EMMA: EFFICIENT VISUAL ALIGNMENT IN MULTI-MODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal Large Language Models (MLLMs) have recently exhibited impressive general-purpose capabilities by leveraging vision foundation models to encode the core concepts of images into representations. These are then combined with instructions and processed by the language model to generate high-quality responses. Despite significant progress in enhancing the language component, challenges persist in optimally fusing visual encodings within the language model for task-specific adaptability. Recent research has focused on improving this fusion through modality adaptation modules but at the cost of significantly increased model complexity and training data needs. In this paper, we propose **EMMA** (**E**fficient **M**ulti-**M**odal **A**daptation), a lightweight cross-modality module designed to efficiently fuse visual and textual encodings, generating instruction-aware visual representations for the language model. Our key contributions include: (1) an efficient early fusion mechanism that integrates vision and language representations with minimal added parameters (less than 0.2% increase in model size), (2) an in-depth interpretability analysis that sheds light on the internal mechanisms of the proposed method; (3) comprehensive experiments that demonstrate notable improvements on both specialized and general benchmarks for MLLMs. Empirical results show that EMMA boosts performance across multiple tasks by up to **9.3%** while significantly improving robustness against hallucinations.

1 INTRODUCTION

Over the past years, Large Language Models (LLMs) have transformed natural language processing (NLP) by demonstrating exceptional abilities in understanding, generating, and reasoning with text across a wide range of tasks; from machine translation and summarization to complex problem-solving and conversational agents (Touvron et al., 2023; Zheng et al., 2023). However, many real-world applications require the ability to process more than just text, such as understanding visual content or synthesizing information from different modalities. This has led to the development of multi-modal LLMs [1], which combine the linguistic strengths of LLMs with vision foundation models, enabling cross-modal understanding and reasoning. By integrating textual and visual information, these models extend the capabilities of traditional LLMs to address tasks like image captioning, visual question answering, and text-to-image generation(Liu et al., 2024b; Alayrac et al., 2022; Achiam et al., 2023).

Current state-of-the-art multi-modal models typically rely on fixed visual feature encodings extracted from vision foundation models, which are projected into the text space and passed to the language model along with instructions (Driess et al., 2023b; Liu et al., 2024a;b). However, the static nature of these encodings, formed without considering the instruction, limits the model's ability to adapt dynamically to specific tasks or contexts. This disconnect between visual and textual components reduces flexibility, making the model less responsive to task-specific nuances. To address this, BLIP-2 (Li et al., 2023b) introduced a cross-attention-based module (called Q-former) to integrate the visual and instruction encodings, a design later adopted by others (Li et al., 2023b; Huang et al., 2023; Ye et al., 2023b). The current state-of-art model, mPLUG-Owl2 (Ye et al., 2024), introduces a *modality adaptation module* that employs attention modules to embed the two modalities into

[1] In this paper, we refer use multi-modality specifically for combination of image and text modalities.

a shared semantic space and thus enhances cross-modality collaboration. mPLUG-Owl2's performance improvement comes with several limitations compared to baseline models. First, mPLUG-Owl2 leverages LLaMA's text embeddings and CLIP's vision encoder to generate the instruction and visual encodings respectively. Therefore the encodings are generated using two distinct models with no initial multi-modality alignment. Second, mPLUG-Owl2's modality-adaptive module introduces roughly 1B more parameters, $3\times$ more than its vision encoder. The modularity adaptation module is then trained from *scratch*, requiring 348 million image-text pairs for pertaining, $300\times$ more than the baseline. Third, the vision encoder requires training during both the pretraining and instruction-tuning stages, which increases the overall training cost and makes the vision encoder more susceptible to loss of generality. Finally, except for a few benchmarks, the model offers only marginal improvements, and in some cases, performs worse than the baseline.

The challenges outlined above led us to explore a more efficient method for modality adaptation. We hypothesized that the need for a complex module for modality adaptation arises from the fact that visual and textual encodings are produced by two entirely separate modules, trained independently. As a result, these complex modules attempt to integrate two distinct spaces, which is inherently difficult. To address this issue, we introduce **EMMA** (**E**fficient **M**ulti-**M**odal **A**daptation), which performs modality fusion through a lightweight modality adaptation mechanism. EMMA integrates CLIP's text encoder with its visual encoder and leverages the pre-trained alignment to adapt visual representations with the instruction via an efficient modality adaptation module (adding less than 0.03% parameters to the model). Our modality adaptation module generates instruction-aware visual representations by attending to more informative, instruction-related tokens, leading to improvements in MLLM-specialized and general benchmarks. A comprehensive set of experiments on benchmarks demonstrates that EMMA significantly enhances cross-modal alignment, improves performance across a range of vision-language tasks, and strengthens the robustness of MLLMs against hallucination. Our contributions can be summarized as follows:

- **Efficient Modality Adaptation**: We introduce a lightweight modality adaptation mechanism that refines visual representations with less than a 0.2% increase in model size, maintaining high efficiency without compromising performance.
- **Comprehensive Analysis of Visual Alignment**: We conduct an in-depth investigation of the Visual Alignment module to provide (1) a detailed understanding of how visual and textual tokens are integrated and (2) an analysis of how effectively the aligned visual representations attend to instructions compared to the initial raw visual encodings.
- **Extensive Empirical Evaluation**: We perform a comprehensive evaluation on both general and MLLM-specialized benchmarks, demonstrating that EMMA significantly improves cross-modal alignment, boosts task performance, and enhances the robustness of multi-modal LLMs.
- **EMMA Outperforms Larger Models**: Compared to mPLUG-Owl2 which has $50\times$ larger modality adaptation module and is trained on $300\times$ more data, EMMA outperforms on 7 of 8 benchmarks. Additionally, compared with BRAVE which has $24\times$ larger vision encoder and is trained on $100\times$ more data EMMA outperforms on all benchmarks.

2 Related Work

Multi-modal Large Language Models (MLLMs). In recent years, there has been significant progress in the development of multi-modal large language models (LLMs) that integrate vision and language to handle tasks requiring both modalities (Zhang et al., 2024; 2023; Wu et al., 2023; Sun et al., 2024; Alayrac et al., 2022; Lai et al., 2023; Li et al., 2023a;g;d; Lin et al., 2024; Liu et al., 2023c;d; Tian et al., 2024; Wang et al., 2024b;c; Chen et al., 2023a). By combining the language understanding of LLMs with the perceptual abilities of vision foundation models, multi-modal LLMs are able to tackle a wide array of tasks that require cross-modal alignment and understanding. We can divide these models into two general categories based on the way that the two visual and textual modalities are integrated. In the first category, including LLaVA (Liu et al., 2024b;a), PaLM-E (Driess et al., 2023a), Shikra (Chen et al., 2023b), etc., the vision encodings are projected to the text space with a few linear layers, and then concatenated with the instruction tokens and passed to the LLMs. In the second category, a more complex module is used for cross-modality adaptation, where both visual and textual encodings are processed through the adaptation module. First introduced with Flamingo (Alayrac et al., 2022) and later adopted by BLIP-2 (Li et al., 2023c),

Method	Vision Encoder	Modality Adapter	LLM	Total
InstructBLIP	1.3B ×3.3	200M×10	7B	7.91B
BLIP-2/InstructBLIP	1.3B×3.3	200M×10	13B	14.2B
Qwen-VL-Chat	1.9B×6.3	80M×4	7.7B	9.6B
BRAVE	**7.2B×24**	100M×5	2.6B	10B
mPLUG-Owl2	0.3B×1	**1000M×50**	7B	8.2B
LLaVA	0.3B×1	20M×0.9	7B	7.3B
EMMA	0.3B	22M	7B	7.3B

Table 1: Paremeter sizes of EMMA compared with state-of-art MLLMs across three components: Vision Encoder, Modality Adapter, and LLM. EMMA uses the same-sized Vision Encoder and LLM as mPLUG-Owl2 but its Modality Adapter, our main contribution, is 50× smaller.

InstructBLIP (Dai et al., 2024), Qwen-VL (Bai et al., 2023), mPLUG-Owl Ye et al. (2023a), and MiniGPT-4 (Zhu et al., 2024), the use of a Q-former (a cross-attention based module) has become a prominent technique.

Enhancing Visual Alignment in MLLMs. Since the emergence of multi-modal LLMs, aligning visual and textual modalities has remained a significant challenge in achieving robust and seamless integration. Previous works (Alayrac et al., 2022; Li et al., 2023c; Ye et al., 2023a;c; Kar et al., 2024) have predominantly focused on utilizing Q-formers and similar cross-attention modules as modality adaptation components to integrate visual and textual embeddings. Recent advancements in modality adaptation include mPLUG-Owl2 (Ye et al., 2023c) and BRAVE (Kar et al., 2024). mPLUG-Owl2 (Ye et al., 2023c) introduces 1B modality adaptation module that employs distinct parameters to project multiple modalities into a unified semantic space for enhanced modality adaptation. On the other hand, BRAVE (Kar et al., 2024) leverages a concatenation of various visual encodings, summing up to 7B, directly feeding into the Q-former. These modality adaptation modules rely on intricate architectures, introducing millions to billions of additional parameters to the model, which significantly increases the computational costs. Recently, proposed a Question-Aware Vision Transformer approach that integrates question-awareness directly into the vision encoder. While this design enhances task-specific performance, it compromises the generality of the visual representations. This added complexity not only demands vast amounts of training data but also imposes considerable overhead during inference. Furthermore, the complication of these systems makes it difficult to discern the primary drivers behind performance improvements — whether they stem from the model complexity, the early fusion of vision and text encodings, or the sheer volume of additional training data. This leads us to the central focus of our work: (1) addressing the inefficiencies in current multi-modal models by proposing a more streamlined approach. Our goal is to achieve an efficient early fusion of visual and textual encodings without significantly increasing the number of parameters or computational costs. (2) Conducting a thorough analysis to determine the key drivers behind performance improvements and the dynamics of the modality adaptation module to generate the multi-modal encodings. This investigation will help isolate the most effective factors and guide future advancements in multi-modal LLMs.

Multi-modal LLM's Benchmarks. The evaluation of multi-modal LLMs relies on a mix of traditional academic benchmarks and newer ones tailored to instruction-following MLLMs. Established benchmarks like VQA-v2 (Goyal et al., 2017) and GQA Hudson & Manning (2019) gauge a model's ability to interpret visuals through open-ended, short-answer questions. ScienceQA (Lu et al., 2022) tests zero-shot generalization in scientific question answering, while VizWiz (Gurari et al., 2018) offers real-world images captured by visually impaired users, challenging models with poor framing, blur, and other quality issues typical of non-professional photos. Additionally, newer benchmarks target instruction-following MLLMs. MathVista (Lu et al., 2023) introduces diverse challenges from mathematical and visual tasks. MMMU (Yue et al., 2024) evaluates multi-modal models on a broad range of college-level tasks requiring deep subject knowledge and reasoning. To assess multi-image reasoning, MUIRBENCH (Wang et al., 2024a) provides a comprehensive benchmark with 12 diverse tasks to evaluate the multi-image understanding abilities of MLLMs.

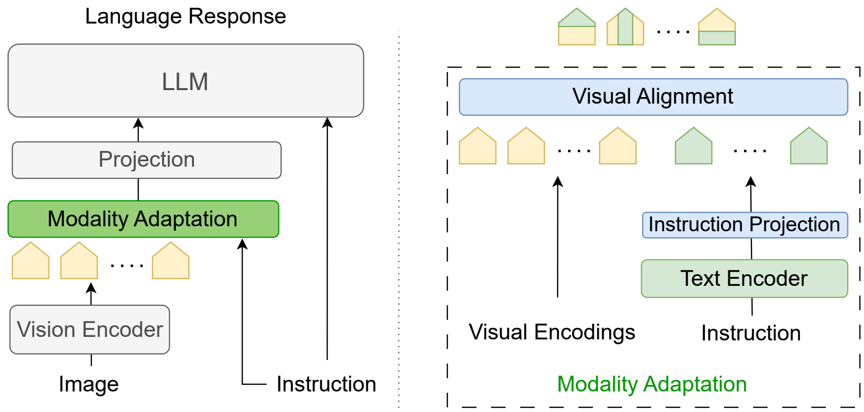


Figure 1: This figure presents the **EMMA's Architecture** where Modality Adaptation is introduced to the standard MLLM architecture. The Modality Adaptation module consists of the Text Encoder, Instruction Projection, and Visual alignment modules. EMMA enhances the multi-modality alignment by incorporating the instruction encodings generated by the Text Encoder which are then projected to the joint space by the Instruction Projection module. The concatenating visual and textual encodings are then passed through the Visual Alignment module, resulting in multi-modal, instruction-aware representations.

For general robustness, MMBench (Liu et al., 2023a) offers a multiple-choice visual question answering benchmark in both English and Chinese, suggesting shuffling the choices to test the model's robustness to the order of options. MMVP (Tong et al., 2024) evaluates robustness by identifying similar images with minute differences and manually pinpointing the visual details the CLIP vision encoder overlooks, which leads to incorrect responses from MLLMs. For hallucination, POPE (Li et al., 2023f) examines the extent of hallucinations across three COCO subsets, and AMBER (Wang et al., 2023), a multi-dimensional benchmark, evaluates generative and discriminative tasks. Additionally, three benchmarks were utilized to assess the proposed method's robustness against hallucination: FOIL (Shekhar et al., 2017), MMRel (Nie et al., 2024), and R-Bench (Wu et al., 2024).

3 Efficient Visual Alignment in Multi-Modal LLMs

In this section, we explain our proposed method which addresses the inefficiencies in current multi-modal models by an efficient early fusion of visual and textual encodings without significantly increasing the number of parameters or computational costs. Moreover, a detailed interpretability analysis is offered to provide insights into the internal mechanisms of the proposed method.

3.1 EMMA: Efficient Multi-Modal Adaptation

Recent progress in multi-modal models has been largely driven by the robust reasoning capabilities of large language models. Therefore, a persistent challenge is to effectively align these two modalities to ensure seamless fusion and task-specific adaptability. Current approaches often rely on complex cross-modality modules, which introduce a significant number of parameters and thus require large amounts of training data. We hypothesize that the need for a complex modality adaptation module arises from the fact that visual and textual encodings are produced by two independently trained components that are themselves unaligned. This is precisely the case for mPLUG-Owl2, which uses CLIP as its vision encoder and LLaMA's text embeddings for its text encoder. As a result, the multi-modality module must, in addition to incorporating text information into the visual embedding, also align the two embeddings.

To address this issue, we propose a simple but surprisingly effective idea—we use both CLIP's vision encoder *and* its text encoder in our multi-modality alignment module. By directly incorporating CLIP's text encoder into its visual encodings, Since CLIP's vision and text encoders were originally jointly trained, multi-modal adaptation is inherently embedded in their encodings, making its text encoder an ideal choice for encoding instructions. The strong, inherent alignment between the two modalities allows for seamless integration, minimizing the need for complex cross-modal modules or extensive training to achieve alignment. Furthermore, CLIP has demonstrated strong performance across diverse tasks, making it a reliable foundation for multi-modal applications.

We refer to our proposed architecture as **EMMA— E**fficient **M**ulti-**M**odal **A**daptation. Figure 1 illustrates EMMA's architecture. The left side highlights the high-level structure, where the standard modules of a multi-modal LLM are shaded in gray, while EMMA's newly introduced Modality Adaptation module is shown in green. On the right, the details of the Modality Adaptation module are depicted. More specifically, let $v(\cdot)$ and $t(\cdot)$ represent the vision encoder and text encoder of CLIP respectively. The visual encodings are defined as $\mathbf{v} = v(\mathbf{x}) \in \mathbb{R}^{n\times d}$, where $\mathbf{x}$ is the input image, and the text encodings as $\mathbf{t} = t(\mathbf{i}) \in \mathbb{R}^{m\times d'}$, where $\mathbf{i}$ is the instruction. Later, the instruction encodings are then processed through the Instruction Projection module $p : \mathbb{R}^{d'} \rightarrow \mathbb{R}^{d}$ which maps the instruction tokens to the same dimensional space as the visual tokens. Once the visual and textual representations are generated, we introduce early fusion via a lightweight module called the Visual Alignment module. This component, consisting of a simple linear layer, combines visual and textual tokens to create the model's multi-modal encoding. The Visual Alignment module, which consists of a linear layer, forms the core of our proposed method. Let $f : \mathbb{R}^{(n+m)\times d} \rightarrow \mathbb{R}^{n\times d}$ represent the Visual Alignment module, where n and m denote the number of visual and textual tokens, respectively. The Visual Alignment module takes the concatenation of the visual and textual tokens as input to generate n refined visual tokens $\tilde{\mathbf{v}}$. The complete pipeline can be expressed as follows:

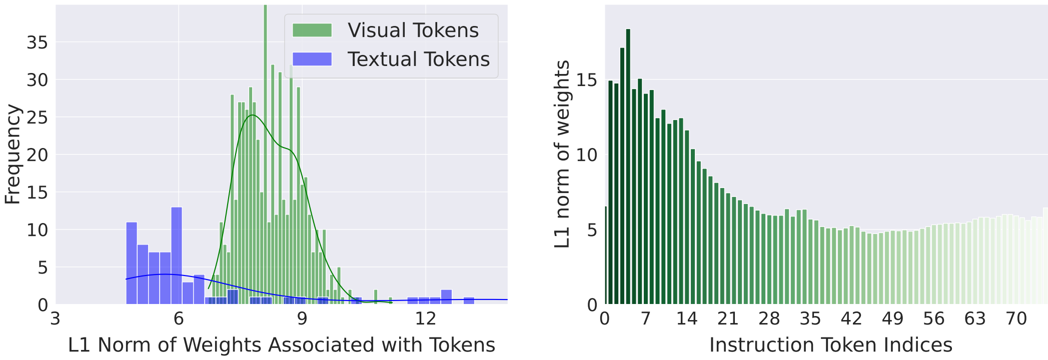
$$\tilde{\mathbf{v}} = f(v(\mathbf{x}),\ p(t(\mathbf{i}))$$

Note that the dimensions of this alignment layer are designed to maintain the same number of visual tokens as the baseline model, ensuring consistency in the number of visual tokens passed to the language model. The Visual Alignment module plays a critical role in ensuring effective alignment between visual and instructions encodings, highlighting the most relevant tokens from the visual encodings in response to the instruction, and thereby delivering more precise visual information to the language model. Its lightweight design facilitates (1) easier interpretability and analysis and more importantly, (2) its simplicity mitigates overfitting when training on small datasets and (3) reduces the training and inference time while outperforming the state-of-the-art models with $10\times$ fewer parameters. Moreover, by leveraging the inherent alignment of CLIP's vision and text encoders, both encoders remain frozen during training. This prevents overfitting to the training data and preserves the generalization capabilities of the original encoders, enabling EMMA to deliver robust performance across diverse tasks.

3.2 Analysis on Modality Adaptation by EMMA

In this section, we analyze the impact of the modality adaptation module introduced by EMMA.

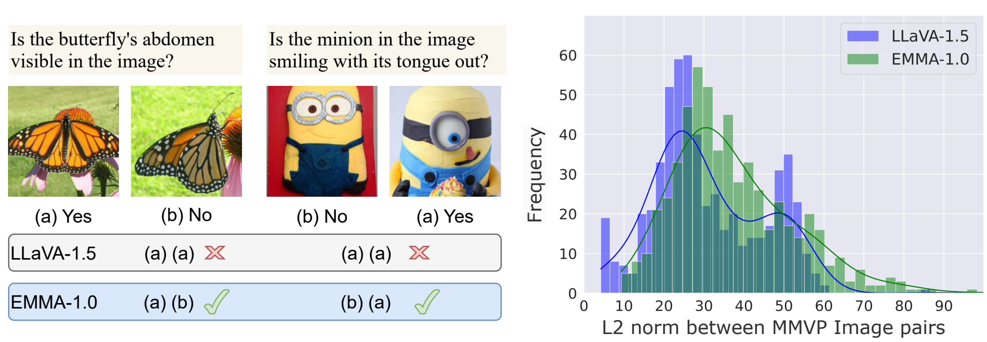
The Mechanics of the Visual Alignment Module. The Visual Alignment module takes the concatenation of the visual and textual tokens ($n + m$ tokens) as input to generate the n refined visual tokens. We begin our examination by scrutinizing the matrix $W \in \mathbb{R}^{(n+m)\times n}$ associated with it. By analyzing the norms of the weights corresponding to each token, we can identify which tokens, visual or textual, are most impactful. The histogram of ℓ_2 and ℓ_1 norms of visual and textual tokens are demonstrated in Figure 2. As the weights of textual tokens are below 1 and the weights of visual tokens are above 1, the ℓ_1 norm of the weights serves as a more indicative measure of token importance. As expected, the visual tokens exhibit higher weights within the Visual Alignment module, signifying their greater influence on the generated multi-modal representation. Another key observation is that certain textual tokens exert more influence than others. To highlight the relative importance of each textual token, Figure 2b presents a bar chart illustrating their respective weights. CLIP's text encoder generates 77 textual tokens, shown along the x-axis. As depicted in the figure, the earlier tokens tend to have a more significant impact. This finding indicates that the alignment module effectively identifies the most informative textual tokens, as instructions typically consist of brief prompts, with the crucial information concentrated in the early tokens and the remaining tokens often masked.



(a) (b)

Figure 2: To evaluate the **contribution of visual and textual tokens to the Aligned Representations**, we analyze the weight matrix of the Visual Alignment module, which consists of a linear layer. **Left**: Fig. 2a presents the ℓ_2 norms of the weights for each token, revealing that (1) visual tokens have a stronger influence on the aligned representations and (2) the impact of textual tokens varies, prompting further investigation in Fig. 2b to identify which specific textual tokens contribute more significantly to the final representations. **Right**: The bar plot reveals that the early tokens are assigned higher weights in the Visual Alignment module, placing greater emphasis on them.

Enhancement in Modality Alignment. The primary objective of EMMA's modality adaptation is to align visual representations with the instructions, ensuring they emphasize the aspects of the image directed by the instructions. Our method achieves this by incorporating instruction encodings into the refinement process of the visual representations. In this section, we explore the alignment capabilities of EMMA by examining the visual representations before and after alignment. To perform this analysis, we utilize the MMVP Tong et al. (2024) benchmark, which is designed to expose the visual shortcomings of MLLMs. We focus on images with visually similar encodings but subtle differences, as shown in Figure 3a. EMMA demonstrates a 9.3% improvement on this benchmark, underscoring its ability to generate more distinguishable visual representations for such images. To empirically validate this, we compute the ℓ_2 norm of visual representations between randomly selected pairs of MMVP images, comparing the results for pre-aligned and post-aligned representations. The histogram of the norms, shown in Figure 3b, reveals a clear shift, indicating that the aligned representations are better at distinguishing between these images by focusing on instruction-relevant tokens.



(a) Instances of MMVP Benchmark (b) ℓ_2 distance between the MMVP image pairs.

Figure 3: **Is EMMA capable of discerning images by focusing on the aspect specified in the instruction?** To address this, we utilize the MMVP benchmark, which contains closely resembling images, as shown in this figure. By leveraging EMMA's visual representations of the image pairs and calculating the ℓ_2 norm, we observe an increase in the distance between them, as illustrated in Figure 3b, indicating EMMA's ability to differentiate between similar images.

Mutual Information between Aligned Visual Tokens and Response Tokens. Another primary objective of the Modality Adaptation module is to adapt the visual representations with the language modality, ensuring they are well-aligned with the language model and encapsulate the necessary information to accurately respond to the given prompt.

To evaluate the contribution of visual tokens to the language model's output, we use the Mutual Information which quantifies the amount of information obtained about one random variable through the other. In this analysis, the LLaVA-In-Wild Liu et al. (2024b) benchmark is employed, which has a set of 24 images with 60 challenging questions in novel domains. For each of the 60 samples, visual representations are generated using the visual modules of both LLaVA and EMMA. It is important to note that EMMA's visual module processes the image and prompt to produce instruction-aware representations, whereas LLaVA generates instruction-agnostic encodings. Additionally, for each sample, the corresponding answer is encoded using the Text Encoder. The Mutual Information between the visual and response encodings is then calculated, with the results shown in Figure 4. As illustrated, the mean mutual information for EMMA is 1.5 times higher than that of LLaVA, underscoring the effectiveness of EMMA's Visual Alignment in steering the model toward accurate language responses.

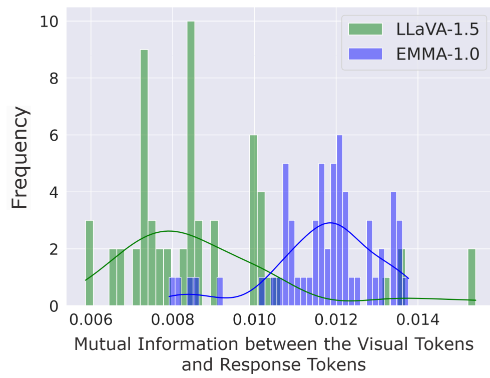


Figure 4: We evaluate the effectiveness of the Visual Alignment module by analyzing how well it refines visual representations to capture instruction-specific details. This is measured by calculating the mutual information between the visual representations and the response encodings produced by the text encoder. As illustrated in the figure, **EMMA's mutual information is 1.5 times higher than LLaVA's**, demonstrating its superior alignment with the instruction.

4 Experimental Evaluation

In this section, we begin by comparing EMMA with state-of-the-art multi-modal LLMs, using the benchmarks introduced earlier. Following this, we conduct a robustness analysis with a focus on hallucination. We conduct an ablation study to identify the optimal layer output from the text encoder for use as textual representations.

Implementation Details. For training, we follow the same two-stage instruction fine-tuning process as LLaVA. In the pretraining stage, only the Visual Alignment and Projection modules are trained, while the language model remains frozen. During the fine-tuning stage, the LLM is unfrozen and fine-tuned along with the two aforementioned modules.

We employ CLIP-ViT-L-14, trained on 336^2 pixel images, as the base image encoder and text encoder. The Visual Alignment module is initialized with the identity matrix for the visual tokens and all zeros for the instruction tokens to transfer all the visual tokens at the beginning of training. Moreover, the Visual Alignment module is designed to maintain the same number of visual tokens as the baseline model. The latest Vicuna v1.5 Zheng et al. (2023) is used as the base LLM. EMMA uses the same set of hyperparameters as the LLaVA-1.5. For all the analysis performed in the Section 3, we use the same dataset as the baseline model, which is 558K and 665K samples for the pretraining and fine-tuning stages respectively. In the Evaluation setting, we have preserved the same pertaining data but scaled the fine-tuning data to 1.2M samples, which is the most efficient data compared to all of the state-of-the-art methods (except for LLaVA) as shown in Table 4.

Benchmarks. We evaluate EMMA across 10 benchmarks that span a diverse set of tasks, including scientific question answering, visual question answering with poor image quality, integrated perception and reasoning tasks, visual dialogue, and general reasoning. The results of recent benchmarks designed specifically for instruction-following large multi-modal models (LMMs) are presented in Table 4, while Table 2 outlines the performance on academic task-oriented benchmarks. EMMA emerges as the state-of-the-art model for 4 out of 5 academic task-oriented benchmarks, outper-

Method	LLM	VE	#Params	PT + IT	VQA^{v2}	VisWiz	SQA^{I}	GQA	OkVQA
BLIP-2	Vicuna-13B	ViT-g/14	13B	129M	65.0	19.6	61	41	45.9
InstructBLIP	Vicuna-7B	ViT-g/14	7.91B	130.2M	-	34.5	60.5	49.2	-
InstructBLIP	Vicuna-13B	ViT-g/14	14.2B	130.2M	-	33.4	63.1	49.5	-
Shikra	Vicuna-13B	ViT-L/14	13b	6.1M	77.4	-	-	-	47.2
IDEFICS	LLaMA-7B	ViT-H/14	9B	354M	50.9	35.5	-	38.4	-
IDEFICS	LLaMA-65B	ViT-H/14	80B	354M	60.0	36.0	-	45.2	-
Qwen-VL-Chat	Qwen-7B	ViT-G/14	9.6B	1.4B	78.2	38.9	68.2	57	56.6
LLaVA-1.5	Vicuna-7B	ViT-L/14	7B	1.2M	78.5	50.0	66.8	**62.0**	-
QA-ViT	Vicuna-7B	ViT-L/14	7B	1.2M	80.5	36.5	-	-	-
mPLUG-Owl2	LLaMA-7B	ViT-L/14	8.2B	348M	79.4	54.5	68.7	56.1	57.7
BRAVE	FlanT5-XL	♣	10B	100M	82.5	54.2	-	52.7	66.0
EMMA	Vicuna-7B	ViT-L/14	7B	1.8M	**89.42**	**56.03**	**73.14**	56.01	**68.57**

Table 2: Comparison between EMMA and previous methods on academic task-oriented datasets. EMMA achieves state-of-the-art performance on 4/5 of the benchmarks while using the fewest number of parameters and training data compared to the previous approaches, utilizing modality adaptation(PT and IT stand for Pretraining and Instruction fine-tuning respectively). The ♣ symbol denotes the collection of vision encoders utilized by BRAVE, including EVA-CLIP-g Fang et al. (2023), ViT-L/14 Radford et al. (2021), SILC-G/16 Naeem et al. (2025), ViT-e Chen et al. (2022), and DINOv2-L/14 Oquab et al. (2023).

forming mPLUG-Owl2 which has $50\times$ larger modality adaptation module and is trained on $300\times$ more data, and BRAVE which has $24\times$ larger vision encoder and is trained on $100\times$ more data. On MLLM specialized benchmarks, EMMA delivers the best performance on 3 out of 5 benchmarks and second best on the two others with less than 0.5% difference. In conclusion, EMMA achieves the best performance in 7 out of 10 benchmarks, despite utilizing the simplest architecture, outperforming other MLLMs that rely on complex modality adaptation modules.

Model	**AMBER**			**R-Bench**		**MMRel**		**FOIL**
	Attr	Rel	Exis	IMG-LVL	INS-LVL	DallE	VG	
LLaVA-1.5	20.23	37.1	91.43	79.89	67.53	59.75	59.45	50.08
EMMA-1.0	**26.15**	**37.34**	**94.50**	**80.78**	**68.12**	**67.7**	**69.27**	**66.13**

Table 3: **Robustness to Hallucination.** This table compares EMMA's robustness against hallucinations with LLaVA, demonstrating consistent improvements over the baseline in avoiding hallucinations.

Robustness & Hallucination. Robustness and the ability to avoid hallucination are essential for multi-modal large language models (MLLMs), which are increasingly applied in areas like medical diagnostics to interpret complex text and image data. Hallucination, a significant security threat to MLLMs, occurs when the model generates information that does not accurately represent the provided images or text. As a result, evaluating and mitigating hallucinations is a critical step in ensuring the reliability of MLLMs before real-world deployment, making it a key focus in model performance assessments. The hallucination evaluations in this section are performed utilizing two benchmarks, AMBER and FOIL, which do not rely on additional LLMs, thereby providing a direct and controlled means of assessing the model's ability to avoid hallucinations. These benchmarks (Wang et al., 2023; Shekhar et al., 2017; Nie et al., 2024; Wu et al., 2024) focus on specific challenges in multi-modal reasoning, testing the model's accuracy in aligning textual and visual content without introducing erroneous information. AMBER consists of 7628, 4924, and 1663 samples for the Attribute, Relation, and Existence categories, respectively. FOIL contains a total of 99,480 test samples, of which 92,705 are straightforward, allowing both LLaVA and our method to successfully avoid hallucinations. However, 6,775 samples present more challenging cases. We compare LLaVA-1.5, the baseline for our approach, with EMMA. The results, shown in Table 3, indicate that EMMA outperforms the baseline, with significant performance gaps in two of the four benchmarks.

Evaluation on OCR Benchmarks. In this section, we evaluate EMMA's performance on a range of Optical Character Recognition (OCR) benchmarks, including OCRbench (Liu et al., 2023b), TextVQA (Singh et al., 2019), InfoVQA (Mathew et al., 2022), ChartQA (Masry et al., 2022),

Method	LLM	VE	#Params	PT + IT	MMB^{EN}	MMB^{CN}	MMMU	MathVista	Muirbench
BLIP-2	Vicuna-13B	ViT-g/14	14.2B	129M	-	-	-	-	-
InstructBLIP	Vicuna-7B	ViT-g/14	7.91B	130.2M	36	23.7	-	-	-
InstructBLIP	Vicuna-13B	ViT-g/14	14.2B	130.2M	-	-	-	**25.3**	-
Shikra	Vicuna-13B	ViT-L/14	13B	6.1M	58.8	-	-	-	-
IDEFICS	LLaMA-7B	ViT-H/14	9B	354M	48.2	25.2	-	19.8	-
IDEFICS	LLaMA-65B	ViT-H/14	80B	354M	54.5	38.1	-	-	-
Qwen-VL-Chat	Qwen-7B	ViT-G/14	9.6B	1.4B	60.6	56.7	**35.9**	-	-
LLaVA-1.5	Vicuna-7B	ViT-L/14	7B	1.2M	64.3	58.3	35.11	21.1	23.46
mPLUG-Owl2	LLaMA-7B	ViT-L/14	8.2B	348M	64.5	-	32.7	22.2	-
BRAVE	FlanT5-XL	♣	10B	100M	-	-	-	-	
EMMA	Vicuna-7B	7B	ViT-L/14	1.8M	**66.44**	**60.15**	35.44	25.1	**32.0**

Table 4: This table compares EMMA with previous methods on specialized MLLM benchmarks. EMMA delivers the best performance on 3/5 benchmarks and secures second place on the remaining two with less than 0.5% difference. The ♣ symbol denotes the collection of vision encoders utilized by BRAVE, including EVA-CLIP-g Fang et al. (2023), ViT-L/14 Radford et al. (2021), SILC-G/16 Naeem et al. (2025), ViT-e Chen et al. (2022), and DINOv2-L/14 Oquab et al. (2023).

and DocVQA Mathew et al. (2021). These benchmarks are specifically designed to assess various aspects of text understanding within visual contexts. OCRbench (Liu et al., 2023b) and TextVQA (Singh et al., 2019) evaluate the ability to extract and interpret text from complex visual inputs, while InfoVQA (Mathew et al., 2022) focuses on understanding textual information in document-like visual contexts. ChartQA (Masry et al., 2022), on the other hand, tests a model's capability to interpret textual and numerical data from charts and graphs. Lastly, DocVQA assesses a model's proficiency in answering questions about scanned documents. The comparison between LLaVA-1.5 and EMMA is presented in Table 5. Notably, EMMA's visual token refinement process does not yield improvements on OCR-specific benchmarks. This finding suggests that the proposed method is most effective when the raw visual encodings contain rich information about the images, allowing the refinement process to remove redundant and unnecessary details while providing instruction-aware representations.

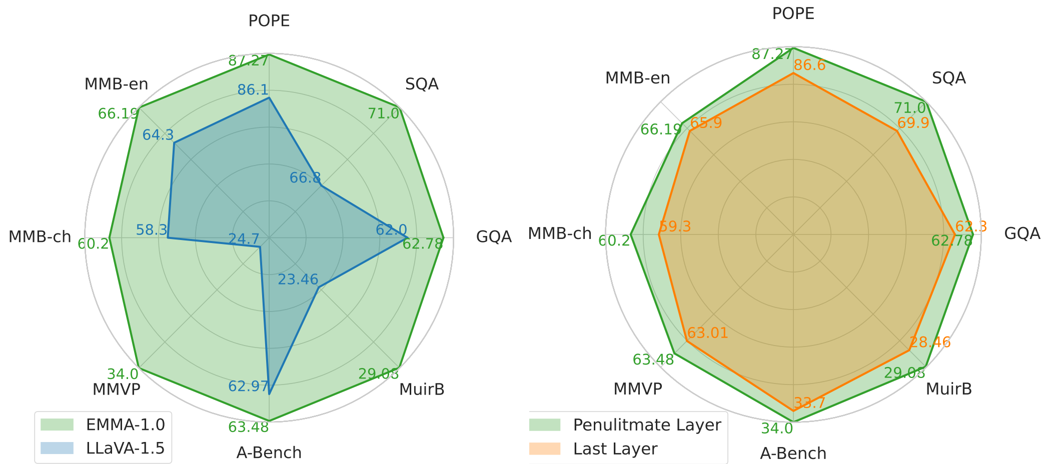
Method	OCRbench	TextVQA	InfoVQA	ChartQA	DocQA
LLaVA-1.5	34.00	**58.20**	14.80	**10.21**	**18.93**
EMMA	**34.67**	57.00	**15.00**	9.90	18.27

Table 5: Performance comparison of LLaVA-1.5 and EMMA on OCR-related benchmarks.

5 Ablation Study.

Training Data. To ensure that any performance improvements are not simply due to the addition of more data, we use the same dataset as the baseline model. Figure 5a demonstrates the improvements achieved by our proposed method, EMMA, across several benchmarks compared to the baseline model, LLaVA-1.5. EMMA consistently outperforms LLaVA-1.5 on tasks such as MMVP (Tong et al., 2024) (+9.3%), MuirBench (Wang et al., 2024a)(+5.62%), SQA (Lu et al., 2022) (+4.2%), MMBench (Liu et al., 2023a) (+1.9%), and POPE (Li et al., 2023e) (+1.17%) showing consistent gains in cross-modal tasks requiring visual and textual understanding.

Ablations on Text Encoder. We ablate the text feature's abstraction level in this section. The textual features can be optionally derived from either the final layer or the penultimate layer of the CLIP Text Encoder. A comparison between the two methods of extracting textual features is presented in Figure 5b, demonstrating the clear superiority of features derived from the penultimate layer. We hypothesize that the final layer of CLIP captures more global and abstract semantics of the instruction, whereas the penultimate layer focuses on finer details. Additionally, since visual features are extracted from the corresponding layer in the visual encoder, using the same layer in the text encoder ensures both modalities operate at a similar level of abstraction, promoting better alignment between them.



(a) (b)

Figure 5: Left: Comparing EMMA with the state-of-the-art model LLaVA-1.5 demonstrates that EMMA surpasses the baseline across a range of benchmarks up to 9.3%. Right: This radar chart compares EMMA's performance when leveraging textual features generated by either the penultimate or last layer of the text encoder, highlighting the advantage of features associated with the penultimate layer.

6 Conclusion

In this work, we addressed the inefficiencies in the modality adaptation modules employed by current multi-modal large language models. We hypothesized that the initial alignment between visual and textual encodings plays a critical role in determining the complexity level of the modality adaptation module and the necessary amount of training data. Our lightweight approach, EMMA (Enhanced Multi-Modal Adaptation), leverages CLIP's text encoder to generate instruction encodings and, by exploiting this initial alignment, demonstrates that the modality adaptation module can be simple while still enhancing the alignment between visual and textual modalities. Through extensive analysis, we demonstrated that EMMA effectively produces instruction-aware visual representations aligned with the language model. Our experiments, evaluated across multiple benchmarks, show that EMMA significantly outperforms state-of-the-art models that use modality adaptation modules $50\times$ larger. Finally, our robustness analysis, particularly in hallucination avoidance, confirmed EMMA's superior ability to accurately process multi-modal data, even in the challenging scenarios.

References

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

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023b.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2022.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 36, 2024.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023a.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023b.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pp. 3608–3617, 2018.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36: 72096–72109, 2023.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pp. 6700–6709, 2019.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pp. 19730–19742. PMLR, 2023c.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023d.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pp. 292–305, 2023f.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023g.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023a.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023b.

Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023c.

Zhaoyang Liu, Zeqiang Lai, Zhangwei Gao, Erfei Cui, Xizhou Zhu, Lewei Lu, Qifeng Chen, Yu Qiao, Jifeng Dai, and Wenhai Wang. Controlllm: Augment language models with tools by searching on graphs. *arXiv preprint arXiv:2310.17796*, 2023d.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pp. 1697–1706, 2022.

Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2025.

Jiahao Nie, Gongjie Zhang, Wenbin An, Yap-Peng Tan, Alex C Kot, and Shijian Lu. Mmrel: A relation understanding dataset and benchmark in the mllm era. *arXiv preprint arXiv:2406.09121*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023.

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

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. In *ICLR*, 2024.

Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024b.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024c.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in lvlms. *arXiv e-prints*, pp. arXiv–2406, 2024.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-docowl: Modularized multimodal large language model for document understanding. *CoRR*, abs/2307.02499, 2023a. doi: 10.48550/ARXIV.2307.02499. URL `https://doi.org/10.48550/arXiv.2307.02499`.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023b.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023c.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13040–13051, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In *ICLR*, 2024.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.