mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration

Qinghao Ye* Haiyang Xu* Jiabo Ye* Ming Yan† Anwen Hu
Haowei Liu Qi Qian Ji Zhang Fei Huang
Alibaba Group
{yeqinghao.yqh, shuofeng.xhy, yejiabo.yjb, ym119608}@alibaba-inc.com

Code & Demo & Models: https://github.com/X-PLUG/mPLUG-Owl/tree/main/mPLUG-Owl2

Abstract

Multi-modal Large Language Models (MLLMs) have demonstrated impressive instruction abilities across various open-ended tasks. However, previous methods primarily focus on enhancing multi-modal capabilities. In this work, we introduce a versatile multi-modal large language model, mPLUG-Owl2, which effectively leverages modality collaboration to improve performance in both text and multi-modal tasks. mPLUG-Owl2 utilizes a modularized network design, with the language decoder acting as a universal interface for managing different modalities. Specifically, mPLUG-Owl2 incorporates shared functional modules to facilitate modality collaboration and introduces a modality-adaptive module that preserves modality-specific features. Extensive experiments reveal that mPLUG-Owl2 is capable of generalizing both text tasks and multi-modal tasks and achieving state-of-the-art performances with a single generic model. Notably, mPLUG-Owl2 is the first MLLM model that demonstrates the modality collaboration phenomenon in both pure-text and multi-modal scenarios, setting a pioneering path in the development of future multi-modal foundation models.

1. Introduction

Large Language Models (LLMs) such as GPT-3 [5], LLaMA [52, 53], and GPT-4 [43] have garnered significant attention due to their exceptional generalization abilities in text understanding and generation. To facilitate the vision-language applications, GPT-4V[1] [42] has recently demonstrated impressive multi-modal capabilities in diverse tasks, e.g., description , question answering, etc., sparking interest among researchers in the potential convergence of the vision-language field. This has led to the emergence of a group of Multi-modal Large Language Models (MLLMs) [4, 14, 28, 35, 61, 62, 64, 71], which aim to enhance LLMs with the ability to understand and handle visual problems.

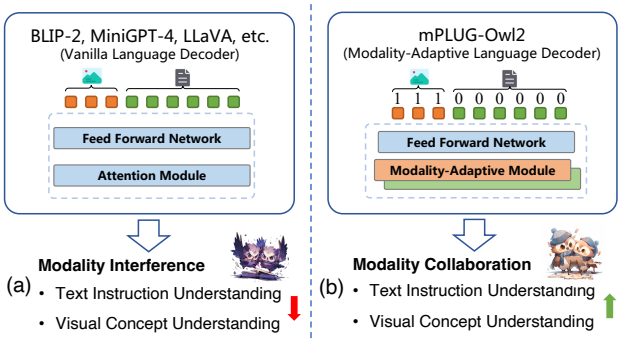

Figure 1. Comparison between existing MLLMs and our proposed model. (a) Previous approaches utilize a standard language decoder (i.e., LLM) to manage different types of instructions, leading to modality interference and performance degradation. (b) We introduce mPLUG-Owl2, which uses a modality-adaptive language decoder to handle different modalities within distinct modules while sharing some parameters for modality collaboration. This approach mitigates the issue of modality interference.

Previous studies [25, 58] in multi-modal learning suggest that different modalities can effectively collaborate, thereby enhancing the performance of both text and multi-modal tasks simultaneously. However, MLLMs is a unified model that supports different modalities and tasks without fine-tuning for specific tasks. Recent works utilize cross-modal alignment modules (e.g., Q-former [14, 28, 71] and linear layer [9, 35]) to map visual features from the vision encoder into the frozen LLMs to carry out multi-modal tasks by leveraging preserved language capabilities. This strategy, unfortunately, restricts the potential of modality collaboration. As a result, some researchers [35, 64] opt to fine-tune LLMs during multi-modal instruction tuning. While fine-tuning significantly improves multi-modal tasks, it risks weakening text task performance [15]. As illustrated in Figure 1,

*Equal contribution
†Corresponding author
[1]https://openai.com/research/gpt-4v-system-card

the challenge of modality collaboration in MLLMs is from applying a single module to balance the gain of modality collaboration and modality interference, where modalities may interfere with each other on a large number of instruction datasets across multiple modalities.

To mitigate this challenge, we present a new general-purpose multi-modal foundation model, mPLUG-Owl2, in this work. Our model features a modularized network design that takes both modality collaboration and modality interference into account, using the language decoder as a universal interface for managing multi-modal signals. Specifically, mPLUG-Owl2 incorporates certain shared functional modules to promote modality collaboration and introduces a modality-adaptive module that serves as a pivot across different modalities. Therefore, vision and language modalities are projected into a shared semantic space for cross-modality interaction, while the proposed module helps preserve modality-specific features. With our novel architecture, modalities with varying information densities are shielded from modality interference due to the modality-adaptive module and can collaborate effectively in capturing shared information. Furthermore, we introduce an innovative two-stage training paradigm that consists of vision-language pre-training and joint vision-language instruction tuning. This paradigm trains the vision encoder across two stages, enabling it to capture both low-level and high-level semantic visual information more effectively.

Extensive experiments illustrate the effectiveness and generalization abilities of mPLUG-Owl2, which achieves state-of-the-art performance on 8 classic vision-language benchmarks using **a single generic model**. Furthermore, it either first or second in performance on 5 recent zero-shot multi-modal benchmarks, underscoring its adaptability and proficiency in multi-modal instruction comprehension and generation. In addition to its cutting-edge performance in multi-modal tasks, mPLUG-Owl2 also achieves state-of-the-art results on multiple pure-text benchmarks. Moreover, we provide in-depth analysis to demonstrate and validate the impact of modality collaboration through our proposed modality-adaptive module, especially in enhancing text tasks, including understanding, knowledge, and reasoning. Finally, comprehensive ablation studies validate the effectiveness of the proposed MLLM training paradigm, which can help inspire the development of future multi-modal foundation models.

2. Related Work

Multi-Modal Large Language Foundation Models. The successful application of Large Language Models (LLMs) has paved the way for developing several approaches aiming to augment the perceptual capacities of LLMs with additional modalities, all within a unified model. There are three primary methods for constructing multi-modal large language foundational models, each showing promise for robust zero-shot generalization capabilities in the vision-language domain. For instance, Flamingo [2] is a forerunner in this area, using a frozen vision encoder and a large language model equipped with gated cross-attention for cross-modality alignment. In contrast, PaLM-E [15] integrates extracted visual features directly through linear layers into the pre-trained PaLM [11] model, which boasts 520 billion parameters, thereby leading to robust performance across numerous real-world applications. This approach has been broadly adopted by models such as LLaVA [35], Shikra [9], etc. One significant limitation of this method, however, is the creation of lengthy visual sequences. To address this, BLIP-2 [28], drawing inspiration from DETR [7], developed a Q-former to reduce the sequence length of visual features efficiently. This design has been mirrored by Kosmos-1 [21], mPLUG-Owl [64], and MiniGPT-4 [71]. Nevertheless, it should be noted that these methods directly align the visual features with the LLMs, treating vision and language signals as equivalent, thereby overlooking the unique granularities between vision and language modalities. To alleviate this problem, we introduce modality-adaptive module. Our proposed model leads to superior performance in both zero-shot and fine-tuning evaluation settings in terms of both image and video.

Instruction Tuning with MLLMs. Instruction tuning optimizes pre-trained large language models to comprehend and adhere to natural instructions, thereby enhancing their ability to generalize unseen tasks in a zero-shot manner. Researchers often employ models such as ChatGPT and GPT-4 [43] to generate diverse and expansive instruction datasets, including those like Alpaca [51], ShareGPT [1], and WizardLM [56]. As multi-modal large language models emerge, research communities are beginning to create high-quality, diverse multi-modal datasets. For instance, MiniGPT-4 [71] utilizes GPT-3.5 to rephrase captions generated by pre-trained models. Concurrently, LLaVA [35], SVIT [68], and LRV-Instruction [33] take advantage of image annotations, such as bounding boxes of objects, image captions, and region descriptions, to prompt GPT-4 to generate instructions and responses using self-instruction methods. Peng et al. [44], Yang et al. [60] leverage in-context learning to improve the generated data. Li et al. [30], Marino et al. [40] directly improve the MLLMs by in-context learning. Models such as mPLUG-Owl [64] and LLaVA-1.5 [34] further advance this area by undergoing joint training with language-only and vision-and-language instruction data, thereby mitigating the risk of catastrophic forgetting of language knowledge. Rather than merely preventing this phenomenon of catastrophic forgetting, mPLUG-Owl2, with the help of the modality-adaptive module, can gain from the collaborative efforts of modalities by being jointly trained with language-only and multi-modal instruction data, thus enhancing both multi-modal and language-only performance.

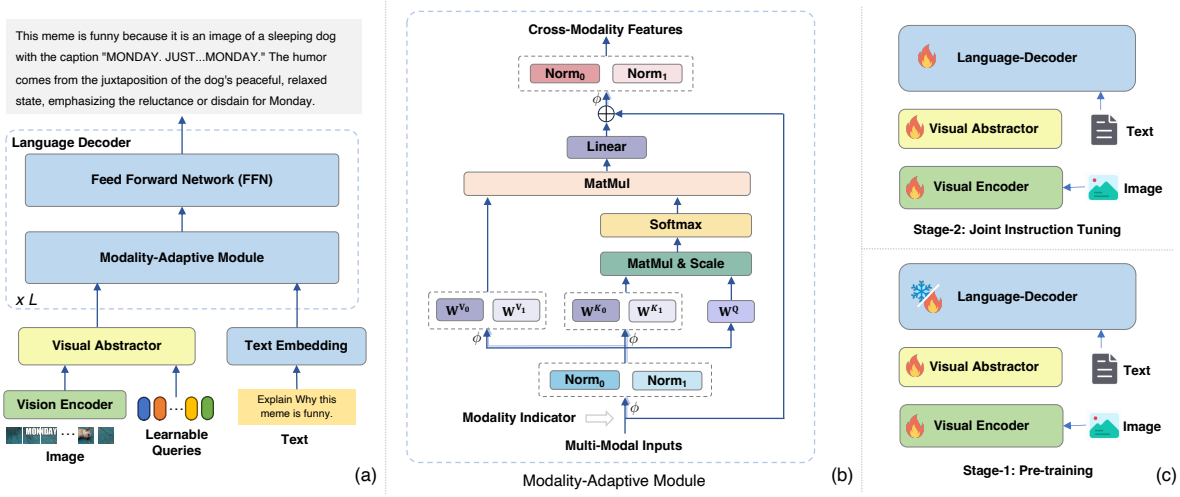

Figure 2. Illustration of the proposed mPLUG-Owl2 and its training paradigm. (a) An overview of mPLUG-Owl2, which consists of a vision encoder, visual abstractor, text embedding layer, and a language decoder. (b) Details of the proposed modality-adaptive module, which takes multi-modal inputs and employs different parameters to project various modalities into a shared semantic space for relational learning while preserving modality-specific features, thereby enabling modality collaboration. (c) The training paradigm of mPLUG-Owl2 involves first pre-training the visual-related modules. Simultaneously, newly added parameters in the language decoder are also learned during the pre-training stage. During the instruction tuning stage, both language instructions and multi-modal instructions are used to jointly train the entire model.

3. Methodology

3.1. Overview

Figure 2 (a) sketches the overview of the mPLUG-Owl2. Specifically, our model comprises a vision encoder, a visual abstractor, a text embedding layer, and a language decoder. Notably, the standard implementation of the text embedding layer and language decoder involves the use of a large language model, such as GPT [5] or LLaMA [52]. We first briefly introduce our model's architecture in Section 3.2. Furthermore, we handle different types of modalities by introducing the modality-adaptive module in Section 3.3. Lastly, we introduce the training paradigm for training mPLUG-Owl2 with modality collaboration in Section 3.4.

3.2. Model Architecture

As depicted in Figure 2, our model, referred to as mPLUG-Owl2, is composed of three main components: a fundamental vision encoder [45], a visual abstractor, and a language decoder. Specifically, we utilize ViT-L/14 as the vision encoder and LLaMA-2-7B [53] as the language decoder. The vision encoder processes an input image with an $H \times W$ resolution and produces a sequence of $\frac{H}{14} \times \frac{W}{14}$ tokens. These visual token features are then combined with text token embeddings and fed into the language decoder that serves as a universal interface that converts various vision-language tasks into text-generation tasks. However, with the increase in image resolution, the encoded visual token sequences can exponentially lengthen. Additionally, the presence of abundant redundancy in the images (e.g., background, similar patches) leads to computational waste and introduces considerable noise. To address this, we propose a visual abstractor equipped with a fixed set of learnable queries to extract higher semantic features from images. Its structure is consistent with the design in mPLUG-Owl [64], which introduces additional modifications on the basis of Perceiver [2]. Specifically, we feed the extracted visual token sequence $\mathcal{I} = [I_1, I_2, \cdots, I_P] \in \mathbb{R}^{P \times d}$ and a fixed number of K learnable queries $\mathcal{Q} \in \mathbb{R}^{K \times d}$ into the visual abstractor. Here, $P = \frac{H}{14} \times \frac{W}{14}$ represents the number of visual patches, and D is the hidden dimension. The visual abstractor consists of a series of visual abstractor layers. In the i-th layer of the visual abstractor, the compressed visual representations $\mathcal{V}^{i+1}$ are computed as follows:

$$\mathcal{C}^i = Attn(\mathcal{V}^i, [\mathcal{I}; \mathcal{V}^i], [\mathcal{I}; \mathcal{V}^i]), \quad (1)$$

$$\mathcal{V}^{i+1} = SwiGLU(\mathcal{C}^i W_1) W_2. \quad (2)$$

Here, $Attn(\cdot, \cdot, \cdot)$ represents the self-attention operation, while $W_1 \in \mathbb{R}^{d \times d'}$ and $W_2 \in \mathbb{R}^{d' \times d}$ are learnable parameters. The function $SwiGLU(\cdots)$ refers to the SwiGLU activation function [48]. We designate $\mathcal{V}^0 = \mathcal{Q}$ to initiate the process. Moreover, to augment the fine-grained perception ability, we integrate sinusoidal positional embeddings with the image feature $\mathcal{I}$ and $\mathcal{V}^i$, thereby preserving positional information, which has been proven essential in [7]. Hence, the computation required by the language decoder decreases from $O((P+L)^2)$ to $O((K+L)^2)$, significantly reducing computational load when $P \gg K$, particularly in scenarios involving multiple images and when the text length L is relatively short. Once the compressed visual

feature is obtained, it is concatenated with text token embeddings and then processed by the language decoder to generate the prediction.

3.3. Modality-Adaptive Module

Prior approaches [14, 35, 64, 71] typically attempt to align visual features with language features by projecting image features into the language semantic space. However, this strategy can cause a mismatch in granularity , where image features often contain fruitful semantic information compared to the discrete semantic information within text embedding features. Those methods disregard the unique characteristics of visual and textual information, thus potentially limiting the model's performance. To this end, we propose a new approach, namely, the Modality-Adaptive Module (MAM), which decouples vision-language representations by projecting visual features and language features into a shared semantic space while preserving the distinctive properties of each modality.

Formally, given a vision-language sequence $X \in \mathbb{R}^{(L_V+L_T)\times d}$ and modality indicators $M \in \{0,1\}^{(L_v+L_T)}$, we first define modality separated operation ϕ as:

$$\phi(X, M, m) = X \odot \mathbb{1}_{\{M=m\}}, \tag{3}$$

where $m \in \{0,1\}$ is the type of modalities (i.e., vision or language). Given the previous layer's output vectors $H_{l-1}, l \in [1, L]$, where L is the number of language decoder layers, we first normalized different modalities into the same magnitude as follows:

$$\tilde{H}_{l-1} = LN_V(\phi(H_{l-1}, M, 0)) + LN_T(\phi(H_{l-1}, M, 1)), \tag{4}$$

where LN_V and LN_T are layer normalization [3] for visual features and language features respectively. Then, we reformulate the self-attention operation by leveraging separated linear projection layers for key projection matrix and value projection matrix while preserving query projection matrix shared as follows:

$$H_l^Q = \tilde{H}_{l-1} W_l^Q, \tag{5}$$

$$H_l^K = \phi(\tilde{H}_{l-1}, M, 0) W_l^{K_0} + \phi(\tilde{H}_{l-1}, M, 1) W_l^{K_1}, \tag{6}$$

$$H_l^V = \phi(\tilde{H}_{l-1}, M, 0) W_l^{V_0} + \phi(\tilde{H}_{l-1}, M, 1) W_l^{V_1}, \tag{7}$$

$$C_l = Softmax\left(\frac{H_l^Q {H_l^K}^\top}{\sqrt{d}}\right) H_l^V, \tag{8}$$

where $W_l^Q, W_l^{K_0}, W_l^{K_1}, W_l^{V_0}, W_l^{V_1} \in \mathbb{R}^{d\times d}$ are the learnable projection matrices, and $C_l \in \mathbb{R}^{(L_V+L_T)\times d}$ is the context features of l-th layer. In this manner, we can calculate the similarities between these two modalities within a shared semantic space, while also preserving the unique characteristics of each modality through different value projection layers. Moreover, by decoupling the key and value projection matrix, we can avoid interference between the two modalities, particularly in relation to granularity mismatch. In a similar vein, we also aim to model these characteristics by using different layer normalization layers. Finally, in order to promote modality collaboration within the same feature space, we maintain a shared FFN for both modalities. As a consequence, the model is able to preserve modality characteristics while achieving modality collaboration via the proposed modality-adaptive module.

3.4. Training Paradigm

As depicted in Figure 2 (c), we employ a two-stage approach in training mPLUG-Owl2, comprising pre-training and visual instruction tuning similar to [35, 64], which aims to align the pre-trained vision encoder and language model during the pre-training phase, and then fine-tune the language model with language modeling loss during the instruction tuning phase. However, we find that simply freezing a pre-trained vision encoder and training a vision-language projector to align visual data with language models can limit their capacity to interpret complex visual information, such as scene text and visual knowledge. To address the issue, we make the vision encoder trainable throughout both the pre-training and instruction tuning stages. This strategy allows the model to capture both low-level and high-level semantic visual information more effectively. Specifically, for the pre-training stage, we enable the vision encoder, visual abstractor, and a part of the modality-adaptive module to be trainable, while keeping the pre-trained language model frozen. Meanwhile, prior research in multi-modal learning [58] has indicated that significant enhancements can be achieved through the collaborative learning of uni-modal and multi-modal sources. Based on this, we adopt a joint training approach by tuning the whole model during the instruction tuning stage, incorporating both text and multi-modal instructions. This methodology enhances the model's comprehension of visual concepts embedded within the text by the multi-modal instructions. Concurrently, the text instruction data augments the model's understanding of intricate natural instructions, thereby ensuring the preservation of its linguistic capabilities.

4. Experiments

4.1. Implementation

Data sets mPLUG-Owl2 is first pre-trained on image-text pairs and fine-tunes on mono-modal and multi-modal instruction data. For pre-training data, we randomly pick about 400 million image-text pairs from five public datasets: Conceptual Captions (CC3M/CC12M) [8], COCO [32], Laion-en [46], COYO [6], DataComp [17]. For instruction data, we collect 5 types of datasets including 1) image captioning (i.e., TextCaps [49], COCO [32]); 2) image question answering (i.e., VQAv2 [19], OKVQA [40], OCR-VQA [41], GQA [22], and A-OKVQA [47]); 3) region-aware QA (i.e., Ref-

Model Type	Method	#Params	Image Caption		General VQA			General VQA (Zero-shot)		
			COCO	Flickr30K (Zero-Shot)	VQAv2	OKVQA	GQA	VizWizQA	TextVQA	SciQA (IMG)
Generalists	BLIP-2 [28]	8.2B	-	74.9	65.0	45.9	41.0	19.6	42.5	61.0
	InstructBLIP [14]	8.2B	102.2	82.4	-	-	49.2	34.5	50.1†	60.5
	Unified-IO$_{XL}$ [38]	2.9B	122.3	-	77.9	54.0	-	57.4‡	-	-
	PaLM-E-12B [15]	12B	135.0	-	76.2	55.5	-	-	-	-
	Shikra [9]	7.2B	117.5	73.9	77.4	47.2	-	-	-	-
	LLaVA-1.5 [34]	7.2B	-	-	78.5	-	**62.0**	50.0	46.1/58.2†	66.8
	Qwen-VL-Chat [4]	9.6B	131.9	81.0	78.2	56.6	57.5	38.9	61.5‡	68.2
	mPLUG-Owl2	8.2B	**137.3**	**85.1**	**79.4**	**57.7**	56.1	**54.5**	**54.3/58.2†**	**68.7**
Specialists	GIT [54]	0.7B	114.8	49.6	78.6	-	-	68.0	59.8	-
	GIT2 [54]	5.1B	145.0	50.7	81.7	-	-	71.0	59.8	-
	PaLI-17B [10]	17B	149.1	-	84.3	64.5	-	71.6	58.8	-

Table 1. **Performance comparison on image caption and visual question answering.** For image caption, CIDEr is reported for evaluation, and accuracy is reported for VQA. Note that specialists are fine-tuned on each individual dataset. † denotes OCR inputs are utilized. ‡ indicates the model has trained on the dataset. We gray out those specialists' methods which are individually fine-tuned on the dataset as well as those fine-tuned results of generalists.

Method	Vision Encoder	Language Model	MME	MMBench	MM-Vet	SEED-Bench	Q-Bench
BLIP-2 [28]	ViT-g (1.3B)	Vicuna (7B)	1293.84	-	22.4	46.4	-
MiniGPT-4 [71]	ViT-g (1.3B)	Vicuna (7B)	581.67	23.0	22.1	42.8	-
LLaVA [35]	ViT-L (0.3B)	Vicuna (7B)	502.82	36.2	28.1	33.5	54.7
mPLUG-Owl [64]	ViT-L (0.3B)	LLaMA (7B)	967.34	46.6	-	34.0	58.9
InstructBLIP [14]	ViT-g (1.3B)	Vicuna (7B)	1212.82	36.0	26.2	53.4	55.8
LLaMA-Adapter-v2 [18]	ViT-L (0.3B)	LLaMA (7B)	1328.40	39.5	31.4	32.7	58.1
Otter [27]	ViT-L (0.3B)	LLaMA (7B)	1292.26	48.3	24.6	32.9	47.2
Qwen-VL-Chat [4]	ViT-G (1.9B)	Qwen (7B)	1487.58	60.6	-	58.2	61.6
LLaVA-1.5 [34]	ViT-L (0.3B)	Vicuna (7B)	**1510.70**	64.3	30.5	**58.6**	60.7
mPLUG-Owl2	ViT-L (0.3B)	LLaMA (7B)	1450.19	**64.5**	**36.2**	57.8	**62.9**

Table 2. **Zero-shot multi-modal evaluation on multi-modal benchmarks** including MME [16], MMBench [36], MM-Vet [66], SEED-Bench [26], and Q-Bench [55]. The overall scores are reported for evaluation. For MMBench and Q-Bench, we report test results.

COCO [65], VisualGenome [24]); 4) multi-modal instruct data (i.e., LLaVA-instruct-150K [35]); 5) text-only instruct data (i.e., ShareGPT-80K [1], SlimOrca [31]). Details can be found in the Appendix.

Training Settings We pre-train the model for 42,500 iterations with a batch size 8,192 for about 348 million image-text pairs. Since we adopt the language modeling loss, the large batch size can be easily achieved by the gradient accumulation technique. mPLUG-Owl2 adopts ViT-L [45] with patch size 14×14 and pre-trained at resolution 224×224. We use the same data augmentation in BLIP-2 [28], including random resized cropping, and horizontal flipping with a probability of 0.5. The number of layers in the visual abstractor is set to 6 and it is randomly initialized. The number of learnable queries is set to 64. For the language model, LLaMA-2 [53] is employed for handling multi-modal features with 7B parameters, and the parameters of modality-adaptive modules are initialized from the language model. We use the AdamW [37] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and ϵ =1e-6 for optimization. The cosine learning rate decay scheduler with a peak learning rate of 1e-4 and with warmup steps 1k. For the learning rate of the vision encoder, we employ layer-wise learning rate decay with a factor of 0.9 to retain the low-level visual representation. For the instruction tuning stage, we train the whole model for 1 epoch with a learning rate of 2e-5 and batch size 256. Besides, we increase the resolution from 224×224 to 448×448. The layer-wise learning rate decay is also employed which is crucial for retaining good visual representation in our experiments.

4.2. Main Results

Image Caption and Visual Question Answering. We assess mPLUG-Owl2 using a wide range of academic benchmarks for evaluating vision-language models. Our evaluation includes eight popular benchmarks, as summarized in Table 1. As the results show, our mPLUG-Owl2 surpasses previous generalist models in both captioning and question answering tasks. Specifically, mPLUG-Owl2 achieves state-of-the-art performance on the Flickr30K datasets, even compared with models with more powerful backbones (e.g., Qwen-VL-Chat [4] and InstructBLIP [14]). Moreover, mPLUG-Owl2 exhibits distinct advantages in visual question answering, especially in OCR-free scenarios, where mPLUG-Owl2 achieves 54.3% accuracy on the TextVQA

dataset in a zero-shot manner, demonstrating the benefits of our training strategy. Also worth noting is that mPLUG-Owl2 shows strong zero-shot performance on the ScienceQA (Image Set) and VizWizQA datasets.

MLLM-oriented Multi-modal Benchmarks. Given the robust zero-shot capabilities of Multi-Modal Language Models (MLLMs), traditional evaluation metrics often fall short in providing a detailed ability assessment. This problem is further exacerbated by their inability to match the given answer accurately, leading to significant robustness issues. To address these challenges, research communities have introduced a series of benchmarks including MME [16], MMBench [36], MM-Vet [66], SEED-Bench [26], and Q-Bench [55]. These benchmarks systematically structure and evaluate complex multi-modal tasks. We applied our model, in a zero-shot manner, to five recently popular multi-modal benchmarks. For a fair comparison, we select models with similar language model sizes, particularly those from the LLaMA family, and detail their differences in the vision encoder. The results of our evaluation are listed in Table 2. In the table, mPLUG-Owl2 achieves higher zero-shot performance in terms of MMBench, MM-Vet, and Q-Bench. Conversely, the performance on MME is lower because of the limited number of test samples in MME, which could potentially lead to sensitive fluctuations in performance. Particularly, it exhibits significant improvement on Q-Bench, a benchmark for examining the low-level visual perception of MLLMs. This improvement occurs when applying a smaller visual backbone (i.e., ViT-L), leading to enhanced low-level visual perception. This demonstrates the effectiveness of our training strategy for training visual backbone.

Method	MMLU	BBH	AGIEval	ARC-c	ARC-e
LLaMA-2 [53]	46.8	38.2	21.8	40.3	56.1
WizardLM [56]	38.1	34.7	23.2	47.5	59.6
LLaMA-2-Chat [53]	46.2	35.6	28.5	54.9	71.6
Vicuna-v1.5 [69]	51.1	41.2	21.2	56.6	72.8
mPLUG-Owl2	**53.4**	**45.0**	**32.7**	**65.8**	**79.9**

Table 3. **Performance on pure-text benchmarks of mPLUG-Owl2** compared to LLaMA-2 (7B) family variants. We adopt 5-shot for MMLU and 0-shot for BBH, AGIEval, and ARC as [13].

Natural Language Understanding and Generation. Current MLLMs often outperform in various multi-modal downstream tasks by leveraging the power of large language models. Nevertheless, the intrinsic capabilities of these models often play a significant role in determining the performance of MLLMs, an aspect that has often been overlooked in prior multi-modal language model studies. Accordingly, we have also assessed the performance of our model in the context of natural language understanding and generation. We perform the evaluation on MMLU [20], BBH [50], AGIEval [70] and ARC [12]. The results are illustrated in Table 3. As observed in the table, mPLUG-Owl2 excels in examination and reasoning, showing a significant improvement on MMLU and BBH by 2.3% and 3.8% respectively. This indicates that mPLUG-Owl2 not only performs well on multi-modal tasks but also achieves better performance compared to the other instruction-tuned LLMs, showing the promising way for developing strong MLLMs.

Method	MSRVTT-QA		MSVD-QA		TGIF-QA	
	Accuracy	Score	Accuracy	Score	Accuracy	Score
Exacting Match						
Flamingo-80B [2]	17.4	-	35.6	-	-	-
FrozenBiLM [59]	16.8	-	32.2	-	41.0	-
BLIP-2 [28]	9.2	-	18.3	-	-	-
HiTeA [63]	21.7	-	37.4	-	-	-
InstructBLIP [14]	22.1	-	41.8	-	-	-
mPLUG-Owl2	**23.6**	-	**42.4**	-	**61.6**	-
GPT-Assisted						
Video Chat [29]	45.0	2.5	56.3	2.8	34.4	2.3
LLaMA-Adapter [18]	43.8	2.7	54.9	3.1	-	-
Video-LLaMA [67]	29.6	1.8	51.6	2.5	-	-
Video-ChatGPT [39]	**49.3**	2.8	64.9	3.3	51.4	3.0
mPLUG-Owl2	46.7	**2.9**	**65.4**	**3.5**	**67.1**	**3.7**

Table 4. **Zero-shot evaluation on video question answering.** Accuracy and relevance score are reported.

Zero-Shot Video Question Answering. Given that videos can be viewed as a sequence of images, we conducted a comprehensive quantitative evaluation using several commonly employed video question-answering datasets, including MSRVTT-QA [57], MSVD-QA [57], and TGIF-QA [23]. These datasets aided in the zero-shot evaluation of the model's ability to understand video content, with the results summarized in Table 4. We employed two types of evaluations: 1) Exact matching, which is commonly used in previous video question-answering evaluations; and 2) GPT-assisted evaluation [39] that assesses the model's capabilities by measuring the accuracy of the model's generated predictions and providing a relative score on a scale of 1-5. We observe that our model achieves superior results on all three video datasets under a zero-shot setting. Furthermore, in terms of relevancy, our model generates more accurate answers than other video MLLMs, thereby demonstrating its superiority and excellent generalization capabilities.

4.3. Discussion

Modality Collaboration for Text Performance. To demonstrate how modality collaboration enhances not only the multi-modal performance but also the text capability of MLLMs, we evaluate the performance of text benchmarks in terms of various abilities including examination, knowledge, understanding, and reasoning. As observed in Figure 3, both examination and knowledge capabilities of MLLMs have significantly improved thanks to the benefits of modality collaboration facilitated by the modality-adaptive module. This improvement arises because multi-modal data

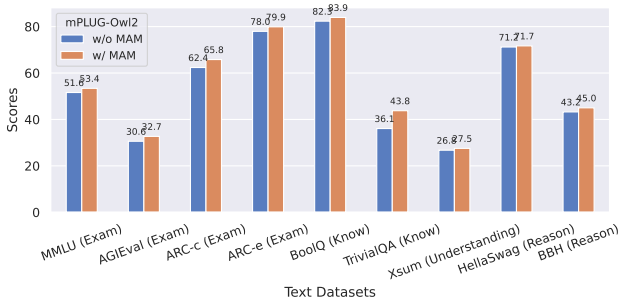

Figure 3. Performance of text benchmarks across various capabilities under modality collaboration.

MAM	Text Inst.	MM Inst.	VQAv2	Q-Bench	MMLU	BBH
	✓		58.2	54.4	51.8	43.6
		✓	76.3	61.3	45.4	25.7
	✓	✓	76.2	60.3	51.6	43.2
✓	✓		60.5	55.6	51.8	44.0
✓		✓	76.5	60.2	46.1	30.6
✓	✓	✓	**76.8**	**62.2**	**52.8**	**45.0**

Table 5. Performance comparison among different types of instruction data and structures.

allows the model to utilize visual information to understand concepts that cannot be described through language. Similarly, the model can generate richer and more substantial responses due to a more concrete understanding of these concepts. Additionally, multi-modal data enhances the reasoning ability of the model because images contain rich information (such as relationships and spatial aspects). The model learns from these aspects and associates them with the text, thereby indirectly enhancing the reasoning ability of the text.

Impact of Joint Vision-Language Instruction Tuning. Table 5 presents the results of instruction tuning with various types of data as well as whether using modality-adaptive module. These results show that even without multi-modal instruction data, the model's performance on multi-modal benchmarks is respectable due to the effective vision-language alignment achieved during pre-training. However, when solely using multi-modal instruction data, we observe an increase in performance on multi-modal datasets, while performance on text tasks decreases by about 5.7%. This phenomenon can be counterbalanced by the joint vision-language tuning proposed, as shown in the table's third row, where the multi-modal performance begins to slightly decrease due to modality interference. To counter this drawback, we apply our proposed modality-adaptive module to the model. Results show that the performance on both multi-modal and text benchmarks improves, with a minimum increase of 0.6% on the VQAv2 dataset and 1.6% on MMLU.

Impact of Trainable Vision Encoder. Table 6 delivers the performance of the training vision encoder during instruction tuning with modality collaboration. It can be observed

Unfreeze	Layer-wise lr.	VQAv2	TextVQA	MMBench	Q-Bench
		74.8	39.8	63.8	60.7
✓		76.2 **(+1.4)**	40.3 **(+0.5)**	62.7 **(-1.1)**	61.6 **(+0.9)**
✓	✓	76.8 **(+2.0)**	42.5 **(+2.7)**	64.5 **(+0.7)**	62.2 **(+1.5)**

Table 6. Influence of learning strategies for visual encoder.

# Learnable Queries	VQAv2	TextVQA	MMBench	Q-Bench
8	58.3	18.6	47.6	52.4
16	66.2	28.5	52.9	54.9
32	72.4	36.3	60.2	57.8
64	**76.8**	42.5	**64.5**	**62.2**
128	76.7	**44.4**	63.6	61.6

Table 7. Performance in terms of number of learnable queries.

that enabling the vision encoder to be trainable improves performance on VQAv2 and Q-Bench by at least 1.4% and 0.9%, respectively, suggesting the benefits of modality collaboration. Conversely, it results in a 1.1% performance drop in MM-Bench, indicating a degree of forgetting and damage to the general visual representation due to the limited diversity of instruction data. To mitigate this challenge, we apply layer-wise learning rate decay with an exponential decay factor of 0.9, which preserves the representation of lower layers while modifying higher semantic representations. By applying the layer-wise learning rate decay, we can notice that performance on TextVQA has increased further with 2.2%, showing the effectiveness of our training strategy.

Impact of Number of Learnable Queries. To investigate the effect of the number of learnable queries $\mathcal{Q}$, we conduct experiments using different numbers of queries in the visual abstractor, as shown in Table 7. It can be observed that the model consistently exhibits improvement as the number of learnable queries increases until it reaches a saturation point, suggesting that 64 may be the optimal number for representing an image. Notably, there is a significant performance boost observed when the number is increased from 8 to 64, e.g., the performance of VQAv2 is increased 18.5%. These findings suggest that a higher number of learnable queries can capture image information more comprehensively, thereby enhancing the model's image comprehension capabilities.

Resolution	VQAv2	TextVQA	MMBench	MM-Vet	Q-Bench
224×224	76.8	42.5	64.5	34.0	62.2
336×336	78.5 **(+1.7)**	49.8 **(+7.3)**	65.2 **(+0.7)**	34.6 **(+0.6)**	62.4 **(+0.2)**
448×448	79.4 **(+2.6)**	54.3 **(+11.8)**	65.4 **(+0.9)**	36.2 **(+2.2)**	62.6 **(+0.4)**

Table 8. Influence of different input image resolutions.

Impact of Image Resolution. Image resolution plays a crucial role in vision-language tasks, as a higher resolution can reduce image blur and improve understanding of fine-grained details. To explore the impact of image resolution on performance across different benchmarks, we adjust the image resolution from 224×224 to 448×448 and the results

are listed in Table 8. As observed in the table, using a higher resolution proves advantageous for multi-modal tasks, particularly in the question answering scenario. Specifically, the performance of VQAv2 has increased from 76.8 to 79.4, representing a 2.6% boost. Simultaneously, there is an 11.8 point lift in the TextVQA benchmark when enlarging the resolution from 224×224 to 448×448. This suggests that OCR-related tasks benefit significantly from increasing the resolution.

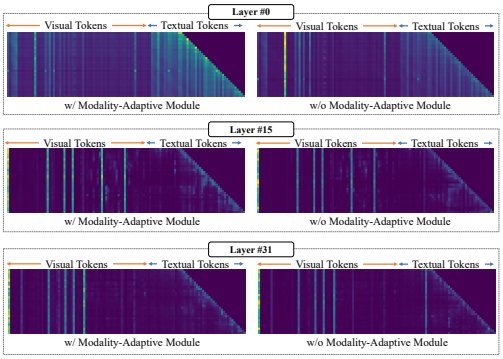

Figure 4. Visualization of the attention maps with and without the Modality-Adaptive Module. We demonstrate the attention maps for the 0-th, 15-th, and 31-st layers.

4.4. Qualitative Analysis

Impact of Modality-Adaptive Module in Multi-Modal Scenario. We investigate the impact of the Modality-Adaptive Module in multi-modal scenarios by visualizing the attention maps of mPLUG-Owl2 with and without this module using image caption input, as shown in Figure 4. Each attention map illustrates the attention scores of generated tokens on the input sequence during the generation process. It can be observed that regardless of whether the MAM is incorporated or not, the model focuses more on the textual tokens in the earlier layers while paying more attention to the visual tokens in the later layers. This suggests that the modeling of visual and textual information plays different roles in the collaboration of MLLMs. An intuitive explanation is that MLLMs initially use syntactic information to comprehend instructions and then identify relevant visual content tokens by considering the textual input. When using the MAM, it can be observed that the model explicitly pays more attention to the textual content in the earlier stages and focuses more on the visual content in the later stages. The MAM prevents visual and textual tokens from being treated as the same and encourages collaboration between different modalities.

Impact of Modality-Adaptive Module in Unrelated-Modality Scenarios. In this example, we feed a question: "What are the seven colors of the rainbow?" along with a randomly selected image to investigate the impact of our module on data that contains unrelated modalities. The responses and attention maps of the model are shown in Figure 5. Our proposed model, mPLUG-Owl2, which incorporates the MAM, accurately identifies all seven colors. During the generation process, it can be observed that the model primarily focuses on the textual input. On the other hand, when the MAM is not utilized, mPLUG-Owl2 only identifies six colors. The model's ability to comprehend text instructions is disrupted, and it is also evident that it places more emphasis on the image during generation. Thanks to the MAM, mPLUG-Owl2 is better able to capture modality-specific features when modeling multimodal inputs. This enhances the adaptability of modality collaboration, resulting in reduced disturbance when the text and image are unrelated.

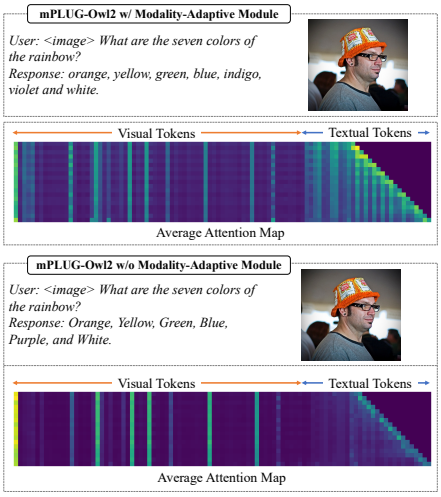

Figure 5. Visualization of the attention maps with and without the Modality-Adaptive Module.

5. Conclusion

In this paper, we present mPLUG-Owl2, a highly capable generalist model by leveraging modality collaboration for enhancing performance across both text and multi-modal tasks. The inclusion of shared functional modules and a modality-adaptive module in mPLUG-Owl2 strengthens the model's ability to harmonize modality collaboration and preserve modality-specific characteristics. The extensive experimental evaluations highlight mPLUG-Owl2's proficiency in generalizing across various tasks, thereby achieving state-of-the-art performances with a singular, generalized model. Most notably, mPLUG-Owl2 stands as the first MLLM model to exhibit the phenomena of modality collaboration in both pure-text and multi-modal contexts. This not only enhances the model's vision-language understanding but also improves its language capabilities in terms of understanding, knowledge, and reasoning. This represents a significant contribution to the field and opens up exciting opportunities for the future development of multi-modal foundation models.

References

[1] Sharegpt. http://sharegpt.com, 2023. 2, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2022. 2, 3, 6

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016. 4

[4] 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, abs/2308.12966, 2023. 1, 5

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. ArXiv, abs/2005.14165, 2020. 1, 3

[6] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 4

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020. 2, 3

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3558–3568, 2021. 4

[9] Ke Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. ArXiv, abs/2306.15195, 2023. 1, 2, 5

[10] 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. arXiv preprint arXiv:2209.06794, 2022. 5

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. J. Mach. Learn. Res., 24:240:1–240:113, 2022. 2

[12] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018. 6

[13] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023. 6

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. ArXiv, abs/2305.06500, 2023. 1, 4, 5, 6

[15] Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. In International Conference on Machine Learning, 2023. 1, 2, 5

[16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023. 5, 6

[17] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108, 2023. 4

[18] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, W. Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Jiao Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. ArXiv, abs/2304.15010, 2023. 5, 6

[19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 4

[20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measur-

ing massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2020. 6

[21] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. ArXiv, abs/2302.14045, 2023. 2

[22] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6700–6709, 2019. 4

[23] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2758–2766, 2017. 6

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123:32–73, 2017. 5

[25] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. Masked vision and language modeling for multi-modal representation learning. arXiv preprint arXiv:2208.02131, 2022. 1

[26] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125, 2023. 5, 6

[27] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. ArXiv, abs/2305.03726, 2023. 5

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. ArXiv, abs/2301.12597, 2023. 1, 2, 5, 6

[29] 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, 2023. 6

[30] Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. arXiv preprint arXiv:2312.01571, 2023. 2

[31] Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification, 2023. 5

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 4

[33] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565, 2023. 2

[34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. ArXiv, abs/2310.03744, 2023. 2, 5

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. ArXiv, abs/2304.08485, 2023. 1, 2, 4, 5

[36] 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, 2023. 5, 6

[37] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 5

[38] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. ArXiv, abs/2206.08916, 2022. 5

[39] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. ArXiv, abs/2306.05424, 2023. 6

[40] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pages 3195–3204, 2019. 2, 4

[41] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In 2019 international conference on document analysis and recognition (ICDAR), pages 947–952. IEEE, 2019. 4

[42] OpenAI. Gpt-4v(ision) system card. 2023. 1

[43] OpenAI. Gpt-4 technical report. ArXiv, abs/2303.08774, 2023. 1, 2

[44] Yingzhe Peng, Xu Yang, Haoxuan Ma, Shuo Xu, Chi Zhang, Yucheng Han, and Hanwang Zhang. Icd-lm: Configuring vision-language in-context demonstrations by language modeling. arXiv preprint arXiv:2312.10104, 2023. 2

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021. 3, 5

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems, 35:25278–25294, 2022. 4

[47] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge.

In European Conference on Computer Vision, pages 146–162. Springer, 2022. 4

[48] Noam Shazeer. Glu variants improve transformer. arXiv preprint arXiv:2002.05202, 2020. 3

[49] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pages 742–758. Springer, 2020. 4

[50] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. arXiv preprint arXiv:2210.09261, 2022. 6

[51] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 2

[52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. ArXiv, abs/2302.13971, 2023. 1, 3

[53] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. ArXiv, abs/2307.09288, 2023. 1, 3, 5, 6

[54] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100, 2022. 5

[55] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. arXiv preprint arXiv:2309.14181, 2023. 5, 6

[56] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. ArXiv, abs/2304.12244, 2023. 2, 6

[57] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In Proceedings of the 25th ACM international conference on Multimedia, pages 1645–1653, 2017. 6

[58] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In International Conference on Machine Learning, 2023. 1, 4

[59] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In NeurIPS, 2022. 6

[60] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Geng Xin. Exploring diverse in-context configurations for image captioning. arXiv preprint arXiv:2305.14800, 2023. 2

[61] 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, 2023. 1

[62] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In The 2023 Conference on Empirical Methods in Natural Language Processing, 2023. 1

[63] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15405–15416, 2023. 6

[64] 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, 2023. 1, 2, 3, 4, 5

[65] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 69–85. Springer, 2016. 5

[66] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490, 2023. 5, 6

[67] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. ArXiv, abs/2306.02858, 2023. 6

[68] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. arXiv preprint arXiv:2307.04087, 2023. 2

[69] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan

Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 6

[70] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. arXiv preprint arXiv:2304.06364, 2023. 6

[71] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. ArXiv, abs/2304.10592, 2023. 1, 2, 4, 5