

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>

A. Additional Experimental Results

In this section, we provide more experimental results for the completeness of our proposed method.

Method	LLM Tuning Method	VQAv2	MMBench	Q-Bench	MMLU
LLaVA-1.5 [18]	Full	78.5	62.7	58.7	50.4
LLaVA-1.5 [18]	Freeze	70.6 (-7.9)	48.2 (-14.5)	32.6 (-26.1)	51.1 (+0.7)
LLaVA-1.5 [18]	LoRA	79.1 (+0.6)	65.0 (+2.3)	58.4 (-0.3)	50.2 (-0.2)
LLaVA-1.5 [18]	Full w/ MAM	79.2 (+0.7)	66.1 (+3.4)	59.5 (+0.8)	52.5 (+2.1)
LLaVA-1.5 [18]	LoRA w/ MAM	79.2 (+0.7)	66.0 (+3.3)	59.7 (+1.0)	52.0 (+1.6)

Table 1. Illustration the phenomenon of modality collaboration and interference. For the LoRA version, we directly utilize the official checkpoint¹, which might tuned carefully. We follow the original implementation using the input resolution of 336×336 .

A.1. Modality Collaboration and Interference.

We illustrate the phenomena of modality collaboration and interference using a recent robust benchmark, LLaVA-1.5 [18]. We adhere to the official instructions to train the model under different LLM tuning conditions, as well as with our proposed MAM module. As displayed in Table 1, freezing the LLM helps preserve language capabilities without any loss but is relatively weaker in multi-modal benchmark performance compared to fully fine-tuning the LLM during the instruction tuning stage. Additionally, we employ LoRA [9], an efficient instruction tuning technique, to balance the performance between freezing and fully fine-tuning. As observed in the third row of Table 1, multi-modal performance improves while text performance decreases, illustrating the phenomenon of modality interference. Conversely, incorporating MAM consistently enhances performance across both multi-modal and pure-text benchmarks. Specifically, we see performance gains of 0.3, 1.3, and 2.1 in VQAv2, MMBench, and MMLU respectively, which underscores the

benefits of our proposed method in modality collaboration and the reduction of modality interference.

Additionally, we investigate the impact of using MAM in different modules within the Attention method in Table 2. It can be observed that applying MAM to only one linear projection can actually harm the model’s performance. It suggests that solely influencing the attention map (applying MAM on Q or K), or just affecting the value (applying MAM on V), is insufficient. When MAM is applied on two linear projections, The results show that applying MAM on Q K achieves the poorest performance as it still only affects the attention map. Conversely, applying MAM on K V demonstrates the best performance, followed by applying MAM on Q V. It is because that the visual tokens are predominantly located at the beginning of the sequence, resulting in a smaller impact from applying MAM on Q. When MAM is applied on Q K V, it achieves comparable effects compared to K V. To trade off performance and cost, we consider only applying MAM on K V.

Module w/ MAM	VQAv2	MMBench	Q-Bench
Q	77.0	64.3	56.1
K	77.1	64.8	57.1
V	78.2	65.4	57.2
Q V	78.6	65.8	58.9
Q K	77.4	64.7	55.3
K V	79.2	66.1	59.5
Q K V	79.1	65.8	59.8

Table 2. The impact of MAM on various module combinations.

A.2. Hallucination Evaluation

We measure the hallucination of our model on image description using MMHal-Bench [29] and compare the results with other recent vision-language models, including Kosmos-2 [24], IDEFICS [12], InstructBLIP [5], LLaVA [19], and LLaVA-RLHF [29]. Following [29], we use GPT-

*Equal contribution

†Corresponding author

¹The checkpoint is recently released on October 27th, 2023.

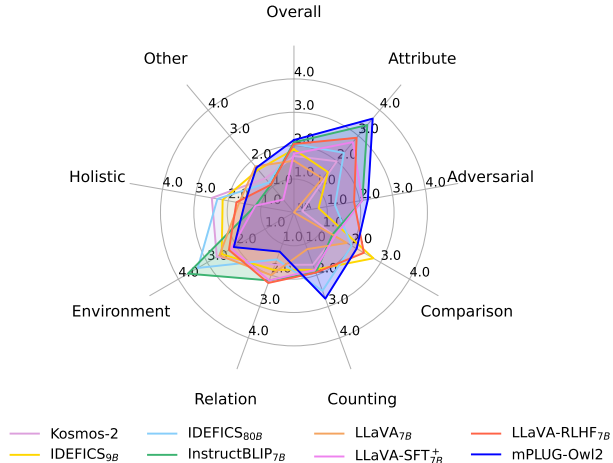


Figure 1. Detailed performance of various models across the eight categories in MMHal-Bench [29], where "Overall" represents the average performance across all categories.

4 to evaluate the overall score and hallucination rate of different MLLMs. As depicted in Figure 1, we find that our mPLUG-Owl2 tends to generate the response with reduced hallucination compared to other methods, especially surpassing IDEFICS [12] with 80 billion parameters, showing the superiority of our methods. Besides, we can notice that our model excels at attribute and counting because the visual abstractor can effectively identify the main parts of the image, which reduces the hallucination.

We also study the hallucination of recent popular MLLMs and present the results in Figure 2. In the first example, the query asks the models to recognize the pattern on the wall. However, the pattern is not clearly visible in the image, causing other models to mistakenly perceive it as a solid color. Our model, on the other hand, accurately notices the white pattern on the wall and correctly answers the question. In the second example, there are only a few trees in the image. However, InstructBLIP incorrectly considers that there are no trees in the image. LLaVA and LLaVA-1.5, on the other hand, hallucinate and consider the tree in the image to be dense. MiniGPT-4 gives the correct answer, but with minimal explanation. Our mPLUG-Owl2, however, answers the question correctly and provides a more detailed explanation.

A.3. POPE Evaluation

We also conduct the hallucination evaluation using POPE [15], the results are shown in Table 3. As we can observe in the table, we can find mPLUG-Owl2 achieves higher F1 scores on the popular and adversarial split, showing the robustness of our model in terms of object hallucination compared to other MLLMs.

A.4. Detailed Evaluation Results on MMBench

MMBench [20] is a meticulously designed benchmark that comprehensively assesses the diverse skills of vision-language models. The results from the test set for various MLLMs are presented in Table 4.

A.5. Detailed Evaluation Results on MM-Vet

We provide the detailed results of MM-Vet in Table 5. It can be observed that by training the visual encoder of mPLUG-Owl2, it exhibits stronger OCR capability compared to the model with the same backbone (i.e., LLaVA, Otter). Besides, mPLUG-Owl2 surpasses models with stronger language decoders such as LLaVA-13B which equips LLM with 13 billion parameters.

A.6. Detailed Evaluation Results on Q-Bench

For evaluating the low-level visual perception abilities, we have included the results of Q-Bench [31] on the test set. By training the visual encoder, the ability of mPLUG-Owl2 in terms of low-level perception has been improved significantly, as it outperforms the model with a stronger visual encoder (i.e., ViT-L (0.3B) v.s. ViT-G (1.9B)), showing the effectiveness of our training paradigm.

A.7. Detailed Evaluation Results on MMHal-Bench

We include Table 7 for the full evaluation results on MMHal-Bench [29].


B. Implementation

B.1. Data Mixture





In this section, we detail our final training data mixture used during the instruction tuning stage in Table 8. Specifically, we process the VQAv2 [8] data by selecting the answer with the highest confidence and combining question-answer pairs that share the same image. This combining strategy is also applied to GQA [10], OKVQA [21], and OCRVQA [22] datasets. Additionally, for multiple-choice questions in A-OKVQA [26], we augment the dataset by switching the order of options to enhance robustness in terms of multiple choices. For caption datasets like COCO [17] and TextCaps [28], we randomly select one caption from the ground truth for each image. Concurrently, some regional-VQA [11, 33] datasets are also used to improve regional abilities.

B.2. Training Hyper-parameters


We report the detailed training hyper-parameter settings of mPLUG-Owl2 in Table 9. Specifically, we leverage the model parallelism with Megatron [27] distributed training framework to ensure a larger resolution training while maintaining efficiency.

Input Image: 





Query Is the wall solid color in this image?

InstructBLIP 	Yes, the wall in the image is solid color.
LLaVA 	Yes, the wall in the image is solid color, which is tan or beige.
LLaVA-1.5 	Yes, the wall in the image is solid color, which is pink.
MiniGPT-4	The wall in this image appears to be solid color.
mPLUG-Owl2 	No, the wall in this image is not solid color. It has a pink and white pattern.

Example #1

Input Image: 

Query Is the tree dense in this image?

InstructBLIP 	No, there is no tree in the image.
LLaVA 	Yes, the tree in the image is dense, with a large number of branches and leaves.
LLaVA-1.5 	Yes, the tree in this image is dense.
MiniGPT-4	The tree is not dense in this image.
mPLUG-Owl2 	No, the tree in the image is not dense. It appears to be a small tree in the background.

Example #2

Figure 2. Example cases compared with InstructBLIP [5], LLaVA [19], LLaVA-1.5 [18], MiniGPT-4 [35] and our mPLUG-Owl2.

Table 3. **Object hallucination benchmark using POPE evaluation pipeline** . "Yes" signifies the likelihood of the model producing a positive response.

Datasets	Metrics	mPLUG-Owl2	Shikra [4]	InstructBLIP [5]	MiniGPT-4 [35]	LLaVA [19]	MM-GPT [7]	mPLUG-Owl [32]
Random	Accuracy (↑)	88.28	86.90	88.57	79.67	50.37	50.10	53.97
	Precision (↑)	94.34	94.40	84.09	78.24	50.19	50.05	52.07
	Recall (↑)	82.20	79.27	95.13	82.20	99.13	100.00	99.60
	F1-Score (↑)	87.85	86.19	89.27	80.17	66.64	66.71	68.39
	Yes (→ 50%)	44.91	43.26	56.57	52.53	98.77	99.90	95.63
Popular	Accuracy (↑)	86.20	83.97	82.77	69.73	49.87	50.00	50.90
	Precision (↑)	89.46	87.55	76.27	65.86	49.93	50.00	50.46
	Recall (↑)	82.06	79.20	95.13	81.93	99.27	100.00	99.40
	F1-Score (↑)	85.60	83.16	84.66	73.02	66.44	66.67	66.94
	Yes (→ 50%)	45.86	45.23	62.37	62.20	99.40	100.00	98.57
Adversarial	Accuracy (↑)	84.12	83.10	72.10	65.17	49.70	50.00	50.67
	Precision (↑)	85.54	85.60	65.13	61.19	49.85	50.00	50.34
	Recall (↑)	82.13	79.60	95.13	82.93	99.07	100.00	99.33
	F1-Score (↑)	83.80	82.49	77.32	70.42	66.32	66.67	66.82
	Yes (→ 50%)	48.00	46.50	73.03	67.77	99.37	100.00	98.67

Method	Language Model	Vision Model	Overall	LR	AR	RR	FP-S	FP-C	CP
MMGPT [7]	LLaMA-7B	CLIP ViT-L/14	16.0	1.1	23.8	20.7	18.3	5.2	18.3
MiniGPT-4 [35]	Vicuna-7B	EVA-G	12.0	13.6	32.9	8.9	28.8	11.2	28.3
InstructBLIP [5]	Vicuna-7B	EVA-G	33.9	21.6	47.4	22.5	33.0	24.4	41.1
LLaMA-Adapter-v2 [6]	LLaMA-7B	CLIP ViT-L/14	38.9	7.4	45.3	19.2	45.0	32.0	54.0
LLaVA [29]	Vicuna-7B	CLIP ViT-L/14	36.2	15.9	53.6	28.6	41.8	20.0	40.4
G2PT [20]	Vicuna-7B	ViT-G	39.8	14.8	46.7	31.5	41.8	34.4	49.8
Otter-I [13]	LLaMA-7B	CLIP ViT-L/14	48.3	22.2	63.3	39.4	46.8	36.4	60.6
mPLUG-Owl [†] [32]	LLaMA-7B	CLIP ViT-L/14	62.3	37.5	75.4	56.8	67.3	52.4	67.2
Shikra [4]	Vicuna-7B	CLIP ViT-L/14	60.2	33.5	69.6	53.1	61.8	50.4	71.7
mPLUG-Owl2	LLaMA2-7B	CLIP ViT-L/14	65.4	29.2	69.7	61.7	67.0	60.0	79.5

Table 4. CircularEval multi-choice accuracy results on MMBench [20] dev set. We adopt the following abbreviations: LR for Logical Reasoning; AR for Attribute Reasoning; RR for Relation Reasoning; FP-C for Fine-grained Perception (Cross Instance); FP-S for Fine-grained Perception (Single Instance); CP for Coarse Perception. Baseline results are taken from [20]. [†] denotes the model is carefully optimized for MMBench.

C. Summary of the Evaluation Benchmarks

We provide a detailed summary of the used evaluation benchmarks and corresponding metrics in Table 10.

D. Broader Impact

mPLUG-Owl2 employs off-the-shelf LLM and web-sourced data. Consequently, it inherits some of the weaknesses of the original LLM and web-crawled data, such as generating uncensored text or producing biased outputs. We address these shortcomings by enhancing the model’s grounding on the visual and instructional input and executing joint vision-language instruction tuning on a diverse range of high-quality datasets. However, we advise against deploying mPLUG-Owl2 models for any downstream applications without prior evaluation of safety and fairness specific to the respective application.

References

- [1] Sharegpt. <http://sharegpt.com>, 2023. 6
- [2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 5
- [3] 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. 5
- [4] 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. 4, 5
- [5] 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

Model	Rec	OCR	Know	Gen	Spat	Math	Total
Transformers Agent (GPT-4) [23]	18.2	3.9	2.2	3.2	12.4	4.0	13.4±0.5
MiniGPT-4-7B [35]	27.4	15.0	12.8	13.9	20.3	7.7	22.1±0.1
BLIP-2-12B [14]	27.5	11.1	11.8	7.0	16.2	5.8	22.4±0.2
LLaVA-7B [19]	28.0	17.1	16.3	18.9	21.2	11.5	23.8±0.6
MiniGPT-4-13B [35]	29.9	16.1	20.4	22.1	22.2	3.8	24.4±0.4
Otter-9B [13]	27.3	17.8	14.2	13.8	24.4	3.8	24.7±0.3
OpenFlamingo-9B [2]	28.7	16.7	16.4	13.1	21.0	7.7	24.8±0.2
InstructBLIP-13B [5]	30.8	16.0	9.8	9.0	21.1	10.5	25.6±0.3
InstructBLIP-7B [5]	32.4	14.6	16.5	18.2	18.6	7.7	26.2±0.2
LLaVA-7B (LLaMA-2) [19]	32.9	20.1	19.0	20.1	25.7	5.2	28.1±0.4
LLaMA-Adapter v2-7B [6]	38.5	20.3	31.4	33.4	22.9	3.8	31.4±0.1
LLaVA-13B (V1.3) [19]	38.1	22.3	25.2	25.8	31.3	11.2	32.5±0.1
LLaVA-13B (LLaMA-2) [19]	39.2	22.7	26.5	29.3	29.6	7.7	32.9±0.1
mPLUG-Owl2	41.3	27.4	27.5	27.9	30.3	7.7	36.2±0.1

Table 5. Evaluation results on various MLLMs regarding each core VL capability on MM-Vet [34]. Rec stands for recognition; Know indicates knowledge; Gen is generation; Spat means spatial. All the numbers are presented in % and the full score is 100%.

Method	Yes-or-no	What	How	Distortion	Others	In-context Distortion	In-context Others	Overall
IDEFICS [12]	0.6004	0.4642	0.4671	0.4038	0.5990	0.4726	0.6477	0.5151
InstructBLIP [5]	0.7099	0.5141	0.4300	0.4500	0.6301	0.5719	0.6439	0.5585
Kosmos-2 [24]	0.6058	0.3124	0.3539	0.3865	0.4654	0.4349	0.4735	0.4334
LLaMA-Adapter-v2 [6]	0.6661	0.5466	0.5165	0.5615	0.6181	0.5925	0.5455	0.5806
LLaVA-1.5 [18]	0.6460	0.5922	0.5576	0.4798	0.6730	0.5890	0.7376	0.6007
LLaVA [19]	0.5712	0.5488	0.5185	0.4558	0.5800	0.5719	0.6477	0.5472
MiniGPT-4 [35]	0.6077	0.5033	0.4300	0.4558	0.5251	0.5342	0.6098	0.5177
mPLUG-Owl [32]	0.7245	0.5488	0.4753	0.4962	0.6301	0.6267	0.6667	0.5893
Otter [13]	0.5766	0.3970	0.4259	0.4212	0.4893	0.4760	0.5417	0.4722
Qwen-VL [3]	0.6533	0.6074	0.5844	0.5413	0.6635	0.5822	0.7300	0.6167
Shikra [4]	0.6909	0.4793	0.4671	0.4731	0.6086	0.5308	0.6477	0.5532
mPLUG-Owl2	0.7318	0.5531	0.5864	0.5374	0.7136	0.5788	0.7338	0.6294

Table 6. Detailed evaluation results for different MLLMs on the test set of Q-Bench [31].

- general-purpose vision-language models with instruction tuning. [ArXiv](#), abs/2305.06500, 2023. **1, 3, 4, 5, 6**
- [6] 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. **4, 5**
- [7] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. [arXiv preprint arXiv:2305.04790](#), 2023. **4**
- [8] 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. **2, 6**
- [9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. [arXiv preprint arXiv:2106.09685](#), 2021. **1**
- [10] 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. **2, 6**
- [11] 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. **2, 6**
- [12] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. **1, 2, 5, 6**
- [13] 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. **4, 5**

Method	Overall Score \uparrow	Hallucination Rate \downarrow	Score in Each Question Type \uparrow							
			Attribute	Adversarial	Comparison	Counting	Relation	Environment	Holistic	Other
Kosmos-2 [24]	1.69	0.68	2.00	0.25	1.42	1.67	1.67	2.67	2.50	1.33
IDEFICS _{9B} [12]	1.89	0.64	1.58	0.75	2.75	1.83	1.83	2.50	2.17	1.67
IDEFICS _{80B} [12]	2.05	0.61	2.33	1.25	2.00	2.50	1.50	3.33	2.33	1.17
InstructBLIP _{7B} [5]	2.10	0.58	3.42	2.08	1.33	1.92	2.17	3.67	1.17	1.08
InstructBLIP _{13B} [5]	2.14	0.58	2.75	1.75	1.25	2.08	2.50	4.08	1.50	1.17
LLaVA _{7B} [19]	1.55	0.76	1.33	0.00	1.83	1.17	2.00	2.58	1.67	1.83
LLaVA-RLHF _{7B} [29]	2.05	0.68	2.92	1.83	2.42	1.92	2.25	2.25	1.75	1.08
mPLUG-Owl2	2.17	0.56	3.67	2.25	2.17	2.75	1.25	2.08	1.50	1.75

Table 7. Detailed evaluation results for different MLMs on MMHal-Bench.

Data Type	Data Name	Size
Text	ShareGPT [1]	40K
	SlimOrca [16]	518K
Dialogue	LLaVA [19]	158K
Caption	COCO [17]	82K
	TextCaps [28]	22K
VQA	VQAv2 [8]	83K
	GQA [10]	72K
	OKVQA [21]	9K
	OCRvQA [22]	80K
	A-OKVQA [26]	50K
Regional-VQA	RefCOCO [33]	30K
	VisualGenome [11]	86K
Total		1.23M

Table 8. Instruction-following Data Mixture of mPLUG-Owl2.

- [14] 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. **5**
- [15] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. Evaluating object hallucination in large vision-language models. *ArXiv*, abs/2305.10355, 2023. **2**
- [16] Wing Lian, Guan Wang, Bleyds Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification, 2023. **6**
- [17] 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. **2, 6**
- [18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744, 2023. **1, 3, 5**
- [19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. **1, 3, 4, 5, 6**
- [20] 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

Configuration	Pre-training	Instruction Tuning
ViT init.	CLIP-L/14 [25]	Pre-train stage
LLM init.	LLaMA-2 [30]	LLaMA-2 [30]
Visual Abstractor init.	Random	Pre-train stage
Image resolution	224 × 224	448 × 448
ViT sequence length	256	1024
LLM sequence length	256	2048
Learnable query numbers	64	64
Optimizer	AdamW	
Optimizer hyperparameter	$\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1e^{-6}$	
Peak learning rate	$1e^{-4}$	$2e^{-5}$
Minimum learning rate	$1e^{-6}$	$1e^{-7}$
ViT learning rate decay		0
ViT Drop path rate		0
Learning rate schedule		Cosine
Weight decay	0.05	0
Gradient clip		1.0
Training steps	42,500	4,800
Warm-up steps	1,000	250
Global batch size	8,192	256
Gradient Acc.		16
Numerical precision		bfloat16
Optimizer sharding		✓
Activation checkpointing		✓
Model parallelism	1	2
Pipeline parallelism		1

Table 9. Training hyper-parameters of mPLUG-Owl2.

- all-around player? *arXiv preprint arXiv:2307.06281*, 2023. **2, 4**
- [21] 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, 6**
- [22] 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. **2, 6**
- [23] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. **5**

Task	Dataset	Description	Split	Metric
Image Caption	COCO Flickr30K	Captioning of natural images Captioning of natural images	karpathy-test karpathy-test	CIDEr (†) CIDEr (†)
General VQA	VQAv2 OKVQA GQA VizWizQA TextVQA SciQA-Img	VQA on natural images VQA on natural images requiring outside knowledge VQA on scene understanding and reasoning VQA on photos taken by people who are blind VQA on natural images containing text Multi-choice VQA on a diverse set of science topics	test-dev val test-balanced test-dev val test	VQA Score (†) VQA Score (†) EM (†) VQA Score (†) VQA Score (†) Accuracy (†)
VideoQA	MSRVTT-QA MSVD-QA TGIF-QA	Video Question Answering Video Question Answering GIF Question Answering	test test test	Accuracy (†) / Relevance Score (†) Accuracy (†) / Relevance Score (†) Accuracy (†) / Relevance Score (†)
Text Benchmark	MMLU BBH AGIEval ARC-c ARC-e	A benchmark designed to measure knowledge acquirement A suite of 23 challenging BIG-Bench tasks A human-centric benchmark specifically designed to evaluate the general abilities of foundation model A multiple-choice question-answering dataset, containing questions from science exams from grade 3 to grade 9. A multiple-choice question-answering dataset, containing questions from science exams from grade 3 to grade 9.	dev test test test test	Accuracy (†) Accuracy (†) Accuracy (†) Accuracy (†) Accuracy (†)
Instruction Following	MME MMBench MM-Vet SEED-Bench Q-Bench	Open-ended VL Benchmark by yes/no questions Open-ended VL Benchmark by Multi-choice VQA with Circular Evaluation Open-ended VL Benchmark with Various Abilities Open-ended VL Benchmark by Multi-choice VQA Open-ended Low-level Vision Benchmark by Multi-choice VQA	Perception test test Image & Video test	Accuracy (†) Accuracy (†) GPT-4 Score (†) Accuracy (†) Accuracy (†)
Hallucination	POPE MMHal-Bench	Object existence by yes/no questions Open-ended hallucination benchmarks	random/popular/adversarial test	Accuracy / Precision / Recall / F1 (†) GPT-4 Score (†)

Table 10. Summary of the evaluation benchmarks of mPLUG-Owl2. EM stands for exacting matching.

- [24] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.14824, 2023. 1, 5, 6
- [25] 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. 6
- [26] 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. 2, 6
- [27] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019. 2
- [28] 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. 2, 6
- [29] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. *ArXiv*, abs/2309.14525, 2023. 1, 2, 4, 6
- [30] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esioibu, 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. 6
- [31] 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. 2, 5
- [32] 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. 4, 5
- [33] 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. 2, 6
- [34] 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
- [35] 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. 3, 4, 5