Retrieval Heads Meet Vision: Uncovering How VLMs Locate and Extract Visual Information

Chanho Park [1] **Daehyeon Choi** [1] **Jihyun Lee** [2] **Minhyuk Sung** [1]

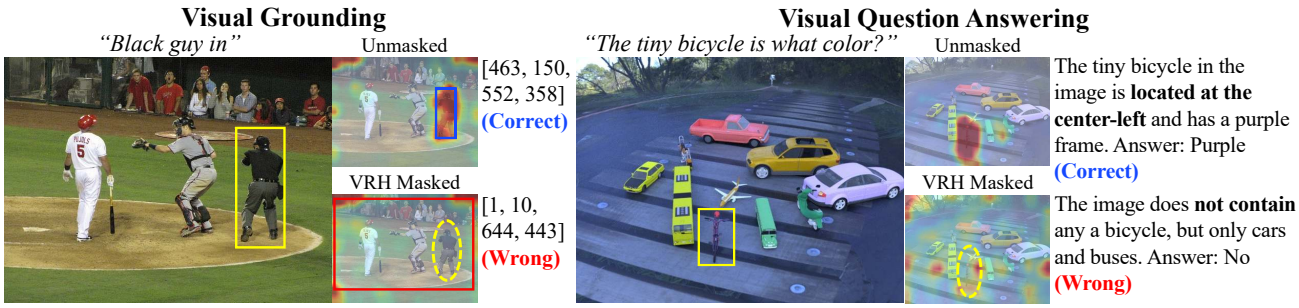


Figure 1. **Visual Retrieval Heads mediate visual reference resolution in VLMs.** We identify a sparse set of attention heads, *Visual Retrieval Heads* (VRHs), that route visual evidence from the image region referred to by the text prompt to the output. On visual grounding (**left**), masking the VRHs redirects prediction-stage attention away from the ground-truth referent and causes the model to output a delocalized bounding box. The same heads, discovered only from grounding, also transfer to visual question answering (**right**): masking them removes attention to the referred object and leads to a fluent but visually unfaithful answer. Yellow boxes denote ground-truth referents; dashed yellow ellipses indicate regions where attention vanishes after VRH masking.

Abstract

Vision-language models (VLMs) can locate an image region referred to by a text prompt and route the corresponding visual evidence to the output, yet the internal mechanism behind this behavior is not understood. Inspired by retrieval heads in large language models, we ask whether VLMs contain an analogous mechanism for visual retrieval. We answer affirmatively by introducing *Visual Retrieval Heads* (VRHs), a small subset of attention heads (about 1.7–2.6%) that are causally responsible for grounding text descriptions to image regions. To find them, we recast existing head-scoring methods under a unified design space over query tokens, key aggregation, and cross-sample aggregation. We then show that scoring attention from output prediction tokens with a sum over the ground-truth referent region most reliably identifies causal heads. Across four VLMs and five referring-expression benchmarks, masking only the top 20 VRHs reduces grounding accuracy by up to 80 percentage points, while masking the same number of random heads has little effect. Beyond replicating the causal-sparse-universal triad established for text retrieval heads, VRHs exhibit several properties not previously reported: they *generalize across visual reference tasks*, remaining causal on attribute, spatial, counting, and visual-math benchmarks despite being discovered through bounding-box prediction; they are *functionally specific*, preserving output format while corrupting localization; and they are *architecturally shared*, transferring causally across VLMs that share an LLM backbone but differ in vision encoder, projector, and instruction tuning.

1. Introduction

How do foundation models locate a specific piece of information buried in a long input context? This "needle-in-a-haystack" question has recently been answered, at least partially, for large language models (LLMs): a small set of attention heads called *retrieval heads* (Wu et al., 2025) is largely responsible for routing relevant input tokens to the output. These heads exhibit three properties that together make them a meaningful unit of analysis. They are

[1]KAIST [2]Independent Researcher. Correspondence to: Minhyuk Sung <mhsung@kaist.ac.kr>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

causal: ablating them collapses long-context retrieval while leaving fluency intact. They are *sparse*: fewer than 5% of all heads carry the behavior. And they are *universal*: they appear across model families and scales, suggesting that long-context retrieval converges onto a shared internal mechanism rather than an idiosyncratic implementation.

This observation motivates our key question: whether *Vision-Language Models (VLMs)* contain an analogous mechanism for *visual* retrieval—locating an image region referred to by a text prompt and routing it to the output. Existing extensions of retrieval-head analysis to VLMs do not yet answer it. OCR Heads (Baek et al., 2025) and VERA (Pei et al., 2026) identify heads that retrieve textual evidence from images, but only when the image is itself a rendering of text; the retrieval target is still a string of characters, not a general visual concept. Localization Heads (Kang et al., 2025) move beyond rendered text and aim to identify heads whose attention maps align with referred objects, but the analysis stops at attention shape: the heads are *not* characterized by causal necessity.

In this work we ask whether causal, universal, and sparse retrieval heads exist in VLMs for general image retrieval, and how to find them. We answer affirmatively by introducing *Visual Retrieval Heads* (VRHs), a small subset of attention heads that are causally responsible for grounding text descriptions to image regions. We show that these heads share the defining properties of their text-only counterparts.

We use referring-expression visual grounding (Yu et al., 2016) as our probing task: given an image and a referring expression, the model must output a bounding box for the referent. This task makes the retrieval target spatially explicit; the ground-truth bounding box defines exactly which visual tokens the model needs to consult. This lets us define a head's contribution to retrieval by how strongly it attends from the model's prediction-stage tokens to the visual tokens inside that box.

Beyond identifying VRHs, our contributions are three-fold. First, on the *methodology* side, we observe that existing head-scoring methods can all be written in a single form, with a per-head score obtained by aggregating attention weights over a set of *key tokens* (the tokens to be retrieved) for each *query token*, and then aggregating these per-query scores across all queries. Under this view, prior methods differ along three axes: the query token set (input prompt tokens vs. output prediction tokens), the key aggregation (an argmax indicator for whether the most-attended key lies in the retrieval region vs. a summation of attention weights over the region), and the cross-sample aggregation (averaging scores across samples vs. voting). For visual grounding, we let the key tokens be the visual tokens that overlap the ground-truth bounding box, and we search this design space directly, scoring each variant by the drop in grounding accuracy when its top-ranked heads are masked. We find that the choice of key aggregation and cross-sample aggregation matters little, but the choice of query tokens matters a great deal: scoring at the *output* prediction stage with a sum over the retrieval region consistently identifies the most causally important heads. We adopt it as our default scoring function.

Second, on the *empirical* side, we demonstrate the causality, sparsity, and universality of the VRHs identified by our scoring function through comprehensive experiments across four VLMs (Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), InternVL3-8B (Zhu et al., 2025), InternVL3.5-8B (Wang et al., 2025b)) and five grounding benchmarks (RefCOCO/+/g (Yu et al., 2016; Mao et al., 2016), RefSpatialBench (Zhou et al., 2025), Toloka (Ustalov et al., 2023)). Masking the top 20 VRHs—about 1.7–2.6% of all heads—reduces grounding accuracy by up to 80 percentage points, while masking the same number of random heads has negligible effect, demonstrating their causal role.

Third, most importantly, our *characterization* of VRHs reveals four properties that distinguish them from previously studied heads. (i) VRHs *generalize across visual reference tasks*: although we discover them through bounding-box prediction, they remain causal on attribute, spatial reasoning, counting, and visual math benchmarks (VAW, Spatial457, SpatialRGPT, CV-Bench, MMStar, CountBenchQA, MathVista) (Pham et al., 2021; Wang et al., 2025c; Cheng et al., 2024; Tong et al., 2024; Chen et al., 2024; Paiss et al., 2023; Lu et al., 2024), suggesting that the mechanism they implement is visual reference resolution in general, not object grounding in particular. (ii) VRHs are *functionally specific*: masking them preserves output-format validity but the predicted box is delocalized, the visual analogue of the "fluent-but-not-factual" failure mode reported for text retrieval heads. (iii) VRHs are *architecturally shared*: VRHs found in one VLM remain causally critical when masked in another VLM that shares the same LLM backbone but differs in vision encoder, projector, and instruction-tuning recipe. (iv) VRH detection are *stably detectable*: results are consistent across output-token choices and scoring samples, with as few as 5 to 10 grounding examples recovering nearly the same top-ranked set as the full 200-example set.

2. Related Work

We discuss prior works on functionally specialized attention heads in large language models (LLMs) and vision-language models (VLMs). Specifically, we focus on methods for identifying heads specialized for *retrieval*-related tasks, which are most closely related to our setting.

Most existing methods in this domain identify such specialized heads by quantifying the importance of each head

Table 1. **Comparison of prior works on retrieval-related specialized heads.**

	Retrieval Heads (Wu et al., 2025)	QRHead (Zhang et al., 2025)	OCR Heads (Baek et al., 2025)	VERA (Pei et al., 2026)	Loc. Heads (Kang et al., 2025)	**VRH (Ours)**
Input Source	Text	Text	Visual document	Visual document	General image	General image
Causality	✓	✓	✓	✓	✗	✓

based on inference-time attention patterns (Wu et al., 2025; Zhang et al., 2025; Baek et al., 2025; Pei et al., 2026). A pioneering work along this direction is Retrieval Heads (Wu et al., 2025), which detects heads specialized for text retrieval in LLMs by measuring how often the output answer token most attends to the correct input token to be retrieved, i.e., the *needle* token. QRHead (Zhang et al., 2025) is a follow-up work showing that analyzing attention patterns from input prompt tokens (*cf.* output tokens in Retrieval Heads) to the input retrieval source tokens is more effective for detecting specialized heads in real-world scenarios.

Beyond text-only retrieval in LLMs, OCR Heads (Baek et al., 2025) is one of the earliest works to extend the head-identification method of Retrieval Heads to vision-language models (VLMs), discovering specialized heads for optical character recognition. Similarly, VERA (Pei et al., 2026) identifies VLM heads specialized for long-context visual document retrieval. However, both of these VLM-based specialized heads are limited to *visual text documents* as the input retrieval source. The first work on specialized heads for visual retrieval from *general images* is Localization Heads (Kang et al., 2025), but it does not verify the causality of the detected heads, i.e., their causal effect on retrieval performance, which is a key property of retrieval specialized heads emphasized in all the aforementioned works (Wu et al., 2025; Zhang et al., 2025; Baek et al., 2025; Pei et al., 2026).

In Tab. 1, we summarize the characteristics of these prior works. To the best of our knowledge, our work is the first to identify a set of attention heads specialized for visual retrieval from *general images* while exhibiting *causality* in the model's visual retrieval performance. Note that we further discuss and compare the methodological differences among these prior works in Sec. 3.2.

3. Toward Identifying Visual Retrieval Heads

We introduce our method for identifying *Visual Retrieval Heads* (VRHs), a subset of attention heads within vision-language models (VLMs) specialized for retrieving visual information relevant to an input text prompt. Our detected VRHs exhibit important desired properties of functionally specialized heads: (1) *causality*: VRHs are causally responsible for visual retrieval performance, such that masking these heads significantly degrades the VLM's retrieval behavior; (2) *universality*: VRHs can be found across multiple VLM families and benchmarks; and (3) *sparsity*: VRHs constitute only a small subset of attention heads, revealing that visual retrieval behavior in VLMs is concentrated in a few heads. To the best of our knowledge, ours is the first work to discover specialized heads in VLMs that are causally responsible for visual retrieval from general images.

In what follows, we first introduce our setup, where visual grounding is used as our probing task (Sec. 3.1). We then show that existing *head-scoring methods*, i.e., methods for quantifying the importance of individual heads for a target functionality, can be expressed under a unified formulation. This formulation allows us to systematically explore the design space of scoring methods for detecting VRHs (Sec. 3.2). Based on this analysis, we propose our VRH scoring method by holistically examining different scoring method variants through causal validation, *i.e.*, by measuring how much visual grounding accuracy degrades when the detected heads are masked (Sec. 3.3).

3.1. Visual Grounding as Probing Task

We use visual grounding as a probing task for analyzing the visual retrieval behavior of VLMs and detecting VRHs. In this task, a VLM is asked to localize the object in the input image referred to by the input text prompt. More specifically, each sample in a visual grounding dataset is represented as $s = (I_s, q_s, B_s)$, where I_s is an image, q_s is a referring expression, and B_s is the ground-truth bounding box of the referred object. Given (I_s, q_s), the VLM generates a text token sequence that is parsed into a predicted box $\hat{B}_s$. A prediction is judged correct if its intersection-over-union (IoU) with the ground-truth box exceeds a threshold, $\mathrm{IoU}(\hat{B}_s, B_s) \geq \tau$, where τ is typically set to 0.5.

3.2. Unified Formulation of Head-Scoring Methods

To effectively design a method to detect VRHs responsible for this visual grounding task, we first review prior works on detecting functionally specialized heads in other retrieval-related tasks in LLMs and VLMs (e.g., text retrieval (Wu et al., 2025; Zhang et al., 2025), optical character recognition (Baek et al., 2025; Pei et al., 2026)). The key idea in these works is to introduce a *head-scoring method* to quantify the importance of each attention head for the target task and detect the specialized heads based on it. Here, the importance of each head is typically scored based on its attention weights computed between: (1) *a set of query to-*

Table 2. **Design choices in head-scoring methods.** We summarize how existing methods instantiate the main components of Eq. 1. For $\mathcal{Q}_s$, some methods use only a few representative tokens from the indicated input or output token set, i.e., $\mathcal{Q}_s^{\text{in}}$ or $\mathcal{Q}_s^{\text{out}}$; we slightly abuse notation for abstraction. †Localization Heads is a special case that does not directly use key aggregation, but instead selects heads based on the spatial entropy of attention weights.

	Retrieval Heads (Wu et al., 2025)	**QRHead** (Zhang et al., 2025)	**OCR Heads** (Baek et al., 2025)	**VERA** (Pei et al., 2026)	**Loc. Head** (Kang et al., 2025)
Query Token Set $\mathcal{Q}_s$	$\mathcal{Q}_s^{\text{out}}$	$\mathcal{Q}_s^{\text{in}}$	$\mathcal{Q}_s^{\text{out}}$	$\mathcal{Q}_s^{\text{out}}$	$\mathcal{Q}_s^{\text{in}}$
Key Aggregation Φ	Φ^{argmax}	Φ^{sum}	Φ^{argmax}	Φ^{sum}	N/A†
Sample Aggregation Ω	Ω^{avg}	Ω^{avg}	Ω^{avg}	Ω^{avg}	Ω^{freq}

kens ($\mathcal{Q}_s$) where retrieval behavior is expected to occur, and (2) *a set of key tokens* ($\mathcal{K}_s$) that correspond to the correct evidence region to be retrieved from the input. In our visual grounding setting, this evidence region is the annotated target region; therefore, we instantiate $\mathcal{K}_s$ as the set of visual tokens whose patches overlap with the annotated bounding box B_s.

For example, existing text retrieval head detection methods in LLMs (Wu et al., 2025; Zhang et al., 2025) assign high importance to heads that place strong attention from either *output tokens* (Wu et al., 2025) (which express the retrieval prediction) or *input prompt tokens* (Zhang et al., 2025) (which specify what should be retrieved) onto input tokens corresponding to the correct retrieval region.

More specifically, let $\bar{\Theta}_{l,h}$ denote the importance score of head h at layer l, computed by aggregating the per-sample head importance scores $\Theta_{l,h}(s)$ over samples $s \in \mathcal{S}$ in the scoring dataset. In most of the existing methods (Wu et al., 2025; Zhang et al., 2025; Baek et al., 2025; Pei et al., 2026), both the final head score $\bar{\Theta}_{l,h}$ and the per-sample head score $\Theta_{l,h}(s)$ can be expressed under the following unified formulations:

$$\begin{aligned} \bar{\Theta}_{l,h} &= \Omega_{s\in\mathcal{S}} \underbrace{\left[\Theta_{l,h}(s)\right]}_{\text{per-sample score}}, \\ \Theta_{l,h}(s) &= \frac{1}{|\mathcal{Q}_s|} \sum_{t\in\mathcal{Q}_s} \Phi_{j\in\mathcal{K}_s}\left[\alpha_{l,h}(t,j)\right], \end{aligned} \tag{1}$$

where $\Omega_{s\in\mathcal{S}}$ denotes a cross-sample aggregation function that aggregates the per-sample scores $\Theta_{l,h}(s)$ over all samples $s \in \mathcal{S}$. In the above equation, each per-sample score $\Theta_{l,h}(s)$ is computed by summarizing the attention weights $\alpha_{l,h}(t,j)$, which reflect how strongly head (l,h) attends from a set of query tokens $t \in \mathcal{Q}_s$ to a set of key tokens $j \in \mathcal{K}_s$. To this end, the key aggregation function $\Phi_{j\in\mathcal{K}_s}$ first aggregates the attention weights over the set of key tokens into a scalar per-query score, and these per-query scores are then averaged over all selected query tokens to finally obtain the per-sample score $\Theta_{l,h}(s)$.

Under this unified formulation, existing methods (Wu et al., 2025; Zhang et al., 2025; Baek et al., 2025; Pei et al., 2026) mainly differ in their choices of query token set $\mathcal{Q}_s$, key aggregation function Φ, and cross-sample aggregation function Ω. We next briefly introduce the common design choices for these variables used in prior work (Wu et al., 2025; Zhang et al., 2025; Pei et al., 2026; Kang et al., 2025; Baek et al., 2025).

- **Query token set** $\mathcal{Q}_s$: (i) *Input query tokens* $\mathcal{Q}_s^{\text{in}}$, which specify the retrieval target, or (ii) *output tokens* $\mathcal{Q}_s^{\text{out}}$, where the retrieved information is generated.
- **Key aggregation function** Φ: (i) *An argmax indicator* Φ^{argmax}, which returns a binary score for whether the most-attended visible key token falls within $\mathcal{K}_s$, or (ii) *a region-sum score* Φ^{sum}, which returns the total attention mass assigned to key tokens in $\mathcal{K}_s$.
- **Cross-sample aggregation function** Ω: (i) *average aggregation* Ω^{avg}, which averages per-sample scores across the dataset, and (ii) *selection-frequency aggregation* Ω^{freq}, which ranks heads by how often they are selected across samples.

Tab. 2 summarizes how existing methods instantiate these choices. Based on this unified view, we define a design space for detecting VRHs: we construct 2^3 head-scoring variants by varying the query-token set $\mathcal{Q}_s$, key aggregation Φ, and cross-sample aggregation Ω according to the common choices above.

3.3. Causal Validation of Head-Scoring Variants

The unified scoring formulation above defines a family of candidate scoring methods for identifying VRHs, depending on the choice of the set of query tokens, key aggregation, and cross-sample aggregation. While these scores measure how strongly a head's attention aligns with the target visual region, attention alignment alone does not establish that the head is functionally necessary for visual retrieval. As emphasized in prior work (Zhang et al., 2025; Wu et al., 2025; Baek et al., 2025; Pei et al., 2026), *causality* is essential for interpreting a head as a meaningful functional unit rather than an attention-correlated artifact. We therefore perform

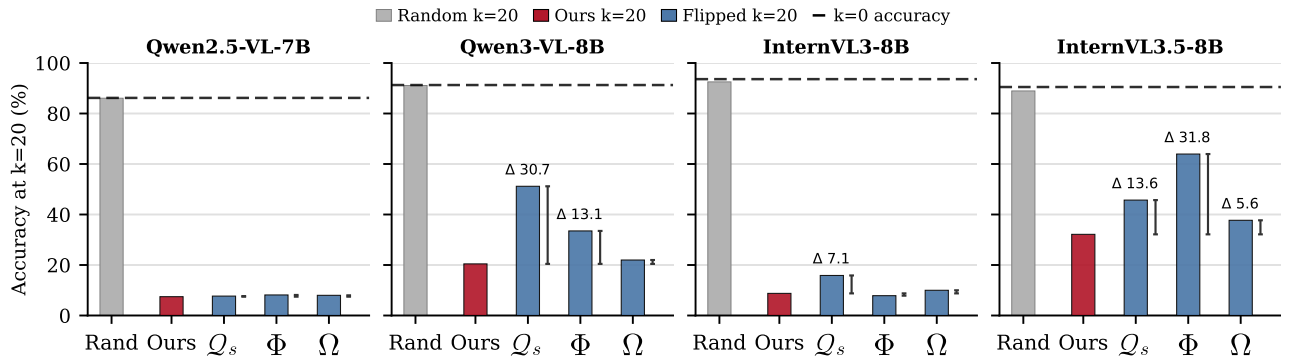


Figure 2. **Causal validation of head-scoring variants on RefCOCO (Yu et al., 2016) across four VLMs (Bai et al., 2025b;a; Zhu et al., 2025; Wang et al., 2025b).** We mask the top-20 heads detected by each variant and report visual grounding accuracy; lower accuracy indicates that the detected heads are more causally important. *Ours* denotes our default variant ($\mathcal{Q}_s^{\text{out}}, \Phi^{\text{sum}}, \Omega^{\text{mean}}$). The $\mathcal{Q}_s$, Φ, and Ω ablations each change one component of this default design at a time, while *Random* masks 20 randomly selected heads. The dashed line shows the unmasked model ($k = 0$), and Δ denotes the accuracy gap relative to *Ours*.

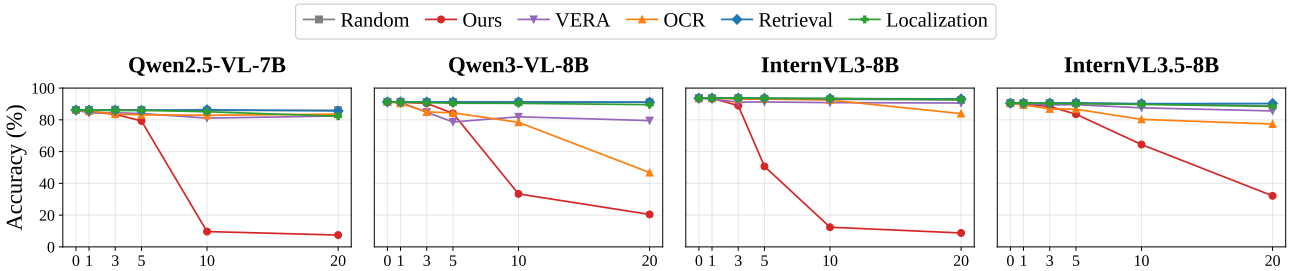


Figure 3. **Comparison with existing heads.** We compare VRHs with Retrieval Heads (Wu et al., 2025), OCR Heads (Baek et al., 2025), VERA (Pei et al., 2026), Localization Heads (Kang et al., 2025), and randomly selected heads on the RefCOCO visual grounding benchmark (Yu et al., 2016). The x-axis denotes the number of masked heads (k). Masking the top-k VRHs causes substantially larger drops in grounding accuracy than masking these baselines, indicating that VRHs capture a distinct set of heads causally important for visual grounding.

causal validation: for each scoring variant, we select the top-ranked heads, mask them during inference, and measure the resulting drop in visual grounding accuracy.

We perform this causal validation and empirically choose ($\mathcal{Q}_s^{\text{out}}, \Phi^{\text{sum}}, \Omega^{\text{mean}}$) as the default design for VRH detection. Fig. 2 reports this validation across four VLMs (Bai et al., 2025b;a; Zhu et al., 2025; Wang et al., 2025b) on RefCOCO (Yu et al., 2016), ablating one design component at a time and measuring whether the resulting top-ranked heads are causally important for visual grounding. Masking heads detected by the default variant consistently causes large accuracy drops relative to the unmasked model ($k = 0$), showing that this score identifies heads that are causally important for visual grounding. The ablations further show that the query-token set is the most important design choice: variants that score attention from output query tokens identify heads whose removal causes substantially larger and more consistent accuracy drops than variants based on input query tokens. This suggests that the strongest causal retrieval signal emerges when the model is autoregressively producing the grounded answer, rather than when it only encodes the referring expression in the prompt.

Comparison with existing specialized heads. We next ask whether VRHs simply overlap with specialized heads identified by prior methods, or instead reveal a distinct mechanism for visual reference resolution in general images. As discussed in Sec. 2, most existing methods focus on text retrieval (Wu et al., 2025; Zhang et al., 2025) or visual retrieval from document-like images (Baek et al., 2025; Pei et al., 2026). Localization Heads (Kang et al., 2025) move beyond rendered text toward general images, but characterize heads by spatial concentration rather than by their functional role. In contrast, VRHs are identified by directly scoring how strongly each head attends from output tokens to the target image region, and are validated by their *causal* effect on visual retrieval performance. Using the same causal validation protocol on RefCOCO (Yu et al., 2016), Fig. 3 shows that masking VRHs causes the largest degradation in grounding accuracy across four VLMs, substantially exceeding the effect of masking random heads or those identified by prior methods.

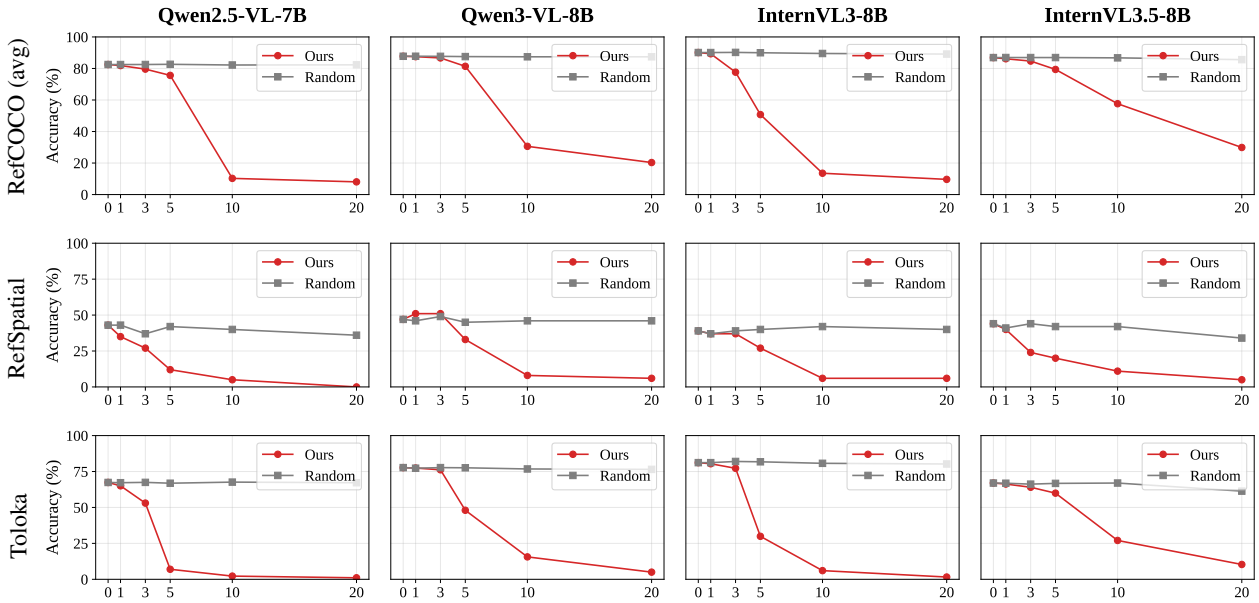


Figure 4. **Universality of VRHs across models and benchmarks.** We report visual grounding accuracy after masking the top-k VRHs across four VLMs and three benchmark families: RefCOCO/+/g (Yu et al., 2016; Mao et al., 2016), RefSpatialBench (Zhou et al., 2025), and Toloka (Ustalov et al., 2023). The x-axis denotes the number of masked heads (k). Masking only a few VRHs consistently causes large accuracy drops, while masking random heads has little effect. This broad causal effect shows that VRHs are not architecture- or dataset-specific, but instead capture a shared attention-head mechanism for visual grounding.

4. Visual Retrieval Heads Are Universal and Sparse

Sec. 3.3 establishes that our scoring method identifies heads that are causally important for visual grounding. We now ask a stronger question: are these heads merely an artifact of a particular model, dataset, or scoring setup, or do they reveal a general mechanism by which VLMs perform visual retrieval? Following the criteria used to characterize retrieval heads in LLMs (Wu et al., 2025; Zhang et al., 2025), we evaluate whether VRHs are both *universal* and *sparse*.

Universality. VRHs are universal, consistently emerging as a causal visual-retrieval unit across diverse model families and grounding benchmarks. We evaluate four VLMs (Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), InternVL3-8B (Zhu et al., 2025), and InternVL3.5-8B (Wang et al., 2025b)) on five grounding datasets spanning three benchmark families: (RefCOCO/+/g (Yu et al., 2016; Mao et al., 2016), RefSpatialBench (Zhou et al., 2025), and Toloka (Ustalov et al., 2023)). As shown in Fig. 4, masking the top-ranked VRHs consistently causes large drops in grounding accuracy across all model–benchmark pairs, whereas masking the same number of random heads has little effect. This indicates that, despite differences in architecture, training recipe, and benchmark distribution, diverse VLMs rely on a sparse set of attention heads to route visual evidence from the referred image region to the output.

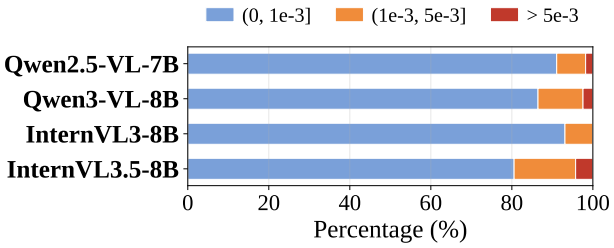


Figure 5. **Sparsity of VRHs.** Distribution of head scores for VRH detection on RefCOCO (Yu et al., 2016) across four VLMs (Bai et al., 2025b;a; Zhu et al., 2025; Wang et al., 2025b). Only a small fraction of heads receive non-negligible scores.

Sparsity. VRHs are highly sparse, suggesting that visual retrieval is concentrated in a small set of specialized heads rather than diffusely implemented throughout the transformer. Fig. 5 shows the distribution of VRH scores on RefCOCO across the four VLMs. Fewer than 15% of heads receive a non-negligible score (above 10^{-3}), and only 0–4% receive a high score (above 5×10^{-3}), depending on the model. Despite occupying only a small fraction of each model, these heads have a large causal effect: masking them is sufficient to substantially disrupt grounding. Thus, visual retrieval in VLMs mirrors the sparse-but-causal organization of retrieval heads in LLMs (Wu et al., 2025).

Table 3. **Generalization across visual reference tasks.** Masking the top-20 VRHs detected on RefCOCO (Yu et al., 2016) consistently degrades performance on seven VQA benchmarks (Pham et al., 2021; Wang et al., 2025c; Cheng et al., 2024; Tong et al., 2024; Chen et al., 2024; Paiss et al., 2023; Lu et al., 2024) spanning attributes, spatial reasoning, counting, vision-centric understanding, and visual math, while random-head masking has little effect.

Model	Condition	VAW	Spatial457	SpatialRGPT	CV-Bench	MMStar	CountBenchQA	MathVista
Qwen2.5-VL-7B	Baseline (k=0)	63.8	68.2	36.1	75.1	58.3	83.9	71.5
	Random (k=20)	63.7 (-0.1)	67.2 (-1.0)	39.5 (+3.4)	71.9 (-3.2)	56.8 (-1.5)	84.1 (+0.2)	72.0 (+0.5)
	VRH (k=20)	40.5 (-23.3)	36.0 (-32.2)	21.8 (-14.3)	53.3 (-21.8)	47.9 (-10.4)	61.5 (-22.4)	61.7 (-9.8)
Qwen3-VL-8B	Baseline (k=0)	74.8	78.2	50.8	87.6	63.0	88.4	66.5
	Random (k=20)	74.3 (-0.5)	76.4 (-1.8)	46.2 (-4.6)	87.1 (-0.5)	63.3 (+0.3)	89.6 (+1.2)	69.4 (+2.9)
	VRH (k=20)	51.0 (-23.8)	36.9 (-41.3)	22.9 (-27.9)	68.7 (-18.9)	47.7 (-15.3)	74.5 (-13.9)	43.0 (-23.5)
InternVL3-8B	Baseline (k=0)	73.1	73.1	50.0	85.9	67.9	81.3	80.0
	Random (k=20)	71.9 (-1.2)	72.4 (-0.7)	49.6 (-0.4)	86.7 (+0.8)	67.7 (-0.2)	81.9 (+0.6)	79.8 (-0.2)
	VRH (k=20)	47.6 (-25.5)	37.3 (-35.8)	38.0 (-12.0)	69.3 (-16.6)	47.6 (-20.3)	60.1 (-21.2)	65.7 (-14.3)
InternVL3.5-8B	Baseline (k=0)	72.6	66.3	50.4	80.3	64.9	78.0	78.9
	Random (k=20)	72.0 (-0.6)	64.3 (-2.0)	51.5 (+1.1)	82.8 (+2.5)	64.4 (-0.5)	79.0 (+1.0)	79.4 (+0.5)
	VRH (k=20)	54.4 (-18.2)	39.3 (-27.0)	41.0 (-9.4)	68.6 (-11.7)	51.5 (-13.4)	65.2 (-12.8)	66.3 (-12.6)

5. Characterizing Visual Retrieval Heads

In previous sections (Sec. 3–4), we established that VRHs are causal, universal, and sparse—properties shared with existing retrieval heads for text or visual text (Wu et al., 2025; Baek et al., 2025; Pei et al., 2026; Zhang et al., 2025). We now show that VRHs exhibit important additional properties: (i) VRHs *generalize across visual reference tasks* beyond the grounding task used to detect them (Sec. 5.1); (ii) VRHs are *functionally specific* to visual retrieval, selectively disrupting reference resolution without degrading general generation (Sec. 5.2); (iii) VRHs are *architecturally shared* across VLMs built on the same LLM backbone, indicating they reside in a reusable language-model circuit (Sec. 5.3); and (iv) VRHs are *stably detectable*, remaining consistent across output-token choices and requiring only a handful of annotated examples (Sec. 5.4). Note that, while we mainly focus on characterizing the properties of VRHs in this section, we kindly refer readers to Sec. B in the Appendix, where we also show that VRHs can also be utilized to improve visual grounding accuracy via direct preference optimization (DPO).

5.1. VRHs Generalize Across Visual Reference Tasks

We first show that VRHs effectively generalize beyond visual grounding: they implement a more general mechanism for visual reference resolution required by diverse vision question-answering tasks involving attributes, spatial reasoning, counting, vision-centric understanding, and visual math. This cross-task generalization importantly distinguishes VRHs from specialized heads studied in prior work on LLMs and VLMs (Wu et al., 2025; Zhang et al., 2025; Baek et al., 2025; Pei et al., 2026), which are typically identified and evaluated within a single task, leaving open whether the discovered heads support a broader cognitive function or only the particular evaluation format used to find them.

We validate this by taking VRHs detected on a visual grounding benchmark, RefCOCO (Yu et al., 2016), and testing their causal effects on seven VQA benchmarks spanning the aforementioned tasks: VAW (Pham et al., 2021), Spatial457 (Wang et al., 2025c), SpatialRGPT (Cheng et al., 2024), CV-Bench (Tong et al., 2024), MMStar (Chen et al., 2024), CountBenchQA (Paiss et al., 2023), and MathVista (Lu et al., 2024). As shown in Tab. 3, masking VRHs consistently degrades performance across all seven benchmarks and all four VLMs, while masking the same number of random heads has little effect. These results reveal that, although VRHs are discovered from visual grounding prompts that require generating *bounding-box coordinates* for the referred object, they capture a general visual reference mechanism required across diverse tasks.

5.2. VRHs are Functionally Specific

We next show that VRHs are functionally specific: masking them does not simply degrade generation, but selectively disrupts visual reference resolution.

For example, on the visual grounding task, we observe that masking VRHs leads to mislocalized bounding-box predictions, while the bounding-box syntax remains correct. In Fig. 7, we further decompose visual-grounding failures under VRH masking versus random-head masking into two types: (a) *parse failure*, where the output does not preserve the bounding-box syntax and cannot be parsed, and (b) *mislocation*, where the output is a syntactically valid bounding box but is spatially mislocalized.

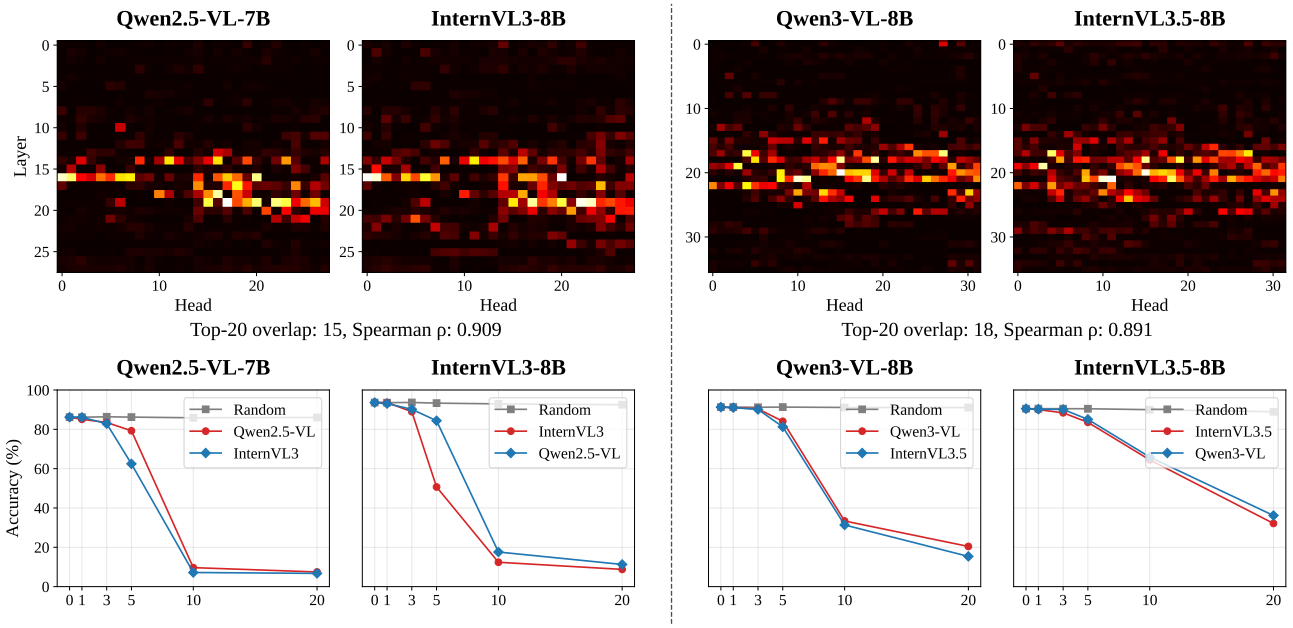


Figure 6. **VRH detection consistency across same-backbone VLMs. Top row:** Layer-head score maps for same-backbone VLMs, where each cell denotes the VRH score $\bar{\Theta}_{l,h}$ at layer l and head h; high-scoring heads appear at similar layer-head locations across same-backbone VLMs. **Bottom row:** Cross-model masking results, where the top-k VRHs identified in one model are masked at the same layer-head indices in the other model. The x-axis denotes the number of masked heads (k).

We find that masking VRHs significantly increases mislocation rates while parse failures remain rare. This contrast mirrors the "fluent-but-not-factual" failure mode reported for text retrieval heads (Wu et al., 2025; Zhang et al., 2025): masking the specialized heads preserves the surface form of the output, but corrupts the factuality of the evidence-grounded content. Fig. 1 illustrates this failure mode qualitatively: masked models produce syntactically valid bounding boxes that localize incorrect regions.

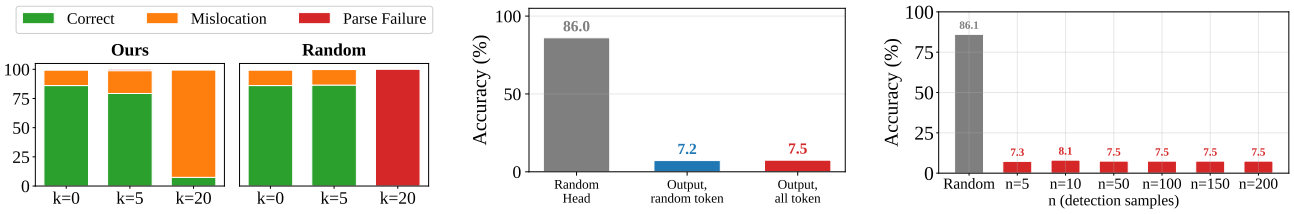
5.3. VRHs are Architecturally Shared

We next show that VRHs are not specific to a single VLM instantiation, but reflect an architecturally shared mechanism within the LLM backbone. In particular, VRHs detected in one VLM remain meaningful in another VLM that shares the same LLM backbone, even when the vision encoder, projector, and instruction tuning differ. We test this with pairs of same-backbone VLMs in two ways. First, we compare the *layer-head score map*, where each entry at layer l and head h records the dataset-level VRH score $\bar{\Theta}_{l,h}$. As shown in Fig. 6(a), same-backbone VLMs exhibit similar high-scoring layer-head locations, with substantial top-20 overlap and high rank correlation. Second, we perform *cross-model masking*: we detect the top-20 VRHs in one VLM and mask the same layer-head indices in another. As shown in Fig. 6(b), transferred VRHs cause a large grounding-accuracy drop in the target model, approaching the effect of masking its own VRHs and far exceeding random-head masking. In summary, these results show that VRHs are architecturally shared across same-backbone VLMs, suggesting that visual retrieval relies on a reusable circuit in the language backbone rather than an idiosyncratic VLM-specific implementation.

5.4. VRHs are Stably Detectable

We finally show that VRHs can be detected reliably across two key sources of variation in the scoring variables: the choice of output query tokens and the number of annotated grounding examples.

Robustness across output tokens. VRH detection is stable across output-token query choices: heads identified using different subsets of output tokens are equally causal. To test this, we select the top-20 VRHs under different query-token choices—a single random output token or all output tokens—and measure grounding accuracy after masking them. As shown in Fig. 7, both choices cause a comparable accuracy drop (7.2% and 7.5%, respectively), while masking the same number of random heads barely affects performance (86.0%). This confirms that our head-scoring method consistently identifies causally important VRHs regardless of the specific output-token query set.



(a) Functional specificity (b) Output-token robustness (c) Sample-size robustness

Figure 7. **VRHs are functionally specific and stably detectable.** All experiments use Qwen2.5-VL-7B (Bai et al., 2025b) on RefCOCO (Yu et al., 2016) with k=20 masked heads unless otherwise noted. **(a)** Visual grounding failure decomposition under VRH masking vs. random-head masking. *Mislocation* denotes syntactically valid but mislocalized box prediction, while *parse failure* denotes output that cannot be parsed as a bounding box. **(b)** Grounding accuracy when masking VRHs detected from a single random output token, all output tokens, or random heads. **(c)** Grounding accuracy when masking VRHs detected with varying numbers of scoring samples n.

Robustness across sample size. VRH detection is highly sample-efficient: only a few grounding annotations are needed to recover the same top-ranked heads as a much larger scoring set. This indicates that VRHs exhibit a strong, repeatable attention signature, rather than a weak pattern that emerges only after extensive averaging. To quantify this, we vary the number of samples used to estimate $\bar{\Theta}_{l,h}$ and compare each resulting top-20 head set with the reference set computed from 200 examples. As shown in Fig. 7, as few as 5–10 examples recover nearly the same top-ranked VRHs as the full scoring set. This makes VRH detection practical: a small number of bounding-box annotations is sufficient to reliably identify VRHs.

6. Conclusion

We introduce Visual Retrieval Heads (VRHs), a sparse set of attention heads that route visual evidence from text-referred image regions to the model output. Using visual grounding as a probing task, we identify VRHs by scoring attention from output tokens to ground-truth visual regions, and validate their causal role through targeted masking across multiple VLMs and grounding benchmarks. These analyses show that VRHs are sparse, causal, and universal. Importantly, their role extends beyond grounding: the same heads are causally important for attribute recognition, spatial reasoning, counting, and visual math, suggesting that VRHs support visual reference resolution generally.

Limitation and Future Work. **(i)** Our analysis identifies VRHs through attention weights, following prior retrieval-head methods. While sufficient to reveal sparse, causal, and general visual-reference heads, this signal primarily captures *where* visual evidence is routed from, rather than *what* information is read out or transformed. Future work could combine VRH detection with activation patching (Goldowsky-Dill et al., 2023), or causal mediation (Li et al., 2026) for a finer-grained mechanistic account. **(ii)** Our experiments focus on single-image VLM inputs as a controlled setting for isolating visual reference behavior. Extending VRH analysis to multi-image, video, and interleaved image-text contexts could further reveal how VLMs route visual evidence across richer multimodal inputs, with potential applications in efficient visual-context selection (Wang et al., 2025a) and multimodal cache compression (Ge et al., 2024).

References

Baek, I., Chang, H., Ryu, S., and Lee, H. How do large vision-language models see text in image? unveiling the distinctive role of ocr heads. In *EMNLP*, 2025.

Bai, S., Cai, Y., Chen, R., et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024.

Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024.

Ge, S., Zhang, Y., Liu, L., Zhang, M., Han, J., and Gao, J. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *ICLR*, 2024. URL https://openreview.net/forum?id=uNrFpDPMyo.

Goldowsky-Dill, N., MacLeod, C., Sato, L., and Arora, A. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023. URL https://arxiv.org/abs/2304.05969.

Kang, S., Kim, J., Kim, J., and Hwang, S. J. Your large vision-language model only needs a few attention heads for visual grounding. In *CVPR*, 2025.

Li, Q., Ye, Z., Feng, X., Zhong, W., Ma, W., and Feng, X. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *AAAI*, 2026.

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Ma, Y. and Okazaki, N. From interpretability to performance: Optimizing retrieval heads for long-context language models. In *Findings of ACL*, 2026.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., and Murphy, K. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

Paiss, R., Ephrat, A., Tov, O., Zada, S., Mosseri, I., Irani, M., and Dekel, T. Teaching clip to count to ten. In *CVPR*, 2023.

Pei, R., Li, H., Guo, F., and Zhu, Q. Vera: Identifying and leveraging visual evidence retrieval heads in long-context understanding. *arXiv preprint arXiv:2602.10146*, 2026.

Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., and Shrivastava, A. Learning to predict visual attributes in the wild. In *CVPR*, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., Iyer, A., Pan, X., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024.

Ustalov, D., Pavlichenko, N., Koshelev, S., Likhobaba, D., and Smirnova, A. Toloka visual question answering benchmark. *arXiv preprint arXiv:2309.16511*, 2023. URL https://arxiv.org/abs/2309.16511.

Wang, F., Yu, S., Wu, J., Tang, J., Zhang, H., and Sun, Q. 3d question answering via only 2d vision-language models. In *ICML*, 2025a.

Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Wang, X., Ma, W., Zhang, T., de Melo, C. M., Chen, J., and Yuille, A. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large mutimodal models. In *CVPR*, 2025c.

Wu, W., Wang, Y., Xiao, G., Peng, H., and Fu, Y. Retrieval head mechanistically explains long-context factuality. In *ICLR*, 2025. URL https://openreview.net/forum?id=qVyeW-grC2k.

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *ECCV*, 2016.

Zhang, W., Yin, F., Yen, H., Chen, D., and Ye, X. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *EMNLP*, 2025.

Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. In *NeurIPS*, 2025.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., Su, W., Shao, J., et al. 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. Additional Qualitative Results

A.1. Visual Grounding

In Figure A1, we provide additional visualizations that complement Fig. 1 in the main paper. Specifically, we visualize the average attention map under three attention-head settings: (i) all heads, (ii) VRHs, and (iii) all heads excluding VRHs, on two grounding benchmarks: RefCOCO (Yu et al., 2016) and Toloka (Ustalov et al., 2023). Across all examples, the all-heads attention concentrates on the ground-truth referent(column ii), and this concentration is preserved when only VRHs are kept (column iii) but largely lost when VRHs are masked (column iv). VRHs are therefore the heads responsible for routing attention to the referent: under VRH masking, attention on the referent vanishes and the predicted bounding box becomes mislocalized.

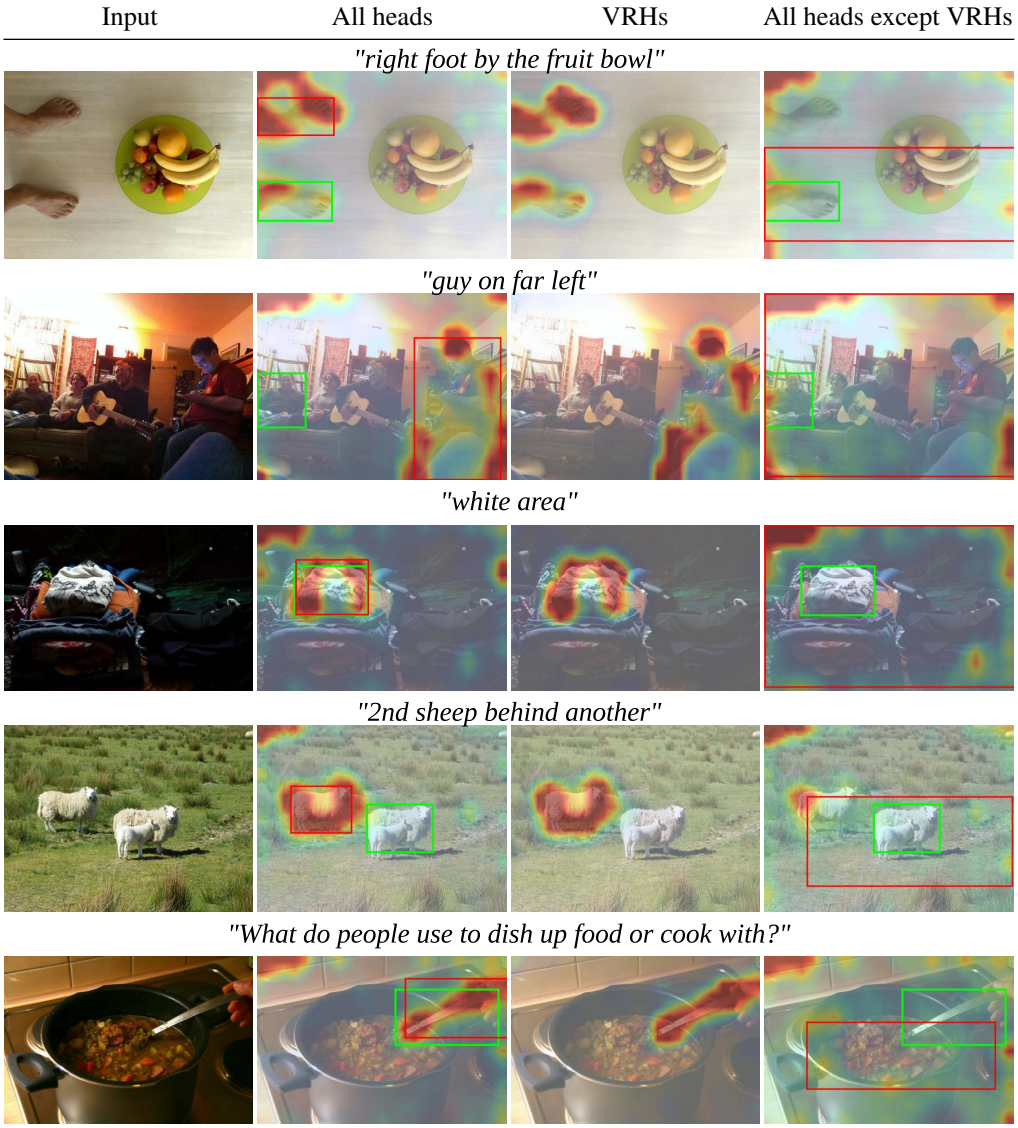
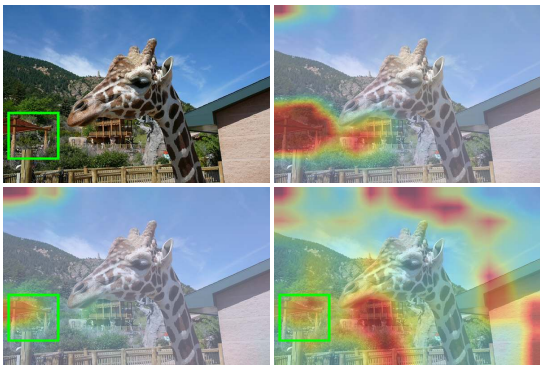


Figure A1. **Attention map visualizations on visual grounding tasks.** From left to right, the columns show the input image and the mean attention maps over: (i) all heads, (ii) the top-20 VRHs, and (iii) all heads excluding VRHs. Green and red boxes indicate ground-truth and predicted bounding boxes, respectively.

A.2. Visual Question Answering

Figure A2 shows analogous visualizations for visual question-answering (VQA) benchmarks: VAW (Pham et al., 2021) and Spatial457 (Wang et al., 2025c). As in the grounding case, attention on the visual evidence is preserved when only VRHs are kept but lost when they are masked: the model's answer remains fluent but becomes visually unfaithful, the "fluent-but-not-factual" failure mode discussed in Sec. 5.2.

"What color is the tarp?"



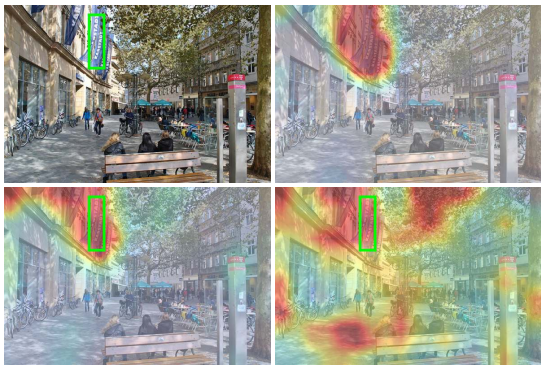
Reasoning: : The tarp appears to be a shade of brown or tan, which is not one of the provided color options.
Answer: : orange

(Correct)

Reasoning: The giraffe is standing in front of a building with a wooden structure that appears to be a part of a zoo or wildlife park. The tarp is not visible in the image.
Answer: brown

(Wrong)

"What color is the banner?"



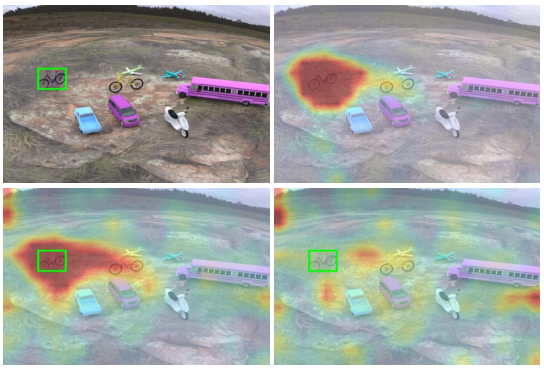
Reasoning: The banners in the image appears to be blue
Answer: blue

(Correct)

Reasoning: The image shows a street scene with people walking and sitting. There's a banner hanging from a building, but which the text is not clearly visible.
Answer: brown

(Wrong)

"What color is the utility bike?"



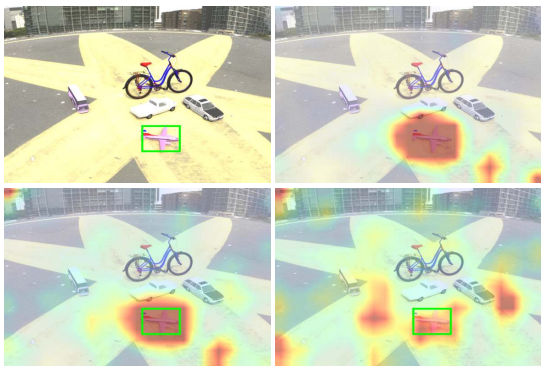
Reasoning: The utility bike is located on the left side of the image. It has a purple frame and black wheels. The color of the utility bike is purple.
Answer: Purple

(Correct)

Reasoning: The image shows various vehicles and objects, including but the utility bike is not clearly visible.
Answer: The utility bike is not clearly visible in the image.

(Wrong)

"Are any purple jets visible?"



Reasoning: The image contains a small airplane with a purple body and pink wings, which matches the description purple jets visible? query
Answer: Yes

(Correct)

Reasoning: The image shows various vehicles and objectss, including but a purple jet is not present.
Answer: No

(Wrong)

Figure A2. **Attention map visualizations on visual question-answering tasks.** Each panel shows the input image (upper left), and mean attention maps over: (i) all heads (lower left), (ii) the top-20 VRHs (upper right), and (iii) all heads excluding VRHs (lower right). Green boxes indicate the ground-truth target regions.

B. Application: Direct Preference Optimization Utilizing Visual Retrieval Heads

While the main paper focuses on *discovering* VRHs and characterizing their properties, the heads we identify are also actionable: they can be used as targets for downstream optimization. Specifically, we follow the recipe of Ma et al. (Ma & Okazaki, 2026), who propose optimizing retrieval heads via DPO (Rafailov et al., 2023) using positive/negative policies obtained from masked versus unmasked models.

B.1. Method

We adapt their recipe to the visual setting:

- **Positive policy** π^+: the original (unmasked) VLM.
- **Negative policy** π^-: the same VLM with the top-20 VRHs masked.
- **DPO data**: for each grounding example $s = (I_s, q_s) \in \mathcal{D}_{\text{train}}$, we sample $y_w \sim \pi^+(\cdot \mid I_s, q_s)$ as the chosen response and $y_l \sim \pi^-(\cdot \mid I_s, q_s)$ as the rejected response.
- **Training**: we fine-tune the unmasked VLM on these pairs with the standard DPO loss (Rafailov et al., 2023) with full parameter update, lr$= 2e-7$, for 4 epochs on 6,000 pairs from RefCOCO/+/g train split.

As a control, we train a separate model with random-head-masked negatives, sampling 20 random heads (excluding first-layer heads) for each example.

The intuition mirrors the LLM case: the negative policy concentrates exactly the failure mode we want to suppress (visually-unfaithful grounding driven by ablated VRHs), so DPO sharpens the model in the direction of stronger reliance on those same heads.

B.2. Results

As shown in Fig. A3, we observe modest but consistent gains, with an average improvement of 0.78 percentage-points across three visual grounding benchmarks (Yu et al., 2016; Mao et al., 2016). This suggests that VRHs are not merely an analytic abstraction, but a mechanistic substrate that can be directly optimized. We view this as an initial application; using VRHs for routing decisions, KV-cache compression analogous to the LLM setting (Wu et al., 2025), or test-time intervention are promising directions for future work.

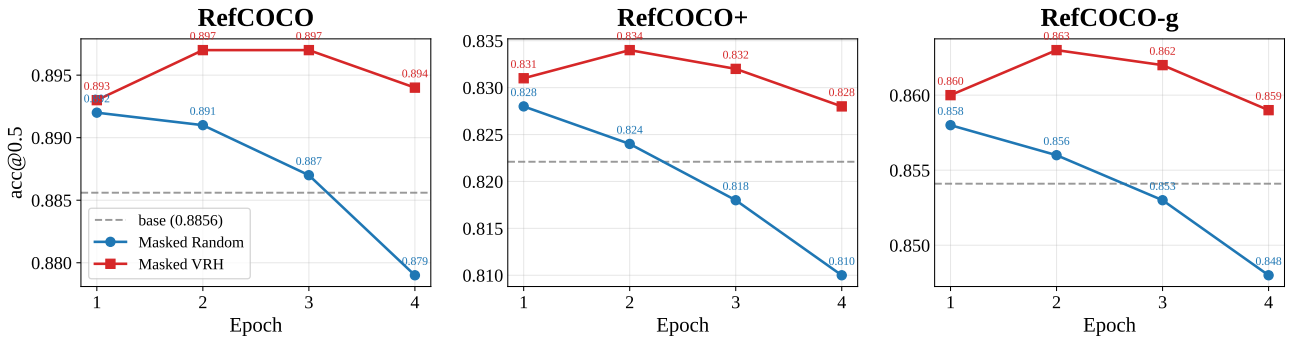


Figure A3. **Grounding accuracy comparisons with DPO under varying negative policies.** We compare two DPO variants that use either random-head dropping or VRH dropping to construct negative responses. Using VRH dropping as the negative policy yields stronger downstream improvements than random-head dropping, suggesting that perturbing causally relevant heads provides a more informative preference signal for optimization.

C. Full Quantitative Results

The main paper reports averaged results over RefCOCO/+/g (Yu et al., 2016; Mao et al., 2016). Here, we provide the full per-dataset breakdown and per-seed variability.

C.1. Per-dataset Visual Grounding Results under Head Masking

Table A1. **Per-dataset visual grounding accuracy.**

Model	Method	k=0	k=1	k=3	k=5	k=10	k=20
RefCOCO (Yu et al., 2016)							
Qwen2.5-VL-7B	Random	86.2	86.2	86.4	86.2	85.9	86.1
	VRH	86.2	85.1	83.5	79.3	9.7	7.5
Qwen3-VL-8B	Random	91.3	91.3	91.2	91.3	91.1	91.1
	VRH	91.3	91.0	90.4	84.1	33.4	20.4
InternVL3-8B	Random	93.6	93.6	93.7	93.3	92.9	92.5
	VRH	93.6	93.6	89.0	50.7	12.4	8.8
InternVL3.5-8B	Random	90.5	90.5	90.5	90.5	90.0	88.9
	VRH	90.5	90.0	88.4	83.6	64.5	32.1
RefCOCOg (Mao et al., 2016)							
Qwen2.5-VL-7B	Random	82.8	82.9	82.9	83.1	82.7	82.5
	VRH	82.8	82.5	79.9	76.1	11.8	9.2
Qwen3-VL-8B	Random	87.7	87.8	87.7	87.2	87.4	87.3
	VRH	87.7	87.5	86.6	81.8	30.6	23.5
InternVL3-8B	Random	90.0	89.9	90.0	89.8	89.4	88.9
	VRH	90.0	90.0	62.4	56.1	17.0	12.1
InternVL3.5-8B	Random	86.8	87.0	86.9	86.9	86.8	85.7
	VRH	86.8	86.1	84.7	79.5	56.3	32.1
RefCOCO+ (Yu et al., 2016)							
Qwen2.5-VL-7B	Random	78.5	78.4	78.3	78.7	78.0	78.3
	VRH	78.5	77.9	75.3	71.5	9.4	7.5
Qwen3-VL-8B	Random	84.2	84.2	84.2	84.0	83.7	83.8
	VRH	84.2	83.9	83.0	78.2	27.8	17.0
InternVL3-8B	Random	86.8	86.8	87.0	86.8	86.2	86.0
	VRH	86.8	84.4	81.5	45.4	11.2	8.0
InternVL3.5-8B	Random	83.4	83.3	83.4	83.4	83.4	82.2
	VRH	83.4	82.4	81.0	75.1	52.2	25.4
Toloka (Ustalov et al., 2023)							
Qwen2.5-VL-7B	Random	67.4	67.2	67.4	66.9	67.6	67.2
	VRH	67.4	65.0	53.0	7.0	2.2	1.1
Qwen3-VL-8B	Random	77.7	77.3	77.7	77.6	76.8	76.5
	VRH	77.7	77.4	76.2	48.0	15.6	5.0
InternVL3-8B	Random	81.2	81.2	82.0	81.8	80.7	80.2
	VRH	81.2	80.4	77.2	29.9	6.0	1.6
InternVL3.5-8B	Random	67.0	66.9	66.2	66.7	67.0	61.3
	VRH	67.0	66.3	64.0	59.9	27.0	10.3
RefSpatialBench (Zhou et al., 2025)							
Qwen2.5-VL-7B	Random	43.0	43.0	37.0	42.0	40.0	36.0
	VRH	43.0	35.0	27.0	12.0	5.0	0.0
Qwen3-VL-8B	Random	47.0	46.0	49.0	45.0	46.0	46.0
	VRH	47.0	51.0	51.0	33.0	8.0	6.0
InternVL3-8B	Random	39.0	37.0	39.0	40.0	42.0	40.0
	VRH	39.0	37.0	37.0	27.0	6.0	6.0
InternVL3.5-8B	Random	44.0	41.0	44.0	42.0	42.0	34.0
	VRH	44.0	40.0	24.0	20.0	11.0	5.0

Table A1 reports visual grounding accuracy separately on five referring-expression visual grounding benchmarks: RefCOCO, RefCOCO+, RefCOCOg, Toloka, and RefSpatialBench (Yu et al., 2016; Mao et al., 2016; Ustalov et al., 2023; Zhou et al., 2025), for all four VLMs and $k \in \{0, 1, 3, 5, 10, 20\}$ under VRH masking and random-head masking.

D. Detected Lists of Visual Retrieval Heads

For reproducibility, we list the top-20 VRH (layer, head) indices identified by our scoring method for each (VLM, grounding-benchmark) pair. The lists are ordered by descending head score $\bar{\Theta}_{l,h}$.

D.1. Top-20 VRHs On RefCOCO (Yu et al., 2016)

Table A2. Top-20 VRHs detected on RefCOCO (Yu et al., 2016). Each entry is (l, h).

Model	Top-20 VRH (layer, head)
Qwen2.5-VL-7B	(19,17), (16,0), (18,16), (16,20), (19,22), (19,20), (16,1), (19,23), (19,15), (17,18), (16,7), (14,11), (14,16), (16,6), (16,17), (19,26), (19,19), (14,23), (17,14), (18,15)
Qwen3-VL-8B	(20,15), (21,10), (18,19), (18,15), (21,11), (21,18), (17,27), (17,4), (19,3), (22,4), (21,8), (20,8), (18,5), (23,13), (20,17), (20,18), (23,30), (21,25), (20,16), (23,10)
InternVL3-8B	(16,20), (16,0), (19,22), (16,1), (19,23), (19,17), (19,20), (16,6), (14,11), (19,15), (16,4), (17,24), (18,16), (16,16), (16,17), (20,21), (16,7), (19,19), (14,16), (14,8)
InternVL3.5-8B	(21,11), (20,15), (21,10), (19,3), (18,15), (17,4), (21,18), (18,19), (17,27), (20,17), (24,13), (21,8), (20,18), (20,16), (21,25), (20,8), (18,5), (22,4), (18,21), (23,30)

D.2. Top-20 VRHs On RefCOCO+ (Yu et al., 2016)

Table A3. Top-20 VRHs detected on RefCOCO+ (Yu et al., 2016). Each entry is (l, h).

Model	Top-20 VRH (layer, head)
Qwen2.5-VL-7B	(19,17), (16,0), (16,20), (18,16), (19,22), (19,20), (16,1), (19,15), (19,23), (17,18), (14,11), (16,7), (14,16), (16,6), (16,17), (19,19), (19,26), (17,24), (14,23), (17,14)
Qwen3-VL-8B	(20,15), (21,10), (18,19), (18,15), (17,27), (21,11), (21,18), (17,4), (19,3), (20,8), (22,4), (21,8), (20,17), (20,18), (23,13), (21,25), (18,5), (23,30), (20,16), (23,10)
InternVL3-8B	(16,0), (16,20), (19,22), (19,23), (16,1), (19,17), (19,20), (16,6), (19,15), (14,11), (16,4), (17,24), (18,16), (20,21), (16,7), (16,16), (16,17), (14,8), (14,16), (19,19)
InternVL3.5-8B	(21,11), (21,10), (20,15), (19,3), (18,15), (21,18), (17,4), (18,19), (17,27), (20,17), (21,8), (20,18), (20,16), (24,13), (20,8), (22,4), (21,25), (17,26), (23,30), (23,10)

D.3. Top-20 VRHs On RefCOCOg (Mao et al., 2016)

Table A4. Top-20 VRHs detected on RefCOCOg (Mao et al., 2016). Each entry is (l, h).

Model	Top-20 VRH (layer, head)
Qwen2.5-VL-7B	(19,17), (16,0), (16,20), (19,22), (18,16), (16,1), (19,20), (19,23), (19,15), (14,11), (17,18), (16,7), (14,16), (16,6), (16,17), (19,26), (14,23), (19,19), (17,14), (18,15)
Qwen3-VL-8B	(20,15), (18,19), (21,10), (17,27), (18,15), (21,18), (21,11), (17,4), (19,3), (22,4), (21,8), (20,8), (18,5), (20,17), (20,18), (21,25), (23,30), (23,13), (20,16), (17,24)
InternVL3-8B	(16,20), (16,0), (16,1), (19,22), (19,17), (19,23), (14,11), (16,6), (19,20), (16,4), (19,15), (18,16), (16,17), (16,16), (17,24), (14,8), (14,16), (16,7), (20,21), (14,14)
InternVL3.5-8B	(21,11), (20,15), (21,10), (18,15), (19,3), (17,4), (21,18), (18,19), (17,27), (20,17), (20,18), (20,16), (21,8), (24,13), (19,13), (20,8), (18,21), (17,26), (22,4), (18,5)

D.4. Top-20 VRHs On Toloka (Ustalov et al., 2023)

Table A5. Top-20 VRHs detected on Toloka (Ustalov et al., 2023). Each entry is (l, h).

Model	Top-20 VRH (layer, head)
Qwen2.5-VL-7B	(19,17), (16,0), (19,22), (16,20), (19,15), (19,20), (19,23), (16,1), (18,16), (17,18), (19,19), (16,6), (19,26), (14,11), (19,25), (16,17), (16,16), (16,18), (16,4), (20,21)
Qwen3-VL-8B	(21,10), (20,15), (18,19), (21,11), (17,4), (18,15), (23,30), (21,8), (23,13), (22,4), (21,18), (20,17), (17,27), (23,10), (20,16), (20,8), (20,18), (19,3), (24,16), (21,31)
InternVL3-8B	(16,20), (19,17), (16,1), (19,20), (16,0), (19,15), (19,22), (16,17), (14,11), (19,23), (18,16), (16,16), (16,6), (20,21), (14,8), (19,19), (16,7), (17,24), (16,18), (21,25)
InternVL3.5-8B	(18,19), (21,11), (17,27), (21,10), (20,15), (18,15), (20,17), (17,4), (24,13), (23,30), (21,8), (19,3), (20,16), (21,18), (20,8), (23,10), (20,23), (18,21), (19,13), (21,9)

D.5. Top-20 VRHs On RefSpatialBench (Zhou et al., 2025)

Table A6. Top-20 VRHs detected on RefSpatialBench (Zhou et al., 2025). Each entry is (l, h).

Model	Top-20 VRH (layer, head)
Qwen2.5-VL-7B	(19,17), (19,22), (19,20), (19,15), (16,0), (16,20), (19,23), (16,1), (19,19), (18,16), (17,18), (19,26), (20,21), (16,16), (16,6), (16,4), (14,11), (19,25), (16,7), (20,23)
Qwen3-VL-8B	(21,10), (21,11), (20,15), (21,8), (23,30), (17,4), (18,19), (18,15), (21,18), (23,13), (23,10), (22,4), (20,17), (19,3), (20,16), (17,27), (20,8), (23,14), (25,10), (21,31)
InternVL3-8B	(19,17), (16,1), (16,20), (19,22), (19,15), (19,20), (19,23), (16,0), (14,11), (20,21), (17,24), (19,19), (16,6), (18,16), (16,17), (16,7), (16,16), (21,25), (21,5), (16,4)
InternVL3.5-8B	(21,11), (21,10), (21,8), (18,15), (19,3), (18,19), (21,18), (20,17), (17,4), (20,15), (17,27), (20,8), (24,13), (23,30), (20,23), (20,16), (21,25), (23,10), (20,21), (21,9)

E. Experimental Setup

We provide additional details on the VLMs, benchmarks, and head-masking procedure used throughout the paper.

E.1. Vision-Language Models

We evaluate our method on four open-source VLMs spanning two model families: Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), InternVL3-8B (Zhu et al., 2025), and InternVL3.5-8B (Wang et al., 2025b). These four models form two same-backbone pairs that differ in vision encoder, projector, and instruction-tuning recipe. This controlled pairing is what enables the cross-model VRH transfer analysis in Sec. 5.3 of the main paper.

Table A7 summarizes the architectural specifications relevant to our analysis: the number of layers L, attention heads per layer H, and total head count $L \cdot H$ that appears in the denominator of our sparsity ratios (e.g., "20 / 784 = 2.6%" for Qwen2.5-VL-7B).

Table A7. Architectural details of the four evaluated VLMs.

Model	# Layers L	# Heads/layer H	Total heads	Language backbone	Vision encoder	Projector
Qwen2.5-VL-7B	28	28	784	Qwen2.5-7B	Qwen ViT	MLP merger
Qwen3-VL-8B	36	32	1152	Qwen3-8B	Qwen ViT	MLP merger + DeepStack
InternVL3-8B	28	28	784	Qwen2.5-7B	InternViT-300M	2-layer MLP
InternVL3.5-8B	36	32	1152	Qwen3-8B	InternViT-300M	2-layer MLP

E.2. Benchmarks

Visual grounding benchmarks (head detection & main causality evaluation). We use five referring-expression visual grounding datasets, grouped into three families:

- **RefCOCO/+/g** (Yu et al., 2016; Mao et al., 2016): standard COCO-derived referring-expression benchmarks. We evaluate on the validation splits and use 8,811, 3,804, 7,573 examples per split.

- **RefSpatialBench** (Zhou et al., 2025): a referring-expression benchmark emphasizing spatial-relation expressions. We use 100 examples (full set).
- **Toloka VQA** (Ustalov et al., 2023): a crowd-sourced visual referring benchmark with diverse expression styles. We use 1705 examples (full set).

For all grounding datasets, a prediction is judged correct if $\mathrm{IoU}(\hat{B}_s, B_s) \geq \tau = 0.5$ (consistent with the main paper).

VQA benchmarks (cross-task generalization in Sec. 5.1). We further evaluate masked VLMs on seven VQA benchmarks spanning attribute prediction (VAW (Pham et al., 2021), Spatial457 (Wang et al., 2025c)), spatial reasoning (Spatial-RGPT (Cheng et al., 2024), CV-Bench (Tong et al., 2024)), vision-centric understanding (MMStar (Chen et al., 2024)), counting (CountBenchQA (Paiss et al., 2023)), and visual math (MathVista (Lu et al., 2024)).

- **VAW** (Pham et al., 2021): We manually convert VAW into a multiple-choice question (MCQ) format and evaluate on 6,309 samples.
- **Spatial457** (Wang et al., 2025c): We use the L1 (Single Object) portion and evaluate on 670 samples.
- **SpatialRGPT** (Cheng et al., 2024): We use the quantitative object-size estimation subset and evaluate on 266 samples.
- **CV-Bench** (Tong et al., 2024): We use the full benchmark and evaluate on 2,638 samples.
- **MMStar** (Chen et al., 2024): We use the minitest split and evaluate on 1,500 samples.
- **CountBenchQA** (Paiss et al., 2023): We use the full set of converted QA and evaluate on 491 samples.
- **MathVista** (Lu et al., 2024): We use the MCQ split and evaluate on 1,000 samples.

E.3. Head Masking Procedure

To evaluate the causal contribution of a head (l, h), we ablate it during the forward pass by zeroing the attention output of head (l, h) before it is added to the residual stream, leaving all other heads, MLPs, and layer norms untouched. This is the same masking convention used by Wu et al. (Wu et al., 2025) for text retrieval heads, ensuring our results are directly comparable.

Generation settings. All evaluations use *greedy decoding* and the model's default chat template. For grounding, the bounding box is parsed from the generated text using the model-specific output format (e.g. "bbox_2d": [x1, y1, x2, y2] for Qwen-VL families).

E.4. Compute

All experiments were conducted on a single NVIDIA RTX A6000 (48 GB VRAM).