

Beyond Semantics: Rediscovering Spatial Awareness in Vision and Language Models

Jianing Qi^{1*}, Jiawei Liu^{1*}, Hao Tang¹, Zhigang Zhu^{1,2}

¹CUNY Graduate Center, ²The City College of New York
{jq@gc, Jiawei.Liu85@gc, htang@gc, zzhu@ccny}.cuny.edu

Abstract

Vision and Language Models (VLMs) recognize objects well but often remain weak on spatial reasoning. In VLMs with a projector, spatial information can reach the decoder either through the content of vision tokens or through decoder positional encoding (RoPE), yet random permutation of vision tokens often causes only minor performance change. To measure positional use more directly, we introduce the Position Sensitivity Index (PSI), a RoPE sensitivity probe, and the 2D Synthetic Spatial (2DS) benchmark, a controlled benchmark without semantic shortcuts. We trace the weak use of RoPE to a large norm imbalance between vision and text at the projector interface, and find that this imbalance suppresses positional effects in the decoder. A broader survey of twelve models shows that this imbalance is common across VLMs with a projector. Our targeted interventions then reveal a dissociation: balancing norms increases positional reliance, while enriching the content of vision tokens yields even higher spatial accuracy, with lower PSI. These results show that RoPE in VLMs with a projector is suppressed rather than totally absent, and that positional reliance and spatial accuracy can be improved through independent mechanisms. Website is at: user074.github.io/respatialaware/.

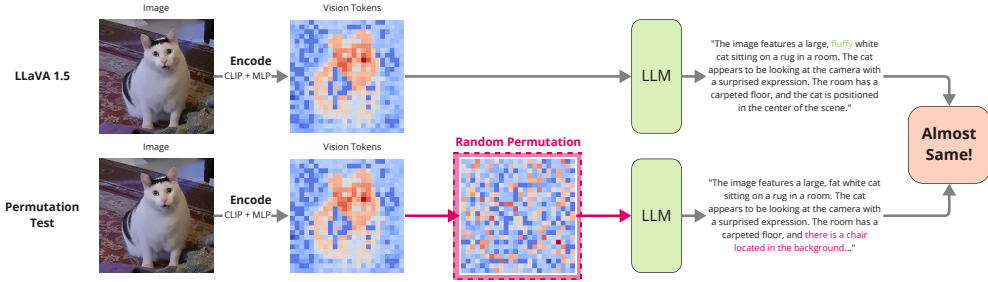


Figure 1: **Permutation Test:** Original (top) vs. randomly permuted vision tokens (bottom). Despite losing spatial ordering, the LLM accurately responds to the prompt “Describe the image,” demonstrating strong robustness and a notable “bag of tokens” tendency, losing spatial relationships. Token embeddings are visualized using cosine similarity relative to a reference token.

1 Introduction

Humans process visual content through two pathways: the ventral stream for object recognition and the dorsal stream for spatial reasoning. Damage to the parietal lobe can preserve recognition while impairing the ability to reason about where objects are and how they relate to each other (Goodale & Milner, 1992).

VLMs with a projector show a similar asymmetry. They perform well on object focused tasks such as recognition and captioning (Goyal et al., 2017; Antol et al., 2015; Gurari et al.,

*Equal contribution.

2018), yet remain weak on basic spatial queries such as left/right judgments (Kamath et al., 2023; Tong et al., 2024b). This is surprising because such models have access to spatial information through two distinct sources: The vision encoder can place spatial structure inside the contents of vision tokens, giving a semantic form of spatial information. The decoder can also use positional encoding (RoPE), which injects order information over the sequence of vision tokens and lets the LLM reason over spatial arrangement.

A simple experiment we have designed reveals that one of these sources is only weakly exploited. When we randomly permute vision tokens after the projector but before the decoder, predictions on multiple standard benchmarks barely change (Figure 1). Each token now receives a different positional encoding, yet the model’s answers remain almost the same. This suggests an important distinction: positional information may be structurally *available* to the decoder without being strongly *used* by the model’s final predictions. What determines this gap?

To study this reliably, we need measurements that separate decoder positional use from vision token content. We therefore introduce two tools and a controlled benchmark. First, the **Position Sensitivity Index (PSI)** quantifies how much a model relies on the order of vision tokens by comparing accuracy before and after a permutation. Second, the **RoPE sensitivity probe** measures how strongly each decoder layer responds to positional encoding on vision tokens, indicating whether this signal is available within the decoder. Finally, the **2D Synthetic Spatial (2DS)** benchmark uses synthetic scenes with controlled layouts where every question requires spatial reasoning, eliminating semantic shortcuts.

Using these tools on LLaVA 1.5 7B, we trace the model’s weak positional sensitivity to a norm imbalance between vision and text. After the projection layer, vision token norms are substantially larger than text token norms. In LLaMA decoders, which apply RMSNorm before each sublayer, these large vision activations persist in the residual stream and suppress the effect of positional encoding on attention (Section 4). A survey of twelve models shows that large interface norm gaps are common across VLMs with a projector (Section 4.1). The key distinction is between *availability* and *usage*.

We then test two targeted interventions. **+Normalize** applies RMS scaling to match the scale of vision embeddings after the projector, directly reducing the norm gap. **+Normalize+Multilayer** combines the same normalization with features from intermediate vision encoder layers, enriching the spatial content of vision tokens. The results reveal a key dissociation. Reducing the norm gap increases positional reliance, but enriching vision token content achieves even higher spatial accuracy through the other spatial source, with less dependence on token order. This indicates that positional sensitivity and spatial accuracy are related but not identical, and can improve through different mechanisms. We validate on both synthetic benchmarks and spatial benchmarks on real images. Our aim is understanding and measurement, rather than chasing the best benchmark performance.

Contributions. (1) *Norm imbalance between vision and text:* We identify a large norm gap between vision and text tokens at the projection interface, and show that it is common across VLMs with a projector. (2) *Diagnostic tools:* The Position Sensitivity Index (PSI), the RoPE sensitivity probe, and the 2DS benchmark provide complementary measurements of positional use. (3) *Targeted interventions:* We show that balancing norms increases positional use, while adding intermediate vision features improves spatial accuracy through the other spatial source.

2 Related Work

Vision and Language Models. Recent Vision and Language Models (VLMs) combine pretrained vision encoders and large language models (LLMs) for image understanding tasks such as captioning and visual question answering (Liu et al., 2023; 2024; Li et al., 2023a; Alayrac et al., 2022; Zhu et al., 2024; Shao et al., 2023; Ding et al., 2022; Hu et al., 2022). A common design uses a pretrained visual encoder, typically CLIP (Radford et al., 2021), with a projection module that maps vision embeddings into a pretrained LLM such as Vicuna (Chiang et al., 2023; Liu et al., 2023; Zhu et al., 2024). These models perform well on

descriptive tasks but often struggle with spatial reasoning (Tong et al., 2024b; Yuksekgonul et al., 2023). Most remedies augment training data with spatial annotations (Tong et al., 2024a; Chen et al., 2024; Cheng et al., 2024; Trabelsi et al., 2025). Our approach instead works at the representation level, addressing a structural cause of spatial weakness without additional spatial labels or data.

VLM Interpretability. Interpretability research aims to explain the internal mechanisms behind VLM successes and failures. Neo et al. (2025) show that vision tokens in LLaVA (Liu et al., 2023) carry strong semantic properties, and they are highly redundant (Chai et al., 2025). Multiple studies have documented consistent VLM failures on simple spatial relations such as “above,” “below,” “left,” and “right” (Yuksekgonul et al., 2023; Kamath et al., 2023; Tong et al., 2024b; Chen et al., 2025). These failures are well characterized at the behavioral level, but the internal cause remains unclear. In pure vision models, Kutscher et al. (2025) show that transformers are sensitive to patch order and that learning a better order improves accuracy. This is complementary to our work: we also hypothesize that token order matters, but we find that it is underutilized in VLMs with a projector. Our work investigates *why* spatial information is suppressed inside the model, identifying embedding norm imbalance as a concrete mechanism and offering targeted interventions.

Modality Imbalance in VLMs. Recent work has identified representation-level imbalances in VLMs. Li et al. (2025a) observe a norm discrepancy between vision and text tokens at the projector interface, attributing performance degradation to representational inertia in Pre-Norm architectures. Li et al. (2025b) show that visual information becomes progressively underutilized in deeper decoder layers. At the task level, IsoBench (Fu et al., 2024) reveals systematic performance gaps when the same problem is presented visually versus textually. Our work identifies a specific consequence of the norm gap that these studies do not address: the imbalance suppresses positional encoding (RoPE) effectiveness at the residual stream level, contributing to spatial reasoning failures. We further show that positional sensitivity and spatial accuracy are dissociable, suggesting the norm gap affects distinct information routes differently.

High Norm Tokens and Attention. Large norms have also been studied in the LLM literature. A few tokens can develop massive activations and act as attention sinks, attracting a disproportionate share of attention (Sun et al., 2024; de Llano et al., 2025), and a similar sink behavior appears in VLMs (Zhang et al., 2026). In vision encoders, Darcet et al. (2024) show that a small set of high norm outlier tokens carries little positional information but yet acts as registers. These effects come from a few outlier tokens and concern where attention is placed. The effect that we study in this work comes from the norm gap between two groups: vision tokens and text tokens. It acts after the attention output is computed, and its consequence is that the decoder under-utilizes token orders (Appendix B). These two phenomena can be present at the same time.

3 Weak Positional Sensitivity in VLMs with a Projector

Dataset	Type	Original	Permutation	Difference
VQAv2	General VQA	78.2	77.35	-0.85
POPE	General VQA	87.3	87.10	-0.2
GQA	Spatial VQA	61.36	58.62	-2.74
CV-Bench 2D	Spatial VQA	56.59	56.26	-0.33

Table 1: Impact of random vision token permutation on accuracy (%).

encoding (RoPE) over the sequence of vision tokens. The first source comes from content. The second comes from token order. A direct way to test the second source is to permute vision tokens after the projector. This changes the positional encoding each token receives while leaving token content intact. If predictions barely change, the model is not using positional encoding strongly.

To understand the positional sensitivity, we use LLaVA 1.5 7B (Liu et al., 2023; 2024) to randomly permute vision token embeddings after processing by the vision encoder and the MLP projection layers, but right before they are fed into the LLM (Figure 1). It effectively

Prior work has documented that VLMs struggle with spatial reasoning, but the internal reason remains unclear. In VLMs with a projector, spatial information can reach the decoder through two sources: spatial structure already encoded in vision token content, and the decoder’s positional

changes the position embedding for each token. The performance is then compared against the original unpermuted baseline. We evaluate performance on benchmarks spanning general VQA (VQAv2 (Goyal et al., 2017), POPE (Li et al., 2023b)), and spatial VQA (GQA (Hudson & Manning, 2019), CV-Bench 2D (Tong et al., 2024a)). As shown in Table 1, random token permutation leads only to minor performance drops (ranging from 0.2% to 2.74%). This surprisingly small decline reveals that current VLMs evaluated with the corresponding benchmarks are largely insensitive to positional information.

This result admits two interpretations. First, the vision encoder may already encode sufficient spatial information within token features, making token order redundant for these tasks. Second, these benchmarks may be solvable from vision token content alone, so a model can score well without relying on token order. To distinguish between these explanations, we need (a) a metric that quantifies how much a model relies on token order, and (b) a benchmark where token order is necessary and semantic shortcuts are removed. We introduce both below.

3.1 Position Sensitivity Index (PSI)

We formalize the permutation test as a general metric. The **Position Sensitivity Index** is defined as:

$$\text{PSI} = \frac{\text{Acc}(\text{original order}) - \text{Acc}(\text{permuted order})}{\text{Acc}(\text{original order})}. \quad (1)$$

PSI is easy to interpret: $\text{PSI} = 0$ means complete order invariance, and larger values indicate greater reliance on token order. We report PSI at both the model level and the dataset level, since positional sensitivity is a joint property of both. Fixing one allows comparison across the other.

Interpretation caveat. Low PSI can arise from genuine order invariance, where the model does not need token order for the task. It can also arise from global model degradation, where both original and permuted orders produce poor accuracy. PSI should therefore be interpreted jointly with accuracy on the original order.

What the permutation controls for. PSI compares the same model before and after permuting the vision embeddings across their positions. The permutation leaves the set of token norms unchanged, so the norm gap is the same in both runs. The only thing that changes is which content sits at which position, so the difference between the runs isolates the effect of reordering. PSI therefore measures reliance on token order, not overall attention quality.

3.2 A 2D Synthetic Spatial Benchmark (2DS)

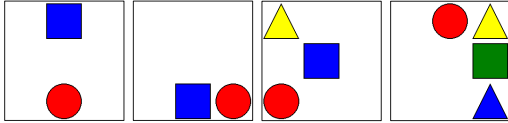


Figure 2: **2DS** examples. Left: layouts with two objects; right: layouts with three or four objects. Semantics are minimal so that answering requires spatial relations.

The permutation test on standard benchmarks is ambiguous: low sensitivity could mean the model ignores RoPE, or it could mean the benchmark can be solved through semantic shortcuts. We construct a **2D Synthetic Spatial (2DS)** Benchmark to isolate benchmark weakness from model weakness. It is a controlled diagnostic, unlike standard benchmarks, which often admit shortcuts such as object co-occurrence and language priors, and can show reasonably high accuracy even when token order matters little.

Recent study (Asadi et al., 2026) shows that VLMs can even score surprisingly high on multimodal benchmarks without any image input, which shows how easily they can be solved from language priors alone. In 2DS, scenes contain colored shapes placed at controlled locations. Queries require either absolute relations (e.g., “bottom”) or relative relations (e.g., “left of”). Color and shape serve only as identifiers for referring to objects, not as cues for the answer. We generate several categories by number of objects (2–6), randomize positions within each, and enumerate questions across $\{\text{color, shape, color+shape}\} \times \{\text{absolute, relative}\}$, leaving no room for language priors. Full construction details are in Appendix C.

3.3 Separating Benchmark Weakness from Model Weakness

Table 2 compares PSI for LLaVA 1.5 across five benchmarks. General VQA benchmarks (VQAv2, POPE) show near-zero PSI because spatial reasoning is rarely the bottleneck. Spatial VQA benchmarks (GQA, CV-Bench 2D) show slightly higher but still low PSI, likely because models can exploit semantic shortcuts to bypass token order.

Only 2DS, which removes these shortcuts by construction, reveals substantial positional sensitivity (41.07%). This pattern reflects the two sources of spatial information. On standard benchmarks, vision token content is usually enough, so permuting token order changes accuracy only slightly. On 2DS, content alone cannot resolve the layout, so the model must use token order. The size of the permutation drop measures how much a benchmark depends on token order rather than on vision content. We use 2DS as the main controlled benchmark when asking whether later interventions actually improve positional use. Detailed permutation results on 2DS are in Appendix C.2.

Dataset	Type	PSI (%)	Original Acc. (%)
VQAv2	General VQA	1.09	78.20
POPE	General VQA	0.34	87.30
GQA	Spatial VQA	4.47	61.36
CV-Bench 2D	Spatial VQA	0.57	56.59
2DS	Controlled spatial	41.07	56.63

Table 2: Position Sensitivity Index (%) for LLaVA 1.5 7B across benchmarks.

4 Norm Imbalance Suppresses Positional Encoding

The previous section showed a striking failure of positional encoding: permuting vision tokens often leaves predictions nearly unchanged, producing an unintended “bag of words” behavior despite the presence of positional encodings (Vaswani et al., 2017; Yuksekgonul et al., 2023).

4.1 Norm Imbalance at the Interface

To probe *why*, we conduct an empirical norm analysis on the COCO validation set (5k image-text pairs) (Lin et al., 2015). COCO is a standard, diverse image and text dataset, and this analysis concerns internal representation statistics rather than spatial labels specific to a benchmark. We extract vision embeddings after the MLP projector and before the language model, compute their L_2 norms, and compare them to text token norms. Figure 3 shows the distribution comparison. We also verified the same pattern on 2DS, indicating that the finding does not depend on the choice of dataset.

In LLaVA 1.5, vision embeddings are typically 1–2 orders of magnitude larger than text embeddings, with an upper tail approaching 3 orders. While alignment work often emphasizes direction (cosine similarity) because logits are inner products (Hessel et al., 2021; Huh et al., 2024; Maniparambil et al., 2024), such large magnitude gaps can dominate the scaled dot product and suppress positional signals.

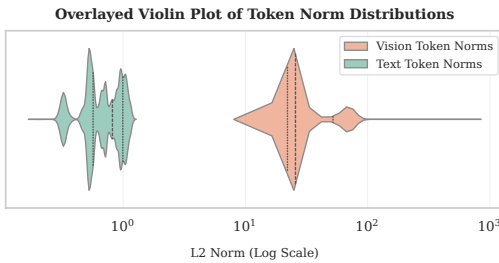


Figure 3: Distribution of L_2 norms for tokens in COCO validation dataset (log scale). Vision token norms range between 10^1 and 10^3 , while text token 3×10^{-1} and 10^0 .

Across models. To further verify whether this norm imbalance is specific to LLaVA or a broader phenomenon, we measure the same interface statistics on twelve models spanning five VLM families that use a projector, all on the COCO validation set. Figure 4 summarizes two norm statistics: panel (a) compares interface norm gap with PSI, and panel (b) shows the residual vision/text norm ratio across layers. For Qwen models, which use M-RoPE, we keep the position IDs unchanged and only shuffle the vision embeddings into the same image token slots, the same procedure that we use for the 1D RoPE models. We mainly compare PSI within each family, because models from different families also differ in position

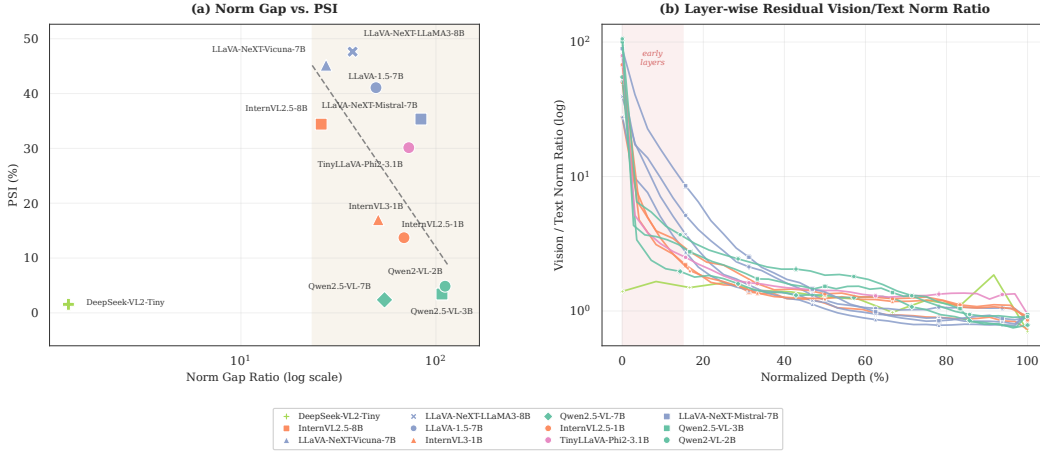


Figure 4: **Across models.** (a) Interface norm gap versus PSI across twelve VLMs with a projector. (b) Residual vision/text norm ratio across decoder depth.

encodings, training data, and architectures, so a cross-family PSI gap cannot be traced to the norm gap alone. Eleven of twelve models exhibit a large interface gap. Most also show the strongest residual imbalance in the early decoder layers, where positional encoding sensitivity should be most suppressed. Only DeepSeek-VL2 Tiny, which jointly trains the vision encoder with the decoder, is close to balanced at the interface and across depth. Exact per-model values are listed in Appendix F.¹

4.2 How Norms Suppress RoPE: Theoretical Mechanism

Our analysis focuses on the LLaVA architecture, which uses a LLaMA decoder (Grattafiori et al., 2024). LLaMA uses a design that applies RMSNorm before each sublayer: each sublayer first normalizes the state (Zhang & Sennrich, 2019), then computes the attention or MLP update on that normalized state, and finally adds that update back to the original residual stream. However, vision token inputs are unnormalized. RoPE is applied inside attention to normalized queries and keys (Su et al., 2024). This means RoPE acts on normalized activations, but its effect must still compete with the unnormalized residual after the update is added back. Figure 4(b) shows the imbalance of vision and text hidden states across layers, and without normalization vision token input ratio can be 2 orders of magnitude larger than text norm.

Our hypothesis is that the large vision norms at the interface make this competition unfavorable to RoPE. Even if RoPE changes attention scores, the resulting update can remain small relative to the large vision residual carried forward by early layers. In that case, positional information is structurally present, but it is not fully available for downstream use. This suppression acts on the attention output as a whole. A large vision norm reduces the effect of the whole attention output on the token representation, whether that output is driven by position or by content. Token order depends on this output more than content does, because token order reaches the token state only through attention, while vision token content is also carried forward in the residual stream. If this account is correct, the ratio between vision and text norms should remain especially high in the early decoder layers, precisely where RoPE sensitivity is weakest.

More detailed explanation is given in Appendix B. To connect the mechanism to a directly measurable quantity, we analyze how attention changes with the RoPE phase. Let logits be $\ell_j = \langle q', k'_j \rangle / \sqrt{d}$ and $\alpha_k = \text{Softmax}(\ell)_k$ be attention weights, and partition tokens into vision (V) and text (T). Define the group masses $\alpha_V = \sum_{v \in V} \alpha_v$, $\alpha_T = \sum_{t \in T} \alpha_t$, and the

¹On 2DS we observe the same imbalance (vision mean = 47.71, text mean = 0.74, ratio = 64×), confirming that the choice of analysis dataset does not affect the finding.

average logit derivatives for each group

$$g_V = \sum_{v \in V} \frac{\alpha_v}{\alpha_V} \frac{\partial \ell_v}{\partial \phi}, \quad g_T = \sum_{t \in T} \frac{\alpha_t}{\alpha_T} \frac{\partial \ell_t}{\partial \phi}.$$

Then (derivation in Appendix B) the change of total vision attention mass w.r.t. the positional phase ϕ is

$$\frac{\partial \alpha_V}{\partial \phi} = \alpha_V \alpha_T (g_V - g_T). \quad (2)$$

This derivation establishes a direct link between the shift in vision attention mass ($\frac{\partial \alpha_V}{\partial \phi}$) and the underlying positional sensitivity of the vision logits (g_V). A very small attention shift implies weak positional sensitivity, while a large shift indicates that RoPE is having a meaningful effect. This provides the mathematical basis for the RoPE sensitivity probe used later.

4.3 Connecting Norm Imbalance to Positional Underutilization

The PSI results from Section 3 provide behavioral evidence of the same problem. On standard benchmarks, permuting vision tokens changes accuracy only slightly. The norm analysis in this section gives a mechanism for that behavior: large vision norms make updates driven by RoPE weak relative to the residual stream. Taken together, these results suggest that positional information in VLMs with a projector is structurally present but functionally suppressed.

Figure 4(a) provides supporting descriptive context for the same point when we plot PSI against norm gaps. Within the LLaVA and InternVL families, we see that larger norm gaps consistently coincide with lower PSI. Across families, the same norm gap can produce very different PSI: Qwen models show uniformly low PSI despite norm gaps comparable to those where LLaVA and InternVL models retain moderate positional reliance. DeepSeek-VL2 Tiny, the only model with a nearly balanced interface, also shows low PSI. These differences across families point to a distinction between *availability* and *usage*. The norm gap determines whether positional encoding is available to the decoder, but availability alone does not guarantee that the model will use it. Usage also depends on factors such as the quality of vision features and training. The comparison across models suggests this distinction but cannot confirm it, because families differ in many ways beyond the norm gap. In Section 5, we test the link directly by reducing the norm gap on a single architecture and measuring whether increased availability leads to increased usage.

5 Targeted Interventions: Restoring and Dissociating Spatial Sources

Section 4 showed that the norm gap controls the availability of positional encoding. Two questions follow: *does reducing the gap increase positional reliance?* And *can enriching vision token content improve spatial accuracy through the other spatial source?* We design two minimal interventions to answer these questions

5.1 Hypotheses and Setup

We formulate two hypotheses, each paired with an intervention that targets one spatial source.

H1 (RoPE contribution). If norm imbalance suppresses RoPE through the residual stream, then reducing the vision/text norm gap should let the LLM use positional encoding more strongly. This should increase positional sensitivity and RoPE probe responses.

H2 (Contribution of vision content). Spatial information already encoded in vision token content can also support spatial reasoning. If intermediate vision features retain richer geometry than final layer features (Wang et al., 2023; Jiang et al., 2024), then adding them should improve spatial accuracy even if PSI does not increase further.

We use LLaVA-1.5 (7B) (Liu et al., 2023) as the baseline and propose two interventions. Both variants are trained from scratch using the standard LLaVA two stage pipeline (558K pretrain pairs, then 665K instruction tuning examples). They are not applied after training to a fully trained model. The intervention therefore requires a model with an open training recipe and open training data, where LLaVA-1.5 provides both.

+Normalize (tests H1). RMS normalization (Zhang & Sennrich, 2019) of vision embeddings *after* the linear projector and *before* feeding the LLM, calibrated to typical text token norms (empirically: mean ≈ 0.83 , max ≈ 1.22).

+Normalize+Multilayer (tests H1+H2). Concatenate intermediate vision features along the token dimension (layers 12, 16, 20, 24) following (Jiang et al., 2024), then apply the same normalization. This injects more detailed spatial signals while controlling scale. The additional compute overhead is empirically negligible (Appendix E).

5.2 Results on 2DS: Prediction Confirmed and Dissociation Revealed

Both interventions improve overall accuracy, but that alone does not tell us whether the gain comes from better use of RoPE or from richer spatial content in vision tokens. 2DS and PSI let us separate these possibilities. Table 3 reports PSI across benchmarks.

+Normalize substantially increases PSI, consistent with H1: reducing the norm gap restores RoPE leverage and raises reliance on token order. The effect is strongest on the controlled spatial benchmark, where PSI rises from 41.07% to 61.21% on 2DS. Spatial VQA benchmarks also respond, with CV-Bench 2D jumping from 0.57% to 11.31%. General VQA benchmarks (VQAv2, POPE) show minimal change, consistent with the benchmark hierarchy in Section 3. Statistical tests on the underlying original versus permuted accuracies support the same pattern (Appendix G): 2DS and GQA are highly significant, CV-Bench 2D becomes significant after the interventions, and POPE remains not significant.

The striking result is that +Normalize+Multilayer does not continue this PSI trend. On 2DS it reaches 54.21%, below +Normalize, even though it achieves the highest accuracy in Table 10. This means that adding more spatial information to vision tokens helps end-to-end spatial reasoning, but it makes the model rely less on RoPE than the norm-only variant.

The dissociation. On 2DS, +Normalize achieves the highest PSI (61.21%) but not the highest accuracy (59.30%). +Normalize+Multilayer achieves the highest accuracy (64.80%) with a lower PSI (54.21%). Both variants apply the same normalization and differ only in vision token content. Accuracy and PSI move in opposite directions between the two treatments, which means that extra accuracy does not require extra reliance on token order. Section 4 distinguished *availability* from *usage* of positional encoding. The intervention data adds a third dimension, *task outcome*, which depends on both spatial sources together. +Normalize mainly increases the availability and usage of positional encoding. +Normalize+Multilayer strengthens vision-token content instead, so the model achieves better accuracy while relying less on RoPE. The category breakdown (Appendix H.1) confirms this: +Normalize+Multilayer gains most on relation-heavy 2DS categories, consistent with stronger local geometry.

The same dissociation appears in a model we did not train. DeepSeek-VL2 Tiny (Section 4) has a nearly balanced interface, the lowest PSI among the surveyed models (1.56%), and the highest 2DS accuracy (85.73%). This matches +Normalize+Multilayer, where strong spatial accuracy comes with low positional reliance. Note that DeepSeek-VL2 Tiny also differs from the other surveyed models in joint training, backbone design, corpus size, and image tiling. These factors are confounded, so the pattern cannot be traced to vision content alone.

The RoPE sensitivity probe confirms that the two interventions produce different internal configurations: both increase early-layer $\Delta\alpha_V$ and Δg_V , but through different layer-wise patterns (Figure 5). Attention analyses show the same split: +Normalize yields more bal-

Dataset	Type	LLaVA 1.5	+ Normalize	+ Normalize + Multilayer
VQAv2	General VQA	1.09	2.03	2.08
POPE	General VQA	0.34	0.35	0.46
GQA	Spatial VQA	4.47	6.10	5.64
CV-Bench 2D	Spatial VQA	0.57	11.31	9.72
2DS	Controlled spatial	41.07	61.21	54.21

Table 3: Position Sensitivity Index (%).

Benchmark	Type	LLaVA 1.5	+ Normalize	+ Normalize + Multilayer
What’sUp Controlled A	Real spatial	58.66	76.94	72.82
What’sUp Controlled B	Real spatial	69.20	73.69	80.31
VQAv2	General VQA	78.20	78.76	79.17
POPE	General VQA	87.30	87.30	87.70
GQA	Spatial VQA	61.36	62.04	62.52
CV-Bench 2D	Spatial VQA	56.59	59.91	58.69

Table 4: Accuracy (%) on real-image spatial and standard benchmarks. What’s Up results are averages over three seeds.

anced vision/text allocation, while +Normalize+Multilayer shows selective, target-centered attention (Appendices I and K).

5.3 Real-Image and Standard Benchmark Results

We next evaluate on real-image spatial benchmarks and standard VLM benchmarks to check whether the 2DS findings generalize and whether general capabilities are preserved.

Table 4 shows that the 2DS findings transfer to real images. On What’s Up Controlled A/B, both interventions yield large gains, but the best model differs by split, preserving the same dissociation pattern. On the remaining four What’s Up splits, semantic cues lead to mixed results, consistent with the benchmark hierarchy (Appendix H.3). On standard benchmarks, both interventions preserve or modestly improve general capabilities. The interventions target how much the decoder uses positional information, and VQAv2 and POPE rarely depend on it (Section 3).

5.4 How Norm Balance Changes RoPE Sensitivity

We now examine the internal mechanism directly to confirm that the behavioral changes reflect the RoPE suppression predicted in Section 4.

Norm ratio across layers. Section 4 predicted that vision token norms should dominate the residual stream most strongly in early layers. Figure 4(b) confirms this across models. The ratio is highest in early layers and decays through mid-depth, with DeepSeek-VL2 Tiny showing a flat profile consistent with its balanced interface.

RoPE sensitivity probe. The derivation in Section 4.2 predicts that large vision norms suppress the positional sensitivity of attention. To measure this directly, we perturb only the RoPE phase of vision keys while keeping the query, text keys, and causal masking fixed. We then compare attention before and after the perturbation. We report two finite difference quantities: $\Delta\alpha_V$, which approximates $\frac{\partial\alpha_V}{\partial\phi}$ and measures how much attention mass moves over vision tokens, and Δg_V , which approximates the average logit derivative for the vision group and measures intrinsic sensitivity to the RoPE angle.

Sensitivity across layers. Figure 5 shows that baseline LLaVA has the weakest RoPE sensitivity in the first few layers, exactly where the norm ratio is largest. Both intervention variants increase $\Delta\alpha_V$ and Δg_V in early layers, with the clearest gains in layers 0–5. This is the layer range where balancing norms should matter most, and it matches the PSI improvements in Table 3.

Dose response. The preceding analyses show that large vision norms co-occur with low RoPE sensitivity, and that normalization restores sensitivity. A continuous scaling study strengthens this link. We multiply the vision norms by a scalar α at the inference time and measure the resulting $\Delta\alpha_V$ and Δg_V (Figure 5). As α decreases toward the text norm scale, RoPE sensitivity rises smoothly and peaks around $\alpha = 0.02$ – 0.05 , close to the reciprocal of the measured LLaVA-1.5 interface gap ($1/49.2584 \approx 0.02$). This is the matched scale regime where the positional encoding becomes most effective.

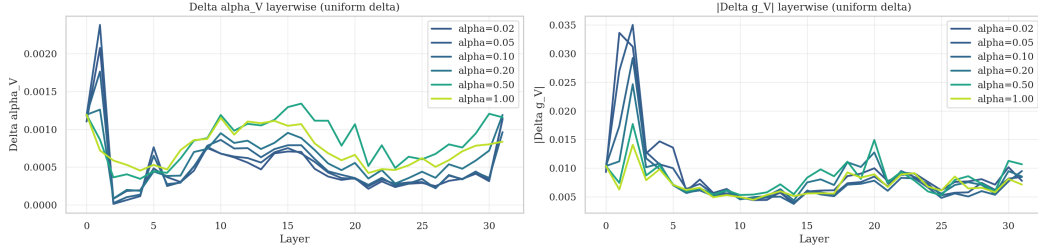


Figure 5: RoPE sensitivity (higher = larger change under a phase shift) across layers.

6 Limitations and Conclusion

Limitations. This study has the following limitations. First, interventions are validated only on LLaVA-1.5 7B, although the descriptive survey suggests that interface norm imbalance is more widespread. Since each intervention variant is trained from scratch, extending the comparison requires another model with an open training recipe and open training data. However, unfortunately, few models release both. Second, the contribution of vision token content is inferred indirectly from the PSI-accuracy dissociation rather than measured with a dedicated probe. Third, our scope is limited to architectures with a projector. Models that jointly optimize the vision encoder and the LLM, such as DeepSeek-VL2, may reduce the norm gap through training alone. Additionally, models that tokenize images into the language model’s own vocabulary, such as Chameleon (Team, 2025), have no projector interface where a norm gap can form. How these models carry positional information is left to future work. Fourth, 2DS is a synthetic diagnostic. It isolates positional use but does not capture the full difficulty of real-world spatial reasoning.

Conclusion. Spatial evaluation of VLMs often conflates three questions. Is positional encoding available? Does the model use it? And how does the final task outcome reflect both spatial sources? This work separates these questions. The norm gap at the projector interface suppresses positional encoding in the residual stream, limiting its availability. Whether the model uses positional encoding, even when available, depends on additional factors including vision feature quality and training. Task outcome reflects both sources together, which is why the highest positional reliance and the highest spatial accuracy come from different interventions.

For model design, norm balance at the projector interface deserves explicit attention. A large scale gap is not just an alignment nuisance. It structurally limits what the decoder can extract from positional encoding. The survey of twelve models shows that this imbalance is widespread (eleven of twelve models), and only the jointly trained DeepSeek-VL2 Tiny achieves a balanced interface. For evaluation, PSI should be interpreted as a diagnostic of positional reliance rather than a full measure of spatial competence. A model can achieve strong spatial accuracy with low PSI if its vision features carry sufficient spatial information. The +Normalize+Multilayer results show this under controlled training, and DeepSeek-VL2 Tiny shows the same pattern in a model we did not train. Controlled benchmarks such as 2DS remain important because they isolate positional reliance from vision token content in ways that standard benchmarks cannot.

Acknowledgments

The work is supported by the National Science Foundation (NSF) through Awards #DMR-2433348 (AI Research Institutes), #2131186 (CISE-MSI), #1827505 (PFI), and the US Air Force Office of Scientific Research (AFOSR) via Award #FA9550-21-1-0082. The work is also supported by a CCNY College-wide Research Vision (CRV) Fund (2022-2025) and the Google CyberNYC Initiative (2025-2030). This work used Google Cloud through the CloudBank project, which is supported by NSF Award #1925001. We specifically thank Dr. Ping Ji’s generous support for computing resources.

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177cccb411a7d800-Paper-Conference.pdf.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- Mohammad Asadi, Jack W O’Sullivan, Fang Cao, Tahoura Nedaei, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.
- Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=tDUrseRRU>.
- Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024.
- Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=k7vcuqLK4X>.
- An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=JKEIYQUSUc>.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=2dn03LLiJ1>.
- Enrique Queipo de Llano, Alvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin, 2025. URL <https://arxiv.org/abs/2510.06477>.
- Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5089–5098, June 2022.
- Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*, 2024.

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. ISSN 0166-2236. doi: [https://doi.org/10.1016/0166-2236\(92\)90344-8](https://doi.org/10.1016/0166-2236(92)90344-8). URL <https://www.sciencedirect.com/science/article/pii/0166223692903448>.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Celebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath R-parthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic,

Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshv, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018. doi: 10.1109/CVPR.2018.00380.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi:

- 10.18653/v1/2021.emnlp-main.595. URL <https://aclanthology.org/2021.emnlp-main.595/>.
- Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17980–17989, June 2022.
- Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. URL <https://arxiv.org/abs/1902.09506>.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/huh24a.html>.
- Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin’e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models, 2024. URL <https://arxiv.org/abs/2310.08825>.
- Amita Kamath, Jack Hessel, and Kai-Wei Chang. What’s “up” with vision-language models? investigating their struggle with spatial reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=RN5KLywT11>.
- Declan Kutscher, David M. Chan, Yutong Bai, Trevor Darrell, and Ritwik Gupta. Reordering patches improves vision models, 2025. URL <https://arxiv.org/abs/2505.23751>.
- Bozhou Li, Xinda Xue, Sihan Yang, Yang Shi, Xinlong Chen, Yushuo Guan, Yuanxing Zhang, and Wentao Zhang. The unseen bias: How norm discrepancy in pre-norm mlms leads to visual information loss. *arXiv preprint arXiv:2512.08374*, 2025a.
- Hengzhuang Li, Xinsong Zhang, Qiming Peng, Bin Luo, Han Hu, Dengyang Jiang, Han-Jia Ye, Teng Zhang, and Hai Jin. Unleashing the intrinsic visual representation capability of multimodal large language models. *arXiv preprint arXiv:2512.06281*, 2025b.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023a. URL <https://proceedings.mlr.press/v202/li23q.html>.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023b. URL <https://arxiv.org/abs/2305.10355>.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL <https://arxiv.org/abs/1405.0312>.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024.

- Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed El Amine Seddik, Sanath Narayan, Karttikeya Mangalam, and Noel E. O'Connor. Do vision and language encoders represent the world similarly? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14334–14343, 2024. doi: 10.1109/CVPR52733.2024.01359.
- Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=chanJGoa7f>.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>.
- Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14974–14983, June 2023.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2023.127063>. URL <https://www.sciencedirect.com/science/article/pii/S0925231223011864>.
- Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=F7aAhfitX6>.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL <https://arxiv.org/abs/2405.09818>.
- Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL <https://openreview.net/forum?id=Vi8AepAXGy>.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024b. URL <https://arxiv.org/abs/2401.06209>.
- Ameni Trabelsi, Maria Zontak, Yiming Qian, Brian Jackson, Suleiman Khan, and Umit Batur. What matters when building vision language models for product image analysis? 2025. URL <https://www.amazon.science/publications/what-matters-when-building-vision-language-models-for-product-image-analysis>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. What makes for good visual tokenizers for large language models?, 2023. URL <https://arxiv.org/abs/2305.12223>.
- Mert Yuksekogunul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=KRLUvxh8uaX>.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf.

Jingyang Zhang. Vlm-visualizer. <https://github.com/zjysteven/VLM-Visualizer>, 2024.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Jiawei Cao, Hao Cheng, and Kaijie Wu. What drives attention sinks? a study of massive activations and rotational positional encoding in large vision-language models. *Information Processing & Management*, 63(2): 104431, 2026.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=1tZbq88f27>.

A Disclosure

A.1 LLM Usage Disclosure

We used large language models as research assistants throughout the project. Specifically, LLMs were used to support code implementation, assist in refining hypotheses and experimental design, aid in literature discovery, and help improve writing clarity and organization.

The core research ideas, problem formulation, and overall methodology were initiated and guided by the authors. All generated code, experimental results, and analyses were carefully reviewed, validated, and iterated upon by the authors. The authors take full responsibility for the correctness, originality, and integrity of all content in this paper.

A.2 Reproducibility Disclosure

We include code, scripts, and instructions necessary to reproduce our results.

The codebase includes data processing pipelines, model loading, training and evaluation scripts, and analysis code for reported metrics and figures. In particular, we provide scripts to recompute all weight-space analyses, including SVD-based decompositions and alignment measures, directly from model checkpoints.

B Theoretical Analysis of Large Norm Effects on Attention Updates and RoPE

We consider a Transformer block that applies RMSNorm before each sublayer (as used in the classic LLaVA decoder). For token state $h^{(l)} \in \mathbb{R}^d$ at layer l ,

$$\underbrace{a^{(l)}(\phi)}_{\text{attn out}} = \text{Attn}(\text{RMSNorm}(h^{(l)}), \phi), \quad h^{(l+1)} = \underbrace{r^{(l)}}_{\text{residual carry}} + a^{(l)}(\phi),$$

where ϕ denotes the RoPE phase and $r^{(l)} = h^{(l)}$ (the input to attention before the update).

Keys and queries are formed from *per token RMS normalized* inputs; RoPE applies a norm preserving rotation to q, k (Su et al., 2024). We write q', k' for the query and key after this rotation. The residual is added back after each sublayer, so a large vision token norm keeps passing forward through the layers.

B.1 Suppression from residual scale.

We first state the result for a general attention update. Let θ be any quantity that the attention output depends on, so that $h^{(l+1)} = r^{(l)} + a^{(l)}(\theta)$, and let $u(\theta) = h^{(l+1)} / \|h^{(l+1)}\|$ be the unit normalization. RMSNorm removes the global scale, so the next sublayer reads rescaled u rather than $h^{(l+1)}$ itself. With some quotient rule and replacing $\frac{\partial \|h^{(l+1)}\|}{\partial \theta} = \frac{(h^{(l+1)})^T}{\|h^{(l+1)}\|} \frac{\partial h^{(l+1)}}{\partial \theta}$ we can have

$$\frac{\partial u}{\partial \theta} = \frac{(I - uu^\top)}{\|h^{(l+1)}\|} \frac{\partial a^{(l)}}{\partial \theta}.$$

When $\|r^{(l)}\| \gg \|a^{(l)}\|$ the magnitude of $\partial u / \partial \theta$ is suppressed by $\sim 1 / \|r^{(l)}\|$ in the vision token part.

Thus, for a fixed attention-level sensitivity, a larger residual norm reduces the effect of the attention update on the normalized representation presented to the next sublayer. The factor $1 / \|h^{(l+1)}\|$ suppression is independent of the particular choice of (θ) : it applies whether the attention update is driven by positional information, token content, or another attention-dependent quantity. A persistent vision–text norm imbalance may make high-norm vision tokens less responsive to attention-mediated updates than lower-norm text tokens. This may contribute to weaker visual integration and stronger reliance on textual priors, although this result by itself does not imply that the underlying vision representations are intrinsically lower quality.

The positional case. Setting $\theta = \phi$ gives the positional case, where a large vision token norm means much less effective position sensitivity. The residual carries over the large vision token norms through layers, and large norm reduce the position attention sensitivity. The reduced attention sensitivity might explain why even nontrivial attention shifts may yield small representational changes when the residual dominates.

B.2 Attention Level Phase Sensitivity (Group Form)

Within a given head, let logits be $\ell_j = \langle q', k'_j \rangle / \sqrt{d}$, and define attention weights $\alpha_j = \text{Softmax}(\ell)_j = \frac{\exp(\ell_j)}{\sum_i \exp(\ell_i)}$. Partition keys into vision V and text T , and define group masses

$$\alpha_V = \sum_{v \in V} \alpha_v, \quad \alpha_T = 1 - \alpha_V.$$

Let the average logit derivatives for the groups be

$$g_V = \sum_{v \in V} \frac{\alpha_v}{\alpha_V} \frac{\partial \ell_v}{\partial \phi}, \quad g_T = \sum_{t \in T} \frac{\alpha_t}{\alpha_T} \frac{\partial \ell_t}{\partial \phi}.$$

A standard calculation with some quotient rule from softmax yields the derivative for a single vision key given a fixed query:

$$\frac{\partial \alpha_v}{\partial \phi} = \alpha_v \left(\frac{\partial \ell_v}{\partial \phi} - \sum_k \alpha_k \frac{\partial \ell_k}{\partial \phi} \right) \quad (3)$$

We can use the above notation and get the group level sensitivity by aggregating all vision attentions and then substituting $\alpha_V, \alpha_T, g_V, g_T$

$$\frac{\partial \alpha_V}{\partial \phi} = \alpha_V \left(g_V - \sum_k \alpha_k \frac{\partial \ell_k}{\partial \phi} \right) \quad (4)$$

$$= \alpha_V \left(g_V - \left(\sum_{v \in V} \alpha_v \frac{\partial \ell_v}{\partial \phi} + \sum_{t \in T} \alpha_t \frac{\partial \ell_t}{\partial \phi} \right) \right) \quad (5)$$

$$= \alpha_V (g_V - (\alpha_V g_V + \alpha_T g_T)) \quad (6)$$

$$= \alpha_V \alpha_T (g_V - g_T) \quad (7)$$

Eq. (7) is a statement about the attention level, and we can directly measure the change without the residual scale entering.

Small phase probe. If we rotate *only vision keys* by a small phase Δ (queries fixed), then g_T is unchanged and to first order

$$\Delta\alpha_V \approx \alpha_V(1 - \alpha_V) \Delta g_V, \quad (8)$$

This factorization underlies our probe: $\Delta\alpha_V$ grows when (i) the *balance factor* $\alpha_V(1 - \alpha_V)$ is large (less saturation) and/or (ii) the *intrinsic* phase response Δg_V is large.

Empirically we can capture q, k inside attention and conduct position perturbation, so we can measure $\Delta\alpha_V$ and Δg_V at the attention level. The *effective* positional change is modulated by the residual. Reducing the interface scale (our +Normalize) decreases $\|r^{(l)}\|$ for vision tokens in early layers, mitigating suppression from residual scale and making phase shifts at the attention level consequential for subsequent computation. It also agrees with our observation that +Normalize increases position sensitivity in early layers in Section 5.4.

What each level measures. The factor $1/\|h^{(l+1)}\|$ scales the position-driven and the content-driven parts of the attention update together. A sensitivity measurement inside attention therefore cannot determine whether it is position or the content. However, PSI provides a complementary behavioral measurement. It compares the same model on the same set of vision embeddings before and after those embeddings are reassigned to different visual positions. The visual content, the token norms, and all model parameters remain unchanged; only the correspondence between token content and position is disrupted. PSI therefore measures how much model performance depends on the original content–position arrangement, rather than on a change in the available visual content itself. Eq. (7) measures the local sensitivity of attention to positional perturbations, whereas PSI measures whether disrupting positional organization degrade the performance. It measures how much model rely on the vision token positions.

Relation to high norm outlier tokens. The residual effect studied here is distinct from previously observed attention sinks. Massive activations can cause a small number of tokens to receive or attract a disproportionate share of attention, thereby changing the attention logits (Sun et al., 2024; de Llano et al., 2025; Zhang et al., 2026). Similarly in vision encoders, a small set of high norm tokens acts as registers (Darcet et al., 2024). Those effects act inside the softmax, and in contrast the factor derived in Section B.1 acts after the attention output has been computed and added to the residual. Our empirical observation also concerns a broad norm gap between vision and text tokens as groups, rather than only a few extreme outlier tokens. These mechanisms operate at different stages and granularity and both can coexist.

C 2DS Dataset Details

We constructed a synthetic dataset that: (1) Ensures spatial cues are essential for correctly answering queries. (2) Isolates and systematically controls semantic aspects, enabling analysis of their contributions to spatial understanding. The dataset combines semantic attributes (color, shape, or both) with spatial properties (relative or absolute positions). We denote the **2D Synthetic Dataset** as **2DS**

Example Queries including:

- “What color is at the bottom of the image?” (Color, absolute position)
- “Is the circle below the square?” (Shape, relative position)

These queries illustrate that object attributes (color, shape) serve primarily as identifiers, with positional relationships being crucial for correct responses.

C.1 Dataset Construction Details

We created categories containing images with 2, 3, 4, 5, and 6 objects. Within each category, we employed the same set of objects while randomly varying their *spatial positions*. Each category comprises 100 images, resulting in a total of 500 images. For every image, we systematically asked questions spanning combinations of semantic (color, shape, color and shape) *and* spatial (absolute, relative) properties, yielding 3000 total questions. Note the emphasis is on spatial relationship reasoning.

C.2 How effective can 2DS measure the position embedding sensitivity

As the preliminary experiments in Section 3 showed that most of the data retain the performance even with random permutation, we can use the same experiments to understand the 2DS results.

Permutation Table 5 shows the 2DS results with permutation of the input tokens. We also report the difference which is used to compute the PSI in Table 3. We can see the +Normalize version is more susceptible to the token permutation. As the baseline, we also test with the Vicuna-7B to show the model performance is well above the random guess.

Model	2DS (%)		
	Orig.	Perm.	Diff.
Vicuna-7B	15.97	–	–
LLaVA 1.5	56.63	33.37	-23.26
+ Norm	59.30	23.00	-36.30
+ Norm + Multi	64.80	29.67	-35.13

Table 5: For 2DS, we report accuracy on original (Orig.) and permuted (Perm.) token arrangements with difference (Diff.).

Compression Another way to test 2DS is through compression, which progressively weakens the positional route together with spatial resolution. If the vision encoder alone carried most of the needed spatial information, compression would not hurt performance much. If positional encoding is important, performance should drop sharply. Figure 6 shows that 2DS behaves like the second case.

The sharper drop on 2DS contrasts with the much smaller degradation on standard benchmarks and again suggests that positional encoding plays a much more important role for this controlled benchmark.

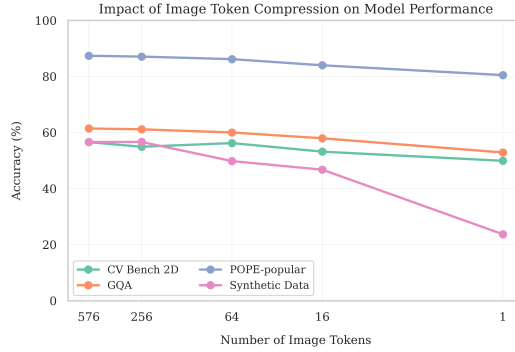


Figure 6: Performance impact of vision token compression on 2DS. Accuracy drops sharply as the vision sequence is compressed, indicating that 2DS depends much more strongly on positional information than the standard benchmarks.

D Training Details

We follow the training recipe described in LLaVA 1.5 7B (Liu et al., 2023), consisting of two stages: pretraining and instruction tuning. All experiments utilize 8 Nvidia A100 GPUs. Learning rates for each training stage are detailed in Table 6, with all other hyperparameters and datasets identical to (Liu et al., 2023).

Model	Pre-train	Finetune
LLaVA 1.5 7B	1e-3	2e-5
+Normalize	1e-3	2e-5
+Normalize+Multilayer	1e-3	2e-5

Table 6: Learning rate of training

E Analysis of Computational Cost and Inference Time

In this section, we analyze the computational overhead of our two proposed methods to evaluate their practical efficiency.

Method 1: +Normalize. Our first method, which applies L2 normalization to the final layer vision tokens, introduces zero additional parameters or floating point operations (FLOPs) during inference. The normalization is a lightweight mathematical operation on the existing embedding vector and its computational cost is negligible.

Method 2: +Normalize+Multilayer. Our second method enriches the visual representation by concatenating features from the intermediate and final layers of the vision encoder. This modification affects the initial visual projection layer. In the original LLaVA, the projector is an MLP that maps a final 1024-dimensional feature vector to a 4096-dimensional LLM embedding space. Our proposed method concatenates four 1024-dimensional feature vectors and feeds them to the projector, producing a 4096-dimensional vector. Although this moderately increases the number of parameters in this single projection layer, this component still represents a very small fraction of the total parameters in the multibillion parameter VLM.

Empirical Inference Time. To assess the real world impact on performance, we benchmarked the inference speed of the baseline model against our two proposed methods. The experiment was conducted on a single NVIDIA RTX 4090 GPU, measuring the average time to process 500 samples from the GQA dataset validation split. All three models are within the same inference time ($0.137s \pm 0.002s$) / sample, so the compute cost differences are trivial in practice.

Model	Family	Size	Text Mean	Vision Mean	Ratio
DeepSeek-VL2 Tiny	DeepSeek-VL2	Tiny	6.96	9.09	1.3×
InternVL2.5 8B	InternVL	8B	0.82	21.02	25.8×
LLaVA-NeXT Vicuna 7B	LLaVA-NeXT	7B	0.79	21.49	27.3×
LLaVA-NeXT LLaMA3 8B	LLaVA-NeXT	8B	0.45	16.96	37.4×
LLaVA-1.5 7B	LLaVA	7B	0.79	38.86	49.3×
InternVL3 1B	InternVL	1B	0.43	21.80	50.6×
Qwen2.5-VL 7B	Qwen-VL	7B	1.04	56.69	54.5×
InternVL2.5 1B	InternVL	1B	0.45	30.71	68.6×
TinyLLaVA 3.1B	TinyLLaVA	3.1B	1.51	109.78	72.6×
LLaVA-NeXT Mistral 7B	LLaVA-NeXT	7B	0.13	11.26	83.7×
Qwen2.5-VL 3B	Qwen-VL	3B	0.80	86.25	107.4×
Qwen2-VL 2B	Qwen-VL	2B	0.54	59.92	111.4×

Table 7: Vision–text interface norm gap across models. L_2 norms are measured on COCO val (5k). Models are sorted by norm gap ratio. LLaVA-1.5, LLaVA-NeXT, and TinyLLaVA follow the classic LLaVA projector design. Qwen-VL and InternVL share the same broad paradigm but differ in connector architecture. DeepSeek-VL2 uses a distinct design with a mixture of experts vision and language connector.

Model	Family	Size	Norm Gap	Orig. Acc (%)	Perm. Acc (%)	PSI (%)
DeepSeek-VL2 Tiny	DeepSeek-VL2	Tiny	1.3×	85.73	84.40	1.56
InternVL2.5 8B	InternVL	8B	25.8×	76.60	50.23	34.42
LLaVA-NeXT Vicuna 7B	LLaVA-NeXT	7B	27.3×	69.73	38.23	45.17
LLaVA-NeXT LLaMA3 8B	LLaVA-NeXT	8B	37.4×	67.80	35.50	47.64
LLaVA-1.5 7B	LLaVA	7B	49.3×	56.63	33.37	41.07
InternVL3 1B	InternVL	1B	50.6×	53.97	44.80	16.99
Qwen2.5-VL 7B	Qwen-VL	7B	54.5×	84.07	82.07	2.38
InternVL2.5 1B	InternVL	1B	68.6×	79.03	68.20	13.71
TinyLLaVA 3.1B	TinyLLaVA	3.1B	72.6×	67.37	47.07	30.13
LLaVA-NeXT Mistral 7B	LLaVA-NeXT	7B	83.7×	55.73	36.03	35.35
Qwen2.5-VL 3B	Qwen-VL	3B	107.4×	80.47	77.70	3.44
Qwen2-VL 2B	Qwen-VL	2B	111.4×	66.47	63.23	4.86

Table 8: Position Sensitivity Index across models on 2DS. Models are sorted by norm gap ratio. Original and permuted accuracies are reported alongside PSI (relative drop). All PSI values use rule based scoring on the full 2DS benchmark.

F Norm and PSI Statistics Across Models

To assess whether the norm imbalance between vision and text is specific to LLaVA or a broader phenomenon, we measure interface norm statistics on twelve models spanning five VLM families that use a projector. For each model, we extract vision and text embeddings at the interface between the projector and the LLM on the COCO validation set (5k images) and compute their mean L_2 norms. Table 7 reports these values. Eleven of twelve models show norm gap ratios between 25.8×

Table 8 reports the corresponding Position Sensitivity Index (PSI) measured on 2DS for the same twelve models. PSI is computed by comparing accuracy with original versus randomly permuted vision tokens. Within the LLaVA and InternVL families, models with larger norm gaps tend to show lower PSI. Qwen models show uniformly low PSI (2–5%) despite high norm gaps (54–111×). DeepSeek-VL2 Tiny has the lowest PSI (1.56%) alongside the highest original accuracy (85.73%), consistent with the dissociation discussed in Section 5. These values underlie Figure 4(a) in the main text.

Dataset	LLaVA 1.5	+ Normalize	+ Normalize + Multilayer
VQAv2	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$
POPE	0.61	0.40	0.23
GQA	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$
CV-Bench 2D	0.66	$p < 0.0001$	$p < 0.0001$
2DS	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$

Table 9: P-values for original order versus permuted order accuracy.

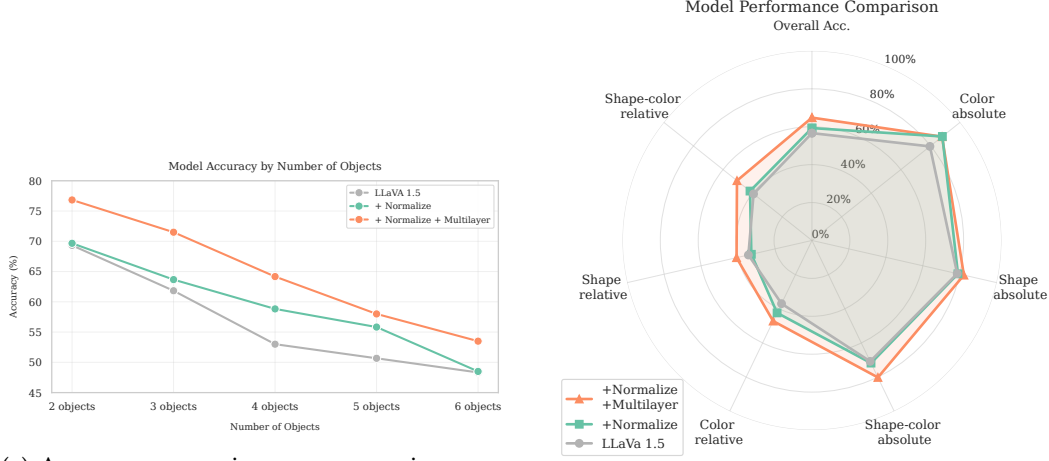
G Statistical Significance of PSI

PSI is computed from accuracy before and after vision token permutation, so significance should be tested on the underlying accuracy changes rather than on the ratio itself. For POPE, GQA, CV-Bench 2D, and 2DS, we have paired predictions at the example level and therefore use McNemar tests. For VQAv2, only aggregate server scores are available, so we use a two proportion Z-test.

The pattern matches the interpretation in the main text. POPE and baseline CV-Bench 2D show no significant permutation effect, while the more spatially focused settings 2DS and GQA do. On CV-Bench 2D, significance appears only after the interventions, consistent with the increase in PSI.

H Benchmark Performance

H.1 2DS Performance



(a) Accuracy comparison across varying numbers of objects. Our targeted adjustments yield consistent improvements, especially as spatial complexity increases.

(b) Accuracy comparison across different categories of 2DS dataset. We see consistent improvements across categories.

Figure 7: Model performance as the number of objects increases across different subcategories

Category breakdown. Table 10 reports the full 2DS breakdown by attribute and relation type. +Normalize chiefly helps categories where cleaner positional use can work with strong color cues, while +Normalize+Multilayer brings larger gains on relative categories, consistent with richer local geometry and object detail.

Category	LLaVA 1.5	+ Normalize	+ Normalize + Multilayer
Color_abs. ↑	79.60	88.00 (+8.40)	87.80 (+8.20)
Color_rel. ↑	37.00	42.40 (+5.40)	47.20 (+10.20)
Shape_abs. ↑	78.60	79.00 (+0.40)	82.20 (+3.60)
Shape_rel. ↑	34.40	32.80 (-1.60)	40.80 (+6.40)
Shape_color_abs. ↑	70.80	71.80 (+1.00)	80.20 (+9.40)
Shape_color_rel. ↑	39.40	41.80 (+2.40)	50.60 (+11.20)
Overall Acc. ↑	56.63	59.30 (+2.67)	64.80 (+8.17)

Table 10: 2DS accuracy (%) by attribute/relation. ↑ higher is better. Δ vs. baseline in parentheses.

Table 10 summarizes accuracy across distinct spatial query categories, and Figure 7b illustrates performance across varying numbers of objects, clearly highlighting the effectiveness of our interventions.

Vision Embedding Normalization: Normalization alone improves overall accuracy from 56.63% to 59.30%, strongly supporting our interpretability insight that balancing embedding norms significantly enhances spatial information utilization. Notably, improvements are particularly pronounced in color-related spatial relationship queries.

Normalization + Intermediate Layer Features: This approach achieves the highest overall accuracy (64.80%, +8.17% improvement over baseline), underscoring the importance of explicitly incorporating local spatial details. This method substantially enhances accuracy, especially in shape and color combined relative position queries.

Both methods consistently outperform the original LLaVA 1.5 model, particularly as spatial complexity (the number of objects) increases (Figure 7a), validating our hypothesis about the critical roles of embedding normalization and local spatial detail preservation.

H.2 Results on Standard Benchmarks

Category	LLaVA 1.5	+ Normalize	+ Normalize + Multilayer
VQAv2 $\uparrow$	78.20	78.76 (+0.56)	79.17 (+0.97)
POPE $\uparrow$	87.30	87.30 (0.00)	87.70 (+0.40)
GQA $\uparrow$	61.46	62.04 (+0.58)	62.52 (+1.05)
CV-Bench2D $\uparrow$	56.59	59.91 (+3.32)	58.69 (+2.10)

Table 11: Performance (accuracy %) on standard vision-language benchmarks. $\uparrow$ indicates higher is better. Values in parentheses show difference from LLaVA 1.5 baseline.

To assess generalization, we evaluate our methods on established benchmarks (Table 11) that are used in recent works (Liu et al., 2023; Tong et al., 2024a; Zhu et al., 2024; Liu et al., 2024).

Both targeted interventions yield moderate yet consistent improvements. Normalization and Multilayer Normalization consistently achieve small but meaningful gains (up to 3.32%), confirming that improved spatial representation benefits overall reasoning without sacrificing semantic capabilities. Notably, normalization shows more pronounced gains on CV-Bench 2D, particularly for spatial relationship tasks.

Specifically, we investigate the subcategories of the CV-Bench2D as shown in Table 12. We are curious why the permutation drop was not significant. As shown in Table 12, we can check the PSI index to see whether the model is sensitive to that task, and Count is not really a spatial task since there are marginal effects when changing spatial information. In comparison, we see the Relation task is more spatial, and the +Normalize has a much better performance in that category.

H.3 What’s Up Full Results

Table 4 in the main text reports What’sUp Controlled A/B results. Table 13 reports the full results for all six splits.

What’sUp (Kamath et al., 2023) contains six splits that differ in image source and spatial complexity. The **Controlled** splits (A and B) use synthetic images with simple objects placed at controlled positions. Semantic cues are minimal, so answering requires genuine spatial reasoning. The **real image** splits use images from COCO and GQA, each with one version containing one object (ONE) and another containing two objects (TWO). These real images carry rich semantic context, including object co-occurrence patterns and scene layout priors, that can support spatial answers without relying on positional encoding.

The Controlled splits show the largest gains from both interventions (up to +18.28 on Controlled A). This is expected because these splits isolate positional reasoning in the same way that 2DS does. The real image splits show mixed results for two reasons. First, the ONE splits already have near ceiling baseline accuracy (COCO-ONE 94.42%, GQA-ONE 96.44%), leaving little room for improvement. Second, the TWO splits allow semantic shortcuts that bypass positional encoding, so interventions targeting the positional source have weaker or inconsistent effects. Notably, +Normalize+Multilayer shows a larger drop on COCO Spatial-TWO (-7.88), which may reflect a training distribution shift when intermediate features change the representation that the model learned to rely on for these natural image queries.

These patterns are consistent with the main text findings. Interventions targeting positional encoding show the strongest effects on tasks that genuinely require spatial reasoning and cannot be solved through semantic cues.

Model	CV-Bench 2D (%)								
	Overall			Count			Relation		
	Orig.	Perm.	PSI	Orig.	Perm.	PSI	Orig.	Perm.	PSI
Vicuna-7B	31.36	–	–	12.31	–	–	54.46	–	–
LLaVA 1.5	56.59	56.26	0.58	52.07	52.37	-0.58	62.05	60.97	1.74
+ Norm	59.91	53.13	11.32	52.81	51.14	3.16	68.51	55.54	18.93
+ Norm + Multi	58.69	52.99	9.71	52.08	49.49	4.97	66.72	57.23	14.22

Table 12: Performance comparison on CV-Bench 2D benchmarks. We report accuracy on original (Orig.) and permuted (Perm.) token arrangements and PSI for each of the three tasks: overall 2D spatial reasoning, counting, and relation. We also report evaluation using Vicuna-7B on top without image input but just text input as the random baseline.

Split	Type	LLaVA 1.5	+ Normalize	+ Normalize + Multilayer
Controlled A	Controlled	58.66	76.94 (+18.28)	72.82 (+14.16)
Controlled B	Controlled	69.20	73.69 (+4.49)	80.31 (+11.11)
COCO Spatial-ONE	Natural	94.42	92.81 (-1.61)	92.66 (-1.76)
COCO Spatial-TWO	Natural	88.79	89.55 (+0.76)	80.91 (-7.88)
GQA Spatial-ONE	Natural	96.44	97.01 (+0.57)	96.81 (+0.37)
GQA Spatial-TWO	Natural	89.46	85.68 (-3.78)	86.02 (-3.44)

Table 13: What’sUp accuracy (%) across all six splits. Results are averages over three seeds. Controlled A/B appear in the main text. Δ vs. baseline in parentheses.

I Cross Modality Balance and System Prompt Effects

We retain the Cross Modality Balance (CMB) analysis in the appendix because it provides a complementary view of how the three models use vision and text. For a decoding step and attention head h , we define

$$\text{CMB}_h = \frac{\sum_{v \in V} a_{v,h}}{\sum_{i \in V \cup T} a_{i,h}},$$

where V and T denote the vision and user text token sets, and $a_{i,h}$ is the attention weight of head h . Larger CMB means that the head allocates more attention to vision tokens than to user text. We use this quantity only as a descriptive appendix analysis, not as a main metric.

One important confound is the system prompt. Figure 8 shows the attention share with and without the system prompt. The system prompt absorbs a large and mostly image invariant portion of attention mass. If it is included, the apparent vision/text balance can be misleadingly compressed. For this reason, the CMB analyses below focus on vision versus user text, and we report the prompt inclusive case only to show the confound explicitly.

Figure 9 shows that the three models organize attention across modalities differently. The baseline has many vision dominant heads scattered across layers, which is consistent with the bias from large norms discussed in the main text: vision tokens attract attention strongly, but not necessarily in a way conditioned on the question. +Normalize shows a more balanced progression from vision dominant early layers toward more text conditioned later layers, which matches the interpretation that balancing norms makes RoPE and question conditioning more usable. +Normalize+Multilayer exhibits a different pattern again: attention becomes more selective and more clearly staged, which is consistent with a model that receives stronger spatial content directly from vision tokens and therefore needs less broad positional search.

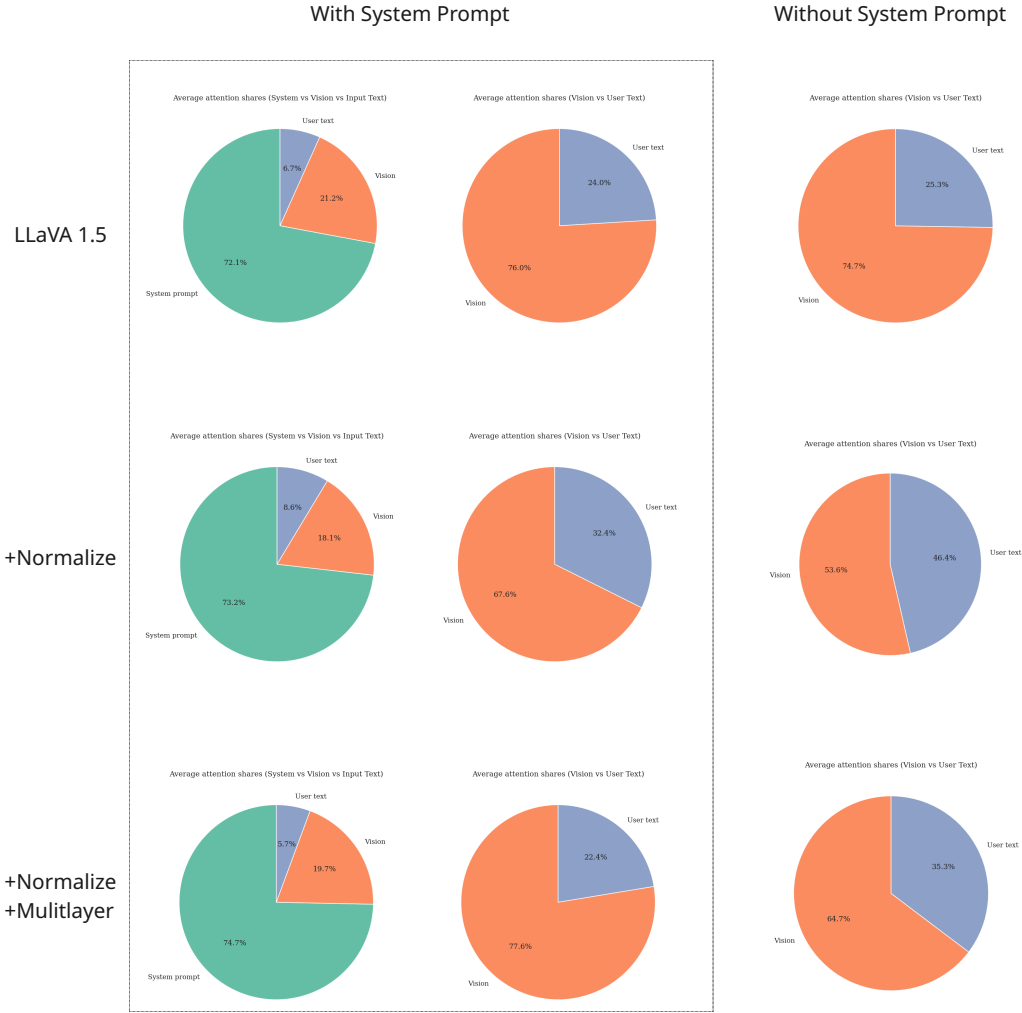


Figure 8: Visualization of system prompt, vision, and user input text attention shares. The left two columns include the system prompt when computing attention shares. The right column removes the system prompt and compares only vision and user text. The normalized variants show a more balanced vision/text allocation once the prompt confound is removed.

J Attention Head Visualization

The head maps for individual images provide a more detailed version of the averaged CMB heatmaps. The baseline often shows strong but diffuse vision preference. +Normalize tends to redistribute attention more broadly across relevant spatial surroundings before settling, which is consistent with a stronger use of positional information. +Normalize+Multilayer tends to produce narrower, more tightly focused patterns, suggesting that the needed spatial cue is already encoded in the vision token content and can be retrieved more directly.

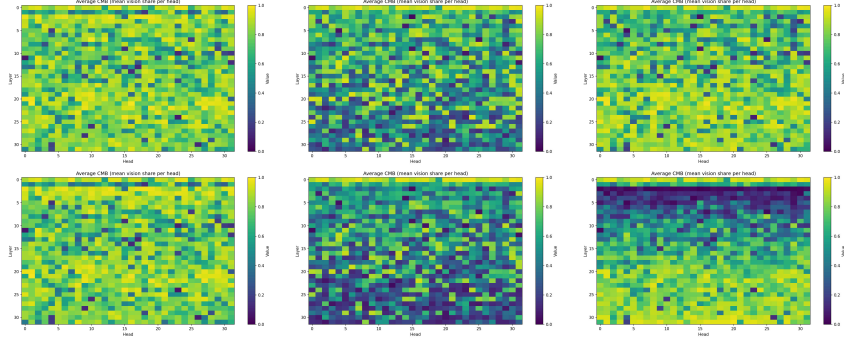


Figure 9: CMB heatmaps average (vision share) by layer $\times$ head on COCO val. Top row: inference with system prompt; bottom row: inference without system prompt. Columns: LLaVA 1.5, +Normalize, and +Normalize+Multilayer. Green indicates stronger vision preference; blue indicates stronger user text preference.

K Attention Visualization

We visualize self-attention from the LLM to interpret how the modifications change spatial reasoning. Figure 12 provides representative attention maps across four illustrative queries. Each query compares three scenarios: the baseline (Original LLaVA 1.5), normalized embeddings, and normalized embeddings combined with multilayer features. We visualize the first text token of the target object in the answer, which is marked in bold italics in Figure 12. More examples are in Appendix K.2.

Baseline (Original LLaVA 1.5): The baseline model shows diffuse and scattered attention patterns, possibly influenced by large magnitude vision embeddings that overshadow positional cues. This diffuseness makes it difficult for the model to clearly focus on relevant spatial tokens, leading to suboptimal spatial reasoning.

Embedding Normalization: Embedding normalization leads to more structured attention distributions over relevant spatial tokens. Compared to the multilayer variant, it often attends to a broader set of surrounding regions around the target. We interpret this as the model actively checking nearby spatial context through the positional channel: when RoPE becomes more usable, the model does not only look at the target object itself, but also explores neighboring regions needed to resolve relations.

Embedding Normalization + Multilayer Features: The multilayer feature variant produces sparser and more selective attention distributions. In many cases it focuses on the target region more directly, with less need to spread attention over the broader surroundings. This is consistent with the interpretation in the main text: features from intermediate layers of the vision encoder already carry more local spatial structure, reducing the need for the LLM to recover that structure through search driven by RoPE.

K.1 Interpreting Attention Patterns and Performance

To quantitatively interpret these observations, we computed the normalized attention entropy (Table 14). We average all text tokens’ attention over the vision tokens to compute the entropy. Lower entropy indicates more targeted attention distributions, and high scores mean more diffused attention distributions. More examples are in Appendix K.2.

The normalized embeddings exhibit higher entropy than the baseline, reflecting broader exploration of spatial regions while still remaining tied to the relevant object. This is consistent with a model that uses RoPE more actively: it reads more of the surrounding layout before deciding. In contrast, the multilayer variant with normalization demonstrates the lowest entropy, indicating a more concentrated focus on a few precise regions. This supports the view that the two improvements come from different mechanisms: +Normalize

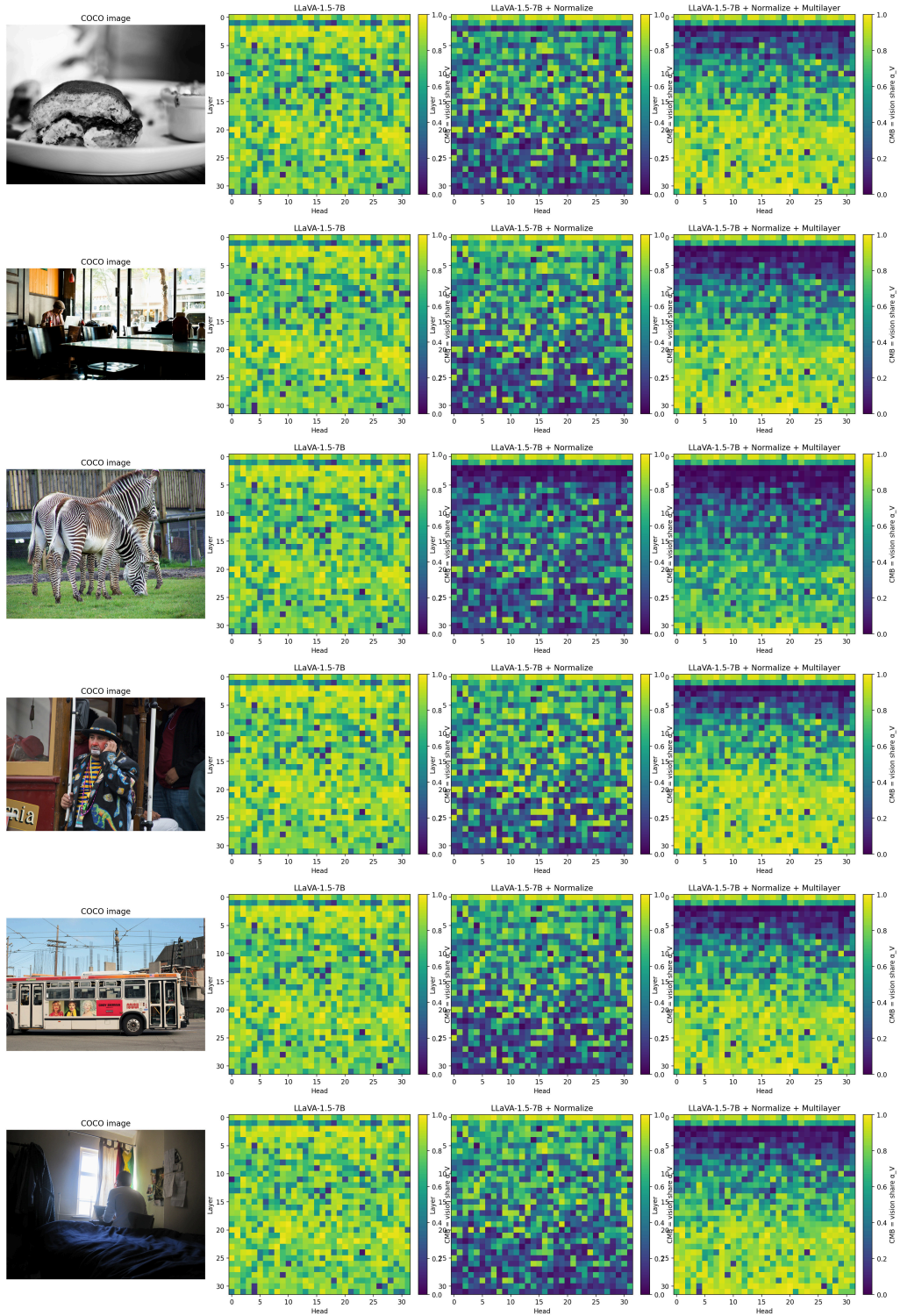


Figure 10: Visualization of each image’s attention head over layers. We run inference without the system prompt. CMB directly calculated from vision and text attention.

increases spatial use through the LLM positional channel, while +Normalize+Multilayer benefits more from spatial information already encoded in the vision encoder outputs.

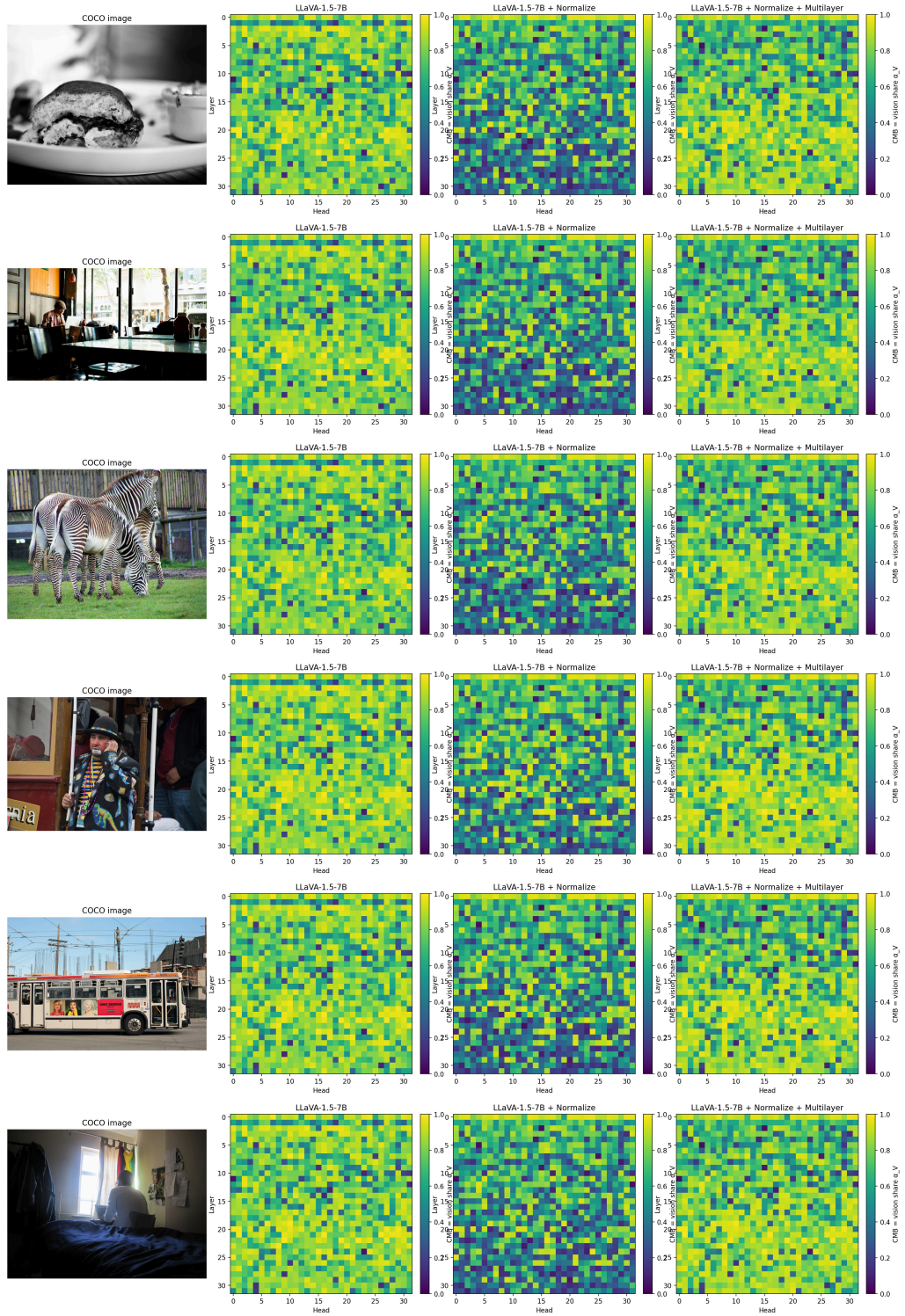


Figure 11: Visualization of each image’s attention head over layers. We run inference with the system prompt. CMB take values from vision and text attentions and normalize them with the sum of two.

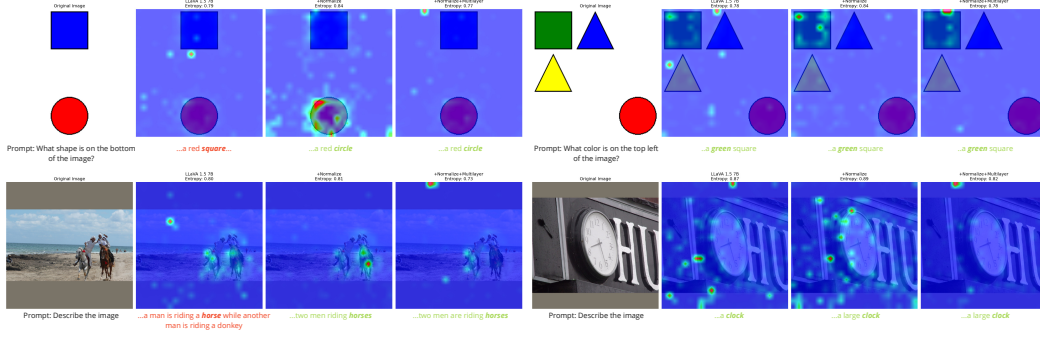


Figure 12: Visualization of self-attention patterns. We overlay the attention map on top of the image. The question for the model is under each row. We use the first target word of the response for attention map, for example the attention map is based on ‘square’ and ‘circle’ for top left rows. Entropy values are on top of each image. **Red answers are wrong. Green answers are correct.** Each image has entropy on top. Lower means more concentrated, and higher means more diffused.

Vision Token Attention Entropy	LLaVA 1.5	+ Normalize	+ Normalize + Multilayer
2DS ↓	0.76	0.90	0.72
COCO Val ↓	0.81	0.89	0.72
CV-Bench2D ↓	0.80	0.90	0.72

Table 14: Average attention entropy over different datasets. Lower entropy indicates more selective attention.

Providing features from early layers of vision encoders potentially offers a shortcut for LLMs to directly leverage spatial cues without extensive positional encoding training. Future research should explore strategies to optimally balance spatial detail from intermediate layers and positional encodings.

K.2 Attention Entropy

We adapt the visualization method from (Zhang, 2024) to analyze the LLM attention specifically on vision tokens. Attention maps correspond to particular, meaningful text tokens, as labeled above each image in 13. This is a more refined version compared with average text tokens. Notably, we observe consistently lower entropy in the models that use multilayer normalization.



Figure 13: More visualizations