CSRA: DECOMPOSING NATIVE VISUAL QK ALIGNMENT IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual-language understanding in vision–language models (VLMs) relies on effective cross-modal alignment. However, existing analyses based on attention maps primarily reveal where models attend, offering limited insight into how alignment forms and where failures originate. We show that cross-modal alignment is directly encoded in native query-key (QK) interactions and distinguish three sources of alignment failure: loss of textual semantic information during query projection, insufficient discriminability of visual target keys, and incorrect matching between queries and keys.Further analysis shows that even when queries preserve semantic structure and keys support target discrimination, their interactions can be dominated by a low-rank shared subspace associated with attention sinks, leading to deviations from semantic alignment. Based on this finding, we propose Common-Space Residual Attribution (CSRA), using singular value decomposition (SVD) to identify and remove the projection components responsible for this interference in QK interactions, thereby suppressing the influence of the sink subspace.Experiments on Qwen2.5-VL, LLaVA-1.5, and InternVL3 demonstrate improved target localization, with RefCOCO mIoU gains of up to 27.1 % over native attention. The estimated subspace also exhibits cross-dataset transferability. Our findings reveal how QK interactions shape cross-modal alignment and provide a unified framework for diagnosing potential alignment failures and guiding mechanism-driven alignment enhancement.

1 INTRODUCTION

Vision–language models (VLMs) have made rapid progress in visual–language understanding and reasoning, where cross-modal alignment provides the basis for relating textual concepts to image content (Liu et al., 2023b;a; Bai et al., 2025; Wang et al., 2025). Existing studies analyze this alignment using attention maps: they extract attention distributions over image regions conditioned on text tokens and visualize these distributions to assess whether the tokens attend to semantically corresponding visual regions (Abnar & Zuidema, 2020; Chefer et al., 2020). Such maps provide useful spatial readouts, yet leave a more fundamental question unresolved: how models align different modalities during inference and at which inference stage misalignment arises. Without a thorough understanding of the mechanisms underlying alignment, existing methods are limited in their ability to diagnose and improve alignment.

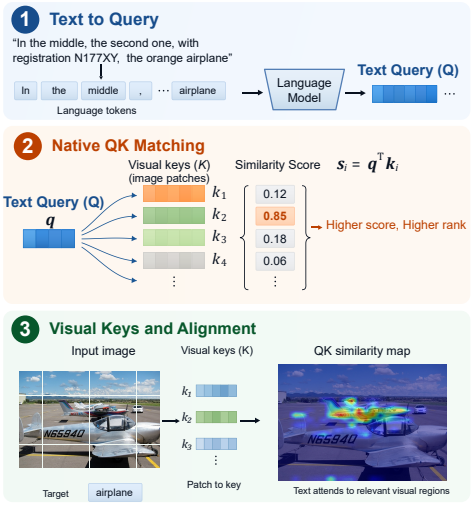


Figure 1: Native QK interactions encode cross-modal alignment.

In this paper, we show that native query–key (QK) interactions contain extractable cross-modal alignment signals. Mainstream generative VLMs employ multi-head attention(MHA), projecting the hidden state of target token into lower-dimensional queries across multiple heads and computing QK dot products. The query is conditioned on the textual context, while visual keys encode features

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.

of candidate image regions. Their scaled dot products determine the relative ranking of these regions and reflect visualtextual relevance. Visual attention sinks manifest as excessive attention to visual tokens unrelated to the query, causing apparent misalignment in raw attention maps (Kang et al., 2025).

Based on this observation, we further identify three potential sources of cross-modal misalignment. First, W_q projects the hidden states of target text token into a much lower-dimensional query space. Therefore, the quality of the query reflects the model's ability to encode and preserve textual semantics. Second, the discriminability of target keys determines the model's ability to accurately identify target regions in the visual representation space. Third, even when queries and keys encode similar semantics, anomalous alignment along certain directions may induce attention sinks and impair alignment in attention maps. We therefore hypothesize that native QK maps reflect two geometric components: a common component shared across targets which induces sink-like responses, and a target-dependent residual component which preserves the spatial variation required for target-selective alignment. This common–residual decomposition provides a unified view of attention sinks and extractable alignment signal in native QK interactions.

We further compare semantic structure before and after the W_Q projection using representational similarity analysis (RSA) and spectrum-matched random projections, then assess the separability of visual keys through target-prototype similarity and ridge-LDA. To study QK interaction, we fix visual keys and estimate a common subspace from specific tokens,typically the first token after the visual prefix (<vision_end> in Qwen2.5-VL). This token precedes the target text, so causal masking prevents target text token sequence from affecting it. For each layer and head, we use singular value decomposition (SVD) to identify the directions of these special-token queries as a basis for the common space. We extract the component of the target query orthogonal to the common space and compute Common-Space Residual Attribution (CSRA) along the native QK path, suppressing interference from sink-like common responses in cross-modal alignment. We then evaluate whether this readout localizes the target.

Experiments conducted on Qwen2.5-VL, LLaVA-1.5, and InternVL3 demonstrate that common-space and boundary-probe maps have Pearson correlations of 0.761–0.959. After common-space removal, the original dominant background peaks were reduced by 83.8%–96.6%. At the 50% layers, CSRA improves RefCOCO mIoU by 16.1 % on Qwen2.5-VL and 27.1 points on LLaVA-1.5. Regarding the low-rank subspace extracted from RefCOCO, after transferring it to COCO Caption, the best results were achieved across multiple backbones and various intermediate layers. These findings not only reveal the internal mechanisms of cross-modal alignment in multimodal models but also provide a new analytical framework for diagnosing different types of alignment failure and guiding targeted, mechanism-driven alignment enhancement.

Our contributions are:

1. We formulate query semantic retention, visual-key target structure, and QK spatial compatibility as an empirical framework for native attention alignment.
2. We establish query-side category structure and key-side target accessibility through complementary geometric probes.
3. We identify a low-rank common query subspace and derive an exact common–residual decomposition of native QK logits.
4. We introduce CSRA and demonstrate improved localization and cross-dataset transfer as diagnostics of native alignment.

2 Related Work

2.1 Cross-Modal Alignment in Vision-Language Models

Cross-modal alignment connects textual concepts to visual content. CLIP learns visual representations through image–text matching (Radford et al., 2021), while FILIP uses word–patch interactions to guide fine-grained contrastive pretraining (Yao et al., 2021). For trained models, CLIP Surgery studies inconsistent visual responses and redundant features, and modifies the inference architecture to improve visual explanations (Li et al., 2023). Attention rollout aggregates attention across

layers (Abnar & Zuidema, 2020), while Transformer relevance-propagation methods include conservative propagation (CP-LRP) (Chefer et al., 2020; Ali et al., 2022). In multimodal LLMs, Token Activation Map (TAM) estimates interference from preceding context and applies rank Gaussian filtering to activation maps (Li et al., 2025). These methods provide ways to inspect visual evidence. Interpreting such evidence requires specifying whether the goal is spatial correspondence or explaining a final prediction (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Therefore, our research focuses on the alignment signal encoded in the native QK scores.

2.2 Representation Learning and Semantic Probing

A central question in representation learning is which distinctions are captured in learned features. Studies of contextualized word representations characterize context dependence and anisotropy (Ethayarajh, 2019), while All-but-the-Top improves word representations by removing the common mean and dominant directions (Mu et al., 2017). Linear probes test what information is accessible from intermediate representations (Alain & Bengio, 2016); control tasks help distinguish encoded structure from what a probe can learn itself (Hewitt & Liang, 2019). Representational similarity analysis complements these approaches by comparing relational structure across representations (Kriegeskorte et al., 2008). Singular value decomposition (SVD) of the querykey projection interaction matrix enables the analysis of interactions between representations in ViT Pan et al. (2024).

2.3 Attention Sinks in Multimodal Models

Attention sinks show that concentrated responses need not track semantic relevance (Xiao et al., 2023). Studies of massive activations and high-norm visual tokens further characterize concentrated attention and internal computation (Sun et al., 2024; Darcet et al., 2023). Spectral analysis has also connected attention sinks to directions in embedding and unembedding spaces (Cancedda, 2024). In multimodal models, Visual Attention Redistribution (VAR) reallocates attention from sink tokens within selected image-centric heads (Kang et al., 2025), whereas Enhancing Vision Attention Sinks (EVAS) strengthens selected shallow-layer visual sink patterns to mitigate hallucination (Zhang et al., 2025). These findings motivate evaluating the role of a concentrated response in its specific computational setting.

3 Semantic Readability in QK Geometry

3.1 Semantic Structure After Query Projection

Along the native QK interaction path, we first analyze the geometric structure of the query representation space. Since $W_Q^{(h)}$ maps a high-dimensional hidden state to a much smaller query vector, we examine whether this reduction preserves semantic category structure and relationships among term representations. We probe models on PartImageNet He et al. (2021) using a balanced set of 48 terms from six categories. For each image, we extract the hidden state at the final token of each term and compare its semantic structure with that of the per-head query projection.

Let x_p be the normalized hidden state entering W_Q for prompt p, and let $y_p = W_Q^{(h)} x_p$ be its per-head projection. For each image, we construct the pairwise Euclidean-distance RDMs D_x and D_y from $\{x_p\}$ and $\{y_p\}$, and define the category reference $R_{pq} = \mathbf{1}[\mathrm{category}(p) \neq \mathrm{category}(q)]$. Hidden RSA measures the Spearman correlation between D_x and R, while query RSA measures the correlation between D_y and R, quantifying category structure before and after projection. Distance RSA measures the Spearman correlation between D_x and D_y. All correlations use corresponding upper-triangular entries, excluding the diagonal. The kNN overlap measures the average fraction of nearest neighbors shared across the two spaces. A spectrum-matched null randomizes the input singular vectors of $W_Q^{(h)}$ while preserving its singular values. Appendix A provides the detailed definitions and reference values.

Table 1 shows category structure in both spaces: query RSA is 0.396–0.434, compared with hidden RSA of 0.449–0.509. Distance RSA of 0.628–0.687 and neighborhood overlap of 0.567–0.600 show that projection retains substantial relational structure. The spectrum-matched reference yields

Backbone	Hidden RSA	Query RSA	Distance RSA	kNN overlap
Qwen2.5-VL-7B	0.449	0.396	0.687	0.567
InternVL3-8B	0.509	0.434	0.685	0.590
LLaVA-1.5-7B	0.501	0.406	0.628	0.600

Table 1: Query-projection geometry aggregated over images, layers, and heads. Semantic RSA compares representation distances with category reference; distance RSA and kNN overlap compare hidden and query geometry. RSA score $\in$[-1,1], kNN overlap $\in$ [0, 1]. Higher values indicate stronger category-related structure or greater geometric preservation.

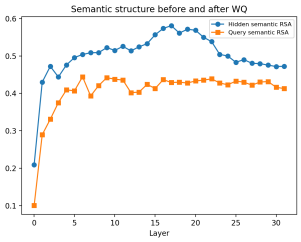


Figure 2: Layer-wise RSA scores for LLaVA-1.5-7B.

similar category RSA (Appendix A.4). The central result is that semantic distinctions survive per-head compression and remain available to the QK computation. We next examine whether visual keys support corresponding target distinctions.

3.2 Semantic Structure in Visual Keys

We examine target-related semantic structure in visual keys through two properties: similarity among keys from the same target region and linear separability between target and background keys.

Similarity among visual keys. Let $\mathcal{V}$ denote the visual-token set, $\Omega \subset \mathcal{V}$ the tokens in the annotated target region, and $\Omega^c = \mathcal{V} \setminus \Omega$ the background tokens. We use unit-normalized pre-RoPE keys, $\widehat{k}_i = k_i/\|k_i\|_2$, and assign each token its average cosine similarity to keys in the target region. For target tokens, we score each key by its similarity:

$$c_i = \widehat{k}_i^\top \left(\frac{1}{|\Omega_{-i}|} \sum_{t \in \Omega_{-i}} \widehat{k}_t \right), \qquad \Omega_{-i} = \begin{cases} \Omega \setminus \{i\}, & i \in \Omega, \\ \Omega, & i \notin \Omega. \end{cases} \tag{1}$$

This nonparametric score summarizes the target columns of the pairwise cosine-similarity matrix, excluding self-similarity. We use AUC to assess whether target tokens receive higher c_i scores than background tokens. AUC measures the fraction of target–background token pairs in which the target token scores higher. A high AUC indicates greater similarity among target keys than between target and background keys. We compute AUC separately for each sample, layer, and KV head, then report the mean within each layer.

Backbone	InternVL3-8B			Qwen2.5-VL-7B			LLaVA-1.5-7B		
Layer	14	19	27	14	19	27	14	22	31
$\text{AUC}_{\text{target}}$	0.936	0.921	0.892	0.927	0.912	0.901	0.906	0.861	0.804
$\text{AUC}_{\text{global}}$	0.948	0.937	0.930	0.942	0.937	0.922	0.935	0.909	0.886
AUC_{norm}	0.584	0.583	0.585	0.591	0.587	0.584	0.541	0.526	0.576

Table 2: AUC for distinguishing annotated target tokens from background tokens on RefCOCO. $\text{AUC}_{\text{target}}$ uses average cosine similarity to keys in the target region. $\text{AUC}_{\text{global}}$ uses average cosine similarity to all other visual keys in global region. AUC_{norm} uses key norms. Additional controls appear in Appendix B.

Target-region similarity distinguishes target from background keys, with AUCs of 0.804–0.936. Global similarity also achieves high AUCs of 0.886–0.948, whereas key norms alone yield only 0.526–0.591. Both similarity measures use unit-normalized keys, so these results support target-related structure in key directions that key magnitude alone does not capture as effectively. These observations support target-related directional structure beyond global centrality.

Linear separability of target and background keys. We test whether a single linear direction can distinguish target keys from background keys. For each example, layer, and KV head, we use

the target annotation to divide the visual keys into target and background groups, Ω and Ω^c. We use the native post-RoPE keys $k_i \in \mathbb{R}^{d_h}$ without additional unit-length normalization, keeping the representations used by attention. We compute the mean of each group:

$$\mu_C = \frac{1}{|C|} \sum_{i \in C} k_i, \qquad C \in \{\Omega, \Omega^c\}. \tag{2}$$

Let $\widehat{\Sigma}_C$ denote the sample covariance of keys within group C. We average the two covariances with equal weight, $\Sigma_w = (\widehat{\Sigma}_\Omega + \widehat{\Sigma}_{\Omega^c})/2$. Ridge-regularized linear discriminant analysis uses this covariance to find a direction with a large difference between the two group means relative to the variation within each group (Fisher, 1936). Adding λI_{d_h} stabilizes the covariance inverse. We then compute the ridge-LDA direction:

$$w = (\Sigma_w + \lambda I_{d_h})^{-1}(\mu_\Omega - \mu_{\Omega^c}), \qquad v^\star = \frac{w}{\|w\|_2}, \qquad s_i = (v^\star)^\top k_i. \tag{3}$$

The score s_i measures each key's projection onto this direction and is used to compute AUC and average precision (AP).

For comparisons with native attention, we instead use $q^\star = \|q\|_2 v^\star$. This replaces the native query's direction while preserving its norm and keeping the visual keys fixed, so that changes in the attention distribution reflect the change in direction. For Qwen2.5-VL-7B, the readout attains mean AUC 0.998 and AP 0.977. These results show that the fixed visual keys admit a linear direction that distinguishes the annotated target from its background. The direction is constructed and evaluated on the same labeled key field, providing a constructive, annotation-conditioned test of within-image linear accessibility. Detailed results are reported in the appendix B.5.

3.3 From Semantic Structure to Reliable Alignment

The preceding analyses show that queries preserve semantic distinctions and that visual keys contain target-related structure. Reliable alignment additionally requires their interaction to assign higher scores to the referenced region than to the background. The linear probe illustrates this distinction: changing the query direction using target annotations can improve localization while leaving the visual keys unchanged. The presence of semantic structure therefore does not ensure that the native query uses it effectively.

We next examine how different components of the native query contribute to visual distribution. Specifically, we test whether a common contribution obscures the correspondence between the target text and its image region. Separating this contribution allows us to assess whether the residual scores provide more reliable target alignment.

4 Common-Space Residual Attribution (CSRA)

CSRA forms a spatial alignment map from the residual contribution to native QK scores. It separates a target query into common and residual components, scores them against the original visual keys. Semantically structured queries and discriminative keys can still give poor target region scores if shared contribution dominates. We examine how the common query component affects the scores assigned to target and background regions, and whether removing its contribution improves target localization. We compute the common and residual scores using the same visual keys. Section 5 evaluates the common responses and the localization accuracy of the residual maps.

4.1 Query-Side Probe and Shared Subspace Estimation

For each layer ℓ and query head h, we extract the query at the first token following the visual sequence and preceding the target text, such as <vision_end> in Qwen2.5-VL. This boundary position that determined by the input template,receives the image and preceding context, while causal masking prevents access to later target words.Its role is to probe recurring query directions independently of those words. The semantic query probe in Section 3.1 instead receives the term being compared.

We capture queries after the normalization and RoPE operations used by attention. Let N be the number of calibration queries, indexed by $n = 1, \ldots, N$, and $\widehat{q}^{\ell,h}_{\mathrm{probe},n} = q^{\ell,h}_{\mathrm{probe},n} / \|q^{\ell,h}_{\mathrm{probe},n}\|_2$ their

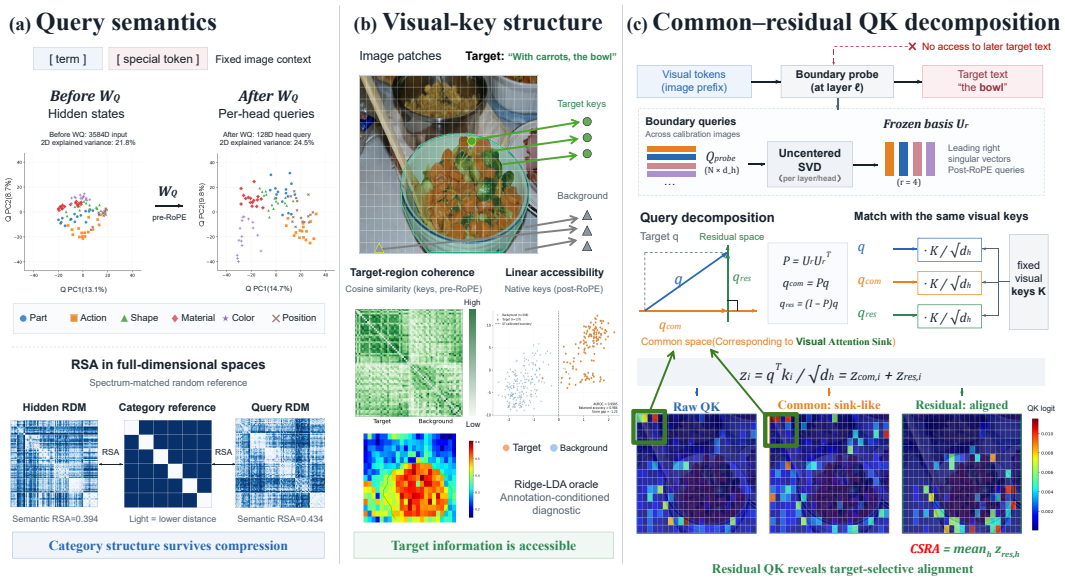


Figure 3: Overview of CSRA: From Query-Key Geometry to Cross-Modal Alignment.

unit-normalized vectors. Stacking them as rows gives $Q^{\ell,h}_{\text{probe}} \in \mathbb{R}^{N\times d_h}$ and the uncentered SVD

$$Q^{\ell,h}_{\text{probe}} = \begin{bmatrix} (\widehat{q}^{\ell,h}_{\text{probe},1})^\top \\ \vdots \\ (\widehat{q}^{\ell,h}_{\text{probe},N})^\top \end{bmatrix} = L\Sigma V^\top, \qquad U_{\ell,h} = V_{[:,1:r]}, \quad P_{\ell,h} = U_{\ell,h}U^\top_{\ell,h}. \tag{4}$$

Among subspaces of the same dimension, the selected subspace maximizes the total squared length of the projected queries. Using uncentered SVD keeps the mean query in this calculation. We then keep each basis fixed and use it for the remaining examples.

Although the basis is fixed, a target query's projection onto it can vary across inputs, and the resulting scores also depend on the visual keys. The common component can therefore produce different spatial maps for different inputs. Under grouped-query attention, each query head uses its paired KV head.

4.2 Orthogonal Common–Residual Decomposition of Native QK Logits

For a target query at position j, fix layer ℓ and query head h, and write $q = q^{\ell,h}_j$, $k_i = k^{\ell,g(h)}_i$, and $P = P_{\ell,h}$. Here $i \in \mathcal{V}$ indexes visual tokens, and $g(h)$ maps the query head to its corresponding KV head. Queries and keys are taken in the post-normalization, post-RoPE coordinates used by attention. The projector decomposes the query as

$$q = q_{\text{com}} + q_{\text{res}}, \qquad q_{\text{com}} = Pq, \qquad q_{\text{res}} = (I-P)q. \tag{5}$$

Keeping the visual keys fixed, linearity of the dot product gives

$$z_i = \frac{\langle q, k_i\rangle}{\sqrt{d_h}} = \frac{\langle Pq, k_i\rangle}{\sqrt{d_h}} + \frac{\langle (I-P)q, k_i\rangle}{\sqrt{d_h}} = z_{\text{com},i} + z_{\text{res},i}, \tag{6}$$

where $z_i = z^{\ell,h}_{j,i}$. The two query components retain their projected scales in this computation.

The SVD basis satisfies $U^\top_{\ell,h}U_{\ell,h} = I$, so $P = P^\top = P^2$ and $P^\top(I-P) = 0$. These orthogonality identities further imply

$$\begin{aligned} \langle Pq, k_i\rangle &= \langle Pq, Pk_i\rangle, \\ \langle (I-P)q, k_i\rangle &= \langle (I-P)q, (I-P)k_i\rangle. \end{aligned} \tag{7}$$

Thus, projecting the query alone computes the contributions of the two complementary subspaces to the native QK logit against the original visual keys. Section 5 evaluates the spatial responses and localization performance of the resulting readouts.

4.3 COMMON-SPACE RESIDUAL ATTRIBUTION

The original QK score combines the common and residual contributions. If the common contribution assigns high scores to irrelevant image regions, it can obscure the target in the resulting map. We therefore subtract the common contribution before combining scores across heads. We call the resulting spatial map Common-Space Residual Attribution (CSRA). Let H be the number of query heads at layer ℓ. The CSRA score for visual token i is

$$D_i^{\mathrm{CSRA}} = \frac{1}{H}\sum_{h=1}^{H}\left(z_i^{\ell,h} - z_{\mathrm{com},i}^{\ell,h}\right) = \frac{1}{H}\sum_{h=1}^{H} z_{\mathrm{res},i}^{\ell,h}. \tag{8}$$

This gives one score for each visual token. Arranging these scores on the image grid produces a spatial map for the target text. The map is computed directly from the residual QK scores. For the attention-mass diagnostics, we also apply softmax to the residual scores over visual tokens separately within each head.

CSRA uses the model's query and key vectors and the common basis constructed above, without target annotations or model parameter updates. Section 5 tests whether the common-component maps reproduce the spatial responses of queries extracted before the target text, and whether the residual maps locate the target more accurately.

5 EXPERIMENTS

5.1 SETUP

We conduct experiments on RefCOCO (Yu et al., 2016) and COCO Caption (Chen et al., 2015) using Qwen2.5-VL-7B (Bai et al., 2025), LLaVA-1.5-7B (Liu et al., 2023b), and InternVL3-8B (Zhu et al., 2025). Table 3 summarizes the five maps used in the experiments. We denote the QK score averaged across query heads by $\mathcal{M}_i(v) = H^{-1}\sum_h v_h^\top k_i^{\ell,g(h)}/\sqrt{d_h}$, where v_h is the query vector used by head h.

Map	Definition	Query contribution
Raw QK map	$\mathcal{M}_i(q)$	Original target query
Boundary-probe map	$\mathcal{M}_i(q_{\mathrm{probe}})$	Boundary query
Common-space map	$\mathcal{M}_i(Pq)$	Target-query projection
CSRA map	$\mathcal{M}_i((I-P)q)$	Signed query residual
CSRA+ map	$[D_i^{\mathrm{CSRA}}]_+$	Positive part after averaging

Table 3: Definitions of the maps. The first four maps average QK scores across query heads. CSRA+ sets negative values of the averaged CSRA map to zero.

Spatial Pearson correlation measures the similarity between the common-space and boundary-probe maps. For each QK map a, softmax over visual tokens converts its scores into attention probabilities, $p_a^h = \mathrm{softmax}_{\mathcal{V}}(z_a^h)$. The probability at the boundary-probe peak and the total probability within the target region are compared before and after common-space removal. We evaluate spatial alignment through target localization. The evaluation follows the protocol of TAM (Li et al., 2025) and use its implementation settings for CP-LRP (Ali et al., 2022). We obtain foreground masks using Otsu thresholding and report mIoU against the ground-truth masks. Details are provided in Appendix C.3.

5.2 THE COMMON SPACE RECOVERS SINK-LIKE SHARED ALIGNMENT

The common component closely matches the spatial response of the boundary probe, which has no access to the target text. Their QK maps have within-head spatial Pearson correlations of 0.761–0.959 (Table 4). This correspondence is obtained from all common directions, which capture 30.4%–71.3% of the target query's squared norm. Removing the common contribution reduces the attention probability at the fixed boundary-probe peak by 83.8%–96.6%, while increasing the total probability within the target region by 1.1%–15.1%.

Model	Layer	Query energy	Pearson corr.	Peak reduction	Δ Target mass
Qwen2.5-VL-7B	14	.554	.934	94.2%	+6.5%
	19	.571	.959	95.0%	+11.6%
	27	.713	.938	92.9%	+15.1%
LLaVA-1.5-7B	14	.492	.907	95.5%	+5.9%
	22	.539	.925	95.6%	+6.7%
	31	.510	.939	95.4%	+3.3%
InternVL3-8B	14	.304	.862	85.8%	+2.9%
	19	.514	.861	96.6%	+4.0%
	27	.506	.761	83.8%	+1.1%

Table 4: Properties of the common component and attention changes after its removal on RefCOCO. Query energy is the fraction of the target query's squared norm captured by the common subspace. Peak reduction measures the relative decrease in attention probability at the fixed boundary-probe peak. Δ Target mass measures the change in total attention probability within the target region. The target region is defined in Appendix C.2.

5.3 CSRA Improves Target Localization

CSRA+ improves target localization over native attention in all nine RefCOCO settings (Table 5). At Qwen2.5-VL layer 19, mIoU increases from 20.86 to 36.97; at LLaVA-1.5 layer 14, it increases from 6.96 to 34.01. These gains are obtained with unchanged visual keys and common basis constructed without target annotations. Figure 4 illustrates the resulting localization maps. The effect vary with depth. At LLaVA-1.5 layer 31, CSRA achieves 3.80 mIoU, below CP-LRP and TAM. TAM also performs better at InternVL3 layer 27. In particular, strong suppression of the shared peak at LLaVA-1.5 layer 31 still leaves the residual map poorly localized. Reducing a dominant shared response therefore does not ensure that the remaining map highlights the correct target. This depth dependence may reflect differences in how individual layers encode and retrieve target-related visual information.

Cross-dataset Transfer. The common basis constructed on RefCOCO remain useful when applied to COCO Caption without refitting. The CSRA+(transfer) in Table 5 reports results using these rank-4 basis within each backbone. The transferred version exceeds native attention in all nine settings and achieves the highest mIoU in seven, including 32.93 at LLaVA-1.5 layer 14 and 26.62 at Qwen2.5-VL layer 19. It also outperforms CSRA+ in eight settings.

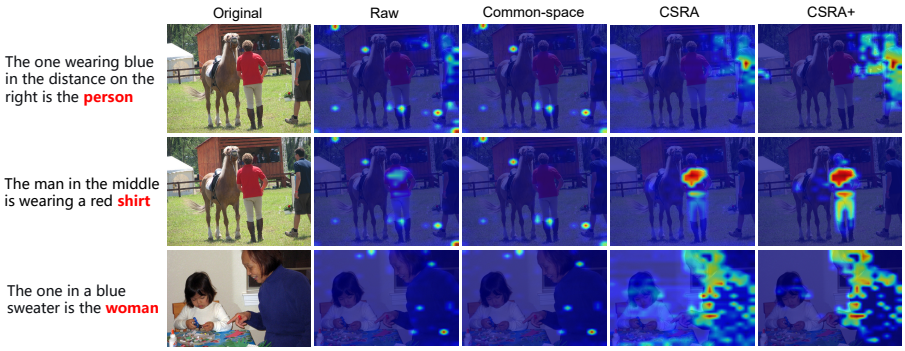


Figure 4: LLaVA-1.5-7B examples on RefCOCO. Rows show the image, native attention, and CSRA. Common removal shifts the response toward the referenced regions.

5.4 Sensitivity to Common-Subspace Rank

Table 6 evaluates rank sensitivity on RefCOCO using Qwen2.5-VL-7B. At layers 14 and 19, mIoU varies by only 0.99 and 1.26 % across ranks 1–16, respectively. Layer 27 is more sensitive: ranks 2 and 4 yield similar results (21.52 and 21.54), whereas ranks 8 and 16 reduce mIoU to 14.64 and 8.60. Rank 4 remains within 0.49 % of the best result at each evaluated layer, supporting its use as a shared

Datasets	Backbone	Layer	Raw↑	CP-LRP↑	TAM↑	CSRA+↑	CSRA+(transfer)↑
RefCOCO	LLaVA-1.5-7B	14	6.96	13.64	8.18	**34.01**	–
		22	1.03	10.21	6.37	**15.40**	–
		31	0.52	**7.46**	6.97	3.80	–
	Qwen2.5-VL-7B	14	19.01	8.79	8.58	**29.42**	–
		19	20.86	9.10	8.80	**36.97**	–
		27	1.37	11.50	12.54	**21.54**	–
	InternVL3-8B	14	17.61	5.10	9.63	**21.17**	–
		19	16.23	7.14	10.14	**18.20**	–
		27	6.94	9.21	**11.90**	9.41	–
COCO Caption	LLaVA-1.5-7B	14	4.98	16.00	4.91	28.01	**32.93**
		22	1.32	10.59	5.09	17.56	**19.88**
		31	0.33	5.48	4.00	**11.52**	10.92
	Qwen2.5-VL-7B	14	16.40	7.64	8.17	20.52	**25.56**
		19	5.62	7.45	8.17	24.08	**26.62**
		27	3.74	10.38	11.40	17.47	**22.84**
	InternVL3-8B	14	14.54	7.42	11.32	15.41	**17.44**
		19	5.27	4.04	9.88	12.93	**13.52**
		27	7.94	6.58	**10.93**	8.77	9.95

Table 5: mIoU (%) under identical visual normalization and Otsu thresholding. CSRA+(transfer) uses a rank-4 basis transferred from RefCOCO without COCO-side fitting.

Data	Backbone	Layer	Rank 1	Rank 2	Rank 4	Rank 8	Rank 16
RefCOCO	Qwen2.5-VL-7B	14	28.43	28.54	**29.42**	29.12	28.74
		19	36.20	36.36	36.97	37.02	**37.46**
		27	19.61	21.52	**21.54**	14.64	8.60

Table 6: Effect of subspace rank on RefCOCO localization. Values are mIoU (%) under Otsu thresholding, evaluated on 2000 records per setting. The best result for each layer is highlighted in bold.

setting. Increasing the removed rank does not uniformly improve localization. The deterioration at layer 27 is consistent with additional removed directions contributing to useful target alignment, suggesting that the functional separation between common and residual responses depends on both rank and depth.

6 Conclusion and Discussion

We examined how semantic structure in queries and visual keys relates to target localization. Semantic category structure survives per-head query projection, and visual keys support within-image target discrimination. CSRA separates common and residual query contributions and uses the residual scores to form a spatial map with the original visual keys. The resulting maps improve on native attention across the evaluated RefCOCO settings. The COCO category-localization results also support reusing a RefCOCO basis within the same backbone.

These findings distinguish the presence of semantic structure from its effective use in query–key matching. A poorly localized attention map can coexist with representations that support more accurate target localization. Understanding alignment therefore requires examining how different query components contribute to the scores assigned to image regions. By separating these contributions while keeping the visual keys fixed, CSRA provides a way to investigate the gap between the information available in the representations and the correspondence expressed by native attention.

AI Use Statement Generative AI tools were used to assist with manuscript drafting, translation, and language refinement, as well as the implementation and debugging of experimental code. They also supported literature exploration related to selected analysis methods. These tools were not used to formulate the research hypotheses, process the experimental datasets, or interpret the experimental results. The authors take full responsibility for the selection and application of the analysis methods, the correctness of the implementation, and the final content of this work, including all AI-assisted text and code.

Ethics Statement This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used, including PartImageNet, RefCOCO, and COCO Caption, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

Reproducibility Statement We have made every effort to ensure that the results presented in this paper are reproducible. The pt file of the common space fitted by Qwen2.5-VL-7B on RefCOCO, along with the CSRA calculation and visualization code, is recorded in supplementary materials for easy replication and verification. The experimental setup, model configurations, and hardware details, is described in the appendix. In addition, the public datasets used in the paper, such as PartImageNet, RefCOCO, and COCO Caption, are publicly available, ensuring consistency and reproducibility of the evaluation results.We believe these measures will enable other researchers to replicate our work and further advance the field.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. XAI for transformers: Better explanations through conservative propagation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 435–451, 2022. URL https://proceedings.mlr.press/v162/ali22a.html.

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

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4792–4808, 2024. doi: 10.18653/v1/2024.acl-long.263. URL https://aclanthology.org/2024.acl-long.263/.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. *arXiv preprint arXiv:2012.09838*, 2020. URL https://arxiv.org/abs/2012.09838.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. URL https://arxiv.org/abs/2309.16588.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. *arXiv preprint arXiv:2112.00933*, 2021. URL https://arxiv.org/abs/2112.00933.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Kang et al. See what you are told: Visual attention sink in large multimodal models, 2025.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *arXiv preprint arXiv:2304.05653*, 2023. URL https://arxiv.org/abs/2304.05653.

Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. Token activation map to visually explain multimodal llms, 2025. URL https://arxiv.org/abs/2506.23270.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a. URL https://arxiv.org/abs/2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b. URL https://arxiv.org/abs/2304.08485.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017. URL https://arxiv.org/abs/1702.01417.

Xu Pan, Aaron Philip, Ziqian Xie, and Odelia Schwartz. Dissecting query-key interaction in vision transformers. *arXiv preprint arXiv:2405.14880*, 2024. URL https://arxiv.org/abs/2405.14880.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. URL https://arxiv.org/abs/2309.17453.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. URL https://arxiv.org/abs/2111.07783.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. *arXiv preprint arXiv:1608.00272*, 2016. URL https://arxiv.org/abs/1608.00272.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in LVLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3512–3534, 2025. doi: 10.18653/v1/2025.emnlp-main.174. URL https://aclanthology.org/2025.emnlp-main.174/.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

A QUERY SEMANTIC-GEOMETRY ANALYSIS

We examine two aspects of query compression: whether semantic categories remain distinguishable, and whether relationships among term representations are preserved. The following definitions apply to one image, one layer, and one query head at a time. We suppress image and layer indices; p, q index terms and h indexes query heads.

A.1 EXPERIMENT CONFIG

Runtime Environment. The experiments were conducted using Python 3.10 and PyTorch 2.8.0 with CUDA 12.8 on a single GPU instance with 32 GB of available GPU memory.

Dataset and preprocessing. We evaluate query-projection geometry on a fixed subset of the PartImageNet test split restricted to synset `n03594945`, used in our car-side-mirror analysis. The experiment uses **100** images; the final image manifest specifies the exact evaluation set. The probe operates on complete RGB images without part cropping or segmentation-mask inputs. Each image is resized directly to 672×672 pixels using bicubic interpolation. The semantic labels used for RSA describe the prompt terms rather than image or part annotations.The detailed image names are recorded in the attachment.

Model, prompts, and probe position. We use Qwen2.5-VL-7B, LLaVA-1.5-7B, InternVL3-8B in evaluation mode with frozen parameters and its native multimodal chat template. Specifically, Qwen2.5-VL-7B utilizes the Instruct version weights from "Qwen/Qwen2.5-VL-7B-Instruct". InternVL3-8B's weights are sampled from "OpenGVLab/InternVL3-8B-hf", and LLaVA-1.5-7B's weights are sampled from "llava-hf/llava-1.5-7b-hf".

The candidate inventory contains terms from six categories: part, action, shape, material, color, and position. Terms are filtered to a common token length measured within the rendered template. Up to eight terms are retained per eligible category, with categories containing fewer than two eligible terms excluded.

Numerical configuration. The default random seed is 0 for both PyTorch and NumPy. Model forward passes use `bfloat16` and scaled dot-product attention; captured states are stored in `float16`, converted to the layer's normalization dtype for RMSNorm. The default forward batch size is 8, and projection metrics are processed in chunks of 4 query heads.

A.2 REPRESENTATIONS AND DISTANCE MATRICES

For each image, we construct one prompt per term while keeping the image, surrounding text, and the identity of an anchor token following the term fixed. We extract the representation at this anchor in evaluation mode. Let $x_p \in \mathbb{R}^d$ be the normalized hidden state entering the query projection and let

$$y_p = W_Q^{(h)} x_p \in \mathbb{R}^{d_h} \tag{9}$$

be its per-head query projection. Here, normalization follows the model's native attention input normalization: if the model applies such normalization (including its learned parameters), we use it; otherwise, we do not normalize. Queries are measured before RoPE, without any additional unit-length normalization.

The default set contains the first eight terms in each row of Table 7, giving P = 48 terms in six equally sized categories. Terms after each semicolon form a disjoint replication inventory, evaluated separately. Category labels describe the terms, independently of image content. Prompt indices follow the order of terms in the table.

For each pair of terms, we compute the hidden-space distance, query-space distance, and category difference:

$$\begin{aligned} D^x_{pq} &= \|x_p - x_q\|_2, \\ D^y_{pq} &= \|y_p - y_q\|_2, \\ R_{pq} &= \mathbf{1}[\mathrm{cat}(p) \neq \mathrm{cat}(q)]. \end{aligned} \tag{10}$$

Category	Terms (default inventory; replication inventory)
Part	wheel, engine, window, door, mirror, handle, seat, antenna; tire, bumper, pedal, dashboard, brake, cabin, headlight, taillight
Action	running, jumping, standing, walking, flying, swimming, sleeping, climbing; rolling, sitting, turning, stopping, moving, spinning, falling, rising
Shape	circle, square, curve, spiral, oval, triangle, rectangle, sphere; cube, cone, cylinder, arc, ring, line, star, diamond
Material	metal, plastic, glass, leather, stone, ceramic, rubber, silk; wood, paper, cotton, steel, alloy, foam, fabric, wool
Color	red, blue, green, yellow, black, white, purple, gray; orange, brown, pink, cyan, silver, gold, violet, beige
Position	left, right, above, below, center, front, behind, beside; top, bottom, middle, inside, outside, upper, lower, corner

Table 7: Balanced term inventory. The reported analysis uses the default inventory.

Thus D^x and D^y are Euclidean-distance RDMs, while R is the binary category reference. Let $\mathrm{ut}(\cdot)$ extract the strict upper triangle in row-major order, excluding the diagonal. Define

$$d_x = \mathrm{ut}(D^x), \quad d_y = \mathrm{ut}(D^y), \quad r = \mathrm{ut}(R). \tag{11}$$

Thus D^x and D^y are Euclidean-distance RDMs, while R is the binary category reference. Let $\mathrm{ut}(\cdot)$ extract the strict upper triangle in row-major order, excluding the diagonal. Define

$$d_x = \mathrm{ut}(D^x), \quad d_y = \mathrm{ut}(D^y), \quad r = \mathrm{ut}(R). \tag{12}$$

Each vector contains $P(P-1)/2 = 1128$ corresponding term pairs. The reference r is fixed across images, layers, and heads.

A.3 Category structure and geometry preservation

Category structure. Hidden and query semantic RSA compare each representation space with the same category reference:

$$\begin{aligned} \mathrm{RSA}_{\mathrm{hidden}} &= \rho_s(d_x, r), \\ \mathrm{RSA}_{\mathrm{query}} &= \rho_s(d_y, r). \end{aligned} \tag{13}$$

Here ρ_s is Spearman correlation, computed as the Pearson correlation of ascending ranks with average ranks for ties in either argument. A positive score indicates that between-category term pairs tend to be farther apart than within-category pairs.

Geometry preservation. Distance RSA compares the ordering of all pairwise distances before and after projection:

$$\mathrm{RSA}_{\mathrm{distance}} = \rho_s(d_x, d_y). \tag{14}$$

To measure local preservation, let $\mathcal{N}_k^x(p)$ and $\mathcal{N}_k^y(p)$ be the k nearest other terms to p in D^x and D^y, respectively. With $k = 3$, the neighborhood overlap is

$$\mathrm{Overlap}_k = \frac{1}{P} \sum_{p=1}^{P} \frac{|\mathcal{N}_k^x(p) \cap \mathcal{N}_k^y(p)|}{k}. \tag{15}$$

These two metrics compare hidden and query geometry without category labels. Distance RSA measures global ordering, whereas overlap measures retention of local neighbors.

A.4 Spectrum-matched reference and aggregation

For each layer and head, write the thin singular-value decomposition as $W_Q^{(h)} = USV^\top$. Within each head, we draw $B = 200$ independent Haar-distributed orthonormal frames $V_b \in \mathbb{R}^{d \times d_h}$ and define

$$W_{\mathrm{null}}^{(h,b)} = USV_b^\top, \qquad b = 1, \ldots, B. \tag{16}$$

Metric	Qwen2.5-VL Learned	Qwen2.5-VL Null	InternVL3.5 Learned	InternVL3.5 Null	LLaVA-1.5 Learned	LLaVA-1.5 Null
Query semantic RSA	.396	.379	.434	.424	.406	.397
Distance RSA	.687	.743	.685	.720	.628	.621
kNN overlap	.567	.578	.590	.591	.600	.575

Table 8: Mean geometry statistics for learned and spectrum-matched random-orientation projections, aggregated as specified above.

Each frame is generated from a $d \times d_h$ matrix of independent standard normal entries using reduced QR decomposition, with column signs adjusted to make the triangular factor's diagonal nonnegative. The null therefore preserves all singular values, including zeros, while randomizing input orientation. Each sampled matrix is reused across all images.

For each draw, replace y_p by $y_p^{(b)} = W_{\text{null}}^{(h,b)} x_p$ and recompute query RSA, distance RSA, and overlap using the same hidden representations and category reference. Hidden RSA is unchanged. Because U is orthogonal, $SV_b^\top x_p$ gives the same distance-based metrics as $USV_b^\top x_p$.

Aggregation. Each metric is computed separately within each image before averaging. For the learned projection, average over images, then heads, then layers. For the null, average over images within each draw, then draws, heads, and layers. Hidden RSA is averaged over images and layers only. These are arithmetic means with equal weights at each stage, without a Fisher transform. Invalid scores are excluded at the stage where they occur; an empty mean remains invalid, and valid/excluded counts are retained. Layer-wise curves omit the final averaging over layers.

Both learned and reference projections retain positive category RSA (Table 8). Distance RSA and overlap characterize the accompanying preservation of global and local geometry. These results concern structure in query space; its relation to visual readout is tested separately in the fixed-key experiments.

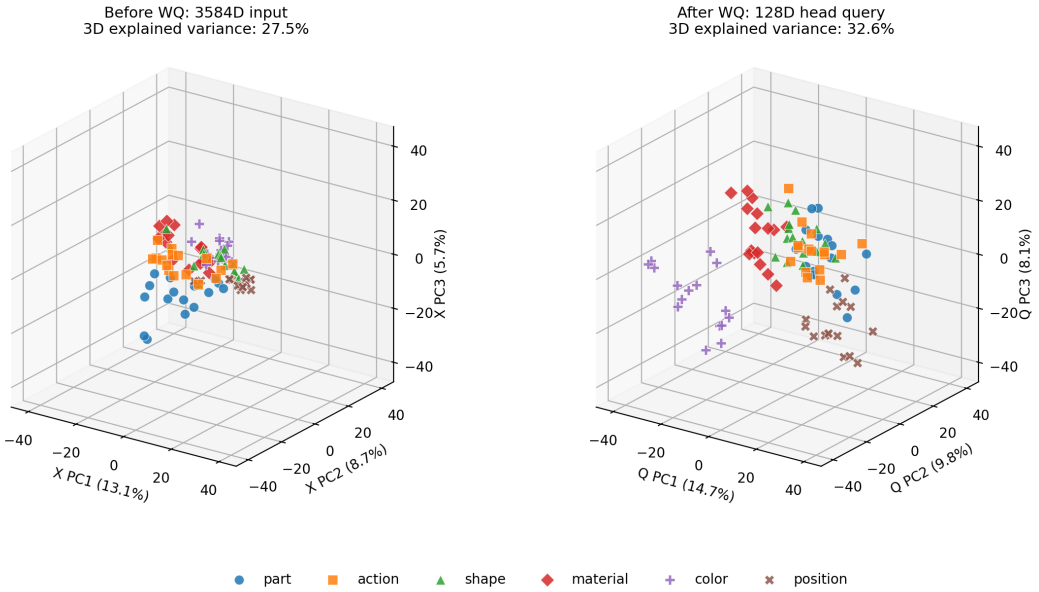


Figure 5: Three-dimensional PCA views of attention-input representations before (W_Q) (left, 3584 dimensions) and the corresponding per-head queries before RoPE (right, 128 dimensions). Each point represents one term, with colors and markers indicating semantic categories. This illustrative example combines the default and replication inventories (96 terms). PCA is fitted separately in each space; the first three components explain 27.5% and 32.6% of the respective total variance. The figure provides a qualitative view of category-related organization.

Qualitative visualization of per-head query representations. The input representations already exhibit category-related organization, which remains visible after projection into the 128-dimensional query space. In the query view, color and position terms occupy relatively distinct regions, and material terms exhibit a recognizable grouping. Part, action, and shape terms show greater overlap. Thus, the displayed head retains visible category-related structure without producing uniformly compact or fully separated category clusters.

B KEY-GEOMETRY STATISTICS AND ORACLE READOUT

B.1 EXPERIMENTAL SETTINGS

Runtime Environment. The experiments were conducted using Python 3.10 and PyTorch 2.8.0 with CUDA 12.8 on a single GPU instance with 32 GB of available GPU memory.

Dataset and annotation preprocessing. We evaluate target-key geometry and oracle readout on a fixed set of $2{,}200$ image–expression examples from the RefCOCO validation split. Each example contains an RGB image, a referring expression, and a target segmentation annotation. The model receives only the image and text; segmentation masks are used offline to define the target-token set and construct the annotation-conditioned diagnostics. We align each mask with the corresponding image transformation and area-average it onto the model's visual-token grid. Tokens with target coverage of at least 0.5 form Ω. If fewer than two tokens meet this threshold, we select the two tokens with the highest coverage. The analysis requires at least two target and two background tokens. The image identifier and referring expression are retained in the evaluation records.The detailed image names are recorded in the attachment.

Models and image preprocessing. The main backbone comparisons use Qwen2.5-VL-7B, LLaVA-1.5-7B, and InternVL3-8B, with checkpoints `Qwen/Qwen2.5-VL-7B-Instruct`, `llava-hf/llava-1.5-7b-hf`, and `OpenGVLab/InternVL3-8B-hf`, respectively. All models run in evaluation mode with fixed parameters and gradient recording disabled.

Oracle construction and attention comparison. The oracle results reported here use Qwen2.5-VL-7B at layers 14, 19, and 27, retaining all four KV heads at each layer. We capture the post-RoPE queries and keys used by attention. For each example–layer–KV-head condition, we construct the ridge-LDA direction from the target and background keys using the class-balanced within-class covariance. The ridge coefficient is $\lambda = 0.05 \max(\mathrm{tr}(\Sigma_w)/d, 10^{-6})$. We keep the visual keys fixed and replace each native query direction with the oracle direction while preserving that query's original norm. Each KV head serves seven query heads. We include all query positions after the vision-end marker in the input sequence, with no last-token restriction. For both native and oracle queries, visual softmax normalization is performed separately for each query and head before averaging within the KV-head group.

B.2 AUC DEFINITIONS

This subsection specifies the three AUC metrics reported in Table 2. Let $\mathcal{V}$ denote the visual-token set, $N = |\mathcal{V}|$, $\Omega \subset \mathcal{V}$ the annotated target-token set, and $\Omega^c = \mathcal{V} \setminus \Omega$ the background-token set. The key-geometry diagnostics use pre-RoPE keys; the oracle analysis below uses post-RoPE keys. Each metric is computed within an individual example, layer, and KV head before aggregation.

Similarity among visual keys and ROC-AUC. For normalized keys $\widehat{k}_i = k_i / \|k_i\|_2$, define the pairwise cosine similarity $M_{it} = \widehat{k}_i^\top \widehat{k}_t$. The target-prototype score is the average similarity to target keys, excluding self-similarity:

$$c_i = \frac{1}{|\Omega_{-i}|} \sum_{t \in \Omega_{-i}} M_{it}, \qquad \Omega_{-i} = \begin{cases} \Omega \setminus \{i\}, & i \in \Omega, \\ \Omega, & i \notin \Omega. \end{cases}$$

For any scalar token score $s = (s_i)_{i \in \mathcal{V}}$, define the empirical ROC-AUC functional

$$\mathcal{A}(s; \Omega) = \frac{1}{|\Omega|\,|\Omega^c|} \sum_{i \in \Omega} \sum_{j \in \Omega^c} \left(\mathbf{1}[s_i > s_j] + \frac{1}{2} \mathbf{1}[s_i = s_j] \right).$$

Target tokens are positive examples and background tokens are negative examples; ties receive half credit. The target-prototype metric is

$$\mathrm{AUC}_{\mathrm{target}} = \mathcal{A}(c; \Omega).$$

Because the annotations define both the prototype reference set and the evaluation labels, this is an annotation-conditioned measure of target-region coherence.

Global-centrality-adjusted AUC. Global key centrality is the average cosine similarity to all other visual keys:

$$g_i = \frac{1}{N-1} \sum_{t \in \mathcal{V} \setminus \{i\}} M_{it}.$$

We regress c_i on g_i with an intercept using all visual tokens in the same example, layer, and KV head. With $\bar{c} = N^{-1} \sum_{i \in \mathcal{V}} c_i$ and $\bar{g} = N^{-1} \sum_{i \in \mathcal{V}} g_i$, the ordinary least-squares coefficient, intercept, and residual score are

$$\widehat{\beta} = \frac{\sum_{i \in \mathcal{V}} (g_i - \bar{g})(c_i - \bar{c})}{\sum_{i \in \mathcal{V}} (g_i - \bar{g})^2},$$
$$\widehat{\alpha} = \bar{c} - \widehat{\beta} \bar{g},$$
$$r_i = c_i - \widehat{\alpha} - \widehat{\beta} g_i.$$

The adjusted metric reported in Table 2 is

$$\mathrm{AUC}_{\mathrm{global}} = \mathcal{A}(r; \Omega).$$

If $\sum_{i \in \mathcal{V}} (g_i - \bar{g})^2 < 10^{-12}$, we set $\widehat{\beta} = 0$ and use $r_i = c_i - \bar{c}$. This metric evaluates separation after removing the scalar linear association between the target-prototype score and global key centrality. The adjustment acts on scalar scores and is separate from the query-space projection used by CSRA.

Key-norm AUC. Using the original keys before unit-length normalization, define

$$n_i = \|k_i\|_2, \qquad \mathrm{AUC}_{\mathrm{norm}} = \mathcal{A}(n; \Omega).$$

Larger key norms receive higher scores. Values above 0.5 indicate that target keys tend to have larger norms than background keys; values below 0.5 indicate the reverse ordering. The original key norms are used because unit-length normalization removes norm variation. RoPE preserves Euclidean norms, so this metric is unchanged by the rotation. For all three AUC metrics, score orientation is fixed by the definitions above: no label-dependent sign reversal or $\max(\mathrm{AUC}, 1 - \mathrm{AUC})$ transformation is applied.

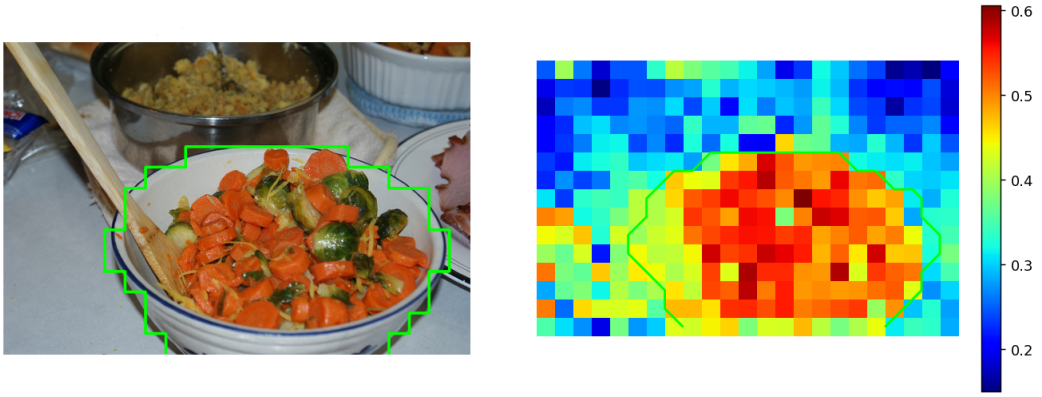


Figure 6: Cosine similarity among visual keys

B.3 Regularized Fisher direction

The oracle readout uses the original, unnormalized post-RoPE keys in the coordinates used by attention. For each image, layer, and KV head, let $\delta = \mu_\Omega - \mu_{\Omega^c}$ and define the class-balanced within-class covariance

$$\Sigma_w = \frac{1}{2}\left(\widehat{\Sigma}_\Omega + \widehat{\Sigma}_{\Omega^c}\right), \qquad \widehat{\Sigma}_C = \frac{1}{|C|-1}\sum_{i\in C}(k_i-\mu_C)(k_i-\mu_C)^\top.$$

Each class covariance therefore uses its own sample-count-minus-one denominator. We scale the ridge to the covariance magnitude:

$$\lambda = \rho \max\left(\frac{\mathrm{tr}(\Sigma_w)}{d}, 10^{-6}\right), \qquad \rho = 0.05,$$

where d is the key dimension.

For positive-definite $S = \Sigma_w + \lambda I$, the regularized Fisher criterion is

$$J(v) = \frac{(v^\top \delta)^2}{v^\top S v}.$$

Writing $u = S^{1/2}v$ and applying Cauchy–Schwarz gives

$$J(v) = \frac{(u^\top S^{-1/2}\delta)^2}{u^\top u} \le \delta^\top S^{-1}\delta,$$

with equality for $v \propto S^{-1}\delta$. This identifies the optimization criterion underlying Eq. 3. We normalize this direction to obtain $v^\star$ and choose its sign so that the target mean score exceeds the background mean score. The direction is constructed and evaluated on the same labeled key field, giving a constructive, annotation-conditioned diagnostic of within-image linear accessibility. The Fisher optimum does not imply that this direction maximizes AUC or IoU over all possible queries.

B.4 Oracle ranking and thresholded localization

We use ROC-AUC and average precision (AP) as the primary ranking metrics, evaluated on the oracle linear scores

$$s_i = (v^\star)^\top k_i.$$

AUC measures target–background ordering, whereas AP summarizes how strongly target tokens are concentrated near the top of the ranking and depends on target prevalence. Both metrics are computed within each image–layer–KV-head condition before aggregation. Multiplication of all scores by a positive scalar leaves their rankings unchanged.

For attention comparisons, we hold the visual keys fixed and replace each selected native query by

$$q^\star_{hj} = \|q_{hj}\|_2 v^\star,$$

preserving the norm of that individual query. For either the native or oracle condition, the visual attention map is

$$a_i = \frac{1}{HJ}\sum_{h=1}^{H}\sum_{j=1}^{J}\frac{\exp(q_{hj}^\top k_i/\sqrt{d})}{\sum_{t\in\mathcal{V}}\exp(q_{hj}^\top k_t/\sqrt{d})}.$$

In Qwen2.5-VL-7B, each KV head is shared by $H = 7$ query heads. The experiment uses the selected post-vision query positions, with J ranging from 6 to 28 across examples. Visual normalization is performed separately for every query and head before their distributions are averaged.

As an auxiliary localization metric, we threshold each attention map using Otsu's method with 128 histogram bins over that map's value range. The predicted token set and its overlap with the annotation are

$$P = \{i \in \mathcal{V} : a_i > \tau_{\mathrm{Otsu}}(a)\}, \qquad \mathrm{IoU}_{\mathrm{Otsu}} = \frac{|P\cap\Omega|}{|P\cup\Omega|}.$$

Native and oracle maps use the same thresholding algorithm, with a threshold computed independently for each map. The threshold is determined from the attention histogram rather than selected to maximize overlap with the annotation. Otsu IoU measures thresholded attention-map localization and depends on the score distribution as well as the ranking. We therefore interpret it as auxiliary evidence alongside AUC and AP.

Layer	AUC	AP	Otsu IoU
14	0.9981	0.9826	0.8099
19	0.9974	0.9763	0.7804
27	0.9969	0.9719	0.7241
Overall	**0.9975**	**0.9770**	**0.7715**

Table 9: Layer-wise ridge-LDA oracle results for Qwen2.5-VL-7B on RefCOCO val. Each layer averages four KV-head groups within each of 2,200 examples and then across examples. The overall row also averages the three layers. AUC and AP evaluate the oracle linear ranking; Otsu IoU evaluates the thresholded norm-matched oracle attention map.

Layer	KV head	AUC	AP	Otsu IoU
14	0	0.9984	0.9848	0.7992
	1	0.9978	0.9788	0.7426
	2	0.9980	0.9817	0.8321
	3	0.9984	0.9853	0.8658
19	0	0.9977	0.9788	0.6456
	1	0.9987	0.9889	0.9179
	2	0.9965	0.9672	0.7568
	3	0.9968	0.9702	0.8012
27	0	0.9964	0.9670	0.8278
	1	0.9978	0.9826	0.3899
	2	0.9967	0.9697	0.8417
	3	0.9966	0.9684	0.8369

Table 10: Ridge-LDA oracle results by layer and KV head for Qwen2.5-VL-7B. Each row averages 2,200 RefCOCO examples. KV-head indices are zero-based; each KV head corresponds to a group of seven query heads. All four KV heads are retained at every layer.

B.5 Aggregation and oracle results

For each layer and KV head, we average the metrics over 2,200 RefCOCO image–expression examples. Layer-level summaries first average the four KV-head groups within each example and then average examples. The overall summary additionally averages layers 14, 19, and 27. Since every group contains the same examples, this is equivalent to an unweighted mean over 26,400 image–layer–KV-head records. The oracle analysis retains all four KV heads at all three layers, including layer 27, KV head 1;

The oracle attains mean AUC 0.9975, AP 0.9770, and Otsu IoU 0.7715. Under the same attention aggregation and Otsu procedure, the native queries attain mean Otsu IoU 0.0343. Table 9 reports the layer-level summaries, and Table 10 gives the KV-head breakdown. The high oracle AUC and AP support the existence of target-selective linear directions in these fixed key fields, while the native–oracle overlap difference shows that weak native localization can coexist with this accessible structure. These results characterize the labeled instances used to construct the oracle directions; they do not establish generalization to unseen images or targets.

Ranking and thresholded overlap need not vary together. For example, layer 27, KV head 1 attains AUC 0.9978 and AP 0.9826, but Otsu IoU 0.3899. This variation reinforces the use of ranking metrics as the primary evidence for linear accessibility and Otsu IoU as an auxiliary localization diagnostic.

C Common-Space Diagnostics and Localization Protocol

C.1 Common-space experimental setup

Runtime Environment. The experiments were conducted using Python 3.10 and PyTorch 2.8.0 with CUDA 12.8 on a single GPU instance with 32 GB of available GPU memory.

Dataset and preprocessing. For Qwen2.5-VL-7B, InternVL3-8B, and LLaVA-1.5-7B, the reference configuration uses the `val-00000-of-00004.parquet` shard of the RefCOCO validation split. Each dataset record supplies an image, a referring expression, and, when available, a bounding box in COCO (x, y, w, h) format. Images are converted to RGB and processed according to each checkpoint's native preprocessing. The detailed image names are recorded in the attachment.

Fitting and evaluation records. For the reported common-space diagnostics on Qwen2.5-VL-7B, InternVL3-8B, and LLaVA-1.5-7B, the run configuration specifies 200 fitting records and 2,000 evaluation records per backbone, with common-space rank $r = 4$. Records are traversed in stored Parquet order from row 0. Fitting collects 200 successful records; evaluation uses the same candidate order after removing fitting indices, so row indices are disjoint. Records are not grouped by image identity; different expressions for the same image may cross partitions, so evaluation is held out at the record level. The common basis is frozen before evaluation readout, and bounding-box annotations are not used for common-basis fitting.

Model, prompts, and probe positions. We evaluate Qwen2.5-VL-7B, InternVL3-8B, and LLaVA-1.5-7B in evaluation mode with all model parameters frozen. Each input consists of the image followed by its referring expression in the model's native multimodal chat template, rendered with the assistant generation prefix. The expression is read from the `answer` field; when list-valued, the first entry is selected, and surrounding brackets, quotes, and a trailing Chinese full stop are removed by the input helper.

Token locations are determined after tokenization and visual-token expansion. For Qwen2.5-VL-7B, visual tokens lie strictly between `<|vision_start|>` and `<|vision_end|>`; with the default target offset of zero, the target query is extracted at the last content token before the following `<|im_end|>`, and the boundary probe is extracted at `<|vision_end|>`. For InternVL3-8B, we force a single 448×448 tile, giving 256 visual tokens (16×16) as the contiguous `<IMG_CONTEXT>` block between `<img>` and `</img>`; the target query is extracted at the last content token before `<|im_end|>`, and the boundary probe is extracted at `</img>`. For LLaVA-1.5-7B, model transferthe full image to 336×336, giving 576 visual tokens (24×24) from the expanded `<image>` placeholder; there is no explicit closing vision token. The generation-prompt length is computed by subtracting the token count of the chat template without a generation prompt from that with one, so the target query is the last content token before the generation prompt, and the boundary probe is the first token after the visual block. In all cases, the boundary probe can attend to the image and preceding context, while causal masking prevents access to the later referring expression.

Common-space estimation. For each layer and query head, let $q_{\mathrm{probe},n} \in \mathbb{R}^{d_h}$ be the boundary query of calibration record n. The default fit normalizes each calibration vector to unit length, with a denominator floor $\varepsilon = 10^{-8}$:

$$\widehat{q}_{\mathrm{probe},n} = \frac{q_{\mathrm{probe},n}}{\max(\|q_{\mathrm{probe},n}\|_2, \varepsilon)}.$$

Stacking these vectors as rows gives

$$Q_{\mathrm{probe}} = \begin{bmatrix} \widehat{q}_{\mathrm{probe},1}^{\top} \\ \vdots \\ \widehat{q}_{\mathrm{probe},N_{\mathrm{fit}}}^{\top} \end{bmatrix} = L\Sigma V^{\top}, \qquad U = V_{[:,1:r]}, \quad P = UU^{\top}. \tag{17}$$

The SVD is uncentered: the cross-record mean is retained. A separate orthonormal basis is fitted for each layer and query head. This calibration normalization is used to estimate U; evaluated queries and their projected components retain their native scales.

Frozen-key readout comparison. For each evaluation record, all readout conditions use the same captured target query, boundary query, and visual key field. Only the query vector used in the offline dot product is changed. The projector is neither refitted on evaluation records nor optimized against their annotations. The readout comparison does not modify model parameters, attention values, or the model's subsequent forward computation. Native, common-only, and residual logits therefore differ only through the query component specified below.

Numerical configuration. The reference Qwen configuration uses `bfloat16` model weights, `eager` attention, inference mode, and a batch size of one. Post-RoPE vectors are detached and stored on CPU in `float32`; SVD, projections, reconstructed logits, softmax distributions, and per-head diagnostics are computed from those `float32` tensors. The default `seed=1234` controls the optional Python-based row shuffle only. Shuffling is disabled by default; the script does not use that argument to seed PyTorch or NumPy.

C.2 METRICS AND AGGREGATION

Evaluation units and readouts. We define each diagnostic for an evaluated record n, layer ℓ, and query head h before aggregation. These indices are suppressed below. Let $\mathcal{V}$ denote the visual-token set, q the target query, q_{probe} the boundary query, and $P = UU^\top$ the frozen common-space projector. Each k_i is the post-RoPE visual key from the KV head paired with the current query head. Define

$$\begin{aligned} q_{\text{full}} &= q, \qquad & q_{\text{com}} &= Pq, \qquad & q_{\text{res}} &= (I-P)q, \\ z_{a,i} &= \frac{\langle q_a, k_i \rangle}{\sqrt{d_h}}, \qquad & p_{a,i} &= \frac{\exp(z_{a,i})}{\sum_{v\in\mathcal{V}} \exp(z_{a,v})}, \end{aligned} \tag{18}$$

where $a \in \{\text{full}, \text{com}, \text{res}, \text{probe}\}$. The projected components are not individually normalized. Every probability map is normalized over visual positions only. In particular, $z_{\text{res}} = z_{\text{full}} - z_{\text{com}}$, and p_{res} is the softmax of that residual logit map. It is not a difference between probability maps. These diagnostics use signed logits without CSRA+ clipping, image-space interpolation, or Otsu thresholding.

Query energy. The common-query energy fraction is

$$e = \frac{\|Pq\|_2^2}{\max(\|q\|_2^2, \varepsilon)}, \qquad \varepsilon = 10^{-8}. \tag{19}$$

This is the fraction of the evaluated target query's squared norm captured by the frozen projector, rather than the explained singular-value energy of the calibration matrix.

Probe correlation. The probe correlation is the spatial Pearson correlation between common-space and boundary-probe map within a query head. Writing $\bar{z}_a = |\mathcal{V}|^{-1} \sum_{i\in\mathcal{V}} z_{a,i}$ and $\widetilde{z}_{a,i} = z_{a,i} - \bar{z}_a$, we compute

$$\rho = \frac{\sum_{i\in\mathcal{V}} \widetilde{z}_{\text{com},i} \widetilde{z}_{\text{probe},i}}{\max\left\{ \sqrt{\sum_{i\in\mathcal{V}} \widetilde{z}_{\text{com},i}^2} \sqrt{\sum_{i\in\mathcal{V}} \widetilde{z}_{\text{probe},i}^2}, \varepsilon \right\}}. \tag{20}$$

The correlation is computed before averaging across heads or records, using logits rather than softmax probabilities or rendered heatmaps. The denominator clamp yields zero for a spatially constant map.

Reference position and peak mass. The default `struct_argmax` rule selects

$$s = \arg\max_{i\in\mathcal{V}} z_{\text{probe},i}, \qquad S_a = p_{a,s}. \tag{21}$$

The reference position is selected separately for each record, layer, and head, then held fixed across its full and residual readouts. Ties follow the first maximizing token in the flattened visual order. The reference is not constrained to lie outside the target region. Accordingly, S_{full} and S_{res} compare probability mass at the same probe-selected location, not separately selected maxima.

Target-token set. Let $\Omega = \Omega_{\text{mass}}$ be the merged visual-token cells whose centers lie inside the ground-truth bounding box. The box is clipped to the original image bounds, and grid-cell coordinates span the full original image extent. Centers on the box boundary are included. If no center lies inside the box, cells with positive-area intersection with the box are used instead. This definition does not use polygon segmentation, a coverage threshold, or a minimum-two-token rule. Missing boxes or empty resulting sets make the corresponding target-region diagnostics undefined, while the remaining diagnostics can still be evaluated.

Target-region mass. The probability mass assigned to the target-token set is

$$T_a = \sum_{i \in \Omega} p_{a,i}, \qquad \delta T = 100(T_{\text{res}} - T_{\text{full}}). \tag{22}$$

The change is measured in percentage points, not relative percent. It is evaluated directly on the merged visual-token grid. It does not require image-space upsampling or foreground-mask thresholding and is distinct from the mIoU evaluation in the localization experiment.

Aggregation and record counts. Let $\mathcal{I}_\ell(f)$ contain the records with a defined metric f at layer ℓ, including their query-head indices. We write

$$\langle f \rangle_\ell = \frac{1}{|\mathcal{I}_\ell(f)|} \sum_{(n,h) \in \mathcal{I}_\ell(f)} f_{n\ell h}. \tag{23}$$

Each defined example–head record receives equal weight. The layer-level query-energy and probe-correlation entries are $\langle e \rangle_\ell$ and $\langle \rho \rangle_\ell$. Target-mass change is $100(\langle T_{\text{res}} \rangle_\ell - \langle T_{\text{full}} \rangle_\ell)$, using the same records with defined target masses for both conditions. Images are not weighted by their visual-token counts or target areas. Per-layer summaries pool records across heads; per-head summaries average records for a fixed layer and head. Exported standard deviations are sample standard deviations across the grouped records, not standard errors or bootstrap confidence intervals. The export's generic `count` field is the maximum non-missing count across its metrics, so it need not equal the number of defined target-mass observations.

Peak-reduction reporting convention. The peak-reduction entries in Table 4 use the ratio of the exported layer-level mean masses:

$$R_\ell = 100 \left(1 - \frac{\langle S_{\text{res}} \rangle_\ell}{\langle S_{\text{full}} \rangle_\ell} \right) \quad (\%). \tag{24}$$

This is a relative decrease in mean mass at the fixed probe-selected position. It differs from averaging per-record percentage decreases. The ratio is defined when $\langle S_{\text{full}} \rangle_\ell > 0$; all layer means used in the reported table satisfy this condition. For example, Qwen layer 19 has full and residual mean peak masses 0.0403011872 and 0.0020028625, respectively, giving a 95.0303% decrease, reported as 95.0%. Its target masses change from 0.3258272461 to 0.4417657700, giving 11.5939 percentage points, reported as $+11.6$ pp.

C.3 Localization experiments and frozen transfer

This subsection details the experimental setup and foreground-mask evaluation for Section 5.3. We use the method names and map definitions in Section 5.1 and report localization results as foreground mIoU.

C.3.1 Experimental setup

RefCOCO records and preprocessing. We follow the RefCOCO dataset and image-preprocessing protocol in Appendix C.1.

COCO image selection and target construction. The COCO Caption setting uses the Karpathy-style validation image list and COCO 2014 instance-segmentation annotations. Duplicate image paths are removed while preserving their order, and the first $2{,}200$ unique images form the evaluation set. For each image, instance masks belonging to the same category are combined by union. Each image–category pair is evaluated once, using the complete category name as the text prompt. All categories with nonempty merged masks are included, with a minimum foreground area of one pixel and no additional area-fraction cutoff or crowd exclusion. Thus, the caption dataset specifies the image set, while category names and instance masks define the localization task.The detailed image names are recorded in the attachment.

Model state and numerical configuration. We evaluate Qwen2.5-VL-7B, InternVL3-8B, and LLaVA-1.5-7B in evaluation mode with frozen parameters. Model-specific image preprocessing, multimodal templates, and target-query positions follow Appendix C.1. Queries and keys are taken in the post-normalization, post-RoPE coordinates used by attention, with the native pairing between query heads and KV heads. Offline projections and readout calculations use `float32`. Evaluated queries and their projected components retain their native scales, and the original visual keys are held fixed. Results are reported separately for the layers in Table 5.

C.3.2 Localization readouts

Native attention. Let $p^h_{\mathrm{full},i}$ denote the target query's attention probability over visual tokens, as defined in Eq. equation 18. The Raw localization map is

$$A_{\mathrm{raw},i} = \frac{1}{H}\sum_{h=1}^{H} p^h_{\mathrm{full},i}. \tag{25}$$

When extracting this map from the full attention tensor, we restrict each head's target-query row to visual positions and normalize it over those positions before averaging across heads.

CSRA+. We use the CSRA+ terminology and definition in Table 3. The common-space estimation procedure is described in Appendix C.1, with 200 RefCOCO calibration records and rank $r = 4$. A separate basis is estimated for each backbone, layer, and query head and frozen before evaluation. Head aggregation precedes the foreground-mask processing described below.

Frozen RefCOCO-to-COCO transfer. CSRA$^+$ (transfer) uses the same readout and foreground-mask processing as CSRA$^+$, differing only in the source of the common-subspace basis. Specifically, the rank-4 basis is fitted on RefCOCO and transferred to COCO category localization within the same backbone. For each layer and query head, let $P^h_{\mathrm{RefCOCO}} = U^h_{\mathrm{RefCOCO}}(U^h_{\mathrm{RefCOCO}})^\top$ be the frozen projector. The transferred readout is

$$\begin{aligned} q^h_{\mathrm{transfer}} &= (I - P^h_{\mathrm{RefCOCO}})q^h, \\ A_{\mathrm{transfer},i} &= \left[\frac{1}{H}\sum_{h=1}^{H}\frac{\langle q^h_{\mathrm{transfer}}, k_i^{g(h)}\rangle}{\sqrt{d_h}}\right]_+, \end{aligned} \tag{26}$$

where $[x]_+ = \max(x, 0)$. Signed residual logits are averaged across heads before clipping at zero. The resulting map undergoes the same maximum normalization, bilinear interpolation, and Otsu thresholding as CSRA$^+$. The RefCOCO basis remains fixed throughout COCO evaluation, with no COCO-side fitting or model parameter updates. The original visual keys are held fixed, and projected queries retain their scales without additional normalization.

Attribution baselines. We compare with CP-LRP and TAM, following the implementation settings adopted from TAM Li et al. (2025) in the main text. The resulting localization maps are evaluated using the foreground normalization, Otsu thresholding, and IoU aggregation described below.

C.3.3 Foreground masks and mIoU

Heatmap normalization and interpolation. Let $G \in \mathbb{R}^{H_p \times W_p}$ denote a localization map on the visual-token grid. We retain its nonnegative part and normalize by its maximum:

$$G^+ = [G]_+, \qquad \widehat{G} = \begin{cases} G^+ / \max_i G^+_i, & \max_i G^+_i > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

The normalized map is then bilinearly interpolated to the original image height and width with `align_corners=False`, yielding a full-resolution score map F. This processing is applied before foreground thresholding.

Ground-truth foreground. For RefCOCO, polygon coordinates are clipped to the original image bounds and rasterized with PIL to obtain the target mask. For COCO category localization, instance masks are decoded with `pycocotools` and combined by union for each image and category. Ground-truth masks and predicted foreground masks are compared at the original image resolution.

Otsu thresholding. We estimate a separate threshold for each nonconstant full-resolution score map using Otsu's method, implemented with the default settings of `skimage.filters.threshold_otsu`:

$$\tau = \mathrm{Otsu}(F), \qquad B(u,v) = \mathbf{1}[F(u,v) > \tau]. \tag{28}$$

The threshold is determined from the score map alone. A constant map produces an empty foreground prediction. The binary mask is evaluated directly without morphological postprocessing or connected-component selection.

Foreground IoU and averaging. For ground-truth mask M_n and predicted mask $B_{n\ell m}$ for target n, layer ℓ, and method m, we compute

$$\mathrm{IoU}_{n\ell m} = \frac{|B_{n\ell m} \cap M_n|}{|B_{n\ell m} \cup M_n|}, \qquad \mathrm{mIoU}_{\ell m} = \frac{100}{|\mathcal{E}_{\ell m}|} \sum_{n \in \mathcal{E}_{\ell m}} \mathrm{IoU}_{n\ell m}\,. \tag{29}$$

Here, $\mathcal{E}_{\ell m}$ contains the successfully evaluated targets for the corresponding layer and method. An empty union receives an IoU of zero. Each RefCOCO image–expression record receives equal weight; on COCO, each image–category pair receives equal weight. The reported mIoU is the arithmetic mean of these foreground IoUs, expressed as a percentage. Zero and constant heatmaps remain in the evaluation as empty predictions.

D Layer-wise Spatial Alignment and the Analysis Depth Range

Full-depth attention analysis. We examine spatial grounding across decoder depth to contextualize the layers used in our common-space analysis. For a decoder with L layers, we index layers by $\ell \in \{0, \ldots, L-1\}$ and define normalized depth as $d_\ell = \ell/(L-1)$. We use average precision (AP) to characterize how well visual-attention scores rank annotated target positions above background positions. This ranking-based measure complements thresholded localization: it assesses spatial ordering without constructing a foreground mask using Otsu's method.

Spatial alignment varies with depth. Figure 7 shows that attention-based grounding is generally weaker in the earliest layers and becomes stronger at intermediate depths. The high-response region is backbone-dependent: several curves attain strong responses at normalized depths around 0.6–0.7, while LLaVA exhibits a strong response closer to 0.4–0.45. The curves are non-monotonic, and their final layers generally fall below their intermediate-layer peaks. These observations motivate examining a range of depths that spans both stronger spatial alignment and its subsequent attenuation.

Choice of the 50%–100% depth range. We focus our intermediate-to-late-layer diagnostics on the latter half of the decoder, $d_\ell \in [0.5, 1]$. This provides a common analysis window that covers the strong-alignment region of several backbones together with the late-layer decline. Including both designs allows us to examine where removing a common query contribution exposes useful target correspondence and where the residual readout remains limited. The midpoint serves as a shared depth reference; the transition to stronger alignment can begin before it, as illustrated by LLaVA. Near-midpoint reference layers account for this variation.

Head-wise heterogeneity. The head-resolved view in Figure 8 complements the cross-model curves. High-AP responses are concentrated in a subset of layer–head pairs, with many stronger responses appearing around and after the middle of the decoder. Individual high-AP heads also persist at later layers. Consequently, a decline in a layer-level summary can coexist with useful spatial signals in particular heads. This heterogeneity is relevant to CSRA, whose common basis is estimated separately for each layer and query head, and supports examining multiple layers within the selected depth range.

Interpretation of late-layer behavior. The full-depth profiles provide an empirical reason to retain the deepest layers in the analysis: they expose settings in which spatial readout weakens despite stronger alignment at earlier depths. These native-attention measurements contextualize the layer dependence of CSRA; its localization results directly evaluate the residual readout. One possible

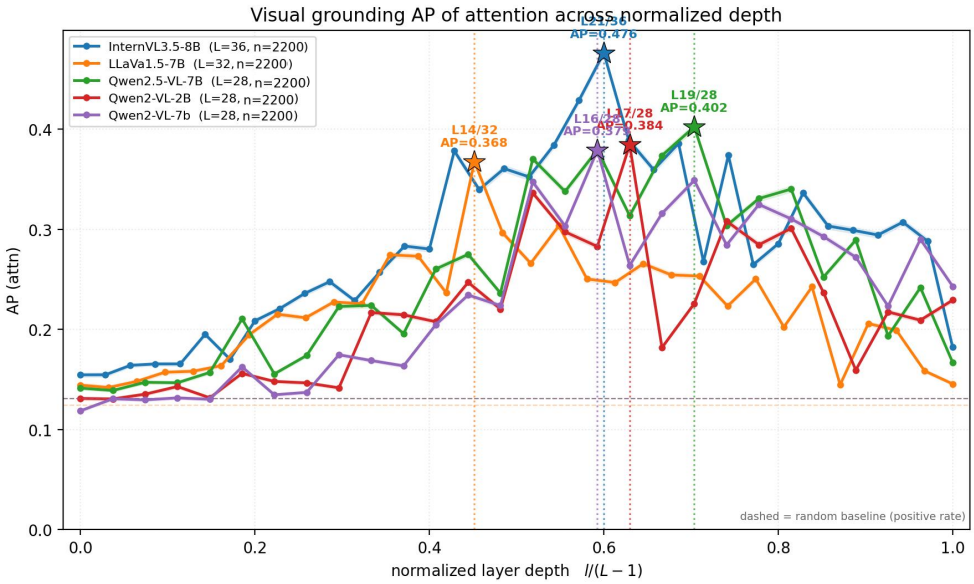


Figure 7: Spatial grounding AP of native visual attention across normalized decoder depth. The five checkpoints, decoder depths, and sample counts are listed in the legend. Stars annotate the displayed layer–AP pairs, and horizontal dashed lines show the positive-rate references used in the plot. Stronger grounding occurs at intermediate to later depths, with backbone-dependent peaks and a subsequent decline toward the output end.

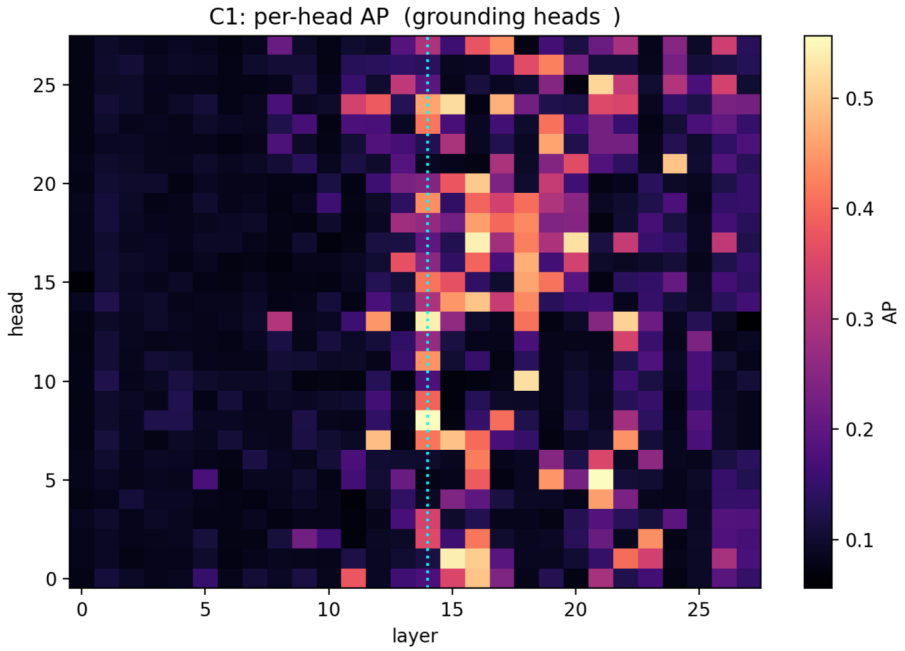


Figure 8: Head-wise spatial grounding AP across decoder layers in the illustrated configuration. Columns index layers, rows index attention heads, and color indicates the AP of each layer–head pair. The cyan vertical guide highlights layer index 14. The distribution reveals substantial head-wise heterogeneity: stronger responses become more frequent around intermediate depths, while some individual heads retain high AP at later depths.

interpretation is that representations become increasingly organized for token prediction near the output, changing how spatial correspondence is expressed in QK interactions. The present AP analysis establishes depth-dependent spatial readability, while this computational explanation remains a hypothesis.