THE UNSEEN BIAS: HOW NORM DISCREPANCY IN PRE-NORM MLLMS LEADS TO 'VISUAL INFORMATION LOSS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs), which couple pre-trained vision encoders and language models, have shown remarkable capabilities. However, their reliance on the ubiquitous Pre-Norm architecture introduces a subtle yet critical flaw: a severe norm disparity between the high-norm visual tokens and the low-norm text tokens. In this work, we present a formal theoretical analysis demonstrating that this imbalance is not a static issue. Instead, it induces an "asymmetric update dynamic," where high-norm visual tokens exhibit a "representational inertia," causing them to transform semantically much slower than their textual counterparts. This fundamentally impairs effective cross-modal feature fusion. Our empirical validation across a range of mainstream MLLMs confirms that this theoretical dynamic—the persistence of norm disparity and the resulting asymmetric update rates—is a prevalent phenomenon. Based on this insight, we propose a remarkably simple yet effective solution: inserting a single, carefully initialized LayerNorm layer after the visual projector to enforce norm alignment. Experiments conducted on the LLaVA-1.5 architecture show that this intervention yields significant performance gains not only on a wide suite of multimodal benchmarks but also, notably, on text-only evaluations such as MMLU, suggesting that resolving the architectural imbalance leads to a more holistically capable model.

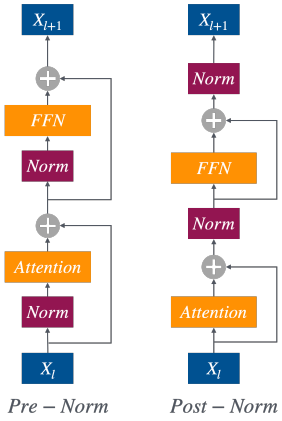
1 INTRODUCTION

In recent years, Multimodal Large Language Models (MLLMs) have achieved significant progress, demonstrating robust performance across a wide range of cross-modal tasks (Comanici et al., 2025; Hurst et al., 2024; Wu et al., 2024; Bai et al., 2025). A prevailing architectural paradigm involves augmenting a pre-trained Large Language Model (LLM) with visual capabilities by coupling it with a pre-trained Vision Encoder (VE). The VE, typically a Vision Transformer (ViT) (Dosovitskiy et al., 2020), first partitions an image into a sequence of patches and encodes them into a series of feature vectors, or "visual tokens." To bridge the modality gap, a lightweight adapter module is then introduced. This module's core function is to act as a translator, projecting these visual tokens into the LLM's word embedding space, thereby making visual information comprehensible to a model originally designed for text (Zhang et al., 2024).

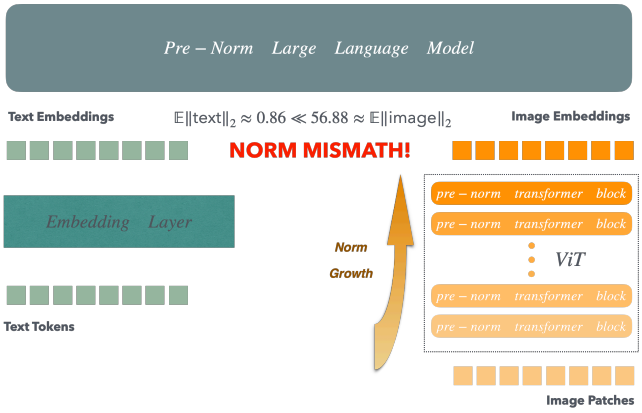
Despite their powerful general-purpose capabilities, emerging research has revealed inherent limitations in MLLMs. For instance, many models struggle with the perception of fine-grained visual details (Rahmanzadehgervi et al., 2024). Furthermore, within their self-attention mechanisms—the core component for weighing the importance of different inputs—visual tokens often receive less focus than their textual counterparts (Chen et al., 2024a). To address these challenges, we identify a more fundamental problem rooted in the now-ubiquitous Pre-Norm Xiong et al. (2020) architectural design. In this paradigm, normalization is applied before the main computational block (F), with the residual update defined as:

$$\mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + F(\text{Norm}(\mathbf{h}^{(l)})) \tag{1}$$

This architecture is widely adopted because it is easier to train. By leaving the residual path $\mathbf{h}^{(l)}$ unaltered, it creates an identity-like connection that ensures smooth gradient flow, preventing vanishing gradients in deep networks. However, this design has a critical side effect: since the output



(a) Pre-Norm vs Post-Norm



(b) The Norm Mismatch Problem Induced by MLLM Architectures

of the residual sum is never re-normalized, the variance—and consequently, the L_2 norm—of the hidden states tends to accumulate and grow with network depth (Kim et al., 2025). As is shown in Figure 1b, it creates a particularly acute imbalance in MLLMs where high-norm visual tokens and lower-norm text tokens are processed together within a shared Pre-Norm LLM backbone—as the visual tokens themselves are generated by a deep, Pre-Norm ViT.

Our formal theoretical analysis reveals a critical dynamic: a fundamental asymmetry in the evolutionary pace of visual and textual representations through the LLM's layers. We demonstrate that for high-norm visual tokens, the Pre-Norm update mechanism induces a high "representational inertia", causing them to undergo a much slower semantic transformation. In contrast, lower-norm textual tokens adapt their representations more readily, leading to a mismatched rate of convergence towards a unified multimodal space. Notably, this dynamic divergence arises not from an intrinsic property of visual versus textual information, but from an architectural artifact: the interplay between the Pre-Norm design and the prevailing MLLM paradigm.

Bridging theory and practice, we first analyzed a range of mainstream open-source VL models, finding that the norm disparities and asymmetric update rates are consistent with our theoretical predictions. Based on this validation, we conducted experiments on the LLaVA architecture, revealing that a remarkably simple intervention—inserting a single LayerNorm layer for visual tokens after the adapter—is sufficient to yield significant performance gains across both multimodal and pure text evaluations, indicating a more holistic improvement in the model's capabilities.

In this work, our key contributions are threefold:

- **Theoretical Identification of Asymmetric Dynamics.** We are the first to identify and theoretically formalize the issue of cross-modal norm disparity in Pre-Norm MLLMs. Our analysis reveals an "asymmetric update dynamic" where high-norm visual tokens exhibit "representational inertia," leading to a slower semantic evolution compared to text tokens .

- **Extensive Empirical Validation.** We provide extensive empirical validation across a suite of mainstream open-source MLLMs, demonstrating that the predicted norm disparities and asymmetric update rates exist, confirming our theoretical model in practice.

- **A Simple and Effective Solution.** We propose a simple, effective, and computationally inexpensive solution: a single, carefully initialized LayerNorm layer to enforce norm alignment. Our experiments show this method yields significant performance gains not only on multimodal tasks but also, unexpectedly, on text-only benchmarks, indicating a more holistic improvement to the model's capabilities.

2 PRELIMINARIES

Our analysis is grounded in the core components of modern Transformer architectures. We briefly review the self-attention mechanism, the role and types of normalization layers, and the critical design choice between Pre-Norm and Post-Norm architectures.

2.1 SELF-ATTENTION

The self-attention mechanism is the computational core of the Transformer. For an input sequence of hidden states $\boldsymbol{H} \in \mathbb{R}^{N \times D}$, it first linearly projects the sequence into queries ($\boldsymbol{Q}$), keys ($\boldsymbol{K}$), and values ($\boldsymbol{V}$) using learned weight matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D \times d_k}$. The unnormalized dot-product scores are computed as $\boldsymbol{Q}\boldsymbol{K}^T$.

2.2 NORMALIZATION LAYERS IN TRANSFORMERS

Normalization layers are a critical component for stabilizing the training of deep networks by controlling the distribution of activations. In Transformers, they ensure that the inputs to each sublayer (self-attention and FFN) remain well-behaved, preventing the magnitude of activations from exploding or vanishing. This is particularly important in MLLMs where features from different modalities, with potentially different statistical properties, are processed together. Two common normalization schemes are:

Layer Normalization (LayerNorm). This technique normalizes activations across the feature dimension for each token independently (Ba et al., 2016). For an input vector $\boldsymbol{x}$, it is defined as:

$$\text{LayerNorm}(\boldsymbol{x}) = \frac{\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]}{\sqrt{\text{Var}[\boldsymbol{x}] + \epsilon}} \odot \boldsymbol{g} + \boldsymbol{\beta} \tag{2}$$

where $\boldsymbol{g}$ (gain) and $\boldsymbol{\beta}$ (bias) are learnable parameters that restore expressive power.

Root Mean Square Norm (RMSNorm). A simplified and computationally efficient variant of LayerNorm that forgoes re-centering (subtracting the mean) (Zhang & Sennrich, 2019). It normalizes by the root mean square of the vector, proving effective in many modern LLMs:

$$\text{RMSNorm}(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\sqrt{\frac{1}{D}\|\boldsymbol{x}\|_2^2 + \epsilon}} \odot \boldsymbol{g} \tag{3}$$

2.3 THE PRE-NORM VS. POST-NORM RESIDUAL ARCHITECTURE

A Transformer block is functionally an additive update mechanism that refines a token's representation (Vaswani et al., 2017; He et al., 2016). For our analysis, it is useful to interpret the components of this update geometrically. The core operation is:

$$\boldsymbol{h}^{(l+1)} = \boldsymbol{h}^{(l)} + \Delta\boldsymbol{h}^{(l)} \tag{4}$$

In this view, we can consider $\boldsymbol{h}^{(l)}$ as the **Previous State**, representing the token's current position in a semantic space, which arrives via the skip connection. The term $\Delta\boldsymbol{h}^{(l)}$, computed by the residual branch (e.g., the self-attention sublayer), can be seen as the **Update Vector** that adjusts this position. The sum, $\boldsymbol{h}^{(l+1)}$, is therefore the resulting **New State**.

The critical design choice is where to place the normalization operation relative to this residual sum. The structural difference between the Pre-Norm and Post-Norm architectures is illustrated in Figure 1a. This defines two architectural families with distinct trade-offs:

Post-Norm Architecture. This was the original Transformer design, which applies normalization after the residual connection:

$$\boldsymbol{h}^{(l+1)} = \text{Norm}(\boldsymbol{h}^{(l)} + \text{Sublayer}(\boldsymbol{h}^{(l)})) \tag{5}$$

While its direct normalization of the output path can preserve strong representational fidelity, the gradients must pass through a normalization layer at every block. This can impede gradient flow in very deep networks, making them harder to train. However, Post-Norm does not lead to network depth degradation and exhibits stronger representational capabilities.

Pre-Norm Architecture. This design, now widely adopted, applies normalization before the sublayer, within the residual branch:

$$\Delta \boldsymbol{h}^{(l)} = \text{Sublayer}(\text{Norm}(\boldsymbol{h}^{(l)})) \tag{6}$$

$$\boldsymbol{h}^{(l+1)} = \boldsymbol{h}^{(l)} + \Delta \boldsymbol{h}^{(l)} \tag{7}$$

Its primary advantage is improved training dynamics. The skip connection path is an uninterrupted, identity-like connection, which ensures smooth gradient flow and makes training deep models significantly easier. However, this design has a critical side effect: since the final output $\boldsymbol{h}^{(l+1)}$ is never re-normalized, the variance—and thus the L2 norm—of the hidden states tends to accumulate across layers. This creates the vulnerability we analyze, especially in multimodal contexts where initial norms are already disparate.

Building upon the architectural components defined in the Preliminaries, we now formalize our central argument: the Pre-Norm architecture, when applied to MLLMs, inherently creates a dynamic imbalance that impairs cross-modal fusion. The issue originates from the standard MLLM paradigm, which injects features from a **pre-trained** vision encoder into a language model. It is an established property that deep Pre-Norm networks, like those used in modern vision encoders, accumulate variance as signals propagate through the layers, resulting in high-norm outputs (Kim et al., 2025). Consequently, when these pre-computed, high-norm visual tokens are introduced into the relatively lower-norm embedding space of the LLM, a significant initial norm disparity is established at the modality interface.

3 Theoretical Analysis of Norm-Induced Decoupling Effect

In this section, we present a theoretical proof that this initial norm imbalance is not a static issue but rather the catalyst for an accelerated geometric divergence between the two modalities, ultimately suppressing the cross-modal attention signal. The full mathematical derivation is provided in the Appendix.

3.1 Analytical Framework and Assumptions

Our proof is predicated on a set of simplifying assumptions that capture the core dynamics of the Pre-Norm architecture:

- **Modality Norm Imbalance**: We analyze two cases: the **imbalanced case** ($k = \frac{\|\boldsymbol{h}_{\text{vis}}\|_2}{\|\boldsymbol{h}_{\text{txt}}\|_2} > 1$) and the ideal **balanced case** ($k = 1$).
- **Uniform Update Magnitude**: Due to the Pre-Norm design, the magnitude of the update vector, $\|\Delta \boldsymbol{h}^{(l)}\|_2$, is decoupled from the input norm $\|\boldsymbol{h}^{(l)}\|_2$. We denote this uniform magnitude as $C^{(l)}$ for a given layer.
- **Consistent Update Geometry**: We assume the update vector $\Delta \boldsymbol{h}$ forms a consistent expected angle, ϕ, with the hidden state $\boldsymbol{h}$ for all tokens within a given layer.
- **Random Rotational Direction**: We assume the direction of the rotational component of the update is drawn from a symmetric distribution over the relevant subspace.

3.2 Asymmetric Angular Velocity and Geometric Divergence

To quantify the rate of directional change, we introduce the concept of **effective angular velocity**. The update vector $\Delta \boldsymbol{h}$ can be decomposed into a component parallel to the hidden state $\boldsymbol{h}$ (which only scales its length) and a component orthogonal to it (which causes rotation). The effective angular velocity, measured by the angle of pure rotation θ_{eff}, is driven solely by this orthogonal component. As derived in Appendix B, its tangent is given by:

$$\tan(\theta_{\text{eff}}) = \frac{C^{(l)} \sin(\phi)}{\|\boldsymbol{h}\|_2 + C^{(l)} \cos(\phi)} \tag{8}$$

A direct and critical consequence of our framework is that this angular velocity becomes asymmetric in the imbalanced case. Because a uniform update magnitude $C^{(l)}$ is applied to hidden states of

different norms, the high-norm vision tokens exhibit a lower effective angular velocity than the low-norm text tokens. Formally, for $\|\boldsymbol{h}_{\text{vis}}\|_2 > \|\boldsymbol{h}_{\text{txt}}\|_2$, it follows that:

$$\tan(\theta_{\text{eff, vis}}) < \tan(\theta_{\text{eff, txt}}) \tag{9}$$

This disparity imparts a higher "representational inertia" to visual tokens. In Appendix B, we rigorously prove that this asymmetry leads to an accelerated geometric divergence between the representations of the two modalities, which in turn weakens the underlying similarity signal available to the attention mechanism.

3.3 Suppression of the Cross-Modal Attention Signal

This weakened geometric signal fundamentally limits the attention mechanism's ability to learn an effective similarity metric. The attention mechanism learns a metric based on the dot product between queries and keys; if the foundational similarity between these vectors is systematically eroded layer by layer due to geometric divergence, the gradient signal for learning this metric becomes weaker and noisier.

As rigorously detailed in the Appendix, this results in a systematically suppressed final attention score. Let S_{imb} and S_{bal} denote the unnormalized attention scores in the imbalanced and balanced cases, respectively. We conclude that their expected values are related by:

$$\mathbb{E}[S_{\text{imb}}] < \mathbb{E}[S_{\text{bal}}] \tag{10}$$

This provides a formal, first-principles explanation for the experimentally observed phenomenon of poor cross-modal fusion. The norm imbalance creates a vicious cycle: it accelerates geometric divergence, which weakens the gradient signal for learning the attention metric, leading to a less effective metric and, ultimately, lower cross-modal attention.

4 Empirical Validation: Probing the Dynamics of Norm Imbalance

Our theoretical analysis provides a formal, first-principles explanation for how norm imbalance can impair multimodal fusion. However, this framework relies on a set of simplifying assumptions to ensure analytical tractability, while the dynamics of large-scale MLLMs are considerably more complex. Therefore, to bridge the gap between our idealized model and real-world behavior, we conduct a series of empirical investigations. These experiments are designed to probe whether the core consequences predicted by our theory—namely, the persistence of norm imbalance and the resulting asymmetric update dynamics—manifest in state-of-the-art Pre-Norm MLLMs. Our investigation is guided by the following research questions:

- **RQ1: Existence of Initial Norm Disparity.** Do visual and text tokens exhibit a significant norm mismatch at the modality interface?

  To answer this, we benchmarked the L_2 norms from both sides of the modality interface. For the visual modality, we measured the output norms of four representative vision encoders—CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), SigLIP-v2 (Tschannen et al., 2025), and MoonViT (Team et al., 2025). For the text modality, we established a baseline by computing the average L_2 norm of the text embedding layers from prominent LLMs: Qwen2.5 (Bai et al., 2025), Qwen3 (Yang et al., 2025), and Llama3.2 (Grattafiori et al., 2024). The analysis of vision encoders was conducted on a dataset of 1000 samples drawn from the MMBench, POPE, and MM-Star benchmarks, which serves as the foundation for all subsequent experiments in this section. The combined results are presented in Table 1.

  As shown in 1, vision encoder output norms are substantially larger than those of text embeddings. This disparity persists because the encoders' contrastive pre-training—even with a final post-norm—is not designed to align with the norm scale of an external LLM's embedding space.

- **RQ2: The Efficacy of the Adapter.** Does the projection adapter harmonize the initial norm disparity before the tokens enter the LLM backbone?

  Within MLLMs, the projector's role is to map visual tokens into the LLM's textual embedding space. A critical question is whether this process also serves to align their norms. To investigate

Table 1: L_2 norms and hidden dimensions at the modality interface: Vision Encoder outputs vs. LLM text embeddings (mean $\pm$ std).

Modality	Model	Dimension	Average L_2 Norm
Visual	CLIP-ViT-large-patch14	1024	29.30
	SigLIP-SO-400m-patch14-384	1152	71.78
	SigLIP2-SO-400m-patch14-384	1152	59.37
	MoonViT-SO-400M	1152	72.17
Text	Qwen2.5-7B-Instruct	3584	0.80
	Qwen3-8B-Instruct	4096	1.38
	Llama3.2-3B-Instruct	3072	1.09

this, we analyzed a suite of prominent models: LLaVA-v1.5 (Li et al., 2024a), Qwen-2.5-VL (Bai et al., 2025), KimiVL (Team et al., 2025), and GLM-4.1V (Hong et al., 2025). For each model, we measured the L_2 norm of visual tokens both before and after the projector and compared them to the text token norm. The results are summarized in Table 2.

Table 2: L_2 norms of visual tokens (before and after projector) vs. text tokens.

Model	Visual (Before Proj.)	Visual (After Proj.)	Text (Embedding)
LLaVA-v1.5	39.96	39.96	1.08
Qwen-2.5-VL	3484.24	56.88	0.86
KimiVL	137.93	4.78	0.85
GLM-4.1V	47.44	4.58	0.80

The results in Table 2 reveal a clear spectrum of effectiveness across different projector designs. While sophisticated projectors like those in KimiVL and GLM-4.1V demonstrate a significant capability for norm compression, a substantial disparity between visual and text token norms persists in all analyzed models. This varied effectiveness highlights a key finding: simply inserting a normalization layer within the projector is not a guaranteed solution. With the exception of LLaVA-v1.5, all other models incorporate internal norm layers, yet their final output norms differ by an order of magnitude.

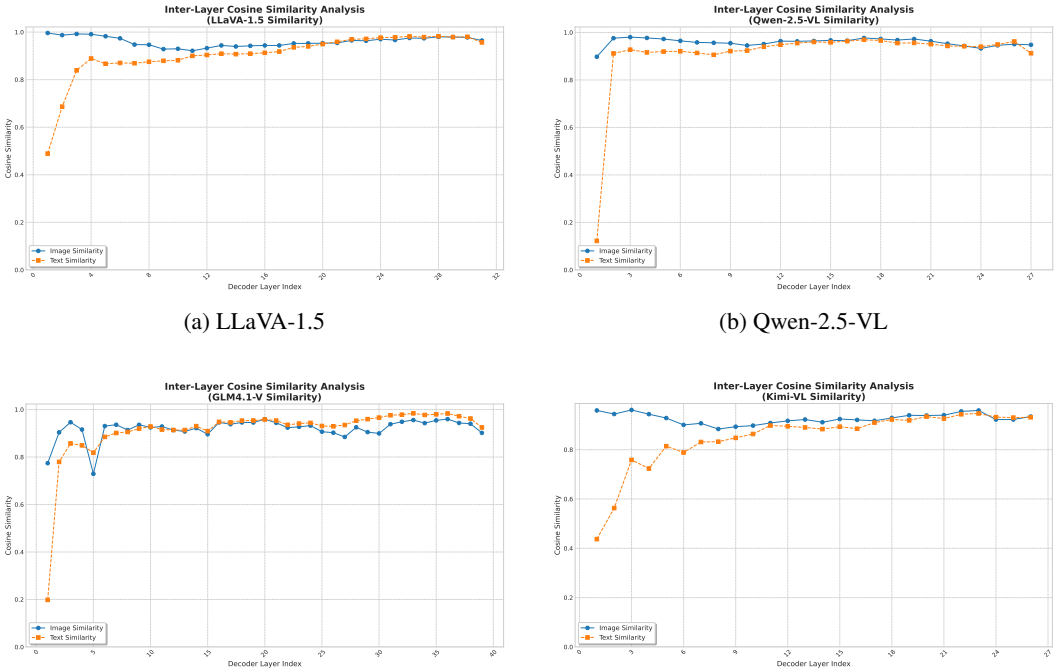
This leads to a broader discussion on current design practices. We note that these architectural choices and their impact on cross-modal norm alignment are seldom, if ever, addressed in the models' respective technical reports.

Notably, the multi-modal training increase in the text embedding norm of Qwen-2.5-VL corroborates that the norm discrepancy is such a fundamental issue that the model inefficiently adjusts static parameters to passively compensate.

- **RQ3: Asymmetry in Update Dynamics.** Do visual and textual hidden states exhibit different update rates, as predicted by our theory of asymmetric angular velocity?

  This question serves as the most direct empirical test of our theory's core mechanism. We use the cosine similarity between consecutive layers ($l-1$ and l) as a proxy for the rate of representational change, a metric conceptually linked to angular velocity. A higher similarity score implies a smaller angular change and thus a slower update rate. We computed this metric for both modalities across all layers to determine if a systematic divergence in their update rates exists, as shown in Figure 2.

  The results in Figure 2 confirm our theoretical predictions, revealing a consistent divergence in update rates between visual and text tokens across all analyzed models. Notably, the magnitude of this dynamic asymmetry appears to be directly correlated with the initial norm disparity identified in RQ1 and RQ2. Models with a smaller initial norm gap, such as Kimi-VL and GLM-4.5V, exhibit a less pronounced difference in update rates. Conversely, models with a more severe norm imbalance, like LLaVA-1.5 and Qwen-2.5-VL, demonstrate a significantly larger gap in their update dynamics, providing strong correlational evidence for our theory.



(a) LLaVA-1.5

(b) Qwen-2.5-VL

(c) GLM4.1-V

(d) Kimi-VL

Figure 2: Inter-layer cosine similarity of hidden states for visual vs. text tokens.

5 Experiments

Our theoretical analysis in Section 3 posited a mechanism whereby norm disparity leads to update asymmetry and, consequently, suppressed visual attention. A critical question remains, however: do these internal dynamics translate into a tangible degradation of the model's downstream capabilities? To investigate this link between internal mechanics and practical performance, we conducted a series of comparative experiments.

5.1 Method: Norm Alignment

To enforce norm alignment between visual and text tokens, we introduce a straightforward intervention. Our approach involves inserting an additional LayerNorm layer immediately after the projector for the visual tokens. Crucially, the learnable gain parameter of this new LN layer is initialized to match the average L_2 norm of the text tokens at the input of the LLM.

To achieve this, we first compute the target L_2 norm, T, by averaging the norms of all non-zero vectors from the language model's text embedding matrix, $\mathbf{W}_e$:

$$T = \frac{1}{|\mathcal{W}^*|} \sum_{\mathbf{w} \in \mathcal{W}^*} \|\mathbf{w}\|_2, \quad \text{where } \mathcal{W}^* = \{\mathbf{w} \in \mathbf{W}_e \mid \|\mathbf{w}\|_2 > \epsilon\} \tag{11}$$

This target norm T is then used to initialize the gain $\mathbf{g}$ and shift $\boldsymbol{\beta}$ parameters of the additional LayerNorm layer. The shift is set to zero, while the gain is uniformly initialized to a scalar value, g_{scalar}, calculated as:

$$g_{\text{scalar}} = \frac{T}{\sqrt{D}}, \quad \text{and} \quad \boldsymbol{\beta}_{\text{init}} = \mathbf{0} \tag{12}$$

where D is the hidden size of the large language model.

5.2 Experimental Setup

Our experiments are conducted within the LLaVA-1.5 architectural framework. Specifically, we employ Llama-3.2-3B-Instruct as the base language model and SigLIP-SO400M-Patch14-384 as the vision encoder. Further details are provided in Appendix C.

A detailed list of the evaluation benchmarks is provided in the Appendix C; for all tasks, we employed a greedy decoding strategy.

5.3 Results and Analysis

5.3.1 Main Performance Gains

The results, summarized in Table 3, underscore the profound impact of our norm alignment strategy. The model equipped with our method consistently outperforms the baseline across a wide array of multimodal tasks. Notably, it also shows marked improvement on pure text evaluations like MMLU (+8.02) and HellaSwag. This latter finding is particularly significant, as it suggests that rectifying the cross-modal dynamic imbalance does not merely improve feature fusion but also leads to a more robust and holistically capable language model, likely by freeing up model capacity that was previously spent compensating for the norm disparity.

Table 3: Performance comparison on various benchmarks. Our method with Norm Alignment consistently outperforms the baseline.

Model	**MMBench$_{dev}$**	**MM-Star**	**POPE**	**SEED-Bench-2**	**OCRBench**
w/o Norm	71.39	37.72	88.14	42.86	40.70
w/ Norm	**72.16** (+0.77)	**41.19** (+3.47)	**88.88** (+0.74)	**47.26** (+4.40)	**45.60** (+4.90)
	ScienceQA	**AI2D**	**HellaSwag**	**MMLU**	**Avg**
w/o Norm	78.99	60.17	65.96	45.19	59.01
w/ Norm	**80.83** (+1.84)	**63.24** (+3.07)	**66.01** (+0.05)	**53.21** (+8.02)	**62.62** (+3.61)

We visualized the attention matrices in Appendix D. The analysis reveals that in the baseline model, text-to-image attention is inappropriately and broadly concentrated on the bottom regions of the image. This suggests a failure in semantic fusion, caused by the positional proximity bias introduced by RoPE's distance-decay property. In stark contrast, our norm-aligned model's text-to-image attention correctly converges on the specific image regions that are semantically relevant to the text query. This visual evidence provides direct confirmation that our method successfully restores meaningful cross-modal attention by correcting the underlying dynamic imbalance, thus enabling true feature fusion.

5.3.2 Ablation Study: The Critical Role of Initialization

To isolate the effect of our proposed initialization strategy, we conducted a crucial ablation study. We compared our method against a baseline where the added LayerNorm layer was initialized with default parameters (gain=1, bias=0). We analyzed the learned parameters immediately after the LLaVA Stage 1 pre-training phase. As shown in Table 4, the parameters of the default-initialized layer remained largely unchanged from their initial state, indicating that the optimization process failed to begin effectively without a reasonable starting point. In contrast, our method shows meaningful parameter updates even after this initial stage. This demonstrates that simply adding a norm layer is insufficient; our targeted initialization is essential to place the parameters in a gradient-rich region of the loss landscape, enabling effective learning.

5.3.3 Diagnostic Analysis: Verifying the Mechanism of Improvement

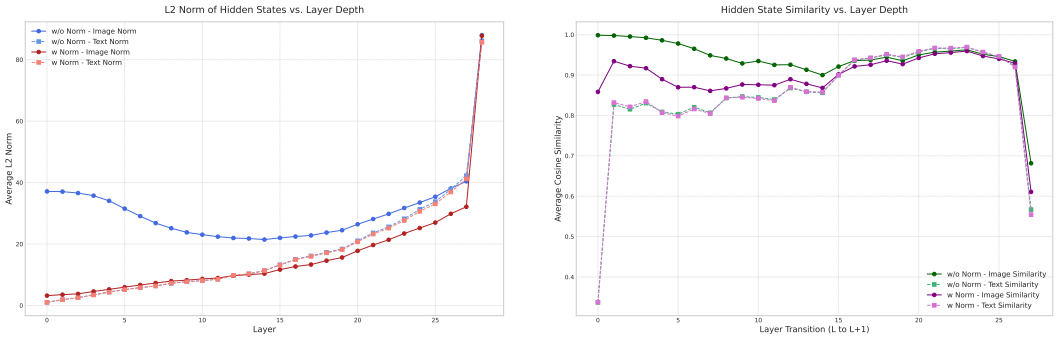
Finally, we performed a diagnostic analysis to confirm that the performance gains are indeed rooted in the successful mitigation of the dynamic imbalance we identified. We analyzed the internal states of the fully trained model (after Stage 2) with and without our norm alignment method. Figure 3 visualizes two key metrics:

Table 4: Learned parameters of the added LayerNorm layer after Stage 1 pre-training, comparing default initialization with our proposed strategy.

Parameter	Metric	Default Init (After Stage 1)	Our Init (After Stage 1)
Gain (g)	L_2 Norm	53.2500	2.2812
	Mean of Abs.	0.9609 ($\pm$ 0.0005)	0.0400 ($\pm$ 0.0001)
Bias (β)	Mean of Abs.	0.0175 ($\pm$ 0.0002)	0.0152 ($\pm$ 0.0001)

- **Layer-wise L2 Norms (Fig. 3a):** The left panel shows that our method successfully aligns the visual token norms with the text token norms from the very first layer and maintains this alignment throughout the model's depth. The baseline model, in contrast, exhibits a persistent and large norm gap.
- **Inter-layer Cosine Similarity (Fig. 3b):** The right panel demonstrates the direct consequence of this alignment. In our model, the update rates (proxied by cosine similarity) of visual and text tokens are nearly identical. This resolves the asymmetric update dynamic present in the baseline, where the high similarity of visual tokens indicates their slower "representational inertia."

Together, these results provide strong evidence that our method works precisely as intended: it corrects the norm disparity, which in turn fixes the asymmetric update rates, leading to the observed performance improvements.



(a) Caption for the left image (e.g., Layer-wise Norms).

(b) Caption for the right image (e.g., Inter-layer Cosine Similarity).

Figure 3: A comparison of token dynamics with and without our norm alignment method. (a) shows the layer-wise L2 norm evolution, while (b) shows the inter-layer cosine similarity, which acts as a proxy for update rate.

6 CONCLUSION

Our analysis reveals a critical, previously undiscovered dynamic within Pre-Norm MLLMs: an "asymmetric update." We have formalized this dynamic theoretically and validated it empirically, showing it to be a direct consequence of the severe norm disparity between visual and text tokens. This analysis demonstrates that the dynamic manifests as "representational inertia" in high-norm visual tokens, fundamentally impairing cross-modal fusion at an architectural level. It was this deep analysis of the mechanism that motivated our targeted solution of enforcing norm alignment via a single LayerNorm. The resulting significant performance gains on both multimodal and, critically, text-only tasks, serve as compelling validation for our core analysis, confirming that resolving this dynamic imbalance unlocks the model's full potential.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

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

Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, et al. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*, 2024.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13817–13827, 2024.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yingjie Fu, Bozhou Li, Linyi Li, Wentao Zhang, and Tao Xie. The first prompt counts the most! an evaluation of large language models on iterative example-based code generation. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):1583–1606, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pp. arXiv–2507, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Jeonghoon Kim, Byeongchan Lee, Cheonbok Park, Yeontaek Oh, Beomjun Kim, Taehwan Yoo, Seongjin Shin, Dongyoon Han, Jinwoo Shin, and Kang Min Yoo. Peri-ln: Revisiting normalization layer in the transformer architecture. *arXiv preprint arXiv:2502.02732*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024b.

Bozhou Li and Wentao Zhang. Id-align: Rope-conscious position remapping for dynamic high-resolution adaptation in vision-language models. *arXiv preprint arXiv:2505.21465*, 2025.

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 *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b.

Zheng Liu, Hao Liang, Bozhou Li, Tianyi Bai, Wentao Xiong, Chong Chen, Conghui He, Wentao Zhang, and Bin Cui. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*, 2024c.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? *arXiv preprint arXiv:2502.05173*, 2025.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International conference on machine learning*, pp. 10524–10533. PMLR, 2020.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

Zhijian Zhuo, Yutao Zeng, Ya Wang, Sijun Zhang, Jian Yang, Xiaoqing Li, Xun Zhou, and Jinwen Ma. Hybridnorm: Towards stable and efficient transformer training via hybrid normalization. *arXiv preprint arXiv:2503.04598*, 2025.

A Background & Related work

A.1 Multimodal Large Language Models

The remarkable success and emergent capabilities of Large Language Models (LLMs) in natural language processing have catalyzed efforts to generalize their powerful abilities to other modalities (Achiam et al., 2023; Hurst et al., 2024; Comanici et al., 2025; Fu et al., 2025; Bai et al., 2025; Yang et al., 2025; Wu et al., 2024). In the multimodal domain, this trend has spurred the rapid development of Multimodal Large Language Models (MLLMs).

Early explorations in MLLMs, such as Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023a), primarily relied on the cross-attention mechanism for modality fusion. A subsequent evolution witnessed a paradigm shift toward a simpler and more efficient approach, an approach popularized by LLaVA (Liu et al., 2023) that has now become the undisputed mainstream. The LLaVA-style architecture eschews the complexity of cross-attention in favor of a more direct solution: it employs a simple projection module, typically a Multi-Layer Perceptron (MLP), to map visual token features directly into the LLM's word embedding space. Conceptually, this treats the image as a sequence of special "visual words" prepended to the text input, which are then processed uniformly by the LLM in an auto-regressive manner. The simplicity, scalability, and powerful performance of this paradigm—particularly when combined with visual instruction tuning—have firmly established it as the foundational blueprint for the vast majority of today's advanced MLLMs

Despite the dominance of the LLaVA paradigm, the pursuit of optimal cross-modal fusion remains an active area of research. Investigators continue to experiment with more sophisticated projector designs (Team et al., 2025; Hong et al., 2025; Cha et al., 2024), alternative representation schemes like visual vocabularies (Lu et al., 2024), or deeper fusion strategies (Meng et al., 2024), novel methods for adapting the core architectures of large models for multimodal scenarios (Deng et al., 2025; Wei et al., 2025; Li & Zhang, 2025). Beyond these m, other researchers have approached the challenge from a data-centric perspective (Bai et al., 2024; Liu et al., 2024c).

In line with this latter direction,

A.2 Normalization

Normalization layers are a cornerstone of modern deep learning, designed to stabilize the training process and accelerate model convergence. By re-scaling the distribution of activations between layers, normalization effectively mitigates the internal covariate shift problem and ensures smooth gradient propagation in deep networks. While Batch Normalization (BN) (Ioffe & Szegedy, 2015) was a seminal work in this area, its dependency on batch size makes it less suitable for natural language processing tasks with variable sequence lengths. Layer Normalization (LayerNorm) was therefore introduced, performing normalization along the feature dimension independently of the batch, and it quickly became the standard for Transformer architectures (Vaswani et al., 2017). This paradigm was further refined by RMSNorm, which improves computational efficiency by removing the mean re-centering step while maintaining performance, leading to its widespread adoption in many modern LLMs such as Llama.

A critical design axis in Transformer architectures is the placement of the normalization layer relative to the residual connection, giving rise to the Pre-Norm and Post-Norm paradigms. The original Post-Norm design applies normalization after the residual addition, which can help preserve strong representational fidelity but is often prone to training instability in deep models. In contrast, the Pre-Norm approach places normalization within the residual branch, greatly improving gradient flow and training stability by maintaining a "clean" skip-connection path. This has made it the de facto standard for large-scale language models. However, the Pre-Norm architecture has a well-documented side effect: because the hidden states on the main path are never re-normalized, their L2 norm tends to accumulate and grow with network depth.

Recently, the community has begun to re-evaluate this classic dichotomy, spurring research into alternative placement strategies. Recent works, like Peri-Norm (Kim et al., 2025)and Hybrid-Norm (Zhuo et al., 2025), have begun to explore combining normalization at different points of the residual connection to merge the benefits of both paradigms. These efforts, however, aim to find a universally optimal static design for unimodal models. In contrast, our work takes a diagnostic perspective:

rather than proposing a new general architecture, we are the first to deeply analyze and reveal how the de facto standard Pre-Norm design itself directly induces a destructive dynamic imbalance within the multimodal context.

B Appendix: Detailed Derivation and Proofs

This appendix provides the full mathematical derivation for the claims made in Section 3, arguing from asymmetric velocity to the final suppression of the attention score.

B.1 Step 1: The General Update Model and Effective Angular Velocity

We begin by defining the geometry of a general update. Any update vector $\Delta \boldsymbol{h}$ can be uniquely decomposed into a component parallel to the hidden state $\boldsymbol{h}$, denoted $\Delta \boldsymbol{h}_{\parallel}$, and a component orthogonal to it, $\Delta \boldsymbol{h}_{\perp}$.

$$\Delta \boldsymbol{h} = \Delta \boldsymbol{h}_{\parallel} + \Delta \boldsymbol{h}_{\perp} \tag{13}$$

The new hidden state is $\boldsymbol{h}' = \boldsymbol{h} + \Delta \boldsymbol{h} = (\boldsymbol{h} + \Delta \boldsymbol{h}_{\parallel}) + \Delta \boldsymbol{h}_{\perp}$. Here, $\Delta \boldsymbol{h}_{\parallel}$ only scales the original vector's magnitude, while $\Delta \boldsymbol{h}_{\perp}$ is solely responsible for the change in direction (rotation).

The rotation is caused by the orthogonal component $\Delta \boldsymbol{h}_{\perp}$ acting on the scaled hidden state $(\boldsymbol{h} + \Delta \boldsymbol{h}_{\parallel})$. The tangent of the effective angle of rotation, θ_{eff}, is therefore:

$$\tan(\theta_{\text{eff}}) = \frac{\|\Delta \boldsymbol{h}_{\perp}\|_2}{\|\boldsymbol{h} + \Delta \boldsymbol{h}_{\parallel}\|_2} \tag{14}$$

Under our **Consistent Update Geometry** assumption, the angle ϕ between $\Delta \boldsymbol{h}$ and $\boldsymbol{h}$ is consistent, which implies $\|\Delta \boldsymbol{h}_{\perp}\| = \|\Delta \boldsymbol{h}\| \sin(\phi)$ and $\|\Delta \boldsymbol{h}_{\parallel}\| = \|\Delta \boldsymbol{h}\| \cos(\phi)$ (assuming ϕ is acute). Substituting this and the **Uniform Update Magnitude** $\|\Delta \boldsymbol{h}\| = C^{(l)}$, we get:

$$\tan(\theta_{\text{eff}}) = \frac{C^{(l)} \sin(\phi)}{\|\boldsymbol{h}\|_2 + C^{(l)} \cos(\phi)} \tag{15}$$

This is the general formula for the effective angular velocity. For visual and text tokens:

$$\tan(\theta_{\text{eff, vis}}) = \frac{C^{(l)} \sin(\phi)}{\|\boldsymbol{h}_{\text{vis}}^{(l)}\|_2 + C^{(l)} \cos(\phi)} \tag{16}$$

$$\tan(\theta_{\text{eff, txt}}) = \frac{C^{(l)} \sin(\phi)}{\|\boldsymbol{h}_{\text{txt}}^{(l)}\|_2 + C^{(l)} \cos(\phi)} \tag{17}$$

Since $\|\boldsymbol{h}_{\text{vis}}^{(l)}\|_2 > \|\boldsymbol{h}_{\text{txt}}^{(l)}\|_2$, the denominator for the visual token is strictly larger. Therefore, the core asymmetry is proven: $\tan(\theta_{\text{eff, vis}}) < \tan(\theta_{\text{eff, txt}})$.

B.2 Step 2: Proof of Recursive Similarity Decay (Theorem 1)

The evolution of cosine similarity is governed by the effective angular velocities.

Theorem 1: Recursive Decay of Cross-Modal Similarity.

> The expected cosine similarity evolves according to $\mathbb{E}[\cos(\Theta^{(l+1)}) \mid \dots] = \gamma_{\text{eff}}^{(l)} \cdot \cos(\Theta^{(l)})$, where the effective decay factor is $\gamma_{\text{eff}}^{(l)} = \cos(\theta_{\text{eff, vis}}^{(l)}) \cos(\theta_{\text{eff, txt}}^{(l)})$.

Proof of Theorem 1. The proof structure is identical to the simpler orthogonal case, as the geometric rotation is driven only by the orthogonal component of the update. Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be the hidden states. The updated unit vector $\hat{\boldsymbol{u}}'$ undergoes an effective rotation $\theta_{\text{eff, u}}$ and can be written as $\hat{\boldsymbol{u}}' = \cos(\theta_{\text{eff, u}})\hat{\boldsymbol{u}} + \sin(\theta_{\text{eff, u}})\hat{\boldsymbol{p}}_u$, where $\hat{\boldsymbol{p}}_u$ is a random direction in the orthogonal subspace. The expectation of the new dot product $\mathbb{E}[\hat{\boldsymbol{u}}' \cdot \hat{\boldsymbol{v}}']$ is computed. The three cross-terms involving random vectors $\hat{\boldsymbol{p}}_u$ and $\hat{\boldsymbol{p}}_v$ vanish in expectation due to the symmetric distribution assumption, leaving only the deterministic term:

$$\mathbb{E}[\cos(\Theta^{(l+1)}) \mid \boldsymbol{u}, \boldsymbol{v}] = \cos(\theta_{\text{eff, u}}) \cos(\theta_{\text{eff, v}})(\hat{\boldsymbol{u}} \cdot \hat{\boldsymbol{v}}) = \gamma_{\text{eff}}^{(l)} \cdot \cos(\Theta^{(l)}) \tag{18}$$

This completes the proof.

■

B.3 Step 3: Proof that Asymmetry Maximizes Decay Rate (Lemma 1)

The lemma is a general mathematical statement about angles and is independent of the underlying model.

Lemma 1: Asymmetry Maximizes Decay Rate.

> For a fixed geometric mean of effective angular velocities, $T = \sqrt{\tan(\theta_{\text{eff},1})\tan(\theta_{\text{eff},2})}$, the decay factor $\gamma_{\text{eff}} = \cos(\theta_{\text{eff},1})\cos(\theta_{\text{eff},2})$ is minimized when $\theta_{\text{eff},1} \neq \theta_{\text{eff},2}$.

Proof of Lemma 1. The proof follows by maximizing the inverse squared of the decay factor, $1/\gamma_{\text{eff}}^2 = (1+\tan^2(\theta_{\text{eff},1}))(1+\tan^2(\theta_{\text{eff},2}))$. Using the AM-GM inequality on the term $\tan^2(\theta_{\text{eff},1})+\tan^2(\theta_{\text{eff},2})$ shows it is minimized in the symmetric case. Thus, $1/\gamma_{\text{eff}}^2$ is minimized, and γ_{eff} is maximized, when the velocities are symmetric. Asymmetry therefore accelerates decay.

■

B.4 Step 4: From Accelerated Divergence to a Suppressed Learned Score

This final step proves that the weaker geometric signal in the norm-imbalanced case necessitates a lower final attention score.

1. From Geometric Divergence to Weaker Input Correlation. First, we establish that the inputs to the attention projections, $\boldsymbol{u} = \text{RMSNorm}(\boldsymbol{h}_{\text{txt}}^{(L)})$ and $\boldsymbol{v} = \text{RMSNorm}(\boldsymbol{h}_{\text{vis}}^{(L)})$, are less correlated in the imbalanced case. From Theorem 1 and Lemma 1, the expected cosine similarity between the final hidden states is systematically lower in the norm-imbalanced scenario. Let Θ_{imb} and Θ_{bal} be the final angles between the hidden states in their respective cases. We have $\mathbb{E}[\cos(\Theta_{\text{imb}})] < \mathbb{E}[\cos(\Theta_{\text{bal}})]$. The inputs to the shared projection matrices $\mathbf{W}_Q$ and $\mathbf{W}_K$ are $\boldsymbol{u} = \sqrt{D}\cdot(\boldsymbol{g}\odot\hat{\boldsymbol{h}}_{\text{txt}})$ and $\boldsymbol{v} = \sqrt{D}\cdot(\boldsymbol{g}\odot\hat{\boldsymbol{h}}_{\text{vis}})$. Their dot product is a positively weighted sum of the component-wise products of the underlying unit vectors: $\boldsymbol{u}\cdot\boldsymbol{v} = D\cdot\sum_{i=1}^{D} g_i^2(\hat{h}_{\text{txt},i}\hat{h}_{\text{vis},i})$. Since the unweighted sum is $\cos(\Theta)$, and the weights are positive, a lower expected cosine similarity directly implies a lower expected dot product between the inputs to the attention mechanism.

$$\mathbb{E}[\boldsymbol{u}\cdot\boldsymbol{v}]_{\text{imb}} < \mathbb{E}[\boldsymbol{u}\cdot\boldsymbol{v}]_{\text{bal}} \tag{19}$$

This rigorously establishes that the foundational geometric signal is weaker in the norm-imbalanced case.

2. The Inescapable Conclusion: Suppressed Scores. The attention mechanism cannot invent correlations where none exist; it can only discover and amplify statistical patterns present in its input data. The statistical object containing all learnable second-order correlation information is the cross-covariance matrix, $\mathbf{C}_{\boldsymbol{uv}} = \mathbb{E}[\boldsymbol{u}\boldsymbol{v}^T]$. A lower expected dot product implies that the trace of this matrix, $\text{Tr}(\mathbf{C}_{\boldsymbol{uv}})$, is smaller, indicating a spectrally weaker matrix. The maximum achievable expected attention score is mathematically bounded by the singular values of this matrix. Since the cross-covariance matrix for the imbalanced case ($\mathbf{C}_{\text{imb}}$) is spectrally weaker than for the balanced case ($\mathbf{C}_{\text{bal}}$), it places a lower mathematical ceiling on the maximum possible attention score the model can learn. The model does its best to find correlation, but there is simply less correlation to be found. Let S_{imb} and S_{bal} denote the final scores. We can thus conclude:

$$\mathbb{E}[S_{\text{imb}}] < \mathbb{E}[S_{\text{bal}}] \tag{20}$$

This completes the proof, showing that the suppressed attention score is a direct mathematical consequence of the impoverished statistical signal caused by the initial norm imbalance.

C Training Details

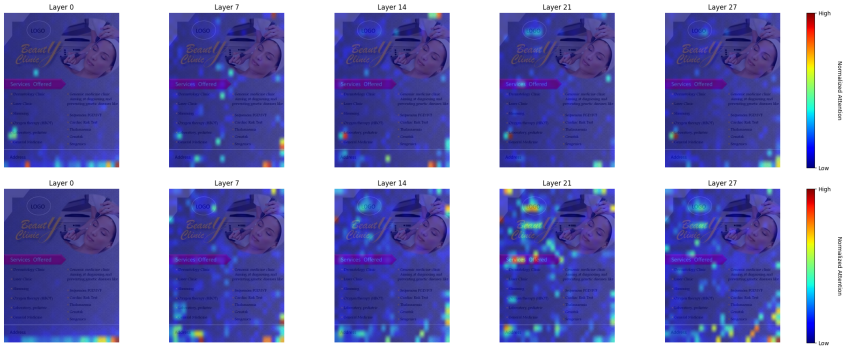
Our experiments are conducted within the LLaVA-1.5 architectural framework. Specifically, we employ Llama-3.2-3B-Instruct as the base language model and SigLIP-So400M-Patch14-384 as the vision encoder. We follow a two-stage training protocol: the first stage consists of one epoch of feature alignment pre-training on the LLaVA-558K dataset, using a learning rate of 1e-3, a per-device batch size of 2, and 2 gradient accumulation steps. This is followed by one epoch of full-model instruction tuning on the LLaVA-NeXT instruction-tuning dataset, for which the learning rate was decreased to 1e-5 for the language model and 2e-6 for the vision encoder, with a per-device batch size of 1 and 4 gradient accumulation steps. Across both stages, we utilized a cosine learning rate scheduler with a warmup ratio of 0.03 and set weight decay to 0. Notably, we do not employ dynamic high-resolution strategies; all images are uniformly resized to 384x384. To ensure reproducibility, we set the random seed to 42 for all experiments.

To comprehensively evaluate the model's performance, we assessed its capabilities on both multimodal and text-only tasks. The model's multimodal abilities were benchmarked against a comprehensive suite of benchmarks, including MMBench-EN (Liu et al., 2024a), MM-Star (Chen et al., 2024b), OCRBench (Liu et al., 2024b), SEED-Bench-2-Plus (Li et al., 2024b), ScienceQA (Lu et al., 2022), AI2D (Kembhavi et al., 2016), and POPE (Li et al., 2023b). Furthermore, to gauge its core language understanding and commonsense reasoning skills, we evaluated its performance on the HellaSwag (Zellers et al., 2019) and MMLU (Hendrycks et al., 2020) benchmarks.

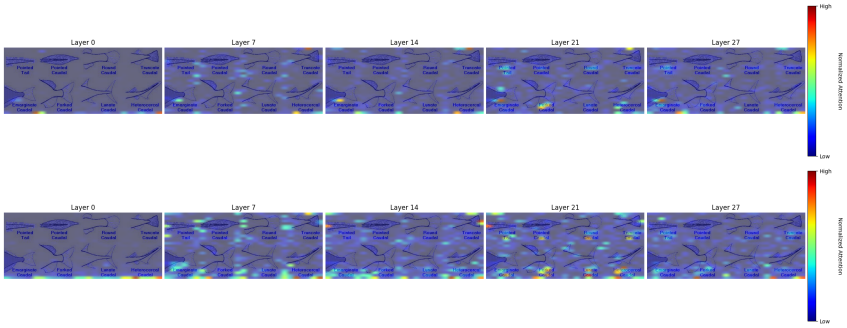
D Appendix: Attention Visualization

In each pair of heatmaps, the bottom image shows the model with norm applied, while the top image shows the baseline model. The caption for each pair corresponds to the text query used.

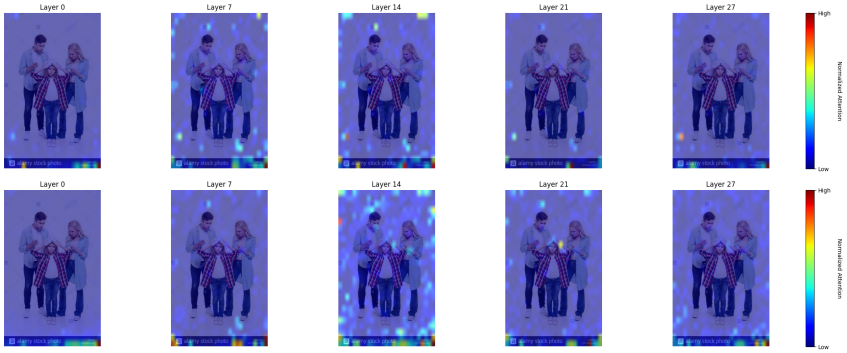
(a) Which mood does this image convey?



(b) What is the main subject of the flyer seen in the image?



(c) How many different pointedkinds are there?



(d) What type of family is shown in the image