BEFORE ATTENTION: FUNCTIONAL VISUAL TOKEN ROLES AT THE VISION–LANGUAGE INTERFACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision–language models connect visual encoders to pretrained language models through a vision–language interface, yet visual-token function is typically characterized only after decoder processing begins. We ask whether functionally predictive differences among visual tokens are already accessible at this interface. Across projector- and merger-based LVLMs, we observe a recurring post-interface organization: activation mass becomes less concentrated within individual tokens, while a restricted set of hidden coordinates is repeatedly reused across tokens. To test whether this structure is functionally meaningful, we develop **PRISM**, a lightweight role-recovery procedure that uses paired pre/post concentration responses to construct interface-response anchors, then recovers image-level *contributive* and *residual* roles from dominant coordinate identities and relative ranks. Controlled interventions reveal a consistent functional asymmetry: removing contributive tokens is substantially more disruptive than removing residual tokens, even after matching removal count and subsequent decoder attention. The recovered distinction also extends beyond the initial structural contrast, indicating that PRISM captures structure not available from scalar concentration change alone. Because the partition is computed before query-conditioned decoder processing, the same image-level assignment can be reused across questions. These results show that functional differences among visual tokens are already reflected in representations entering the language model, establishing the vision–language interface as an earlier stage for multimodal analysis.

1 INTRODUCTION

Large vision–language models (LVLMs) connect a visual encoder to a pretrained language model through a vision–language interface (Liu et al., 2024a;b; Lin et al., 2024a; Bai et al., 2025). Before multimodal decoder processing begins, this interface maps visual representations into the language-model hidden space, typically through a modality projector or visual merger. Prior work has primarily studied this interface for visual–language alignment, architectural adaptation, or information transformation and compression (Cha et al., 2024; Lin et al., 2024b; McKinzie et al., 2024; Li et al., 2025; Yao et al., 2024; Jiang et al., 2025; Fan et al., 2026). Much less is known about whether post-interface representations already expose systematic functional differences among individual visual tokens before decoder-side multimodal computation begins.

If such differences are already present, the representations entering the language model contain structure predictive of subsequent token behavior before query-conditioned attention and multimodal interaction. This would position the vision–language interface as an earlier stage for token-level multimodal analysis. We therefore ask: *are functionally predictive visual-token roles already accessible at the vision–language interface?*

We investigate this question by examining visual representations immediately after the vision–language interface. Across diverse LVLM families, we find a recurring coordinate-level organization: although activation mass becomes less concentrated within individual tokens, a restricted set of hidden coordinates is repeatedly reused across tokens. We treat this organization as a structural regularity rather than evidence that individual coordinates are intrinsically important, and test whether it predicts subsequent token behavior.

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.

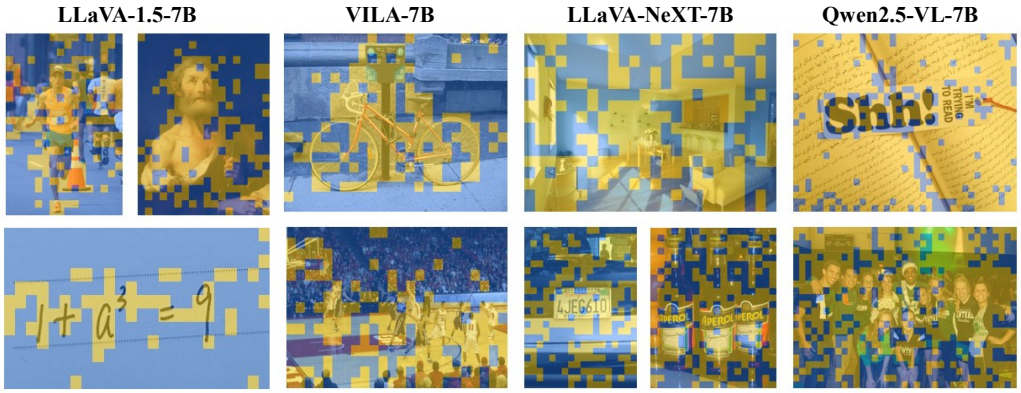

Figure 1: **PRISM identifies visual-token roles directly at the vision–language interface across LVLM families.** From post-interface coordinate–rank structure, PRISM recovers image-level *contributive* (yellow) and *residual* (blue) roles before query-conditioned decoder processing across LLaVA-1.5-7B, VILA-7B, LLaVA-NeXT-7B, and Qwen2.5-VL-7B.

To operationalize this idea, we introduce **PRISM** (**P**ost-interface **R**ole **I**dentification via **S**tructural **M**atching), a lightweight framework for recovering image-level visual-token roles from post-interface representations. PRISM uses a paired pre/post concentration contrast to construct interface-response anchors, then recovers recurring role patterns from the identities and relative ranks of dominant post-interface coordinates. The anchors provide an initial structural contrast rather than assumed functional labels, separating the interface-response signal from the richer coordinate–rank representation used for role recovery. By default, PRISM partitions visual tokens into *contributive* and *residual* roles, where residual denotes the unmatched complement and may still contain useful or complementary visual evidence. When a reproducible external sink criterion is available, the same coordinate–rank representation can optionally support an additional *sink* role; we study this model-specific extension in Appendix N.

We establish the functional meaning of this partition through controlled interventions. Removing contributive tokens is consistently more damaging than removing residual tokens across LVLM architectures. On LLaVA-1.5 and VILA, this asymmetry persists under count- and attention-matched interventions, showing that neither removal count nor decoder-attention magnitude explains the effect. Additional controls show that the initial interface-response contrast contains useful information compared with simpler post-interface scalar heuristics, but direct scalar ranking remains weaker and less consistent than the recovered coordinate–rank partition. A dense linear probe likewise fails to reproduce PRISM's behavior. Because the partition is computed before query-conditioned decoder processing, the same image-level assignment can be reused across questions and used as a prior for fixed-budget visual-token selection. Together, these results show that functional differences among visual tokens are already reflected in representations entering the language model, establishing the vision–language interface as an earlier stage for token-level multimodal analysis.

Main findings. Our main findings are summarized as follows:

- We identify a recurring post-interface organization across diverse LVLM families: activation concentration decreases within individual tokens, while a restricted set of dominant coordinates is strongly reused across tokens.
- We introduce **PRISM**, which uses paired pre/post concentration responses to construct interface-response anchors and recovers image-level *contributive* and *residual* roles from recurring post-interface coordinate–rank structure.
- Controlled interventions show that contributive tokens affect downstream behavior more than residual tokens. The distinction persists in count- and attention-matched controls, extends beyond the initial scalar contrast, and yields a reusable image-level prior across questions and fixed-budget token selection.

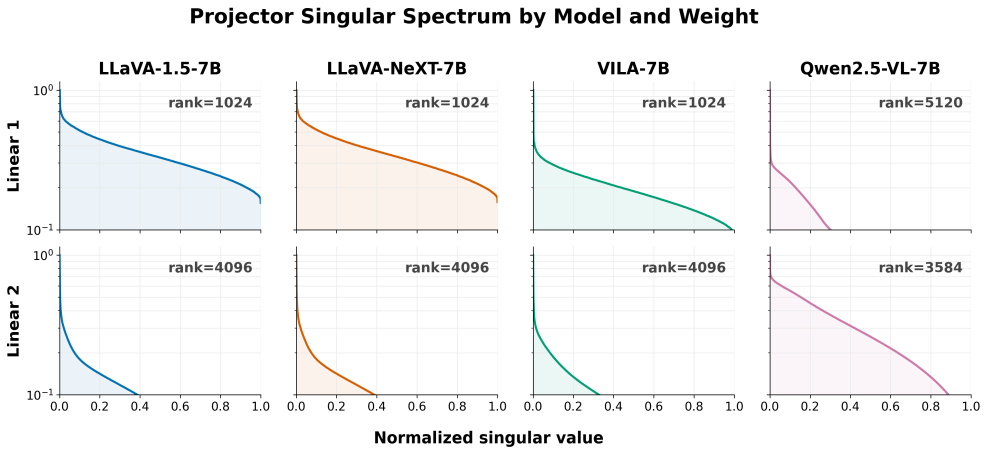

Figure 2: **Directional selectivity of vision–language interfaces.** Normalized singular spectra are shown across four LVLM interfaces. The horizontal axis indexes normalized rank and the vertical axis normalized singular value. Linear 1/2 denote projector layers for MLP-based interfaces and merger transformations for Qwen2.5-VL. Additional spectral statistics are reported in Appendix D.

2 Structure at the Vision–Language Interface

We begin by characterizing visual representations at the vision–language interface before decoder processing. Given an image I, the vision encoder produces visual tokens $V = f_{\text{enc}}(I) \in \mathbb{R}^{P \times d_v}$, which the interface maps to the language-model hidden space, $H = f_{\text{int}}(V) \in \mathbb{R}^{P \times d}$. We denote the pre- and post-interface representations of token i by $\mathbf{v}_i \in \mathbb{R}^{d_v}$ and $\mathbf{h}_i \in \mathbb{R}^{d}$, respectively. We analyze this stage at two complementary levels: the spectral structure of the constituent interface transformations and the coordinate organization of their output representations.

2.1 Directional Selectivity

Figure 2 shows rapidly decaying singular spectra across the evaluated vision–language interfaces. Rather than acting approximately uniformly across the available representation space, the constituent linear transformations exhibit substantial anisotropy. For the MLP projectors in LLaVA-1.5, LLaVA-NeXT, and VILA (Liu et al., 2024a;b; Lin et al., 2024a), the second transformation shows markedly stronger spectral concentration than the first. Qwen2.5-VL uses a different merger-based interface (Bai et al., 2025) and does not follow the same two-stage pattern, but its analyzed merger transformations likewise exhibit substantial spectral anisotropy. Complementary spectral diagnostics in Appendix D support the same conclusion.

These spectra are weight-level diagnostics of the constituent linear transformations rather than a complete characterization of the nonlinear interface. Accordingly, the observed anisotropy does not imply that particular singular directions correspond to semantic concepts or functional token roles. It instead shows that visual representations enter the language model through a structured, non-uniform transformation. We next examine whether recurring organization is also visible directly in post-interface token activations.

2.2 Shared Coordinate Organization

To characterize how activation mass is distributed within each token, we define the top-K_c concentration of a representation $\mathbf{x} \in \mathbb{R}^m$ as

$$C(\mathbf{x}) = \frac{\sum_{j \in \text{TopK}(|\mathbf{x}|, K_c)} |x_j|}{\sum_{j=1}^{m} |x_j|}. \tag{1}$$

This statistic measures the fraction of activation mass carried by the K_c largest-magnitude coordinates.

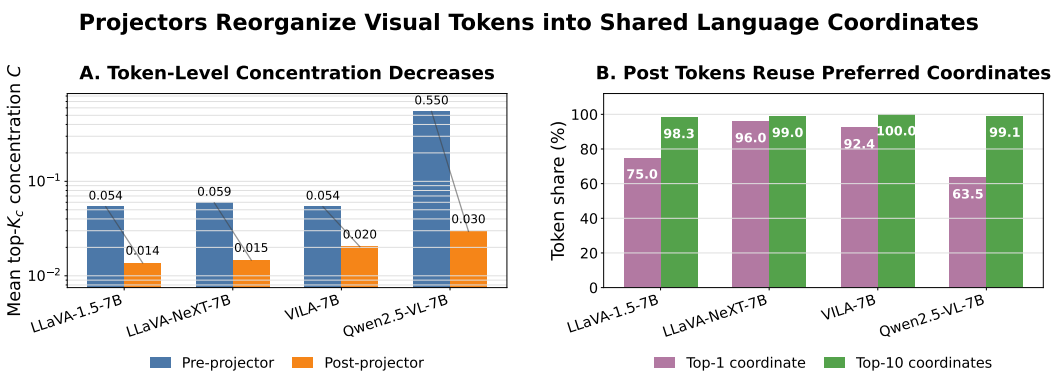

Figure 3: **Post-interface visual tokens exhibit shared coordinate organization.** (A) Mean top-K_c concentration before and after the vision–language interface. Under the same fixed-K_c statistic, post-interface representations exhibit lower concentration in their largest-magnitude coordinates. (B) Reuse of dominant post-interface coordinates. For each token $\mathbf{h}_i$, we identify its largest-magnitude coordinate and report the fraction of tokens covered by the most frequently occurring one and ten coordinates. The ten most frequent coordinates cover 98.3–100.0% of tokens across the evaluated models. Panel B uses 50 sampled VQAv2 validation images (Goyal et al., 2017).

Figure 3A shows lower fixed-K_c concentration after the vision–language interface across the evaluated model families. Thus, under this statistic, post-interface representations do not exhibit increased concentration within their largest-magnitude coordinates. Instead, activation mass is less concentrated in this restricted set of dimensions.

A different pattern emerges when comparing coordinates *across* tokens. As shown in Fig. 3B, the identities of the largest-magnitude post-interface coordinates are highly recurrent: the single most frequently dominant coordinate accounts for 63.5–96.0% of tokens depending on the model, while the ten most frequent coordinates cover 98.3–100.0%. Thus, lower concentration *within* individual tokens coexists with strong reuse of dominant coordinates *across* tokens. Although such recurrence alone does not establish semantic or functional importance, particularly because high-magnitude dimensions can also arise in pretrained language models (Sun et al., 2024), it reveals a compact coordinate-level regularity that cannot be described by token-wise scalar magnitude alone.

Key observation. Post-interface visual tokens exhibit lower concentration *within* individual tokens while strongly reusing a restricted set of dominant coordinates *across* tokens. This contrast motivates separating two aspects of interface-level organization: a token-wise scalar response to the interface and the coordinate identities and relative ranks that characterize its post-interface representation.

2.3 Interface-Induced Structural Contrast

The preceding analysis reveals two complementary aspects of interface-level organization: how a token's concentration changes across the interface and which coordinates dominate its post-interface representation. Because a post-interface statistic alone provides no reference to the token's pre-interface state, we define the paired contrast

$$\Delta_i = C(\mathbf{h}_i) - C(\mathbf{v}_i). \tag{2}$$

We interpret Δ_i as an *interface-relative structural contrast*: it expresses each token's post-interface concentration relative to its own pre-interface concentration, rather than decomposing the interface transformation itself.

Importantly, Δ_i is a signed structural contrast, not a token-importance score. Figure 3A shows that concentration decreases on average after the interface across all evaluated model families. Within an image, the upper tail therefore captures tokens whose concentration is relatively better preserved or increased across the interface, whereas the lower tail captures relatively larger decreases. We use these tails only as contrasting regimes of paired interface response: we neither interpret the absolute magnitude of Δ_i as semantic or functional importance nor compare its values across model families.

The scalar contrast is deliberately incomplete: it records a relative change in concentration while discarding which post-interface coordinates characterize the resulting representation. PRISM therefore uses the image-wise extremes of Δ_i to construct an initial structural partition over contrasting interface-response regimes, while leaving intermediate tokens unassigned. This partition carries no assumed functional semantics. After this initial calibration, Δ_i is discarded, and PRISM instead recovers recurring patterns in the identities and relative ranks of dominant post-interface coordinates.

The paired reference to the pre-interface state is also empirically consequential. Under the same fixed token budget, ranking by post-interface concentration $C(\mathbf{h}_i)$ alone performs substantially worse than ranking by the paired contrast Δ_i across all three reported metrics (Table 19). For example, on POPE, the F1 score increases from 22.66 with top post-interface concentration to 69.56 with top-Δ_i. Thus, referencing each token to its own pre-interface state yields a markedly different and more informative scalar signal than the endpoint concentration alone. At the same time, direct Δ_i ranking remains inconsistent as a standalone selector—on POPE, for example, random selection reaches 77.22 F1—and remains weaker than the recovered PRISM partition. The contrast is therefore useful for constructing the initial structural partition, but insufficient as a token-importance score by itself.

The functional meaning of the recovered partition is established independently through downstream interventions. Section 3 formalizes this procedure.

3 Recovering Visual-Token Roles with PRISM

We introduce **PRISM** (**P**ost-interface **R**ole **I**dentification via **S**tructural **M**atching), a lightweight framework that converts the interface-level structure identified in Sec. 2 into a query-agnostic visual-token partition. PRISM consists of three steps: constructing an initial structural partition from paired interface-response anchors, representing each token by its dominant post-interface coordinate ranks, and learning a compact rule that captures recurring coordinate–rank patterns. The resulting partition is computed entirely before decoder processing.

3.1 Pretext Structural Partition from Interface Responses

Visual-token roles are not directly annotated. Rather than assigning surrogate labels to latent functional roles, we use the interface-relative contrast Δ_i from Eq. 2 to construct a *pretext structural partition*. For each image, we define lower- and upper-response anchor sets as

$$\mathcal{A}^- = \{i \mid \Delta_i \le Q(q_{\text{low}})\}, \qquad \mathcal{A}^+ = \{i \mid \Delta_i \ge Q(q_{\text{high}})\}, \tag{3}$$

where $Q(q)$ denotes the image-wise q-quantile. We use $q_{\text{low}} = 0.3$ and $q_{\text{high}} = 0.9$ by default. Tokens outside $\mathcal{A}^- \cup \mathcal{A}^+$ remain unassigned.

Only the relative ordering within each image is used. The two anchor sets define contrasting regimes of signed interface response: $\mathcal{A}^+$ contains tokens whose concentration is relatively better preserved or increased across the interface, whereas $\mathcal{A}^-$ captures relatively larger decreases. The superscripts $+$ and $-$ denote the upper and lower response tails only and carry no functional semantics. The intermediate 60% of tokens remains unassigned. Unlike confidence-based pseudo-labeling, these anchors are not intended to estimate the correctness of a latent functional role. They instead provide an initial structural contrast for discovering recurring post-interface organization. Restricting calibration to the two anchor sets also avoids imposing the full Δ_i ordering on the recovered partition.

After this initial calibration, Δ_i is discarded. PRISM instead recovers recurring structure from post-interface coordinate identity and relative rank. Calibration visualizations and sensitivity analyses are provided in Appendix B.

3.2 Coordinate–Rank Representation

The scalar contrast in Eq. 2 records only a one-dimensional interface-relative response and discards which hidden coordinates characterize the post-interface representation. PRISM therefore represents each post-interface token $\mathbf{h}_i$ by the ordered identities of its $K_{\max}$ largest-magnitude coordinates:

$$\boldsymbol{\pi}_i = (\pi_{i,1}, \ldots, \pi_{i,K_{\max}}) = \text{TopKIndices}(|\mathbf{h}_i|, K_{\max}), \tag{4}$$

where coordinates are ordered by decreasing absolute activation magnitude.

Let $\rho_i(d)$ denote the rank of coordinate d in $\boldsymbol{\pi}_i$, with $\rho_i(d) = \infty$ if d is absent. We define a coordinate–rank predicate

$$p_{d,r}(\mathbf{h}_i) = \mathbb{I}[\rho_i(d) \le r]\,, \tag{5}$$

which tests whether coordinate d appears among the top-r largest-magnitude coordinates of token i. This representation deliberately discards absolute activation values and retains only coordinate identity and relative rank, reflecting the shared-coordinate organization observed in Sec. 2.2.

3.3 Rule Induction and Role Assignment

PRISM learns a compact disjunctive rule $\mathcal{R} = \{(d_1, r_1), \ldots, (d_M, r_M)\}$ over coordinate–rank predicates:

$$R(\mathbf{h}_i; \mathcal{R}) = \bigvee_{(d,r)\in\mathcal{R}} p_{d,r}(\mathbf{h}_i). \tag{6}$$

The rule fires when any selected coordinate–rank condition is satisfied; predicates are neither weighted nor averaged.

Starting from $\mathcal{R}_0 = \varnothing$, we greedily add the predicate that best recovers the two-set response-anchor partition. Let $f(\mathcal{R})$ denote the macro-F1 of rule $\mathcal{R}$ on $\mathcal{A}^- \cup \mathcal{A}^+$, with the two anchor sets defining the reference partition. At step t, we select

$$p_t^\star = \arg\max_{p\notin\mathcal{R}_{t-1}} f(\mathcal{R}_{t-1} \cup \{p\})\,. \tag{7}$$

A predicate is retained only if it strictly improves f. Rule growth terminates when no remaining predicate improves the objective or when 32 predicates have been selected. This objective recovers the pretext structural contrast; it uses neither downstream task labels nor functional-role annotations.

At inference time, the learned rule induces a binary structural partition:

$$g(\mathbf{h}_i) = \begin{cases} \text{matched}, & R(\mathbf{h}_i;\mathcal{R}) = 1, \\ \text{unmatched}, & \text{otherwise.} \end{cases} \tag{8}$$

Controlled interventions in Sec. 5 establish a consistent functional asymmetry between these groups: removing rule-matched tokens is more damaging than removing the unmatched complement. We therefore refer to the matched and unmatched groups as *contributive* and *residual*, respectively. These terms describe a relative functional distinction rather than absolute token utility; residual tokens may still preserve useful or complementary visual evidence.

Optional sink-aware extension. When a reproducible visual-sink criterion is externally specified (Kang et al., 2025), the same coordinate–rank representation can support an additional sink rule evaluated before the contributive rule. We restrict this extension to backbones with a reproducible sink criterion, avoiding transfer of model-specific dimensions or thresholds across architectures. Sink supervision is not required by default PRISM; details are provided in Appendix N.

Calibration and inference. PRISM is calibrated once per LVLM backbone and reused across datasets and questions. At inference, role assignment requires only the post-interface representation, top-$K_{\max}$ coordinate extraction, and evaluation of a small predicate set. It requires no decoder attention, gradients, iterative token scoring, or additional language-model forward passes. Calibration and hyperparameter details are provided in Appendix B.

4 Related Work

Representations at the vision–language interface. Prior work has studied the vision–language interface from architectural and representational perspectives (Cha et al., 2024; Lin et al., 2024b; McKinzie et al., 2024; Li et al., 2025). Recent analyses examine semantic abstraction across the projector (Yao et al., 2024), fine-grained vision–language alignment (Jiang et al., 2025), and information and redundancy in projected visual tokens (Fan et al., 2026). Most closely related, Fan et al.

(2026) characterize projected tokens through sink, dead, and alive categories based on semantic content. PRISM instead asks whether post-interface coordinate structure exposes differences predictive of downstream functional sensitivity, and recovers a reusable image-level partition from coordinate identity and relative rank.

Decoder-side visual-token analysis. Visual-token importance is commonly analyzed after multimodal decoder processing has begun, particularly through attention-based signals (Chen et al., 2024; Kang et al., 2025). Such signals incorporate the current language context and have been used to characterize or select visual tokens during decoder computation. PRISM studies a complementary pre-decoder regime: it recovers a query-agnostic partition directly from post-interface representations, before decoder attention is available. Controlled interventions test whether the recovered distinction can be explained by subsequent decoder-attention magnitude.

Visual-token selection and pruning. A growing body of work improves LVLM efficiency by reducing visual tokens processed by the language model (Chen et al., 2024; Shang et al., 2025; Zhang et al., 2024; Yang et al., 2025; Alvar et al., 2025). Existing methods exploit decoder attention, redundancy, sparsity, or diversity to merge, prune, or select token subsets. These methods primarily address which tokens to retain under a computational budget. PRISM instead recovers a pre-decoder role prior that can be combined with downstream selectors under a fixed token budget.

5 Experiments

5.1 Experimental Setup

We evaluate PRISM on four LVLM families: LLaVA-1.5-7B (Liu et al., 2024a), VILA-7B (Lin et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025), covering projector- and merger-based vision–language interfaces. Evaluation focuses on POPE (Li et al., 2023), TextVQA (Singh et al., 2019), and MME (Fu et al., 2025). Each rule is calibrated once per backbone and reused across datasets and questions without downstream benchmark labels. Controlled interventions operate on each model's native visual-token sequence under its standard evaluation preprocessing; fixed token counts are imposed only for the fixed-budget selection experiments. Within each comparison, the checkpoint, preprocessing, prompts, decoding, and evaluation protocols are fixed. Full settings are provided in Appendix A.

Our experiments test whether (1) the recovered roles exhibit functional asymmetry, (2) this distinction survives simpler controls, and (3) the image-level prior remains reusable across questions and token selectors.

Table 1: **Functional validation of recovered roles. (A)** Percentage change from full-token performance after complete role removal. **(B)** Count- and attention-matched intervention. Contrib. attn. is the relative mean-attention difference between matched contributive and residual tokens; positive gaps indicate greater damage from contributive removal.

A. Complete role removal				
Backbone	Removed	POPE F1	TextVQA	MME
LLaVA-1.5	Residual	+0.23	−2.57	−4.46
	Contributive	**−12.24**	**−19.28**	**−12.16**
VILA	Residual	−2.52	−8.02	−4.21
	Contributive	**−15.56**	**−16.40**	**−9.44**
LLaVA-NeXT	Residual	+0.72	−4.08	−3.64
	Contributive	**−12.40**	**−24.06**	**−5.84**
Qwen2.5-VL	Residual	−1.54	−0.39	+0.80
	Contributive	**−11.33**	**−40.62**	**−17.02**
B. Count- and attention-matched intervention				
Backbone	Contrib. attn.	POPE gap	TextVQA gap	MME gap
LLaVA-1.5	−47.2%	+10.56	+10.45	+214.9
VILA	−55.9%	+12.71	+6.29	+150.1

5.2 Functional Validation by Controlled Intervention

We first test whether the recovered roles predict subsequent model behavior. For each input, we intervene on one role while leaving the other unchanged; retained tokens within the intervened role are sampled uniformly at random. Across all four backbones, removing contributive tokens is consistently more damaging than removing residual tokens (Table 1A). Full retention curves are provided in Appendix C, confirming a relative functional asymmetry rather than a useful-versus-useless distinction.

Because the roles differ in cardinality (Appendix O), we additionally count-match contributive and residual tokens within decoder-attention quartiles on LLaVA-1.5 and VILA. Within these matched subsets, contributive tokens have 47.2% and 55.9% lower mean decoder attention, respectively, yet their removal remains more damaging on every metric (Table 1B). Thus, neither removal count nor subsequent decoder-attention magnitude explains the functional asymmetry.

Table 2: **PRISM extends beyond the initial interface-response contrast.** PRISM and direct Δ_i use the same top-10% budget ($K = 58$). Mid. 60% is the fraction of PRISM-selected tokens originating from the unassigned middle region of the Δ_i distribution. PRISM results are mean $\pm$ standard deviation over three seeds.

Backbone	Selector	TextVQA	POPE F1	MME-P	MME Total	Mid. 60%
LLaVA-1.5	Direct Δ_i	**53.07**	61.69	1250.7	1566.7	–
LLaVA-1.5	**PRISM**	52.71 ± 0.17	**81.99±0.39**	**1385.0±11.2**	**1707.5±33.3**	**58.34**
VILA	Direct Δ_i	51.68	67.48	1177.8	1535.0	–
VILA	**PRISM**	**53.18±0.57**	**80.27±0.28**	**1337.2±0.9**	**1658.0±8.9**	**56.28**

Table 3: **Additional controls.** **(A)** Dense probe versus PRISM using the same response-anchor assignments. **(B)** Overlap with size-matched high-activation axes of the base LLM. **(C)** Functional gaps after calibration on different image sources; positive values indicate greater damage from contributive removal.

A. Dense-probe control

Selector	TextVQA	POPE	MME-P
Linear probe	55.23	74.14	1407.3
PRISM + rand.	**55.87±0.10**	**86.39±0.10**	**1430.2±4.0**

B. Base-LLM overlap

Comparison	MME	WikiText
Content top-26	0/26	0/26
Stopword top-26	0/26	0/26

C. Calibration-source robustness

Calibration	POPE	TextVQA	MME
LLaVA-1.5			
MME	+10.71	+9.70	+143.6
TextVQA	+11.29	+11.01	+202.8
POPE	+10.80	+7.65	+122.2
VILA			
MME	+11.25	+5.20	+93.6
TextVQA	+10.71	+5.10	+95.9
POPE	+9.08	+3.62	+45.5

5.3 Separating PRISM from Simpler Explanations

Beyond the initial structural contrast. We first clarify what information is carried by the interface-relative contrast. Under matched fixed-budget selection on LLaVA-1.5, top-Δ_i substantially outperforms post-interface norm and concentration across TextVQA, POPE, and MME-P, showing that the paired contrast contains useful information beyond these simpler endpoint scalars (Appendix M). At the same time, direct Δ_i ranking is not consistently strong relative to other selectors, indicating that the scalar contrast is insufficient as a standalone importance score.

At the original top-10% budget ($K = 58$), direct Δ_i ranking is substantially weaker than PRISM on POPE and MME while remaining comparable on TextVQA (Table 2). Moreover, 58.34% and 56.28% of PRISM-selected contributive tokens on LLaVA-1.5 and VILA originate from the *unassigned middle 60%* of the Δ_i distribution. Thus, the initial interface-response contrast provides useful structural anchors, while the learned coordinate–rank rule generalizes beyond its scalar ordering. Additional direct-Δ_i controls are provided in Appendix F.

Additional controls. The recovered distinction is not explained by generic supervised access to post-interface activations. Using the same response-anchor assignments, PRISM outperforms a

Table 4: **Reuse and downstream utility.** **(A)** One fixed PRISM mask is reused across questions; Gain denotes PRISM minus FastV. **(B)** Performance retained relative to full-token inference at approximately 10% visual-token budget.

A. Same-image reuse across questions

Subset	Full	FastV	PRISM	Gain
LLaVA / ≥ 2 Q	76.72	71.31	**74.73**	+3.42
LLaVA / ≥ 2 types	76.35	70.92	**74.36**	+3.44
VILA / ≥ 2 Q	79.48	73.32	**76.02**	+2.70
VILA / ≥ 2 types	79.12	73.00	**75.70**	+2.70

B. Fixed-budget role-guided selection

Backbone	FastV	Residual	Contributive
LLaVA-1.5	88.3	78.5	**94.4**
VILA	91.8	83.0	**93.7**
LLaVA-NeXT	90.9	90.3	**94.5**
Qwen2.5-VL	90.3	85.7	**92.5**
Average	90.3	84.4	**93.8**

logistic-regression probe over the full 4096-D representation on TextVQA, POPE, and MME-P (Table 3A). Its learned coordinates also show zero overlap with size-matched high-activation axes of the text-only base LLM on MME and WikiText (Table 3B). Finally, recalibration on TextVQA or POPE preserves positive functional gaps across all evaluated metrics on LLaVA-1.5 and VILA (Table 3C). Full protocols are provided in Appendices F.4, J, and H.

5.4 Reuse and Downstream Utility

Reuse across questions. We reuse one fixed PRISM role map across 5,000 VQAv2 questions (Goyal et al., 2017) covering 883 images, including 855 with at least two question types. Under the same approximately 10% token budget, PRISM remains closer to full-token inference than FastV on both LLaVA-1.5 and VILA (Table 4A). Because no role recomputation is required when the question changes, this supports the partition as a reusable image-level prior rather than a query-conditioned importance estimator. Details are provided in Appendix I.

Fixed-budget token selection. We next use the recovered roles as candidate pools for downstream selection. With the token budget and FastV selector fixed, contributive-guided selection improves average retained performance from 90.3 to 93.8, whereas residual-guided selection decreases it to 84.4 (Table 4B). The same trend holds with DivPrune (Alvar et al., 2025) (Appendix K.2); full FastV results are provided in Appendix K.1.

6 Conclusion

We investigate whether functionally predictive differences among visual tokens are already accessible at the vision–language interface before decoder processing. Across diverse LVLM architectures, we find directional selectivity in interface transformations and recurring dominant-coordinate structure in post-interface representations. Building on this structure, **PRISM** uses paired interface-relative concentration responses to construct an initial structural contrast and recovers query-agnostic contributive–residual roles from compact coordinate–rank patterns. Controlled interventions establish a consistent functional asymmetry that is not explained by direct scalar ranking, dense probing, calibration source, or subsequent decoder attention, while the same partition remains reusable across questions and fixed-budget token selection. These results establish the vision–language interface as an earlier stage for token-level multimodal analysis.

7 Limitations and Future Work

Our results establish an empirical relationship between post-interface coordinate–rank organization and downstream functional sensitivity, but do not identify its underlying mechanism. The paired contrast Δ_i is useful but incomplete, and our results do not establish it as the unique or optimal interface-response statistic for recovering such structure. Future work could compare alternative interface-response statistics and trace how these patterns arise from interface training, visual-encoder statistics, and subsequent decoder computation.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, stylistic refinement, and the presentation of author-provided experimental results. They were not used to execute experiments, generate experimental data, or implement the code used in this work. All research decisions, experimental procedures, quantitative analyses, and final scientific claims were independently reviewed and validated by the authors. The authors reviewed and revised all AI-assisted text and take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable or private data, user studies, or the collection of new datasets. Our experiments analyze existing large vision–language models using established multimodal benchmarks and standard evaluation protocols. The proposed method concerns the analysis and intervention of internal visual-token representations and does not introduce a direct high-stakes or user-facing application. As with related work on multimodal models, downstream use remains subject to the limitations and potential biases of the underlying models and evaluation datasets.

REPRODUCIBILITY STATEMENT

We provide the methodological definitions and experimental protocols required to reproduce the main results in the paper and appendix. The appendix documents model configurations, preprocessing and evaluation settings, calibration and rule-induction procedures, hyperparameters, token budgets, controlled-intervention protocols, and additional sensitivity analyses. All experiments were implemented and conducted by the authors following the procedures described in the manuscript. Upon acceptance, we will publicly release the implementation of **PRISM**, including code for calibration, role assignment, controlled interventions, fixed-budget token selection, and evaluation, to facilitate independent reproduction and further research.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

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.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13817–13827. IEEE, 2024.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Yingqi Fan, Junlong Tong, Anhao Zhao, and Xiaoyu Shen. What do visual tokens really encode? uncovering sparsity and redundancy in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11987–11997, June 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2025.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. *Advances in Neural Information Processing Systems*, 38:874–899, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 133(10):6794–6812, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26679–26689. IEEE, 2024a.

Junyan Lin, Haoran Chen, Dawei Zhu, and Xiaoyu Shen. To preserve or to compress: An in-depth study of connector selection in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5666–5680, 2024b.

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. IEEE, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pp. 304–323. Springer, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867. IEEE, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

A Additional Experimental Details

A.1 Models and Evaluation Protocol

Model families. We evaluate PRISM on four LVLM backbones: LLaVA-1.5-7B (Liu et al., 2024a), VILA-7B (Lin et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025). These models cover both MLP-based modality projectors and merger-style vision–language interfaces. Unless otherwise specified, each model uses its native preprocessing and evaluation resolution. Role assignment is performed immediately after the vision–language interface and before decoder traversal.

Visual-token configurations. For controlled role intervention, we operate on each model's full native visual-token sequence under its standard benchmark preprocessing, because the objective is to measure the functional sensitivity of the recovered roles rather than pruning efficiency. Only for fixed-budget selection do we use standardized token configurations: LLaVA-1.5-7B and VILA-7B use 336×336 inputs with 576 visual tokens, LLaVA-NeXT-7B uses 672×672 inputs with 2880 visual tokens, and Qwen2.5-VL-7B uses fixed 1008×1008 inputs yielding 1296 visual tokens. Consequently, native role counts reported elsewhere need not sum to these fixed-budget sequence lengths.

Evaluation protocol. We use `lmms-eval` (Zhang et al., 2025) and official benchmark-specific evaluation procedures whenever supported. Within each comparison, the checkpoint, image preprocessing, resolution, prompt format, decoding configuration, and evaluation procedure are held fixed. Only the PRISM partition or the downstream token-selection strategy is varied. For the broader fixed-budget evaluation, we additionally report results on GQA (Hudson & Manning, 2019), ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2024), and MMBench (Liu et al., 2024c), together with TextVQA (Singh et al., 2019), MME (Fu et al., 2025), and POPE (Li et al., 2023).

Calibration. A PRISM rule is calibrated once for each backbone and subsequently reused across downstream datasets and questions. For the primary LLaVA-1.5-7B calibration, we use 1,187 unique MME (Fu et al., 2025) images, corresponding to approximately 2.1×10^5 visual tokens. MME serves as a broad calibration source spanning diverse perception and reasoning categories; downstream evaluation labels are not used during rule induction.

Rule-learning cost. On a single NVIDIA GeForce RTX 4090 GPU, collecting LLaVA-1.5 interface activations for calibration takes approximately 90 seconds, and greedy rule induction takes approximately 86 seconds. This cost is incurred once per backbone. At inference time, PRISM extracts the top-$K_{\max}$ coordinate indices from post-interface representations and evaluates at most 32 coordinate–rank predicates.

Controlled intervention. When intervening on residual tokens, all contributive tokens are retained; when intervening on contributive tokens, all residual tokens are retained. Within the intervened role, retained tokens are sampled uniformly at random. No secondary importance score is used.

Fixed-budget selection. Unless otherwise specified, fixed-budget experiments retain 58 of 576 tokens for LLaVA-1.5-7B and VILA-7B, 288 of 2880 tokens for LLaVA-NeXT-7B, and 130 of 1296 tokens for Qwen2.5-VL-7B. Selected tokens are restored to their original raster-scan order before decoder processing.

Retained-performance metric. For method m, we define retained performance as

$$\mathrm{Rel.}(m) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{\mathrm{Score}_b(m)}{\mathrm{Score}_b(\mathrm{Full})} \times 100, \tag{9}$$

where $\mathcal{B}$ denotes the benchmark metrics included in the corresponding comparison. MME uses its total score.

B Rule Calibration and Hyperparameter Sensitivity

This section details how the interface-response anchors are constructed and examines whether PRISM depends on a narrow choice of calibration hyperparameters. The four quantities considered here play distinct roles. The quantiles q_{low} and q_{high} determine the lower-response anchor set $\mathcal{A}^-$ and upper-response anchor set $\mathcal{A}^+$, respectively; K_c determines how concentration is summarized when constructing the paired contrast Δ_i; and $K_{\max}$ determines how many dominant post-interface coordinates are exposed to the coordinate–rank rule learner. Importantly, K_c affects only the paired response contrast and anchor construction, whereas $K_{\max}$ controls the representation used for rule induction and inference.

PRISM uses the paired interface-relative structural contrast

$$C(\mathbf{x}) = \frac{\sum_{j \in \text{TopK}(|\mathbf{x}|, K_c)} |x_j|}{\sum_j |x_j|}, \qquad \Delta_i = C(\mathbf{h}_i) - C(\mathbf{v}_i), \tag{10}$$

to construct the response-anchor sets used for initial calibration. Unless otherwise specified, we use $q_{\text{low}} = 0.3$, $q_{\text{high}} = 0.9$, $K_c = 10$, and $K_{\max} = 20$. The intermediate region between the two image-wise quantiles remains unassigned and is excluded from rule induction. When reported below, Anchor F1 denotes macro-F1 evaluated over $\mathcal{A}^- \cup \mathcal{A}^+$.

Tail construction. The image-wise extremes of Δ_i define two deliberately separated regimes of signed interface response. The upper tail forms the upper-response anchor set $\mathcal{A}^+$, whereas the lower tail forms the lower-response anchor set $\mathcal{A}^-$. Tokens outside $\mathcal{A}^- \cup \mathcal{A}^+$ remain unassigned. Because Δ_i is signed, $\mathcal{A}^+$ contains tokens whose concentration is relatively better preserved or increased across the interface, whereas $\mathcal{A}^-$ captures relatively larger decreases. The superscripts $+$ and $-$ denote the two response tails only and carry no functional semantics. We use only these asymmetric tails as response anchors rather than imposing the full ordering of Δ_i on the recovered partition. Unlike confidence-based pseudo-labeling, these anchors do not estimate the correctness of a latent functional role; together they define a pretext structural partition used only for calibration. This design requires the coordinate–rank rule to generalize beyond the scalar anchors while avoiding a global activation threshold across images or model families.

Qualitative interface-response contrast. Figure 4 visualizes the pre-interface concentration $C(\mathbf{v}_i)$, post-interface concentration $C(\mathbf{h}_i)$, and their difference Δ_i for representative LLaVA-1.5-7B examples. The contrast exhibits spatially structured variation across visual tokens, providing qualitative motivation for using image-wise extremes as response anchors. These visualizations are descriptive only: they do not imply that Δ_i is a semantic or functional importance score.

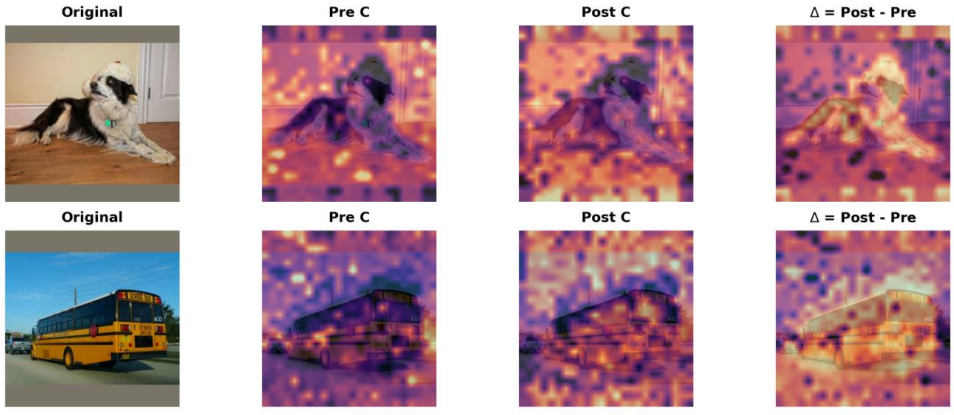

Figure 4: **Qualitative visualization of the paired interface-response contrast on LLaVA-1.5-7B.** From left to right: input image, pre-interface concentration $C(\mathbf{v}_i)$, post-interface concentration $C(\mathbf{h}_i)$, and $\Delta_i = C(\mathbf{h}_i) - C(\mathbf{v}_i)$. The contrast is spatially structured across visual tokens and provides qualitative motivation for using its image-wise extremes as response anchors.

Unified cross-model protocol. We use the same default response-anchor protocol across backbones rather than independently tuning the quantiles for each architecture or downstream benchmark. The rule is calibrated once per backbone and then reused on downstream datasets. Thus, the reported cross-model results do not rely on selecting different q_{low}, q_{high}, K_c, or $K_{\max}$ values for each evaluation benchmark. The sensitivity experiments below vary one component at a time on LLaVA-1.5-7B to measure how strongly downstream behavior depends on these choices.

Sensitivity to q_{high}. The upper-tail threshold controls the strictness of $\mathcal{A}^+$. A larger q_{high} retains fewer, more extreme tokens in $\mathcal{A}^+$, whereas a smaller value admits a broader portion of the Δ_i distribution. Table 5 varies q_{high} from 0.7 to 0.9 while fixing $q_{\text{low}} = 0.3$ and leaving the remaining procedure unchanged.

Table 5: **Sensitivity to q_{high} on LLaVA-1.5-7B.** All rules use the same MME calibration set and fixed-$K = 100$ evaluation.

$(q_{\text{low}}, q_{\text{high}})$	Anchor F1	TextVQA	POPE F1	MME-P
$(0.3, 0.7)$	76.48	57.24	82.19	1482.0
$(0.3, 0.8)$	**79.20**	**57.39**	82.88	**1493.8**
$(0.3, 0.9)$	78.01	57.06	**83.01**	1491.5

No single upper-tail setting dominates all reported quantities. The anchor-fitting macro-F1 is highest at $q_{\text{high}} = 0.8$, POPE is slightly higher at 0.9, and TextVQA and MME-P vary only modestly across the three settings. This is useful for interpreting the default configuration: the downstream effect is not tied to choosing the threshold that maximizes the response-anchor fitting objective. Across the tested range, changing the size of $\mathcal{A}^+$ preserves the same overall level of downstream performance.

Sensitivity to q_{low}. The lower-tail threshold controls the breadth of $\mathcal{A}^-$. With $q_{\text{high}} = 0.9$, we vary $q_{\text{low}} \in \{0.1, 0.2, 0.3\}$, corresponding to progressively broader $\mathcal{A}^-$ sets while leaving the intermediate region unassigned.

Table 6: **Sensitivity to q_{low} on LLaVA-1.5-7B.**

q_{low}	TextVQA	POPE F1	MME-P
0.1	**57.37**	**83.10**	1476.6
0.2	57.05	82.95	**1494.8**
0.3	57.06	83.01	1491.5

The three lower-tail settings yield closely grouped TextVQA and POPE results, while MME-P varies somewhat more but without a consistent monotonic trend. In particular, enlarging $\mathcal{A}^-$ from the bottom 10% to the bottom 30% does not systematically improve or degrade downstream performance. This suggests that the recovered partition is not determined by a narrow choice of the lower-response anchor-set size.

Sensitivity to coordinate cutoffs. We next vary K_c and $K_{\max}$, which affect different stages of PRISM. K_c specifies how many coordinates contribute to the scalar concentration measure used to construct the paired response contrast, whereas $K_{\max}$ determines the depth of ordered coordinate signature available to the rule learner. We vary each while holding the other at its default value.

Table 7: **Sensitivity to K_c and $K_{\max}$ on LLaVA-1.5-7B.**

K_c	$K_{\max}$	TextVQA	POPE F1	MME-P
5	20	57.25	83.00	1488.6
10	20	57.06	**83.01**	1491.5
20	20	**57.40**	82.60	1480.0
10	10	56.63	82.39	**1500.3**
10	30	57.06	**83.01**	1491.5

Changing K_c from 5 to 20 leaves the three downstream metrics in a relatively narrow range, indicating that the response-anchor construction does not depend on one specific concentration cutoff. Changing $K_{\max}$ from 10 to 20 improves TextVQA and POPE while the MME-P score moves in the opposite direction, again showing no single configuration that uniformly optimizes every benchmark. More importantly, increasing $K_{\max}$ from 20 to 30 produces effectively identical results, suggesting limited benefit from exposing substantially deeper coordinate ranks once the dominant portion of the signature is available.

Across all sensitivity configurations in Tables 5–7, the maximum ranges are 0.77 TextVQA points, 0.62 POPE F1 points, and 20.3 MME-P points. Taken together, these experiments indicate that the reported behavior is not concentrated at a single narrow hyperparameter setting. We therefore retain the same default configuration throughout the main experiments rather than tuning PRISM separately for individual datasets.

C Full Controlled-Intervention Results

The main paper reports the complete-removal endpoint in Table 1A for compactness. Here we report the full retention curves to test whether the functional asymmetry is specific to complete role removal or appears progressively as a larger fraction of a role is removed. For each input, only one role is reduced at a time while every token from the other role is preserved. Tokens retained within the intervened role are sampled uniformly at random, so the curves do not introduce an additional within-role importance score. Table 8 reports 50%, 25%, and 0% retention for each role, together with the full-token reference.

Table 8: **Full controlled intervention curves.** For each role, we retain the indicated fraction of its tokens while keeping all tokens from the other role unchanged. Contributive-token reduction produces substantially larger degradation across all four backbones.

		LLaVA-1.5-7B			**VILA-7B**		
Setting	Retain	POPE F1	TextVQA	MME Total	POPE F1	TextVQA	MME Total
Full	100%	85.85	58.3	1864.8	86.25	62.2	1788.7
Residual	50%	86.23	57.6	1825.6	86.23	60.8	1765.0
	25%	86.34	57.1	1786.7	85.75	58.9	1762.8
	0%	86.05	56.8	1781.6	84.08	57.2	1713.4
Contributive	50%	84.60	55.1	1857.5	84.22	58.1	1742.0
	25%	82.40	52.1	1808.3	81.58	56.1	1741.5
	0%	75.34	47.1	1638.0	72.83	52.0	1619.8
		LLaVA-NeXT-7B			**Qwen2.5-VL-7B**		
Setting	Retain	POPE F1	TextVQA	MME Total	POPE F1	TextVQA	MME Total
Full	100%	86.39	61.3	1853.6	86.64	77.8	2316.9
Residual	50%	87.02	60.1	1809.2	86.38	77.9	2325.0
	25%	87.12	59.8	1817.5	86.28	77.8	2327.2
	0%	87.01	58.8	1786.1	85.30	77.5	2335.4
Contributive	50%	85.74	57.3	1781.2	85.26	74.5	2271.9
	25%	83.76	53.0	1801.1	83.21	68.4	2249.3
	0%	75.67	46.6	1745.3	76.82	46.2	1922.6

The full curves show that the contributive–residual asymmetry is not created only by the extreme 0% endpoint. As contributive retention is reduced, performance generally degrades more strongly, and the separation becomes especially clear at aggressive removal. By contrast, substantial residual-token removal often produces much smaller changes; several metrics remain close to the full-token

reference even when the entire residual role is removed. This pattern appears across all four backbones despite their substantially different native visual-token counts.

Residual removal can nevertheless reduce performance, particularly because the residual role may contain a large fraction of the visual sequence and can preserve useful or complementary context. The intervention therefore does not support a useful-versus-useless interpretation. Instead, it supports a relative functional asymmetry: under comparable interventions, tokens recovered as contributive are more consequential on average than the unmatched residual complement. Because native role cardinalities differ, the complete-removal curves are complemented by the count- and attention-matched intervention in Appendix G.

D Additional Analysis of Interface Directional Selectivity

The main paper visualizes the complete singular spectra of the constituent interface transformations. Here we summarize the same geometry with complementary rank- and energy-based statistics. Stable rank and entropy effective rank quantify how broadly transformation energy is distributed across singular directions, while the 95%-energy statistic measures how many directions are required to recover most of the spectrum. Top-k energy dominance provides the complementary view of how much of the transformation is concentrated in only a few leading directions. These quantities are used only as weight-level diagnostics; they are not inputs to PRISM and do not assign token roles.

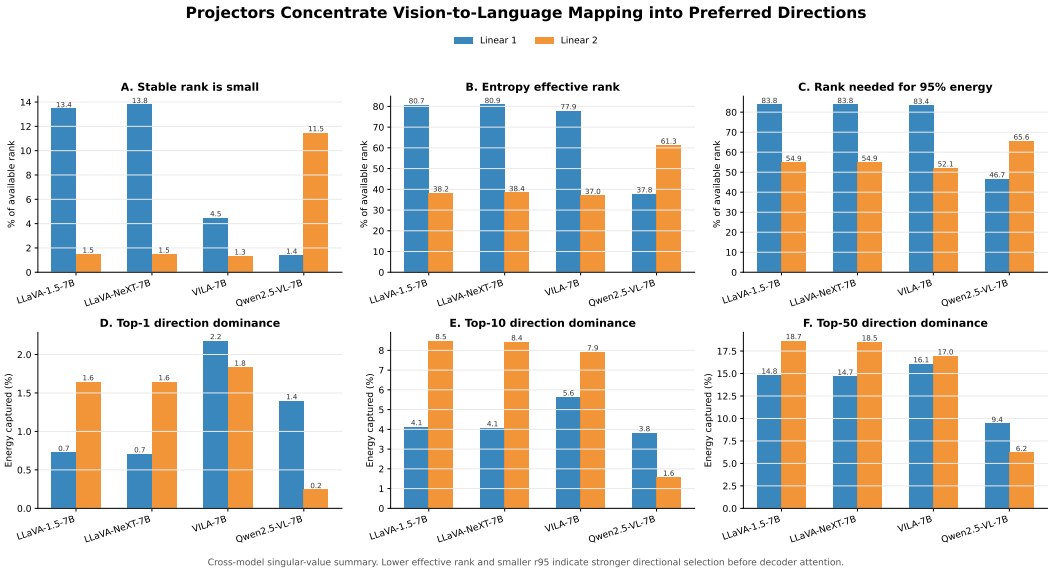

Figure 5: **Quantitative summary of vision–language interface singular-value geometry.** We report stable rank, entropy effective rank, the fraction of rank required to recover 95% of spectral energy, and top-1/top-10/top-50 energy dominance across the interface weights of LLaVA-1.5-7B, LLaVA-NeXT-7B, VILA-7B, and Qwen2.5-VL-7B.

For LLaVA-1.5, LLaVA-NeXT, and VILA, we analyze the two linear transformations of the MLP interface. For Qwen2.5-VL, we analyze the corresponding linear weights in the visual merger. The main paper presents the complete singular spectra; Fig. 5 reports complementary summary statistics.

Across the MLP-based interfaces, the second transformation is substantially more directionally concentrated than the first. In both LLaVA models, for example, the stable rank of the second linear transformation is approximately 1.5% of the available rank, compared with roughly 13–14% for the first transformation. The second transformation also requires fewer singular directions to recover 95% of spectral energy and allocates a larger fraction of its energy to the leading directions.

Qwen2.5-VL follows a different merger architecture and therefore need not exhibit the same two-stage pattern. Nevertheless, its analyzed merger weights remain directionally concentrated under effective-rank and energy-dominance measures. The agreement between the complete spectra in the

main paper and these summary statistics indicates that the observation is not an artifact of a single spectral measure.

We emphasize that these diagnostics characterize the linear weights that constitute the interface rather than the complete nonlinear mapping experienced by each token. Accordingly, the analysis supports non-uniform transformation geometry but does not assign semantics or functional meaning to individual singular directions. Its role in the paper is to motivate a closer examination of the post-interface representations themselves, where PRISM is defined.

E Additional Decoder-Side Diagnostics

PRISM assigns roles before any decoder computation and does not use decoder-side signals during calibration or inference. The analyses in this section are therefore deliberately post hoc. Their purpose is to ask whether a partition recovered at the interface is reflected in later multimodal computation, not to define or supervise the roles themselves. We consider aggregate decoder attention and low-dimensional trajectories of decoder hidden states; neither analysis is used as the primary functional validation, which is provided by controlled intervention.

E.1 Aggregate Decoder Attention

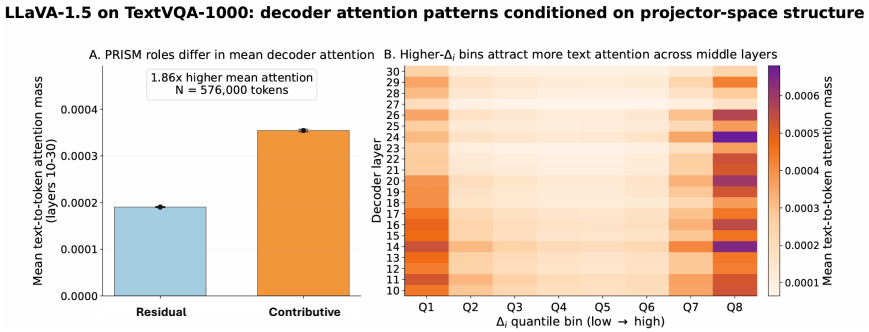

Figure 6: **Interface-level structure anticipates subsequent decoder attention.** We analyze LLaVA-1.5-7B on 1,000 TextVQA examples using text-to-visual attention from decoder layers 10–30. (A) PRISM contributive tokens receive $1.86\times$ higher aggregate mean attention than residual tokens over 576,000 visual tokens. (B) Tokens with larger paired interface-response contrast Δ_i also attract greater attention across intermediate decoder layers. Both the PRISM partition and Δ_i are computed before decoder traversal.

Contributive tokens receive $1.86\times$ higher aggregate text-to-visual attention than residual tokens on average. Likewise, larger values of Δ_i are associated with elevated subsequent attention across intermediate decoder layers. The important temporal ordering is that both Δ_i and the PRISM partition are computed before these decoder-attention values exist. The correlation therefore shows that interface-level structure is associated with a later decoder-side quantity, but it does not establish that attention causes the functional asymmetry or that PRISM merely reproduces an attention score.

For this reason, we do not use aggregate attention as functional validation. Attention can be correlated with many representation-level properties, and the two roles also differ in their native cardinalities. The count- and attention-matched intervention in the main paper and Appendix G explicitly removes this explanation by comparing role-matched tokens within decoder-attention strata.

E.2 Decoder-Space Representation Trajectories

For each evaluated model and selected decoder layer, we collect hidden states after the PRISM partition has already been assigned. Tokens are grouped into text stopwords, text content words, residual visual tokens, and contributive visual tokens. We project these hidden states onto their first two principal components solely for visualization; PCA does not participate in PRISM calibration or inference.

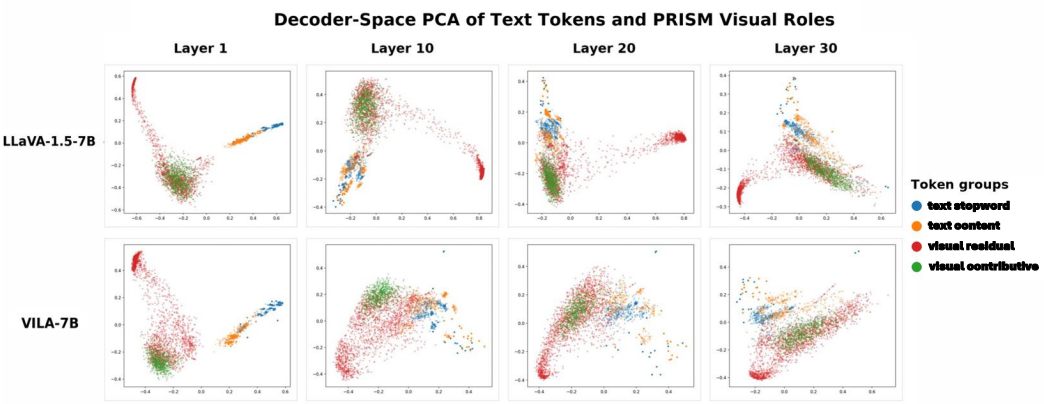

Figure 7: **Decoder-space PCA of text tokens and PRISM visual roles.** Across LLaVA-1.5-7B and VILA-7B, contributive visual tokens increasingly occupy regions associated with content-bearing text tokens across decoder depth, whereas residual tokens remain comparatively diffuse or peripheral. Roles are fixed before decoder traversal; decoder states are used only for post-hoc analysis.

In earlier decoder layers, visual and textual representations remain comparatively separated. With increasing depth, contributive visual tokens move toward regions occupied by content-bearing text tokens, whereas residual visual tokens remain more diffuse. The pattern is observed for both LLaVA-1.5 and VILA and is consistent with a pre-decoder distinction that remains visible after subsequent multimodal processing.

We do not interpret geometric proximity in a two-dimensional PCA projection as a semantic equivalence between visual and text tokens. The projection depends on the sampled examples and on the dominant variance directions of the selected hidden states. Accordingly, this analysis is qualitative supporting evidence about representation trajectories, not a standalone token-importance metric or a component of PRISM.

F Controls on Interface-Response Anchors and the Learned Representation

The main paper shows that PRISM extends beyond the scalar contrast used to construct its response anchors. This section separates several possible explanations for that result. First, we ask whether direct use of Δ_i itself can reproduce the contributive–residual functional asymmetry. Second, we repeat the comparison while preserving the exact quantiles used to define the response anchors, eliminating differences in effective selection budget. Third, we measure where PRISM-selected tokens originate relative to the two response-anchor sets, including the unassigned middle 60%. Finally, we compare the compact coordinate–rank rule with a dense linear classifier that receives the complete post-interface vector under the same anchor-set membership. Together, these controls distinguish the initial interface-response contrast from the representation and rule used by PRISM at inference.

F.1 Count-Matched Direct-Δ_i Intervention

We first test whether the functional distinction can be obtained simply by using the interface-response statistic itself. For each image, we form high- and low-Δ_i sets whose cardinalities match the corresponding PRISM contributive and residual sets. We then apply the same removal intervention. This comparison is intentionally favorable to the scalar baseline in terms of token count: any difference therefore reflects which tokens are selected, rather than how many tokens are removed.

PRISM preserves a consistent contributive–residual removal-effect gap across all metrics, whereas direct Δ_i is less consistent. In particular, selecting tokens directly from the scalar extremes does not reproduce the same two-sided functional separation obtained from the learned coordinate–rank

Table 9: **Count-matched direct-Δ_i removal versus PRISM role removal.**

Backbone	Setting	POPE F1	TextVQA	MME Total
LLaVA-1.5-7B	Full	85.85	58.30	1864.8
	PRISM contributive removed	75.34	47.10	1638.0
	Direct Δ_i high removed	78.91	47.10	1695.8
	PRISM residual removed	86.05	56.80	1781.6
	Direct Δ_i low removed	78.19	56.23	1718.0
VILA-7B	Full	86.25	62.20	1788.7
	PRISM contributive removed	72.83	52.00	1619.8
	Direct Δ_i high removed	79.89	54.34	1700.6
	PRISM residual removed	84.08	57.20	1713.4
	Direct Δ_i low removed	77.54	55.69	1673.7

partition. The result is compatible with Δ_i containing useful structural information, but shows that this information is insufficient when used directly as the final token-selection or role score. This is precisely the role assigned to Δ_i in PRISM: it defines the initial response anchors and is discarded after calibration.

F.2 Quantile-Preserving Direct-Δ_i Comparison

The previous comparison count-matches the PRISM role sizes and therefore changes the effective direct-Δ_i quantiles on images where the recovered role sizes differ from the calibration tails. We therefore perform a second comparison that preserves the original response-anchor quantiles exactly: the top 10% ($K = 58$) and bottom 30% ($K = 173$) for the 576-token LLaVA and VILA configurations. For PRISM, we sample the same number of tokens from the corresponding recovered role pool. Thus, both methods operate at the same fixed budget while the direct baseline uses exactly the regions defining $\mathcal{A}^+$ and $\mathcal{A}^-$.

Table 10: **Quantile-preserving fixed-budget comparison with direct Δ_i.** PRISM rows report mean $\pm$ standard deviation over seeds 42/43/44.

Backbone	Retained pool	TextVQA	POPE F1	MME-P	MME Total
LLaVA-1.5	Direct Δ_i top 10% ($K = 58$)	53.07	61.69	1250.7	1566.7
	PRISM contributive ($K = 58$)	52.71±0.17	81.99±0.39	1385.0±11.2	1707.5±33.3
	Direct Δ_i bottom 30% ($K = 173$)	45.15	66.76	1110.7	1434.3
	PRISM residual ($K = 173$)	46.25±0.16	66.78±0.66	1182.9±1.5	1490.6±20.3
VILA	Direct Δ_i top 10% ($K = 58$)	51.68	67.48	1177.8	1535.0
	PRISM contributive ($K = 58$)	53.18±0.57	80.27±0.28	1337.2±0.9	1658.0±8.9
	Direct Δ_i bottom 30% ($K = 173$)	47.56	70.03	1239.5	1570.5
	PRISM residual ($K = 173$)	50.32±0.19	70.07±0.13	1240.6±27.8	1558.1±27.2

At the original response-anchor quantiles, PRISM contributive tokens substantially outperform direct top-decile Δ_i retention on POPE and MME and remain similar or better on TextVQA for both backbones. The gap is therefore not an artifact of changing the effective quantile in the count-matched comparison. The residual-side comparison is intentionally reported as well: it shows that the relationship is not a simple inversion of the upper-response tail and reinforces the interpretation of residual as the unmatched complement of the learned rule rather than a direct bottom-Δ_i class.

F.3 Relation to the Response-Anchor Tails

Across 1,187 MME calibration images, a $K = 58$ PRISM contributive subset on LLaVA-1.5 contains $23.67 \pm 0.09\%$ direct top-decile tokens and $58.34 \pm 0.13\%$ tokens from the unassigned middle 60%. For VILA, the corresponding fractions are $28.21 \pm 0.14\%$ and $56.28 \pm 0.17\%$. Thus, contributive assignments are enriched for $\mathcal{A}^+$, as expected from the initial structural partition, but most selected contributive tokens do not come from the upper-response anchor set. More than half instead

originate from tokens that were unassigned during rule induction. This provides a direct measure of generalization beyond the scalar response anchors.

For residual subsets, $34.28 \pm 0.05\%$ on LLaVA-1.5 and $35.14 \pm 0.02\%$ on VILA lie in the direct bottom 30%. The residual role therefore should not be identified with $\mathcal{A}^-$ either. This asymmetry follows from the form of the learned rule: contributive requires a coordinate–rank rule match, whereas residual denotes the unmatched complement after the learned rule is applied to all tokens.

F.4 Dense Linear-Probe Control

For the dense-probe control reported in Table 3A of the main paper, we train a logistic-regression probe on the full 4096-dimensional post-interface vector using the same anchor-set membership, calibration examples, and visual tokens as PRISM. This baseline tests whether PRISM's result follows simply from giving any supervised classifier access to post-interface activations. Unlike PRISM's sparse coordinate–rank predicates, the probe can use every post-interface coordinate and its continuous value.

Table 11: **Dense-probe control at $K = 100$ on LLaVA-1.5-7B.** PRISM and the linear probe use identical anchor-set membership and calibration data. Random sampling within the PRISM contributive pool removes downstream FastV ranking as a confound.

Selector	TextVQA	POPE F1	MME-P
Linear probe	55.23	74.14	1407.3
PRISM + random	55.87 ± 0.10	$\mathbf{86.39 \pm 0.10}$	1430.2 ± 4.0
PRISM + FastV	**57.06**	83.01	**1491.5**

Despite receiving dense access to all 4096 post-interface coordinates, the linear probe underperforms the compact coordinate–rank rule on all three reported metrics. This control does not imply that dense representations are generally inferior; rather, under the same anchor-set membership, unrestricted linear access is not sufficient to reproduce the downstream behavior of the coordinate–rank partition. The PRISM+random row is especially important because it removes downstream FastV ranking: tokens are sampled uniformly from the contributive pool, yet the result still exceeds the dense probe. The advantage is therefore not solely introduced by the downstream attention-based selector.

G Attention-Matched Functional Intervention

The main paper reports the residual-minus-contributive score gaps from the matched intervention in Table 1B. Here we provide the matching protocol and the corresponding absolute benchmark scores.

Aggregate analysis in Appendix E.1 shows that contributive tokens receive higher decoder attention on average. That correlation creates an important alternative explanation: PRISM might simply identify tokens that the decoder will later attend to more strongly. To control for this possibility while also removing role-cardinality differences, we divide tokens into four text-to-visual attention-mass quartiles using decoder layers 10–30 and count-match contributive and residual tokens within each quartile. The intervention is then performed on these matched subsets rather than on the full native roles.

After matching, the direction of the attention difference reverses: the selected contributive tokens receive 47.2% lower mean attention than the matched residual tokens on LLaVA-1.5 and 55.9% lower mean attention on VILA. This creates a deliberately stringent comparison. If subsequent decoder-attention magnitude were sufficient to explain the functional effect, removing the higher-attention matched residual tokens should be at least as damaging.

Contributive-token removal remains substantially more damaging on every reported metric despite the reversed mean-attention ordering. On both backbones, residual removal leaves performance close to the full-token reference, whereas contributive removal produces large drops in POPE, TextVQA, and MME. Because the comparison matches role counts within attention strata, the result is not explained by removing more tokens from one role. Because the matched contributive subset

Table 12: **Count-matched intervention within decoder-attention quartiles.**

Backbone	Setting	POPE F1	TextVQA	MME Total
LLaVA-1.5	Full	85.85	58.30	1864.8
	Contributive removal	75.40	47.05	1636.0
	Residual removal	85.96	57.50	1850.9
VILA	Full	86.25	62.20	1788.7
	Contributive removal	72.66	52.81	1612.9
	Residual removal	85.37	59.10	1763.0

has lower mean decoder attention, it is also not reducible to subsequent decoder-attention magnitude. Together with the aggregate correlation in Appendix E.1, this shows that PRISM and decoder attention are related but not interchangeable signals.

H Cross-Dataset Rule Calibration

The default experiments calibrate one PRISM rule per backbone using MME images. A natural concern is that the recovered coordinate–rank structure might depend narrowly on that particular image source. We therefore recalibrate PRISM from scratch using TextVQA and POPE images and repeat the controlled intervention on LLaVA-1.5 and VILA. TextVQA calibration uses 1,187 unique images, matching the primary MME calibration set size, while POPE calibration uses 500 images. No downstream benchmark labels are used for rule induction.

The main paper summarizes the resulting residual-minus-contributive removal gaps in Table 3C. Here we report the corresponding absolute scores and compare the token assignments produced by rules learned from different calibration sources.

Table 13: **Controlled intervention under different calibration sources.**

Backbone	Calibration	Removed role	POPE	TextVQA	MME
LLaVA-1.5	MME	Contributive	75.34	47.10	1638.0
		Residual	86.05	56.80	1781.6
	TextVQA	Contributive	74.86	45.86	1603.7
		Residual	86.15	56.87	1806.5
	POPE	Contributive	74.77	48.39	1662.1
		Residual	85.57	56.04	1784.3
VILA	MME	Contributive	72.83	52.00	1619.8
		Residual	84.08	57.20	1713.4
	TextVQA	Contributive	72.98	52.69	1634.0
		Residual	83.69	57.79	1729.9
	POPE	Contributive	74.48	53.31	1675.7
		Residual	83.56	56.93	1721.2

Compared with MME calibration, contributive assignments obtained from TextVQA/POPE calibration have global Jaccard scores of 0.777/0.677 on LLaVA-1.5 and 0.836/0.772 on VILA. Intersection-over-smaller-set values range from 0.822 to 0.980. The rules are therefore not identical across calibration sources, which is expected because the response anchors are constructed from the image-wise Δ_i distribution of each calibration set. At the same time, the substantial overlap shows that a large portion of the recovered partition is stable.

The more important test is functional rather than combinatorial: contributive-token removal remains more damaging than residual-token removal for every calibration source and every reported metric on both backbones. Thus, the functional asymmetry does not require MME-specific calibration, even though the exact set of matched tokens can vary with the images used to induce the rule.

I Query-Agnostic Role Reuse Across Questions

Same-image protocol. For the same-image evaluation reported in the main paper, PRISM computes the role map and retained K-token mask once per image and reuses them unchanged across all questions associated with that image. We evaluate 5,000 VQAv2 (Goyal et al., 2017) questions over 883 images, of which 855 contain at least two question types. The role assignment and retained-token set are therefore fixed before any individual question is processed. Only the subsequent language-model computation changes with the question. This design isolates the property of interest: whether an image-level pre-decoder prior remains useful across different queries without recomputing visual-token roles.

Breakdown by question type. The fixed image-level mask remains effective across qualitatively different question categories. Table 14 reports the LLaVA-1.5 breakdown used to complement the aggregate same-image results in the main paper.

Table 14: **Same-image PRISM reuse across VQAv2 question types on LLaVA-1.5-7B.** The same image-level mask is reused without question-specific recomputation.

Type	N	Full	FastV	PRISM
Counting	660	57.33	53.03	**55.27**
Color	524	84.85	69.96	**80.97**
Yes/No	2000	91.48	87.31	**89.32**
Spatial	118	47.97	43.73	**47.12**
Object/Attr.	1439	67.30	64.52	**66.50**

Across all five reported categories, PRISM remains closer to full-token inference than FastV while reusing the same pre-decoder mask. The effect is therefore not confined to one question format such as yes/no or object-centric queries; it remains visible across counting, color, spatial, and object/attribute questions as well.

Interpretation and scope. Query-agnostic role assignment does not imply query-independent multimodal reasoning. PRISM provides an image-level prior before decoder processing, while the decoder continues to condition the retained visual evidence on each language query. The same-image experiment tests reuse of the prior, not invariance of the final model response to language context. The consistent gains across question types indicate that one pre-decoder partition can remain useful under varied queries without requiring question-specific role recomputation. This distinction is important: PRISM is intended to identify reusable structure in the visual representation, while query-specific reasoning remains the responsibility of subsequent decoder computation.

J Base-LLM High-Activation Axis Analysis

A potential alternative explanation is that PRISM simply rediscovers high-activation dimensions already present in the underlying language model. If so, the learned coordinate predicates might reflect generic dominant axes of the text model rather than structure associated with visual representations at the interface. We therefore compare PRISM contributive-rule coordinates with high-activation axes measured directly from the text-only Vicuna-7B backbone (Chiang et al., 2023).

Dimensions are ranked by token-RMS-normalized mean absolute activation. We separately form rankings from content tokens and stopwords so that the comparison is not tied to only one class of text tokens. Using 1,000 seed-42 MME image–question pairs, we perform a size-matched comparison against the 26 contributive-rule coordinates. The same analysis is repeated on 1,000 WikiText (Merity et al., 2016) chunks to avoid relying only on text paired with the multimodal evaluation set. As summarized in Table 3B of the main paper, the matched pre-decoder intersections are empty for both content-token and stopword-derived top-26 axes on both corpora.

We additionally examine whether coordinate overlap appears only after decoder computation has begun.

Table 15: **Contributive-coordinate overlap across text-representation depth.**

Text representation	Content/contributive	Stopword/contributive
Input embedding	0/26	0/26
Decoder layer 1	5/26	4/26
Decoder layers 16–31	4/26	5/26

Partial overlap appears only after decoder traversal. At the input stage, where the relevant inheritance hypothesis is most direct, none of the 26 PRISM coordinates appears in the size-matched content- or stopword-derived high-activation sets. The same zero-overlap result on both MME text and WikiText makes a corpus-specific coincidence less likely.

These results do not support simple inheritance of PRISM coordinates from the base LLM's dominant input-stage axes. The small overlaps observed after decoder computation has begun are interpreted more conservatively as depth-dependent representational correspondence. They do not show that the pre-decoder PRISM partition is inherited from generic language-model high-activation dimensions.

K Full Fixed-Budget Token-Selection Results

K.1 FastV

The main paper reports the retained-performance summary in Table 4B. Here we provide the complete benchmark-level results. Standard FastV (Chen et al., 2024) ranks the full visual-token set using decoder attention. For the PRISM variants, we change only the candidate pool: the same FastV ranking is applied after restricting candidates to either contributive or residual tokens. The final number of retained tokens is unchanged, and all methods use the same approximately 10% visual-token budget. This isolates whether the pre-decoder role partition provides useful information before the downstream selector is applied.

Table 16: **Full fixed-budget token-selection results with FastV.**

Method	TextVQA	GQA	SQA^{IMG}	MME	POPE	MMMU	MMB^{EN}	MMB^{CN}	Rel.
LLaVA-1.5-7B									
FastV	54.7	51.2	69.9	1539.4	57.4	35.2	58.7	49.8	88.3
Residual	45.6	49.7	68.2	1325.0	54.8	33.7	47.9	37.5	78.5
Contributive	55.0	56.0	68.4	1731.8	80.5	35.6	61.2	50.3	**94.4**
VILA-7B									
FastV	55.7	53.4	71.4	1568.3	70.2	37.3	63.1	51.7	91.8
Residual	49.3	49.3	66.9	1482.2	61.3	35.3	55.2	43.1	83.0
Contributive	55.7	55.9	70.3	1688.6	78.1	35.0	64.9	51.7	**93.7**
LLaVA-NeXT-7B									
FastV	54.0	55.0	67.8	1608.3	74.2	34.7	62.3	54.2	90.9
Residual	50.2	55.8	67.9	1666.2	76.8	34.8	60.9	51.6	90.3
Contributive	55.0	58.9	67.0	1726.3	80.9	36.8	63.1	54.9	**94.5**
Qwen2.5-VL-7B									
FastV	70.8	51.2	80.9	2073.6	78.1	47.8	76.2	72.5	90.3
Residual	46.0	54.7	83.0	1824.5	78.3	47.6	75.5	71.8	85.7
Contributive	72.2	52.8	82.8	2143.5	80.6	48.7	77.3	74.1	**92.5**

Restricting FastV to contributive candidates improves the aggregate retained performance on every evaluated backbone, whereas residual-restricted selection is consistently weaker. The gain is not required to appear on every individual benchmark: different tasks can benefit from different visual

evidence, and the role prior is deliberately query-agnostic. The consistent aggregate pattern across four architectures is the relevant result. Since the final token budget and downstream ranking criterion are unchanged, the difference isolates the utility of the interface-level role prior as a candidate filter rather than a replacement scoring function.

K.2 DivPrune

DivPrune (Alvar et al., 2025) provides a complementary test because its selection criterion is diversity-oriented rather than decoder-attention-based. A hard role restriction could interfere directly with that coverage objective, so we use a softer role-biased allocation instead: 80% of the token budget is allocated to one role and the remaining 20% to the other. DivPrune then performs its usual selection within the resulting allocation. The comparison therefore asks whether favoring one PRISM role is useful even when the downstream selector is based on a qualitatively different principle.

Table 17: **Fixed-budget role-biased DivPrune results.**

Method	TextVQA	GQA	SQA^{IMG}	MME	POPE	MMMU	MMB^{EN}	MMB^{CN}	Rel.
LLaVA-1.5-7B									
DivPrune	54.7	57.6	68.4	1619.7	85.3	34.0	59.1	48.9	93.3
80% Residual	52.4	56.5	68.1	1603.8	81.4	35.1	59.5	47.8	92.1
80% Contributive	56.0	57.8	68.6	1663.5	85.4	35.6	59.9	49.9	**94.9**
VILA-7B									
DivPrune	56.6	56.9	69.5	1610.6	83.3	36.1	63.0	49.4	93.6
80% Residual	57.0	56.9	69.7	1630.8	81.5	35.7	64.1	50.2	93.8
80% Contributive	56.9	56.8	68.9	1682.8	82.8	34.9	63.5	50.2	**93.8**
LLaVA-NeXT-7B									
DivPrune	53.6	60.3	67.7	1730.4	84.1	36.2	64.0	56.4	95.4
80% Residual	53.6	59.5	67.4	1768.9	82.0	36.8	64.7	55.2	95.3
80% Contributive	55.7	60.4	67.3	1725.6	84.3	37.3	64.3	55.2	**96.0**
Qwen2.5-VL-7B									
DivPrune	60.8	54.2	75.8	2018.4	81.9	43.7	74.0	73.3	87.6
80% Residual	48.9	54.2	74.3	1870.1	80.5	42.3	73.2	72.9	83.9
80% Contributive	60.4	54.7	75.4	1976.4	82.2	44.7	73.4	74.3	**87.7**

The role effect is smaller under the diversity-oriented selector than under FastV, which is reasonable because DivPrune explicitly preserves coverage and the 80/20 allocation never removes either role entirely. Nevertheless, increasing the share allocated to contributive tokens is generally preferable to increasing the residual share across the evaluated backbones. The VILA aggregate retained-performance values are effectively tied, illustrating that the prior need not improve every selector/model combination by a large margin. Overall, the result supports PRISM as an early candidate prior that can complement different downstream selection rules rather than as a replacement for one particular pruning algorithm.

L Supplementary Application: Decode-Time Reweighting

As a supplementary application, we test whether the pre-decoder role partition can guide visual-attention allocation during generation without removing visual tokens. This experiment is not used to establish the core functional distinction. Let $A_{t,j}^{(\ell,h)}$ denote the attention from generated text position t to visual token j at decoder layer ℓ and head h. For the contributive-token set S_{contrib}, we define

$$\widetilde{A}_{t,j}^{(\ell,h)} = \begin{cases} \gamma A_{t,j}^{(\ell,h)}, & j \in S_{\text{contrib}}, \\ A_{t,j}^{(\ell,h)}, & j \in S_{\text{vis}} \setminus S_{\text{contrib}}. \end{cases} \tag{11}$$

We then renormalize only over the visual-token span:

$$A'^{(\ell,h)}_{t,j} = \widetilde{A}^{(\ell,h)}_{t,j} \frac{\sum_{k\in S_{\text{vis}}} A^{(\ell,h)}_{t,k}}{\sum_{k\in S_{\text{vis}}} \widetilde{A}^{(\ell,h)}_{t,k} + \epsilon}. \tag{12}$$

Attention outside the visual-token span remains unchanged. The intervention is applied only during autoregressive decoding; the prefill pass, model parameters, and complete visual-token sequence remain unchanged. We use a fixed $\gamma = 1.2$ for all evaluated backbones. Consequently, the procedure preserves total visual attention mass and changes only its allocation among visual tokens. We evaluate this intervention using CHAIR (Rohrbach et al., 2018), which measures object hallucination in generated image descriptions.

Table 18: **Role-aware decode-time reweighting on CHAIR.** Lower is better for both metrics.

Model	Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$
LLaVA-1.5	Base	51.00	14.14
	PRISM	**50.00**	**13.55**
LLaVA-NeXT	Base	33.20	9.72
	PRISM	**31.60**	**8.92**
Qwen2.5-VL	Base	34.80	10.18
	PRISM	**28.60**	**9.59**

Both CHAIR metrics improve for all three evaluated backbones. Because no token is removed and total visual attention mass is preserved, the gains cannot be attributed to a shorter visual sequence or a larger total amount of visual attention. We treat this experiment as supplementary evidence that the recovered role prior can also be used to redistribute visual evidence during generation, rather than as part of PRISM's core functional validation.

M Additional Scalar Interface-Space Baselines

We further compare PRISM against scalar alternatives computed at the vision–language interface. These baselines test whether a single per-token magnitude is sufficient for the fixed-budget selection behavior attributed to coordinate identity and rank. Top-Δ_i and bottom-Δ_i rank tokens directly by the paired response contrast, while post-norm and post-concentration use scalar statistics of the post-interface representation itself. Random selection provides a budget-matched reference. All methods retain the same $K = 100$ tokens on LLaVA-1.5-7B.

Table 19: **Scalar interface-space baselines on LLaVA-1.5-7B at $K = 100$.**

Selector	TextVQA	MME-P	POPE F1
PRISM	**57.06**	**1491.5**	**83.01**
Top-Δ_i	55.60	1398.7	69.56
Random	50.43	1398.4	77.22
Bottom-Δ_i	43.67	982.7	51.94
Top post-norm	43.02	664.2	17.28
Top post-concentration	42.42	730.3	22.66

Top-Δ_i is the strongest scalar alternative among those evaluated on TextVQA and MME-P, while random selection is higher than Top-Δ_i on POPE. Thus, the paired interface-response contrast contains useful information but does not define a consistently strong token ordering by itself. Post-interface norm and concentration are substantially weaker, showing that large activation magnitude or concentration alone does not recover the same candidate set.

These results are consistent with the motivation for the coordinate–rank representation. A scalar preserves only one magnitude per token and discards which coordinates are responsible for that value. PRISM instead retains the identities and relative ranks of dominant coordinates and uses Δ_i only to define the initial response anchors. The scalar baselines therefore provide an additional control separating the paired response statistic from the representation used for final role assignment.

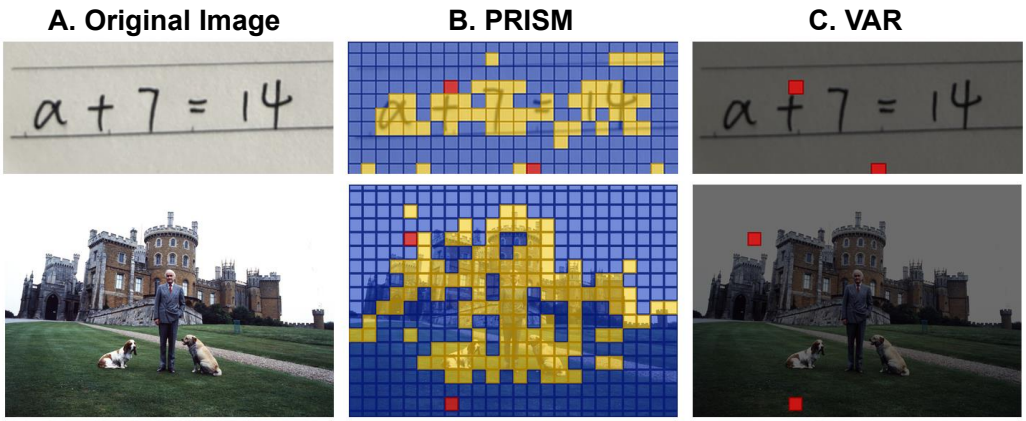

Figure 8: **Qualitative sink-aware PRISM assignments.** Column A shows the input image, Column B shows the PRISM contributive, residual, and optional sink assignments, and Column C shows sink tokens identified by the external activation-based criterion. The overlays are diagnostic visualizations and should not be interpreted as semantic segmentation masks.

N Model-Specific Sink-Aware Extension

The default PRISM formulation is the binary contributive–residual partition used throughout the main paper. We additionally examine whether the same post-interface coordinate–rank representation can recover an externally defined visual-sink state when a reproducible sink specification is available.

Scope and reproducibility. We evaluate this extension on LLaVA-1.5-7B, for which an activation-based sink criterion, including the relevant hidden dimensions and threshold, is explicitly specified in prior work (Kang et al., 2025). A comparably reproducible, backbone-specific criterion is not publicly available for the other evaluated models. Rather than transfer LLaVA-specific dimensions or introduce new model-dependent thresholds, we therefore restrict quantitative sink recovery to the backbone for which the external definition can be reproduced exactly.

This restriction concerns only the availability of an external sink definition, not the PRISM machinery itself. Given a reproducible backbone-specific sink labeling criterion, the same coordinate–rank rule-induction procedure can in principle be applied to other architectures without changing the core method. Default PRISM requires no sink supervision, and we make no claim about the presence or absence of analogous sink states in backbones for which such labels are unavailable.

External sink criterion. For representation $\mathbf{x} \in \mathbb{R}^d$ and sink-dimension set $\mathcal{D}_{\text{sink}}$, the external criterion is

$$\phi(\mathbf{x}) = \max_{m \in \mathcal{D}_{\text{sink}}} \left| \frac{x[m]}{\sqrt{\frac{1}{d}\sum_{j=1}^{d} x[j]^2}} \right|. \tag{13}$$

For LLaVA-1.5-7B,

$$\mathcal{D}_{\text{sink}} = \{1415, 2533\}, \qquad \tau_{\text{sink}} = 20. \tag{14}$$

These dimensions and threshold are taken directly from the external specification and are not tuned by PRISM.

Sink-aware role assignment. We learn a sink rule $\mathcal{R}_{\text{sink}}$ using the same coordinate–rank representation as the default method and evaluate it before the contributive rule:

$$g_{\text{sink}}(\mathbf{h}_i) = \begin{cases} \text{sink}, & R(\mathbf{h}_i; \mathcal{R}_{\text{sink}}) = 1, \\ \text{contributive}, & R(\mathbf{h}_i; \mathcal{R}_{\text{sink}}) = 0 \wedge R(\mathbf{h}_i; \mathcal{R}_{\text{contrib}}) = 1, \\ \text{residual}, & \text{otherwise}. \end{cases} \tag{15}$$

The learned LLaVA-1.5 extension contains 2 sink predicates and 23 contributive predicates.

Table 20: **Recovery of externally specified visual-sink states from post-interface coordinate–rank signatures on LLaVA-1.5-7B.**

Metric	Value
Evaluation tokens	216,093
Sink-label accuracy	99.47%
Sink recall	98.60%
Sink F1	83.56%
Tri-class calibration accuracy	81.94%

The 98.60% sink recall shows that the same post-interface coordinate–rank representation can recover the externally specified sink state before decoder attention is evaluated. More broadly, this result suggests that PRISM is not inherently restricted to the default binary partition: when a reproducible external label for an additional role is available, the same coordinate–rank representation and rule-induction framework can be used to recover that role without modifying the core method. We therefore treat the sink experiment as evidence of representational extensibility rather than as part of the core binary formulation, since the additional supervision depends on an external, model-specific criterion.

O Role Composition Across Models and Datasets

PRISM does not impose a fixed contributive-token ratio at inference. The number of matched tokens is determined by whether each post-interface coordinate–rank signature satisfies the learned rule, so role cardinalities can vary across architectures and image distributions. Table 21 reports the resulting average counts on the three primary evaluation datasets. These statistics are useful for interpreting both complete role removal and fixed-budget selection: the two native roles should not be assumed to contain the same number of tokens.

Table 21: **Average PRISM role composition across evaluation datasets.** Counts reflect native visual-token sequences under benchmark-specific preprocessing, not the standardized sequence lengths used only for fixed-budget selection.

Role	LLaVA-1.5			VILA			LLaVA-NeXT			Qwen2.5-VL		
	POPE	TextVQA	MME	POPE	TextVQA	MME	POPE	TextVQA	MME	POPE	TextVQA	MME
Contributive	146.5	148.5	145.9	154.7	142.9	136.8	1007.1	1047.0	961.5	219.6	622.4	468.1
Residual	429.5	427.5	430.1	421.3	433.1	439.2	1206.5	1224.3	1110.6	140.1	351.9	408.1

The role composition differs substantially across model families. LLaVA-1.5 and VILA assign a minority of their 576 visual tokens to the contributive role on average, whereas the higher-resolution LLaVA-NeXT representation contains many more tokens in both roles. Qwen2.5-VL also shows noticeable dataset-dependent variation in the average role counts. This variation is expected because PRISM applies a learned matching rule rather than enforcing a predetermined percentage at inference.

The residual role still constitutes a substantial portion of the visual sequence in most settings. Consequently, removing the entire residual set can eliminate considerable contextual evidence even if the average functional sensitivity of those tokens is lower. This is why the main paper complements complete-role removal with count-matched interventions. The contributive–residual terminology should therefore be read as a relative functional asymmetry under the learned interface-level partition, not as a binary distinction between useful and useless visual evidence.