Why Vision Fails as a Universal Bridge: Rectifying Modality Asynchrony in Multilingual MLLMs

Yihang Du[1,2,*] **Juhao Liang**[1,*] **Zhengzhao Lai**[1] **Siyu Li**[1] **Yan Hu**[1,3,†]
[1]The Chinese University of Hong Kong, Shenzhen
[2]Shenzhen Loop Area Institute
[3]National Health Data Institute (Shenzhen)
yihangdu@slai.edu.cn, juhaoliang1@link.cuhk.edu.cn, huyan@cuhk.edu.cn
* Equal contribution, † Corresponding author

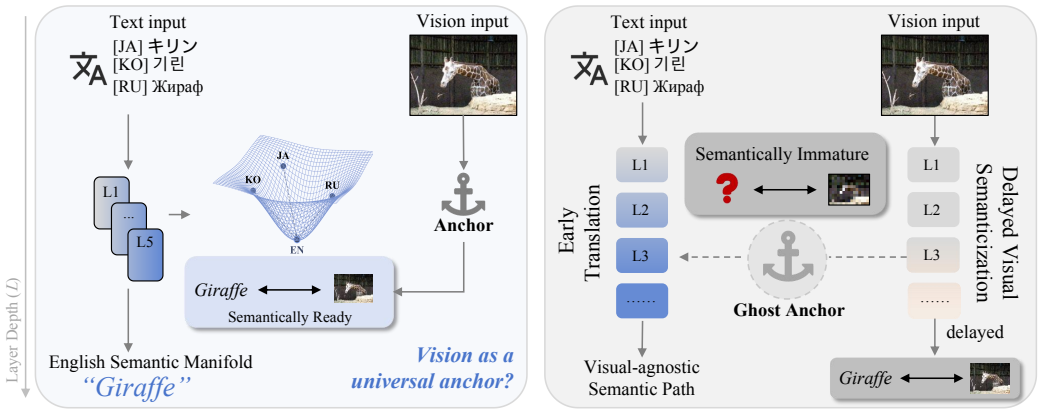
Abstract

Multimodal large language models (MLLMs) exhibit substantial performance degradation in non-English visual reasoning, despite the strong multilingual competence of their text-only backbones. While mechanistic evidence from text-only models suggests that non-English inputs are routed through an English-centric latent space, the multimodal implications of this phenomenon remain unexplored. Through layer-wise mechanistic analysis, we identify the **Ghost Anchor** phenomenon: a temporal modality asynchrony where linguistic representations largely converge toward the English semantic manifold in early layers, while visual semanticization remains immature. Consequently, visual signals are physically present yet exert limited influence during the early alignment window. To address this, we propose **ANCHOR**, a training framework employing Proactive Visual Anchoring (PVA) to accelerate early visual semantic emergence and encourage visual representations to guide linguistic translation. Mechanistic interventions indicate that ANCHOR increases the influence of visual signals during early translation. Furthermore, experiments on xMMMU, MaXM, and CVQA show that ANCHOR improves aggregate performance over standard baselines across both fine-tuned and zero-shot languages.

1 Introduction

Multimodal Large Language Models (MLLMs) have significantly advanced vision-language understanding by integrating LLMs with visual perception (Tong et al., 2026; Bai et al., 2025a). However, while MLLMs excel in English, achieving equitable understanding across diverse languages remains a fundamental challenge (Yue et al., 2024; Pfeiffer et al., 2022). Empirical evaluations reveal a substantial performance degradation in non-English contexts, indicating that current multimodal alignment remains highly English-dependent and fails to ground visual concepts within a truly language-agnostic space (Zhong et al., 2024; Changpinyo et al., 2023).

A prevailing intuition attributes this disparity to data scarcity, motivating efforts to curate massive multilingual multimodal datasets (Dash et al., 2025; Yue et al., 2024). Yet, this data-centric approach incurs prohibitive costs and overlooks the internal representational mechanisms. Instead of forcing alignment externally, an emerging consensus acknowledges that MLLMs inherently develop a shared semantic space where diverse languages converge (Zeng et al., 2025). Ideally, heterogeneous inputs would converge toward a language-neutral semantic space through synchronized cross-modal interactions. In contemporary MLLMs, however, this shared space often empirically manifests as English-dominated because of language-centric architectures and imbalanced pre-training data (Wu et al., 2024). Within this empirically observed geometry, visual signals—grounded in physical reality—should act as active anchors that ground representations across linguistic boundaries (Li et al., 2025; Ni et al., 2021; Zhou et al., 2021).



(a) Ideal: Synchronized Alignment (b) Reality: Modality Asynchrony

Figure 1: **The Ghost Anchor phenomenon.** (a) Ideal synchronized cross-modal alignment. (b) The reality of modality asynchrony: rapid linguistic translation bypasses delayed visual semantic emergence in early layers.

However, the empirical reality is far from this synchronized ideal. Despite the theoretical promise of vision as a universal anchor, visual signals provide limited assistance for cross-lingual alignment: non-English languages continue to exhibit systematic deficits even when visual grounding should theoretically facilitate comprehension (Bugliarello et al., 2022; Zheng et al., 2025). This disconnect raises a critical question: *Why does the visual modality fail to serve as the anticipated bridge?*

We hypothesize that the answer lies in the temporal dynamics of the Transformer architecture. Through a rigorous layer-wise mechanistic analysis, we uncover a profound **Modality Asynchrony**: linguistic representations exhibit substantial convergence toward the English semantic manifold in early layers, while visual grounding remains semantically immature until much deeper in the network. As illustrated in Figure 1, this early linguistic convergence emerges well before visual semantics become available. Consequently, the linguistic stream is forced to map to the English space "blindly", bypassing visual guidance.

Modality Asynchrony creates what we term the **Ghost Anchor** phenomenon: the visual modality is physically present in the computation yet remains semantically invisible during the critical early alignment phase (Figure 1). Consequently, the model defaults to a visual-agnostic translation path, which fundamentally degrades its non-English visual reasoning performance.

To rectify this asynchrony, we propose **ANCHOR**, a training framework designed to synchronize dual-stream convergence. By accelerating visual semantic emergence through Proactive Visual Anchoring (PVA), ANCHOR encourages visual signals to become grounded early and influence the linguistic translation process. Experiments on benchmarks such as xMMMU (Yue et al., 2024), MaXM (Changpinyo et al., 2023), and CVQA (Romero et al., 2024) show that our approach improves cross-lingual visual performance.

In this work, we make the following principal contributions:

- We systematically investigate the internal dynamics of multilingual MLLMs and formalize the **Ghost Anchor** phenomenon, where **Modality Asynchrony** renders visual signals semantically inaccessible during early alignment.
- We provide mechanistic evidence through layer-wise analysis and controlled interventions, showing that visual information exerts negligible causal influence on the linguistic translation trajectory during the early alignment window.
- Building on these insights, we propose **ANCHOR**, a training framework that rectifies asynchrony by accelerating visual semantic emergence through Proactive Visual An-

choring (PVA), which explicitly supervises early-layer visual representations using an external visual foundation model.

2 Related Work

Mechanistic Perspectives on Cross-lingual Alignment Cross-lingual alignment has been primarily studied in text-only multilingual models, where semantically equivalent expressions gradually converge in hidden space and support transfer (Zeng et al., 2025; Zhong et al., 2024). Using logit-lens and causal-tracing style probes, mechanistic analyses further show that many multilingual LLMs process non-English inputs through an English-centric internal representation space, where non-English tokens are implicitly translated into an English semantic pivot before being mapped to the target output (Wendler et al., 2024; Schut et al., 2025; Zeng et al., 2025). At the same time, strong global alignment can hurt language-specific competence, especially for low-resource or distant languages (Elshabrawy et al., 2025). This tension is still analyzed almost entirely without vision, leaving unclear how visual signals reshape (or fail to reshape) cross-lingual alignment inside MLLMs (Wu et al., 2024; Sundar et al., 2025). Our work fills this gap by extending mechanistic analysis to multimodal settings and showing that visual signals can fail exactly when modality asynchrony delays visual semantic emergence during the critical alignment window.

Multilingual Large Vision Language Models Early multilingual vision-language models such as M^3P (Ni et al., 2021), UC^2 (Zhou et al., 2021), and related BERT-era frameworks (Huang et al., 2021) explicitly optimized cross-lingual and cross-modal objectives during pretraining. In the LLM era, mainstream multilingual MLLMs rely more on scale-first pipelines, where large translated or synthetic instruction mixtures are expected to induce multilingual visual competence implicitly (Yue et al., 2024; Dash et al., 2025). This data-centric strategy improves benchmarks but incurs high annotation cost and offers limited control over how multilingual text and visual semantics are aligned internally. More targeted methods add alignment modules or objectives, including mBLIP (Geigle et al., 2024), LRM-LLaVA (Li et al., 2025), and M^2-VLP (Ahmat et al., 2025). However, their supervision is still mainly applied at output or global-distribution levels rather than explicitly constraining layer-wise cross-modal interaction during the critical alignment phase. In contrast, our approach starts from a mechanistic diagnosis of modality asynchrony and directly targets the critical layer-wise interaction window.

3 Unveiling Modality Asynchrony: A Mechanistic Analysis

Do visual signals truly function as language-agnostic anchors in multilingual MLLMs? To answer this, we systematically investigate both the independent evolutionary trajectories of multilingual text and visual signals, and how they interact within the model's internal representations.

The English Semantic Manifold To perform joint reasoning, MLLMs inherently project heterogeneous inputs (e.g., multilingual text and visual signals) into a shared representational space (Song et al., 2025). During inference, representations from both diverse languages (Zeng et al., 2025; Zhao et al., 2024) and cross-modal visual inputs (Wu et al., 2024; Venhoff et al., 2025) progressively converge into a unified geometry. Driven by the language-centric architecture of contemporary MLLMs (Shen et al., 2025; Shu et al., 2025) and imbalanced pre-training data, this shared hub pragmatically manifests as an English-dominated space (Schut et al., 2025; Zhong et al., 2024). Thus, we formalize the English Semantic Manifold ($\mathcal{S}_{en}$) as the target coordinate system where intermediate hidden states h_l progressively converge across layers: $h_l \xrightarrow{\text{layers}} h^* \in \mathcal{S}_{en}$.

Dynamics of Multilingual and Modal Convergence. Let $\mathcal{S}_{src}$ denote the source-language representational space and $\mathcal{S}_{vis}$ the initial perceptual space occupied by visual tokens. Within this unified framework, the recurrence $h_l = f_l(h_{l-1})$ traces an incremental trajectory

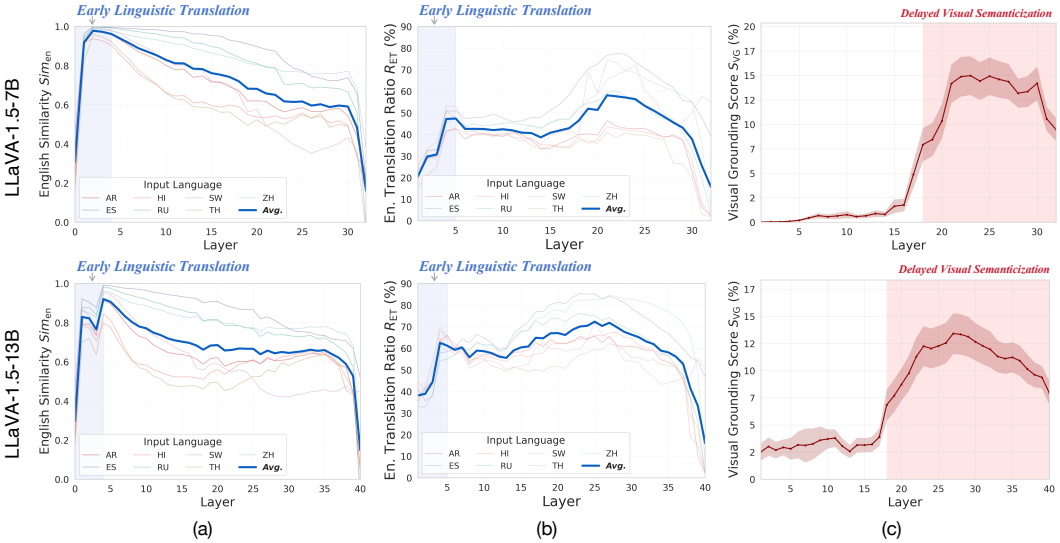


Figure 2: **Modality Asynchrony**. **(a-b)** Layer-wise English Similarity (Sim_{en}) and English Translation Ratio (R_{ET}) in text stream. **(c)** Visual Grounding Score (S_{VG}) in vision stream. Shaded regions indicate the standard error of the mean (SEM).

toward the English semantic manifold, where h_l denotes the hidden state at layer l. We use $F_l = f_l \circ \cdots \circ f_1$ to denote the cumulative transformation through layer l. Both streams exhibit convergence toward $\mathcal{S}_{en}$, yet follow distinct trajectories: the linguistic stream is hypothesized to undergo an implicit transition from $\mathcal{S}_{src}$ toward $\mathcal{S}_{en}$ under F_l, while the visual stream undergoes visual semanticization from $\mathcal{S}_{vis}$ toward $\mathcal{S}_{en}$.

Crucially, for the MLLM to perform true multimodal reasoning, the implicit translation of the text stream should theoretically be grounded in the visual context. Since Transformer architectures aggregate information via layer-wise self-attention, this grounding hinges on the **temporal coordination** between the two paths. If visual tokens have not yet achieved visual semanticization during the specific layer window where translation occurs, the linguistic stream could default to a purely text-based transformation. Such a mechanistic asynchrony would result in a bypass of visual evidence, potentially explaining the reasoning failures observed in multilingual contexts. In the following section, we empirically measure the precise layer-wise timing of these two processes to test this asynchrony hypothesis.

3.1 Modality Asynchrony: Early Translation, Delayed Semanticization

To investigate whether these pathways evolve synchronously, we extract intermediate hidden states h_l under a controlled input condition where visual tokens are paired with multilingual text descriptions. All of the experimental details and metric definitions can be found in Appendix A. We monitor the dual-stream convergence using three complementary metrics. All vocabulary-level projections employ the *logit lens* technique (Nostalgebraist, 2020), which maps intermediate hidden states to probability distributions over the vocabulary via the pre-trained unembedding matrix.

- **Text Stream:** We track the continuous geometric rotation and discrete lexical transition toward $\mathcal{S}_{en}$ using **English Similarity (Sim_{en})** and the **English Translation Ratio (R_{ET})**, respectively. Specifically, Sim_{en} measures the cosine similarity between the mean hidden states of the source and English reference sequences. R_{ET} extracts the top-k decoded words from each text token via the logit lens and measures the proportion classified as English by a zero-shot language classifier.
- **Vision Stream:** We directly probe the visual semanticization process using the **Visual Grounding Score (S_{VG})**, tracking when visual tokens decode into ground-truth object labels via the logit lens.

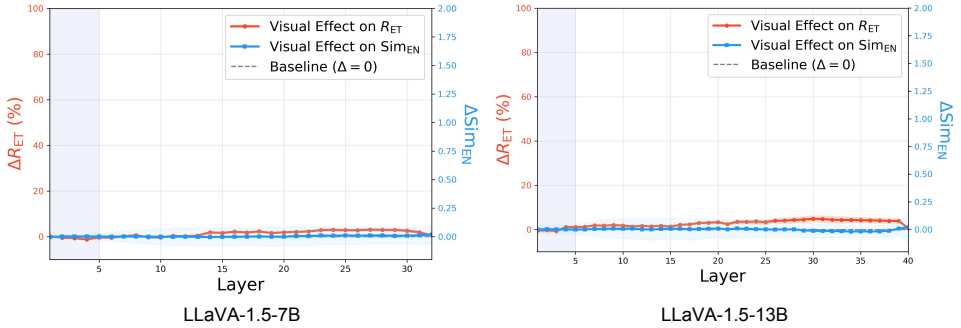


Figure 3: **Visual Causal Effect** (LLaVA-1.5-7B left, 13B right). Difference between Normal and Noise inputs (Δ = Normal − Noise) across layers. Shaded regions indicate the standard error of the mean (SEM).

Finding 1: Early Linguistic Translation. Both Sim_{en} and R_{ET} reveal rapid linguistic convergence within the initial layers. As shown in Figure 2(a-b), source language representations efficiently rotate toward the English representation and decode into English vocabulary as early as layers 1–5. This demonstrates that the English-dominant linguistic transition becomes strongly detectable in the earliest stages of the trajectory.

Finding 2: Delayed Visual Semanticization. In contrast, visual semantics emerge later. Figure 2(c) shows that S_{VG} remains near-zero throughout early-to-middle layers, only rising beyond layer 20. Bridging the cross-modal gap ($h_l : \mathcal{S}_{vis} \rightarrow \mathcal{S}_{en}$) appears more demanding than intra-lingual translation, creating a pronounced Modality Asynchrony.

The Ghost Anchor Phenomenon. The asynchrony suggests a mechanistic misalignment: by the time the linguistic stream settles into $\mathcal{S}_{en}$, visual tokens remain close to their initial perceptual manifold $\mathcal{S}_{vis}$. At this stage, they primarily encode low-level sensory features and provide limited assistance to the translation process. We characterize this temporal separation as the Ghost Anchor phenomenon—visual tokens are physically present but exert limited measurable influence during early alignment. We test this hypothesis through controlled intervention in Section 3.2.

3.2 Causal Validation of the Ghost Anchor

While R_{ET} is already high in early layers (Figure 2(b)), the critical question is whether this translation trajectory is sensitive to visual context or predominantly driven by textual priors. To test this, we replace each visual input with matched Gaussian noise—preserving the input dimensions, patch count, and token positions while substantially disrupting recognizable objects and scene structure (see Appendix A.4). We therefore measure the **Visual Causal Effect** on the text stream via both the continuous geometric shift (ΔSim_{en}) and the discrete lexical transition (ΔR_{ET}). As shown in Figure 3, both metrics remain near-zero across early layers for LLaVA-1.5. These small changes indicate that the measured early translation trajectory has limited sensitivity to high-level visual content under this intervention, providing evidence consistent with the Ghost Anchor phenomenon. We further conduct experiments on a broader range of SOTA MLLMs in Appendix C.1, which show similarly small effects.

4 ANCHOR: Rectifying Modality Asynchrony through Proactive Visual Anchoring

4.1 Method Overview

Motivated by our mechanistic discovery of the Ghost Anchor effect, we propose ANCHOR to structurally synchronize the disparate evolutionary trajectories of text and vision. Unlike

data-centric interventions that force external alignment, ANCHOR explicitly supervises the early-layer evolution of visual representations. Its core component, Proactive Visual Anchoring (PVA), leverages an external visual foundation model (VFM) to enforce visual semantic readiness before the critical alignment window closes, ensuring visual features actively guide linguistic translation.

4.2 Proactive Visual Anchoring

Standard vision-language training treats the vision encoder as a fixed feature extractor and only adapts the language side. However, under text-only supervision, fine-grained visual information may not be preserved as visual tokens propagate through the LLM's early layers. By the time the semantic verbalization process completes, the cross-modal alignment window has already closed. PVA therefore uses an external VFM as a semantic target to encourage visually grounded representations in the early layers.

Alignment Objective. Given the MLLM's intermediate hidden states $h_{i,l} \in \mathbb{R}^{d_{\text{MLLM}}}$ for visual token i at layer l, we pass them through a lightweight alignment network $\text{MLP}_{align}(\cdot)$ and compare the output against the semantic supervisor from a pre-trained VFM. Following the formulation in prior representation alignment work, the Proactive Visual Anchoring loss is defined as the negative mean cosine similarity across all visual tokens:

$$\mathcal{L}_{anc} = -\frac{1}{|\mathcal{V}_{vis}|} \sum_{i \in \mathcal{V}_{vis}} \cos\big(\text{MLP}_{align}(h_{i,L_{early}}), z_{i,VFM}\big), \tag{1}$$

where $|\mathcal{V}_{vis}|$ denotes the number of visual tokens, and $z_{i,VFM} \in \mathbb{R}^{d_{\text{VFM}}}$ is the fine-grained semantic feature extracted from SigLIP-SO400M/14 (Tschannen et al., 2025) for visual token i. The alignment network $\text{MLP}_{align} : \mathbb{R}^{d_{\text{MLLM}}} \to \mathbb{R}^{d_{\text{VFM}}}$ consists of a two-layer perceptron with GELU activation that projects the MLLM hidden states into the VFM feature space.

The key design choices are: (1) we apply supervision at layer L_{early}, where visual semantics should become available for cross-modal alignment; and (2) we use a VFM (SigLIP-SO400M/14) as the supervisor because it provides visually grounded semantic features that can serve as a shared reference across target languages.

Combined Training Objective. The full training objective combines the standard visual instruction tuning loss with the PVA anchoring loss:

$$\mathcal{L} = \mathcal{L}_{SFT} + \alpha \cdot \mathcal{L}_{anc}, \tag{2}$$

where $\mathcal{L}_{SFT}$ is the standard next-token prediction loss on multilingual image-text pairs, and α controls the strength of the anchoring regularization. During training, we apply full-parameter fine-tuning (FFT) to both the multimodal projector and the first L_{early} transformer layers, freeze the remaining layers and the vision encoder, and train the alignment network MLP_{align}. The ablation in Section C.3 suggests that restricting the intervention to this early window better preserves downstream performance than unconstrained full-depth updates.

By supervising the early-layer trajectory of visual tokens, PVA effectively shifts the visual semantic emergence curve earlier in the model's depth, ensuring that when the linguistic stream begins its cross-modal alignment phase, grounded visual evidence is already available and can actively influence the alignment process.

5 Empirical Results and Analysis

To empirically validate the ANCHOR framework, we structure our evaluation around three core research questions: **RQ1** Multilingual Performance: Does ANCHOR improve visual reasoning across fine-tuned and zero-shot languages without degrading native English proficiency? **RQ2** Mechanistic Rectification: Does the framework mitigate the Ghost Anchor phenomenon by increasing the influence of visual signals during early translation? **RQ3** Intervention Design: Does explicit PVA supervision during the early alignment window improve performance over unanchored early-layer adaptation?

(a) Results on the xMMMU

Size	Models	en	SFT Languages 🔥				Zero-shot ❄		Avg.
			ar	fr	hi	pt	id	ja	
7B	LLaVA-1.5	**35.2**	29.9	34.6	28.2	33.7	31.6	32.0	32.2
	+ Std. LoRA	34.9	32.9	36.6	30.2	36.0	32.0	30.9	33.4
	+ Early-Layer FFT	32.6	28.9	32.2	24.1	31.6	**34.3**	32.0	30.8
	+ ANCHOR	35.0	**34.8**	**39.1**	**32.6**	**39.2**	33.9	**34.4**	**35.6**
13B	LLaVA-1.5	34.8	30.9	35.6	29.2	37.0	33.7	30.9	33.2
	+ Std. LoRA	34.8	31.5	35.6	30.2	36.0	33.0	31.2	33.2
	+ Early-Layer FFT	33.2	31.2	35.2	29.6	33.7	30.6	28.3	31.7
	+ ANCHOR	**35.1**	**36.3**	**37.5**	**34.2**	**37.3**	**35.2**	**33.1**	**35.5**

(b) Results on the MaXM

Size	Models	en	SFT Languages 🔥				Zero-shot ❄		Avg.
			fr	hi	th	zh	he	ro	
7B	LLaVA-1.5	49.4	32.6	17.7	17.2	27.8	12.9	15.1	24.7
	+ Std. LoRA	49.0	33.0	37.7	30.2	30.3	16.1	**18.3**	30.7
	+ Early-Layer FFT	40.5	30.7	32.3	26.9	**36.5**	**19.3**	16.9	29.0
	+ ANCHOR	**51.3**	**36.8**	**38.4**	**38.2**	29.3	14.7	**18.3**	**32.4**
13B	LLaVA-1.5	51.8	30.3	15.8	20.1	22.0	14.3	23.6	25.4
	+ Std. LoRA	48.6	34.9	38.3	32.8	30.3	16.1	26.8	32.5
	+ Early-Layer FFT	36.6	35.2	33.5	34.0	**40.0**	13.6	**35.6**	32.6
	+ ANCHOR	**54.1**	**39.1**	**39.3**	**37.2**	36.9	**20.5**	28.1	**36.5**

Table 1: Performance comparison on multilingual multimodal benchmarks: (a) xMMMU and (b) MaXM. Best results are highlighted in **bold**.

5.1 Experiment Settings

Implementation Details We implement ANCHOR using LLaVA-1.5 as our base MLLM, a classic architecture that is widely used for interpretability research. For PVA, we employ SigLIP-SO400M/14 as the external Visual Foundation Model. Following our mechanistic analysis, we set $L_{early} = 10$ as the early alignment window. Models are fine-tuned on 135K image-text pairs (15K × 9 languages) from ShareGPT4V (Chen et al., 2024). See Section B.1 for dataset details.

Baselines To isolate the efficacy of PVA, we compare ANCHOR against two baselines: (1) **Std. LoRA**, which fine-tunes without any visual-side intervention; and (2) **Early-Layer FFT**, which applies full-parameter fine-tuning to the first L_{early} transformer layers but omits the PVA objective.

Benchmarks We evaluate multilingual proficiency across three benchmarks targeting distinct dimensions: (1) **xMMMU** (Yue et al., 2024) for deep semantic reasoning in expert domains; (2) **MaXM** (Changpinyo et al., 2023) for cross-lingual transfer on low-resource languages; and (3) **CVQA** (Romero et al., 2024) for culturally nuanced visual mappings. Crucially, our evaluation covers both languages included in multilingual SFT and languages unseen during training, enabling us to assess generalization across both seen and zero-shot settings.

5.2 Quantitative Results

To systematically assess the macroscopic generalization of our framework, we evaluate ANCHOR on xMMMU, MaXM, and CVQA. Results across all evaluations consistently exhibit three core trends that corroborate our hypotheses:

Size	Models	SFT Languages 🔥			Zero-shot ❄			Avg.
		ar-EG	es-ES	zh-CN	am-ET	ja-JP	ro-RO	
7B	LLaVA-1.5	32.5	67.0	48.6	26.5	35.0	**51.0**	**43.4**
	+ Std. LoRA	31.5	62.0	44.7	26.1	35.5	46.7	41.1
	+ Early-Layer FFT	29.6	50.3	42.8	**27.4**	35.5	44.4	38.3
	+ ANCHOR	**37.3**	**69.3**	**49.1**	26.4	**36.9**	37.4	42.7
13B	LLaVA-1.5	38.4	70.4	54.3	25.2	45.8	57.0	48.5
	+ Std. LoRA	36.0	67.9	50.5	26.1	42.4	52.3	45.9
	+ Early-Layer FFT	28.1	52.2	41.8	**36.8**	37.0	44.7	40.1
	+ ANCHOR	**40.1**	**70.8**	**59.4**	26.0	**47.5**	**58.0**	**50.3**

Table 2: Performance on CVQA. We show six language–region pairs whose languages include both those used in SFT and those unseen during training; full results are reported in Table 8. The Avg. column reports the average accuracy across the six displayed language–region pairs. Best results are highlighted in **bold**.

Retention of the English Semantic Manifold. A persistent challenge in cross-lingual alignment is degradation of the backbone's native English proficiency. Our evaluations indicate that ANCHOR generally preserves—and in certain cases, enhances—this capability. For instance, while maintaining stable English performance on xMMMU, ANCHOR yields improvements on MaXM. These results suggest that early-layer visual grounding via PVA can integrate visual anchors without substantially interfering with the LLM's established English-centric reasoning pathways.

Substantial Gains in Target SFT Languages. For non-English languages explicitly encountered during fine-tuning, ANCHOR demonstrates pronounced improvements. Compared with the degradation observed for the Early-Layer FFT baseline, ANCHOR better preserves performance while updating early layers. For example, on the 13B CVQA benchmark, standard early-layer adaptation reduces Arabic accuracy from 38.4% to 28.1%; in contrast, ANCHOR improves it to 40.1% and raises the average across the six displayed language–region pairs to 50.3%, surpassing both the base model and standard LoRA on this displayed subset. This contrast supports our mechanistic premise and indicates that explicit visual anchoring is beneficial when adapting early layers.

Robust Zero-Shot Cross-Lingual Generalization. The benefits of mitigating modality asynchrony extend to unseen linguistic spaces, reducing the cross-lingual "alignment tax" often incurred during SFT. When transferring to zero-shot languages, standard baselines can degrade because of overfitting to the target SFT distribution. ANCHOR cushions this degradation on CVQA and achieves absolute zero-shot improvements on xMMMU (e.g., zero-shot Japanese improving from 32.0% to 34.4% on the 7B model). These results suggest that mitigating the Ghost Anchor phenomenon can produce a more transferable cross-modal alignment space.

Generalization to Newer MLLMs. To examine whether ANCHOR extends beyond LLaVA-1.5, we additionally evaluate it on LLaVA-NeXT (7B/13B) and InternVL3 (8B/14B), using the same training and evaluation settings as the corresponding Std. LoRA baselines. ANCHOR consistently improves over Std. LoRA across the English, SFT-language, and zero-shot-language groups. Detailed gains on xMMMU, MaXM, and CVQA are reported in Tables 5 to 7.

5.3 Mechanistic Analysis: Validating Ghost Anchor Rectification

We mechanistically validate how the ANCHOR framework rectifies the Ghost Anchor phenomenon through (1) quantifying the visual causal influence on the early translation stage, and (2) examining a layer-wise case study where visual anchoring successfully overrides incorrect linguistic priors.

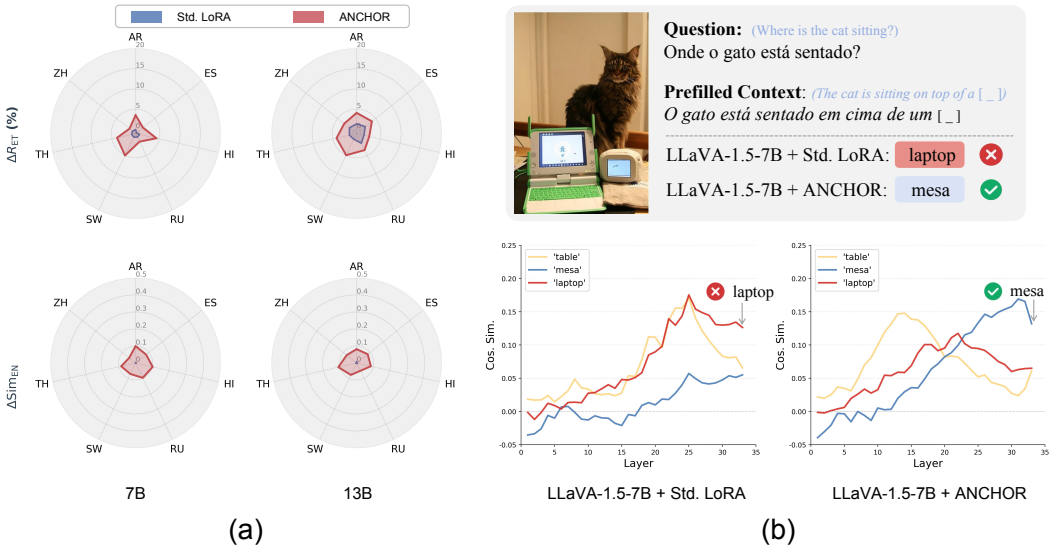


Figure 4: **(a)** Average visual causal effect (ΔR_{ET} and ΔSim_{en}) across 7 languages.**(b)** Layer-wise hidden state dynamics in a cross-lingual VQA case, showing how ANCHOR enables visual grounding to override textual priors during translation.

Enabling Visual Causal Effect Following the methodology in Section 3.2, we measure the visual causal effect (ΔR_{ET} and ΔSim_{en}) across the early layers ($l \in [1, 10]$). As shown in Figure 4(a), standard LLaVA-1.5 exhibits limited visual influence, with baseline ΔR_{ET} averaging near zero (-0.016% for 7B and 1.18% for 13B). This provides evidence consistent with the Ghost Anchor effect. In contrast, ANCHOR produces a positive shift in both ΔR_{ET} and ΔSim_{en} across model scales, indicating increased visual influence on the measured early translation process.

Case Study: Visual Anchoring Overriding Linguistic Priors To contextualize this shift, we examine a cross-lingual VQA case where the baseline follows a text-induced prior and outputs *laptop* rather than the visual target, a *table* (*mesa* in Portuguese). To trace this internal semantic evolution, we compute the layer-wise cosine similarity between the final prompt token's normalized hidden state and the static unembedding weight vectors of specific target concepts. As shown in Figure 4(b), the baseline's early-layer hidden states exhibit stronger alignment with the misleading linguistic prior. In contrast, ANCHOR shifts the internal trajectory toward the visually supported English concept (*table*), which is consistent with its final Portuguese output (*mesa*).

6 Limitations

Our analysis operationalizes the English Semantic Manifold as an empirical property of the English-dominant MLLMs studied in this work, rather than as a universal or normative semantic space. Models trained with different language distributions or architectural designs may exhibit different intermediate semantic geometries. In addition, our mechanistic metrics, including the logit lens, English Translation Ratio, and Visual Grounding Score, measure the decodability of semantic information from intermediate representations, which does not necessarily imply that the model uses all decoded information during generation. Although the Gaussian-noise intervention provides controlled evidence about the influence of visual content, it also introduces a distribution shift and may alter input properties beyond high-level semantics. Therefore, our results characterize modality asynchrony under these operationalizations rather than providing an exhaustive account of all cross-modal causal pathways.

7 Conclusion

In this work, we investigate why visual signals provide limited assistance as cross-lingual bridges for multilingual MLLMs. Mechanistic analysis identifies the Ghost Anchor phenomenon associated with Modality Asynchrony: linguistic streams quickly converge to an English-centric semantic manifold in early layers, whereas visual semantics emerge later, limiting the influence of visual inputs during early cross-modal alignment. To address this, we propose ANCHOR, which leverages Proactive Visual Anchoring (PVA) with an external Visual Foundation Model to accelerate early-layer visual semantic emergence and encourage visual evidence to guide linguistic translation. Experiments on xMMMU, MaXM, and CVQA show that ANCHOR improves cross-lingual visual reasoning and mitigates the performance degradation associated with unanchored early-layer adaptation.

Source Availability

The datasets and code are available at https://github.com/YihDu/MM_align.

Acknowledgements

This work is supported by Longgang District Special Funds for Science and Technology Innovation under Grant LGXXKYJG2024001.

References

Ahtamjan Ahmat, Lei Wang, Yating Yang, Bo Ma, Rui Dong, Kaiwen Lu, Rong Ma, and Xinyue Wang. M2-vlp: Enhancing multilingual vision-language pre-training via multi-grained alignment. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 3438–3450, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714861. URL https://doi.org/10.1145/3696410.3714861.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. In *International Conference on Machine Learning*, pp. 2370–2392. PMLR, 2022.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. Maxm: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2667–2682, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, et al. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*, 2025.

Ahmed Elshabrawy, Hour Kaing, Haiyue Song, Alham Fikri Aji, Hideki Tanaka, Masao Utiyama, and Raj Dabre. When alignment hurts: Decoupling representational spaces in multilingual models. *arXiv preprint arXiv:2508.12803*, 2025.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. mBLIP: Efficient bootstrapping of multilingual vision-LLMs. In Jing Gu, Tsu-Jui (Ray) Fu, Drew Hudson, Asli Celikyilmaz, and William Wang (eds.), *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pp. 7–25, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.alvr-1.2. URL https://aclanthology.org/2024.alvr-1.2/.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander G Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2443–2459, 2021.

Junchen Li, Qing Yang, Bojian Jiang, Shaolin Zhu, and Qingxuan Sun. Lrm-llava: Overcoming the modality gap of multilingual large language-vision model for low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 24449–24457, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3977–3986, 2021.

Nostalgebraist. Interpreting gpt: The logit lens. https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, Aug 2020. Accessed: 5 Jan 2026.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xgqa: Cross-lingual visual question answering. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2497–2511, 2022.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*, 2025.

Shufan Shen, Junshu Sun, Qingming Huang, and Shuhui Wang. Vl-sae: Interpreting and enhancing vision-language alignment with a unified concept set. *arXiv preprint arXiv:2510.21323*, 2025.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. Large vision-language model alignment and misalignment: A survey through the lens of explainability. *arXiv preprint arXiv:2501.01346*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5311–5329, 2025.

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. *arXiv preprint arXiv:2502.15639*, 2025.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, et al. Beyond language modeling: An exploration of multimodal pretraining. *arXiv preprint arXiv:2603.03276*, 2026.

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.

Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip Torr, and Neel Nanda. How visual representations map to language feature space in multimodal llms. *arXiv preprint arXiv:2506.11976*, 2025.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. *arXiv preprint arXiv:2411.04986*, 2024.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2024.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10602–10617, 2025.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319, 2024.

Weihua Zheng, Zhengyuan Liu, Tanmoy Chakraborty, Weiwen Xu, Xiaoxue Gao, Bryan Chen Zhengyu Tan, Bowei Zou, Chang Liu, Yujia Hu, Xing Xie, et al. Mma-asia: A multilingual and multimodal alignment framework for culturally-grounded evaluation. *arXiv preprint arXiv:2510.08608*, 2025.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond english-centric llms: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*, 2024.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4155–4165, 2021.

A Experimental Setup for Mechanistic Analysis

A.1 Input Data Construction

We randomly sample 100 images from the COCO 2017 Validation Set (Lin et al., 2014) as our experimental stimuli. To ensure strict semantic consistency across languages, we develop a high-quality parallel corpus by translating all ground-truth English captions into seven diverse target languages (Table 3) using GPT-5.1 (Singh et al., 2025).

Language	Abbreviations
Chinese	zh
Spanish	es
Hindi	hi
Arabic	ar
Russian	ru
Swahili	sw
Thai	th

Table 3: Selected languages used in the mechanistic experiments and their corresponding ISO 639-1 codes.

To maintain a pure monolingual context and prevent any English prior leakage, the input sequence is formulated entirely in the target language: `[Visual Tokens] + [Target Language Instruction] + [Target Language Caption]`. The instruction prompts the model to verify semantic consistency (e.g., *"Please determine if the following description accurately describes the image."*, translated accordingly). An illustrative example of a constructed input sample across multiple languages is presented in Figure 5.

A.2 Hidden State Extraction

To measure the linguistic and visual streams separately, we extract hidden states at each layer l and isolate two subsets from the full sequence:

- **Text tokens:** All hidden states corresponding to the multilingual instruction and caption.
- **Visual tokens:** Only hidden states from image patches that contain labeled foreground objects (identified via COCO segmentation masks), excluding ambiguous background regions.

Formally, we define the source-language target set as $\mathcal{H}_{text}^{(l)} = \{\mathbf{h}_{src,1,l}, \ldots, \mathbf{h}_{src,M_{src},l}\}$, containing all M_{src} source-language instruction and caption tokens. For English Similarity, the structurally matched English reference set is $\mathcal{H}_{en}^{(l)} = \{\mathbf{h}_{en,1,l}, \ldots, \mathbf{h}_{en,M_{en},l}\}$, containing the corresponding M_{en} English instruction and caption tokens. The visual target set is $\mathcal{H}_{vis}^{(l)} = \{\mathbf{h}_{vis,i,l} \mid i \in \mathcal{N}_{fg}\}$, where $\mathcal{N}_{fg}$ denotes the index set of labeled object tokens. These subsets are directly utilized to compute the metrics defined in Section A.3.

A.3 Metrics Overview

We quantify layer-wise dynamics through three complementary metrics: English Similarity (Sim_{en}) and English Translation Ratio (R_{ET}) for the text stream, and Visual Grounding Score (S_{VG}) for the vision stream. For vocabulary-level projections, we employ the *logit lens* technique (Nostalgebraist, 2020), which maps an intermediate hidden state $\mathbf{h}_l$ to the probability distribution over the vocabulary $\mathcal{V}$ via the pre-trained unembedding matrix $\mathbf{W}_U \in \mathbb{R}^{|\mathcal{V}| \times d}$.

[EN]
Please determine if the following description accurately describes the image Batter preparing to swing at pitch during major game. A baseball player swinging a bat over home plate. A baseball player with a bat in his hand as the pitcher gets ready to throw the ball. A baseball game where the batter is waiting for the pitch. there is a male tennis player that is at base.

[ZH]
请判断以下描述是否准确地描述了图片内容 击球员在重要比赛中准备挥棒击球。一名棒球运动员在本垒板上方挥动球棒。一名棒球运动员手持球棒，投手准备投球。一场棒球比赛中，击球手正等待投球。有一位男性网球运动员在基地位置。

[ES]
Por favor, determine si la descripción describe con precisión el contenido de la imagen Un bateador preparándose para batear un lanzamiento durante un partido importante. Un jugador de béisbol balanceando un bate sobre la base de home. Un jugador de béisbol con un bate en la mano mientras el lanzador se prepara para lanzar la pelota. Un partido de béisbol en el que el bateador espera el lanzamiento. hay un jugador de tenis masculino que está en la base.

[HI]
कृपया निर्धारित करें कि क्या निम्नलिखित विवरण छवि की सामग्री का सटीक वर्णन करता है। एक प्रमुख खेल के दौरान पिच पर घूमने के लिए तैयार होता बल्लेबाज। घर की प्लेट के ऊपर बल्ला घुमाता एक बेसबॉल खिलाड़ी। बल्लेबाजी के लिए तैयार होते हुए बल्ला हाथ में पकड़े एक बेसबॉल खिलाड़ी जबकि पिचर गेंद फेंकने के लिए तैयार है। एक बेसबॉल खेल जहां बल्लेबाज पिच की प्रतीक्षा कर रहा है। एक पुरुष टेनिस खिलाड़ी आधार पर है।

[AR]
ضارب بدقة الصورة محتوى يصف التالي الوصف كان إذا ما تحديد يرجى
يستعد للتأرجح بالمضرب خلال مباراة كبيرة. لاعب بيسبول يلوح بمضرب فوق القاعدة المنزلية. لاعب بيسبول يحمل عصا رمي بينما يستعد الرامي لإلقاء الكرة. مباراة بيسبول يكون فيها الضارب في انتظار الرمية. هناك لاعب تنس ذكر موجود في القاعدة.

[RU]
Пожалуйста, определите, точно ли следующее описание соответствует содержанию Бэттер готовится отбить подачу во время важной игры. Бейсболист, замахивающийся битой над домашней площадкой. Бейсболист с битой в руке, когда питчер готовится бросить мяч. Бейсбольная игра, бьющий игрок ждёт подачи. Мужчина-теннисист находится у базы.

[SW]
Tafadhali tambua ikiwa maelezo yafuatayo yanafafanua kwa usahihi maudhui ya picha Mchezaji anayepanda kipanda kabla ya kupiga soka wakati wa kipindi muhimu. Mchezaji wa baseball anapandisha batani juu ya bao la nyumbani. Mchezaji wa baseball anayeshikilia bati mkononi wakati mpenda anaandaa kumwachia mpira. Kucheza baseball ambapo mchezaji anasubiri mshambulio. kuna mchezaji wa tenisi ambaye ni kwenye msimbo.

[TH]
โปรดพิจารณาว่าคำบรรยายต่อไปนี้อธิบายเนื้อหาของภาพได้อย่างถูกต้องหรือไม่ ผู้ตีกำลังเตรียมตัวสวิงไม้ตีลูกเบสบอลในการแข่งขันสำคัญ นักเบสบอลคนหนึ่งเหวี่ยงไม้ตีบอลเหนือแผ่นบ้าน นักเบสบอลถือไม้ตีขณะที่ผู้ขว้างลูกกำลังเตรียมตัวขว้างลูกบอล การแข่งขันเบสบอลที่ผู้ตีกำลังรอการขว้างลูก มีนักเทนนิสชายคนหนึ่งที่อยู่ที่ฐาน

Figure 5: **Illustrative example of the input construction.** This figure demonstrates the parallel input constructions across 7 structurally diverse target languages. To enforce a strict monolingual context and prevent English prior leakage during inference, both the standardized verification instruction and the ground-truth image caption are translated into the target language and paired with identical visual tokens from the COCO dataset.

A.3.1 English Similarity (Sim_{en})

To measure continuous geometric alignment toward the English semantic manifold $\mathcal{S}_{en}$, we first compute the sequence-level mean representation for the source-language input (using $\mathcal{H}_{text}^{(l)}$) and its structurally matched English reference (using $\mathcal{H}_{en}^{(l)}$). Both inputs contain the corresponding instruction and caption, while their token counts may differ across languages:

$$\overline{\mathbf{h}}_{src,l} = \frac{1}{M_{src}} \sum_{j=1}^{M_{src}} \mathbf{h}_{src,j,l}, \quad \overline{\mathbf{h}}_{en,l} = \frac{1}{M_{en}} \sum_{j=1}^{M_{en}} \mathbf{h}_{en,j,l} \tag{3}$$

The layer-wise English Similarity is then defined as their cosine similarity:

$$\text{Sim}_{en}(l) = \frac{\overline{\mathbf{h}}_{src,l} \cdot \overline{\mathbf{h}}_{en,l}}{\|\overline{\mathbf{h}}_{src,l}\| \|\overline{\mathbf{h}}_{en,l}\|} \tag{4}$$

A low Sim_{en} in initial layers reflects the geometric separation between source language manifolds and the English semantic pivot.

A.3.2 English Translation Ratio (R_{ET})

To quantify vocabulary-level convergence, we analyze the proportion of text tokens in $\mathcal{H}_{text}^{(l)}$ that decode into English. For the j-th source language token, we extract the top-k candidate decoded words via the logit lens to form the set $\mathcal{T}_{j,l}$, setting $k = 5$:

$$\mathcal{T}_{j,l} = \text{Top}_k(\text{Softmax}(\mathbf{W}_U \mathbf{h}_{src,j,l})) \tag{5}$$

To rigorously exclude language-agnostic symbols (e.g., punctuation, numbers), we employ GPT-4o as a zero-shot language classifier, denoted as $\Phi_{lang}(\cdot)$. A token is considered to have transitioned into the English space if at least one of its top-k decoded words is classified as English. The translation ratio is calculated only over tokens that bear actual linguistic content—that is, tokens for which at least one top-k decoded word is classified as a natural language (e.g., English, Chinese, Arabic) rather than as punctuation, numbers, or other symbols. Formally:

$$R_{ET}(l) = \frac{\sum_{j=1}^{M_{src}} \mathbb{I}\Big[\exists w \in \mathcal{T}_{j,l} \text{ s.t. } \Phi_{lang}(w) = \text{English}\Big]}{\sum_{j=1}^{M_{src}} \mathbb{I}\Big[\exists w \in \mathcal{T}_{j,l} \text{ s.t. } \Phi_{lang}(w) \in \{\text{Natural Languages}\}\Big]} \tag{6}$$

This metric directly quantifies the discrete lexical transition into the English space while accounting for the distributional nature of the output logits.

A.3.3 Visual Grounding Score (S_{VG})

We evaluate the emergence of visual semantics by measuring how visual tokens align with concrete linguistic concepts. Using the labeled object hidden states $\mathcal{H}_{vis}^{(l)}$ extracted via COCO annotations, we monitor their convergence toward the set of ground-truth object labels $\mathcal{C}_{GT}$.

For each object token $i \in \mathcal{N}_{fg}$ at layer l, we extract the top-k candidate text tokens via the logit lens: $\mathcal{T}_{i,l} = \text{Top}_k(\text{Softmax}(\mathbf{W}_U \mathbf{h}_{vis,i,l}))$, setting $k = 5$. The grounding score S_{VG} is defined as the successful semantic hit rate across these object tokens:

$$S_{VG}(l) = \frac{1}{|\mathcal{N}_{fg}|} \sum_{i \in \mathcal{N}_{fg}} \mathbb{I}\Big[\exists w \in \mathcal{T}_{i,l}, \exists c \in \mathcal{C}_{GT} \text{ s.t. } w \subseteq c\Big] \tag{7}$$

where $w \subseteq c$ denotes that a decoded candidate word matches or is a valid substring of a ground-truth object label. S_{VG} serves as a direct measurement for the cross-modal semantic verbalization process.

(a) (b)

Figure 6: **(a)** Original segmentation masks and semantic labels from the COCO dataset. **(b)** The mapped 24×24 visual patches and their corresponding labels.

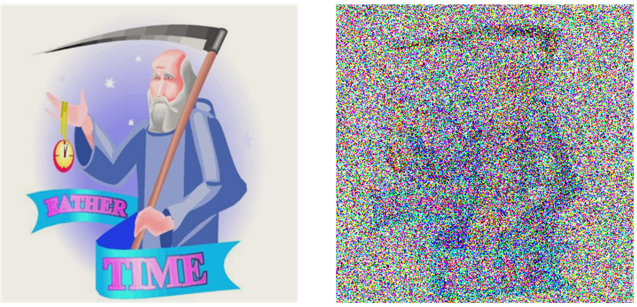
A.4 Noise Intervention: Implementation Details

Motivation. The noise intervention creates a control condition in which visual tokens remain present while recognizable high-level visual content is substantially reduced. This probes the sensitivity of linguistic translation to visual content while holding constant: (i) sequence length, (ii) token positions, and (iii) per-channel first- and second-order pixel statistics.

Implementation. For each image I, we compute per-channel mean μ_c and standard deviation σ_c from the original pixel values. The noise image I_{noise} is generated by sampling from a Gaussian distribution with matching statistics:

$$I_{\text{noise}}^{(c)}(x,y) \sim \mathcal{N}\left(\mu_c, \sigma_c^2\right), \quad \text{where } \mu_c = \frac{1}{HW}\sum_{x,y} I^{(c)}(x,y) \tag{8}$$

The noise tensor is passed to the vision encoder via `pixel_values`, preserving the input dimensions, patch count, and visual-token positions while disrupting recognizable object and scene structure.



(a) Normal input (b) Noise condition

Figure 7: Visual comparison of intervention conditions. **Left:** Normal RGB input with intact semantic content. **Right:** Gaussian noise with matched per-channel statistics (μ, σ) but no recognizable objects or scenes.

Validation. As shown in Figure 7, the noise intervention removes recognizable object contours, textures, and scene structure from the input. Because the matched noise preserves only per-channel first- and second-order pixel statistics, it provides a control condition with substantially reduced high-level visual semantics; we do not assume that it eliminates every decodable visual signal.

B Datasets

B.1 Training Datasets

To construct a high-quality multilingual visual instruction tuning dataset, we systematically sample and filter image-text pairs from the ShareGPT4V dataset (Chen et al., 2024). To minimize potential noise and artifacts introduced during the machine translation process, we implement a strict filtering mechanism based on text length and annotation quality, prioritizing structurally straightforward sentences. Specifically, we extract 9,000 image-text pairs from the COCO subset and 6,000 from the GQA subset within ShareGPT4V. These specific subsets are chosen because their concise and visually-grounded question-answering formats make them ideal anchors for cross-lingual alignment. Subsequently, we translate the selected English questions and answers into nine typologically diverse target languages using GPT-5.1 (Singh et al., 2025). These languages, namely Arabic (ar), Spanish (es), French (fr), Hindi (hi), Portuguese (pt), Russian (ru), Swahili (sw), Thai (th), and Chinese (zh), are specifically chosen to encompass a broad spectrum of language families and resource availability levels. This pipeline ensures that the semantic consistency between the visual inputs and text representations is accurately preserved across different linguistic contexts.

B.2 Details of Evaluated Multilingual Multimodal Benchmarks

To comprehensively assess visual understanding and reasoning capabilities across diverse linguistic and cultural contexts, we evaluate the proposed method on three distinct multilingual multimodal benchmarks: xMMMU (Yue et al., 2024), MaXM (Changpinyo et al., 2023), and CVQA (Romero et al., 2024).

Multilingual MMMU (xMMMU) xMMMU (Yue et al., 2024) focuses on complex multimodal understanding and reasoning across multiple academic subjects, evaluating the capacity to understand specialized content across different languages and modalities. The dataset features 183 subfields and 30 diverse image types, including charts, diagrams, and chemical structures. It consists of 300 English questions randomly sampled from the MMMU validation set, which were subsequently translated into six languages (ar, fr, hi, id, ja, pt) using GPT-4o. The questions are presented in multiple-choice and short-answer formats. We follow the evaluation protocol in (Yue et al., 2024) and report accuracy.

Multilingual Visual Question Answering (MaXM) MaXM (Changpinyo et al., 2023) is a test-only multilingual open-ended visual question-answering benchmark encompassing seven languages (fr, hi, th, zh, en, he, ro). To address the challenge of cultural diversity in multimodal understanding, the images used in MaXM were taken in regions where each language is spoken to better reflect cultural contexts. It contains 2K questions, where answers are in the same language as the question. We use accuracy as the evaluation metric following prior work (Yue et al., 2024; Dash et al., 2025).

Culturally-diverse Multilingual Visual Question Answering (CVQA) CVQA (Romero et al., 2024) evaluates the ability to reason about culturally diverse visual content. It consists of over 21K multiple-choice questions designed to test multimodal reasoning across 39 distinct language-region pairs, covering 31 unique languages and 30 countries. To accurately reflect global cultural nuances, the dataset incorporates a wide spectrum of languages, ranging from high-resource languages like Chinese and regional variations of Spanish, to lower-resource indigenous languages such as Amharic, Javanese, Minangkabau, and Swahili. We evaluate models under the local-language question setting and report accuracy as the evaluation metric.

C Additional Experiment Results

C.1 Mechanistic Results on more MLLMs

Table 4 reports the effect of visual intervention on the English Translation Ratio (ΔR_{ET}) and the English Similarity ($\Delta\text{Sim}_{\text{en}}$), averaged over the early translation stage ($L \in [1, 10]$). Consistent with our main findings, all evaluated models, including Qwen2.5-VL (Bai et al., 2025b), Gemma 3 (Gemma Team, 2025), and LLaVA-NeXT (Liu et al., 2024), exhibit small changes under the intervention, providing evidence that the Ghost Anchor phenomenon extends across the evaluated architectures and scales.

Model	$\Delta R_{\textbf{ET}} \uparrow$	$\Delta\textbf{Sim}_{\textbf{en}} \uparrow$
Qwen2.5-VL 3B	4.12×10^{-3}	-2.31×10^{-4}
Qwen2.5-VL 7B	5.34×10^{-3}	-4.51×10^{-3}
Gemma-3-4B	-1.13×10^{-2}	-4.74×10^{-4}
Gemma-3-13B	6.95×10^{-3}	7.37×10^{-3}
LLaVA-NeXT-7B	-3.19×10^{-3}	-4.71×10^{-4}
LLaVA-NeXT-13B	8.61×10^{-3}	-8.28×10^{-3}

Table 4: Visual Causal Effect on English Translation Ratio (ΔR_{ET}) and English Similarity ($\Delta\text{Sim}_{\text{en}}$).

Backbone	Size	ΔEn	ΔSFT Avg. 🔥	ΔZero-shot Avg. ❄	ΔAvg.
LLaVA-NeXT	7B	+0.4	+3.1	+1.7	**+2.3**
LLaVA-NeXT	13B	+0.5	+4.0	+1.9	**+2.9**
InternVL3	8B	+0.2	+2.3	+1.1	**+1.7**
InternVL3	14B	+0.5	+2.6	+1.3	**+1.9**

Table 5: Generalization to newer MLLM backbones on xMMMU. Each entry reports the accuracy difference in percentage points between ANCHOR and the corresponding Std. LoRA baseline (ANCHOR − Std. LoRA). SFT Avg. and Zero-shot Avg. are macro-averages over their respective language groups, while Avg. is the macro-average over all seven evaluated language subsets, including English.

Backbone	Size	ΔEn	ΔSFT Avg. 🔥	ΔZero-shot Avg. ❄	ΔAvg.
LLaVA-NeXT	7B	+0.8	+3.4	+1.2	**+2.4**
LLaVA-NeXT	13B	+1.0	+3.8	+1.3	**+2.7**
InternVL3	8B	+0.5	+2.6	+0.8	**+1.8**
InternVL3	14B	+0.6	+3.0	+0.9	**+2.1**

Table 6: Generalization to newer MLLM backbones on MaXM. Each entry reports the accuracy difference in percentage points between ANCHOR and the corresponding Std. LoRA baseline (ANCHOR − Std. LoRA). SFT Avg. and Zero-shot Avg. are macro-averages over their respective language groups, while Avg. is the macro-average over all seven evaluated language subsets, including English.

Backbone	Size	ΔSFT Avg. 🔥	ΔZero-shot Avg. ❄	ΔAvg.
LLaVA-NeXT	7B	+3.4	+1.4	**+2.1**
LLaVA-NeXT	13B	+3.8	+1.8	**+2.5**
InternVL3	8B	+2.6	+1.0	**+1.6**
InternVL3	14B	+2.9	+1.1	**+1.7**

Table 7: Generalization to newer MLLM backbones on CVQA. Each entry reports the accuracy difference in percentage points between ANCHOR and the corresponding Std. LoRA baseline (ANCHOR − Std. LoRA). SFT Avg. and Zero-shot Avg. are macro-averages over their respective language–region groups, while Avg. is the macro-average over all 39 language–region pairs.

C.2 Full Results on the CVQA Benchmark

Models	ar-EG	es-AR	es-CL	es-CO	es-EC	es-ES	es-MX	es-UY	hi-IN	pt-BR
LLaVA-1.5-7B	32.5	54.3	60.7	56.4	53.6	67.0	48.3	43.5	40.3	57.4
+ Std. LoRA	31.5	51.7	56.4	55.2	49.5	62.0	46.1	40.6	38.8	56.3
+ Early-Layer FFT	29.6	41.5	44.9	46.5	42.3	50.3	43.7	37.8	37.3	45.8
+ ANCHOR	37.3	55.8	60.3	58.9	53.7	69.3	51.2	46.7	42.5	59.0
LLaVA-1.5-13B	38.4	57.4	62.0	61.4	53.6	70.4	51.1	39.7	49.3	63.4
+ Std. LoRA	36.0	54.7	57.7	61.4	55.0	67.9	50.8	39.7	46.8	59.9
+ Early-Layer FFT	28.1	41.1	44.0	44.0	42.5	52.2	41.8	34.0	40.3	47.5
+ ANCHOR	40.1	59.2	65.3	63.0	56.1	70.8	51.5	42.8	50.3	64.3
Models	**ru-RU**	**sw-KE**	**zh-CN**	**zh-SG**	**am-ET**	**bg-BG**	**bn-IN**	**br-FR**	**fil-PH**	**ga-IE**
LLaVA-1.5-7B	57.5	36.3	48.6	48.6	26.5	39.4	29.7	29.6	46.3	44.2
+ Std. LoRA	57.0	37.4	44.7	47.6	26.1	38.8	33.6	26.9	42.4	43.9
+ Early-Layer FFT	41.5	38.5	42.8	43.9	27.4	34.8	33.2	30.6	39.4	36.8
+ ANCHOR	57.1	36.3	49.1	52.5	26.4	38.0	34.6	28.3	44.5	44.0
LLaVA-1.5-13B	57.0	42.9	54.3	52.8	25.2	44.7	34.6	35.8	48.8	46.6
+ Std. LoRA	53.5	41.8	50.5	52.8	26.1	42.3	32.9	32.1	47.8	42.0
+ Early-Layer FFT	46.0	44.7	41.8	42.0	36.8	36.7	27.6	31.4	38.9	36.8
+ ANCHOR	57.4	46.0	59.4	56.0	26.0	42.8	35.7	34.7	49.1	48.5
Models	**id-ID**	**ig-NG**	**ja-JP**	**jv-ID**	**ko-KR**	**min-ID**	**mn-MN**	**mr-IN**	**ms-MY**	**no-NO**
LLaVA-1.5-7B	42.2	32.5	35.0	36.4	51.0	36.3	29.8	36.1	47.9	52.5
+ Std. LoRA	40.1	32.0	35.5	35.0	47.2	36.2	31.1	34.2	45.4	51.5
+ Early-Layer FFT	35.7	27.5	35.5	31.0	43.5	31.1	32.1	32.7	35.9	42.8
+ ANCHOR	40.4	31.9	36.9	37.3	49.3	37.1	33.2	37.1	47.2	53.1
LLaVA-1.5-13B	46.6	36.5	45.8	36.7	54.1	35.9	30.8	38.6	48.6	58.2
+ Std. LoRA	40.8	31.5	42.4	37.7	48.6	36.7	31.0	39.1	44.8	52.2
+ Early-Layer FFT	39.6	30.5	37.0	32.3	47.2	35.1	27.2	35.2	39.4	45.2
+ ANCHOR	41.2	34.8	47.5	39.0	53.2	37.7	33.9	40.5	46.7	56.6
Models	**om-ET**	**ro-RO**	**rw-RW**	**si-LK**	**su-ID**	**ta-IN**	**te-IN**	**ur-IN**	**ur-PK**	**Avg.**
LLaVA-1.5-7B	37.4	51.0	34.5	24.4	37.5	31.8	34.0	27.7	24.1	41.6
+ Std. LoRA	35.5	46.7	34.9	23.6	35.0	26.2	27.0	29.5	25.9	40.0
+ Early-Layer FFT	22.0	44.4	33.6	25.3	34.0	24.3	27.0	28.6	26.4	36.0
+ ANCHOR	34.9	37.4	35.3	24.1	36.4	29.0	33.2	32.3	26.6	42.0
LLaVA-1.5-13B	34.6	57.0	34.5	28.0	39.0	26.2	34.0	41.4	34.3	44.9
+ Std. LoRA	32.2	52.3	36.6	30.2	38.0	32.2	31.5	37.7	34.7	43.1
+ Early-Layer FFT	30.4	44.7	27.2	26.7	34.0	22.9	30.0	30.5	31.0	37.0
+ ANCHOR	32.0	58.0	37.1	30.9	38.7	33.3	32.6	39.9	38.3	45.9

Table 8: CVQA benchmark performance. Pink columns denote language–region pairs whose languages are included in SFT, while Cyan columns denote languages unseen during training. The **Avg.** column reports the macro-average accuracy across all 39 pairs.

C.3 Ablation Study

Target Layer (L_{early}) The choice of L_{early} is critical in our ANCHOR framework, as it defines the precise early translation window during which cross-modal alignment must occur. We conduct an ablation over different values of L_{early} to understand its impact on downstream reasoning performance.

Table 9 presents the results of varying L_{early} from 6 to 16 layers. We evaluate on xMMMU and MaXM, reporting the macro-averaged accuracy across all evaluated languages.

The results reveal an inverted U-shaped trend: choosing L_{early} too small (e.g., 6 layers) may provide insufficient time for visual semantics to emerge. Conversely, setting L_{early} too large (e.g., 16 layers) may allow the linguistic trajectory to become strongly English-dominant before effective visual guidance is available. The best evaluated setting, $L_{early} = 10$, is consistent with our mechanistic analysis of visual semantic emergence and achieves the highest downstream performance among the tested values. These results suggest that the timing of early-layer intervention is important for cross-modal alignment.

Intervention Scope: Early-Layer vs. Full SFT To further validate our layer-specific design, we compare our early-layer intervention ($L_{early} = 10$) against a full-parameter fine-tuning baseline augmented with the PVA objective. Crucially, in this baseline, the PVA constraint remains applied at layer 10 to ensure a strictly controlled comparison. As shown in Table 10, applying Full SFT degrades the overall accuracy across both benchmarks compared to our targeted early-layer approach. This performance gap suggests that full-depth updates may overfit to the SFT language distribution or interfere with representations that support cross-lingual generalization. Confining parameter updates to the early alignment window yields better aggregate performance in this comparison while preserving more of the LLM's multilingual capability.

L_{early}	xMMMU	MaXM
6	32.8	28.5
8	34.2	30.7
10	**35.6**	**32.4**
12	34.7	31.8
14	33.5	30.2
16	32.1	28.9

Table 9: Ablation study on the target layer L_{early}. xMMMU and MaXM results are macro-averaged across all evaluated languages. All experiments are conducted on LLaVA-1.5-7B.

Intervention Scope	xMMMU	MaXM
Full SFT + PVA	32.8	27.5
ANCHOR (Ours)	**35.6**	**32.4**

Table 10: Ablation on intervention scope using LLaVA-1.5-7B. Full-parameter fine-tuning augmented with PVA at layer 10 yields lower aggregate performance than the early-layer intervention on both benchmarks, suggesting that targeted early-layer updates better preserve cross-lingual generalization in this setting.

D Complete Implementation Details

Model Architecture The alignment network ($\text{MLP}_{\text{align}}$) consists of a two-layer perceptron with GELU activation.

Training Configuration We train for 1 epoch on 135K image-text pairs. We use the AdamW optimizer with a peak learning rate of 2×10^{-5} and a cosine decay schedule with a 3% warmup period. The PVA loss weight is set to $\alpha = 0.4$. All experiments are conducted on NVIDIA A100 (80GB) GPUs with a global batch size of 128. Input images are resized to 336×336 resolution, and the maximum sequence length is 2048 tokens. During training, we apply full-parameter fine-tuning to the multimodal projector and the first $L_{early} = 10$ transformer layers, while freezing layers 11–32 (for 7B) or 11–40 (for 13B) and the vision encoder. The trainable parameters thus include the vision-language projector, the alignment network $\text{MLP}_{\text{align}}$, and the initial FFT transformer layers. For the standard LoRA baseline, we employ a rank $r = 128$ and a scaling factor $\alpha_{\text{lora}} = 256$, applying the adaptation to all linear layers (including q, k, v, o and MLP gates) of the LLM backbone to ensure a competitive and fair comparison.