WHEN MODALITY GOES MISSING: THE SPURIOUS ALIGNMENT TRAP IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) are generally assumed to degrade predictably when visual input is missing or corrupted, yet existing evaluations focus mainly on downstream performance drops, leaving the internal geometric degradation of representation spaces largely unexplored. In this work, we conduct a subspace geometric analysis of MLLM inference under visually degraded inputs. By measuring the effective rank and principal angles between vision and text subspaces across Transformer layers, we reveal a phenomenon termed the **Spurious Alignment Trap**: when visual input degrades to a blank image, the vision subspace loses dimensionality sharply, while conventional angle-based alignment metrics may counterintuitively indicate improved cross-modal alignment. To distinguish genuine alignment from this illusion, we introduce Gaussian noise images as a diagnostic control. This contrast suggests that the angles under blank conditions are artifacts of dimensional collapse. More importantly, further analysis reveals that under collapse, such metrics decouple from cross-modal semantic alignment and instead reflect the model's modality coupling structure and the landing direction of collapse. Building on these observations, we propose the **Rank–Angle Diagnostic (RAD)** to detect whether a sample's visual input has degraded during inference. Experiments on LLaVA-1.5 and Qwen2.5-VL across VQA and Image Captioning show that blank inputs consistently induce vision-subspace collapse and suppress the expansion of effective rank in deeper layers; across three types of degradation (15%–50% area masking, Gaussian blur, and complete degradation), RAD reliably detects complete degradation, while its sensitivity to partial degradation varies across architectures. Experiments on SmolVLM further validate the coupling–direction relationship on a third model. Our work reveals when and why angular alignment metrics become unreliable, while RAD-based detection requires neither labels nor downstream evaluation. Code will be released.

1 INTRODUCTION

Significant progress has been made in adapting Multimodal Large Language Models (MLLMs) to missing modalities, with existing strategies achieving notable success in maintaining task accuracy and robustness (Zhao et al., 2026; Dai et al., 2026; Pipoli et al., 2025; Zhang et al., 2025; Ke et al., 2025; Reza et al., 2025; Dai et al., 2025; Lang et al., 2025a; Hu et al., 2024; Lee et al., 2023). However, these advances have been evaluated almost exclusively through the lens of downstream task metrics. The internal geometric consequences of modality removal (e.g., how representations collapse or shift) remain largely unexplored, creating a gap between what we can achieve behaviorally and what we understand structurally. This naturally raises two questions: *When modality data is missing or corrupted, what really happens inside the representation space of MLLMs? And, most critically, can conventional angle-based alignment metrics still be trusted?* These questions are particularly critical for the visual modality, as vision serves as the primary channel through which MLLMs perceive the physical world, and its failure directly impacts the model's ability to understand and reason about scenes.

To answer the above questions, we examine the internal representational dynamics of MLLMs under visually degraded inputs, from a subspace geometry perspective. Specifically, we introduce two

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

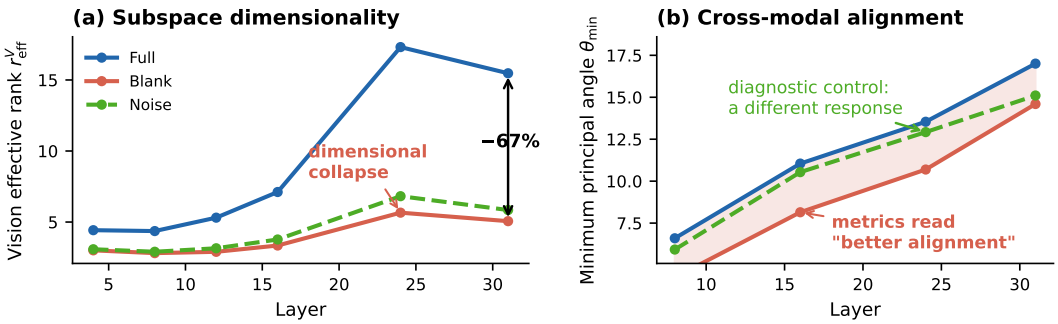


Figure 1: The Spurious Alignment Trap. (a) Vision-subspace effective rank across layers: blank inputs induce dimensional collapse. (b) Minimum principal angle: the metric reads *better* alignment for collapsed inputs, while the noise control, equally devoid of semantics, elicits a markedly different angular response.

geometric diagnostic tools: (1) *effective rank* (Wei et al., 2024; Roy & Vetterli, 2007), to quantify the effective dimensionality of each modality's representation subspace; and (2) *principal angles* (Andrew et al., 2013), to measure the alignment between vision and text subspaces and characterize their cross-modal geometric relationship. Both are computed across Transformer layers.

Under this analytical framework, we discover an intriguing yet counterintuitive phenomenon, as shown in Figure 1: when visual input degrades to a blank image, the dimensionality of the vision subspace collapses sharply, yet conventional angle-based alignment metrics may misleadingly suggest "better alignment." We term this the **Spurious Alignment Trap**. To distinguish genuine alignment from this illusion, we introduce Gaussian noise images as a key diagnostic control: noise images likewise carry no semantic content, yet elicit a markedly different angular response, indicating that the alignment signals observed under blank conditions arise from dimensional collapse rather than from the absence of semantics. More fundamentally, we find that under dimensional collapse, such alignment metrics decouple from cross-modal semantic alignment and instead reflect the model's modality coupling structure and the landing direction of collapse. Matched-rank random-subspace null tests provide strong evidence for this structural interpretation ($p < 10^{-4}$ at every layer). Building on these observations, we propose the **Rank–Angle Diagnostic (RAD)**, a layer-wise metric combining subspace dimensionality with cross-modal correlation, to detect whether a sample's visual input has degraded during inference; unlike inspecting model outputs, RAD remains reliable when MLLMs fall back on language priors to produce plausible answers despite missing visual cues (see Appendix K for a case study). Experiments on LLaVA-1.5 (Liu et al., 2024) and Qwen2.5-VL (Bai et al., 2025) across VQA and Image Captioning tasks substantiate these findings, and show that under a range of degraded inputs (15%–50% area masking, Gaussian blur, and complete degradation), RAD reliably detects complete degradation while its sensitivity to partial degradation varies across architectures. We further validate this coupling–direction relationship on a third model, SmolVLM (Marafioti et al., 2025), where the observed pattern is consistent.

2 Related Works

Multimodal Learning with Missing Modalities. Significant progress has been made in adapting MLLMs to missing modalities, with existing methods broadly falling into two categories. (1) Prompt-based methods introduce a small number of learnable prompts to enable models to adapt to incomplete inputs. Representative works include missing-aware prompting (Lee et al., 2023), Deep Correlated Prompting (Hu et al., 2024), Synergistic Prompting (Zhang et al., 2025), and Memory-Driven Prompting (Zhao et al., 2025). (2) Generation and imputation methods attempt to explicitly reconstruct missing information through retrieval augmentation or cross-modal interaction (Dai et al., 2026; Lang et al., 2025a;b). For instance, MissRAG (Pipoli et al., 2025) constructs a memory bank of complete modalities and leverages retrieval mechanisms to generate missing modalities. However, these methods are primarily evaluated from the perspective of task performance (e.g., accuracy, F1 score), and lack analysis of the geometric changes in MLLMs' internal representations under modality-deficient conditions.

Language Priors under Missing Vision. A complementary line of work has diagnosed this problem behaviorally: prior studies in the VQA and MLLM evaluation literature have shown that models can bypass visual information entirely and answer questions based solely on language priors (Chen et al., 2024; Guan et al., 2024; Agrawal et al., 2018; Goyal et al., 2017). These works treat the model as a black box and diagnose behavioral reliance on prior knowledge. Our work, however, focuses on a different scenario: the visual signal is always present during inference, but the input content is corrupted or incomplete. We examine how this degradation reshapes the internal geometric structure of representations, and whether conventional alignment metrics remain reliable under such conditions.

Geometric Analysis of Multimodal Representations. In contrast to the aforementioned works that evaluate robustness solely through downstream task metrics, another line of research has gradually shifted toward the internal geometric structure of representation spaces to gain a deeper understanding of alignment mechanisms in multimodal models. Existing studies primarily rely on two types of tools: (1) scalar similarity metrics, most notably CLIP and its successors (Zhai et al., 2022; Jia et al., 2021; Radford et al., 2021), which encode vision and text into a shared embedding space and measure cross-modal alignment via cosine similarity; and (2) subspace-geometric metrics (Zeng et al., 2026; Kornblith et al., 2019; Raghu et al., 2017; Andrew et al., 2013), which assess alignment quality by analyzing the geometric relationships between subspaces (e.g., through principal angles). However, these metrics rely on an implicit assumption: both the vision and text subspaces maintain sufficient dimensional structure for similarity or angle computations to be semantically meaningful.

Most closely related to our work is Diff-eRank (Wei et al., 2024), which extends effective rank to multimodal alignment by comparing rank values across modalities. While insightful, it treats alignment as a scalar comparison between rank values, without considering the geometric relationship between subspaces. Our work differs fundamentally in two key aspects: (1) Diff-eRank examines rank evolution during training with full modalities, whereas we analyze geometric degradation during inference under visually degraded inputs; and (2) Diff-eRank assesses alignment through scalar rank comparison alone, whereas we jointly consider subspace dimensionality and geometric angles to reveal how conventional metrics mislead when vision is degraded. These distinctions render the two works complementary rather than overlapping.

3 METHODOLOGY

3.1 EFFECTIVE RANK

Definition 1 (Effective Rank). Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with singular value decomposition $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, where $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ and $r = \min(m, n)$, the effective rank is defined as the exponential of the Shannon entropy of the singular value distribution:

$$\mathrm{eRank}(\mathbf{A}) = \exp\left(-\sum_{i=1}^{r} p_i \log p_i\right), \quad p_i = \frac{\sigma_i}{\sum_{j=1}^{r}\sigma_j}. \tag{1}$$

The Effective Rank (eRank) is a metric that measures the information richness of a representation space. It quantifies the effective dimensionality of representations by jointly considering the number of directions and how energy is distributed across those directions.

Definition 2 (Covariance Matrix of Representations). Let $\mathbf{H}^{(\ell)} \in \mathbb{R}^{n\times d}$ denote the hidden-state matrix at layer ℓ of the language-model backbone, where n is the sequence length and d is the hidden dimension. Based on the positions of image tokens, the sequence is partitioned into visual and textual token sets, whose hidden states form $\mathbf{V}^{(\ell)} \in \mathbb{R}^{n_v\times d}$ and $\mathbf{T}^{(\ell)} \in \mathbb{R}^{n_t\times d}$ with $n_v+n_t = n$. For any representation matrix $\mathbf{X} \in \mathbb{R}^{m\times d}$ with $m \geq 2$, we define its uncentered covariance as:

$$\Sigma_{\mathbf{X}} = \frac{1}{m-1}\mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{d\times d}, \tag{2}$$

whose eigenvalues we denote by $\lambda_1(\mathbf{X}) \geq \cdots \geq \lambda_d(\mathbf{X}) \geq 0$. Note that $\Sigma_{\mathbf{X}}$ is symmetric positive semi-definite, so its singular values coincide with its eigenvalues, $\lambda_i = \sigma_i^2/(m-1)$, i.e., the squared singular values of $\mathbf{X}$ up to a constant factor.

Implementation. In practice, we apply eRank to the covariance matrix rather than to the hidden-state matrix itself, i.e., $\mathrm{eRank}(\Sigma_{\mathbf{X}})$, with a spectral threshold of 10^{-4} filtering numerically near-zero singular values. The rationale and derivations are provided in Appendix A. Finally, we obtain three levels of effective rank: global $r_{\text{eff}}^{(g)}(\ell)$, vision $r_{\text{eff}}^{V}(\ell)$, and text $r_{\text{eff}}^{T}(\ell)$:

$$r_{\text{eff}}^{(g)}(\ell) = \mathrm{eRank}(\Sigma_{\mathbf{H}^{(\ell)}}), \quad r_{\text{eff}}^{V}(\ell) = \mathrm{eRank}(\Sigma_{\mathbf{V}^{(\ell)}}), \quad r_{\text{eff}}^{T}(\ell) = \mathrm{eRank}(\Sigma_{\mathbf{T}^{(\ell)}}). \tag{3}$$

3.2 Principal Angle

Definition 3 (Modality Subspace). At layer ℓ, the visual and text modality representations span subspaces:

$$\mathcal{S}_V^{(\ell)} = \mathrm{span}(\mathbf{V}^{(\ell)}) \subseteq \mathbb{R}^d, \qquad \mathcal{S}_T^{(\ell)} = \mathrm{span}(\mathbf{T}^{(\ell)}) \subseteq \mathbb{R}^d, \tag{4}$$

with dimensions $r_v = \dim \mathcal{S}_V^{(\ell)} \leq \min(n_v, d)$ and $r_t = \dim \mathcal{S}_T^{(\ell)} \leq \min(n_t, d)$. Let $\mathbf{Q}_v^{(\ell)} \in \mathbb{R}^{d \times r_v}$ and $\mathbf{Q}_t^{(\ell)} \in \mathbb{R}^{d \times r_t}$ denote orthonormal bases of the two subspaces.

Definition 4 (Principal Angles). The principal angles $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$ (with $k = \min(r_v, r_t)$) between $\mathcal{S}_V^{(\ell)}$ and $\mathcal{S}_T^{(\ell)}$ are defined by the singular values of the matrix $\mathbf{Q}_v^{(\ell)\top}\mathbf{Q}_t^{(\ell)}$:

$$\theta_i = \arccos\big(\sigma_i\big(\mathbf{Q}_v^{(\ell)\top}\mathbf{Q}_t^{(\ell)}\big)\big), \qquad i = 1, \ldots, k. \tag{5}$$

where $\sigma_i(\cdot)$ denotes the i-th singular value in descending order.

This work focuses on the *minimum principal angle* $\theta_{\min} = \theta_1$, the angle between the closest pair of directions in the two subspaces. In the subspace-geometric view of cross-modal alignment (Raghu et al., 2017; Andrew et al., 2013), a *small* $\theta_{\min}$ is conventionally interpreted as evidence that the two modalities share a highly correlated direction and are thus geometrically aligned, whereas a *large* $\theta_{\min}$ (approaching $90°$) indicates effective decoupling. However, this interpretation presumes that both subspaces retain sufficient dimensional structure; we revisit this assumption in Section 4.3.

3.3 The Rank–Angle Diagnostic

Definition 5 (Modality Coupling Structure). Let v_i^V denote the i-th principal direction of the vision representations at a layer, and let P_T be the orthogonal projector onto the text subspace $\mathcal{S}_T$. The modality coupling structure of a model is characterized by the redundancy of the cross-modal shared information, i.e., the concentration of $\{\|P_T v_i^V\|\}_{i=1,2,\ldots}$: under full-modal inputs, the structure is *strongly coupled* if these values concentrate on one or a few components (low redundancy), and *weakly coupled* if they are distributed across multiple components (high redundancy). Formally, the redundancy can be quantified by the participation ratio (PR) of the normalized spectrum:

$$\mathrm{PR} = \frac{1}{\sum_i p_i^2}, \qquad p_i = \frac{\|P_T v_i^V\|^2}{\sum_j \|P_T v_j^V\|^2}, \tag{6}$$

where p_i is the energy fraction of the i-th component. PR is an effective-number measure (Chun et al., 2026): when the energy is concentrated on a single component, $\mathrm{PR} \to 1$; when it is uniformly distributed across r components, $\mathrm{PR} \to r$. A larger PR therefore indicates that the shared information is spread over more components, i.e., higher redundancy.

Definition 6 (Rank–Angle Diagnostic). At layer ℓ, let $r = r_{\text{eff}}^{V}(\ell)$ and $c = \cos\theta_{\min}(\ell)$. We define

$$\mathrm{RAD}(\ell) = \log\big(r\big) \cdot c. \tag{7}$$

The logarithmic rank term measures the structural richness of the vision subspace, and the angular term measures cross-modal correlation. As $r \to 1$, RAD vanishes regardless of how small the angle appears, which is precisely the regime in which the Spurious Alignment Trap operates. Its key property is *Pareto consistency*: within the same model, layer, and task, if $r_1 \geq r_2$ and $c_1 \geq c_2$, then $\mathrm{RAD}_1 \geq \mathrm{RAD}_2$. Computed at each layer, RAD forms a *diagnostic trajectory* $\{\mathrm{RAD}(\ell)\}_\ell$. At deployment time, full-modal inputs establish the reference trajectory; when an incoming sample's RAD trajectory drops below the reference and deviates beyond the variability of the reference itself, the sample is flagged as degraded; the depth of the drop indicates the degradation severity. This process requires neither labels nor downstream evaluation. We validate RAD in Sec. 4.4.

4 EXPERIMENTS

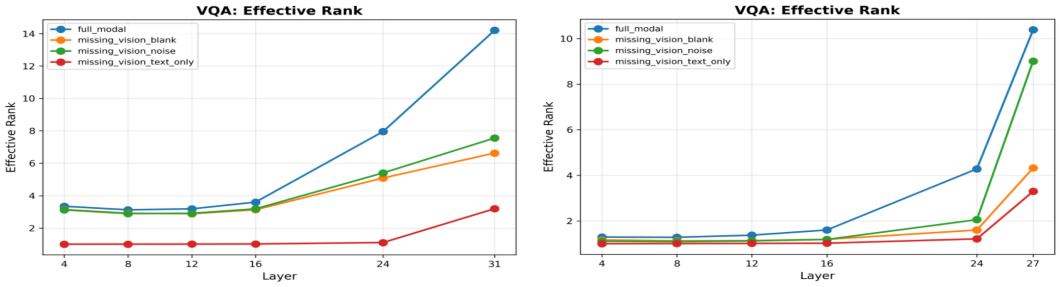
4.1 EXPERIMENTAL SETUP

Models and datasets. We conduct experiments on two mainstream open-source MLLMs: LLaVA-1.5-7B (Liu et al., 2024) and Qwen2.5-VL-7B (Bai et al., 2025). LLaVA-1.5 uses a separate vision encoder and an MLP projector to map visual features into the language model, whereas Qwen2.5-VL adopts a native multimodal design with dynamic-resolution input, forming a cross-architecture comparison. We evaluate on two tasks: visual question answering (VizWiz-VQA (Gurari et al., 2018)) and image captioning (Flickr30K (Plummer et al., 2017; Young et al., 2014)). The former targets real-world scenarios, while the latter examines open-ended generation. For computational efficiency, we randomly sample 500 examples from each test set with a fixed random seed to ensure reproducibility. Implementation details are provided in Appendix C.

Input conditions. For each sample, we perform inference under four input conditions: (1) *full_modal*, where the original image is provided; (2) *blank*, where the image is replaced by a constant image with a fixed gray value (128, 128, 128), so that the visual token stream still passes through the vision encoder; (3) *noise*, where the image is replaced by pixel-level Gaussian noise, which carries no semantic information yet yields a higher-dimensional vision subspace than blank, serving as a diagnostic control; and (4) *text_only*, where no image is provided and only the text input is retained, serving as a reference for the complete absence of visual input.

Layers and metrics. We extract hidden states from Transformer layers spanning shallow to deep: (1) layers $\{8, 16, 24, 31\}$ for LLaVA-1.5 (32 layers in total), and (2) layers $\{8, 16, 24, 27\}$ for Qwen2.5-VL (28 layers in total). All computations are performed layer-wise. For effective rank, we apply a spectral threshold of 10^{-4} to filter numerically near-zero singular values, so that floating-point error does not affect the rank estimate; the collapse conclusions are invariant to this choice across 10^{-2}–10^{-6} (Appendix A). Complete definitions of all metrics are given in Sec. 3, while additional analysis quantities introduced in this section are defined where they are used.

4.2 DIMENSIONAL COLLAPSE IS CONSISTENT ACROSS ARCHITECTURES



(a) LLaVA-1.5 (b) Qwen2.5-VL

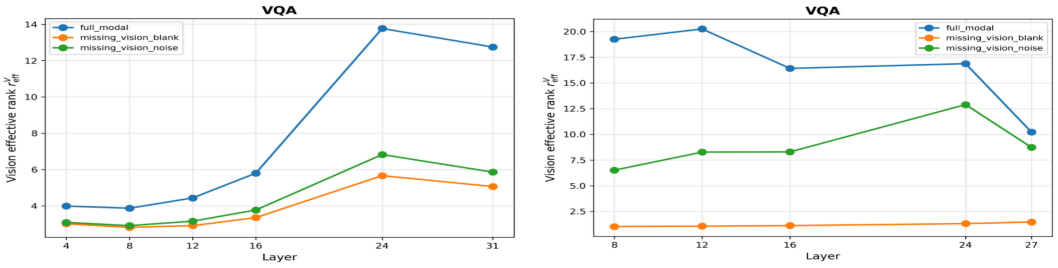
Figure 2: Effective rank trajectories across layers on VQA under four input conditions (full, blank, noise, text_only). Results on image captioning are provided in Appendix D.

Figure 2 shows the layer-wise trajectories of the effective rank of the overall representation $r_{\text{eff}}^{(g)}(\ell)$ under four input conditions for LLaVA-1.5 and Qwen2.5-VL. Under the full-modal condition, the effective rank expands progressively with depth. Under blank images, the effective rank is suppressed to a low level at every layer, with its depth-wise growth markedly limited; noise images occupy an intermediate level. Under text_only inputs, no visual tokens are present and the effective rank is determined entirely by text tokens, serving as a reference for complete visual absence.

The two architectures differ in absolute values but agree in pattern: missing vision progressively suppresses the expansion of effective rank in deeper layers, reflecting a degradation in the model's capacity to construct complex, task-relevant representations. This suppression is primarily driven by the dimensional collapse of the vision subspace (as shown in Figure 3), and provides a geometric perspective for understanding the downstream performance drop under missing modalities.

Table 1: Per-sample classification accuracy (5-fold cross-validation; chance: 33.3%).

Layer	LLaVA-1.5 / VQA	LLaVA-1.5 / Caption	Qwen2.5-VL / VQA	Qwen2.5-VL / Caption
L8	85.3±1.5%	89.3±0.5%	92.7± 1.2%	98.4± 0.6%
L16	92.7±1.2%	97.2±0.6%	94.5± 0.9%	95.9±1.0%
L24	93.9±1.4%	96.9±0.9%	92.5± 1.2%	67.7±1.2%
Last	93.8±1.6%	95.7±0.3%	88.8± 1.1%	67.2±2.1%



(a) LLaVA-1.5 (b) Qwen2.5-VL

Figure 3: Vision-subspace effective rank across layers on VQA under full, blank, and noise conditions. Blank images collapse the rank to a low level; noise occupies an intermediate level.

Per-sample separability. Each trajectory above is based on the mean over 500 samples. To rule out the possibility that the separation is merely an artifact of averaging (i.e., that samples from different conditions overlap substantially and the separation emerges after averaging), we test whether the above pattern still holds at the per-sample level: for each layer, we take each sample's vision-subspace effective rank as its feature, use linear discriminant analysis (LDA) to classify it into one of the three conditions, and evaluate accuracy with 5-fold cross-validation, where all conditions of the same source image are kept within the same fold (1500 data points per layer).

Table 1 reports the classification accuracy at each layer. On LLaVA-1.5, accuracy ranges from 85.3% to 97.2% across all layers for both tasks. On Qwen2.5-VL, accuracy ranges from 88.8% to 94.5% on VQA and remains high in the shallow and middle layers of Caption (98.4% and 95.9%). The only notable exception is Qwen2.5-VL on Caption at deeper layers (L24 and the last layer), where accuracy drops to around 67%, indicating partial overlap between the full and noise classes under this condition (see Figure 3 (b), where the full and noise trajectories gradually approach each other in deeper layers). Nevertheless, all accuracies are markedly above the 33.3% chance baseline, and the blank condition remains separable from both throughout, showing that the separation in Figure 3 is not an artifact of averaging.

4.3 Angular readouts decouple from alignment under collapse

Having confirmed that degraded inputs induce dimensional collapse, we now examine whether principal angles remain trustworthy after collapse. Figure 4 shows the layer-wise trajectories of the minimum principal angle across two architectures and two tasks, yielding an unexpected result: they give different answers. Within each model, the direction of blank relative to full is exactly opposite. In LLaVA-1.5, the blank-condition trajectory lies below the full-modal one at almost every layer. Under the interpretation of Sec. 3.2, this means cross-modal alignment appears *better* under blank images, which is precisely the Spurious Alignment Trap. In Qwen2.5-VL, the situation is reversed: the blank trajectory overtakes the full-modal one, which by the same interpretation means that alignment appears *worse* under blank images. This direction is more consistent with intuition.

The key observation is that the blank-to-full direction is reversed *within* both architectures while the input is equally devoid of semantics. If angular readouts are interpreted as evidence of semantic alignment (Sec. 3.2), a semantics-free input cannot point to "better alignment" in one case and "worse alignment" in the other. We therefore argue that under collapse, angular readouts decouple from cross-modal semantic alignment. What they reflect instead is analyzed immediately below.

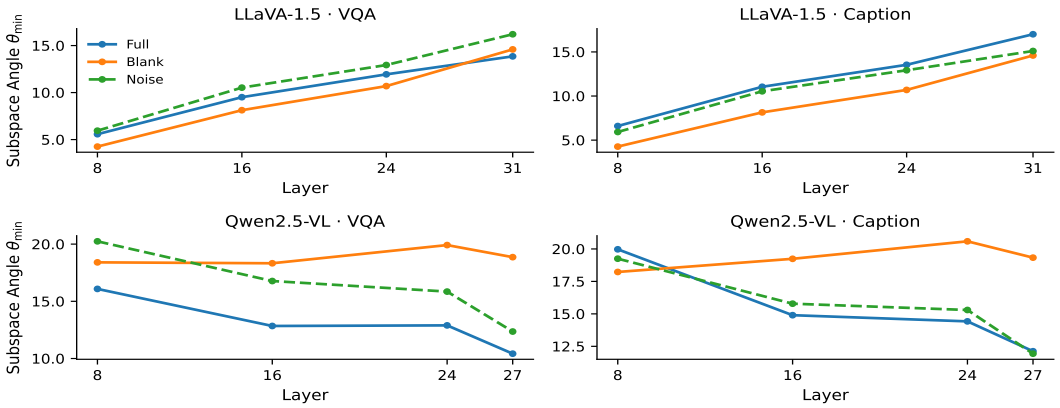
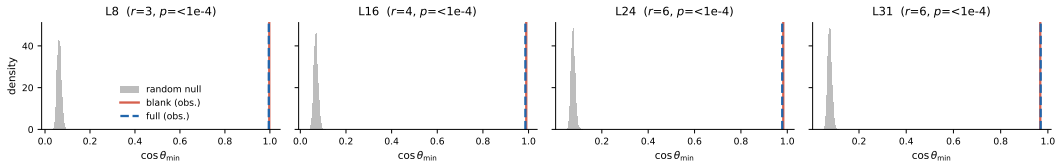
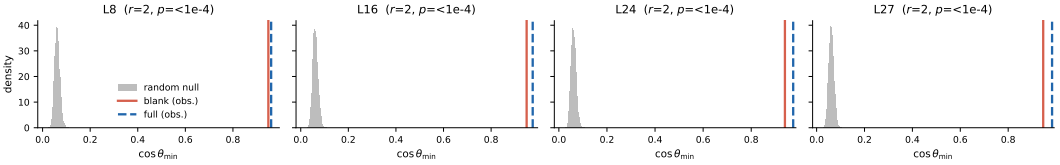


Figure 4: Minimum principal angle $\theta_{\min}$ across layers under full, blank, and noise inputs, for LLaVA-1.5 (top) and Qwen2.5-VL (bottom) on VQA and image captioning. Blank inputs yield *smaller* angles in LLaVA-1.5 but *larger* angles in Qwen2.5-VL, consistently across both tasks.



(a) LLaVA-1.5: matched-rank random-subspace null (K=10000)



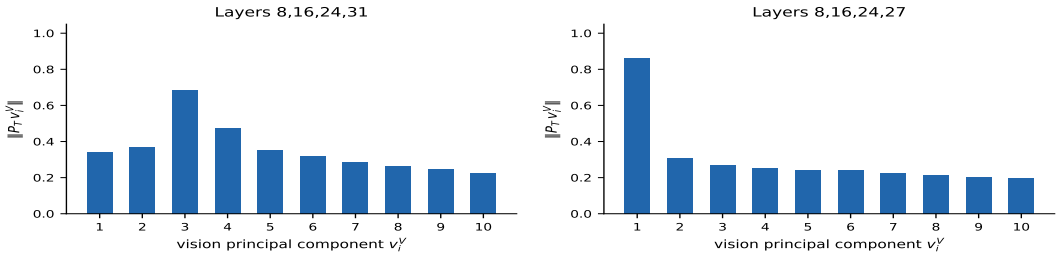
(b) Qwen2.5-VL: matched-rank random-subspace null (K=10000)

Figure 5: Matched-rank random-subspace null distributions for LLaVA-1.5 (top) and Qwen2.5-VL (bottom) at selected layers. Gray histograms show the null distribution of $\cos\theta_{\min}$ between a random r-dimensional subspace and the shared text subspace; red and blue lines mark the observed values for blank and full conditions. The observed blank values lie far in the right tail at every layer.

Collapse directions are structured, not coincidental. The anomalous angular readouts, whether shrinking or expanding, face a fundamental alternative explanation: *could the proximity or distance between the collapsed vision subspace and the text subspace be a mere dimensional coincidence?* If so, the observed angular readouts would carry no information about the model itself.

We test this hypothesis with a matched-rank random-subspace null distribution. The null distribution provides a discriminative test: if the blank-condition angles lie significantly outside the null distribution, the collapse direction follows a systematic pattern rather than a random one; conversely, if they fall within the null, they are indistinguishable from dimensional coincidence and carry no model-specific information. To construct this null distribution, for the blank condition at each layer of each model, we set $r = \lceil r_{\text{eff}}^{V} \rceil$ (matched to the observed collapse), take the shared text subspace estimated from 500 pooled samples as the reference. We then draw 10^4 random r-dimensional subspaces from $\mathbb{R}^d$, computing the cosine of the minimum principal angle between each and the text subspace. The full computational procedure is provided in Appendix E.

The result is unambiguous: as shown in Figure 5, the null distribution concentrates around ≈ 0.07, while the observed blank values of both architectures (0.93–0.97) lie far in the right tail, with $p < 10^{-4}$ at every layer. The collapse direction is therefore not random, but follows a systematic pattern. What determines its orientation, toward or away from the text subspace, is the modality coupling structure and the landing direction of collapse, as analyzed below.



(a) LLaVA-1.5, PR ≈ 6.0 (b) Qwen2.5-VL, PR ≈ 2.7

Figure 6: Distribution of $\|P_T v_i^V\|$ across vision principal components, averaged over layers.

Coupling structure and collapse landing govern the response. We compute $\|P_T v_i^V\|$ per sample and layer on VQA, and average across samples and layers for each principal component index i, obtaining a layer-aggregated spectrum $\{\overline{\|P_T v_i^V\|}\}_{i=1,2,\ldots}$. Within each layer, principal components are ordered by singular value; the i-th bar represents the average closeness of the i-th principal component of each layer to the text subspace (the principal directions of the same index differ across layers in general). Figure 6 shows that the two architectures differ markedly: LLaVA-1.5 is weakly coupled, with $\|P_T v_i^V\|$ spread across multiple principal components (PR ≈ 6.0), whereas Qwen2.5-VL is strongly coupled, with $\|P_T v_i^V\|$ concentrated on the leading component (PR ≈ 2.7).

This difference in coupling structure, together with the collapse direction, determines the direction in which the minimum principal angle changes. (1) **Under weak coupling**, cross-modal shared information is distributed across multiple principal components; even if collapse affects part of them, the collapsed subspace is more likely to contain a direction close to the text subspace, so the minimum principal angle is pulled down, manifesting as spurious shrinking. (2) **Under strong coupling**, shared information is concentrated on a single leading component; once the directions of the collapsed subspace deviate from it, the minimum principal angle tends to increase. **In other words**, the collapse direction itself is not guided by this redundancy. what redundancy determines is the probability that the collapsed subspace contains a direction close to the text subspace. When shared information is spread across multiple components (weak coupling), the collapse outcome is more likely to contain a direction close to the text subspace, and the angular readout tends toward spurious shrinking; when it is concentrated on a single leading component (strong coupling), a collapse that deviates from this component leaves a subspace with little shared information, and the angular readout tends toward reverse increase.

This explanation has a precise algebraic basis. Since the principal directions v_i^V are orthonormal, the minimum principal angle satisfies

$$\max_i \|P_T v_i^V\| \;\le\; \cos\theta_{\min} \;\le\; \sqrt{r}\,\max_i \|P_T v_i^V\|, \tag{8}$$

where r is the number of principal directions. The lower bound follows directly from the definition of $\cos\theta_{\min}$. The upper bound shows that when multiple components carry moderate projections (as in weak coupling), their contributions can add coherently, allowing $\cos\theta_{\min}$ to exceed any individual $\|P_T v_i^V\|$ by up to a factor of $\sqrt{r}$, thus tending toward smaller angles. This mechanism exists only when the projection mass is distributed across multiple components; when it concentrates on one or a few components (as in strong coupling), the optimal direction is that component itself, and $\cos\theta_{\min}$ reduces to $\max_i \|P_T v_i^V\|$. A full derivation is provided in Appendix B. We additionally analyze cosine similarity under the same settings as principal angles; see Appendix G for details. To verify that the coupling–direction relationship is predictive rather than post hoc, we perform additional validation on a third model, SmolVLM(Marafioti et al., 2025); see Appendix I for details.

4.4 RAD-BASED DETECTION OF VISUAL DEGRADATION

Input conditions for RAD detection. Unlike the analysis above, which uses blank and noise as controlled extremes, RAD detection requires a graded series of degradation levels. We therefore use three partial masks (*mask15*, *mask30*, *mask50*), in which a randomly located black box covers 15%, 30%, and 50% of the image area; a Gaussian-blurred variant (*blur4*, Gaussian blur with $\sigma = 4$ pixels); the blank condition as complete degradation; and *full_modal* as the reference.

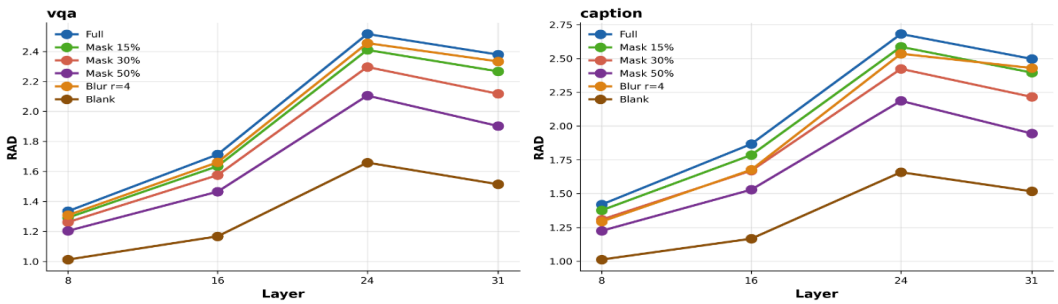


Figure 7: RAD trajectories of LLaVA-1.5 on VQA and image captioning under six input conditions.

RAD trajectories. For each sample, we compute RAD at each layer according to Definition 6, and average over all samples under the same condition to obtain the RAD trajectory. Figure 7 shows the trajectories of LLaVA-1.5 on both tasks: RAD stays highest under full modality and decreases monotonically with the degradation level, from 15% masking through 30% and 50% masking and Gaussian blur, down to the lowest level under complete degradation. This pattern indicates that RAD reflects the degree of degradation and provides the basis for the detection method that follows. For Qwen2.5-VL, the response to partial masking is less sensitive, and we provide a more detailed analysis in Appendix F. We evaluate RAD quantitatively in Appendix H, using AUROC, TPR at a fixed false-positive rate, and comparisons with simple baselines.

Per-sample consistency. The trajectories above are condition means, whereas detection operates on individual samples. We therefore measure the per-sample consistency: for each sample, whether its full-modal RAD exceeds its degraded-condition RAD. Table 2 shows that this consistency increases monotonically with degradation severity, from 57–62% under 15% masking to 100% under complete degradation on VQA, and from 81–93% to 100% on Caption. Masking of 50% and complete degradation are separable from full modality at every layer. The overall lower consistency on VQA is consistent with the heterogeneity of VizWiz images.

Table 2: Per-sample consistency.

Comparison	VQA	Caption
full > mask15	57–62%	81–93%
full > blur4	56–59%	63–100%
full > mask30	70–79%	93–100%
full > mask50	88–94%	99–100%
full > blank	**100%**	**100%**

Inference-time detection. Using the full-modal RAD trajectory as the reference, a new sample is flagged as degraded when its entire RAD trajectory falls below the reference and deviates beyond the reference's variability. At deployment time, we average the RAD trajectories of full-modal samples to obtain the reference trajectory, and record the per-layer standard deviation. Operationally, the trajectory-level condition is evaluated per layer: a sample is flagged at layer ℓ when $\mathrm{RAD}(\ell) < \mu_{\mathrm{full}}(\ell) - k\,\sigma_{\mathrm{full}}(\ell)$, and the depth of the deviation reflects the degradation severity. This method performs better on severe degradation, making it more suitable for severely degraded inputs. The procedure requires neither labeled data nor downstream evaluation: the decision boundary is determined by the variability of the reference itself. A case study is provided in Appendix K.

5 Discussion and Conclusion

In this work, we present a subspace geometric analysis of the internal representations of MLLMs under visual degradation. We find that when the vision subspace collapses, angle-based alignment metrics no longer reflect cross-modal semantic alignment, but instead reflect the model's modality coupling structure and the landing direction of collapse. We term this phenomenon the Spurious Alignment Trap, and support its mechanism through noise controls, null-distribution tests, and component-wise decomposition, as well as an algebraic bound relating the minimum principal angle to the per-component projections. We also propose RAD for detecting visual input degradation at deployment time. Experiments on two representative architectures and a third model across VQA and image captioning validate these findings. Limitations are discussed in Appendix J.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks requiring disclosure: implementing methods (writing parts of the code and reviewing the integrated codebase; all code runs in a local environment and was cross-verified by the authors); assisting in developing the conceptual framework (the core content and ideas were proposed by the authors, with AI tools used for analysis and expansion, e.g., the introduction of the participation ratio); validating the reasonableness of our interpretation of results (AI tools also extended this analysis, but all extensions were rigorously examined and compared by the authors); providing feedback on research methodology or experiments (e.g., the AI suggested introducing multiple input-degradation strategies); assisting in writing proofs (all derivations were rigorously verified by the authors); and assisting with translation. We have not used generative AI tools for the rest of the tasks requiring disclosure, as they are not applicable to this work (e.g., qualitative data analysis).

Additionally, we used generative AI tools for the following tasks with recommended disclosure: creating and editing parts of the code; polishing the language to improve readability; summarizing and analyzing existing literature; searching for relevant information; and identifying relevant literature.

We have reviewed all AI-assisted work. For example, parts of the experimental code generated with AI assistance were run and verified for correctness by the authors in a local environment; paragraphs translated with AI assistance were checked sentence by sentence to ensure consistency with the experimental data and conclusions; and the literature suggested by AI was manually verified to confirm that it exists and is relevant to this work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The complete experimental setup, the computation of all metrics, and the full proofs of our claims are described in detail in the main text and the appendix. To ensure reproducibility of our results, the code will be publicly released.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018. doi: 10.1109/CVPR.2018.00522.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1247–1255, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/andrew13.html.

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, 2025. URL https://arxiv.org/abs/2502.13923.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087. Curran Associates, Inc., 2024. doi: 10.52202/079017-0850.

Chanwoo Chun, Abdulkadir Canatar, SueYeon Chung, and Daniel Lee. Estimating dimensionality of neural representations from finite samples. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, 2026.

Ruiting Dai, Yuqiao Tan, Lisi Mo, Tao He, Ke Qin, and Shuang Liang. Robustpt: Dynamic disentanglement prompt tuning in vision-language models with missing modalities. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, ICMR '25, pp. 164–172. Association for Computing Machinery, 2025. ISBN 9798400718779. doi: 10.1145/3731715.3733423. URL https://doi.org/10.1145/3731715.3733423.

Wei Dai, Haoyu Wang, Honghao Chang, Lijun He, Fan Li, Jian Sun, and Haixia Bi. Enhancing foundation vlm robustness to missing modality: Scalable diffusion for bi-directional feature restoration, 2026. URL https://arxiv.org/abs/2602.03151.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024. doi: 10.1109/CVPR52733.2024.01363.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. In *Advances in Neural Information Processing Systems*, volume 37, pp. 67446–67466. Curran Associates, Inc., 2024. doi: 10.52202/079017-2153.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25864–25873, 2025. doi: 10.1109/CVPR52734.2025.02409.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18035–18043, 2025a. doi: 10.1609/aaai.v39i17.33984.

Jian Lang, Rongpei Hong, Zhangtao Cheng, Ting Zhong, Yong Wang, and Fan Zhou. Redeeming modality information loss: Retrieval-guided conditional generation for severely modality missing learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 1241–1252. Association for Computing Machinery, 2025b. ISBN 9798400714542. doi: 10.1145/3711896.3737101. URL https://doi.org/10.1145/3711896.3737101.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2024. doi: 10.1109/CVPR52733.2024.02484.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models, 2025. URL https://arxiv.org/abs/2504.05299.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. Missrag: Addressing the missing modality challenge in multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3215–3224, 2025. doi: 10.1109/ICCV51701.2025.00308.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and Salman Asif. Robust multimodal learning via cross-modal proxy tokens. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Wtc6wvcYJ0.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610, 2007.

Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Diff-erank: A novel rank-based metric for evaluating large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2: 67–78, 2014.

Zhichen Zeng, Wenxuan Bao, Xiao Lin, Ruizhong Qiu, Tianxin Wei, Xuying Ning, Yuchen Yan, Chen Luo, Monica Xiao Cheng, Jingrui He, and Hanghang Tong. Subspace alignment for vision-language model test-time adaptation, 2026. URL https://arxiv.org/abs/2601.08139.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18102–18112, 2022. doi: 10.1109/CVPR52688.2022.01759.

Zhihui Zhang, Luanyuan Dai, Qika Lin, Yunfeng Diao, Guangyin Jin, Yufei Guo, Jing Zhang, and Xiaoshuai Hao. Synergistic prompting for robust visual recognition with missing modalities. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1881–1890, 2025. doi: 10.1109/ICCV51701.2025.00183.

Shu Zhao, Nilesh Ahuja, Tan Yu, Tianyi Shen, and Vijaykrishnan Narayanan. Mora: Missing modality low-rank adaptation for visual recognition. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yihan Zhao, Wei Xi, Xiao Fu, and Jizhong Zhao. Enhancing multimodal model robustness under missing modalities via memory-driven prompt learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 2458–2466. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/274. URL https://doi.org/10.24963/ijcai.2025/274.

Supplementary Material

CONTENTS

1 Introduction **1**

2 Related Works **2**

3 Methodology **3**

3.1 Effective Rank 3

3.2 Principal Angle 4

3.3 The Rank–Angle Diagnostic 4

4 Experiments **5**

4.1 Experimental Setup 5

4.2 Dimensional collapse is consistent across architectures 5

4.3 Angular readouts decouple from alignment under collapse 6

4.4 RAD-based detection of visual degradation 8

5 Discussion and Conclusion **9**

A The rationale and derivations of the covariance form of eRank **15**

B Derivation of the Principal-Angle Bounds **15**

C Implementation Details **17**

D Effective Rank Trajectories on Image Captioning **17**

E Matched-rank random-subspace null test: computational procedure **17**

F Qwen2.5-VL is more robust to partial masking **18**

G The analysis of cosine similarity under visual degradation **19**

H Quantitative evaluation of RAD **20**

I Validation of the Coupling–Direction Relationship on SmolVLM **22**

J Limitations and Future Work **23**

K Case Study: RAD-Based Detection on Individual Samples **24**

K.1 RAD-based detection on an individual VQA sample 24

K.2 RAD-based detection on an individual Image Captioning sample 24

A THE RATIONALE AND DERIVATIONS OF THE COVARIANCE FORM OF ERANK

In our implementation, we apply eRank to the uncentered covariance matrix rather than to the hidden-state matrix itself, i.e., $\mathrm{eRank}(\mathbf{\Sigma_X})$. There are three reasons.

(1) The eigenvalues of the covariance matrix admit a clear variance interpretation. $\mathbf{\Sigma_X}$ is symmetric positive semi-definite, so its singular values coincide with its eigenvalues:

$$\lambda_i = \frac{\sigma_i^2}{m-1}, \tag{9}$$

where σ_i are the singular values of $\mathbf{X}$. By the standard PCA interpretation, each eigenvalue λ_i measures the variance of the data along the corresponding eigenvector direction, i.e., the dispersion of different tokens along that direction. Therefore, the singular values obtained from the SVD of $\mathbf{\Sigma_X}$ directly reflect the dispersion of the data across directions, thereby quantifying the importance of each direction in the representation space.

(2) Rank consistency. $\mathrm{rank}(\mathbf{\Sigma_X}) = \mathrm{rank}(\mathbf{X})$, so spectral analysis of the covariance matrix does not lose the rank information of the representation space.

(3) Uncentered construction. We do not center the representations when computing the covariance matrix. For a representation matrix $\mathbf{X}$, its mean direction corresponds to the component shared by all token representations. Centering would remove this component and the energy it carries, whereas the uncentered construction preserves it in the spectrum, so that the analysis covers the full energy of the representations rather than only the differences among tokens.

Derivation. Given $\mathbf{\Sigma_X} = \frac{1}{m-1}\mathbf{X}^\top\mathbf{X}$, let the SVD of $\mathbf{X}$ be $\mathbf{X} = \mathbf{U\Sigma V}^\top$. Then

$$\mathbf{\Sigma_X} = \frac{1}{m-1}\mathbf{V\Sigma}^\top\mathbf{U}^\top\mathbf{U\Sigma V}^\top = \mathbf{V}\left(\frac{\mathbf{\Sigma}^2}{m-1}\right)\mathbf{V}^\top. \tag{10}$$

Therefore, the singular values of $\mathbf{\Sigma_X}$ are $\mu_i = \sigma_i^2/(m-1)$, $i = 1, \ldots, r$, where $\sigma_1, \ldots, \sigma_r$ are the singular values of $\mathbf{X}$; since $\mathbf{\Sigma_X}$ is symmetric positive semi-definite, its singular values coincide with its eigenvalues, so $\lambda_i = \mu_i = \sigma_i^2/(m-1)$, and its eigenvectors are the right singular vectors of $\mathbf{X}$. Finally, we show that $\mathrm{rank}(\mathbf{\Sigma_X}) = \mathrm{rank}(\mathbf{X})$. From the decomposition above,

$$\mathbf{\Sigma_X} = \mathbf{V}\,\mathbf{D}\,\mathbf{V}^\top, \qquad \mathbf{D} = \frac{\mathbf{\Sigma}^2}{m-1} = \mathrm{diag}\left(\frac{\sigma_1^2}{m-1}, \ldots, \frac{\sigma_r^2}{m-1}, 0, \ldots, 0\right), \tag{11}$$

where $\mathbf{D}$ is a diagonal matrix whose first r diagonal entries are strictly positive (since $\sigma_1, \ldots, \sigma_r > 0$) and whose remaining entries are zero. The rank of a diagonal matrix equals the number of its nonzero diagonal entries, so

$$\mathrm{rank}(\mathbf{D}) = r. \tag{12}$$

Multiplication by the scalar $1/(m-1)$ does not change the rank, and multiplication by the orthogonal matrix $\mathbf{V}$ on either side also preserves the rank. Therefore,

$$\mathrm{rank}(\mathbf{\Sigma_X}) = \mathrm{rank}(\mathbf{V}\,\mathbf{D}\,\mathbf{V}^\top) = \mathrm{rank}(\mathbf{D}) = r = \mathrm{rank}(\mathbf{X}). \tag{13}$$

Sensitivity to the spectral threshold. The spectral threshold used to filter numerically near-zero singular values does not affect our conclusions. We recompute the vision effective rank under five thresholds, 10^{-2} to 10^{-6}, on both tasks and at layers 8 and 16 (Figure 8). Across all thresholds, layer–task combinations, the ordering of the three conditions and the magnitude of the gap remain unchanged (e.g., at layer 16, full ≈ 6.0–7.5 versus blank ≈ 3.4–3.5), and the curves saturate below 10^{-5}, indicating that 10^{-4} lies in a regime where separation is already substantial. The collapse conclusions are therefore insensitive to this choice.

B DERIVATION OF THE PRINCIPAL-ANGLE BOUNDS

We derive the bounds in Eq. equation 8. Let $v_1^V, \ldots, v_r^V$ be an orthonormal basis of the vision subspace $\mathcal{S}_V$, and let P_T be the orthogonal projector onto the text subspace.

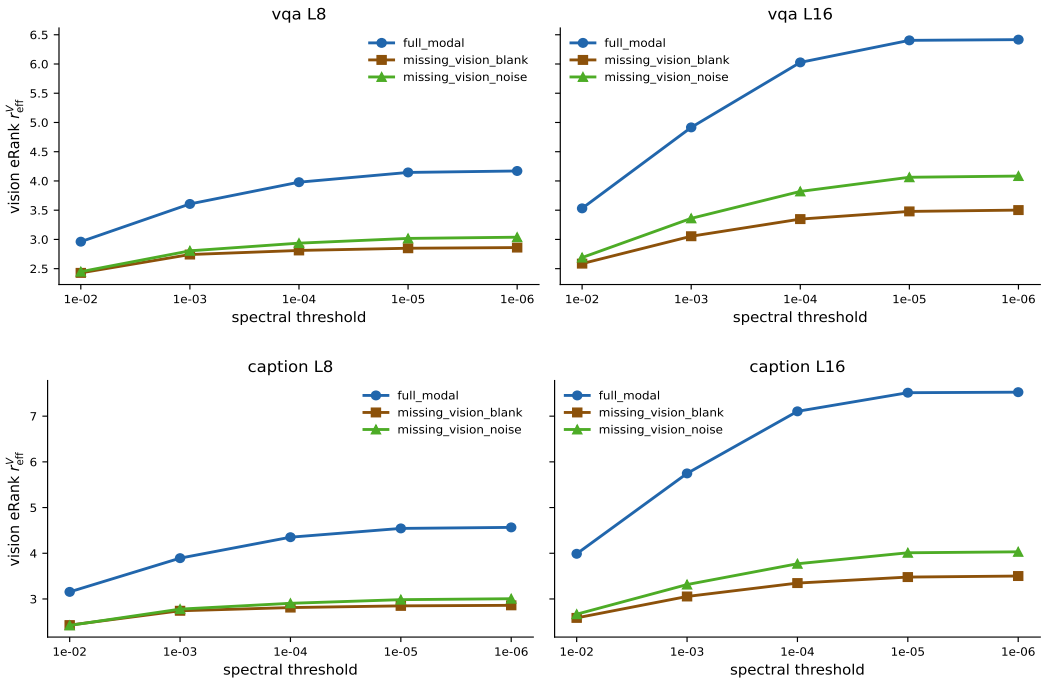


Figure 8: Vision effective rank of the three conditions as a function of the spectral threshold: VQA (top) and image captioning (bottom), at layers 8 and 16.

Lower bound. By the definition of $\theta_{\min}$ (Sec. 3.2), each unit vector $u \in \mathcal{S}_V$ satisfies $\|P_T u\| \leq \cos\theta_{\min}$. Since each v_i^V is a unit vector, it follows that $\|P_T v_i^V\| \leq \cos\theta_{\min}$ for all i, and hence

$$\max_i \|P_T v_i^V\| \leq \cos\theta_{\min}. \tag{14}$$

Upper bound. Any unit vector $u \in \mathcal{S}_V$ can be written as

$$u = \sum_{i=1}^{r} a_i v_i^V, \qquad \sum_{i=1}^{r} a_i^2 = 1, \tag{15}$$

where the second condition follows from $\|u\| = 1$ and the orthonormality of $\{v_i^V\}$. This gives

$$\|P_T u\| = \left\| \sum_{i=1}^{r} a_i \, P_T v_i^V \right\| \leq \sum_{i=1}^{r} |a_i| \, \|P_T v_i^V\| \leq \left(\sum_{i=1}^{r} |a_i| \right) \max_i \|P_T v_i^V\|, \tag{16}$$

where the first inequality is the triangle inequality and the second follows from $\|P_T v_i^V\| \leq \max_j \|P_T v_j^V\|$. By the Cauchy-Schwarz inequality,

$$\left(\sum_{i=1}^{r} |a_i| \right)^2 \leq \left(\sum_{i=1}^{r} a_i^2 \right) \left(\sum_{i=1}^{r} 1^2 \right) = r, \tag{17}$$

hence

$$\sum_{i=1}^{r} |a_i| \leq \sqrt{r}. \tag{18}$$

Therefore $\|P_T u\| \leq \sqrt{r} \max_i \|P_T v_i^V\|$ for every unit vector $u \in \mathcal{S}_V$. Since $\cos\theta_{\min}$ is the maximum of $\|P_T u\|$ over all such u, the same bound holds for $\cos\theta_{\min}$:

$$\cos\theta_{\min} \leq \sqrt{r} \max_i \|P_T v_i^V\|. \tag{19}$$

Combining both bounds. Together, Eqs. equation 14 and equation 19 yield Eq. equation 8.

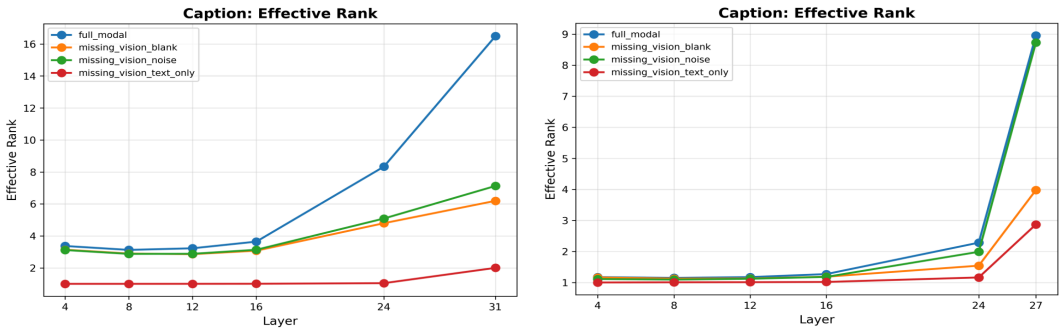
Table 3: Experimental configuration.

Item	Configuration
Models and data	
MLLMs	LLaVA-1.5-7B, Qwen2.5-VL-7B, SmolVLM-2B
Datasets	VizWiz-VQA, Flickr30K
Layers probed	LLaVA: {8, 16, 24, 31}; Qwen: {8, 16, 24, 27}; SmolVLM: {6, 12, 18, 24}
Inference	
Random seed	42
Batch size	LLaVA: 16; Qwen: 4; SmolVLM: 16
Hardware	3 × NVIDIA RTX 3090 (24 GB)
Environment	
LLaVA-1.5	python 3.8.10, torch 2.2.2+cu118, Transformers 4.37.0
Qwen2.5-VL	python 3.10.21, torch 2.5.1+cu121, Transformers 5.17.0
SmolVLM	python 3.10.21, torch 2.5.1+cu121, Transformers 5.17.0

C Implementation Details

All computations are forward-only under `model.eval()`. LLaVA prompts follow the template `USER: <image>\n{question}\nASSISTANT:` or `USER: <image>\n Describe this image in detail.\nASSISTANT:` , and Qwen prompts are built via `apply_chat_template` with the condition image embedded in the message content. Partial masks are drawn as black boxes covering the stated fraction of the image area at a random location; blank is a 336×336 gray image (128, 128, 128), and blur uses a Gaussian filter with $\sigma = 4$. All randomness is controlled by the fixed seed.

D Effective Rank Trajectories on Image Captioning



(a) LLaVA-1.5 (b) Qwen2.5-VL

Figure 9: Effective rank trajectories across layers on Image Captioning under four input conditions.

Figure 9 shows the overall effective rank trajectories on image captioning, under the same four input conditions as in the main text. The pattern on image captioning largely mirrors that on VQA. Under full modality, the overall effective rank expands with depth; under blank and noise, this expansion is suppressed, with blank at the lowest level.

E Matched-rank random-subspace null test: computational procedure

This section describes the procedure for constructing the null distribution used in Sec. 4.3 (Figure 5).

Reference text subspace. For each model and layer, we pool the text-token representations of 500 full-modal samples and take the top-10 right singular vectors of the pooled matrix, yielding a reference basis.

Rank matching. At each layer, the random subspaces have dimension $r = \lceil r_{\text{eff}}^{V} \rceil$, where r_{eff}^{V} is the mean blank-condition vision effective rank at that layer, so that the null matches the observed collapse in dimension.

Null sampling. We draw $K = 10^4$ independent random r-dimensional subspaces of $\mathbb{R}^d$ (each represented by an orthonormal basis Q). For each draw, we compute the cosine of the minimum principal angle against the reference text subspace, i.e., the largest singular value of $Q^\top \tilde{Q}_t$. These K values form the null distribution.

Statistic and p-value. The observed value of each blank sample is computed in the same way, with the sample's own vision-subspace basis in place of Q. The p-value is

$$p = \frac{\#\{\cos\theta_{\min}^{\text{rand}} \geq \cos\theta_{\min}^{\text{obs}}\}}{K}, \tag{20}$$

i.e., the fraction of null draws whose value is at least the observed one. The observed values fall at the extreme right tail of the null at every layer, and we therefore report $p < 10^{-4}$.

F QWEN2.5-VL IS MORE ROBUST TO PARTIAL MASKING

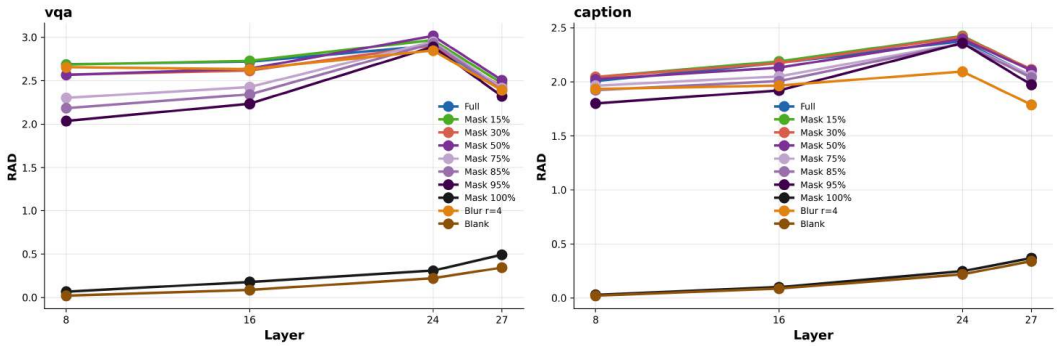


Figure 10: RAD trajectories of Qwen2.5-VL across masking levels on both tasks.

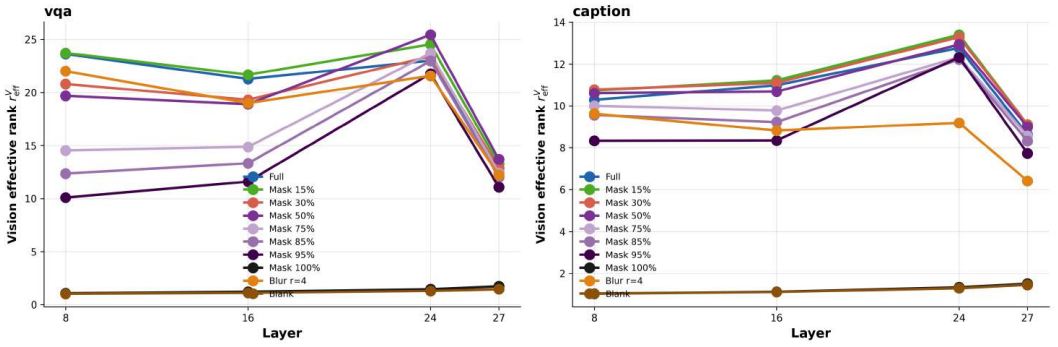


Figure 11: Vision effective rank of Qwen2.5-VL across masking levels on both tasks.

In the main text, we show that RAD decreases monotonically with the degradation level on LLaVA-1.5, and we analyze the response of Qwen2.5-VL to partial masking in this section. To understand its full response profile, we additionally include masking levels of 75%, 85%, and 95%, as well as 100%, forming a complete sequence from 15% to 100%. We first characterize the shape of the

Table 4: Per-sample paired test on Qwen2.5-VL: mean rank difference $\Delta r = r_{\text{eff}}^{V}(\text{mask}) - r_{\text{eff}}^{V}(\text{full})$ and the fraction of samples with $\Delta r > 0$, at layers 8, 16, 24, and 27. Each cell reports the two quantities in the form difference/fraction (fraction in %).

	VQA				Caption			
Condition	L8	L16	L24	L27	L8	L16	L24	L27
mask15	+0.1/50	+0.4/52	+1.5/56	+0.8/57	+0.4/92	+0.2/79	+0.6/87	+0.5/85
mask30	-2.8/39	-2.0/41	+0.4/56	+0.4/56	+0.5/85	+0.1/59	+0.5/70	+0.5/74
mask50	-3.9/39	-2.4/37	+2.4/65	+1.2/61	+0.3/73	-0.3/32	+0.2/55	+0.4/66
mask75	-9.1/17	-6.4/17	+0.7/53	-0.1/49	-0.3/34	-1.2/9	-0.4/37	-0.1/48
mask85	-11.3/3	-8.0/4	-0.0/53	-0.4/47	-0.7/21	-1.8/4	-0.6/43	-0.3/41
mask95	-13.5/4	-9.7/5	-1.2/41	-1.4/33	-2.0/4	-2.6/1	-0.5/44	-0.9/29
mask100	-22.6/0	-20.1/0	-21.6/0	-10.7/0	-9.3/0	-9.9/0	-11.4/0	-7.1/0

response from the RAD trajectories, then examine the vision effective rank, supported by a per-sample paired test, to determine where the response turns, and finally compare this response with that of LLaVA-1.5 and with complete degradation.

As shown in Figure 10, the RAD response of Qwen2.5-VL to partial masking decreases monotonically overall, but remains heavily entangled with the full-modal trajectory at low masking levels. On VQA, the trajectories under *mask15* to *mask50* are entangled with full and nearly indistinguishable; from *mask75* onward the trajectory drops step by step, *mask95* is clearly separated from full, and *mask100* falls directly to the same level as blank, with the two curves completely overlapping at the bottom. On Caption, the RAD trajectories under masking below 75% are likewise entangled with full, high masking levels drop step by step, and *mask100* overlaps with blank. To determine whether this insensitivity arises from a failure of the RAD mechanism or from a property of the architecture itself, we compute the vision effective rank under each masking condition layer by layer.

Figure 11 shows the layer-wise trajectories of vision effective rank under each condition. On VQA, the rank under *mask15* is almost identical to full, or even slightly higher; *mask30* and *mask50* show a mild decrease, a marked drop begins at *mask75*, and *mask100* falls to near the bottom together with blank. On Caption, the pattern is the same but with a gentler magnitude, with the low-mask ranks marginally exceeding the full-modal one. This indicates that the robustness of Qwen2.5-VL to partial masking arises from a property of the architecture itself, rather than from a failure of the RAD mechanism. A per-sample paired test (Table 4) further confirms that this transition is systematic: the shallow-layer fraction crosses the 50% line between mask15 and mask30 on VQA and between mask30 and mask75 on Caption (depending on the layer), and the deep-layer rank increase under moderate masking covers more than half of the samples, ruling out individual fluctuation.

G THE ANALYSIS OF COSINE SIMILARITY UNDER VISUAL DEGRADATION

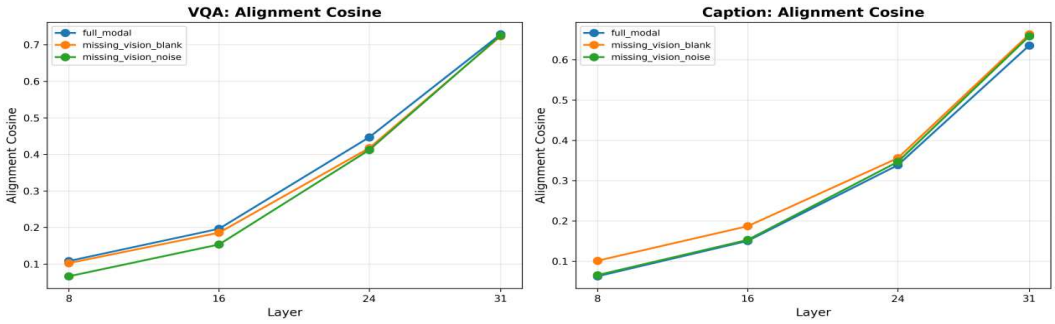
We additionally analyze cosine similarity between the mean-pooled vision and text representations, under the same settings as the principal angles. Given the hidden states of the vision and text tokens at layer ℓ, the two mean representations and their cosine similarity are computed as follows. For each sample and each layer, we compute the mean vision representation and the mean text representation:

$$\bar{\mathbf{v}} = \frac{1}{n_v}\sum_{i=1}^{n_v}\mathbf{v}_i, \qquad \bar{\mathbf{t}} = \frac{1}{n_t}\sum_{j=1}^{n_t}\mathbf{t}_j, \tag{21}$$

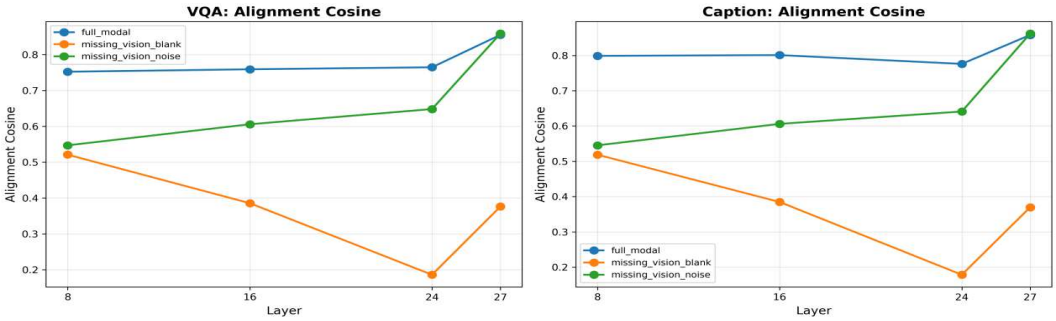
where $\mathbf{v}_i$ and $\mathbf{t}_j$ are the hidden states of the vision and text tokens at that layer, respectively. The cosine similarity is then defined as

$$\cos(\bar{\mathbf{v}}, \bar{\mathbf{t}}) = \frac{\bar{\mathbf{v}} \cdot \bar{\mathbf{t}}}{\|\bar{\mathbf{v}}\|\,\|\bar{\mathbf{t}}\| + \epsilon}, \tag{22}$$

where $\epsilon = 10^{-12}$ prevents division by zero. Unlike principal angles, which compare two subspaces, cosine similarity compares the direction of the two mean representations.



(a) Cosine similarity on LLaVA-1.5



(b) Cosine similarity on Qwen2.5-VL

Figure 12: Cosine similarity between mean-pooled vision and text representations under full, blank, and noise inputs.

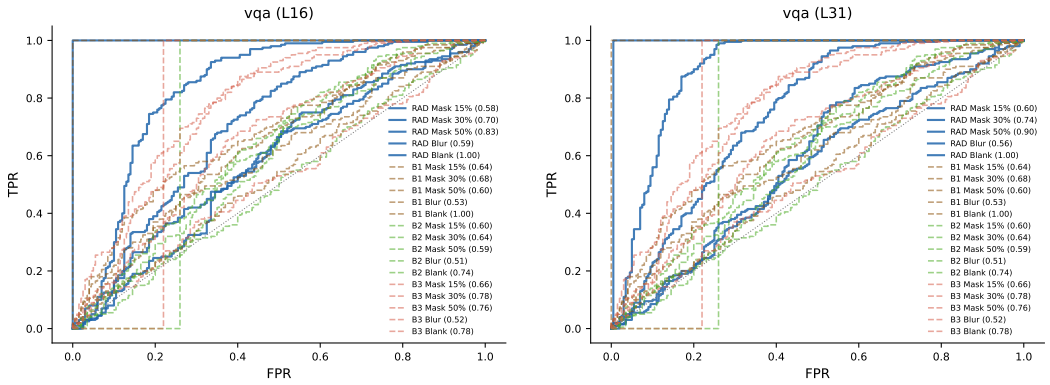
As shown in Figure 12, cosine similarity exhibits opposite behaviors on the two architectures. On LLaVA-1.5, the cosine values under all degraded conditions stay close to those under full modality on both tasks, and even coincide in deeper layers, so this value cannot indicate whether the input has degraded or gone missing, and its discrimination becomes uninformative; on Qwen2.5-VL, by contrast, the values decrease markedly, consistent with the general expectation that degradation weakens alignment. This shows that, under visually degraded inputs, the failure of angular readouts is not specific to the minimum principal angle as a subspace measure: cosine similarity exhibits a similar response.

H Quantitative evaluation of RAD

Detection protocol. We make the detection rule of the main text operational as follows. At each layer ℓ, let $\mu(\ell)$ and $\sigma(\ell)$ be the mean and standard deviation of the RAD values of full-modal samples; the decision boundary is $\tau(\ell) = \mu(\ell) - k\,\sigma(\ell)$, and a sample is judged degraded if its RAD at that layer falls below $\tau(\ell)$. All metrics in this section are computed with each layer evaluated independently: full-modal samples serve as negatives and degraded-condition samples as positives, and we report the AUROC and the TPR@FPR=5% (the detection rate at a 5% false-positive rate), obtained by sweeping the threshold.

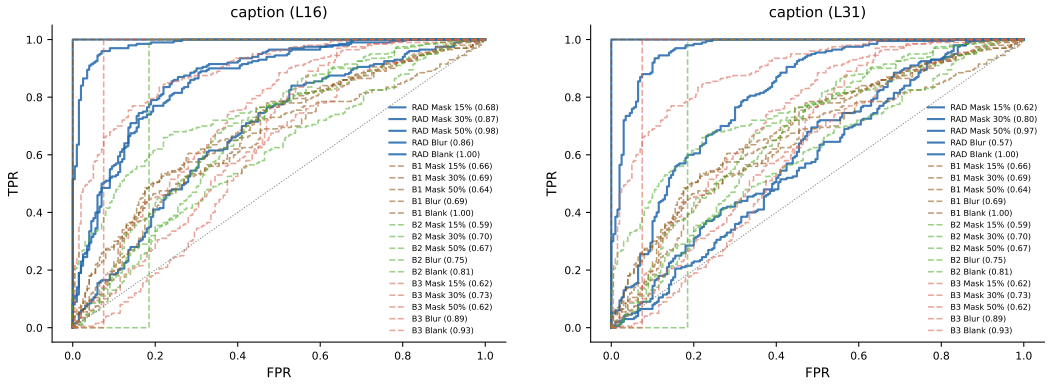
We evaluate RAD against three simple baselines, and for each condition, layer, and baseline, we report AUROC and TPR@FPR=5%:

- B1 (pixel statistics): the standard deviation of image pixels;
- B2 (token variance): the mean variance of vision-encoder output features across dimensions;
- B3 (feature cosine distance): the cosine distance between the pooled vision features and the full-modal reference.



(a) L16 (b) L31

Figure 13: ROC curves of RAD and the three baselines on VQA: (a) layer 16; (b) layer 31. B1: pixel statistics; B2: token variance; B3: feature cosine distance.



(a) L16 (b) L31

Figure 14: ROC curves of RAD and the three baselines on image captioning: (a) layer 16; (b) layer 31. B1: pixel statistics; B2: token variance; B3: feature cosine distance.

Results. Tables 5 and 6 report the full results, and Figures 13 and 14 show the ROC curves at layers 16 and 31. Complete degradation (blank) is detected by RAD and B1 at every layer with AUROC $\approx$ 1.0, whereas the two encoder-space statistics (B2, B3) fail completely at a 5% false-positive rate, with TPR of 0: a constant input does not appear anomalous in encoder statistics. The detection rate for partial masking rises steadily with the masking ratio, and RAD clearly outperforms all baselines in the moderate-to-severe range: on VQA, the AUROC under mask50 reaches 0.90 at layer 31, compared with a maximum of 0.76 among the three baselines; on Caption the gap is larger (0.98 vs. 0.64), with a detection rate of 0.92 vs. 0.19 at FPR=5%. Mild masking (15%) is a weak region shared by all statistics. In this range, B1 and B3 surpass RAD: the black box directly alters the pixel distribution and shallow-layer features, while the change in vision effective rank is still small. Blur lies in between: on Caption, B3 slightly exceeds RAD (0.89 vs. 0.86), while on VQA all statistics are weak.

Summary. Overall, RAD clearly outperforms the three simple baselines under moderate and severe masking; under mild degradation, all methods perform poorly, and some baselines are even better. RAD provides a geometrically interpretable degradation measure as a single scalar, computable without any downstream output. These results support the use of RAD in scenarios with substantial missing content and indicate that detection under mild degradation remains an open problem.

Table 5: Detection evaluation on VQA. Each cell reports AUROC / TPR@FPR=5%. B1: pixel statistics; B2: token variance; B3: feature cosine distance. For each baseline, the orientation yielding the higher AUROC is reported.

	RAD				Baselines		
Condition	L8	L16	L24	L31	B1	B2	B3
Mask 15%	0.56/0.04	0.58/0.03	0.59/0.04	0.60/0.06	0.64/0.11	0.60/0.07	0.66/0.12
Mask 30%	0.68/0.04	0.70/0.04	0.72/0.06	0.74/0.06	0.68/0.18	0.64/0.09	0.78/0.22
Mask 50%	0.80/0.04	0.83/0.05	0.87/0.20	0.90/0.31	0.60/0.15	0.59/0.04	0.76/0.09
Blur	0.58/0.09	0.59/0.08	0.58/0.08	0.56/0.07	0.53/0.06	0.51/0.04	0.52/0.09
Blank	0.99/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	0.74/0.00	0.78/0.00

Table 6: Detection evaluation on image captioning. Each cell reports AUROC / TPR@FPR=5%. B1: pixel statistics; B2: token variance; B3: feature cosine distance. For each baseline, the orientation yielding the higher AUROC is reported.

	RAD				Baselines		
Condition	L8	L16	L24	L31	B1	B2	B3
Mask 15%	0.64/0.04	0.68/0.10	0.63/0.07	0.62/0.05	0.66/0.12	0.59/0.06	0.62/0.07
Mask 30%	0.83/0.20	0.87/0.38	0.82/0.19	0.80/0.14	0.69/0.18	0.70/0.09	0.73/0.09
Mask 50%	0.96/0.78	0.98/0.92	0.97/0.84	0.97/0.76	0.64/0.19	0.67/0.06	0.62/0.01
Blur	0.84/0.29	0.86/0.33	0.69/0.08	0.57/0.03	0.69/0.10	0.75/0.34	0.89/0.57
Blank	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	0.82/0.00	0.93/0.00

I Validation of the Coupling–Direction Relationship on SmolVLM

To verify that the coupling–direction relationship is predictive rather than post hoc, we perform an additional validation on a third model, SmolVLM. Specifically, we first characterize the modality coupling structure of SmolVLM: (1) if it is strongly coupled, the angular response should resemble that of Qwen2.5-VL; (2) if it is weakly coupled, the angular response should resemble that of LLaVA-1.5, i.e., the Spurious Alignment Trap should appear. We then compute the minimum principal angle between the text and vision subspaces at each Transformer layer exactly as in the main text, and compare it with the prediction.

Experimental setup. All experimental settings are consistent with the main text, and the input template follows the official configuration of the model. The analysis follows exactly the same protocol as the main text: VQA and image captioning, layers $\{6, 12, 18, 24\}$.

Modality coupling structure. Fig. 15 shows the coupling spectrum of SmolVLM under full-modal inputs. Compared with LLaVA-1.5 (PR ≈ 6.0) and Qwen2.5-VL (PR ≈ 2.7), the projection mass of SmolVLM spreads even more broadly, with PR ≈ 6.9 on VQA and ≈ 6.4 on image captioning, indicating a **weakly coupled structure** consistent with LLaVA-1.5. Under the mechanism of Section 4.3, weak coupling predicts a spurious shrinking of the angular response under blank inputs, i.e., the Spurious Alignment Trap.

Dimensional collapse. Under blank inputs, the vision effective rank collapses sharply at each layer relative to the full-modal condition (Fig. 17).

Response and prediction test. Fig. 16 shows the layer-wise trajectories of the minimum principal angle $\theta_{\min}$ under the three input conditions. Blank inputs induce the collapse of the vision subspace, and $\theta_{\min}$ lies far below the full-modal trajectory at all four layers on both tasks; the Gaussian-noise control lies between the full and blank conditions. The direction of the response matches the

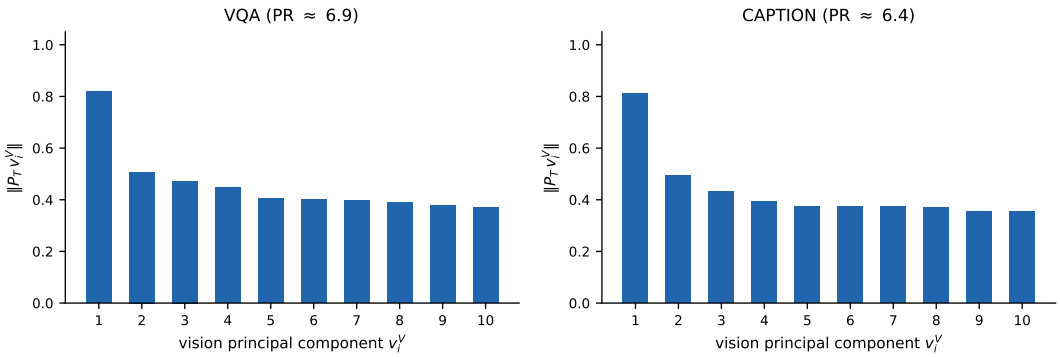


Figure 15: Coupling spectrum of SmolVLM under full-modal inputs, layer-averaged over layers $\{6, 12, 18, 24\}$. The projection mass $\|P_T v_i^V\|$ is spread across vision principal components on both VQA and image captioning, indicating a weakly coupled structure.

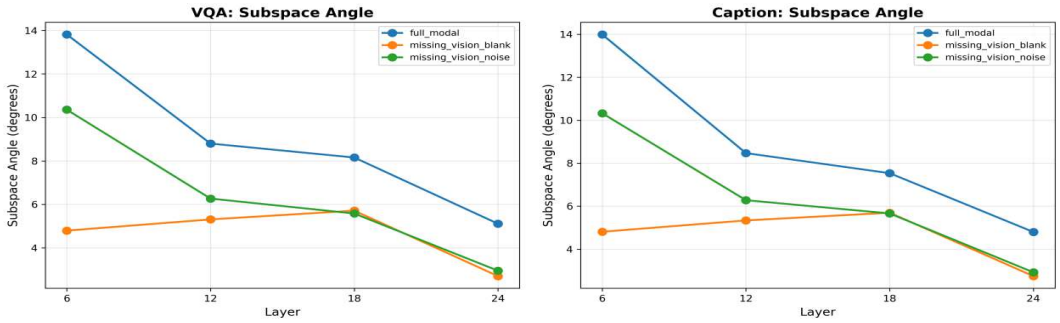


Figure 16: Validation on SmolVLM. Minimum principal angle $\theta_{\min}$ across layers under full, blank, and noise inputs on VQA (left) and image captioning (right).

prediction from the coupling structure: weak coupling corresponds to spurious shrinking, i.e., the Spurious Alignment Trap observed in LLaVA-1.5.

J Limitations and Future Work

Limitations. Several limitations remain to be acknowledged. First, detection is limited for mild degradation (low masking ratios and mild blur), and the method is better suited to cases with substantial missing content. Second, our experiments cover three architectures; extending our conclusions to further families (with different vision-encoder configurations and training paradigms) remains to be verified. Third, the degradation types considered here are synthetic operations (area masking, Gaussian blur, and blank inputs); natural corruption such as transmission errors in real-world deployment is not covered, although the procedure for constructing these conditions is fully compatible with our protocol. Fourth, we focus on the image and text modalities only, and the discussion of scenarios with more modalities is limited, as the latter would also be more complex.

Future work. We see two possible extensions. A particularly promising extension is to use RAD as a detection layer in a missing-aware pipeline and connect it with missing-modality completion methods: for example, triggering a recovery mechanism once visual degradation is detected, forming a detection–recovery loop that places our diagnostic tool upstream of existing work. Finally, we hope to extend the analysis to scenarios with more modalities, and to examine whether these geometric phenomena and the detection protocol hold for more general modality combinations.

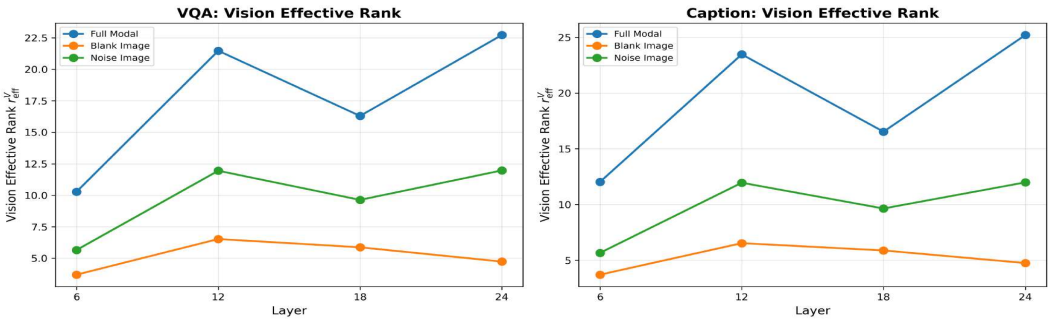


Figure 17: Vision-subspace effective rank r_{eff}^V of SmolVLM across layers under full, blank, and noise inputs on VQA (left) and image captioning (right). Blank inputs collapse the rank at every layer, with noise in between.

K Case Study: RAD-Based Detection on Individual Samples

K.1 RAD-based detection on an individual VQA sample

To illustrate how RAD-based detection behaves on an individual sample, we select one VQA sample and compute its RAD trajectory under four input conditions (full modality, Gaussian blur, 50% masking, and complete degradation), comparing each against the reference trajectory at $k = 1$.

Figure 18 shows the input image and model output for this sample under the four conditions. In all cases, the model produces a semantically plausible answer; even under complete degradation (blank), it still responds with "The camera is very old, as it is described as an old camera." This illustrates a fundamental limitation of behavioral evaluation: when visual input becomes ineffective, the model falls back on language priors and continues to generate fluent, plausible answers, so the generated text alone cannot reveal whether the visual input is still valid. In contrast, as shown in Figure 19, the RAD trajectory clearly separates the 50% masking and complete-degradation conditions. For Gaussian blur, the input image differs only slightly from the original, and the RAD separation is correspondingly limited. Under full modality, the sample's RAD trajectory stays close to the reference; as degradation deepens, the trajectory shifts downward; under complete degradation, it falls far below the reference and beyond the reference's own variability, and is correctly flagged as degraded. This case illustrates that RAD provides an internal diagnostic signal that does not depend on output behavior: even when the model still generates plausible answers under degradation, RAD reflects the degraded state of its visual input.

K.2 RAD-based detection on an individual Image Captioning sample

Figure 20 shows the input image and the caption generated by the model for this sample under the four conditions. Under the full image, the caption is specific and grounded; as degradation deepens, the caption gradually loses detail and becomes generic; under complete degradation, the model still produces a fluent description that depicts a plausible scene despite the absence of any visual evidence. Since the captioning prompt contains no question content, this fluent but ungrounded output arises entirely from the model's internal priors. As in the VQA case, the sample's RAD trajectory lies below the reference trajectory under all degradation conditions (Figure 21); the generated text alone provides no signal of input failure, whereas RAD reveals this phenomenon.

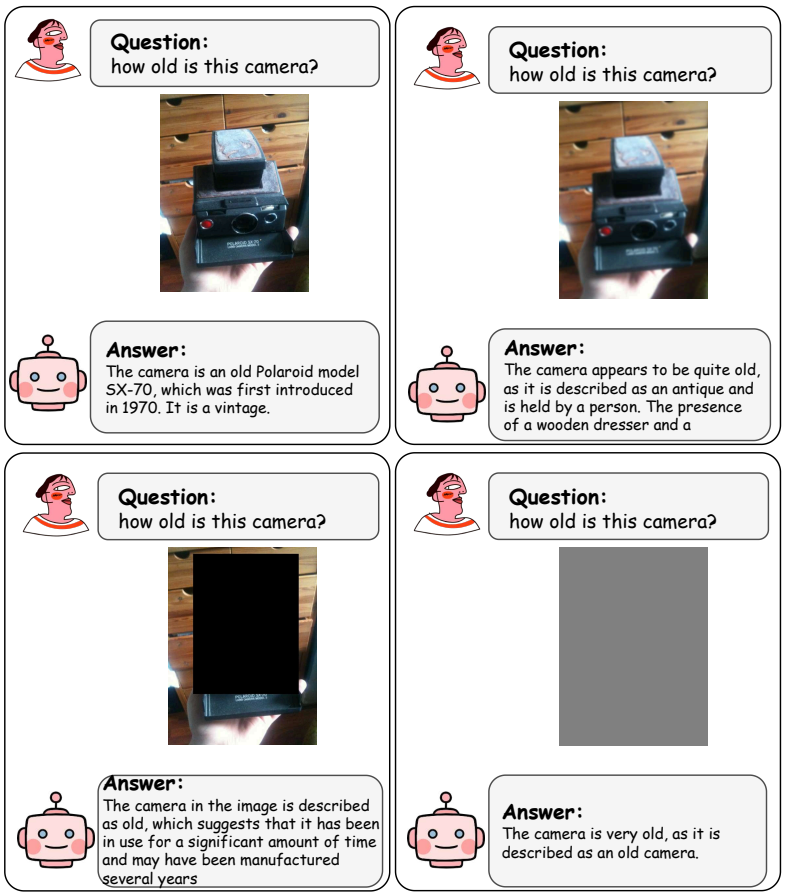
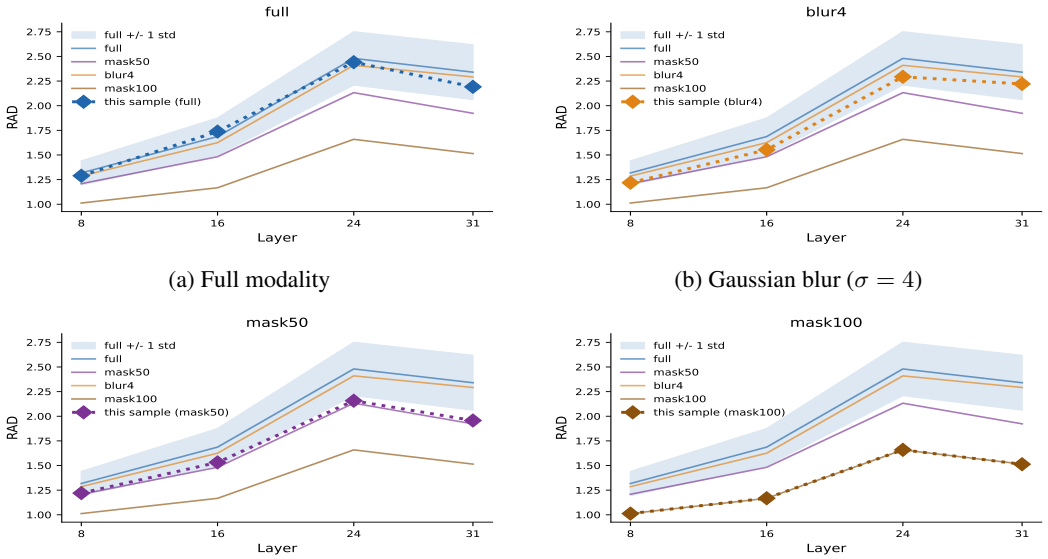


Figure 18: One VQA sample under four input conditions (full, blur, 50% masking, blank) on LLaVA-1.5.



(a) Full modality

(b) Gaussian blur ($\sigma = 4$)

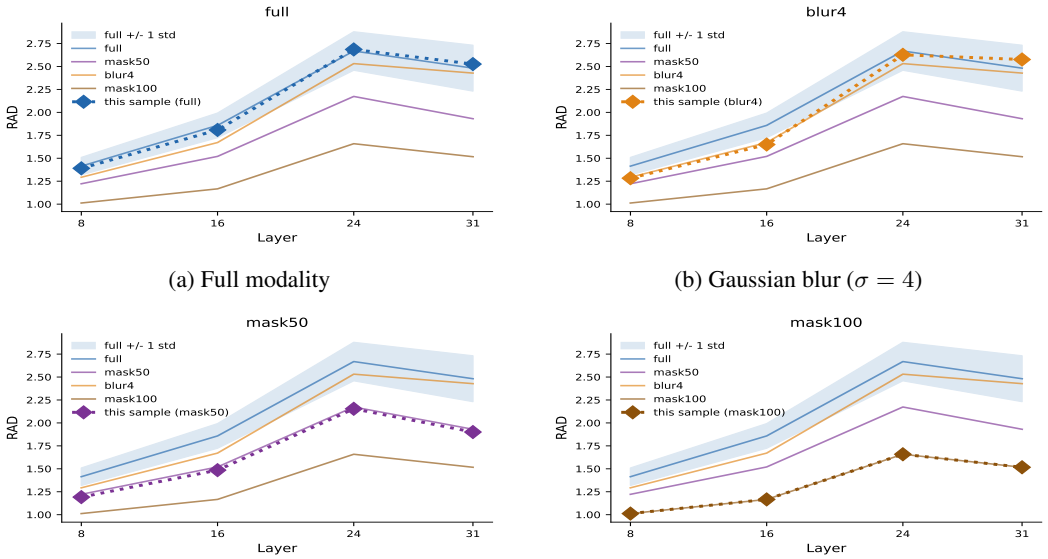
(c) 50% masking

(d) Blank

Figure 19: RAD trajectories of one VQA sample on LLaVA-1.5 under four input conditions: full modality, Gaussian blur ($\sigma = 4$), 50% masking, and blank.



Figure 20: One Image Caption sample under four input conditions (full, blur, 50% masking, blank) on LLaVA-1.5.



(a) Full modality

(b) Gaussian blur ($\sigma = 4$)

(c) 50% masking

(d) Blank

Figure 21: RAD trajectories of one Image Caption sample on LLaVA-1.5 under four input conditions: full modality, Gaussian blur ($\sigma = 4$), 50% masking, and blank.