Mind the Heads: Topological Representation Alignment for Multimodal LLMs

Davide Caffagni[†1] **Alberto Compagnoni**[†1,2] **Federico Melis**[†1] **Sara Sarto**[1]
Pier Luigi Dovesi[3] **Mark Granroth-Wilding**[3] **Marcella Cornia**[1] **Lorenzo Baraldi**[1]

[1]University of Modena and Reggio Emilia [2]University of Pisa [3]AMD Silo AI

aimagelab.github.io/HeRA

Abstract

Representation alignment has emerged as an effective approach to improve Multimodal Large Language Models (MLLMs) by regularizing their internal representations toward those of an external vision encoder. However, existing methods typically align a fixed layer of the language backbone, overlooking the fine-grained structure of Transformer models. In this work, we propose **He**ad-Wise **R**epresentation **A**lignment (**HeRA**), a method that enforces cross-modal alignment at the level of individual attention heads. Our approach is grounded in the Platonic Representation Hypothesis, focusing on preserving the *topological structure* of representations (*i.e.*, their local neighborhood relationships) across modalities. Following the Mutual K-Nearest Neighbor (MKNN) alignment metric, we introduce a contrastive objective that acts as a differentiable proxy for matching local structures. HeRA applies this objective during multimodal training to specific attention heads in the LLM, selected by their alignment score according to the MKNN metric. Counterintuitively, we find that aligning the *least* aligned heads yields the largest gains. Extensive evaluations across multiple MLLMs and 18 benchmarks demonstrate that HeRA consistently improves performance on challenging vision-centric tasks. Besides, according to different benchmarks, we notice a positive trend in most HeRA models in mitigating the tendency to generate visual hallucinations.

1 Introduction

Multimodal Large Language Models (MLLMs) [1, 20, 36, 44] have emerged as powerful systems capable of solving a wide range of vision-language tasks. Despite their rapid progress, improvements are still largely driven by scaling data, model size, and post-training techniques, rather than by principled changes to their internal mechanisms. While current pipelines have proven highly effective for many applications, MLLMs still exhibit notable limitations in foundational visual reasoning scenarios. Tasks such as confirming the presence of specific objects, accurately counting them, understanding spatial relationships, or parsing dense visual information remain surprisingly challenging [7, 36, 38, 46, 47]. This highlights a severe deficit in visual perception, raising a fundamental question: *how can we improve multimodal reasoning by directly intervening on the interaction between vision and language within the model?*

A growing line of work addresses this deficiency through *representation alignment* [3, 41, 49]: during multimodal training, the internal representations of the language model are regularized to match those of an external vision encoder. This can be interpreted as a form of cross-modal distillation, where the MLLM acts as a student and the vision encoder as a teacher, producing aligned representations of the same underlying content across modalities. While this technique has shown promise in improving

[†]Equal contribution. Emails: {name}.{surname}@{[1]unimore.it, [2]phd.unipi.it, [3]amd.com}

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

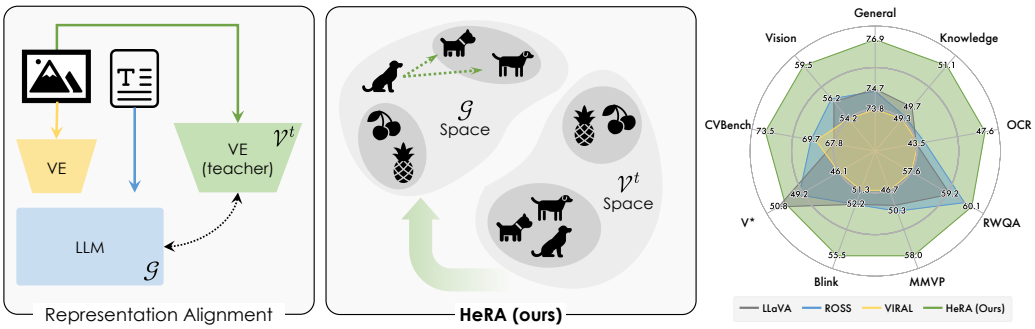

Figure 1: Standard representation alignment imposes strict vision-language feature matching (*left*), while HeRA (*center*) matches cross-modal local neighbors, leading to superior VQA results (*right*).

visual grounding, existing approaches typically align a *fixed* representation within the language backbone [49], such as the middle layer, without accounting for the internal structure of the model.

This limitation is particularly relevant in MLLMs built upon pre-trained LLMs with strong language priors. Unlike diffusion-based models, where representation alignment is applied during training from scratch [16, 50], aligning representations in MLLMs may interact in complex ways with the pre-existing organization of the language model. In this setting, selecting *which* representation to align becomes a critical design choice.

In this work, we pursue a more principled approach to representation selection, grounded in the *Platonic Representation Hypothesis* (PRH) [10, 14]. PRH posits that representations learned across different modalities are *locally consistent*: semantically similar inputs share the same neighborhood structure in their respective latent spaces. This can be interpreted as a form of *topological alignment* across modalities, where the local geometry of the representation space is preserved, and can be quantified by the Mutual K-Nearest Neighbor (MKNN) metric, which measures the agreement between local neighborhoods. While prior work has established a positive correlation between MKNN alignment and downstream language performance [8, 14], it remains unclear whether explicitly enforcing such alignment leads to improvements in MLLMs.

To address this, we propose **He**ad-Wise **R**epresentation **A**lignment (**HeRA**), a method that enforces cross-modal alignment at the level of individual attention heads rather than fixed, coarser layers. We use pre-computed MKNN scores as a diagnostic to guide this selection. Counterintuitively, we find that targeting the *least* aligned heads yields the largest gains, as it strengthens misaligned components of the model while preserving already aligned structures. As outlined in Fig. 1, HeRA applies a contrastive objective to these selected heads, encouraging their representations to match the *local topological structure* induced by an external vision encoder. This serves as a differentiable proxy for MKNN alignment, promoting cross-modal consistency without imposing rigid feature matching that often conflicts with the language modeling objective.

We evaluate HeRA across multiple MLLMs under the popular LLaVA [20] framework. Extensive evaluations across 18 benchmarks [36] demonstrate that HeRA yields consistent improvements on challenging vision-centric tasks without sacrificing (and often improving) general visual question-answering performance. Furthermore, most models register a positive trend in different visual hallucination benchmarks [11, 42], for which we credit the explicit vision-language alignment enforced by HeRA, that reduces the over-reliance on linguistic priors inherited from pre-training.

2 Related Work

The Platonic Representation Hypothesis. The Platonic Representation Hypothesis (PRH) [14] posits that models trained across different architectures, modalities, and objectives converge toward structurally similar latent spaces. Crucially, concurrent work [10] highlights that this structural consistency holds *locally* rather than globally: semantically equivalent inputs preserve their neighborhood relationships across modalities, while the absolute global geometry may differ. While PRH shows that this local cross-modal alignment naturally emerges with scale and correlates with improved capabilities, it is unclear if a *causal* relationship can be established. In this work, we investigate whether explicitly enforcing this local neighborhood consistency can lead to better MLLMs.

Vision-Centric Supervision in MLLMs. Recent efforts to boost visual understanding in MLLMs have focused on introducing explicit vision-centric supervision. Several works attempt this by enforcing representation alignment between the MLLM and a teacher vision encoder. However, these methods typically operate directly on the visual features extracted from a fixed, hard-coded layer of the language backbone. For instance, JARVIS [3] reconstructs visual targets using representations from one-quarter of the LLM depth. VIRAL [49] aligns features from the middle layer, and ROSS [41] trains a denoiser using the final layer outputs. In contrast, HeRA takes a fundamentally different approach: we enforce topological alignment within the *textual* space of the MLLM (conditioned on the multimodal input) rather than strictly matching features from the vision teacher. Furthermore, we abandon the restrictive fixed-layer assumption entirely, instead targeting specific attention heads to preserve local neighborhood structures without conflicting with the language modeling task.

Research on MLLMs is moving fast [1, 44], thanks to massive datasets, post-training, stronger LLM backbones, and natively multimodal models [32, 37]. In this work, we study a novel representation alignment objective on the LLaVA [20] framework to keep experiments computationally tractable, although we also apply it on top of state-of-the-art LLMs, such as the latest Qwen3 [48] family.

3 Proposed Method

3.1 Background

Multimodal Large Language Models (MLLMs). From an architectural perspective (refer to Fig. 2, *left*), an MLLM $\mathcal{M}$ comprises *(i)* an LLM $\mathcal{G}$, which constitutes the reasoning backbone and natural language interface of the model; *(ii)* a pre-trained vision encoder $\mathcal{V}$ to process visual inputs; and *(iii)* a projector proj, which aligns the output embedding space of $\mathcal{V}$ with the input embedding space of $\mathcal{G}$.

$\mathcal{M}$ ingests and generates text x as a sequence of tokens $\mathbf{x}_{1,\dots,T}$ converted into latent vectors by the embedding matrix of $\mathcal{G}$. On the other hand, visual inputs I are first processed by the vision encoder $\mathcal{V}$, then converted into the input embedding space of $\mathcal{G}$ by the projector: $\mathbf{v}_{1,\dots,V} = \mathrm{proj}(\mathcal{V}(I))$, and finally concatenated to the sequence of text embeddings. We train $\mathcal{M}$ to minimize the negative log-likelihood of generating token $\mathbf{x}_j$ given the image I and the preceding text $\mathbf{x}_{1,\dots,j-1}$:

$$\mathcal{L}_{\mathrm{LM}}(I_i, x_i, \mathcal{M}) = -\sum_{j}^{T} \log P\left(\mathbf{x}_j | \mathbf{v}_{1,\dots,V}, \mathbf{x}_{1,\dots,j-1}; \mathcal{M}\right). \tag{1}$$

Mutual K-Nearest Neighbor (MKNN) Alignment Metric. MKNN [14] is a kernel alignment metric enabling comparison between different representation functions. In this work, we measure the alignment between textual and visual representations of the same data point. Given an image-text pair $(I, x)_i \in \mathcal{D}$, where $\mathcal{D}$ is a dataset of aligned image-text pairs, we denote by $\mathcal{G}(x_i) \in \mathbb{R}^{d_{\mathcal{G}}}$ a representation of the text extracted from the language model, and by $\mathcal{V}^t(I_i) \in \mathbb{R}^{d_{\mathcal{V}^t}}$ representation of the corresponding image from a *teacher* vision encoder. Here, $\mathcal{G}(x_i)$ refers to an internal representation (*e.g.*, from intermediate layers or attention heads).

For a dataset $\mathcal{D}$, MKNN measures the agreement between the local neighborhood structures induced by the two representation spaces, by computing the average intersection of their k-nearest neighbor sets. We denote by $\mathcal{N}_k^{\mathcal{F}}(\cdot)$ the operator returning the k-nearest neighbors according to maximum dot product similarity in the L2-normalized latent space of the embedding function $\mathcal{F}$. For instance, in the language space, where we average pool the output embeddings, it is formally defined as follows:

$$\mathcal{N}_k^{\mathcal{G}}(x_i) = \mathrm{argmax}_{j \neq i}^{(k)} \mathcal{G}(x_i)^\top \mathcal{G}(x_j). \tag{2}$$

In the visual space, we pool by taking the `CLS` embeddings at the output of the vision encoder $\mathcal{V}^t$. The MKNN alignment metric between $\mathcal{G}$ and $\mathcal{V}^t$ is thus defined as:

$$m_{k\mathrm{NN}}(\mathcal{G}, \mathcal{V}^t, \mathcal{D}) = \mathop{\mathbb{E}}_{(I,x)_i \in \mathcal{D}} \left[\frac{1}{k} \left| \mathcal{N}_k^{\mathcal{G}}(x_i) \cap \mathcal{N}_k^{\mathcal{V}^t}(I_i) \right| \right] \in [0, 1], \tag{3}$$

where $|\cdot|$ denotes set cardinality.

High scores in Eq. 3 reflect that the *local* topological latent structure generated by $\mathcal{G}$ is preserved in the latent space of $\mathcal{V}^t$.

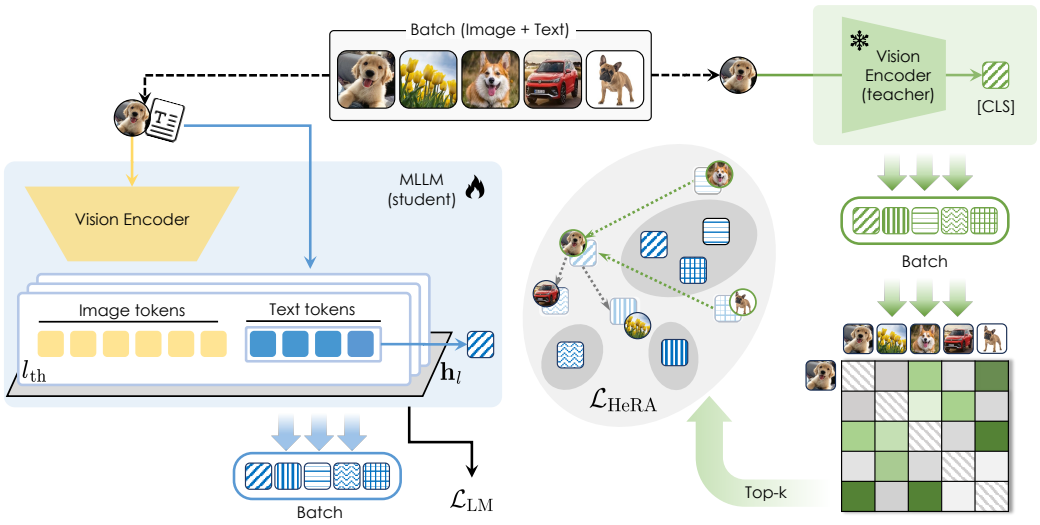

Figure 2: **Overview of HeRA.** Alongside the standard language modeling objective ($\mathcal{L}_{\text{LM}}$), HeRA employs a contrastive loss ($\mathcal{L}_{\text{HeRA}}$) to pull representations from selected LLM attention heads closer to their k-nearest neighbors (Top-k), computed in the latent space of a frozen teacher vision encoder.

3.2 Contrastive Learning as a Proxy for Representation Alignment

For a fixed vision encoder $\mathcal{V}^t$, $m_{k\text{NN}}$ has been positively correlated with better performance on language modeling tasks [14]. We want to probe whether a causal effect could exist in a multimodal scenario: *can we train a better MLLM by explicitly enforcing alignment with the visual domain?*

A natural approach would be to directly maximize $m_{k\text{NN}}$ during training. However, Eq. 3 depends on discrete neighbor indices and is therefore not differentiable. To address this, we propose a contrastive objective that encourages the multimodal representations produced by $\mathcal{M}$ to match the local neighborhood structure induced by the teacher vision encoder.

Given a batch $\mathcal{B} = \{(I, x)_i\}$, let $\mathcal{M}(I_i, x_i) \in \mathbb{R}^{d_{\mathcal{G}}}$ denote the L2-normalized multimodal representation obtained by average pooling the text embeddings[1]. For each sample i, we first identify a set of target neighbors $\mathcal{N}_k^{\mathcal{V}^t}(I_i)$, corresponding to the k nearest neighbors of I_i in the teacher vision space. We then train $\mathcal{M}$ so that its representation $\mathcal{M}(I_i, x_i)$ is close to the representations of these neighbors, while being separated from the rest of the batch. Formally, this can be achieved via a multi-target variant of the InfoNCE [29] loss:

$$\mathcal{L}_{\text{RA}}(I_i, x_i, \mathcal{M}) = -\frac{1}{k} \sum_{j \in \mathcal{N}_k^{\mathcal{V}^t}(I_i)} \log \frac{\exp\left(\frac{\mathcal{M}(I_i,x_i)^\top \mathcal{M}(I_j,x_j)}{\tau}\right)}{\sum_{z \in \mathcal{B}, z \neq i} \exp\left(\frac{\mathcal{M}(I_i,x_i)^\top \mathcal{M}(I_z,x_z)}{\tau}\right)}, \tag{4}$$

where τ is a learnable scalar governing the sharpness of the distribution. Minimizing Eq. 4 teaches the *student* model $\mathcal{M}$ to produce multimodal representations sharing the same *local* neighborhood as the corresponding visual representations from the *teacher* model $\mathcal{V}^t$, which is exactly the property measured by the $m_{k\text{NN}}$ metric.

3.3 Head-Wise Representation Alignment (HeRA)

While the contrastive objective in Eq. 4 enforces alignment at the level of a single pooled representation, it does not account for the internal structure of the language backbone. In particular, $\mathcal{G}$ processes the entire multimodal sequence, suggesting that alignment can be more precisely controlled by operating directly on its internal representations.

In principle, $\mathcal{G}$ generates multiple representations for a given input. Indeed, we can collect a representation from each Transformer layer of the language backbone. For language modeling (*i.e.*, Eq. 1), we care about the last layer to sample the next token (after passing through the unembedding

[1] As text embeddings are conditioned on visual inputs, they can be considered inherently multimodal.

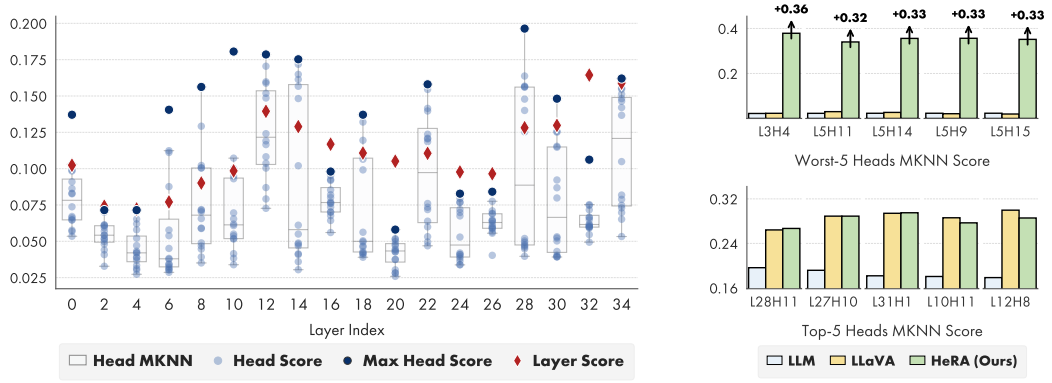

Figure 3: **Left:** Alignment with DINOv2-L, measured with the MKNN metric on each layer and attention head of Qwen2.5-3B. **Right:** MKNN scores of the Worst-5 and Top-5 heads, computed on *(i)* the base LLM; *(ii)* after the LLaVA multimodal training; and *(iii)* after the addition of HeRA.

matrix). Conversely, representation alignment methods [16, 50] typically rely on intermediate layers. However, the choice of which layer(s) to use is typically treated as a fixed hyperparameter (*e.g.*, selecting the middle layer [49]), which does not adapt to the specific structure of a given model.

In this work, we instead probe finer-grained representations within the language model, specifically focusing on the individual attention heads in each multi-head self-attention layer of $\mathcal{G}$. Because different attention heads specialize in different roles within an LLM [27, 28, 43], working at the head level enables a more atomic intervention on the language model, mitigating potential conflicting effects between language modeling and representation alignment.

Head-level Representations. In standard multi-head attention, the final output of a layer is obtained by concatenating the outputs of the individual heads and multiplying them by an output projection matrix $\mathbf{W}_O \in \mathbb{R}^{d_\mathcal{G} \times d_\mathcal{G}}$. Let $\mathbf{h}_{l,h} \in \mathbb{R}^{d_{head}}$ be the output of the h-th attention head in layer l, where $d_{head} = d_\mathcal{G}/H$. The output projection can be written as:

$$\text{MultiHead}(\cdot) = [\mathbf{h}_{l,1}(\cdot), \ldots, \mathbf{h}_{l,H}(\cdot)]\mathbf{W}_O. \tag{5}$$

Because matrix multiplication is a linear operator, we can decompose $\mathbf{W}_O$ into H distinct blocks along its row dimension, such that $\mathbf{W}_O = [\mathbf{W}_{O,1}^\top, \ldots, \mathbf{W}_{O,H}^\top]^\top$, with each $\mathbf{W}_{O,h} \in \mathbb{R}^{d_{head} \times d_\mathcal{G}}$. The multi-head attention output can then be equivalently written as:

$$\text{MultiHead}(\cdot) = \sum_{h=1}^{H} \mathbf{h}_{l,h}(\cdot)\mathbf{W}_{O,h}. \tag{6}$$

This decomposition allows us to isolate the projected contribution of each head before it is summed into the shared residual stream. Given a multimodal input (I_i, x_i), we define:

$$\mathcal{M}^{l,h}(I_i, x_i) = \mathbf{h}_{l,h}(I_i, x_i)\mathbf{W}_{O,h} \in \mathbb{R}^{d_\mathcal{G}}, \tag{7}$$

where $\mathcal{M}^{l,h}$ denotes the representation extracted from the h-th attentive head in the l-th layer of $\mathcal{M}$ during the multimodal forward pass.

Head-wise Alignment Objective. We apply the contrastive alignment loss of Eq. 4 independently to selected head-level representations. For a set of layer-head indices $\mathcal{H} = \{(l, h)\}$, we define our head-wise representation alignment loss as the average alignment loss over the selected heads:

$$\mathcal{L}_{\text{HeRA}}(I_i, x_i, \mathcal{M}, \mathcal{H}) = \mathop{\mathbb{E}}_{(l,h)\in\mathcal{H}} \left[\mathcal{L}_{\text{RA}}(I_i, x_i, \mathcal{M}^{l,h})\right]. \tag{8}$$

The final training objective (illustrated in Fig. 2) is given by the sum of the language modeling and head-wise representation alignment losses:

$$\mathcal{L}(I_i, x_i, \mathcal{M}, \mathcal{H}) = \mathcal{L}_{\text{LM}}(I_i, x_i, \mathcal{M}) + \lambda\mathcal{L}_{\text{HeRA}}(I_i, x_i, \mathcal{M}, \mathcal{H}), \tag{9}$$

where λ is a fixed hyperparameter to balance the two contributions.

Table 1: Ablation study on the choice of *(i)* the objective for representation alignment (feature- vs. contrastive-based), and *(ii)* the granularity of the LLM representation to align (layer- vs. head-level).

Representation Alignment			**LLM:** Qwen2.5-3B					**LLM:** Qwen3-4B				
Objective	Granularity	Selection	**General**	**Knowledge**	**OCR**	**Vision**	**All**	**General**	**Knowledge**	**OCR**	**Vision**	**All**
-	-	-	73.5	46.7	42.2	50.5	54.2	75.6	49.6	43.8	56.3	57.4
Feature	Layer	Middle	72.6	46.4	42.3	50.1	53.8	75.0	48.6	42.8	53.4	56.0
Feature	Head	Worst (5)	74.0	46.5	42.8	51.6	54.7	76.0	49.3	44.7	55.9	57.6
Contrastive	Layer	Middle	72.8	45.7	41.6	51.1	53.8	75.8	49.5	43.7	57.2	57.6
Contrastive	Layer	Worst (5)	73.1	46.2	40.9	51.7	54.0	73.2	47.9	41.2	52.1	54.6
Contrastive	Head	Random (5)	73.7	46.6	42.1	49.7	54.0	**76.1**	49.7	44.0	56.6	57.7
Contrastive	Head	Top (5)	73.8	46.5	42.8	50.2	54.3	75.9	49.9	**45.2**	56.3	57.8
Contrastive	Head	Worst (1)	73.5	46.9	42.9	51.2	54.6	76.0	49.7	44.1	57.0	57.8
Contrastive	Head	Worst (3)	73.9	**47.6**	43.6	51.0	55.0	75.8	49.9	44.7	57.0	57.9
Contrastive	Head	Worst (10)	62.6	45.5	34.6	38.9	46.0	75.8	49.7	44.7	56.3	57.7
Contrastive	**Head**	**Worst (5)**	**74.5**	47.5	**43.8**	**52.9**	**55.7**	76.0	**50.1**	44.5	**58.5**	**58.4**

Heads Selection. For an LLM with L layers and H attention heads, the total number of heads is $L \times H$. In practice, this number is in the order of hundreds, making an extensive search for the optimal $\mathcal{H}$ unfeasible. To this end, we propose to exploit the $m_{k\text{NN}}$ metric of Eq. 3 to rank the heads by their alignment score with the vision encoder $\mathcal{V}^t$. We compute this rank using the language model $\mathcal{G}$ *before* the multimodal training, so that its representations are purely textual. Surprisingly, we find that there always exists a set of heads whose alignment score greatly exceeds that of any layer in the same model (see Fig. 3, left). Once the alignment rank is computed, we posit to select a subset of m heads following two reasonable strategies. Specifically, we select either *(i)* the best aligned heads (*i.e.*, $\mathcal{H}_m^{\text{top}}$), as forcing the alignment should be easier because they start from an already partially aligned latent space, or *(ii)* the least aligned heads (*i.e.*, $\mathcal{H}_m^{\text{worst}}$), so to strengthen the components of the model further away from the visual domain. Empirically, we find that choosing $\mathcal{H}_m^{\text{worst}}$ works best: it boosts the alignment of poorly aligned heads, while preserving the alignment of the strongest heads, as displayed in Fig. 3, right.

4 Experiments

4.1 Experimental Settings

Training Details. We train all models following the two-stage LLaVA-1.5 pipeline [20], with the same training data and protocol. As vision encoder $\mathcal{V}$, we adopt SigLIP2 ViT-SO400M/14@384 [39] across all experiments. For the LLM $\mathcal{G}$, we consider a diverse set of architectures, including Vicuna [4], LLama3 [9], Qwen2.5 [31], and Qwen3 [48], ranging from 3B to 14B parameters. We apply the HeRA loss (cf. Eq. 8) in both training stages, using λ equal to 0.01, and k equal to 10. Unless otherwise specified, we use DINOv2 ViT-L [30] as teacher vision encoder $\mathcal{V}^t$.

Head Selection. We perform head selection *prior* to multimodal training. Specifically, we compute the $m_{k\text{NN}}$ alignment score (Eq. 3) for each head using the LLM $\mathcal{G}$, before any multimodal finetuning. This produces a ranking of heads based on their degree of alignment with the visual domain. The scores are computed on 1,000 samples from the GranD dataset [33], which provides highly detailed captions enabling a reliable estimation of cross-modal neighborhood structure. We select the $m = 5$ least aligned heads and restrict the application of HeRA to this subset throughout training.

Evaluation Benchmarks. We primarily evaluate our method using the Cambrian comprehensive benchmark suite [36] covering General, Knowledge, OCR, and Vision tasks. In addition, we evaluate hallucination robustness on CHAIR-MSCOCO [52], AMBER [42], and HallusionBench [11]. The complete details on the evaluation datasets are reported in Appendix B.

4.2 Ablation Studies and Analyses

We start by presenting a set of ablation studies in Table 1 designed to understand how key architectural and objective choices influence the behavior of HeRA. For these experiments, we employ Qwen2.5-3B [31] and Qwen3-4B [48] as the underlying LLMs. As a baseline, we consider a LLaVA model trained on top of the same LLMs without alignment regularization (*first* row). For all configurations, we use the same training settings used in our approach, as described in Sec. 4.1.

Table 2: VQA results of HeRA applied to the LLaVA training recipe on different LLMs.

	General	Knowledge	OCR	Vision						
Model	Avg	Avg	Avg	RWQA	MMVP	Blink	V*	CVBench	Avg	
Qwen2.5-3B	73.5	46.7	42.2	55.2	46.0	46.8	44.5	**60.2**	50.5	
+ HeRA (Ours)	**74.5**	**47.5**	**43.8**	**56.3**	**48.0**	**49.1**	**51.3**	59.6	**52.9**	Δ+2.4
Qwen3-4B	75.6	49.6	43.8	59.9	48.0	**55.1**	50.3	68.3	56.3	
+ HeRA (Ours)	**76.0**	**50.1**	**44.5**	**61.3**	**56.0**	53.7	**52.4**	**69.1**	**58.5**	Δ+2.2
Vicuna-7B	**72.2**	44.3	**45.7**	56.5	38.7	46.8	44.5	**62.1**	49.7	
+ HeRA (Ours)	72.1	**44.5**	**45.7**	**57.8**	**42.7**	**47.9**	**49.7**	61.9	**52.0**	Δ+2.3
Llama3-8B	73.3	45.0	43.0	60.1	46.0	49.2	44.0	**69.5**	53.8	
+ HeRA (Ours)	**74.6**	**46.3**	**44.7**	**60.4**	**46.7**	**50.2**	**51.8**	66.4	**55.1**	Δ+1.3
Qwen2.5-7B	76.2	50.2	47.9	59.6	51.3	**51.7**	**50.8**	70.2	56.7	
+ HeRA (Ours)	**76.5**	**50.5**	**48.6**	**61.3**	**54.0**	50.2	50.3	**71.3**	**57.4**	Δ+0.7
Qwen3-8B	74.7	49.7	43.5	59.2	49.3	52.2	**50.8**	67.8	55.9	
+ HeRA (Ours)	**76.9**	**51.1**	**47.6**	**60.1**	**58.0**	**55.5**	50.3	**73.5**	**59.5**	Δ+3.6
Vicuna-13B	73.4	45.5	**47.7**	**58.7**	**44.7**	50.6	46.1	63.7	52.7	
+ HeRA (Ours)	**73.6**	**45.7**	47.6	57.3	44.0	**52.3**	**49.2**	**66.5**	**53.9**	Δ+1.2
Qwen2.5-14B	75.6	50.7	44.8	**60.1**	47.3	52.7	46.6	67.9	54.9	
+ HeRA (Ours)	**77.4**	**52.8**	**49.3**	**60.1**	**52.0**	**55.2**	**52.4**	**71.6**	**58.3**	Δ+3.4
Qwen3-14B	77.4	**52.8**	46.1	60.3	57.3	**52.6**	50.3	**70.8**	58.2	
+ HeRA (Ours)	**77.7**	52.6	**47.8**	**62.5**	**58.0**	52.0	**51.8**	70.2	**58.9**	Δ+0.7

Objective and Granularity. First, we consider a standard representation alignment approach [49, 50], where the LLM is trained to maximize the cosine similarity between visual-token features at the middle layer and those from the teacher vision encoder (*second* row), using a trainable projector. While this is ineffective on both LLMs, switching to our contrastive learning objective using the same representation granularity (*fourth* row) shows promising results on vision tasks.

Head-Level Alignment and Selection. Sticking with the contrastive objective, we move to a finer granularity, considering the textual representations from specific attention heads in the LLM. The head selection criterion follows the MKNN alignment score with the vision encoder: we select either the top-5 (*seventh* row) or worst-5 (*last* row) according to this ranking. On both LLMs, we record a striking difference favoring alignment on the worst-5 heads. For instance, compared to the top-5 selection, Qwen2.5-3B boosts its performance on vision-centric tasks by +2.7 points, whereas Qwen3-4B enjoys a +2.2 points gain. As control trials, we also apply the contrastive alignment on a random subset of 5 heads (*sixth* row), yielding no clear benefit, and experiment with the "worst-5" selection criterion at the layer-level (*fifth* row), which actually registers a mild performance regression on Qwen2.5-3B and a severe degradation on Qwen3-4B.

Connection to the Platonic Representation Hypothesis. The superiority of the worst-5 strategy corroborates the positive correlation between alignment and performance reported by the PRH [14]. As shown in Fig. 3 (*right*), after HeRA training, the worst-5 heads drastically increase their vision-language alignment without penalizing the alignment of the top-5 heads. Conversely, explicitly forcing alignment on the top-5 heads has no meaningful collateral impact on the worst-5 heads. Interestingly, aligning the visual features from the worst-5 heads (*third* row) is ineffective, and has little impact on their alignment scores (see the plot in Fig. 5 in Appendix C).

Number of Heads to Align. Finally, we ablate the number of heads to align (Table 1, *bottom*). Using 3 heads, or even a single one, yields modest gains on both LLMs, while the best overall performance is achieved with 5 heads, particularly on challenging vision benchmarks. Conversely, further scaling up the number of heads to 10 leads to a regression, particularly with Qwen2.5-3B, indicating that aligning too many heads begins to conflict with the core language modeling task.

4.3 Main Experimental Results

Results on Cambrian Benchmarks. To assess the generalizability and scalability of our proposed representation alignment regularization, we evaluate HeRA across a diverse suite of language models, as reported in Table 2. We deliberately select models spanning multiple architectural generations to ensure our findings are not isolated to a specific design. This includes established baselines like the Vicuna family, as well as the latest generation of state-of-the-art open-source models, such as Qwen3.

Table 3: Results of HeRA on visual hallucinations benchmarks.

	MSCOCO		AMBER (Generative)				AMBER (Discriminative)				HallusionBench		
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	$CHAIR_i \downarrow$	Cover ↑	HalRate ↓	Cog ↓	Acc ↑	Prec ↑	Rec ↑	F1 ↑	qAcc ↑	Easy ↑	Hard ↑
Qwen2.5-3B	44.2	12.6	5.8	52.0	30.1	3.3	83.6	83.9	**93.2**	88.3	23.7	60.7	55.2
+ HeRA (Ours)	**44.0**	**11.8**	**5.2**	**52.3**	**27.6**	**2.8**	**85.6**	**86.5**	92.8	**89.5**	**24.4**	**61.3**	**60.2**
Qwen3-4B	**42.0**	**11.7**	**5.5**	**53.0**	28.2	3.0	**86.5**	89.9	**89.6**	**89.7**	**22.2**	**63.1**	**54.6**
+ HeRA (Ours)	43.8	11.9	**5.5**	52.5	**27.6**	**2.8**	**86.5**	**90.1**	89.4	**89.7**	18.9	60.7	50.3
Vicuna-7B	**46.6**	**12.6**	6.3	52.2	**30.8**	**3.3**	84.2	90.0	85.8	87.8	15.6	**55.4**	45.7
+ HeRA (Ours)	47.6	12.9	**6.1**	**52.6**	31.9	3.5	**85.9**	**90.8**	**87.7**	**89.2**	**16.3**	55.0	**49.6**
Llama3-8B	44.8	13.0	**5.5**	51.5	27.5	2.9	**85.9**	**89.2**	89.6	89.4	16.9	**58.7**	**49.9**
+ HeRA (Ours)	**40.2**	**11.7**	5.6	**51.7**	**26.6**	**2.7**	**85.9**	88.3	**90.8**	**89.5**	**17.4**	58.0	49.3
Qwen2.5-7B	45.2	**11.9**	5.3	52.3	26.7	**2.7**	87.6	**89.6**	91.9	90.7	21.5	**66.8**	49.4
+ HeRA (Ours)	**44.8**	12.4	**4.9**	**52.5**	**25.1**	**2.7**	**87.7**	89.1	**92.8**	**90.9**	**23.3**	66.6	**50.9**
Qwen3-8B	42.0	12.3	5.8	52.8	28.9	3.2	87.3	90.3	90.6	90.4	21.1	61.1	53.3
+ HeRA (Ours)	**39.2**	**11.0**	**5.2**	**53.1**	**27.9**	**3.0**	**89.0**	**92.4**	**90.9**	**91.6**	**25.7**	**66.2**	**54.5**
Vicuna-13B	**43.0**	**11.9**	**6.1**	52.5	**28.4**	3.3	84.8	**93.9**	82.4	87.8	13.0	55.2	43.9
+ HeRA (Ours)	47.4	12.2	**6.1**	**52.9**	30.3	**3.2**	**85.7**	91.7	**86.2**	**88.9**	**14.7**	**56.0**	**45.6**
Qwen2.5-14B	42.6	12.2	6.1	**52.3**	29.4	3.6	86.2	87.9	91.8	89.8	23.7	61.5	**56.5**
+ HeRA (Ours)	**39.0**	**10.6**	**5.4**	51.9	**28.0**	**3.1**	**89.0**	**90.4**	**93.4**	**91.9**	**25.9**	**66.2**	56.4
Qwen3-14B	43.2	11.7	5.5	**53.1**	28.8	3.2	88.8	**91.6**	91.5	91.5	23.7	66.2	53.0
+ HeRA (Ours)	**41.2**	**11.4**	**5.0**	53.0	**26.7**	**3.0**	**88.9**	90.7	**92.7**	**91.7**	**27.9**	**69.0**	**56.8**

Table 4: VQA comparison of different representation alignment strategies for MLLMs.

	General	Knowledge	OCR	Vision						
Alignment	Avg	Avg	Avg	RWQA	MMVP	Blink	V*	CVBench	Avg	
-	74.7	49.7	43.5	59.2	49.3	52.2	50.8	67.8	55.9	
ROSS [41]	74.6	49.4	44.0	59.8	50.3	52.2	49.2	69.7	56.2	Δ+0.3
VIRAL [49]	73.8	49.3	43.6	57.6	46.7	51.3	46.1	69.3	54.2	Δ-1.7
JARVIS [3]	76.8	49.9	46.2	59.2	54.7	54.2	**55.5**	69.9	58.7	Δ+2.8
CMAR [8]	76.4	51.0	46.0	58.8	50.0	52.1	52.9	70.9	56.9	Δ+1.0
HeRA (Ours)	**76.9**	**51.1**	**47.6**	**60.1**	**58.0**	**55.5**	50.3	**73.5**	**59.5**	Δ+3.6

Furthermore, we scale the parameter count across our experiments, progressing from compact models (3B and 4B) up to larger reasoning engines (13B and 14B). Detailed results are in Appendix C.

A common risk when modifying the internal representations of a pre-trained LLM is the potential degradation of its inherent linguistic and reasoning priors. However, the results demonstrate that our head-wise alignment successfully preserves, and frequently improves, the model's core competencies. Across the General, Knowledge, and OCR task categories, the inclusion of HeRA consistently yields stable or higher average scores compared to the standard LLaVA training recipe. For instance, Qwen2.5-14B sees its General average rise from 75.6 to 77.4, while its Knowledge and OCR averages experience parallel uplifts. This indicates that isolating the alignment to a strategic subset of attention heads successfully mitigates catastrophic interference with the language modeling objective.

The most substantial impact of HeRA is observed in the Vision-Centric benchmarks, which directly measure visual perception, spatial reasoning, and multimodal grounding. Regardless of the underlying architecture, our method systematically drives up visual performance. Earlier models like Vicuna-7B experience a robust +2.3 point gain, proving that proper representation alignment can benefit legacy architectures. Simultaneously, modern models equipped with stronger text priors also reap significant benefits; notably, Qwen3-8B achieves the highest individual leap with a +3.6 average improvement.

This trend persists as we scale the LLM backbone. When applied to the largest models in our suite, the representation alignment remains highly effective, with Qwen2.5-14B securing a $+3.4$ point increase and Qwen3-14B pushing the upper bound of the Vision-Centric average to 58.9.

Results on Hallucination Benchmarks. Although mitigating hallucinations, an open problem in MLLMs, is not explicitly enforced by our contrastive representation alignment loss, we find that HeRA has a positive effect on it, as outlined in Table 3. Across both the CHAIR-MSCOCO and AMBER generative benchmarks, models trained with HeRA consistently lower their hallucination rates (*e.g.*, $CHAIR_s$ and $CHAIR_i$). Crucially, on AMBER, this reduction in hallucinations is achieved while simultaneously improving, or at least maintaining, the cognition (Cog) score, which is a challenging balance, as models often become overly conservative when penalized for hallucinations.

In discriminative settings, HeRA yields steady improvements in accuracy and F1 scores on AMBER, indicating a more robust visual grounding. On HallusionBench, the method drives positive gains

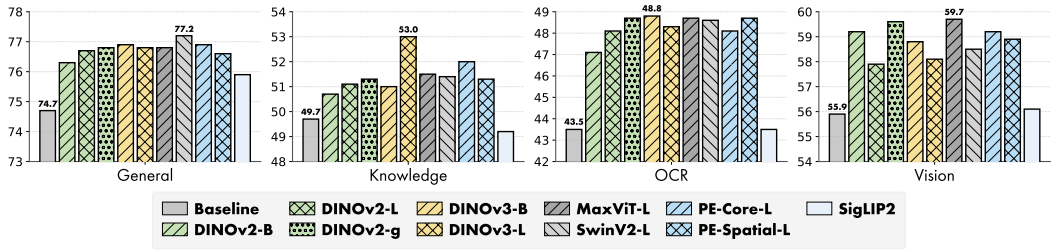

Figure 4: VQA results of HeRA with different teacher vision encoders.

across nearly all models, significantly improving qAcc, Easy, and Hard metrics. The sole exception is Qwen3-4B, with a drop on this specific benchmark; however, this is vastly offset by the superior performance gains across standard VQA and Vision-Centric benchmarks (as detailed in Table 2). Ultimately, we credit the positive trend in hallucination reduction to the explicit vision-language alignment imposed by HeRA, which curbs the over-reliance of the MLLM on linguistic priors.

Comparison with Previous Representation Alignment Methods. In Table 4, we compare HeRA against recent representation alignment strategies: ROSS [41] trains an auxiliary denoiser network conditioned on LLM visual features to recover visual tokens; VIRAL [49] aligns visual features from the LLM middle layer during the instruction tuning stage; JARVIS [3] reconstructs masked image latents using representations from one quarter of the LLM depth; and CMAR [8] optimizes the CKA alignment metric between textual features from the penultimate LLM layer and the teacher encoder. We run each experiment according to its official implementation. All methods are trained on the same LLaVA [20] dataset, feature Qwen3-8B [48] as the LLM, SigLIP2 ViT-SO400M/14@384 [39] as the vision encoder, and DINOv2-L [30] as the teacher for alignment (with the exception of ROSS).

Compared to these fixed-layer approaches, our targeted head-wise alignment proves significantly more effective. Notably, the strict pointwise feature alignment operated by VIRAL is the only method that registers a regression compared to LLaVA (*first* row). Furthermore, while CMAR shares our goal of topological alignment between spaces of different modalities, the CKA metric forces *global* point-wise relationships to match those from vision encoder. By contrast, HeRA focuses strictly on preserving *local* neighborhood structures, without imposing rigid distance constraints between samples. Ultimately, on the demanding Vision-Centric benchmarks, HeRA yields a +3.6 point average improvement, outperforming the next best method, JARVIS (+2.8), and achieves the highest overall scores across the General, Knowledge, and OCR tasks. Detailed results are provided in Appendix C.

4.4 Varying the Teacher Visual Encoder

In Fig. 4, we experiment with different teacher vision encoders used to extract the targets for the HeRA contrastive loss. All models are trained using Qwen3-8B as the language backbone and SigLIP2 as the primary vision encoder. We observe that using SigLIP2 itself as the teacher is mostly ineffective, as its visual representations are already incorporated into the MLLM. Indeed, aligning with DINO-based [30, 34] teachers yields strong and consistent gains, even when using the base models (DINOv2-B and DINOv3-B). We further evaluate other visual backbones [40, 23] and two models from the Perception Encoder family [2] which consistently yield slight improvements over DINOv2, further supporting the generality of HeRA across visual teachers. Despite the slight improvements of these models over DINOv2-L, we retain the latter to ensure a fair comparison with competing methods [3, 49, 8]. However, we note no clear benefits from employing the larger 1B-parameter DINOv2-g, suggesting that base vision encoders may suffice for topological alignment.

4.5 Interpretability Analysis Through Visual Attention

To understand the original role of the worst aligned heads and how HeRA changes it, we analyze the visual attention of LLaVA and three identically trained HeRA variants aligning the Worst-5, Top-5, and Random-5 heads. Specifically, for each attention head h at layer l, we measure the fraction of attention that *answer* tokens assign to image tokens:

$$\text{VisAttn}(l,h) = \mathbb{E}_{(I,x)\in\mathcal{D}}\,\mathbb{E}_{t\in\mathcal{A}}\left[\sum_{v\in\mathcal{V}} A_{t,v}^{(l,h)}\right], \tag{10}$$

Table 5: Visual attention (% of attention mass on image tokens) by head set and the model-wide average across all 576 heads. We also report the ranks of the five aligned heads among all heads.

Model	Aligned Heads	Worst-5	Top-5	Random-5	All heads	Ranks of aligned heads
LLaVA	-	11.9	18.7	12.2	12.2	–
+ **HeRA (Ours)**	Worst-5	**75.0**	18.1	12.7	11.2	1, 2, 3, 4, 14
+ HeRA	Top-5	41.8	**85.4**	18.2	20.6	1, 2, 3, 4, 5
+ HeRA	Random-5	48.8	49.3	**79.3**	25.5	1, 2, 3, 5, 143

where $A^{(l,h)}$ is the post-softmax attention matrix, $\mathcal{V}$ indexes the visual positions and $\mathcal{A}$ the answer tokens. We evaluate all head sets in all four Qwen2.5-3B models. Heads are ranked among the model's 576 heads, where rank 1 means highest visual attention. Results are averaged across the vision-centric benchmarks, which show nearly identical trends.

As shown in Table 5, the contrastive objective reliably converts the aligned heads into visual heads, regardless of which heads are selected. In every variant, the trained heads become the most image-attending heads of the entire backbone (diagonal entries of Table 5), whereas in LLaVA the same heads sit around the middle of the ranking. Since $\mathcal{L}_{\text{HeRA}}$ aligns pooled head outputs and never uses attention maps, this shift toward image tokens emerges implicitly from the alignment.

Table 6 reveals that, in LLaVA, the Worst-5 heads behave as attention sinks, allocating roughly three quarters of their attention to the first sequence position, far above the model average. We therefore posit a connection between low cross-modal alignment (*i.e.*, low MKNN score) and underutilized heads. After training, all three selections are dominated by image attention, but Worst-5 mainly redirects unused sink attention, reducing attention to the other text tokens by only 12.1 points, versus 33.0 and 47.6 points for Top-5 and Random-5.

Table 6: Answer tokens attention budget (%) for each head set in LLaVA and HeRA.

Model	Heads	Image	Sink	Text
LLaVA	Worst-5	11.9	**74.3**	13.8
LLaVA	Top-5	18.7	41.7	39.7
LLaVA	Random-5	12.2	31.8	56.0
LLaVA	All	12.2	28.3	59.6
HeRA (Ours)	Worst-5	**75.0**	23.3	1.7
HeRA	Top-5	**85.4**	7.9	6.7
HeRA	Random-5	**79.3**	12.3	8.4

The three variants also differ sharply in how confined the intervention is. With Worst-5, the increase in visual attention remains localized to the aligned heads: the model-wide average and the two non-aligned head sets stay close to the LLaVA baseline (Table 5, second row). With Top-5 or Random-5, visual attention instead spreads across the whole model: the model-wide average rises from 12.2% to 20.6% and 25.5%, respectively, and head sets that were never aligned become substantially more visual than in LLaVA. This specificity is also what separates the variants downstream: as reported in Table 1, on Qwen2.5-3B the Worst-5 variant is the only one that improves the vision-centric average over the baseline, while the Top-5 and Random-5 selections fail to do so, despite their aligned heads being individually even more visual (Table 5). We hypothesize that broader changes disrupt non-visual functions, whereas Worst-5 repurposes underused heads without affecting the others.

5 Conclusion

In this work, we introduced **He**ad-Wise **R**epresentation **A**lignment (**HeRA**), a novel method to enhance Multimodal Large Language Models through topological representation alignment. Guided by the Platonic Representation Hypothesis, HeRA uses a contrastive proxy for the MKNN metric to align specific attention heads with an external vision encoder, demonstrating that targeting the *least* aligned heads yields the most substantial gains. Evaluations across multiple architectures and benchmarks reveal that our approach significantly benefits demanding vision-centric tasks without compromising, even improving, core linguistic capabilities, while mitigating visual hallucinations.

Acknowledgments

This work has been supported by the EU Horizon project "ELLIOT" (No. 101214398), by the EuroHPC JU project "MINERVA" (GA No. 101182737), and by the PNRR project "ITSERR" (CUP B53C22001770006) funded by the EU - NextGenerationEU. We also acknowledge EuroHPC JU for awarding the project EHPC-AIF-2025SC04-225 access to LUMI at CSC, Finland.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2026.

[3] Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Pier Luigi Dovesi, Shaghayegh Roohi, Mark Granroth-Wilding, and Rita Cucchiara. Seeing Beyond Words: Self-Supervised Visual Learning for Multimodal Large Language Models. *arXiv preprint arXiv:2512.15885*, 2025.

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023.

[5] David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling Language-Free Visual Representation Learning. In *ICCV*, 2025.

[6] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*, 2023.

[7] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal Large Language Models Can See But Not Perceive. In *ECCV*, 2024.

[8] Yulu Gan, Kaiya Ivy Zhao, and Phillip Isola. Cross-Modal Alignment Regularization: Enhancing Language Models with Vision Model Representations. In *ICLR Workshops*, 2025.

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

[10] Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic Representation Hypothesis: An Aristotelian View. In *ICML*, 2026.

[11] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *CVPR*, 2024.

[12] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. VizWiz Grand Challenge: Answering Visual Questions From Blind People. In *CVPR*, 2018.

[13] Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*, 2019.

[14] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic Representation Hypothesis. In *ICML*, 2024.

[15] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

[16] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for End-to-End Tuning with Latent Diffusion Transformers. In *ICCV*, 2025.

[17] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[18] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *EMNLP*, 2023.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *CVPR*, 2024.

[21] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is Your Multi-modal Model an All-around Player? In *ECCV*, 2024.

[22] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. *Sci China Inf Sci*, 67(12):220102, 2024.

[23] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022.

[24] 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.

[25] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *NeurIPS*, 2022.

[26] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *ACL*, 2022.

[27] Andrew Joohun Nam, Henry Conklin, Yukang Yang, Thomas L Griffiths, Jonathan D Cohen, and Sarah-Jane Leslie. Causal Head Gating: A Framework for Interpreting Roles of Attention Heads in Transformers. In *NeurIPS*, 2025.

[28] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-Context Learning and Induction Heads. *arXiv preprint arXiv:2209.11895*, 2022.

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*, 2018.

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *TMLR*, pages 1–31, 2024.

[31] Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024.

[32] Qwen Team. Qwen3.5: Towards Native Multimodal Agents, 2026.

[33] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel Grounding Large Multimodal Model. In *CVPR*, 2024.

[34] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[35] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *CVPR*, 2019.

[36] 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.

[37] Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, et al. Beyond Language Modeling: An Exploration of Multimodal Pretraining. *arXiv preprint arXiv:2603.03276*, 2026.

[38] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *CVPR*, 2024.

[39] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025.

[40] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV*, 2022.

[41] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive Visual Instruction Tuning. In *ICLR*, 2025.

[42] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

[43] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *ICLR*, 2023.

[44] 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.

[45] Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. FineVision: Open Data Is All You Need. *arXiv preprint arXiv:2510.17269*, 2025.

[46] Penghao Wu and Saining Xie. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. In *CVPR*, 2024.

[47] xAI. Grok, 2024.

[48] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

[49] Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, et al. Visual Representation Alignment for Multimodal Large Language Models. *arXiv preprint arXiv:2509.07979*, 2025.

[50] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think. In *ICLR*, 2025.

[51] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *CVPR*, 2024.

[52] Zihao Yue, Liang Zhang, and Qin Jin. Less is More: Mitigating Multimodal Hallucination from an EOS Decision Perspective. In *ACL*, 2024.

A Additional Implementation Details

Training Details. Following the two-stage training recipe of LLaVA-1.5 [20], in the first stage, we train only the projector proj, a two-layer MLP, using 558k image-caption pairs, while keeping the language model $\mathcal{G}$ frozen. In the second stage, we jointly optimize $\mathcal{G}$ and proj for visual instruction tuning on the LLaVA-Instruct-665k dataset. The training settings (*e.g.*, optimizer, learning rate, and batch size) are kept identical to LLaVA-1.5. All experiments are conducted on AMD MI250x devices, each of which comprises 2 GPUs with 64GB of VRAM. The first training stage runs on 16 GPUs for up to 6 hours, depending on the size of the LLM. The second training stage runs on 32 GPUs, up to 14 hours. We find no noticeable difference in training time with the addition of $\mathcal{L}_{\text{HeRA}}$. The MKNN alignment scores can be efficiently computed offline. For instance, for a 7B LLM and DINOv2-L, it takes less than one hour on a single GPU.

Contrastive Learning Details. To generate supervision signals for each batch, we extract the `[CLS]` token representations from the teacher vision encoder $\mathcal{V}^t$. We then compute the pairwise dot products between these representations to form a similarity matrix (*i.e.*, a Gram matrix), which captures the visual neighborhood structure of the batch. For every sample, we identify its top-k nearest neighbors to construct multi-positive contrastive targets. This is achieved by assigning a uniform probability of $\frac{1}{k}$ to these k neighbors, and a probability of zero to all other samples.

For the student model, we extract the head-wise representations from the selected set of heads ($\mathcal{H}$) and apply average pooling across the embeddings corresponding to text tokens. In the first training stage, text tokens correspond to the caption of the input image, and thus we employ all of them. In the second training stage, text tokens represent multi-turn dialogs, and we pool exclusively over the tokens pertaining to the `<ASSISTANT>` turn. These are the same tokens contributing to the language modeling loss of Eq. 1.

The temperature parameter τ of Eq. 4 is learned in logarithmic scale, and it is initialized as 0.07.

LLM Details and Selected Heads. We collect in Table 7 the exact checkpoints of each LLM used in this work. All of them are publicly accessible on the Hugging Face Hub. We also report the specific attention heads of each LLM used to compute $\mathcal{L}_{\text{HeRA}}$, sorted from left to right by increasing value of the MKNN alignment score. We indicate with `LXHY` the index of the `Y`-th head in layer `X`.

Table 7: Checkpoint reference and list of selected heads for each LLM.

		Selected Heads ($\mathcal{H}$)				
LLM	**Hugging Face Page**	H1	H2	H3	H4	H5
Qwen2.5-3B [31]	`Qwen/Qwen2.5-3B-Instruct`	L3H4	L5H11	L5H14	L5H9	L5H15
Qwen3-4B [48]	`Qwen/Qwen3-4B-Instruct-2507`	L7H6	L13H12	L3H6	L3H5	L13H13
Vicuna-7B [4]	`lmsys/vicuna-7b-v1.5`	L0H14	L1H1	L0H1	L0H3	L0H30
Llama3-8B [9]	`meta-llama/Meta-Llama-3-8B-Instruct`	L0H31	L2H21	L0H29	L2H23	L2H25
Qwen2.5-7B [31]	`Qwen/Qwen2.5-7B-Instruct`	L14H5	L4H7	L14H4	L4H10	L4H13
Qwen3-8B [48]	`Qwen/Qwen3-8B`	L13H12	L7H6	L3H5	L13H13	L13H15
Vicuna-13B [4]	`lmsys/vicuna-13b-v1.5`	L0H38	L1H0	L0H11	L0H10	L0H26
Qwen2.5-14B [31]	`Qwen/Qwen2.5-14B-Instruct`	L13H27	L18H13	L0H9	L0H7	L3H8
Qwen3-14B [48]	`Qwen/Qwen3-14B`	L14H24	L9H36	L9H32	L1H37	L1H36

B Evaluation Benchmarks

Cambrian Evaluation Suite [36].[2] It comprises a comprehensive suite of benchmarks designed to evaluate diverse capabilities of MLLMs, including general perception, knowledge reasoning, OCR and chart understanding, and core visual abilities. Accordingly, the benchmarks are grouped into four categories: General, Knowledge, OCR, and Vision. In our experiments, we consider 18 benchmarks: GQA [13], POPE [18], MME [6], MMBench (MMB) [21], and SEED-Bench (SEED) [17] for the General category; ScienceQA (SQA) [25], MMMU [51], MathVista [24], and AI2D [15] for Knowledge; ChartQA [26], OCRBench (OCRB) [22], TextVQA [35], and VizWiz [12] for OCR; and MMVP [38], RealWorldQA (RWQA) [47], Blink [7], V* [46], and CVBench [36] for Vision. When reporting averages, we normalize the MME score by dividing it by 20 to ensure consistency with the scale of the other benchmarks.

[2] `https://github.com/cambrian-mllm/cambrian`

Table 8: Statistical robustness of the head selection criterion. We report mean, standard deviation, and [min, max] range over five training seeds, using Qwen2.5-3B as the LLM.

	General		Knowledge		OCR		Vision		All	
Model	Mean	Range	Mean	Range	Mean	Range	Mean	Range	Mean	Range
LLaVA	73.5±0.1	[73.4, 73.5]	46.6±0.1	[46.4, 46.7]	41.8±0.5	[41.0, 42.2]	50.9±1.0	[49.8, 52.4]	54.2±0.3	[53.8, 54.6]
+ HeRA Top-5	73.8±0.1	[73.7, 74.0]	46.5±0.3	[46.1, 46.8]	42.4±0.3	[42.1, 42.8]	50.4±1.0	[49.1, 51.4]	54.2±0.3	[53.9, 54.6]
+ HeRA Random-5	73.8±0.1	[73.7, 74.1]	46.9±0.2	[46.6, 47.0]	42.6±0.6	[41.8, 43.1]	50.8±0.9	[49.7, 52.1]	54.5±0.4	[54.0, 55.1]
+ HeRA Worst-5 **(Ours)**	**74.5**±0.1	[74.4, 74.6]	**47.4**±0.3	[46.9, 47.7]	**43.6**±0.2	[43.4, 43.8]	**52.2**±0.8	[51.2, 53.1]	**55.4**±0.2	[55.1, 55.7]

Table 9: VQA comparison between DINOv2 and SigLIP2 as the vision encoder for LLaVA.

		General	Knowledge	OCR	Vision					
LLM	**Vision Encoder**	Avg	Avg	Avg	RWQA	MMVP	Blink	V*	CVBench	Avg
Qwen2.5-3B	DINOv2	67.1	43.3	26.3	51.5	30.0	47.8	46.1	58.0	46.7
Qwen2.5-3B	SigLIP2	73.5	46.7	42.2	55.2	46.0	46.8	44.5	**60.2**	50.5
+ HeRA (Ours)	SigLIP2	**74.5**	**47.5**	**43.8**	**56.3**	**48.0**	**49.1**	**51.3**	59.6	**52.9**
Qwen3-4B	DINOv2	70.5	46.1	24.2	54.9	38.7	49.2	49.2	63.1	51.0
Qwen3-4B	SigLIP2	75.6	49.6	43.8	59.9	48.0	**55.1**	50.3	68.3	56.3
+ HeRA (Ours)	SigLIP2	**76.0**	**50.1**	**44.5**	**61.3**	**56.0**	53.7	**52.4**	**69.1**	**58.5**

Hallucination Datasets. We evaluate hallucinatory tendencies on three widely used benchmarks: AMBER [42], CHAIR-MSCOCO [52], and HallusionBench [11]. CHAIR-MSCOCO measures object- and sentence-level hallucination rates (*i.e.*, CHAIR_i and CHAIR_s) on model-generated descriptions for 500 images sampled from the MSCOCO [19] validation set. The AMBER generative task further introduces *cognition* (Cog), which quantifies the overlap between model- and human-hallucinated objects, and *coverage* (Cover), which measures object-level recall. Complementarily, the AMBER discriminative task captures a broader set of hallucination types, including attribute and relation hallucinations in addition to object existence, using a ground truth set of 1,004 manually annotated images. For CHAIR-MSCOCO and the AMBER generative task, we use a maximum generation length of 512 tokens with greedy decoding. For the AMBER discriminative task, we append the instruction *"Answer only with Yes or No. Use exactly one word. Do not use commas, periods, or symbols."* to each query to enforce binary (Yes/No) responses. We evaluate hallucination robustness on HallusionBench [11] using an exact-match protocol. Specifically, we force the model to output unambiguous responses by appending the following instruction to each query: *"Answer the question using a single word or phrase: Yes or No."* We report three standard metrics. *qAcc* (Question Pair Accuracy) measures group-level consistency. A prediction is counted as correct under qAcc only if the model answers all questions within the same group correctly. In addition, we report *Easy* accuracy, computed over unmodified questions, and *Hard* accuracy, computed over adversarially modified or misleading variants designed to induce hallucinations.

C Additional Experiments

C.1 Additional Ablation Studies and Results

Statistical Robustness of Head Selection Criterion. To verify that the advantage of the Worst-5 selection reported in Table 1 does not stem from training stochasticity, we repeat the LLaVA baseline and the Top-5, Random-5, and Worst-5 variants of HeRA across five training seeds (*i.e.*, 42, 1024, 1111, 4096, and 6144), using Qwen2.5-3B as the LLM and keeping all other settings unchanged. Note that, for Random-5, the selected heads vary across seeds. Table 8 reports the mean, standard deviation, and [min, max] range over the five runs for each category of the Cambrian suite. Worst-5 achieves the highest mean in every category, and both its minimum and maximum scores exceed those of all other configurations. On General, OCR, and the overall average, its lowest score even matches or surpasses the highest score of every alternative. We further run one-sided Welch's t-tests comparing Worst-5 against LLaVA, Top-5, and Random-5. Worst-5 significantly outperforms every alternative in all categories, with $p \leq 0.003$ on the overall average, $p \leq 0.008$ on General, Knowledge, and OCR, and $p \leq 0.027$ on Vision. These results confirm that the gains brought by aligning the least aligned heads cannot be explained by seed variability.

DINOv2 as Vision Encoder. In this work, we demonstrate the effectiveness of leveraging a teacher vision encoder, *e.g.*, DINOv2-L [30], as a source of supervision for topological representation

Table 10: Effects of $\mathcal{L}_{\text{HeRA}}$ when applied to the different training stages of LLaVA.

	$\mathcal{L}_{\text{HeRA}}$		General	Knowledge	OCR	Vision					
Model	St.1	St.2	Avg	Avg	Avg	RWQA	MMVP	Blink	V*	CVBench	Avg
Qwen2.5-3B	-	-	73.5	46.7	42.2	55.2	46.0	46.8	44.5	60.2	50.5
+ HeRA	✓	-	74.2	**47.5**	**44.2**	**57.4**	44.0	46.9	**51.3**	**60.7**	52.1
+ HeRA	-	✓	73.5	46.4	43.0	54.5	40.0	48.8	44.5	60.5	49.7
+ HeRA (Ours)	✓	✓	**74.5**	**47.5**	43.8	56.3	**48.0**	**49.1**	**51.3**	59.6	**52.9**
Qwen3-4B	-	-	75.6	49.6	43.8	59.9	48.0	**55.1**	50.3	68.3	56.3
+ HeRA	✓	-	75.6	49.7	43.4	60.7	51.3	53.3	46.1	66.2	55.5
+ HeRA	-	✓	75.7	50.0	**44.6**	59.6	54.7	52.7	49.2	67.8	56.8
+ HeRA (Ours)	✓	✓	**76.0**	**50.1**	44.5	**61.3**	**56.0**	53.7	**52.4**	**69.1**	**58.5**

Table 11: Representation geometry analysis on Qwen2.5-3B. We report anisotropy and effective rank globally and on the Top-5 and Worst-5 heads.

	Global		Top-5 Heads		Worst-5 Heads	
Model	**Anisotropy ↓**	**Effective Rank**	**Anisotropy ↓**	**Effective Rank**	**Anisotropy ↓**	**Effective Rank**
Qwen2.5-3B	0.61	75.87	0.30	93.50	0.85	22.26
+ LLaVA	0.64	68.38	0.34	79.39	0.83	24.05
+ HeRA (Ours)	0.64	70.23	0.33	81.37	**0.71**	**55.37**

alignment. It is natural to ask what if we do not perform representation alignment at all, by directly plugging in DINOv2-L as the vision encoder of an MLLM. Table 9 answers this question by comparing DINOv2-L vs SigLIP2 [39], and clearly demonstrates that DINOv2-L is ineffective on its own as a vision encoder for MLLMs. For fair comparison, we feed DINOv2-L with the same image resolution of 384×384 pixels as SigLIP2. Despite that, DINOv2-L suffers from severe deficits, especially on OCR tasks. These results agree with prior works [5, 36] showing that unsupervised visual encoders alone fall short against language-supervised encoders on VQA benchmarks. On the other hand, as testified by Fig.4, SigLIP2 is largely ineffective as a representation teacher, despite being an excellent MLLM vision encoder. These results suggest that an effective teacher should provide a visual neighborhood structure complementary to the one already available to the MLLM through its own vision encoder.

HeRA in Different Training Stages. We seamlessly apply the $\mathcal{L}_{\text{HeRA}}$ on both training stages of LLaVA. However, there are neat differences between them that are worth discussing. For instance, in the first training stage (*i.e.*, St.1), the MLLM is fed with images and their related captions, which represent aligned image-text pairs, *i.e.*, the same concept is expressed in two different modalities. This appears to be a suitable stage for representation alignment, as the student MLLM and the teacher vision encoder process the same underlying concepts. Conversely, image-text pairs in the second training stage (*i.e.*, St.2) are not aligned the same way: the text corresponds to a multi-turn dialog between a user and the assistant, which focuses on the image, but does not exactly mimic the visual content as an image caption. With that in mind, if one had to select a single training stage for $\mathcal{L}_{\text{HeRA}}$, we would expect a larger impact on St.1 rather than St.2. However, according to Table 10, that is not always the case: with Qwen2.5-3B, $\mathcal{L}_{\text{HeRA}}$ helps more when applied during St.1, while the opposite holds with Qwen3-4B. Ultimately, our original proposal of regularizing both training stages with $\mathcal{L}_{\text{HeRA}}$ works best on both models.

Effects of HeRA on Geometrical Space. In Table 11 we analyze how HeRA affects some geometrical properties of the internal heads of the MLLM. Anisotropy measures the degree to which representations are concentrated along common directions in the embedding space, thus providing interesting insights on the degree of collapse of a representation space. Effective rank, instead, measures how broadly the representation spectrum is distributed across different directions of the embedding space: a low effective rank indicates that few directions dominate the representation, while high rank suggests that a large number of dimensions contribute meaningfully to it. Our analysis highlights that HeRA induces positive changes in the representation space of the Worst-5 heads: it reduces anisotropy and more than doubles the effective rank with respect to LLaVA. Table 11 also shows no significant changes in either anisotropy or effective rank for the Top-5 heads or at the global level, confirming that the localized intervention of HeRA has limited impact on the rest of the model.

Extended Results on All Benchmarks. As we aim to improve MLLMs, we focus on strengthening their visual perception, which is particularly stressed on vision-centric benchmarks. Consequently, in

Table 12: Detailed VQA results of HeRA applied to the LLaVA training recipe on different LLMs.

	General						Knowledge					OCR					Vision					
	Avg	GQA	POPE	MME	MMB	SEED	**Avg**	SQA	MMMU	MathV	AI2D	**Avg**	ChartQA	OCRB	TextVQA	VizWiz	**Avg**	MMVP	RWQA	Blink	V*	CVBench
Qwen2.5-3B	73.5	63.8	87.7	1484.0	70.9	71.2	46.7	75.4	40.9	7.7	62.9	42.2	17.3	37.1	60.6	53.8	50.5	46.0	55.2	46.8	44.5	**60.2**
+ HeRA (Ours)	**74.5**	**64.1**	**87.9**	**1525.5**	**72.0**	**72.0**	**47.5**	**75.5**	**41.3**	**7.9**	**65.1**	**43.8**	**20.2**	**39.1**	**61.7**	**54.3**	**52.9**	**48.0**	**56.3**	**49.1**	**51.3**	59.6
Qwen3-4B	75.6	64.6	87.3	1528.4	75.4	74.1	49.6	77.2	44.6	**8.4**	68.4	43.8	23.3	**42.5**	**65.6**	44.0	56.3	48.0	59.9	**55.1**	50.3	68.3
+ HeRA (Ours)	**76.0**	**64.9**	**87.9**	**1544.6**	**75.6**	**74.3**	**50.1**	**78.8**	**44.7**	8.1	**68.9**	**44.5**	**24.8**	40.5	**65.6**	**47.2**	**58.5**	**56.0**	**61.3**	53.7	**52.4**	**69.1**
Vicuna-7B	**72.2**	64.9	87.5	**1469.0**	**64.6**	70.8	44.3	70.3	**36.3**	11.1	59.4	**45.7**	20.0	**40.6**	64.3	**58.0**	49.7	38.7	56.5	46.8	44.5	**62.1**
+ HeRA (Ours)	72.1	**65.1**	**87.8**	1453.4	64.0	**70.9**	**44.5**	**70.9**	35.0	**12.4**	**59.5**	**45.7**	**20.6**	40.3	**64.4**	57.6	**52.0**	**42.7**	**57.8**	**47.9**	**49.7**	61.9
Llama3-8B	73.3	64.8	87.5	**1506.0**	67.4	71.5	45.0	72.5	37.9	7.2	62.5	43.0	18.8	**40.1**	63.8	49.5	53.8	46.0	60.1	49.2	44.0	**69.5**
+ HeRA (Ours)	**74.6**	**65.6**	**87.6**	1503.3	**71.7**	**72.7**	**46.3**	**75.9**	**38.0**	**8.4**	**63.1**	**44.7**	**19.9**	39.9	**64.6**	**54.2**	**55.1**	**46.7**	**60.4**	**50.2**	**51.8**	66.4
Qwen2.5-7B	76.2	64.9	88.2	**1582.2**	75.3	73.7	50.2	**77.9**	45.1	**8.8**	68.8	47.9	24.2	39.9	64.9	62.5	56.7	51.3	59.6	**51.7**	**50.8**	70.2
+ HeRA (Ours)	**76.5**	**65.3**	**88.3**	1574.8	**76.0**	**74.1**	**50.5**	**77.9**	**46.9**	7.6	**69.6**	**48.6**	**23.9**	**40.7**	**65.6**	**64.1**	**57.4**	**54.0**	**61.3**	50.2	50.3	**71.3**
Qwen3-8B	74.7	64.8	86.5	1472.8	75.4	73.1	49.7	77.5	**47.0**	7.3	67.2	43.5	20.2	37.4	64.3	51.9	55.9	49.3	59.2	52.2	**50.8**	67.8
+ HeRA (Ours)	**76.9**	**66.1**	**87.6**	**1562.3**	**77.6**	**74.9**	**51.1**	**78.8**	46.2	**8.6**	**70.8**	**47.6**	**27.6**	**41.4**	**67.9**	**53.6**	**59.5**	**58.0**	**60.1**	**55.5**	50.3	**73.5**
Vicuna-13B	73.4	65.5	87.7	1491.9	67.1	**72.2**	45.5	72.0	**38.0**	9.9	**62.0**	**47.7**	**22.2**	41.6	**67.2**	**59.5**	52.7	**44.7**	**58.7**	50.6	46.1	63.7
+ HeRA (Ours)	**73.6**	**65.8**	**88.0**	**1504.7**	**67.2**	71.9	**45.7**	**72.2**	36.6	**13.2**	60.8	47.6	22.0	**41.8**	67.1	59.4	**53.9**	44.0	57.3	**52.3**	**49.2**	**66.5**
Qwen2.5-14B	75.6	64.2	**88.0**	1548.8	75.5	72.7	50.7	76.4	48.2	8.3	69.7	44.8	19.6	33.9	63.7	61.9	54.9	47.3	**60.1**	52.7	46.6	67.9
+ HeRA (Ours)	**77.4**	**66.1**	**88.0**	**1600.9**	**77.5**	**75.5**	**52.8**	**78.6**	**49.1**	**9.6**	**74.1**	**49.3**	**27.4**	**39.0**	**67.9**	**62.9**	**58.3**	**52.0**	**60.1**	**55.2**	**52.4**	**71.6**
Qwen3-14B	77.4	65.6	87.6	1609.5	**78.5**	74.6	**52.8**	**79.6**	**49.6**	**10.2**	71.9	46.1	26.8	41.1	69.7	47.0	58.2	57.3	60.3	**52.6**	50.3	**70.8**
+ HeRA (Ours)	**77.7**	**66.3**	**88.0**	**1632.5**	78.0	**74.9**	52.6	79.4	49.1	9.2	**72.6**	**47.8**	**28.1**	**41.8**	**69.9**	**51.4**	**58.9**	**58.0**	**62.5**	52.0	**51.8**	70.2

Table 13: Detailed VQA results of different representation alignment strategies for MLLMs.

	General						Knowledge					OCR					Vision					
Alignment	**Avg**	GQA	POPE	MME	MMB	SEED	**Avg**	SQA	MMMU	MathV	AI2D	**Avg**	ChartQA	OCRB	TextVQA	VizWiz	**Avg**	RWQA	MMVP	Blink	V*	CVBench
-	74.7	64.8	86.5	1472.7	75.4	73.1	49.7	77.5	**47.0**	7.3	67.2	43.5	20.2	37.4	64.3	51.9	55.9	59.2	49.3	52.2	50.8	67.8
ROSS [41]	74.6	64.5	87.2	1472.5	74.9	72.7	49.4	76.4	46.3	7.4	67.4	44.0	17.8	36.8	64.8	**56.8**	56.2	59.8	50.3	52.2	49.2	69.7
VIRAL [49]	73.8	64.2	87.3	1396.6	74.9	72.6	49.3	76.4	45.7	8.0	67.1	43.6	19.3	36.7	64.2	54.2	54.2	57.6	46.7	51.3	46.1	69.3
JARVIS [3]	76.8	64.6	**88.2**	**1605.1**	76.7	74.2	49.9	77.0	45.4	8.4	68.6	46.2	27.1	40.5	66.0	51.2	58.7	59.2	54.7	54.2	**55.5**	69.9
CMAR [8]	76.4	65.6	87.4	1556.6	76.5	74.7	51.0	78.3	45.9	**8.6**	**71.0**	46.0	23.3	**41.6**	67.6	51.6	56.9	58.8	50.0	52.1	52.9	70.9
HeRA (Ours)	**76.9**	**66.1**	87.6	1562.3	**77.6**	**74.9**	**51.1**	**78.8**	46.2	**8.6**	70.8	**47.6**	**27.6**	41.4	**67.9**	53.6	**59.5**	**60.1**	**58.0**	**55.5**	50.3	**73.5**

the main paper, we reported detailed scores on specific vision-centric VQA datasets, such as RealWorldQA, MMVP, Blink, V*, and CVBench, leaving the average score on the General, Knowledge, and OCR categories. Here, we report the full results over the 18 VQA benchmarks considered in our study. Specifically, we refer to Table 12 for the results of different LLM families, and to Table 13 for a detailed comparison between representation alignment methods on the Qwen3-8B LLM.

C.2 Additional Analyses

Additional MKNN Head-Wise Analysis. In Fig. 5, we provide a detailed comparison of the MKNN alignment scores for the Worst-5 (*upper* half) and Top-5 (*bottom* half) attention heads across different training strategies. These specific heads are identified by computing their MKNN alignment scores on the base Qwen2.5-3B LLM prior to any multimodal training. Interestingly, we observe that the relative alignment of these heads is largely preserved after standard multimodal training (*i.e.*, LLaVA): heads that are naturally highly aligned in the base LLM remain highly aligned, whereas poorly aligned heads stay poorly aligned.

When we apply HeRA to the Top-5 heads, their alignment scores further increase, but this intervention has absolutely no impact on the Worst-5 heads. As shown in Tab. 1 (*seventh* row), this translates into suboptimal downstream performance. Conversely, our proposed strategy, that is HeRA applied to the Worst-5 heads, greatly increases the alignment of the targeted components. Crucially, this massive boost does not sacrifice the integrity of the Top-5 heads, which record MKNN alignment scores remarkably similar to the LLaVA baseline.

Finally, we observe that enforcing representation alignment at the feature level, specifically, via cosine similarity maximization between the MLLM visual features and the teacher vision encoder, is ineffective at modifying the local topological structure (nor at improving performance, see Table 1, *third* row). As depicted in the plot, this approach has a very small effect on the MKNN alignment scores of the targeted Worst-5 heads, further highlighting the unique contribution of our topology-aware contrastive objective.



Figure 5: Effect of the multimodal training of LLaVA and representation alignment methods on the Worst-5 and Top-5 heads. Worst-5 and Top-5 heads are selected by the lowest and highest MKNN alignment score with DINOv2-L, computed on the Qwen2.5-3B LLM before multimodal training.

MKNN Alignment With Larger Qwen2.5 Models. In Fig. 6 and Fig. 7, we extend the MKNN alignment analysis of Fig. 3 to larger LLMs within the same family, specifically evaluating Qwen2.5-7B and Qwen2.5-14B against the DINOv2-L teacher. The left parts confirm that our previous observations persist at larger scales: representations of specific individual attention heads consistently exhibit a much higher natural alignment with the visual domain than those obtained out of any layer in the same model. Furthermore, the right parts highlight the atomic nature of our intervention. Applying HeRA to the Worst-5 heads successfully drives a massive boost in their cross-modal alignment, without disrupting the structural alignment of the Top-5 heads that are already naturally aligned in the base LLM.

MKNN Analysis with Different Representation Alignment Methods. In Fig. 8, we compare the MKNN alignment scores of HeRA against existing representation alignment strategies. For this analysis, all methods are trained using Qwen3-8B as the LLM and SigLIP2 as the vision encoder, with the MKNN alignment metric computed with respect to the DINOv2-L teacher. We remind to Table 4 for a quantitative comparison on VQA benchmarks.

Consistent with our previous findings, all methods increase the alignment of the Top-5 heads. However, this is largely a natural consequence of the multimodal training process itself, as most methods achieve scores that closely mirror the unregularized LLaVA baseline. The only method that registers a more significant, distinct impact on the Top-5 heads is CMAR. CMAR shares a conceptual similarity with HeRA in that it enforces cross-modal topological alignment rather than strict feature-level visual matching. However, a key difference lies in their scope: CMAR relies on the CKA metric to match *global* pairwise relationships across all samples within a training batch, whereas HeRA strictly targets the consistency of *local* neighborhoods.

Regarding the analysis on the Worst-5 heads, both CMAR and feature-matching methods fail to induce any meaningful structural changes. Conversely, HeRA is the *only* method capable of significantly increasing the cross-modal alignment of these initially poorly aligned heads.

C.3 Qualitative Results

In Fig. 10, we provide a qualitative comparison between the LLaVA [20] baseline, ROSS [41], and HeRA using Qwen3-8B and SigLIP2. The representative samples demonstrate that HeRA consistently delivers more accurate and better-grounded answers across all evaluated categories (General, Knowledge, OCR, and Vision-Centric), effectively correcting various perceptual errors made by the baselines.

Despite these clear improvements, in Fig. 11, we report a few failure cases where our model still struggles. Specifically, HeRA can occasionally misinterpret fine-grained visual details, such

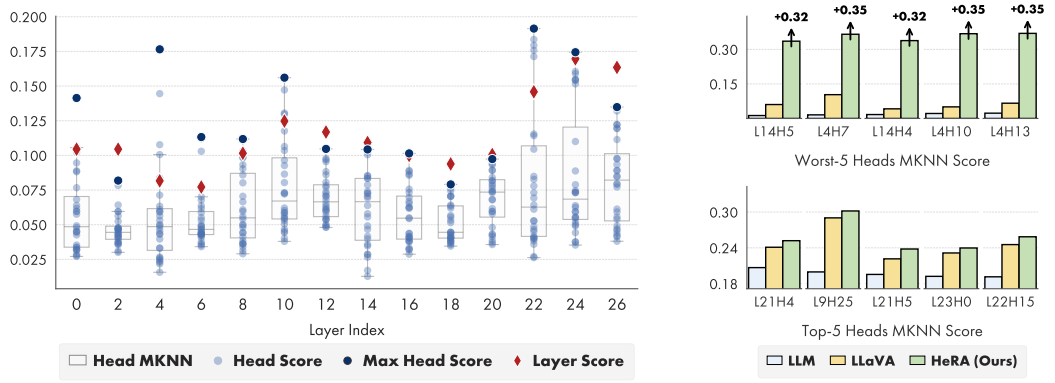

Figure 6: **Left:** Alignment with DINOv2-L, measured with the MKNN metric on each layer and attention head of Qwen2.5-7B. **Right:** MKNN scores of the Worst-5 and Top-5 heads, computed on *(i)* the base LLM; *(ii)* after the LLaVA multimodal training; and *(iii)* after the addition of HeRA.

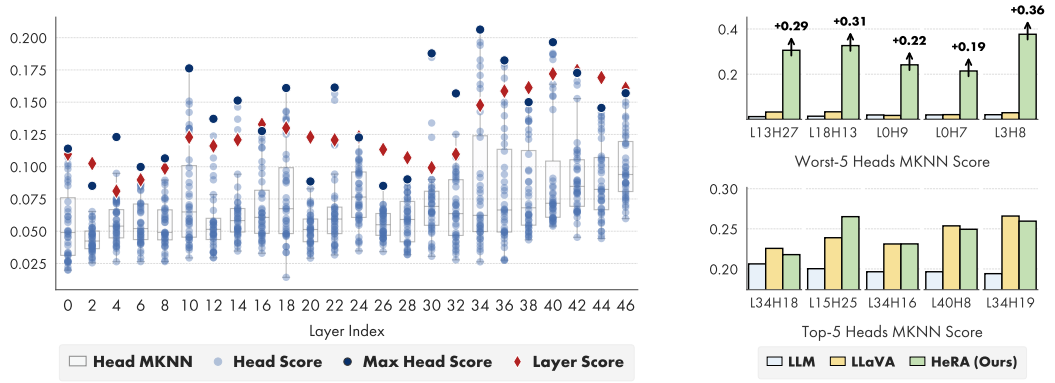

Figure 7: **Left:** Alignment with DINOv2-L, measured with the MKNN metric on each layer and attention head of Qwen2.5-14B. **Right:** MKNN scores of the Worst-5 and Top-5 heads, computed on *(i)* the base LLM; *(ii)* after the LLaVA multimodal training; and *(iii)* after the addition of HeRA.

as accurately counting multiple small instances, identifying ambiguous materials and shapes, or inferring precise spatial relationships in cluttered scenes.

D Limitations and Societal Impacts

We are aware that the landscape of MLLMs has rapidly evolved beyond LLaVA with the introduction of frontier proprietary models (as acknowledged at the end of Sec. 2). However, we adopted the LLaVA pipeline because it remains computationally tractable, allowing for the extensive ablation studies and rigorous evaluations presented in this work.

To bridge this gap and explore the potential of our method on modern architectures, we conducted a pioneering study applying HeRA directly to Qwen3-VL-4B [1], a state-of-the-art multimodal LLM. We fine-tuned the model using an 83k-sample Cambrian split derived from the FineVision [45] dataset. As shown in Fig. 9, HeRA records promising results, particularly on demanding vision-centric tasks. Crucially, the improvements on visual benchmarks are substantially higher with HeRA than with standard fine-tuning alone; moreover, simple fine-tuning actually registers a performance regression in the General category, which does not manifest in HeRA.

Beyond this architectural constraint, we do not foresee direct negative societal impacts arising specifically from our representation alignment technique. Rather, by improving visual grounding and mitigating object hallucinations, HeRA contributes to the development of more reliable and factual vision-language systems.

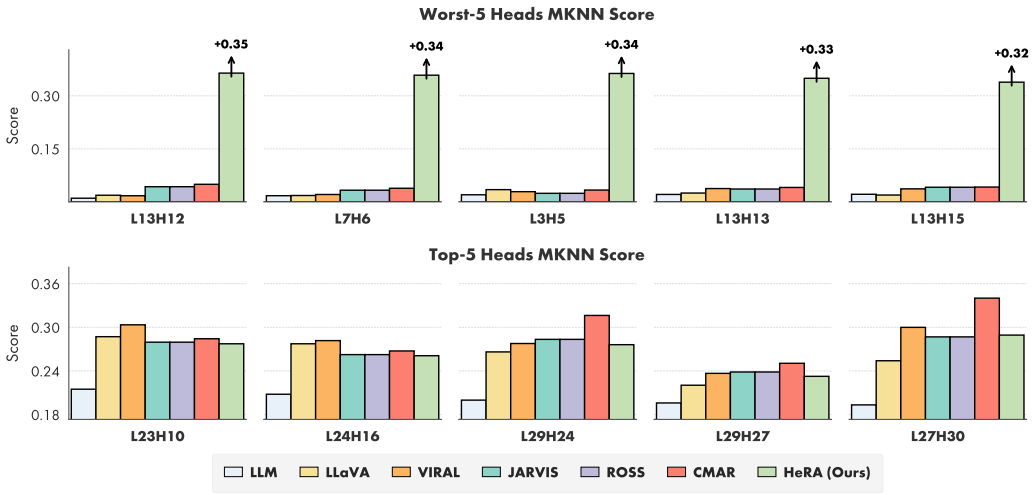

Figure 8: Comparison of MKNN scores of the Worst-5 and Top-5 heads after the second training stage performed with HeRA and our competitors (*i.e.* LLaVA, VIRAL, JARVIS, ROSS, CMAR).

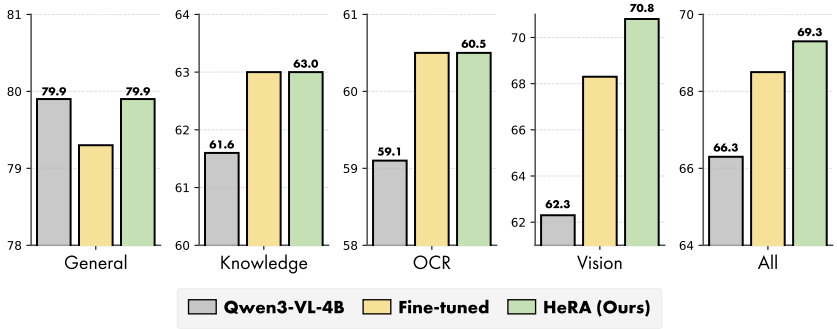

Figure 9: VQA results on Qwen3-VL-4B after fine-tuning, with and without the HeRA objective.

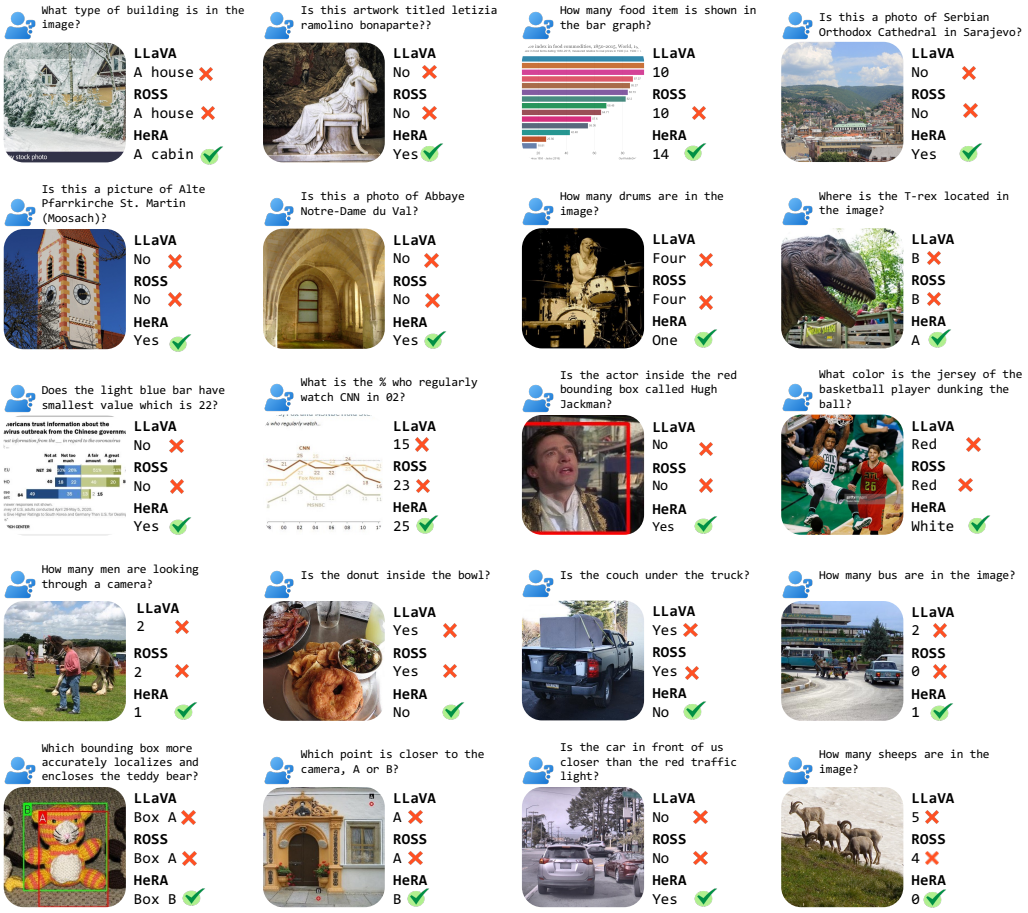

Figure 10: Qualitative comparison of LLaVA [20], ROSS [41], and HeRA using Qwen3-8B and SigLIP2. We present representative samples across all Cambrian categories: General, Knowledge, OCR, and Vision-Centric.

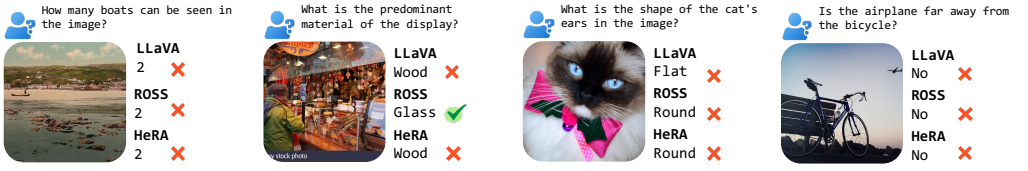

Figure 11: Failure cases of HeRA on VQA tasks.

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: The claim of improving MLLMs via representation alignment is empirically validated through more 18 VQA benchmarks spanning General, Knowledge, OCR, and Vision-Centric categories, as well as hallucination mitigation.

   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: The limitations are discussed in Appenxid D, as well as in the end of Section 2.

   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: The paper does not provide mathematical proofs, although the main findings can be empirically interpreted under the theory of the Platonic Representation Hypothesis.

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 provide specific sections about the implementation details: Section 4 and Appendix A.

   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: [Yes]

Justification: We build upon publicly accessible datasets and pre-trained models, and we disclose all the details required to reproduce our results. Moreover, our codebase will be open-sourced.

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 provide all the required training details in Appendix A.

   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: Due to computational constraints, we cannot afford to repeat experiments with multiple seeds to measure error mean and standard deviation. However, we tested our method on many models pertaining to different architecture and scales to ensure that the original improvements were not spurious.

   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: We detailed the computational resources employed in this work in Appendix A. Moreover, we will release our exact training scripts.

   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: This works complies with 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: [Yes]

    Justification: The potential societal impacts of this work are discussed alongside the limitations in Appendix D.

    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: The release of our codebase does not present direct risks of misuse. Possible indirect risks derived from the usage of LLMs employed in this work are already disclosed in the official technical reports.

    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: The original owners of the data used in the paper are properly credited, and our codebase will be released with a license compatible with those of any asset employed in this work.

    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: [No]

    Justification: At the time of the submission, there is no new asset to be released. Our trained models will be publicly released further in the future.

    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: The paper does not involve crowdsourcing nor research with human subjects.

    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: The paper does not involve crowdsourcing nor research with human subjects.

    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: [Yes]

    Justification: This work studies how to develop more advanced MLLMs. Consequently, LLMs are at the core of our study, in that they constitute the backbone of any MLLM.

    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.