-only and norm-matched random editing in every setting (§4.3.1, Tabs. 6 and S12), yielding the reference branch contrastive decoding requires.

# 3 CONTRASTIVE DECODING WITH VISUAL KEY AND VALUE PERTURBATIONS

Building on the finding in §2.3 that reversing negative reconstruction terms reduces the correct answer's margin, we propose **Key-Perturbed Contrastive Decoding (K-PCD)** and **Value-Perturbed Contrastive Decoding (V-PCD)** (Fig. 2). Unlike methods altering inputs or attention masks, we directly perturb visual representations to construct the degraded reference branch; §4.3.3 compares this choice against an attention-masked reference.

**Selective visual perturbation.** At each decoder layer, earlier-image directions $U_j$ and coefficients $c_t$ (Eqs. 2–6) are computed from current representations immediately before editing; thus, earlier edits propagate naturally to subsequent layers. For perturbation strength $\alpha \geq 0$, we edit the negative reconstruction terms $U_j^\top c_{t,-}$ (§2.3):

$$\widetilde{x}_t = x_t - 2\alpha U_j^\top c_{t,-}, \quad t \in \mathcal{I}_j,\ j > 1, \alpha \geq 0. \quad (9)$$

**Perturbing visual Keys or Values.** K-PCD applies Eq. 9 to visual Keys before positional transformation (Su et al., 2024), while V-PCD applies the same operation in Value space with its corresponding means, directions, and coefficients. Direct edits are confined to later-image visual tokens at the chosen projection site, after which attention and representation updates proceed normally without disrupting cross-image attention. For V-PCD, editing along earlier-image directions rather than scaling norms is motivated by direction changes dominating magnitude variations (§2.1).

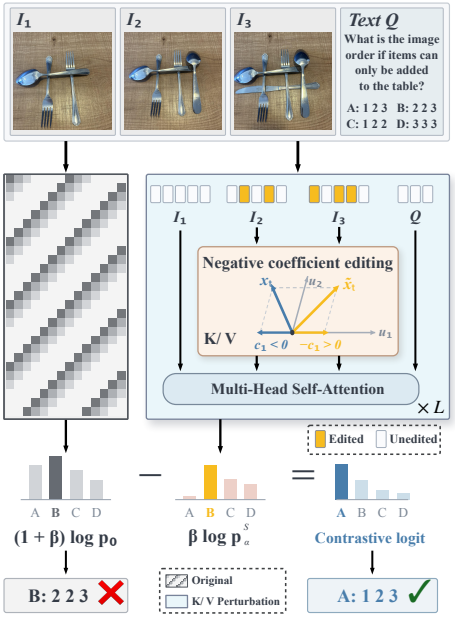



Figure 2: **Overview of K-PCD and V-PCD.** Editing is applied to later-image visual Keys (K-PCD) or Values (V-PCD) by editing reconstruction terms with negative coefficients ($c_t < 0$) along earlier-image directions. Contrastive decoding penalizes predictions from this degraded branch, amplifying correct candidates.

**Contrastive decoding against the perturbed visual branch.** Given input $\xi$, let $p_0(y)$ and $p_\alpha^S(y)$ denote next-token distributions from the original and perturbed branch at site $S \in \{K, V\}$ with strength $\alpha$. Following Leng et al. (2024), decoding proceeds as:

$$y_\tau = \arg\max_{y \in \mathcal{V}} \left[(1+\beta)\log p_0(y) - \beta \log p_\alpha^S(y)\right], \quad (10)$$

where $\beta \geq 0$ controls contrast strength. The score margin between a correct–incorrect candidate pair becomes $m_0 + \beta(m_0 - m_\alpha)$, where $m_\alpha$ is the margin in the perturbed branch (Eq. 8). Since reversal reduces the correct candidate's margin on average (§2.3; $g_\alpha > 0$ on 68.8%–78.1% of examples across both models and sites), this adjustment favors the correct candidate on average. Here, $\alpha$ controls representation degradation, while $\beta$ governs its decoding penalty.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Implementation details.** We evaluate seven pre-trained models from three MLLM families: Qwen2.5-VL-3B/7B (Bai et al., 2025), InternVL3-1B/2B/8B (Zhu et al., 2025), and LLaVA-OneVision-0.5B/7B (Li et al., 2024a), spanning from 0.5B to 8B parameters and different visual-tokenization schemes. We adopt the validation protocol of DelimScaling (Lee et al., 2026) for hyperparameter selection: for each model–benchmark pair, we select the perturbation strength $\alpha$ and contrast strength $\beta$ on a fixed 10% selection subset and freeze them for full-set evaluation. Baselines that require selection are tuned on the same subset; FOCUS uses its published configurations where available (see Appendix S10).

**Baselines.** We compare against three state-of-the-art training-free methods that intervene at different points in multi-image processing: VCD (Leng et al., 2024) (visual-input perturbation), DelimScaling (Lee et al., 2026) (delimiter-based intervention), and FOCUS (Park et al., 2025) (image-wise

Table 3: **Evaluation across four multi-image benchmarks.** Accuracy (%) under our shared harness and the selection protocols in §4.1. Averages weight all seven models equally. Bold and underline mark the best and second-best values in each column; tied entries share the same mark, and when the best value is tied, the next distinct value is underlined.

| Dataset | Method | Qwen2.5-VL | | InternVL3 | | | LLaVA-OV | | Average |
|---|---|---|---|---|---|---|---|---|---|
| | | 3B | 7B | 1B | 2B | 8B | 0.5B | 7B | |
| Mantis | VCD(Leng et al., 2024) | 57.14 | **71.43** | 49.77 | 55.76 | 68.20 | 41.01 | 60.83 | 57.73 |
| | DelimScaling(Lee et al., 2026) | 58.53 | 69.12 | 48.85 | 56.22 | 66.82 | 41.01 | 64.52 | 57.87 |
| | FOCUS(Park et al., 2025) | 58.99 | **71.43** | 47.93 | **57.60** | 67.28 | 43.32 | 63.13 | 58.53 |
| | **K-PCD (ours)** | 58.53 | **71.43** | **53.46** | 56.22 | 70.51 | **47.00** | 64.06 | **60.17** |
| | **V-PCD (ours)** | **60.83** | 70.97 | 49.77 | 53.92 | **70.97** | 43.32 | **64.98** | 59.25 |
| MIRB | VCD(Leng et al., 2024) | 54.39 | **63.16** | 33.44 | 43.24 | **53.97** | 28.79 | 48.50 | 46.50 |
| | DelimScaling(Lee et al., 2026) | 56.66 | 62.64 | **39.53** | 42.11 | 50.36 | **32.09** | 47.88 | 47.32 |
| | FOCUS(Park et al., 2025) | 56.24 | 61.40 | **39.53** | 44.07 | 52.84 | 31.17 | 48.40 | 47.66 |
| | **K-PCD (ours)** | **57.28** | 63.05 | 38.60 | 44.79 | 52.53 | 31.99 | 48.09 | 48.05 |
| | **V-PCD (ours)** | 56.66 | 62.54 | 37.56 | **45.72** | 53.04 | 31.79 | **49.43** | **48.11** |
| MuirBench | VCD(Leng et al., 2024) | 39.31 | 45.42 | 29.62 | **27.62** | 37.00 | 24.58 | 33.85 | 33.91 |
| | DelimScaling(Lee et al., 2026) | **42.00** | **47.62** | 28.85 | 27.12 | 35.50 | **25.50** | **35.27** | **34.55** |
| | FOCUS(Park et al., 2025) | 34.35 | 45.54 | 28.04 | 26.69 | 36.42 | 24.73 | 34.00 | 32.82 |
| | **K-PCD (ours)** | 37.38 | 46.42 | **29.73** | 27.46 | 37.00 | 25.12 | 35.08 | 34.03 |
| | **V-PCD (ours)** | 37.58 | 46.08 | 28.88 | 27.31 | **38.69** | 24.46 | 35.12 | 34.02 |
| QBench2 | VCD(Leng et al., 2024) | **64.20** | **76.00** | 50.20 | **66.20** | 74.90 | 51.50 | 72.70 | 65.10 |
| | DelimScaling(Lee et al., 2026) | 61.70 | 75.70 | 49.90 | 65.30 | 76.40 | 40.20 | 73.70 | 63.27 |
| | FOCUS(Park et al., 2025) | 63.00 | 75.50 | 50.20 | 64.50 | 75.60 | 50.80 | **73.80** | 64.77 |
| | **K-PCD (ours)** | 63.30 | 75.50 | 50.30 | 65.30 | **76.70** | **52.60** | 73.60 | 65.33 |
| | **V-PCD (ours)** | 63.20 | 75.90 | **51.20** | 65.00 | 76.60 | 52.30 | 73.50 | **65.39** |

prediction aggregation). All methods use the shared evaluation protocol; Appendix S10 describes each baseline and its configuration selection.

**Benchmarks.** We evaluate multi-image reasoning on Mantis (Jiang et al., 2024), MIRB (Zhao et al., 2024), MuirBench (Wang et al., 2025), and QBench2 (Zhang et al., 2024b), using the complete Mantis, MIRB, and MuirBench test splits and the QBench2 development split, containing 217, 969, 2,600, and 1,000 examples, respectively. Appendix S10 describes their task coverage.

### 4.2 Main Results

**Comparison with inference-time baselines.** K-PCD and V-PCD achieve the two highest overall accuracies in Tab. 3: averaged equally over the 28 model–benchmark settings, they reach $51.89\%$ and $51.69\%$, compared with $50.95\%$ for FOCUS, the strongest baseline. These gains hold beyond the selection subset: on the remaining held-out samples, paired tests against FOCUS give $+0.80\%$ for K-PCD (95% CI $[+0.28\%, +1.33\%]$, Holm-adjusted $p = 0.0028$) and $+1.17\%$ for V-PCD (95% CI $[+0.67\%, +1.69\%]$, Holm-adjusted $p < 0.0002$; Appendix S10).

The gains are largest on Mantis, where K-PCD reaches $60.17\%$ against $58.53\%$ for FOCUS, and are concentrated on the smallest backbones: LLaVA-OV-0.5B and InternVL3-1B improve by $3.68\%$ and $3.69\%$ over their strongest baseline. Both methods also lead on MIRB and QBench2, by narrower margins of $0.2\%$–$0.5\%$ over the best baseline average. MuirBench is the exception, where DelimScaling attains the highest average ($34.55\%$ vs. $34.03\%$ for K-PCD). This gap comes mainly from Qwen2.5-VL-3B ($42.00\%$ vs. $37.38\%$); over the remaining six models, K-PCD averages higher than DelimScaling, and both of our methods remain ahead of VCD and FOCUS. When Mantis is split by image count, our methods remain best or tied-best for both two-image and three–four-image inputs (Appendix S7).

**Comparison of Key and Value perturbations.** Both variants exceed FOCUS when averaged separately within each model family, and neither site dominates. Their cross-model averages are within $0.06\%$ on MIRB, MuirBench, and QBench2; the clearest gap is on Mantis, where K-PCD leads by $0.92\%$, driven mainly by the two smallest backbones. V-PCD instead gives the best accuracy among all evaluated methods in several individual settings, including InternVL3-8B on Mantis ($70.97\%$), LLaVA-OV-7B on MIRB ($49.43\%$), and InternVL3-8B on MuirBench ($38.69\%$). The preferred site thus varies with the model and benchmark.

Table 4: **Single-site vs. joint editing.**

| Model | Benchmark | K-PCD | V-PCD | Joint |
|---|---|---|---|---|
| LLaVA-OV-7B | Mantis | 64.06 | **64.98** | 64.52 |
| | MuirBench | **34.83** | 34.50 | 33.00 |
| | MIRB | 48.50 | **48.67** | 48.17 |
| InternVL3-8B | Mantis | 70.51 | **70.97** | 70.05 |
| | MuirBench | 37.83 | **39.83** | 38.00 |
| | MIRB | 50.33 | **50.67** | 50.17 |
| Average | | 51.01 | **51.60** | 50.65 |

Table 5: **A-PCD vs. K-PCD and V-PCD.**

| Benchmark | A-PCD | K-PCD | V-PCD |
|---|---|---|---|
| Mantis | 57.21 | **60.17** | 59.25 |
| MIRB | 47.97 | 48.05 | **48.11** |
| MuirBench | 33.77 | **34.03** | 34.02 |
| QBench2 | **65.56** | 65.33 | 65.39 |
| Average | 51.13 | **51.89** | 51.69 |

## 4.3 Ablation Studies

### 4.3.1 Coefficient Selection and Random Editing

To isolate the effect of direction from magnitude, we replace the coefficient-based edit with an isotropic random direction at the same site, matching per-token edit norm and layer coverage at each perturbation strength, using the same contrastive decoding rule. We evaluate on Mantis with LLaVA-OV-7B and InternVL3-8B, with parameters selected on the fixed 10% selection subset. Tab. 6 reports this independently tuned comparison for the two models. Appendix S9 gives the selected configurations and an additional comparison that also includes Qwen2.5-VL-7B, in which the random control uses the parameters selected for PCD (see Tab. S12).

Table 6: **Coefficient selection and random editing** on Mantis using LLaVA-OV-7B and InternVL3-8B.

| Model | Site | PCD | All coef. reversed | Positive coef. | Random | Δ |
|---|---|---|---|---|---|---|
| LLaVA-OV-7B | K | **64.06** | **64.06** | 63.59 | 62.52 | +1.54 |
| | V | **64.98** | **64.98** | 64.06 | 64.06 | +0.92 |
| InternVL3-8B | K | **70.51** | 69.59 | 68.66 | 69.89 | +0.62 |
| | V | **70.97** | 70.05 | 67.28 | 68.97 | +2.00 |
| Average | | **67.63** | 67.17 | 65.90 | 66.36 | +1.27 |

As seen in Tab. 6, PCD (K-PCD or V-PCD at the corresponding site) outperforms random editing for both models, at both K and V, averaging $67.63\%$ against $66.36\%$. Editing only the positive coefficients gives the lowest average accuracy ($65.90\%$), below the four-setting average of random editing, and reversing all coefficients does not exceed PCD in any of the four settings ($67.17\%$ on average). In these settings, the results support selecting negative coefficients along earlier-image directions when constructing the perturbed branch.

### 4.3.2 Single-Site and Joint Editing

Since both Key and Value perturbations yield useful contrasts, we next ask whether combining them improves on either alone. The joint variant applies the negative-coefficient edit to both Keys and Values within a single perturbed branch. We evaluate LLaVA-OV-7B and InternVL3-8B on all Mantis examples and fixed 600-example subsets of MuirBench and MIRB, each variant selecting its parameters on the same 10% selection subset. Joint editing averages $50.65\%$, below K-PCD ($51.01\%$) and V-PCD ($51.60\%$; Tab. 4). It trails V-PCD in all six settings and K-PCD in four, and falls $1.83\%$ short of the better single-site variant on MuirBench for both models. One possible reason is that the two edits interact: the Key edit changes the attention weights with which the edited Values are aggregated. Since each single-site reversal already removes much of the correct answer's margin (Tab. 2), compounding the two may also lower the competing candidates, and the contrastive gain $\beta(m_0 - m_\alpha)$ in Eq. 10 shrinks when the reference degrades all candidates alike. A single-site perturbation thus gives the more selective reference.

### 4.3.3 Attention Masking as the Reference Branch

A simpler way to weaken cross-image evidence is to remove it at the attention mask. **Attention-Perturbed Contrastive Decoding (A-PCD)** runs the disabled condition of §2 as its reference branch, blocking direct attention from later-image to earlier-image visual tokens, and substitutes its distribution $p^{\text{off}}(y)$ for $p^S_\alpha(y)$ in Eq. 10. Inputs, model weights, and the contrastive rule are unchanged, and $\beta$ is selected on the same 10% selection subset. Across all 28 model–benchmark settings (Tab. 5; per-model results in Tab. S7), A-PCD averages $51.13\%$, above FOCUS ($50.95\%$) but below K-PCD ($51.89\%$) and V-PCD ($51.69\%$), and it exceeds the better of the two in only 4 of

28 settings, while K-PCD and V-PCD are higher than A-PCD in 20 and 17 settings, respectively. The gap is concentrated on Mantis (57.21% vs. 60.17% and 59.25%) and is largest on the smallest backbones, where K-PCD leads A-PCD by 5.99% on both InternVL3-1B and LLaVA-OV-0.5B. On the other three benchmarks, the three variants are close (49.10%, 49.13%, and 49.17% on average), with A-PCD slightly ahead on QBench2 (65.56% vs. 65.39% for V-PCD). One explanation is that masking removes all direct cross-image exchange at once, whereas our edit reverses only the negative reconstruction terms shown to carry support for the correct answer (§2.3), giving a more selective reference while leaving the causal mask intact.

# 5 Related Work

## 5.1 Contrastive Decoding with a Degraded Branch

Contrastive decoding contrasts a model's prediction with a degraded reference branch to mitigate hallucination. Hallucination in vision-language models has been surveyed by Bai et al. (2024) and is assessed using protocols such as CHAIR and POPE (Rohrbach et al., 2018; Li et al., 2023c). Existing reference constructions include *model-side* methods using smaller models or earlier layers (Li et al., 2023b; Chuang et al., 2024); *input-side* methods using image noise, disturbance instructions, context removal, text-only predictions, or locally grounded visual regions (Leng et al., 2024; Wang et al., 2024c; Shi et al., 2024; Favero et al., 2024; Chen et al., 2024b); and *token-side* methods such as SID, which retains the least important visual tokens after early layers to form its reference branch (Huo et al., 2025). Related attention-based decoding methods include OPERA, which combines an over-trust penalty with retrospection-allocation (Huang et al., 2024), and PAI, which combines image-attention amplification with text-only logit correction (Liu et al., 2024b). In multi-image settings, recent approaches isolate images against noise (Park et al., 2025) or mask text-to-text attention (Park et al., 2026). We take a complementary route to constructing a contrastive reference: intervening directly on visual Keys or Values. Our Key-Perturbed and Value-Perturbed Contrastive Decoding (K-PCD and V-PCD) keep the original inputs, causal mask, token sequence, and model weights intact, deriving the reference branch by perturbing later-image Keys or Values along earlier-image directions.

## 5.2 Interpreting Visual Representations Inside MLLMs

Our analysis also complements literature on how MLLMs internally represent and route visual content. Basu et al. (2024) extend causal tracing / activation patching from text-only models (Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024) to multimodal models, whereas Zhang et al. (2025) use attention knockout to analyze cross-modal information transfer. Other work analyzes depth-wise token consumption to guide pruning (Chen et al., 2024a) or decomposes visual encoders into text-aligned concepts (Gandelsman et al., 2024). A parallel line of work describes internal states through linear directions, identified by concept-based probing or difference-of-means constructions (Bolukbasi et al., 2016; Kim et al., 2018; Park et al., 2024; Marks & Tegmark, 2024), and steers model behavior by adding such directions to activations at inference time (Subramani et al., 2022; Turner et al., 2023; Li et al., 2023a; Zou et al., 2023; Rimsky et al., 2024). We adopt the same additive, inference-time intervention, but build the directions from the images themselves rather than from labeled concepts, and edit later-image Keys and Values to form a contrastive reference rather than the final prediction.

# 6 Conclusion

In this work, we study how cross-image interaction shapes the visual representations underlying multi-image reasoning, looking beyond attention flow alone. We find that cross-image changes in visual Keys and Values both affect answer accuracy. When evaluated along directions defined by earlier images, these interaction-induced shifts are consistently negative for Keys in both analyzed models, with Values exhibiting the same pattern in almost all settings. These fits isolate negative components that can be edited at inference time; removing or reversing them progressively reduces the correct answer's margin over competing options. Motivated by this insight, Key-Perturbed Contrastive Decoding (K-PCD) and Value-Perturbed Contrastive Decoding (V-PCD) construct targeted degraded references directly from within the visual representations. Across seven models from three families and four multi-image benchmarks, both achieve the two highest overall average accuracies among the evaluated inference-time methods. More broadly, our findings demonstrate that representation-level visual perturbation offers an effective, training-free framework for multi-image reasoning in MLLMs.

### AI USE STATEMENT

AI assistance was used to polish the manuscript text, write code, and help run the experiments for this work. All AI-assisted text, code, and experimental outputs were reviewed and verified by the authors.

### ETHICS STATEMENT

This work studies inference-time methods using existing models and established visual question-answering benchmarks. It does not collect new human-subject data. The methods inherit the capabilities and limitations of the underlying models and datasets; the reported benchmark improvements concern answer accuracy and do not establish suitability for safety-critical decisions.

### REPRODUCIBILITY STATEMENT

K-PCD and V-PCD are implemented in PyTorch (Paszke et al., 2019). Experiments are conducted on NVIDIA RTX 6000 Ada 48GB GPUs. To support reproducibility, we will release the full implementation after acceptance. Implementation details are included in §4.1, and the analysis protocols in Appendices S2–S5.

## REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *EMNLP*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *NeurIPS*, 2024.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, 2016.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024a.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *ICML*, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024c.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *ICLR*, 2024.

Anurag Das, Adrian Bulat, Alberto Baldrati, Ioannis Maniadis Metaxas, Bernt Schiele, Georgios Tzimiropoulos, and Brais Martinez. More images, more problems? a controlled analysis of vlm failure modes. In *ACL Findings*, 2026.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *CVPR*, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *ICLR*, 2024.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, 2024.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *ICLR*, 2025.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *NeurIPS*, 2024.

Minyoung Lee, Yeji Park, Dongjun Hwang, Yejin Kim, Seong Joon Oh, and Junsuk Choe. Enhancing multi-image understanding through delimiter token scaling. In *ICLR*, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, 2024.

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

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*, 2023a.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *ACL*, 2023b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023c.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *ECCV*, 2024b.

Yiyang Liu, James Liang, Ruixiang Tang, Yugyung Lee, Majid Rabbani, Sohail Dianat, Raghuveer Rao, Lifu Huang, Dongfang Liu, Qifan Wang, et al. Re-imagining multimodal instruction tuning: A representation view. In *ICLR*, 2025.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *COLM*, 2024.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *ICLR*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *ICML*, 2024.

Wooje Park, Insu Lee, Soohyun Kim, Jaeyun Jang, Minyoung Noh, Kyuhong Shim, and Byonghyo Shim. Revealing multi-view hallucination in large vision-language models. *arXiv preprint arXiv:2603.23934*, 2026.

Yeji Park, Minyoung Lee, Sanghyuk Chun, and Junsuk Choe. Mitigating cross-image information leakage in LVLMs for multi-image tasks. *arXiv preprint arXiv:2508.13744*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *ACL*, 2024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *NAACL*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In *ACL Findings*, 2022.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *CVPR*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, 2020.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *ICLR*, 2025.

Taowen Wang, Yiyang Liu, James Chenhao Liang, Junhan Zhao, Yiming Cui, Yuning Mao, Shaoliang Nie, Jiahao Liu, Fuli Feng, Zenglin Xu, et al. M2pt: Multimodal prompt tuning for zero-shot instruction learning. In *EMNLP*, 2024a.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. ReTaKe: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*, 2024b.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *ACL Findings*, 2024c.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *ICLR*, 2024.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. EventHallusion: Diagnosing event hallucinations in video LLMs. *arXiv preprint arXiv:2409.16597*, 2024a.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *ICLR*, 2023.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *CVPR*, 2025.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-Bench$^+$: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10404–10418, 2024b.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

## SUMMARY OF THE APPENDIX

This appendix contains additional analyses, experimental details, and results of our ICLR 2027 submission: *Perturbed Visual Keys and Values as a Contrastive Reference for Multi-Image Reasoning*, organized as follows:

- §S1 includes more **related work** discussions on cross-image interaction.
- §S2 provides the **detailed text readout analysis**, decomposing Key-induced changes into attention allocation and Value-induced changes into aggregated visual content.
- §S3 gives **details of disabling the cross-image changes in Keys and Values**, including answer-margin and correctness-transition statistics.
- §S4 details the **measurements along earlier-image directions**, covering the paired projection changes, the fixed-direction coefficient comparison, and the coefficient signs within a single forward pass.
- §S5 reports the **full results for editing negative reconstruction terms**, detailing the effects of strengthening, removing, and reversing these components across all layers.
- §S6 presents **additional design ablations**, covering joint Key and Value perturbations, Query gating, editing the shared hidden representation, and the attention-masked reference branch (A-PCD).
- §S7 reports **performance with more input images**, splitting Mantis by image count.
- §S8 reports **efficiency and sensitivity to** $\lambda$ on LLaVA-OV-7B.
- §S9 lists the **random editing details**, including selected configurations and parameter-transfer results.
- §S10 describes the **evaluation protocol**, including hyperparameter selection, baseline configurations, and benchmark coverage.
- §S11 discusses the **social impact** of this work, covering its potential benefits and dual-use risks.
- §S12 lists the **asset licenses and usage terms** of the models and benchmarks used in this work.
- §S13 outlines **limitations and future work**, discussing computational scalability, dynamic hyperparameter routing, and extensions to video-language tasks.

## S1 Additional Discussion of Prior Work on Cross-Image Interaction in MLLMs

Understanding how images interact inside a shared context is crucial for multi-image reasoning. Existing studies characterize this interaction at the *position-level* to mitigate order bias (Tian et al., 2025), the *flow-level* to measure or regulate transfer between images (Park et al., 2025; Lee et al., 2026), and the *layer-level* to map cross-image mixing depth (Das et al., 2026). Our focus is on how cross-image interaction changes visual Keys and Values, through which text-token Queries access visual tokens. We investigate this functional interface directly: establishing the causal relevance of both components (§2.1), characterizing the geometric shifts relative to earlier images induced by cross-image attention (§2.2), and demonstrating how targeted edits on negative reconstruction terms degrade correct-answer margins (§2.3) to motivate our contrastive decoding framework.

## S2 Detailed Text Readout Analysis

This appendix provides the detailed readout analysis summarized in §2.1. It uses LLaVA-OneVision-7B on the same 200 Mantis examples (a fixed random subset, seed 42) across all 28 decoder layers. The original and disabled conditions are defined in §2.

For the readout analysis, we fix the original text Query at the final prompt position, which predicts the first answer token. Let $\mathcal{R}(Q, K, V)$ denote the multi-head attention output after head concatenation and the output projection $W_O$, before residual addition. Its softmax includes all causally visible tokens, including non-visual tokens. We construct $\bar{K}^{\text{off}}$ and $\bar{V}^{\text{off}}$ by substituting the corresponding visual positions from the disabled condition into the original Key and Value tensors, retaining the original non-visual positions; $\bar{K}^{\text{on}} = K^{\text{on}}$ and $\bar{V}^{\text{on}} = V^{\text{on}}$. At each layer, we isolate the representational changes:

$$\begin{aligned} o^{s,s'} &= \mathcal{R}(Q^{\text{on}}, \bar{K}^{s}, \bar{V}^{s'}), \qquad s, s' \in \{\text{on}, \text{off}\}, \\ \delta_K &= o^{\text{on,on}} - o^{\text{off,on}}, \qquad \delta_V = o^{\text{on,on}} - o^{\text{on,off}}. \end{aligned} \tag{11}$$

Table S1: **Components of the fixed-Query readout change.** Mean normalized magnitudes $\nu$ from Eq. 12, with 95% bootstrap confidence intervals, on 200 examples. All entries report $100\nu$; the signed projections $\rho$ are reported in the text. Component norms are measured after head concatenation and $W_O$. Early, middle, and late are thirds of the 28 decoder layers of LLaVA-OV-7B.

| Component | Early | Middle | Late |
|---|---|---|---|
| K: visual/non-visual mass | 3.81 [3.67, 3.96] | 4.84 [4.64, 5.05] | 4.39 [4.17, 4.63] |
| K: between images | 1.14 [1.09, 1.18] | 2.05 [1.98, 2.12] | 2.28 [2.09, 2.49] |
| K: within-image tokens | 1.40 [1.36, 1.44] | 2.44 [2.35, 2.54] | 2.55 [2.43, 2.68] |
| V: magnitude | 0.188 [0.181, 0.195] | 0.501 [0.479, 0.526] | 0.412 [0.378, 0.448] |
| V: direction | 1.35 [1.31, 1.39] | 3.56 [3.41, 3.73] | 3.07 [2.83, 3.33] |

Thus, $\delta_K$ isolates the effect of changing visual Keys while fixing Queries and Values, whereas $\delta_V$ isolates the effect of changing visual Values under identical attention weights. Here, we use post-RoPE (Su et al., 2024) Queries and Keys, projected Values, and the model's attention scaling and grouped-query head mapping (Ainslie et al., 2023).

The following subsections decompose each $\delta_S$, $S \in \{K, V\}$, into component vectors $\delta_{S,f}$. We report their normalized magnitudes and signed projections onto the corresponding total change as:

$$\nu_{S,f} = \frac{\|\delta_{S,f}\|_2}{\|o^{\mathrm{on,on}}\|_2}, \qquad \rho_{S,f} = \frac{\langle \delta_{S,f}, \delta_S \rangle}{\|\delta_S\|_2^2}. \tag{12}$$

Component vectors sum to the total change; their norms need not add. The signed projections sum to one when defined, but may be negative or exceed one. All vector comparisons are made after $W_O$. We first compute each metric per example and layer, average over valid layers within early, middle, and late depth thirds, and then average equally across examples. Depth thirds are defined by $(\ell + 1/2)/28$ for zero-indexed layer $\ell$. Readout confidence intervals use 2,000 percentile bootstrap resamples of examples (Efron & Tibshirani, 1993).

### S2.1 Changes in Attention Allocation

Visual Keys can change the readout at three levels: the total attention assigned to visual tokens, its distribution between images, and token selection within each image. To separate these effects, consider one attention head and let $a_t^s$ be its attention weight at token $t$ with Query $Q^{\mathrm{on}}$ and Key source $\bar{K}^s$, $s \in \{\mathrm{on}, \mathrm{off}\}$. For image spans $\mathcal{I}_1, \ldots, \mathcal{I}_J$ and their union $\mathcal{I}$, define:

$$\begin{aligned} M^s &= \sum_{t \in \mathcal{I}} a_t^s, \qquad p_j^s = \frac{\sum_{t \in \mathcal{I}_j} a_t^s}{M^s}, \\ \bar{v}_j^s &= \frac{\sum_{t \in \mathcal{I}_j} a_t^s v_t^{\mathrm{on}}}{M^s p_j^s}. \end{aligned} \tag{13}$$

Here $M^s$ is the visual attention mass, $p_j^s$ is image $j$'s share of that mass, and $\bar{v}_j^s$ is its conditional Value readout. Let $\bar{v}_{\mathrm{nv}}$ be the conditional readout over non-visual tokens. Because the Query and non-visual Keys and Values are fixed, this conditional readout is unchanged when visual Keys are replaced with $\bar{K}^{\mathrm{off}}$. The complete head output therefore factors as:

$$F(M^s, p^s, \bar{v}^s) = (1 - M^s)\bar{v}_{\mathrm{nv}} + M^s \sum_j p_j^s \bar{v}_j^s. \tag{14}$$

We average the marginal vector changes over all six substitution orders, sharing interactions symmetrically.

Tab. S1 shows that the allocation of attention between visual and non-visual tokens is the largest component of the Key-induced readout change in every depth band. The corresponding mean signed projections $\rho$ decrease from $0.760$ in early layers to $0.598$ in late layers, while the between-image component grows from $0.046$ to $0.153$. These results locate the Key-side effect at multiple levels of allocation, with changes in the visual attention mass producing the largest average readout component.

### S2.2 Changes in Aggregated Visual Content

We next hold the entire attention distribution fixed and examine how Value changes affect the content it aggregates. At each visual token and Key/Value head, write $v^s = n^s e^s$, where $n^s = \|v^s\|_2$ and $e^s$ is its unit direction. The symmetric decomposition:

$$\begin{aligned} \Delta v_{\text{mag}} &= \tfrac{1}{2}(n^{\text{on}} - n^{\text{off}})(e^{\text{on}} + e^{\text{off}}), \\ \Delta v_{\text{dir}} &= \tfrac{1}{2}(n^{\text{on}} + n^{\text{off}})(e^{\text{on}} - e^{\text{off}}) \end{aligned} \tag{15}$$

satisfies $\Delta v_{\text{mag}} + \Delta v_{\text{dir}} = v^{\text{on}} - v^{\text{off}}$. Linearity after projection gives $\delta_{V,\text{mag}} + \delta_{V,\text{dir}} = \delta_V$.

Directional changes dominate this decomposition across all three depth bands (Tab. S1). Their mean normalized magnitude is approximately seven times that of magnitude changes. Thus, the effect of cross-image attention on Value readout is expressed predominantly through changes in Value direction.

We also examine how contributions from different image positions combine. Under fixed original attention $a^{\text{on}}$, define:

$$\begin{aligned} \mathbf{r}_j^s &= W_O \,\text{Concat}_h \left[ \sum_{t \in \mathcal{I}_j} a_{h,t}^{\text{on}} v_{g(h),t}^s \right], \\ \kappa^s &= \frac{\|\sum_j \mathbf{r}_j^s\|_2}{\sum_j \|\mathbf{r}_j^s\|_2}, \end{aligned} \tag{16}$$

where $g(h)$ maps a Query head to its Key/Value head. The coherence $\kappa^s$ measures how much of the summed contribution norms remain after vector addition. The coherence is lower with original Values than with disabled-condition Values in every depth band: the mean paired difference $\kappa^{\text{on}} - \kappa^{\text{off}}$ is $-0.0030$ $[-0.0031, -0.0028]$ in early, $-0.0293$ $[-0.0307, -0.0280]$ in middle, and $-0.0026$ $[-0.0038, -0.0014]$ in late layers (95% bootstrap CIs over examples). Cross-image interaction therefore also changes the geometry with which image contributions combine, most strongly in middle layers.

## S3 Details of Disabling Cross-Image Changes in Keys and Values

The experiment disabling the component-specific cross-image changes uses LLaVA-OneVision-7B on a fixed random subset of 200 Mantis examples (seed 42), using float16, matched scaled dot-product attention (SDPA) mask implementations, and greedy generation capped at 32 tokens. The condition disabling the Key change supplies post-RoPE visual Keys $K^{\text{off}}$ and visual Values $V^{\text{on}}$; the condition disabling the Value change supplies the converse combination. These tensors are supplied at every decoder layer during both prefill and generation, while non-visual states evolve naturally under the intervention. A joint condition supplying $(K^{\text{off}}, V^{\text{off}})$ reaches $50.5\%$ accuracy, matching the fully disabled condition. For the answer-margin analysis, we use the 184 eligible multiple-choice examples where the option label is a single token at the original prompt boundary. The incorrect competitor is selected under the original condition and held fixed (Eq. 8). Disabling the Key change reduces the mean margin by $0.504$, affecting $66.3\%$ of examples. Disabling the Value change reduces the mean margin by $0.638$, affecting $64.7\%$ of examples, compared with $0.688$ when cross-image interaction is fully disabled. In full generation, disabling the Key change turns 27 correct answers into incorrect ones (and corrects 11), while disabling the Value change yields 34 such errors (and corrects 10).

## S4 Measurements Along Earlier-Image Directions

Tab. 1 uses the direct projection measurements defined in Eqs. 3–5. Earlier-image directions are constructed from the disabled condition and held fixed. For every later-image token $t$ and earlier image $i < j$, we compute $d_{t,i} = \langle x_t^{\text{on}} - x_t^{\text{off}}, u_i^{\text{off}} \rangle$.

For each example and layer, the image-level score averages $\langle \Delta\mu_j, u_i^{\text{off}} \rangle$ over all earlier-image pairs $i < j$ and normalizes it by the mean image-level change magnitude. The token-level score is the fraction of eligible token–earlier-image pairs with $d_{t,i} < 0$. Both model captures contain 216 Mantis examples with two to four images and no failed examples. The two models are run in different precisions: LLaVA-OneVision-7B in float16 with 28 decoder layers (0–27) and InternVL3-1B in

Table S2: **Fixed-direction coefficient comparison.** Percentage of later-image coefficients with $c^{\text{on}} < c^{\text{off}}$, alongside 95% bootstrap confidence intervals over examples, and the mean coefficient change $c^{\text{on}} - c^{\text{off}}$. Both conditions are refitted in the disabled condition's earlier-image directions. Layer 0 is excluded to match the main analysis.

| | | Coefficients decreasing (%) | | | Mean $c^{\text{on}} - c^{\text{off}}$ | | |
|---|---|---|---|---|---|---|---|
| Model | Site | Early | Middle | Late | Early | Middle | Late |
| *All examples* ($n = 216$) | | | | | | | |
| LLaVA-OV-7B | K | 74.2 [71.9, 76.2] | 77.8 [74.7, 80.8] | 72.6 [68.7, 76.3] | $-0.160$ | $-1.203$ | $-1.260$ |
| | V | 61.4 [60.0, 62.7] | 66.8 [64.3, 69.1] | 58.4 [57.0, 59.9] | $-0.016$ | $-0.274$ | $-0.098$ |
| InternVL3-1B | K | 66.7 [65.2, 68.2] | 78.1 [75.8, 80.3] | 75.4 [71.9, 79.1] | $-0.180$ | $-1.065$ | $-1.546$ |
| | V | 40.0 [38.5, 41.6] | 67.9 [65.6, 70.3] | 57.3 [55.5, 59.2] | $+0.015$ | $-0.245$ | $-0.110$ |
| *Two images* ($n = 145$) | | | | | | | |
| LLaVA-OV-7B | K | 75.1 [72.3, 78.0] | 75.8 [71.6, 79.8] | 69.0 [63.4, 74.5] | $-0.138$ | $-0.947$ | $-0.913$ |
| | V | 61.4 [59.5, 63.0] | 65.1 [61.8, 68.3] | 56.4 [54.6, 58.2] | $-0.014$ | $-0.220$ | $-0.064$ |
| InternVL3-1B | K | 67.3 [65.4, 69.2] | 78.8 [75.7, 81.6] | 76.2 [71.9, 80.4] | $-0.158$ | $-0.972$ | $-1.412$ |
| | V | 40.3 [38.4, 42.3] | 68.1 [65.0, 71.1] | 57.1 [54.8, 59.4] | $+0.013$ | $-0.216$ | $-0.076$ |
| *Three or four images* ($n = 71$) | | | | | | | |
| LLaVA-OV-7B | K | 72.2 [68.9, 75.4] | 81.9 [78.7, 85.0] | 79.9 [76.0, 83.6] | $-0.204$ | $-1.728$ | $-1.970$ |
| | V | 61.4 [59.8, 63.1] | 70.1 [67.4, 72.7] | 62.4 [60.6, 64.2] | $-0.020$ | $-0.385$ | $-0.166$ |
| InternVL3-1B | K | 65.4 [63.0, 67.7] | 76.7 [73.5, 79.9] | 73.8 [68.2, 79.1] | $-0.226$ | $-1.254$ | $-1.819$ |
| | V | 39.2 [37.1, 41.4] | 67.5 [64.3, 70.9] | 57.8 [54.7, 60.8] | $+0.017$ | $-0.303$ | $-0.180$ |

bfloat16 with 24 decoder layers (0–23). Captured token-block means agree with the independently recorded image means to within $1.53 \times 10^{-5}$ in both models.

### S4.1 Fixed-Direction Coefficient Comparison

The rightmost panel of Tab. 1 compares the fitted reconstruction coefficients of the two conditions using fixed earlier-image directions. The directions and the ridge fit follow Eqs. 2 and 6, built from the disabled condition and applied unchanged to both conditions. This ensures that the directions remain constant and only the representations differ. For each later-image token $t$ and coefficient $i < j$, we compute $c_t^{\text{on}}$ and $c_t^{\text{off}}$ and report the fraction of coefficients with $c_t^{\text{on}} < c_t^{\text{off}}$.

Since each $u_i$ is a unit vector, $\operatorname{tr}(U_j U_j^\top) = j - 1$, so the ridge constant $\lambda = 0.01$ in Eq. 6 equals the trace-normalized term $0.01 \operatorname{tr}(U_j U_j^\top)/(j-1)$; degenerate directions with $u_i = 0$ simply reduce the effective rank of $U_j$.

Tab. S2 reports this comparison for all three example splits. The two-image split is reported separately because $u_1 = -u_2$ leaves a single coefficient per token, measuring a scalar projection shift rather than a multi-direction reconstruction.

### S4.2 Coefficient Signs of Key and Value Representations

To align with the intervention in §2.3, we additionally fit each condition using earlier-image directions constructed from its own current representations, and report the fraction of negative reconstruction coefficients in Tab. S3. These are the quantities directly available to K-PCD and V-PCD at inference time. Negative coefficients are already common in the disabled condition due to the centered, jointly fitted construction. These frequencies describe how often the intervention selects a negative coefficient.

## S5 Full Results for Editing Negative Reconstruction Terms

The diagnostic in §2.3 evaluates LLaVA-OneVision-7B and Qwen2.5-VL-7B using bfloat16, SDPA, and batch size one. For each model, it selects 64 eligible Mantis multiple-choice examples in a fixed random order (seed 42), spanning two to four images. Each example is evaluated under the original branch and three editing strengths per site, without any contrastive score mixing.

At each layer, the edited branch constructs its own current earlier-image directions and coefficients before applying the editing rule (Eq. 7). Thus, $\alpha$ specifies an all-layer editing rule, with preceding edits naturally contributing to the representations in later layers. Strengthening, removal, and reversal use $\alpha = -0.5$, $0.5$, and $1$, respectively.

Table S3: **Coefficient signs within a single forward pass.** Percentage of later-image token–earlier-image coefficients with a negative coefficient when each condition is fitted using its own earlier-image directions. Cells report original / disabled, followed by their difference in percentage points. Layer 0 is excluded to match the main analysis.

| Model | Site | Early | Middle | Late |
|---|---|---|---|---|
| *All examples* ($n = 216$) | | | | |
| LLaVA-OV-7B | K | 74.9 / 69.0 (+5.9) | 93.9 / 65.0 (+28.9) | 87.3 / 63.5 (+23.9) |
| | V | 73.7 / 71.1 (+2.6) | 93.2 / 71.4 (+21.9) | 85.7 / 74.1 (+11.5) |
| InternVL3-1B | K | 77.9 / 73.2 (+4.6) | 83.4 / 80.6 (+2.8) | 91.6 / 89.4 (+2.2) |
| | V | 73.0 / 73.9 (−0.9) | 89.3 / 85.6 (+3.7) | 82.9 / 80.9 (+1.9) |
| *Two images* ($n = 145$) | | | | |
| LLaVA-OV-7B | K | 77.6 / 70.0 (+7.6) | 99.4 / 65.7 (+33.7) | 95.1 / 64.9 (+30.1) |
| | V | 76.1 / 73.8 (+2.2) | 95.5 / 73.6 (+21.9) | 87.8 / 77.1 (+10.7) |
| InternVL3-1B | K | 79.3 / 73.7 (+5.6) | 89.0 / 86.5 (+2.5) | 99.7 / 98.7 (+1.0) |
| | V | 75.7 / 76.2 (−0.5) | 91.8 / 88.1 (+3.7) | 84.9 / 83.2 (+1.7) |
| *Three or four images* ($n = 71$) | | | | |
| LLaVA-OV-7B | K | 69.4 / 67.0 (+2.5) | 82.7 / 63.6 (+19.1) | 71.6 / 60.4 (+11.2) |
| | V | 68.9 / 65.5 (+3.4) | 88.6 / 66.8 (+21.9) | 81.4 / 68.1 (+13.2) |
| InternVL3-1B | K | 75.0 / 72.3 (+2.7) | 72.0 / 68.7 (+3.3) | 75.0 / 70.4 (+4.6) |
| | V | 67.5 / 69.2 (−1.7) | 84.2 / 80.5 (+3.7) | 78.8 / 76.3 (+2.5) |

Table S4: **Detailed all-layer editing results.** All entries for a model use the same 64 examples. Transition columns count changes in candidate correctness relative to the original branch.

| Model | Site | Operation | Mean $m_\alpha$ | Mean $g_\alpha$ | Median $g_\alpha$ | $g_\alpha > 0$ (%) | Right→wrong | Wrong→right |
|---|---|---|---|---|---|---|---|---|
| LLaVA-OV-7B | K | Strengthen | 0.8778 | 0.0773 | 0.0060 | 50.0 | 7 | 2 |
| | K | Remove negative | 0.7666 | 0.1885 | 0.1250 | 56.3 | 4 | 2 |
| | K | Reverse negative | 0.2266 | 0.7285 | 0.6250 | 68.8 | 15 | 5 |
| | V | Strengthen | 1.0166 | −0.0615 | 0.0000 | 40.6 | 3 | 3 |
| | V | Remove negative | 0.6279 | 0.3271 | 0.1250 | 59.4 | 5 | 3 |
| | V | Reverse negative | 0.3252 | 0.6299 | 0.3438 | 68.8 | 11 | 7 |
| Qwen2.5-VL-7B | K | Strengthen | 1.5321 | 0.6964 | 0.1875 | 57.8 | 10 | 0 |
| | K | Remove negative | 1.3672 | 0.8613 | 0.7500 | 71.9 | 12 | 1 |
| | K | Reverse negative | 0.1562 | 2.0723 | 1.5000 | 73.4 | 25 | 3 |
| | V | Strengthen | 2.1660 | 0.0625 | 0.0000 | 45.3 | 4 | 2 |
| | V | Remove negative | 1.6758 | 0.5527 | 0.3750 | 62.5 | 7 | 1 |
| | V | Reverse negative | 0.0186 | 2.2100 | 1.6875 | 78.1 | 32 | 1 |

Tab. S4 provides the full margin and candidate-transition summary. The original mean margin is $0.9551$ for LLaVA-OV-7B and $2.2285$ for Qwen2.5-VL-7B. The largest reversal-induced losses coincide with more correct-to-incorrect than incorrect-to-correct candidate transitions at both sites in both models, consistent with these negative reconstruction terms supporting the correct answer's margin in this diagnostic.

# S6 Additional Design Ablations

## S6.1 Joint Key and Value Perturbations and Query Gating

Tab. S5 expands the averages in Tab. 4 and adds a Query-gated variant. The comparison uses the complete Mantis split and fixed 600-example subsets of MuirBench and MIRB, shared across all variants. Parameter selection uses the 10% selection subsets of the full benchmarks (21, 260, and 96 examples, respectively) employing the grid and tie-break rules from Appendix S10.

Joint editing applies the negative-coefficient editing rule separately to the current Key and Value representations within a single branch. The Query-gated variant retains the Key edit but adds a binary coefficient filter. This filter retains a selected coefficient only when its reversal increases the attention logit for the final prompt Query, accounting for the model's positional transformation.

On these six settings, K-PCD reaches $51.01\%$ average accuracy, compared with $50.45\%$ after adding the Query gate (Tab. S5). The original rule outperforms the Query gate in five of six comparisons, indicating that query-aware filtering does not improve on the unsupervised coefficient-based selection in these settings.

Table S5: **Full results for joint editing and Query gating.** Generation accuracy (%) on the comparison sets. Average gives equal weight to the six model–benchmark settings.

| Model | Benchmark | $n$ | K-PCD | V-PCD | Joint K+V | K + Query gate |
|---|---|---|---|---|---|---|
| LLaVA-OV-7B | Mantis | 217 | 64.06 | **64.98** | 64.52 | 63.59 |
| | MuirBench | 600 | **34.83** | 34.50 | 33.00 | 33.83 |
| | MIRB | 600 | 48.50 | **48.67** | 48.17 | 47.67 |
| InternVL3-8B | Mantis | 217 | 70.51 | **70.97** | 70.05 | 69.12 |
| | MuirBench | 600 | 37.83 | **39.83** | 38.00 | 37.50 |
| | MIRB | 600 | 50.33 | 50.67 | 50.17 | **51.00** |
| Average | | | 51.01 | **51.60** | 50.65 | 50.45 |

Table S6: **Hidden-state editing.** Generation accuracy (%) in the four-setting control. K-PCD scores are directly sourced from Tab. 3.

| Model | Benchmark | $n$ | K-PCD | Hidden-state edit |
|---|---|---|---|---|
| Qwen2.5-VL-7B | Mantis | 217 | 71.43 | 70.51 |
| | MuirBench | 2600 | 46.42 | 45.38 |
| InternVL3-8B | Mantis | 217 | 70.51 | 69.12 |
| | MuirBench | 2600 | 37.00 | 37.00 |
| Average | | | 56.34 | 55.50 |

### S6.2 Editing the Shared Hidden Representation

We also applied the same coefficient-based editing rule to the normalized hidden representations at each decoder layer, right before the Query, Key, and Value projections. This modified representation feeds all three projections, while the pre-normalization residual connection remains intact. This variant constructs its earlier-image directions and coefficients in the hidden representation space and utilizes the identical two-branch contrastive decoding rule.

Tab. S6 compares these scores with the K-PCD results from Tab. 3. Hidden-state editing obtains an average accuracy of $55.50\%$, trailing the $56.34\%$ achieved by K-PCD. This highlights the benefit of performing perturbations directly at the functional Key or Value sites rather than broadly across the shared hidden states.

### S6.3 Attention-Masked Reference Branch (A-PCD)

A-PCD shares the two-branch contrastive rule of Eq. 10, but its reference branch is the disabled condition of §2: direct attention from later-image to earlier-image visual tokens is masked, while text tokens attend to all images and the inputs, token positions, and model weights are unchanged. It therefore has no perturbation strength $\alpha$; the contrast strength $\beta$ is selected on the same 10% selection subset and frozen for full-set evaluation. Tab. S7 reports the full results next to K-PCD and V-PCD from Tab. 3. Averaged over the 28 settings, A-PCD reaches $51.13\%$, compared with $51.89\%$ for K-PCD and $51.69\%$ for V-PCD. It is below K-PCD in 20 settings and below V-PCD in 17, and it is above the better of the two in only 4 settings: three on QBench2 and one on MIRB.

## S7 Performance with More Input Images

To examine robustness as image count increases, we split Mantis by image count (145 two-image examples; 71 with three or four; the single five-image example is excluded) and compare K-PCD and V-PCD with the three baselines on LLaVA-OV-7B and InternVL3-8B (Tab. S8). Our methods achieve the best or tied-best accuracy in every model and image-count group: K-PCD leads on two-image LLaVA-OV-7B and three–four-image InternVL3-8B, and V-PCD leads on two-image InternVL3-8B and matches the best result on three–four-image LLaVA-OV-7B. Our methods use two decoding branches for every image-count group, whereas image-wise aggregation such as FOCUS requires a number of branches that grows with image count; measured costs are reported in §S8.

Table S7: **Attention-masked vs. representation-perturbed reference branches.** Accuracy (%) on the four benchmarks under the protocol of §4.1. A-PCD masks direct later-to-earlier image attention in the reference branch; K-PCD and V-PCD results are repeated from Tab. 3. Bold marks the best of the three variants in each column; averages weight all seven models equally.

| Dataset | Method | Qwen2.5-VL | | InternVL3 | | | LLaVA-OV | | Average |
|---|---|---|---|---|---|---|---|---|---|
| | | 3B | 7B | 1B | 2B | 8B | 0.5B | 7B | |
| Mantis | A-PCD | 57.14 | 70.05 | 47.47 | 52.07 | 68.66 | 41.01 | 64.06 | 57.21 |
| | K-PCD (ours) | 58.53 | **71.43** | **53.46** | **56.22** | 70.51 | **47.00** | 64.06 | **60.17** |
| | V-PCD (ours) | **60.83** | 70.97 | 49.77 | 53.92 | **70.97** | 43.32 | **64.98** | 59.25 |
| MIRB | A-PCD | 56.66 | **63.78** | 38.49 | 44.79 | 52.32 | 31.58 | 48.19 | 47.97 |
| | K-PCD (ours) | **57.28** | 63.05 | **38.60** | 44.79 | 52.53 | **31.99** | 48.09 | 48.05 |
| | V-PCD (ours) | 56.66 | 62.54 | 37.56 | **45.72** | **53.04** | 31.79 | **49.43** | **48.11** |
| MuirBench | A-PCD | 36.92 | 46.38 | 28.73 | 27.31 | 37.38 | 24.77 | 34.88 | 33.77 |
| | K-PCD (ours) | 37.38 | **46.42** | **29.73** | **27.46** | 37.00 | **25.12** | 35.08 | **34.03** |
| | V-PCD (ours) | **37.58** | 46.08 | 28.88 | 27.31 | **38.69** | 24.46 | **35.12** | 34.02 |
| QBench2 | A-PCD | **64.40** | 75.40 | **51.30** | 65.20 | 76.50 | 52.40 | **73.70** | **65.56** |
| | K-PCD (ours) | 63.30 | 75.50 | 50.30 | **65.30** | **76.70** | **52.60** | 73.60 | 65.33 |
| | V-PCD (ours) | 63.20 | **75.90** | 51.20 | 65.00 | 76.60 | 52.30 | 73.50 | 65.39 |
| Overall average (28 settings): A-PCD 51.13, K-PCD **51.89**, V-PCD 51.69 | | | | | | | | | |

Table S8: **Accuracy by image count** on Mantis (%). Two-image: 145 examples; three–four-image: 71 examples. Bold marks the best value in each column.

| | LLaVA-OV-7B | | InternVL3-8B | |
|---|---|---|---|---|
| Method | 2 img. | 3–4 img. | 2 img. | 3–4 img. |
| VCD | 66.21 | 49.30 | 72.41 | 59.15 |
| DelimScaling | 68.97 | **54.93** | 69.66 | 60.56 |
| FOCUS | 68.97 | 50.70 | 70.34 | 60.56 |
| K-PCD (ours) | **70.34** | 52.11 | 74.48 | **61.97** |
| V-PCD (ours) | 69.66 | **54.93** | **75.86** | 60.56 |

Table S9: **Efficiency comparison** (LLaVA-OV-7B, 32 Mantis samples).

| Method | TTFT (s) | TPOT (s) | Peak mem. (GiB) | Gen. tokens | Total (s) |
|---|---|---|---|---|---|
| Vanilla | 0.200 | 0.023 | 15.47 | 4.6 | 8.8 |
| DelimScaling | 0.202 | 0.023 | 15.47 | 4.6 | 8.9 |
| FOCUS | 0.784 | 0.068 | 15.84 | 4.3 | 31.7 |
| K-PCD (ours) | 0.415 | 0.044 | 15.59 | 4.5 | 17.9 |
| V-PCD (ours) | 0.424 | 0.044 | 15.59 | 4.6 | 18.3 |

## S8 Efficiency and Sensitivity to $\lambda$

These are small-scale checks on LLaVA-OV-7B with fixed Mantis subsets. They are not intended as a comprehensive robustness study.

**Efficiency.** We measure 32 fixed Mantis samples with batch size 1 and greedy decoding. TTFT spans from input preprocessing and branch construction to the first selected token, and all timing points are CUDA-synchronized. TPOT is computed per sample as (last-token time − first-token time)/(tokens − 1); since outputs are short (4.3–4.6 tokens on average), TPOT rests on few decoding steps. Peak memory is reset per sample, includes model weights, and is reported as the maximum over samples; total time sums over samples.

K-PCD and V-PCD roughly double TTFT and total time relative to vanilla decoding, consistent with their two decoding branches, and add about 0.12 GiB of peak memory. On the same samples, FOCUS incurs about $3.9\times$ TTFT, $3.6\times$ total time, and 0.37 GiB additional peak memory (Tab. S9).

**Sensitivity to $\lambda$.** All experiments fix $\lambda = 0.01$ without tuning. Because the earlier-image directions are not orthogonal, the ridge term can affect the fitted coefficients, including which ones are negative

Table S10: **Sensitivity to $\lambda$** (LLaVA-OV-7B, 64 Mantis samples). Transitions and edit differences are relative to $\lambda = 0.01$; relative norm difference averages over nonzero reference edits.

| Method | $\lambda$ | Acc. (%) | W→R / R→W | Sign chg. | Rel. norm diff. | Cosine |
|---|---|---|---|---|---|---|
| K-PCD | 0.001 | 65.6 | 1 / 1 | 0.28% | 0.014 | 1.000 |
| K-PCD | 0.01 | 65.6 | – | – | – | – |
| K-PCD | 0.1 | 65.6 | 1 / 1 | 2.02% | 0.106 | 0.997 |
| V-PCD | 0.001 | 67.2 | 0 / 0 | 0.16% | 0.013 | 1.000 |
| V-PCD | 0.01 | 67.2 | – | – | – | – |
| V-PCD | 0.1 | 65.6 | 0 / 1 | 1.32% | 0.105 | 0.998 |

Table S11: **Selected parameters for PCD and random editing on Mantis.** All configurations use every decoder layer.

| | | PCD | | Random, independent | |
|---|---|---|---|---|---|
| Model | Site | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ |
| LLaVA-OV-7B | K | 1.00 | 0.30 | 0.75 | 0.30 |
| | V | 1.00 | 0.10 | 1.00 | 0.02 |
| Qwen2.5-VL-7B | K | 1.00 | 0.02 | 1.00 | 0.30 |
| | V | 1.00 | 0.02 | 1.00 | 0.02 |
| InternVL3-8B | K | 1.00 | 0.30 | 0.50 | 0.30 |
| | V | 0.75 | 0.30 | 1.00 | 0.30 |

Table S12: **Random editing with transferred PCD parameters.** Generation accuracy (%) on Mantis. $\Delta$ represents the percentage-point difference (PCD minus random).

| Model | Site | PCD | Random | $\Delta$ |
|---|---|---|---|---|
| LLaVA-OV-7B | K | 64.06 | 63.44 | +0.61 |
| | V | 64.98 | 64.21 | +0.77 |
| Qwen2.5-VL-7B | K | 71.43 | 70.81 | +0.61 |
| | V | 70.97 | 70.81 | +0.15 |
| InternVL3-8B | K | 70.51 | 68.36 | +2.15 |
| | V | 70.97 | 68.66 | +2.30 |

and hence which terms are edited. We therefore check whether the edit depends on this choice. On 64 fixed Mantis samples, we vary $\lambda \in \{0.001, 0.01, 0.1\}$ in Eq. 6 while keeping direction construction, edited layers, $\alpha$, and $\beta$ fixed. For the geometric comparison, each layer's cached Key or Value states and earlier-image directions are held fixed and only $\lambda$ changes; accuracy comes from separate end-to-end decodes. Accuracy is on this 64-sample subset and is not comparable to Tab. 3.

Across a 100-fold range of $\lambda$, fewer than 2.1% of coefficients change sign, the edit vectors remain nearly parallel (cosine $\geq 0.997$), and accuracy changes by at most one sample (Tab. S10). Within this range and setting, the perturbation is thus insensitive to $\lambda$; broader validation is left for future work.

## S9 Random Editing Details

Tab. S11 lists the selected parameters for the independent comparison discussed in §4.3.1. Each random candidate configuration matches the per-token edit norm of the PCD forward pass at that candidate's $\alpha$. The independent comparison evaluates separately selected configurations; the transfer comparison uses the identical $\alpha$ and $\beta$ values for PCD and its random counterpart.

Tab. S12 provides the parameter-transfer results. Under this protocol, PCD consistently maintains higher accuracy across all six comparisons, indicating that the coefficient-selected edit provides a more useful contrastive reference than magnitude-matched random editing in these settings.

## S10 Evaluation Protocol and Benchmark Coverage

**Hyperparameter selection and generation.** Following the validation protocol of DelimScaling (Lee et al., 2026), for K-PCD and V-PCD and each model–benchmark pair, we select

$\alpha \in \{0.25, 0.5, 0.75, 1.0\}$ and $\beta \in \{0.02, 0.05, 0.1, 0.3\}$ on a fixed, randomly sampled 10% subset with seed 42, then freeze the selected values for full-set evaluation. Ties prefer smaller $\beta$, followed by $\alpha$ closer to one, and then smaller $\alpha$. All methods use greedy generation in our experiments.

**Baseline configuration selection.** The three baselines are described in §4.1. VCD (Leng et al., 2024) retains its diffusion perturbation and adaptive plausibility constraint and selects among 16 configurations ($\alpha_{\text{VCD}} \in \{0.5, 1.0\}$, $\beta_{\text{VCD}} \in \{0.05, 0.1, 0.2, 0.3\}$, noise step $\in \{500, 999\}$) on the same 10% subset, breaking ties toward its published defaults; it uses greedy decoding for this comparison. For DelimScaling (Lee et al., 2026), model-level layers are fixed across benchmarks, and scaling strengths are selected on the corresponding selection subset. FOCUS (Park et al., 2025) uses the configurations published in its Table C1 where available; for the remaining settings, it selects among 45 configurations ($\alpha_{\text{FOCUS}} \in \{-0.8, -0.5, -0.3, -0.1, 0, 0.1, 0.3, 0.5, 0.8\}$, noise scale $\lambda_{\text{FOCUS}} \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$) on the same 10% subset.

**Benchmark coverage.** We evaluate on four complementary multi-image benchmarks. *Mantis* (Jiang et al., 2024) examines skills such as co-reference, comparison, and reasoning across images. *MIRB* (Zhao et al., 2024) covers perception, visual knowledge, reasoning, and multi-hop reasoning. *MuirBench* (Wang et al., 2025) evaluates diverse multi-image tasks and relations, including minimally modified unanswerable counterparts that test whether models can identify insufficient evidence. *QBench2* (Zhang et al., 2024b) focuses on low-level visual perception and quality comparisons between image pairs.

**Paired significance tests.** To test whether the improvements over the strongest baseline are statistically reliable, we compare K-PCD and V-PCD with FOCUS (Park et al., 2025) using same-harness per-sample logs. For each model–benchmark setting, we remove the fixed 10% selection subset and recompute both scores on the remaining held-out samples with the benchmark's official aggregation. Samples are paired by question identifier. Each setting is weighted equally in the suite-level mean. Confidence intervals are computed from 20,000 paired bootstrap resamples, and two-sided $p$-values from 100,000 paired randomization permutations (seed 42). We apply Holm correction across the two comparisons.

Table S13: **Paired significance tests against FOCUS on held-out samples.** Accuracy (%) is the equal-weight mean over all 28 model–benchmark settings. $\Delta$ is the paired difference (ours minus FOCUS) in %.

| Comparison | Settings | FOCUS | Ours | $\Delta$ | 95% CI | $p$ (Holm) |
|---|---|---|---|---|---|---|
| K-PCD vs. FOCUS | 28 | 50.21 | 51.00 | $+0.80$ | $[+0.28, +1.33]$ | 0.0028 |
| V-PCD vs. FOCUS | 28 | 50.21 | 51.38 | $+1.17$ | $[+0.67, +1.69]$ | $< 0.0002$ |

Both methods improve significantly over FOCUS (see Tab. S13).

**Vanilla comparison.** We also compare K-PCD and V-PCD with vanilla greedy decoding under the same evaluation protocol on Mantis, MIRB, and MuirBench (see Tab. S14). To examine whether the gains persist beyond the 0.5B–8B range of the main evaluation, we additionally include InternVL3-14B, the largest backbone in this study; InternVL3-1B and QBench2 are not included in this comparison, and the K-PCD and V-PCD entries for the remaining six models are taken from Tab. 3. Both methods improve over vanilla decoding in 20 of the 21 model–benchmark settings, with average gains of 4.40 points for K-PCD and 4.36 points for V-PCD. The gains carry over to InternVL3-14B on all three benchmarks, reaching 2.77 and 4.15 points on Mantis for K-PCD and V-PCD, respectively, suggesting that the benefit of representation-level perturbation is not limited to small backbones.

## S11 Social Impact

**Positive impact.** This work clarifies how cross-image interaction shapes the visual Keys and Values used for generation, supporting more interpretable multi-image systems. As a training-free method with a fixed two-branch design, it can be added to existing MLLMs without retraining and may benefit comparison-heavy applications such as change tracking in medical or remote-sensing images and document analysis.

Table S14: **Vanilla greedy decoding results.**

| Dataset | Method | Qwen2.5-VL | | InternVL3 | | | LLaVA-OV | |
|---|---|---|---|---|---|---|---|---|
| | | 3B | 7B | 2B | 8B | 14B | 0.5B | 7B |
| Mantis | Vanilla | 55.30 | 70.05 | 49.77 | 64.98 | 72.81 | 36.41 | 57.14 |
| | K-PCD | 58.53 | **71.43** | **56.22** | 70.51 | 75.58 | **47.00** | 64.06 |
| | V-PCD | **60.83** | 70.97 | 53.92 | **70.97** | **76.96** | 43.32 | **64.98** |
| MIRB | Vanilla | 49.23 | 59.13 | 40.87 | 47.16 | 53.35 | 28.48 | 47.99 |
| | K-PCD | **57.28** | **63.05** | 44.79 | 52.53 | 56.14 | **31.99** | 48.09 |
| | V-PCD | 56.66 | 62.54 | **45.72** | **53.04** | **56.45** | 31.79 | **49.43** |
| MuirBench | Vanilla | 30.38 | 43.08 | **28.04** | 31.62 | 38.88 | 21.42 | 28.73 |
| | K-PCD | 37.38 | **46.42** | 27.46 | 37.00 | **41.50** | **25.12** | 35.08 |
| | V-PCD | **37.58** | 46.08 | 27.31 | **38.69** | 39.58 | 24.46 | **35.12** |

**Potential negative impact.** Stronger cross-image reasoning carries dual-use risks, such as re-identifying individuals across photographs or producing misleading visual comparisons. Since this work introduces no new data or model weights and only perturbs internal representations of public models at inference time, we do not expect it to meaningfully shift their risk profile. Contrastive decoding also does not correct errors or biases in the base model (§S13).

## S12 Asset Licenses and Usage Terms

We use only publicly released models and benchmarks, subject to their respective licenses and usage terms. Qwen2.5-VL-7B (Bai et al., 2025) and LLaVA-OneVision-0.5B/7B (Li et al., 2024a) are released under Apache-2.0. Qwen2.5-VL-3B (Bai et al., 2025) is released under the Qwen Research License Agreement, which permits use for non-commercial research and evaluation purposes; our use falls within these terms. InternVL3-1B/2B/8B (Zhu et al., 2025) are released under the MIT license, and their Qwen2.5 language components are released under Apache-2.0. The Mantis (Jiang et al., 2024) codebase and evaluation resources are released under Apache-2.0. MuirBench (Wang et al., 2025) is released under CC BY 4.0, with instances derived from existing datasets following the licenses of their original sources (*e.g.*, CC BY-NC 4.0 for SEED-Bench- and GeneCIS-derived instances). MIRB (Zhao et al., 2024) and QBench2 (Zhang et al., 2024b) are publicly released by their authors for benchmarking the multi-image and low-level visual abilities of MLLMs, and we use them solely for academic evaluation in accordance with the terms stated in their official releases.

## S13 Limitations and Future Work

While K-PCD and V-PCD improve accuracy across diverse models and benchmarks, our study has several limitations. First, like other contrastive decoding methods, PCD requires an additional decoding branch to obtain the perturbed reference distribution. The number of branches is fixed at two regardless of image count, whereas image-wise aggregation methods such as FOCUS (Park et al., 2025) require a number of branches that grows with image count. Actual runtime and memory still depend on input length and the per-layer perturbation; in a small-scale measurement on LLaVA-OV-7B, PCD roughly doubles TTFT and total decoding time relative to vanilla decoding with about 0.12 GiB additional peak memory, compared with $3.9\times$ TTFT for FOCUS (§S8). Second, the perturbation and contrast strengths ($\alpha$ and $\beta$) are selected on a selection subset and held fixed during evaluation. Inspired by adaptive allocation in AdaLoRA (Zhang et al., 2023), future work could set them per instance based on the complexity of cross-image interactions in the prompt. Finally, our analysis centers on a discrete number of images, but the geometric dynamics we characterize offer a first step toward understanding temporal interactions in video (Lin et al., 2024; Cheng et al., 2024). Extending negative-coefficient editing to video-language tasks, where inter-frame interactions are dense and highly redundant (Wang et al., 2024b), is a promising direction for training-free hallucination mitigation in video question answering (Zhang et al., 2024a; Li et al., 2025).