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

Chanho Park[1] **Daehyeon Choi**[1] **Jihyun Lee**[2] **Minhyuk Sung**[1]
[1]KAIST [2]Independent Researcher
{charlieppark, daehyeonchoi, mhsung}@kaist.ac.kr
jihyunlee1027@gmail.com
Project Page: https://visual-retrieval-heads.github.io/

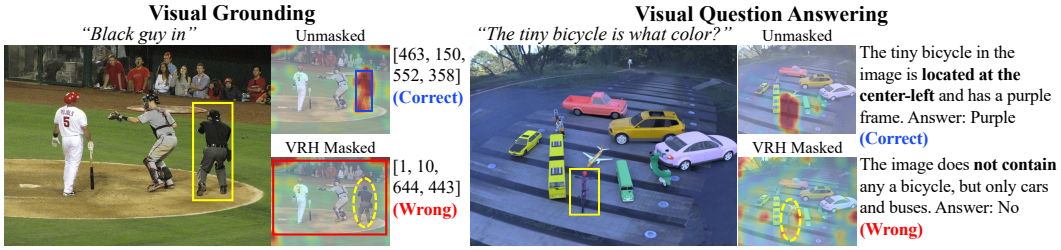


Figure 1: **Visual Retrieval Heads mediate visual reference resolution in VLMs.** We identify a sparse set of attention heads, *Visual Retrieval Heads* (VRHs), that are causally necessary for resolving the image region referred to by the text prompt. On visual grounding (**left**), masking the VRHs redirects 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.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

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* [1] 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 *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 [2] and VERA [3] 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 [4] 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 [5] 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 [6–9] and five grounding benchmarks (RefCOCO/+/g [5, 10], RefSpatialBench [11], Toloka [12]), and on seven further VLMs [27–33]. 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) [13–19], 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 *architec-*

turally 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) VRHs are *stably detectable*: results are consistent across output-token choices and scoring samples, with as few as 5 to 10 grounding examples yielding nearly the same masking effect 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 based on inference-time attention patterns [1, 20, 2, 3]. A pioneering work along this direction is Retrieval Heads [1], 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 [20] 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 [2] 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 [3] 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 [4], 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 [1, 20, 2, 3].

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.

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

	Retrieval Heads [1]	QRHead [20]	OCR Heads [2]	VERA [3]	Loc. Heads [4]	**VRH (Ours)**
Input Source	Text	Text	Visual document	Visual document	General image	General image
Causality	✓	✓	✓	✓	✗	✓

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 [1, 20], optical character recognition [2, 3]). 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 tokens* ($\mathcal{Q}_s$) 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 [1, 20] assign high importance to heads that place strong attention from either *output tokens* [1] (which express the retrieval prediction) or *input prompt tokens* [20] (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 [1, 20, 2, 3], 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:

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

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 [1, 20, 2, 3] 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 [1, 20, 3, 4, 2].

- **Query token set** $\mathcal{Q}_s$: (i) *Input query tokens* $\mathcal{Q}_s^{\mathrm{in}}$, which specify the retrieval target, or (ii) *output tokens* $\mathcal{Q}_s^{\mathrm{out}}$, 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.

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., $\mathcal{Q}_s^{\texttt{in}}$ or $\mathcal{Q}_s^{\texttt{out}}$; 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 [1]	**QRHead [20]**	**OCR Heads [2]**	**VERA [3]**	**Loc. Heads [4]**
Query Token Set $\mathcal{Q}_s$	$\mathcal{Q}_s^{\texttt{out}}$	$\mathcal{Q}_s^{\texttt{in}}$	$\mathcal{Q}_s^{\texttt{out}}$	$\mathcal{Q}_s^{\texttt{out}}$	$\mathcal{Q}_s^{\texttt{in}}$
Key Aggregation Φ	$\Phi^{\texttt{argmax}}$	$\Phi^{\texttt{sum}}$	$\Phi^{\texttt{argmax}}$	$\Phi^{\texttt{sum}}$	N/A†
Sample Aggregation Ω	$\Omega^{\texttt{avg}}$	$\Omega^{\texttt{avg}}$	$\Omega^{\texttt{avg}}$	$\Omega^{\texttt{avg}}$	$\Omega^{\texttt{freq}}$

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 [20, 1–3], *causality* is essential for interpreting a head as a meaningful functional unit rather than an attention-correlated artifact. We therefore perform *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 $(\mathcal{Q}_s^{\text{out}}, \Phi^{\text{sum}}, \Omega^{\text{avg}})$ as the default design for VRH detection, averaging the per-sample scores over the scoring samples that the VLM grounds correctly. Fig. 2 reports this validation across four VLMs [6–9] on RefCOCO [5], 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.

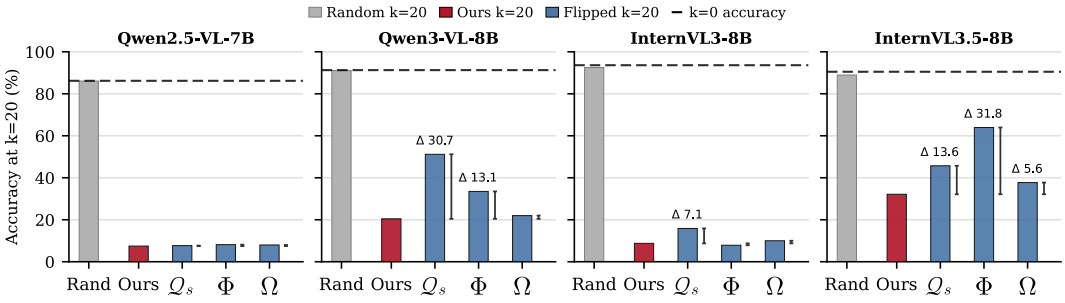


Figure 2: **Causal validation of head-scoring variants on RefCOCO [5] across four VLMs [6–9].** We mask the top-20 heads detected by each variant and report visual grounding accuracy; lower accuracy indicates that the detected heads are more causally important. *Ours* denotes our default variant $(\mathcal{Q}_s^{\text{out}}, \Phi^{\text{sum}}, \Omega^{\text{avg}})$. 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*.

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 [1, 20] or visual retrieval from document-like images [2, 3]. Localization Heads [4] 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 [5], 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.

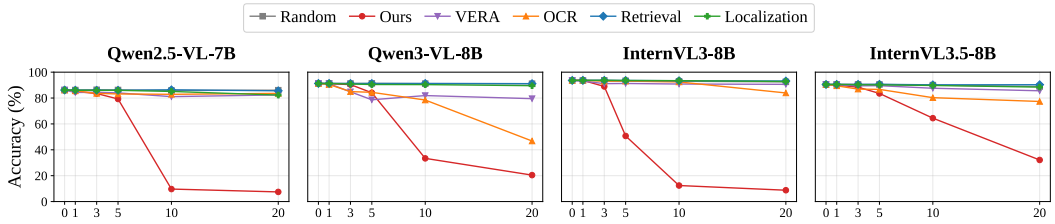


Figure 3: **Comparison with existing heads.** We compare VRHs with Retrieval Heads [1], OCR Heads [2], VERA [3], Localization Heads [4], and randomly selected heads on the RefCOCO visual grounding benchmark [5]. The x-axis denotes the number of masked heads (k). Masking the top-k VRHs causes substantially larger drops in grounding accuracy than masking these baselines, indicating that VRHs capture a distinct set of heads causally important for visual grounding.

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 [1, 20], 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 [6–9] on five grounding datasets spanning three benchmark families: (RefCOCO/+/g [5, 10], RefSpatialBench [11], and Toloka [12]). 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. Fig. 6 extends the same contrast at k=20 on RefCOCO to seven further VLMs built on Llama-3, Gemma 2, DeepSeekMoE, InternLM2.5, GLM, and Phi [27–33], and masking the VRHs collapses grounding in all of them. These seven are evaluated on a separate detection and evaluation split of the benchmark (Sec. K). This indicates that, despite differences in architecture, training recipe, and benchmark distribution, diverse VLMs rely on a sparse set of attention heads for correct visual reference behavior.

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 [1].

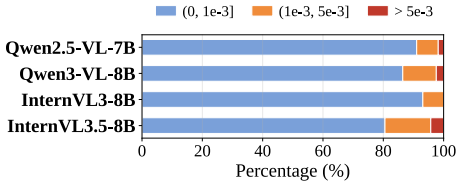


Figure 5: **Sparsity of VRHs.** Distribution of head scores for VRH detection on RefCOCO [5] across four VLMs [6–9]. Only a small fraction of heads receive non-negligible scores.

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 [1–3, 20]. We now show that VRHs exhibit important additional properties: (i) VRHs *generalize across visual reference tasks* beyond the

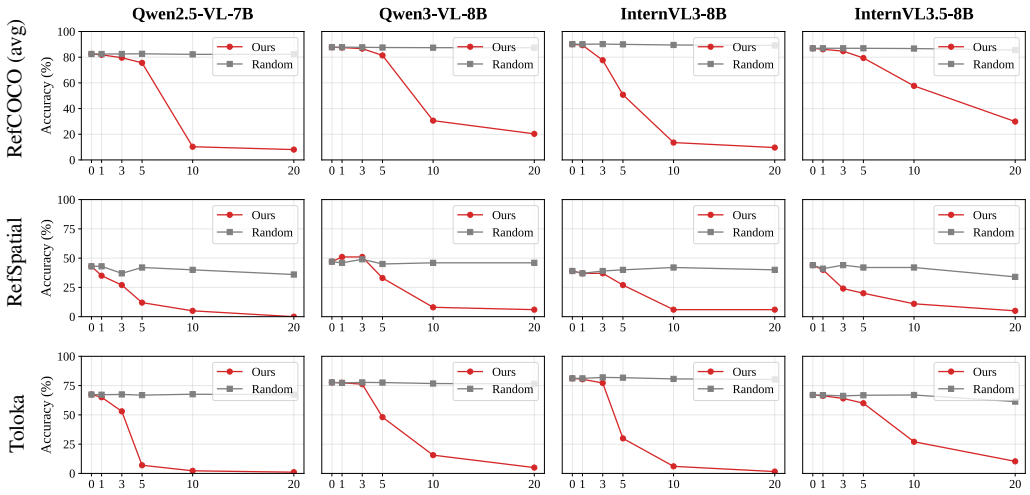


Figure 4: **Universality of VRHs across models and benchmarks.** We report visual grounding accuracy after masking the top-k VRHs across four VLMs and three benchmark families: RefCOCO/+/g [5, 10], RefSpatialBench [11], and Toloka [12]. The x-axis denotes the number of masked heads (k). Masking only 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.

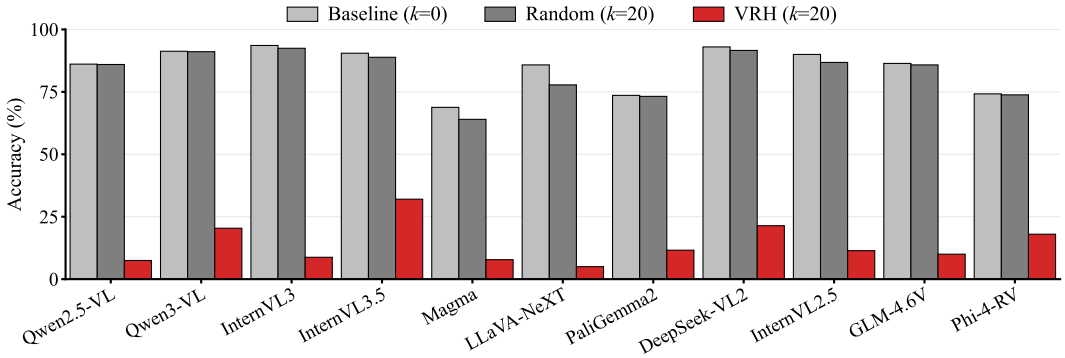


Figure 6: **VRHs are universal across VLMs.** Visual grounding accuracy on RefCOCO [5] for all eleven evaluated VLMs, before masking, after masking 20 random heads, and after masking the top-20 VRHs. Masking the VRHs causes a large accuracy drop in every model, whereas masking the same number of random heads has little effect. Exact values, evaluation splits, and per-model protocols are given in Sec. K.

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 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 [1, 20, 2, 3], 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 [5], and testing their causal effects on seven VQA benchmarks spanning the aforementioned tasks: VAW [13], Spatial457 [14], SpatialRGPT [15], CV-Bench [16], MMStar [17], CountBenchQA [18], and MathVista [19]. 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.

Table 3: **Generalization across visual reference tasks.** Masking the top-20 VRHs detected on RefCOCO [5] consistently degrades performance on seven VQA benchmarks [13–19] 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)

5.2 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. 8a, we further decompose visual-grounding failures under VRH masking versus a control that masks randomly selected heads together with selected first-layer (L0) heads into two types: (i) *parse failure*, where the output does not preserve the bounding-box syntax and cannot be parsed, and (ii) *mislocation*, where the output is a syntactically valid bounding box but is spatially mislocalized.

We find that masking VRHs significantly increases mislocation rates while parse failures remain rare, whereas the control mainly causes parse failures. This contrast mirrors the "fluent-but-not-factual" failure mode reported for text retrieval heads [1, 20]: 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 the top row of Fig. 7, 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 the bottom row of Fig. 7, 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. The same holds for the third pair in Fig. 7, Magma-8B [27] and LLaVA-NeXT-Llama3-8B [28], where cross-masking the top-20 VRHs reduces

grounding accuracy from 68.8% to 0.0% and from 85.8% to 5.8%, respectively. 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.

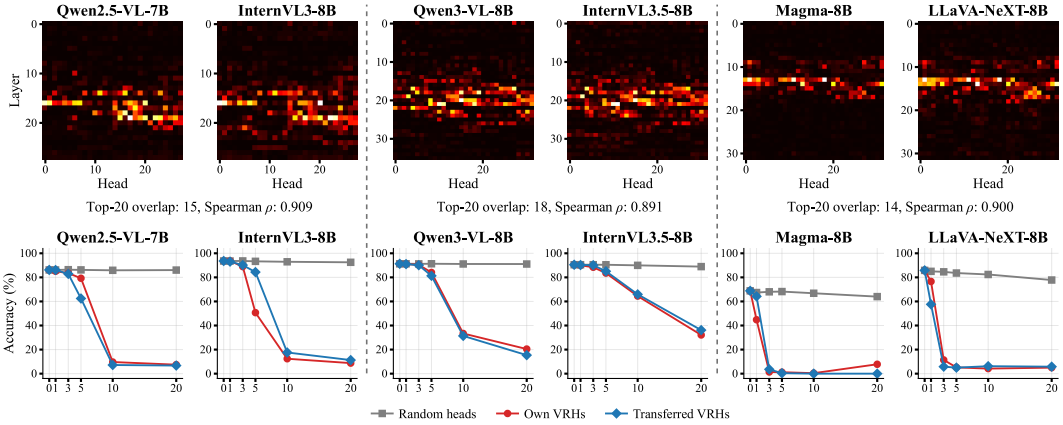
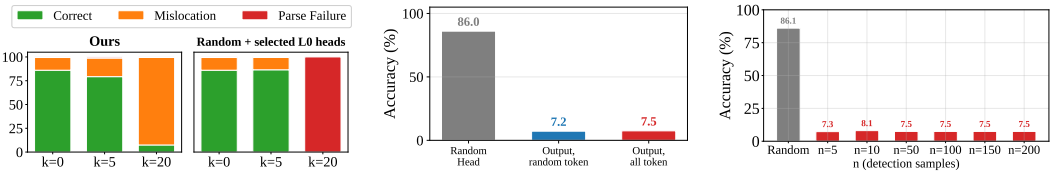


Figure 7: **VRH detection consistency across same-backbone VLMs.** Each block is a pair of VLMs that share a language backbone but differ in vision encoder, projector, and instruction tuning: Qwen2.5-VL-7B [6] with InternVL3-8B [8], which share a Qwen2.5-7B backbone; Qwen3-VL-8B [7] with InternVL3.5-8B [9], which share a Qwen3-8B backbone; and Magma-8B [27] with LLaVA-NeXT-Llama3-8B [28], which share a Llama-3 backbone. **Top row:** Layer-head score maps, 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. *Own* VRHs are the heads detected in the panel's own model and *transferred* VRHs are those detected in its partner model. The x-axis denotes the number of masked heads (k).

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. 8b, both choices cause a comparable accuracy drop (7.2% and 7.5%, respectively), while masking the same number



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

Figure 8: **VRHs are functionally specific and stably detectable.** All experiments use Qwen2.5-VL-7B [6] on RefCOCO [5] with k=20 masked heads unless otherwise noted. **(a)** Visual grounding failure decomposition under VRH masking vs. a control that masks randomly selected heads together with selected first-layer (L0) heads, which disrupts the output format. *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.

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.

Robustness across sample size. VRH detection is highly sample-efficient: only a few grounding annotations are needed to detect heads whose masking is as damaging as that of heads detected from 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 measure grounding accuracy after masking each resulting top-20 head set. As shown in Fig. 8c, heads detected from as few as 5–10 examples reduce grounding accuracy nearly as much as those detected from the full scoring set of 200 examples. This makes VRH detection practical: a small number of bounding-box annotations is sufficient to identify causally important heads.

6 Conclusion

We introduce Visual Retrieval Heads (VRHs), a sparse set of attention heads that are causally necessary for correct visual reference behavior. 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 [21], or causal mediation [22] 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 [23] and multimodal cache compression [24].

Acknowledgments and Disclosure of Funding

This work was supported by the NRF grant (RS-2026-25486000); IITP grants (RS-2022-00156435, RS-2024-00399817, RS-2025-25441313, RS-2025-25443318, RS-2026-25526850, RS-2026-25617720) funded by MSIT, Korea; the Industrial Technology Innovation Program grant (RS-2025-02317326) funded by MOTIE, Korea; the InnoCORE program of MSIT (AI Meta-Scientist, N10260110); and the DRB-KAIST SketchTheFuture Research Center.

References

[1] Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *ICLR*, 2025.

[2] Ingeol Baek, Hwan Chang, Sunghyun Ryu, and Hwanhee Lee. How do large vision-language models see text in image? Unveiling the distinctive role of OCR heads. In *EMNLP*, 2025.

[3] Rongcan Pei, Huan Li, Fang Guo, and Qi Zhu. VERA: Identifying and leveraging visual evidence retrieval heads in long-context understanding. In *IJCAI*, 2026.

[4] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *CVPR*, 2025.

[5] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[7] Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[8] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[9] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[10] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

[11] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. RoboRefer: Towards spatial referring with reasoning in vision-language models for robotics. In *NeurIPS*, 2025.

[12] Dmitry Ustalov, Nikita Pavlichenko, Sergey Koshelev, Daniil Likhobaba, and Alisa Smirnova. Toloka visual question answering benchmark. *arXiv preprint arXiv:2309.16511*, 2023.

[13] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *CVPR*, 2021.

[14] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6D spatial reasoning of large multimodal models. In *CVPR*, 2025.

[15] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024.

[16] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *NeurIPS*, 2024.

[17] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024.

[18] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten. In *ICCV*, 2023.

[19] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

[20] Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *EMNLP*, 2025.

[21] Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

[22] Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *AAAI*, 2026.

[23] Fengyun Wang, Sicheng Yu, Jiawei Wu, Jinhui Tang, Hanwang Zhang, and Qianru Sun. 3D question answering via only 2D vision-language models. In *ICML*, 2025.

[24] Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *ICLR*, 2024.

[25] Youmi Ma and Naoaki Okazaki. From interpretability to performance: Optimizing retrieval heads for long-context language models. In *Findings of ACL*, 2026.

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

[27] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, Yuquan Deng, Lars Liden, and Jianfeng Gao. Magma: A foundation model for multimodal AI agents. In *CVPR*, 2025.

[28] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: Stronger LLMs supercharge multimodal capabilities in the wild, 2024. https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

[29] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. PaliGemma 2: A family of versatile VLMs for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

[30] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[31] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[32] Z.ai. GLM-4.6V: Open source multimodal models with native tool use. Z.ai Technical Blog, 2025. https://z.ai/blog/glm-4.6v.

[33] Jyoti Aneja et al. Phi-4-reasoning-vision-15B technical report. *arXiv preprint arXiv:2603.03975*, 2026.

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 [5] and Toloka [12]. 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.

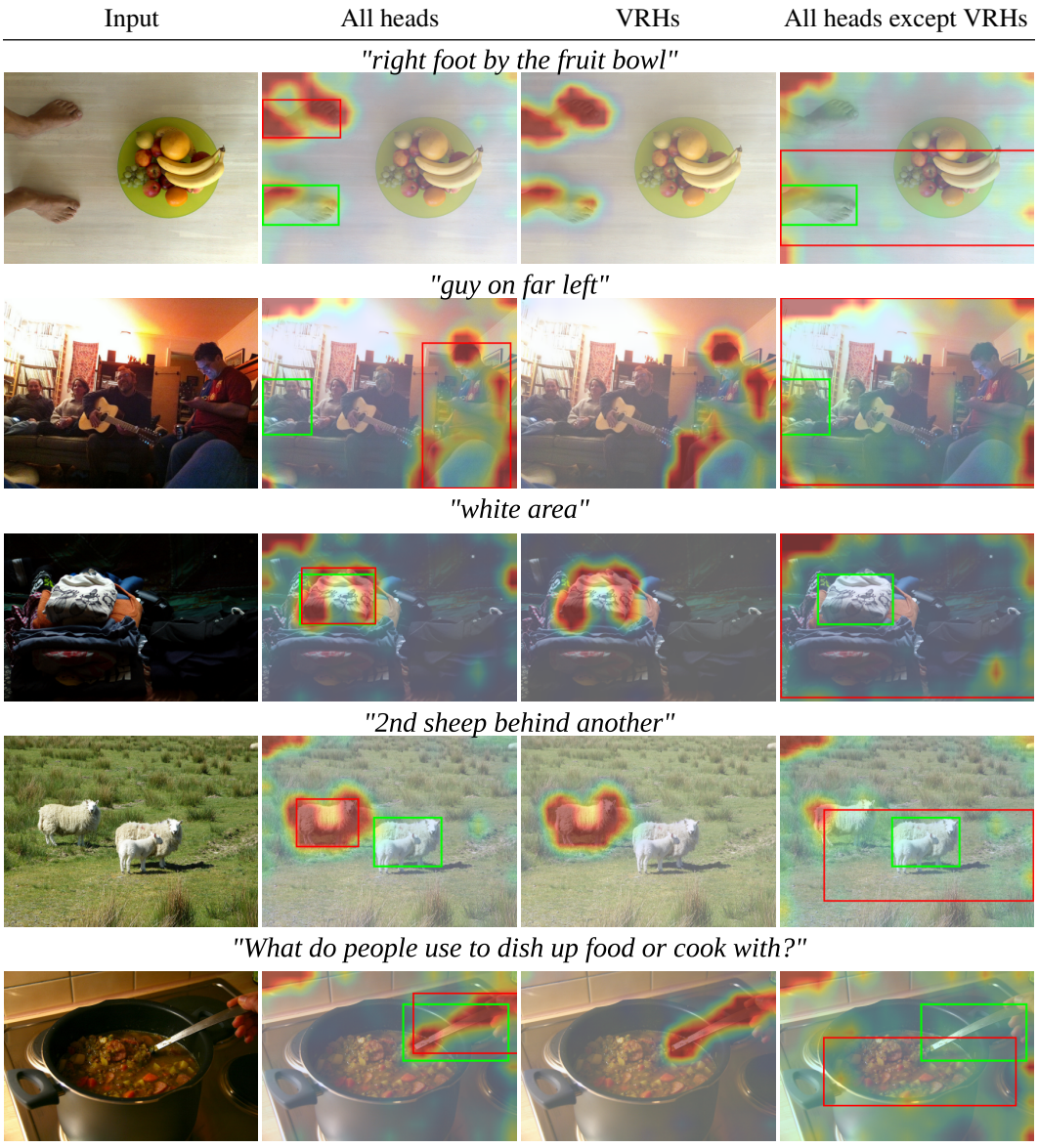
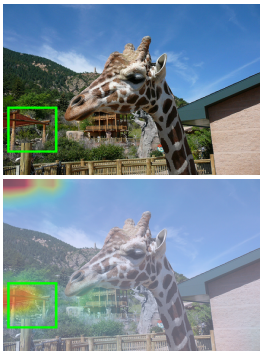


Figure A1: **Attention map visualizations on visual grounding tasks.** From left to right, the columns show the input image and the mean attention maps over: (i) all heads, (ii) the top-20 VRHs, and (iii) all heads excluding VRHs. Green and red boxes indicate ground-truth and predicted bounding boxes,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

A.2 Visual Question Answering

Figure A2 shows analogous visualizations for visual question-answering (VQA) benchmarks: VAW [13] and Spatial457 [14]. 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

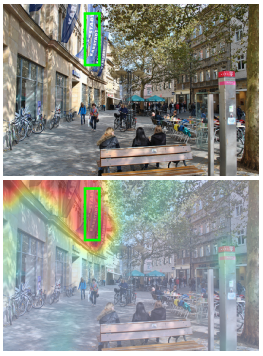
(Correct)



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

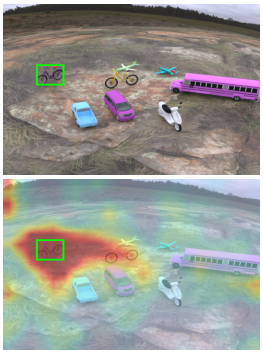
(Correct)



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

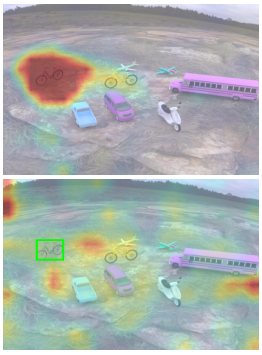
(Wrong)

"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. [25], who propose optimizing retrieval heads via DPO [26] 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 [26] with full parameter update, learning rate 2×10^{-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 [5, 10]. 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 [1], or test-time intervention are promising directions for future work.

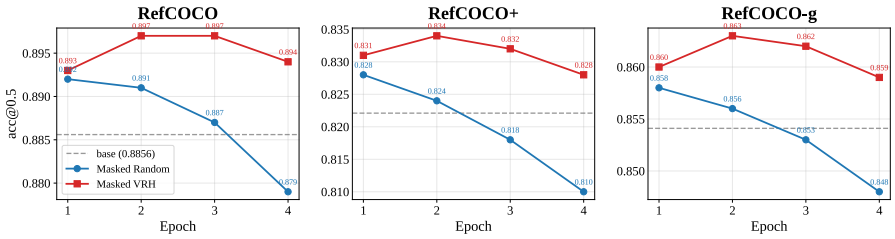


Figure A3: **Grounding accuracy comparisons with DPO under varying negative policies.** We compare two DPO variants that use either random-head dropping or VRH dropping to construct negative responses. Using VRH dropping as the negative policy yields stronger downstream improvements than random-head 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 [5, 10]. Here, we provide the full per-dataset breakdown.

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

Table A1 reports visual grounding accuracy separately on five referring-expression visual grounding benchmarks: RefCOCO, RefCOCO+, RefCOCOg, Toloka, and RefSpatialBench [5, 10, 12, 11], for all four VLMs and $k \in \{0, 1, 3, 5, 10, 20\}$ under VRH masking and random-head masking.

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

Model	Method	k=0	k=1	k=3	k=5	k=10	k=20
RefCOCO [5]							
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 [10]							
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+ [5]							
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 [12]							
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 [11]							
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

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 [5]

Table A2: Top-20 VRHs detected on RefCOCO [5]. 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+ [5]

Table A3: Top-20 VRHs detected on RefCOCO+ [5]. 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 [10]

Table A4: Top-20 VRHs detected on RefCOCOg [10]. 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 [12]

Table A5: Top-20 VRHs detected on Toloka [12]. 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 [11]

Table A6: Top-20 VRHs detected on RefSpatialBench [11]. 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 [6], Qwen3-VL-8B [7], InternVL3-8B [8], and InternVL3.5-8B [9]. 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** [5, 10]: standard COCO-derived referring-expression benchmarks. We evaluate on the validation splits and use 8,811, 3,804, 7,573 examples per split.

- **RefSpatialBench** [11]: a referring-expression benchmark emphasizing spatial-relation expressions. We use 100 examples (full set).
- **Toloka VQA** [12]: 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 [13], Spatial457 [14]), spatial reasoning (SpatialRGPT [15], CV-Bench [16]), vision-centric understanding (MMStar [17]), counting (CountBenchQA [18]), and visual math (MathVista [19]).

- **VAW** [13]: We manually convert VAW into a multiple-choice question (MCQ) format and evaluate on 6,309 samples.
- **Spatial457** [14]: We use the L1 (Single Object) portion and evaluate on 670 samples.
- **SpatialRGPT** [15]: We use the quantitative object-size estimation subset and evaluate on 266 samples.
- **CV-Bench** [16]: We use the full benchmark and evaluate on 2,638 samples.
- **MMStar** [17]: We use the minitest split and evaluate on 1,500 samples.
- **CountBenchQA** [18]: We use the full set of converted QA and evaluate on 491 samples.
- **MathVista** [19]: 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. [1] 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).

F Ablation on the Combined Query Token Set

Sec. 3.3 of the main paper ablates $\mathcal{Q}_s^{\text{in}}$ and $\mathcal{Q}_s^{\text{out}}$ separately. Here we additionally experiment with a VRH detection setting that uses their union $\mathcal{Q}_s^{\text{in}} \cup \mathcal{Q}_s^{\text{out}}$, and compare the causal effect of the resulting heads with that of our default VRHs in Tab. A8. Because this setting uses a superset of the query tokens of our default $\mathcal{Q}_s^{\text{out}}$ setting, it has a comparable causal effect for some models. However, our default setting still performs better on three models, Qwen3-VL, InternVL3, and InternVL3.5, with a particularly noticeable gap on Qwen3-VL. These results suggest that our default setting is more robust across models.

Table A8: **Visual grounding accuracy on RefCOCO [5] after masking the top-20 VRHs detected using $\mathcal{Q}_s^{\text{in}} \cup \mathcal{Q}_s^{\text{out}}$ versus $\mathcal{Q}_s^{\text{out}}$ (ours) across four VLMs.**

	Qwen2.5-VL-7B	Qwen3-VL-8B	InternVL3-8B	InternVL3.5-8B
Baseline (k=0)	86.11	91.27	93.59	90.48
VRH, $\mathcal{Q}_s^{\text{in}} \cup \mathcal{Q}_s^{\text{out}}$ (k=20)	7.14	29.85	8.91	32.94
VRH, $\mathcal{Q}_s^{\text{out}}$ (ours, k=20)	7.48	20.39	8.76	32.03

G Comparison with VERA and Localization Heads

G.1 VERA

Tab. 2 of the main paper summarizes each method by its main design axes, and at this level of abstraction VERA [3] appears close to our default scoring rule. Their implementations, however, differ in three respects that substantially change the selected heads and their causal effect:

- **Cross-sample aggregation.** We average the per-sample scores over the samples whose grounding prediction is correct, whereas VERA averages them over all samples of the scoring set.
- **Query tokens.** We use all output tokens $\mathcal{Q}_s^{\text{out}}$, while VERA uses the attention at a single predefined generation step.
- **Key aggregation.** We sum the attention over the referent tokens $\mathcal{K}_s$, and VERA divides this sum by the head's total attention over all visual tokens. Our score therefore favors heads that put a large amount of attention on the referent, and VERA's favors heads for which the referent takes a large fraction of their visual attention.

Comparison on the same data. Whereas the VERA baseline in Fig. 3 re-implements VERA's scoring rule within our detection framework, here we apply VERA's official scoring implementation to our RefCOCO [5] probing setup by converting the ground-truth box of each sample into VERA's evidence region, and denote the result VERA*. VERA* and our method score the heads on the same 200 detection samples; VERA* averages over all of them, and our method over those answered correctly. The two top-20 head sets are then masked and evaluated with the same code on the same 8,611 evaluation samples, which are disjoint from the detection samples, so the two settings differ only in the masked heads. Because this comparison was run as a separate experiment, its accuracies differ slightly from those in Tab. A12. The VRH sets detected in it share 16–20 of their top-20 heads with Tab. A2. As shown in Tab. A9, masking the VERA* heads reduces grounding accuracy by 18.1 percentage points on average over the four VLMs, compared with 71.5 percentage points for our VRHs, and the two top-20 sets share only 3–9 heads.

Table A9: **Visual grounding accuracy on RefCOCO [5] after masking the top-20 heads detected by VERA's scoring (VERA*) or by our scoring (VRHs) on the same detection samples.** The last row gives the number of heads shared by the two top-20 sets.

	Qwen2.5-VL-7B	Qwen3-VL-8B	InternVL3-8B	InternVL3.5-8B
Baseline (k=0)	85.67	91.23	93.65	90.64
VERA* (k=20)	73.27	56.01	70.70	88.75
VRH (k=20)	7.71	20.13	9.26	37.91
Shared heads (of 20)	9	8	8	3

G.2 Localization Heads

We measure the overlap between the top-20 VRHs (Tab. A2) and the top-20 Localization Heads [4] compared in Fig. 3 of the main paper. As shown in Tab. A10, the Jaccard similarity is 0.081 for the two Qwen models and 0 for the two InternVL models, i.e., the two sets share 3, 3, 0, and 0 heads. The overlap is small for every VLM, suggesting that VRHs and Localization Heads identify largely different subsets of heads rather than different orderings of a shared set of candidates.

Table A10: **Overlap between the top-20 VRHs and the top-20 Localization Heads [4].**

	Qwen2.5-VL-7B	Qwen3-VL-8B	InternVL3-8B	InternVL3.5-8B
Jaccard similarity	0.081	0.081	0	0
Shared heads (of 20)	3	3	0	0

H Sensitivity to the Number of Masked Heads

Fig. 2 of the main paper reports the scoring-variant ablation at k=20. Tab. A11 reports the same ablation for $k \in \{5, 10, 20, 50\}$ across the four VLMs, comparing our design choice $(\mathcal{Q}_s^{\text{out}}, \Phi^{\text{sum}}, \Omega^{\text{avg}})$ against variants obtained by flipping one design component at a time and against random-head masking. At k=5 the masking effect is relatively weak or inconsistent across models, which makes the relative ranking among scoring variants less conclusive. Once k increases beyond 5, our design choice consistently ranks either first or second in terms of causal effect across $k \in \{10, 20, 50\}$. Notably, masking only the top-10 VRHs already causes substantial degradation across all scoring variants, showing that a small subset of VRHs is sufficient to strongly impair grounding. We use k=20 in the main experiments, following the convention established in the retrieval-head literature [1, 2]; Tab. A11 shows that our main conclusion remains largely robust for $k \geq 10$.

Table A11: **Scoring ablation for varying numbers of masked heads $k \in \{5, 10, 20, 50\}$ across four VLMs.** We report RefCOCO [5] visual grounding accuracy after masking random heads, our detected VRHs, or VRHs obtained by flipping each scoring design choice. **Bold** and underline mark the largest and the second-largest causal effect among the four scoring variants in each row.

	Random heads	VRHs (Ours)	Flipped $\mathcal{Q}_s$	Flipped Φ	Flipped Ω
(a) Qwen2.5-VL-7B					
k=0 (baseline)	86.11	86.11	86.11	86.11	86.11
k=5	86.16	79.29	**56.08**	63.28	63.14
k=10	85.79	9.63	10.52	**8.87**	9.63
k=20	85.97	**7.48**	7.68	8.13	8.01
k=50	84.67	**6.63**	9.85	7.21	8.70
(b) Qwen3-VL-8B					
k=0	91.27	91.27	91.27	91.27	91.27
k=5	91.26	84.14	79.90	**57.03**	84.24
k=10	91.12	33.42	80.47	33.60	**32.63**
k=20	91.09	**20.39**	51.10	33.61	21.96
k=50	89.97	9.29	**7.20**	14.17	13.11
(c) InternVL3-8B					
k=0	93.59	93.59	93.59	93.59	93.59
k=5	93.28	**50.62**	72.06	50.62	56.16
k=10	92.90	**12.39**	39.05	14.20	14.20
k=20	92.46	8.76	15.83	**7.89**	9.93
k=50	90.98	8.31	13.90	**7.35**	8.87
(d) InternVL3.5-8B					
k=0	90.48	90.48	90.48	90.48	90.48
k=5	90.45	83.52	86.94	**82.22**	83.52
k=10	89.95	**64.48**	71.50	69.83	64.48
k=20	88.86	**32.03**	45.66	63.96	37.56
k=50	88.09	16.98	**10.42**	23.49	19.73

I Head Detection and Evaluation Splits

Split protocol. On RefCOCO [5], heads are detected on 200 samples drawn from the validation split. In our original RefCOCO setup, used for the RefCOCO results of the main paper and of Tab. A1, these 200 samples are included in the full evaluation set of 8,811 samples. Tab. A12 re-evaluates all four models after excluding them, i.e., on the remaining 8,611 samples, which are disjoint from the detection set by construction. The accuracies differ by at most 0.11 percentage points, so including them has little effect on the results. The additional RefCOCO analyses in Secs. F, G.1, and H use the disjoint 8,611-sample evaluation set.

Tab. 3 of the main paper further evaluates the heads detected on these 200 RefCOCO samples on separate, non-grounding benchmarks.

Table A12: **Visual grounding accuracy on RefCOCO [5] after masking the top-20 VRHs, evaluated with and without the 200 head-detection samples.**

Evaluation set	Qwen2.5-VL-7B	Qwen3-VL-8B	InternVL3-8B	InternVL3.5-8B
With detection samples (n=8,811)	7.46	20.44	8.76	32.14
Without detection samples (n=8,611)	7.48	20.39	8.76	32.03

J Detection Cost and Fused Attention

With our default setting of 200 detection samples, VRH detection takes 5.3–7.6 minutes with a peak memory usage of 15.93–17.83 GiB on a single A6000 GPU across the four VLMs. Fig. 8c of the main paper shows that heads detected from 5–10 samples already have nearly the same masking effect. VRH detection requires explicitly extracting attention weights over visual tokens, so its cost grows with the number of visual tokens. Tab. A13 reports the peak memory usage and computation time of VRH detection while varying the number of visual tokens by changing the input resolution or the number of input images. To examine how memory and computational costs interact with the use of fused-attention kernels, we report results using both eager and fused attention. As the number of visual tokens increases from 256 to 512 and 3,328, both memory usage and computation time increase, with the growth being substantially more pronounced under eager attention. In particular, at 3,328 visual tokens, eager attention requires 34.59 GiB and 4.86 s/sample, compared with 16.01 GiB and 2.31 s/sample using fused attention. These results show that fused-attention kernels substantially mitigate the memory and computational overhead associated with scaling to larger numbers of visual tokens. This is because eager attention materializes dense attention maps whose memory grows approximately quadratically with sequence length, whereas fused attention avoids retaining the full maps by recomputing only the query rows needed for VRH scoring.

Table A13: **Peak memory usage and computation time of VRH detection with varying numbers of visual tokens.** We use InternVL3-8B [8] with dynamic tiling as the base model and conduct all measurements on a single 48 GB A6000 GPU.

Setting	# Visual tokens	Peak memory (GiB) (Eager / Fused Att.)	Time (s/sample) (Eager / Fused Att.)
Low resolution (448×448)	256	15.09 / 14.92	1.77 ± 0.47 / 1.41 ± 0.21
Low resolution (448×448), 2 images	512	15.55 / 15.01	2.33 ± 0.99 / 1.56 ± 0.34
High resolution (1792×1344)	3,328	34.59 / 16.01	4.86 ± 0.77 / 2.31 ± 0.29

K Evaluated VLMs: Protocols and Per-Model Results

This section provides the protocol and the exact values behind Fig. 6 and Fig. 7 of the main paper.

Protocol. The four VLMs studied throughout the paper follow the split of Sec. I. For the seven further VLMs, VRHs are detected on a 200-sample RefCOCO [5] detection set using the same scoring function, and the masked models are evaluated on a disjoint 500-sample evaluation set of the same benchmark. The random-head control masks the same number of heads, k=20, drawn outside the detected set. Tab. A14 lists their exact accuracies as plotted in Fig. 6.

Table A14: **Exact values plotted in Fig. 6** for the seven VLMs introduced in addition to the four studied throughout the paper. Visual grounding accuracy on RefCOCO [5] after masking the top-20 VRHs versus the same number of random heads. GLM-4.6V-Flash and Phi-4-Reasoning-Vision were released after the main experiments of this work were completed.

Condition Language backbone	Magma-8B [27] Llama-3	LLaVA-NeXT-Llama3-8B [28] Llama-3	PaliGemma2-10B [29] Gemma 2	DeepSeek-VL2-Small [30] DeepSeekMoE	InternVL2.5-8B [31] InternLM2.5	GLM-4.6V-Flash [32] GLM	Phi-4-Reasoning-Vision [33] Phi
Baseline (k=0)	68.8	85.8	73.6	93.0	90.0	86.4	74.2
Random (k=20)	64.0	77.8	73.2	91.6	86.8	85.8	73.8
VRH (k=20)	**7.8**	**5.0**	**11.6**	**21.4**	**11.4**	**10.0**	**18.0**

Cross-model transfer for the Llama-3 pair. Magma-8B [27] and LLaVA-NeXT-Llama3-8B [28] share Llama-3-based language backbones, making them a third same-backbone pair for the analysis of Sec. 5.3; their score maps and the full k-sweep are plotted in Fig. 7 of the main paper. As shown in Tab. A15, cross-masking VRHs causes a large drop in visual grounding accuracy in both directions, showing that VRHs are architecturally shared beyond the Qwen family.

Table A15: **Cross-model transferability of VRHs between Magma-8B [27] and LLaVA-NeXT-Llama3-8B [28].** We report RefCOCO [5] visual grounding accuracy before and after masking the top-20 VRHs identified in the other model.

	Magma-8B	LLaVA-NeXT-Llama3-8B
Baseline (k=0)	68.8	85.8
Random (k=20)	64.0	77.8
Own VRHs (k=20)	7.8	5.0
Transferred VRHs (k=20)	0.0	5.8

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims reflect the contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitation of our work in the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: We do not include theoretical result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include experimental setup that to reproduce main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code is not included with the paper. The detected heads are listed in Appendix D, and the experimental setup is described in Appendix E; we plan to release the code on the project page.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We present evaluation setup in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report single runs without error bars. VRH detection and the masking-based evaluation use a fixed detection set and greedy decoding, so they have no run-to-run variance; the random-head control uses a single random draw.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details on compute resources are presented in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We comply the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Foundational research with no direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No high-risk data or models released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models and datasets we use are publicly released and cited, and we use them for research under their original licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets are released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We do not involve LLMs as any important, original, or non-standard component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.