Expanding the Boundaries of Vision Prior Knowledge in Multi-modal Large Language Models

Anonymous ACL submission

Abstract

Does the prior knowledge of the vision encoder constrain the capability boundary of Multi-modal Large Language Models (MLLMs)? While most existing research treats MLLMs as unified systems optimized through end-to-end training, the impact of vision encoder's prior knowledge is seldom investigated. In this work, we introduce a novel metric, $Rank_e$, to quantify the effect of prior knowledge of the vision encoder on MLLM performance. Our analysis reveals a positive correlation between prior knowledge and MLLM performance. Moreover, we find that domain-specific fine-tuning using solely end-to-end visual question answering (VQA) data is insufficient, particularly for entities with low inherent visual prior knowledge. To address this issue, we propose VisPRE (Vision Prior Remediation), a two-stage training framework that explicitly incorporates prior knowledge at the vision encoder level. Experimental results demonstrate that augmenting vision encoder's prior knowledge substantially boosts the visual understanding capabilities of MLLMs, offering a novel and effective strategy for improving performance, especially in scenarios involving uncommon visual entities.

1 Introduction

Multi-modal Large Language Models have emerged as a rapidly growing area of research. Combining the powerful capabilities of Large Language Models with the ability to process visual input, MLLMs excel in tasks such as image understanding, VQA (Agrawal et al., 2016), image captioning, and visual instruction following. The development of models such as GPT-4o (OpenAI, 2024), GPT-4V (OpenAI, 2023), and Claude-3.5 (Anthropic, 2024) have demonstrated remarkable proficiency in advanced multi-modal understanding. Besides, open-source models like LLaVA (Liu et al., 2024b,a; Li et al., 2024a) series, Qwen2-VL (Wang et al., 2024), and InternVL2 (Chen et al., 2024b,a) are making significant strides, bridging the gap in the field.

Figure 1: **Knowledge quadrants of a MLLM.** "Vision known" indicates that the vision encoder recognises the entity in the image, while "Language known" indicates that the language model possesses relevant information about the entity. Only when both vision and language are "known" can the MLLM achieve accurate comprehension and response generation.

A pivotal challenge in advancing MLLMs is forging a seamless and robust alignment between vision and language. One effective approach involves integrating an off-the-shelf external vision encoder with a language model using a modality conversion module (Alayrac et al., 2022; Li et al., 2023a,d; Zhu et al., 2023; Dai et al., 2023; Bai et al., 2023; Liu et al., 2024b; Li et al., 2022), which we refer to as the modular approach. Compared to the monolithic multi-modal approach (Team, 2024a; Luo et al., 2024; Bavishi et al., 2023; Zhan et al., 2024), which is built from scratch using multi-modal data, the modular approach is more data-efficient and achieves comparable performance. Despite these advantages, the modular approach still faces challenges, as the vision and language components are

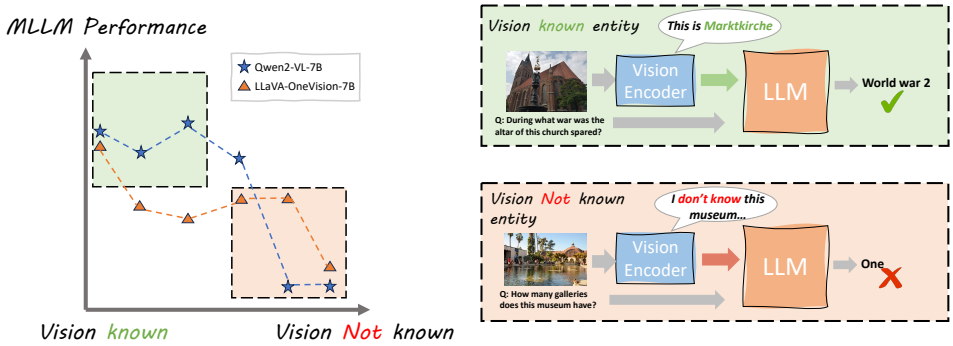


Figure 2: Left: **Current MLLM performance vs. vision prior knowledge.** Current MLLMs demonstrate positive correlation between vision prior knowledge and overall performance. Right: **"Vision Known" and "Vision Not Known" Entities.** (1) For "vision known entities", the vision encoder contains sufficient prior knowledge, enabling MLLM answers correctly; (2) For "vision not known entities", insufficient visual knowledge leads to MLLM failure. We propose the $Rank_e$ metric to quantify vision encoder's prior knowledge about specific entities, along with a two-stage training framework to enhance encoder knowledge, expanding MLLM's performance boundaries.

trained separately from distinct data distributions, leading to an inherent misalignment in their knowledge. To illustrate the importance of knowledge alignment, we present a knowledge quadrant diagram in Fig. 1, with the horizontal axis representing the knowledge held by the language model and the vertical axis representing the knowledge held by the vision encoder. Only when both components possess necessary knowledge (in the "Vision known & Language known" quadrant) can the multi-modal model accurately handle complex cross-modal tasks (Li et al., 2023c; Cheng et al., 2024). Misalignment in knowledge from either the vision or language side introduces limitations to the model's capabilities, making it essential to bridge this gap to enhance the performance of multi-modal models. Many existing studies focus on addressing knowledge misalignment from the language perspective, expanding from "Vision known & Language not known" to "Vision known & Language known". Some studies (Caffagni et al., 2024; Jiang et al., 2024) enhance language model knowledge with external documents related to images, while CVLM (Li et al., 2024b) trains a "Visual Knowledge Aligner" module to enrich text-based knowledge associated with images. However, as a crucial component of MLLM (Collins and Olson, 2014), the vision encoder also possesses varying prior knowledge about the real world, such as entities, textures, and causality (Pinker, 1984; Cavanagh, 2011). But the impact of this vision prior knowledge on MLLM capabilities remains unexplored, leading to a natural question: **How does vision prior knowledge affect MLLM's capability?** In this paper, we attempt to answer this question by investigating the following research questions:

- **Q1:** How to measure prior knowledge in vision encoders?
- **Q2:** Does vision prior knowledge constrain MLLM?
- **Q3:** How to transcend vision prior knowledge limits?

To address these questions, we introduce $Rank_e$ to quantify the vision encoder's prior knowledge. Through experiments with various model combinations, we reveal a positive correlation between MLLM performance and visual prior knowledge. Fig. 2 (left) demonstrates the relationship between current MLLM performance and vision prior. Furthermore, we find that direct fine-tuning with end-to-end VQA data is insufficient for improving MLLM performance on low prior entities. Fig. 2 (right) illustrates the knowledge misalignment on low prior entities. To overcome this limitation, we propose a two-stage training framework that injects vision prior knowledge into the vision encoder, resulting in significant improvements in MLLM performance. In summary, our main contributions are:

- We introduce the $Rank_e$ metric to quantify a vision encoder's prior knowledge, revealing a positive correlation between MLLM performance and the encoder's embedded visual knowledge.
- Our analysis shows that domain-specific fine-tuning with only end-to-end VQA data proves

insufficient, particularly for entities with low vision prior knowledge.

- We propose a two-stage training framework **VisPRE** (**Vis**ion **P**rior **Re**mediation) that injects prior knowledge at the vision encoder level, significantly enhancing MLLM performance, especially for entities with low vision prior knowledge.

2 Vision Prior Measurement

Vision encoders are typically trained on extremely large-scale data (from 400 million to 10 billion samples (Tong et al., 2024a)), often with undisclosed data (e.g., OpenAI CLIP (Radford et al., 2021)), making direct evaluation of vision priors from training data infeasible. Therefore, to answer **Q1**, we shift our focus to evaluating observable behavioral evidence - specifically, how effectively these encoders recognize visual entities. We thus propose the $Rank_e$ metric, which quantifies an encoder's vision prior knowledge for a given entity e.

In this section, we begin by describing the modality alignment process in modular MLLMs, then formulating the definition of vision prior knowledge. Finally, we introduce the $Rank_e$ metric to quantify this knowledge.

Modular MLLMs establish cross-modal understanding through an alignment process that maps visual information to textual representations. Formally, given an input text prompt T_A and target image I_B, where $\mathcal{F}$ represents the MLLM's internal representation function that maps inputs to hidden states, the alignment process can be described as:

$$\begin{aligned} \mathcal{F}(T_A, I_B) &\xrightarrow{\text{align}} \mathcal{F}(T_A, \hat{T}_B) \\ \text{where} \quad \hat{T}_B &\sim P(T|I_B) \end{aligned} \tag{1}$$

Here, $\hat{T}_B$ represents the generated text that preserves the semantic content of I_B. Building upon the Platonic representation hypothesis (Huh et al., 2024), we posit that cross-modal alignment occurs through a shared latent space $\mathcal{Z}$. This allows us to decompose the $P(T|I_B)$ as:

$$P(T|I_B) = \sum_{z \in \mathcal{Z}} \underbrace{P_{\text{vision}}(z|I_B)}_{\text{Vision prior}} \cdot P_{\text{align}}(T|z, I_B) \tag{2}$$

The latent representation z serves as an intermediary that connects the visual and textual domains. While $P_{\text{align}}(T|z, I_B)$ reflects the MLLM's ability to convert latent representation z into textual output T, $P_{\text{vision}}(z|I_B)$ represents the vision encoder's capability to transform image I_B into an appropriate latent representation. $P_{\text{vision}}(z|I_B)$ constitutes what we define as vision prior knowledge—the encoder's pre-existing understanding of visual entities encoded in its parameters.

To quantify the inherent vision prior $P_{\text{vision}}(z|I_B)$, we discretize the continuous latent space $\mathcal{Z}$ into a set of entity-specific latent representations. For a given image I_B, we approximate $P(z|I_B)$ by evaluating the probability that the vision encoder correctly identifies an entity within I_B. To achieve this, we propose the $Rank_e$ metric, which measures how well the encoder identifies a target entity e from visual inputs, thereby evaluating the vision encoder's inherent prior knowledge. As shown in Fig. 3, for an entity e, $Rank_e$ is computed as follows:

- **Similarity scoring:** For an image I_e containing entity e, compute the image-text similarity score $s_j = \phi(I_e, T_j)$ using the vision encoder and its corresponding text encoder, where $\{T_1, ..., T_n\}$ are textual descriptions of n candidate entities.

- **Ranking:** Rank the entities in descending order based on their similarity scores $\{s_j\}_{j=1}^n$, and record the position of the target entity e as $Rank_e$. If multiple images $\{I_e^{(1)}, ..., I_e^{(m)}\}$ are available for single entity e, compute $Rank_e$ for each image separately and take the average:

$$\text{Rank}_e = \frac{1}{m} \sum_{i=1}^{m} rank\big(\phi(I_e^{(i)}, T_e)\big). \tag{3}$$

where $rank(\phi(I_e, T_e))$ denotes the position of $\phi(I_e, T_e))$ in ordered $\{s_j\}_{j=1}^n$. Lower $Rank_e$ values indicate stronger visual prior knowledge, with optimal performance when $\text{Rank}_e = 1$.

3 Experiments

In this section, we explore the three proposed research questions. In Section 3.1, we describe the overall experimental setup. In Section 3.2, we verify the relationship between MLLM and the prior knowledge of its vision encoder. From Section 3.3 to Section 3.4, we show the insufficiency of end-to-end fine-tuning and propose a training framework to transcend vision prior knowledge limits.

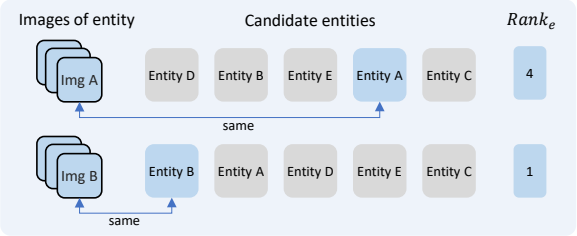


Figure 3: **Illustration of metric Rank$_e$.** For a target entity e, we compute cross-modal similarity scores between its vision representations (extracted by vision encoder) and text representations of all candidate entities (extracted by corresponding text encoder). The rank of entity e among these candidates defines its $Rank_e$. In this example, while Image A depicts Entity A, entity A achieves 4th-highest similarity score, resulting in $Rank_e$ = 4.

3.1 Experiment Setting

Models. To systematically examine the impact of vision encoder's prior knowledge on MLLM performance across different vision encoders and base LLM combinations, we train nine MLLMs from scratch based on an encoder-projector-LLM architecture. For the vision encoder, we use widely adopted encoders in MLLMs, including OpenAI ViT-L-14 (Radford et al., 2021), SigLIP ViT-SO-14 (Zhai et al., 2023), and DFN ViT-H-14 (Fang et al., 2023). For base LLM, we select the LLaVA-1.5 language model, Vicuna-7B-v1.5 (Chiang et al., 2023), and recent open-source models, Llama-3.1-Instruct-7B (Dubey et al., 2024) and Qwen-2.5-Instruct-7B (Team, 2024b).

Datasets. To evaluate MLLMs under different vision priors, we require a VQA dataset that meets two conditions: (1) it provides entity annotations covering a wide range of prior knowledge—from extremely rare to very common entities; (2) it includes entity-centric visual questions and answers for MLLM performance assessment. Here, rare entities refer to those that appear infrequently or not at all in the vision encoder's training data, making them difficult for the vision encoder to recognize accurately. The Encyclopedia-VQA (Mensink et al., 2023) dataset fulfills both requirements. With extensive entity annotations covering up to 16.7k entity categories, it captures both common and rare entities and poses a hard challenge for MLLMs with its knowledge-based VQA questions.

Training. We conducted training on a 8×A800 GPUs. Initially, we pre-trained the model on the LLaVA (Liu et al., 2024b) dataset to develop an MLP projector aligned with selected vision encoder. For fine-tuning phase, we sampled 10% of the LLaVA instruction tuning dataset and integrated it with additional fine-tuning data to optimize computational efficiency while maintaining performance quality.

Metrics and Evaluation. We use Llama-3.1-70B (Dubey et al., 2024) to judge model responses, denoted as a function $g(\cdot)$ that takes the question, entity, ground truth answer, and model output as input, returning *true* if the answer is correct. Using this, we define entity accuracy Acc$_e$ for each entity e as the fraction of correct responses among all related questions:

$$\text{Acc}_e = \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbb{1}\left[g(y_i, \hat{y}_i) = \text{true}\right] \quad (4)$$

where N_e is the number of questions for entity e, y_i is the ground truth answer and other question information, and $\hat{y}_i$ is the model's output. The overall dataset accuracy Acc$_{\text{macro}}$ is calculated as the macro-average of all entity accuracies. Details of the evaluation configurations are in Appendix B.

3.2 Vision Prior Constrains MLLM Performance

To investigate **Q2:** "Does vision prior knowledge constrain MLLM?", we first categorize entities into two types: those "vision encoder knows" and those "vision encoder doesn't know" then observe MLLM performance across both categories. Through our proposed $Rank_e$ metric, we measure the vision encoder's knowledge of entities in Encyclopedia-VQA, where a lower $Rank_e$ indicates greater knowledge. For MLLM performance, we test accuracy in answering entity-related questions in Encyclopedia-VQA.

Our study aims to address knowledge misalignment where MLLM capabilities are limited by the vision encoder. Therefore, we retain only cases where the LLM component possesses adequate entity knowledge, regardless of the vision encoder's knowledge. Specifically, we prompt the MLLM with "This is {entity name}" rather than the actual image; if the MLLM answers correctly, we retain this case. Additionally, we discovered a number of cases where MLLMs provide correct answer without image description or actual image. We attribute this to the MLLM's dependency on question format (Jiang et al., 2024). We eliminated this subset

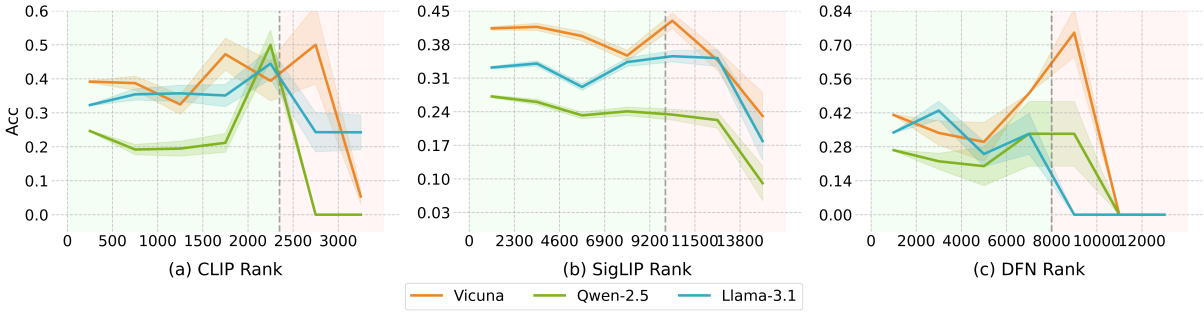


Figure 4: **MLLM Performance distribution across different Rank$_e$ intervals.** Performance of all MLLMs decreases as $Rank_e$ increases across three encoder configurations. The Vicuna-CLIP model shows an 87% performance drop from $0 < Rank_e < 500$ to $Rank_e > 3000$, indicating correlation between performance and vision prior knowledge. This relationship is non-linear with a critical threshold. We marked this threshold by a vertical line in the figure—green on the left indicating sufficient prior knowledge for reasoning, and red on the right showing insufficient knowledge causing sharp performance decline.

from our analysis. Fig. 4 illustrates the relationship between MLLM accuracy and $Rank_e$

Finding 1: MLLM performance correlates positively with vision prior knowledge. As shown in Fig. 4, across all three encoder choices, MLLM performance consistently declines as entity $Rank_e$ increases. For the CLIP encoder, from the interval $0 < Rank_e < 500$ to $Rank_e > 3000$, Vicuna's performance drops by 87%, Llama3.1's by 100%, and Qwen-2.5's by 21%. In SigLIP encoder experiments, overall performance declines by about 50% across all three models from the leftmost to the rightmost interval, while for the DFN encoder, the decline reaches 100%.

Notably, CLIP-Vicuna MLLM does not exhibit a significant performance decline until $Rank_e$ reaches 3000. The phenomenon is also observed in the SigLIP and DFN configurations. This threshold effect suggests that the positive correlation between vision prior knowledge and MLLM performance is not strictly linear, but rather exhibits a mutation beyond a critical point. We posit that this stems from the vision encoder holding a *known* status for entities below a certain $Rank_e$ threshold, meaning it can still provide sufficient prior knowledge for the MLLM to answer entity-related questions. Once $Rank_e$ exceeds this threshold, the vision encoder no longer provides adequate prior knowledge, resulting in a sharp drop in MLLM performance. Considering that LLM part of MLLM possesses adequate knowledge about all entities here, it is the vision encoder of MLLM that constrains the overall performance on entities beyond the threshold.

3.3 Shortcomings of End-to-end Finetuning

To investigate **Q3:** "How to transcend vision prior knowledge limits?", we implement a typical solution as our baseline—finetuning MLLMs on end-to-end domain-specific VQA data. Following established MLLM finetuning approaches (Liu et al., 2024b,a), we freeze the vision encoder parameters and only tune the LLM component. This setup enables the LLM parameters to compensate for limitations in vision prior knowledge.

Vision Encoder	LLM	Number of (Q, A) pairs		Number of entities
		Train	Test	
OpenAI ViT-L-14	Vicuna-7B	1877	531	90
	Llama3.1-8B	2305	624	106
	Qwen2.5-7B	2345	645	109
SigLIP ViT-SO-14	Vicuna-7B	2290	615	106
	Llama3.1-8B	2669	717	123
	Qwen2.5-7B	2614	705	118
DFN ViT-H-14	Vicuna-7B	1914	531	90
	Llama3.1-8B	2339	615	105
	Qwen2.5-7B	2291	618	105

Table 1: **Dataset Statistics.** We report the number of (question, answer) pairs for each dataset split across different encoder-language model combinations. Each corresponding train-test pair shares the same entities.

We constructed our finetuning dataset from Encyclopedia-VQA. Following the method in Section 3.2, we retained questions that MLLMs answered correctly when prompted with "This is {entity_name}" instead of the actual image. After calculating $Rank_e$ across the dataset, we observed naturally different $Rank_e$ distributions across encoders. To balance the distribution of entities with varying levels of prior knowledge, we sampled entities to create more uniform rank distributions for validation. We then divided each subset into training and test sets containing the same entities

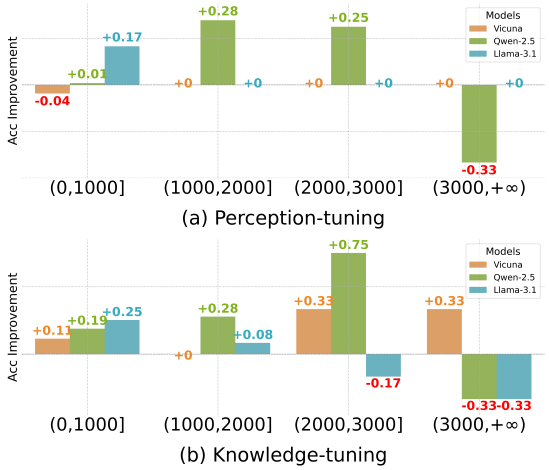


Figure 5: **Perception-tuning and Knowledge-tuning underperform on low-prior (high Rank$_e$) entities.** The figure illustrates performance improvements compared to Zero-shot: Perception-tuning shows a significant drop for Qwen-2.5 when $Rank_e > 3000$. Similarly, Knowledge-tuning leads to notable performance declines for both Qwen-2.5 and Llama-3.1 in the low-prior range ($Rank_e > 3000$).

but with different questions. Dataset statistics are presented in Table 1, with detailed construction methodology in Appendix A.

Successful knowledge-based VQA requires three essential MLLM capabilities: (1) recognizing entities in images; (2) possessing relevant knowledge about these entities; and (3) utilizing this knowledge to answer questions. As the LLM component already contains adequate entity knowledge, MLLM performance can be enhanced through two approaches: (1) improving visual entity recognition and (2) optimizing knowledge utilization for question answering.

To explore these approaches, we develop two distinct types of finetuning data: (1) **Perception-tuning data**, where we transform original Encyclopedia-VQA questions into perception-focused queries such as *What is this image about?* and (2) **Knowledge-tuning data**, which preserves the original questions from Encyclopedia-VQA. Detailed construction methodologies for both datasets are provided in Appendix A.

Finding 2: Domain-specific finetuning with only end-to-end VQA data is insufficient, particularly for entities with low visual prior knowledge. Fig. 5 illustrates the accuracy improvements of Perception-tuning and Knowledge-tuning models compared to Zero-shot baselines under CLIP encoder configuration. As shown in Figure (a), after Perception-tuning, Qwen-2.5 performance decreased in the $Rank_e > 3000$ range, while Vicuna and Llama-3.1 showed no improvement. As shown in Figure (b), after Knowledge-tuning, Qwen-2.5 and Llama3.1's performance decreased for approximately 33% in the $Rank_e > 3000$ range compared to Zero-shot. The comprehensive experimental results across all nine encoder-language model combinations are shown in Table 2.

3.4 Vision Prior Remediation

In previous sections, we established that MLLM performance correlates positively with vision prior knowledge, and that end-to-end fine-tuning yields insufficient. Based on these findings, we propose VisPRE, a training framework that injects entity-related prior knowledge at the vision encoder level to enhance MLLM performance. The specific process of our training framework is illustrated in Fig. 6, which comprises two key stages:

- **Remedy Encoder:** We first reformat the Perception-tuning data into (image, entity_name) pairs, and then fine-tune the vision encoder alongside the text encoder using contrastive loss. This stage enhances the encoder's prior knowledge of entities present in the Perception-tuning data.
- **Instruction Tuning:** We incorporate the fine-tuned encoder into the MLLM architecture and perform end-to-end fine-tuning of the entire model using Knowledge-tuning data. This stage aligns the trained vision encoder with the base LLM and stimulates the model's knowledge of entities.

To systematically evaluate VisPRE, we establish several baselines: Zero-shot, Perception-tuning, and Knowledge-tuning from Section 3.2. Additionally, we include Knowledge-tuning* and Mix-tuning*, where the asterisk (*) denotes unfreezing the vision encoder parameters during fine-tuning. Mix-tuning represents a combination of Knowledge-tuning and Perception-tuning data. The evaluation results are presented in Table 2.

Finding 3: Remediating prior knowledge at the vision encoder level is effective. Perception-tuning shows only marginal improvements over Zero-shot performance, occasionally even degrading results. Knowledge-tuning yields limited gains, with Knowledge-tuning* showing only modest improvement over standard Knowledge-tuning. Mix* doesn't exceed Knowledge* performance. In contrast, our VisPRE framework outperforms all base-

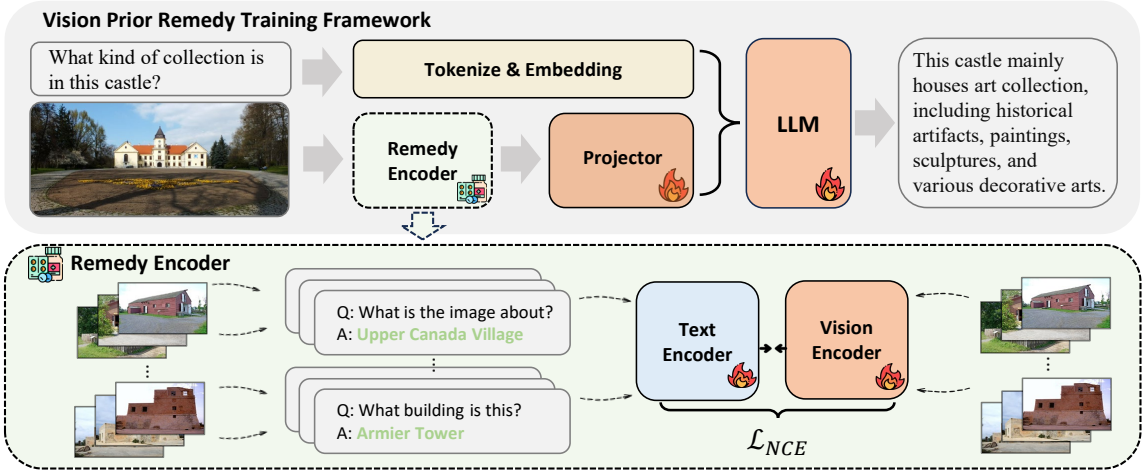


Figure 6: **Overview of our proposed VisPRE framework.** Our framework enriches the vision encoder with entity-specific prior knowledge by first extracting (image, entity_name) pairs from Perception-tuning data and then finetuning the vision encoder using contrastive loss. The enhanced encoder is subsequently integrated into the MLLM, which is further fine-tuned on Knowledge-tuning data.

Vision Encoder	LLM	Zero-shot	Perception	Knowledge	Knowledge*	Mix*	VisPRE(Ours)
OpenAI ViT-L-14	Vicuna-7B	51.22	49.91	54.05	53.48	55.37	**56.31**
	Llama3.1-8B	37.82	39.26	45.67	45.99	44.71	**48.24**
	Qwen2.5-7B	46.05	48.84	54.57	**56.59**	53.49	54.42
SigLIP ViT-SO-14	Vicuna-7B	52.03	53.66	53.66	57.24	57.07	**57.89**
	Llama3.1-8B	38.91	37.66	41.28	**41.84**	41.42	41.28
	Qwen2.5-7B	36.45	36.31	41.13	41.42	42.84	**44.54**
DFN ViT-H-14	Vicuna-7B	59.07	58.70	63.33	64.97	62.90	**66.85**
	Llama3.1-8B	38.70	39.84	45.08	46.99	45.69	**48.29**
	Qwen2.5-7B	40.45	38.10	43.33	44.66	**46.76**	43.69

Table 2: **Results on 9 MLLM combinations.** Our method outperforms finetuning approaches including Perception-tuning, Knowledge-tuning, Knowledge-tuning* and Mix-tuning*, demonstrating that our method significantly enhances MLLM performance through prior remediation. We mark the best result in **bold** for each model, and * indicates unfreezing the vision encoder parameters during fine-tuning.

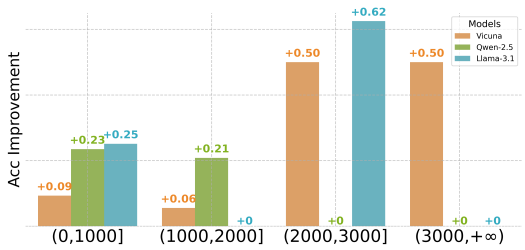


Figure 7: **VisPRE outperforms on all Rank$_e$ levels.** The figure shows performance gains over Zero-shot: With the CLIP encoder, all three models demonstrate improvements across different $Rank_e$ entities, especially for low-prior (high $Rank_e$) entities.

lines, achieving superior results in six of nine model combinations. As shown in Fig. 7, **VisPRE improves MLLM performance across all Rank$_e$ entities, particularly those with low vision priors**, demonstrating clear advantages over alternative tuning approaches in Fig. 5. These results confirm that enhancing encoder prior knowledge substantially expands MLLM capabilities.

4 Case Study

Here we present an illustrative example. As shown in the upper left of Fig. 8, we input an image of the Portuguese Synagogue with the entity-related question: "Where were this synagogue's books sent in 1979?". For **(1) LLM**: The MLLM correctly answers when receiving only the textual description "This is Portuguese Synagogue" instead of the actual image, indicating the LLM component possesses knowledge about this entity. For **(2) MLLM (Original)**: With image input, the MLLM fails to answer correctly. We calculated this entity's $Rank_e$ as 516, indicating low prior knowledge in the visual encoder. **(3) MLLM (SFT)**, despite end-to-end fine-tuning, still fails since the visual encoder's prior knowledge remains unchanged. Our training framework, VisPRE, first injects prior knowledge into the visual encoder, elevating the entity's $Rank_e$ to 10, then conducts end-to-end

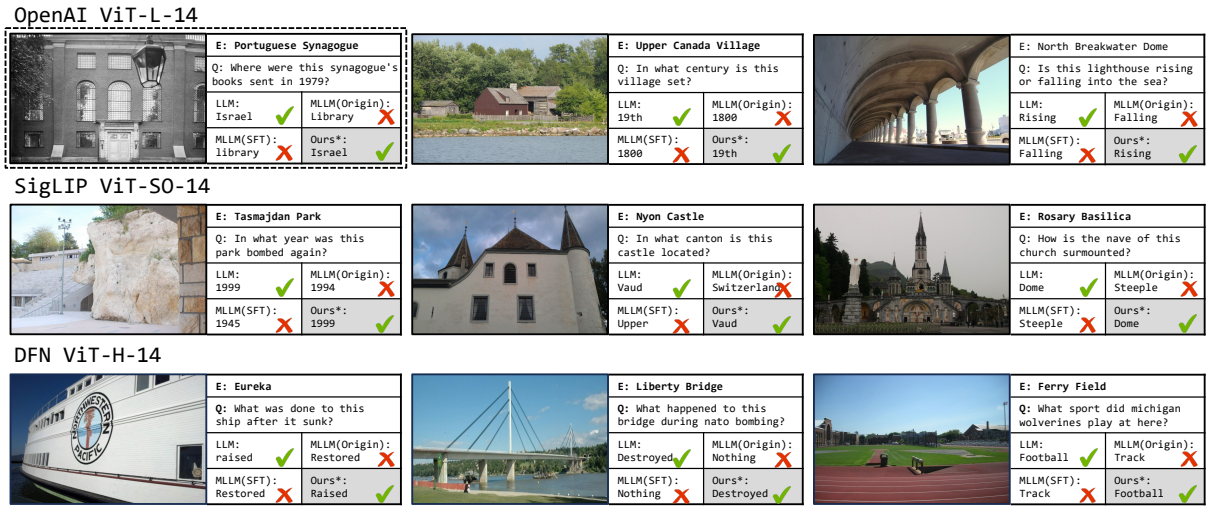


Figure 8: **Examples of Vicuna-7b's responses with different encoders.** When prompted with image description, the LLM answers correctly, demonstrating adequate knowledge of image entities. However, the original (Origin) and fine-tuning with Knowledge-tuning data (SFT) MLLM fails to answer, highlighting the limitations of its vision encoder. With VisPRE(Ours*), the model answer accuratly. For additional cases, refer to Appendix C.

fine-tuning. Consequently, **(4) Ours*** overcomes the visual encoder's limitations and correctly answers the question.

5 Related Works

Multi-modal Large Language Models. MLLMs incorporate visual features into language models, enabling them to perform a wide range of visual tasks. The current MLLM implementations can be classified into two categories. (1) Monolithic MLLMs. Tokenizing different modal inputs uniformly and training the model from scratch (Team, 2024a; Bavishi et al., 2023; Chen et al., 2024b; Zhan et al., 2024), which is computationally expensive. (2) Modular MLLMs. Utilizing pre-trained vision-language models (e.g., CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), DINOv2 (Oquab et al., 2023)) to obtain visual representations of images, and then train MLLMs through cross-modal data, aligning the visual features provided by vision encoder to language model's embedding space. This method is more data-efficient and widely used by open-source MLLMs (e.g., Flamingo (Alayrac et al., 2022), BLIP2 (Li et al., 2023b), LLaVA (Liu et al., 2024b), Qwen-VL (Bai et al., 2023), InternVL2 (Chen et al., 2024b)). Our work focuses on modular multimodal models. While most works treat modular MLLM as a unified system, our research focuses on the impact of vision encoder part on the language model part.

Cross-modality Alignment. With increasing adoption of Modular MLLMs, research focuses on the relationship between vision encoders and MLLM performance. Tong et al. (2024b) found CLIP (Radford et al., 2021) and corresponding MLLMs have similar performance trends across visual modalities, indicating CLIP features cause MLLM deficiencies in these modes, and addressed these by introducing DINOv2 (Oquab et al., 2023) features. Yang et al. (2024) proposed cross-modal alignment metrics to measure vision encoder performance, fitting a binary quadratic polynomial that predicts MLLM performance using that encoder. Different from previous works, our research offers a novel perspective, demonstrating that MLLM performance correlates positively with its vision encoder's prior knowledge.

6 Conclusion

In this paper, we introduce $Rank_e$ to quantify prior knowledge in vision encoder. We find that MLLM's performance is positively correlated with prior knowledge of vision encoder, and end-to-end finetuning MLLM yields insufficient on improving low prior entity performance. To address this issue, we propose VisPRE training framework that enhances MLLM's performance by increasing the prior knowledge within the vision encoder. Our study demonstrates a novel pathway for enhancing MLLM performance, offering substantial value for applications involving uncommon entities.

Limitations

The primary limitation of our study is the current unavailability of VQA datasets with comprehensive rare entity annotations. While our study explores MLLMs' capabilities when confronted with uncommon entities—those inadequately represented in visual encoders' pretraining data, most established entity-annotated datasets like S3VQA (Jain et al., 2021) predominantly feature common entities. To address this challenge, we leveraged the Encyclopedia VQA (Mensink et al., 2023) dataset with its diverse collection of 16.7k entity types, providing a sufficient foundation to identify and analyze less familiar entities. Nevertheless, our findings would benefit from additional specialized datasets explicitly focused on uncommon entities, which would enable a more granular analysis of visual encoders' boundary capabilities and offer complementary insights to our current observations.

Ethics Statement

Our study utilizes MLLMs for knowledge-based VQA tasks. MLLMs may reflect biases present in the training data. Additionally, the VQA data used in our research includes pictures of landscapes and related knowledge questions, which may lead the model to generate offensive content. In this regard, we suggest users to examine the generated outputs cautiously in real-world applications.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering. *Preprint*, arXiv:1505.00468.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2024-06-21.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826.

Patrick Cavanagh. 2011. Visual cognition. *Vision Research*, 51(13):1538–1551. Vision Research 50th Anniversary Issue: Part 2.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? *Preprint*, arXiv:2401.13275.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jessica A Collins and Ingrid R Olson. 2014. Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic bulletin & review*, 21:843–860.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2491–2498.

Botian Jiang, Lei Li, Xiaonan Li, Zhaowei Li, Xiachong Feng, Lingpeng Kong, Qi Liu, and Xipeng Qiu. 2024. Understanding the role of llms in multimodal evaluation benchmarks. *Preprint*, arXiv:2410.12329.

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

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yunxin Li, Xinyu Chen, Baotian Hu, Haoyuan Shi, and Min Zhang. 2024b. Cognitive visual-language mapper: Advancing multimodal comprehension with enhanced visual knowledge alignment. *arXiv preprint arXiv:2402.13561*.

Yunxin Li, Baotian Hu, Xinyu Chen, Yuxin Ding, Lin Ma, and Min Zhang. 2023c. A multi-modal context reasoning approach for conditional inference on joint textual and visual clues. *Preprint*, arXiv:2305.04530.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, Yong Xu, and Min Zhang. 2023d. Lmeye: An interactive perception network for large language models. *Preprint*, arXiv:2305.03701.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. 2024. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124.

OpenAI. 2023. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf. Accessed: 2024-05-26.

OpenAI. 2024. Introducing gpt-4o: our fastest and most affordable flagship model. https://platform.openai.com/docs/guides/vision. Accessed: 2024-05-26.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Steven Pinker. 1984. Visual cognition: An introduction. *Cognition*, 18(1):1–63.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Chameleon Team. 2024a. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 87310–87356. Curran Associates, Inc.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide

shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Shijia Yang, Bohan Zhai, Quanzeng You, Jianbo Yuan, Hongxia Yang, and Chenfeng Xu. 2024. Law of vision representation in mllms. *arXiv preprint arXiv:2408.16357*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. 2024. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9637–9662, Bangkok, Thailand. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

A Datasets

Here we describe the detailed construction process of our dataset. Based on Encyclopedia-VQA (Mensink et al., 2023), we constructed Knowledge-tuning and Perception-tuning datasets for each encoder-language model combination to validate **Finding 2**.

A.1 Preprocess

Question Filtering. First, we focus on improving the parts where MLLM's capabilities are limited by the vision encoder. Therefore, we only retained questions that could be answered by the corresponding LLM when prompted with "This is {entity_name}" instead of the actual image. Next, to ensure that there were no duplicate or similar questions for the same entity across training and test sets, we deduplicated the dataset based on (entity_name, answer) pairs. Finally, we only retained entities with three or more corresponding questions to ensure sufficient questions for dividing into training and validation sets.

Prior Calculation. We calculated $Rank_e$ for all entities in the filtered dataset. We examined the distribution of $Rank_e$ calculated using different types of encoders (CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), DFN (Fang et al., 2023)) across the dataset, as shown in Fig. 10. We found significant variations in $Rank_e$ distributions among different encoders. CLIP's $Rank_e$ values were mostly concentrated in the range of $Rank_e < 400$, with entity counts increasing as $Rank_e$ decreased; In contrast, SigLIP's $Rank_e$ distribution is more uniform, with at least 10 entities present across most $Rank_e$ intervals; DFN's $Rank_e$ distribution was similar to CLIP's, with most values concentrated in the range of $Rank_e < 400$.

Entity Sampling. For SigLIP, we divided $Rank_e$ into intervals of size 1000 and sampled 10 entities from each interval. For CLIP and DFN, using the same sampling strategy as SigLIP would result in insufficient sampling of entities in dense intervals, making it difficult to distinguish different levels of prior knowledge in these regions. Therefore, we adopted a sampling method that approximates the original distributions of CLIP and DFN. We sampled 10 entities from intervals of $0 < Rank_e <= 2$, $2 < Rank_e <= 4$, $4 < Rank_e <= 8$, ..., $512 < Rank_e <= 1024$, $Rank_e > 1024$, ensuring that the sampled distribution approximates the original distribution while retaining all entities with low prior knowledge to reflect the relationship between entity prior knowledge and model performance. Finally, we retained the questions corresponding to the sampled entities and divided the dataset into training and test sets, with statistical information shown in Table 1.

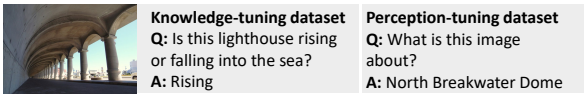


Figure 9: Knowledge-tuning and Perception-tuning datasets

A.2 Construction

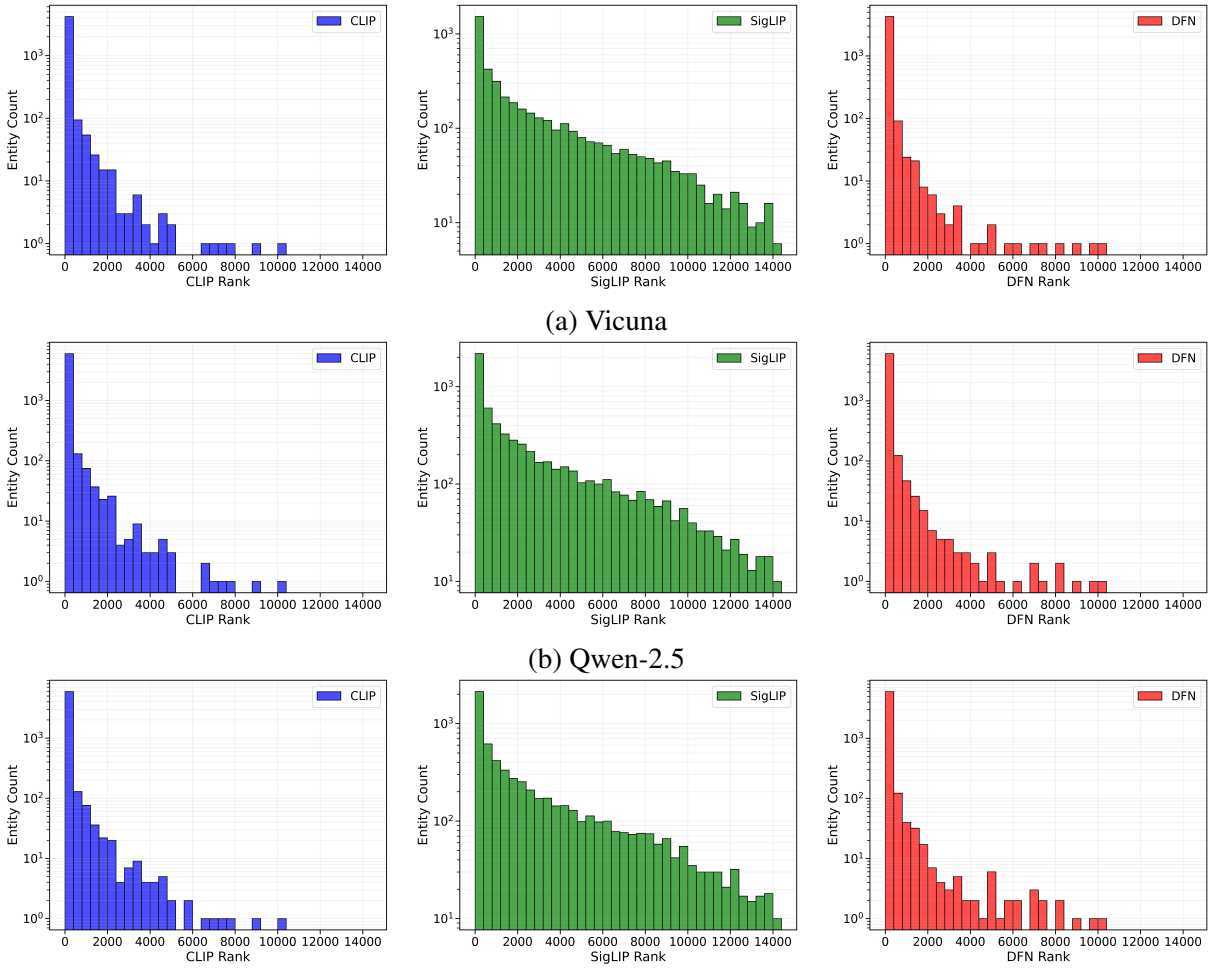
For Knowledge-tuning dataset, we use the original question and answer from the Encyclopedia-VQA dataset. For Perception-tuning dataset, we replace the original question in the Knowledge-tuning dataset with cognitive question like "What is this image about?" and substitute the answers with the entity text corresponding to the image. Examples of Knowledge-tuning and Perception-tuning datasets are shown in Fig. 9.

B Evaluation Settings

We employ Llama-3.1-70B (Dubey et al., 2024) to evaluate the accuracy of MLLM's responses to VQA questions. Specifically, we provide Llama-3.1-70B with the question, entity name (`wikipedia_title` in prompt), ground truth answer, and MLLM's response. The model outputs *true* to indicate a correct answer and *false* to indicate an incorrect answer. The prompt template is shown in Fig. 11, with the `few_shot_examples` shown in Fig. 12.

C More Cases

In Fig. 13, we demonstrated Vicuna-7B's responses under different encoder configurations. Here in Fig. 13, we show examples of responses from Llama-3.1-7B and Qwen-2.5-7B under different encoders.



(a) Vicuna

(b) Qwen-2.5

(c) Llama-3.1

Figure 10: The $Rank_e$ distribution of entities calculated using three different encoders. Here we show the entities that (a)Vicuna, (b)Qwen-2.5 and (c)Llama-3.1 could answer after using text prompts instead of entity images. We can see that the $Rank_e$ distributions for both CLIP and DFN are concentrated in intervals near the left side, while SigLIP's $Rank_e$ distribution is relatively uniform.

Prompt for Llama-3.1 evaluation

You are an expert evaluator tasked with assessing the correctness of model predictions. Your job is to determine if a given prediction is correct based on the provided information. Follow these strict guidelines:

1. You will be given four pieces of information:
- Question: The original question asked
- Wikipedia_title: The title of the Wikipedia article that corresponds to the knowledge base for the question
- Answer: The correct answer(s) to the question, possibly including multiple candidates separated by "|"
- Prediction: The model's prediction to be evaluated

2. Understand that the question is specifically about the entity described in the Wikipedia_title.

3. Compare the prediction to the answer(s), taking into account the context of the question and the Wikipedia_title.

4. Apply these strict criteria:
- The prediction must be accurate and specific.
- If there are multiple candidate answers separated by "|", the prediction must match at least one of them to be considered true.
- For numerical answers, the prediction must be within 10% of at least one correct answer to be considered true.
- For categorical or descriptive answers, the prediction must match the key concepts or categories in at least one of the provided answers.
- Partial or vague answers that don't fully capture the specificity of any correct answer should be considered false.
- Pay close attention to units, specificity, and context provided in the question, Wikipedia_title, and answer(s).

5. Your response must be exactly one word:
- Output "true" if the prediction meets all the criteria for correctness.
- Output "false" if the prediction fails to meet any of the criteria.

6. Do not provide any explanations or additional comments.

`{few_shot_examples}`

Remember, your task is to evaluate the correctness of the prediction based on all the information provided. Be strict in your assessment, but consider all given correct answers. Respond only with "true" or "false".

Question: `{question}`
Wikipedia_title: `{wikipedia_title}`
Answer: `{answer}`
Prediction: `{prediction}`
Evaluation:

Figure 11: Complete prompt for evaluating MLLM responses using Llama-3.1-70B. We prompt the model to determine whether a `prediction` is correct by examining the `question`, `wikipedia_title` (entity name), and `answer`. The model outputs *true* for correct predictions and *false* for incorrect ones. The `few_shot_examples` are shown in Fig. 12

Few-shot examples

Examples:
Question: Along with the mojave desert, in what desert is this plant found?
Wikipedia_title: Acmispon rigidus
Answer: Sonoran Desert
Prediction: Sonoran
Evaluation: true

Question: How many people can this stadium host?
Wikipedia_title: Mercedes-Benz Stadium
Answer: 71,000 | 75,000
Prediction: 73,000
Evaluation: true

Question: When was this novel first published?
Wikipedia_title: To Kill a Mockingbird
Answer: 1960
Prediction: 1962
Evaluation: false

Figure 12: `few_shot_examples` in prompt for Llama-3.1 evaluation. We provide three examples to help the model understand the evaluation requirements.

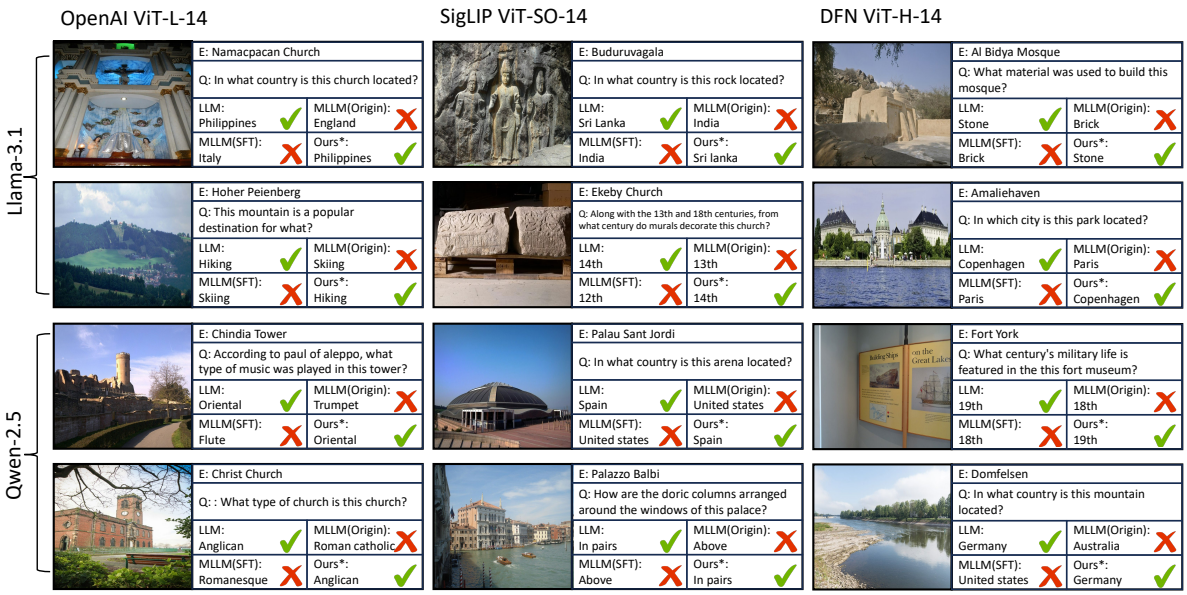


Figure 13: We present examples of Llama-3.1 and Qwen-2.5's responses under three encoder setups. When prompted with text to identify objects in the image, the LLM provides correct answers, demonstrating its knowledge of image entities. In contrast, the MLLM (Origin) fails to respond correctly, highlighting the limitations of its vision encoder. Even after fine-tuning with Knowledge-type VQA data (MLLM SFT), the model still cannot provide accurate answers, revealing the constraints of fine-tuning. Finally, with our Remedy Encoder, the model delivers accurate responses, demonstrating that our method effectively expands the MLLM's visual priors.