Awaker2.5-VL: Stably Scaling MLLMs with Parameter-Efficient Mixture of Experts

Jinqiang Long¹*, Yanqi Dai¹*, Guoxing Yang¹, Hongpeng Lin¹, Nanyi Fei¹, Yizhao Gao¹, and Zhiwu Lu²

Metabrain AGI Lab, Shanghai, China
https://www.metabrainagi.com
Gaoling School of Artificial Intelligence, Renmin University of China

Abstract. As the research of Multimodal Large Language Models (MLLMs) becomes popular, an advancing MLLM model is typically required to handle various textual and visual tasks (e.g., VQA, Detection, OCR, and ChartQA) simultaneously for real-world applications. However, due to the significant differences in representation and distribution among data from various tasks, simply mixing data of all tasks together leads to the well-known "multi-task conflict" issue, resulting in performance degradation across various tasks. To address this issue, we propose Awaker2.5-VL, a Mixture of Experts (MoE) architecture suitable for MLLM, which acquires the multi-task capabilities through multiple sparsely activated experts. To speed up the training and inference of Awaker2.5-VL, each expert in our model is devised as a low-rank adaptation (LoRA) structure. Extensive experiments on multiple latest benchmarks demonstrate the effectiveness of Awaker 2.5-VL. The code and model weight are released in our Project Page: https://github.com/MetabrainAGI/Awaker.

Keywords: Multimodal Large Language Model \cdot Multi-task Conflict \cdot Mixture of Experts

1 Introduction

With the rapid development of Large Language Model (LLM) [1, 2, 13], Multimodal Large Language Model (MLLM) [5, 9–11, 16] have also become a new research hotspot in recent years. Series of MLLMs such as BLIP2 [9], MiniGPT-4 [16], and LLaVA [11] have demonstrated impressive performance in various vision-text tasks (e.g., image captioning, and visual question answering). Qwen-VL-Chat [3] transfers traditional vision tasks (e.g., object detection, and OCR) to vision-text tasks, endowing the model with the ability to perform more tasks by defining specific instructions and output formats.

However, the above mentioned visual-center tasks have significant differences in image input, instructions, and output formats. For example, image captioning

^{*} They have made equal contribution and done this work during internship.

requires the model to perceive the entire image and generate a natural language description of its content. In contrast, object detection requires the model to locate specific objects in the image and output their exact positions in numerical coordinates. The common training strategy is to mix the training data from multiple tasks and feed it uniformly into the model for training. Since the current model architecture does not specifically address the differences among multiple tasks, this simple mixing strategy often leads to the well-known "multi-task conflict" issue, which further results in reduced performance across all tasks.

To address the "multi-task conflict" issue, we propose Awaker2.5-VL, a stable Mixture of Experts (MoE) architecture suitable for large multimodal models. Specifically, we set up multiple expert models to acquire the task-specific capabilities across various tasks, with a gating network automatically controlling the activation and deactivation of experts. In the meantime, we include a global expert that remains always active to ensure the versatility and generalization of the model. To speed up the training and inference of Awaker2.5-VL, each expert in our model is devised as a low-rank adaptation (LoRA) structure. Additionally, we uniformly design the routing strategy for MoE in our model. It is worth noting that during the training of Awaker2.5-VL, the base model is frozen, and only the MoE/Lora modules are trained, resulting in an extremely low training cost. Finally, we implement Awaker2.5-VL with Qwen2-VL-7B-Instruct [14] as the base model, and achieve state-of-the-art results on several recent benchmarks, demonstrating the effectiveness of our Awaker2.5-VL.

The main contributions of this work are summarized as follows:

- (1) We design a stable Mixture of Experts (MoE) architecture, called Awaker 2.5-VL, suitable for large multimodal models.
- (2) We conduct an extensive exploration of MoE routing strategies and design a simple yet effective routing strategy for our proposed Awaker2.5-VL.
- (3) We achieve state-of-the-art results on several recent benchmarks with our proposed Awaker 2.5-VL.

2 Related work

Multimodal Large Language Model. In recent years, with the rapid development of large language models, large multimodal models have also become a new research hotspot, leading to many innovative research outcomes. BLIP-2 [9] introduces the Q-Former module between the vision encoder and the LLM, which enhances the interaction between visual and language features through a query mechanism. LLaVA [11] first generates multimodal image-text instruction data using GPT-4 and then performs instruction fine-tuning to train a large multimodal model for general-purpose visual and language understanding. The Qwen-VL [3] series of models uniformly transform traditional visual tasks (such as object detection and OCR) into vision-language tasks, enabling them to perform a variety of different tasks including image captioning, visual question answering, object detection, and OCR.

Mixture of Experts. Scaling Law [8] indicates that the number of parameters in a model is typically directly related to the model's complexity and expressive power. However, simply increasing the number of parameters often leads to higher resource consumption. To reduce the model training cost, Mixture of Experts (MoE) [4] models have been introduced into large language models. MoE is a sparse, gate-controlled deep learning model primarily composed of a set of expert models and a gating model. The fundamental idea of MoE is to partition the input data into multiple regions based on task types and assign each region to one or more expert models. Each expert model can focus on processing its specific portion of the input data, thereby improving the overall performance of the model. Mistral-8x7B [7] released by Mistral AI is an MoE model composed of 7 billion-parameter sub-models. It outperforms the 70 billion-parameter Llama-2 on multiple benchmarks. MOE-LLaVA [11] proposes a sparse, multimodal large model based on MoE, which uses only about 3 billion sparsely activated parameters. Despite this, MoE-LLaVA performs comparably to LLaVA-1.5-7B on various visual understanding datasets and even surpasses LLaVA-1.5-13B on object hallucination benchmarks.

3 Methodology

3.1 MoE Structure

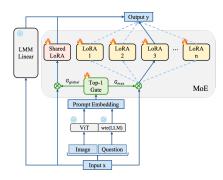
The MoE architecture of our Awaker2.5-VL (Figure 1), following MoCLE [6], consists of a base model parameterized by W_0 with frozen parameters, n experts and a gate layer, which can acquire the model's ability to handle various tasks. For Large Language Models, the Mixture of Experts (MoE) is typically implemented in the Feed-Forward Network layer, where each expert is an FFN layer. Unlike LLMs, each expert in Awaker2.5-VL is divesed as a LoRA structure parameterized with W_E^m ($m \in 1, 2, ..., n$). Additionally, to maintain the model's generalization capability, Awaker2.5-VL includes an always activated expert parameterized with W_E^G , meaning every piece of data passes through this global expert. The gate network is a simple linear layer parameterized with $W_G \in \mathbb{R}^n$ that controls which experts are activated when data is forwarded through the model and assigns weights to the outputs of these experts. Given an input x, the gating vector can be represented as follow:

$$G_{\text{experts}} = \text{top}_k(\text{softmax}(\frac{1}{\tau}(W_G x) + \epsilon))$$
 (1)

We set the number of activated experts to 1, and define G_{max} as the maximum value in G_{experts} , which is used as the weight of the activated expert. The weight of the global expert can be defined as follow:

$$G_{\text{global}} = 1 - G_{\text{max}} \tag{2}$$

4 Long et al.



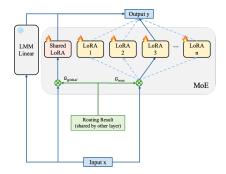


Fig. 1. The Standard MoE Structure in Awaker 2.5-VL.

Fig. 2. The Simplified MoE Structure in Awaker2.5-VL.

The final model output is determined by the combined outputs of the pre-trained model, the global expert, and the mixture of experts:

$$y = \sum_{m=1}^{n} G_{\text{experts}} W_E^m x + G_{\text{global}} W_E^G x + W_0 x \tag{3}$$

We also design a simplified version of the MoE architecture (as shown in Figure 2), where the gate Layer is removed. Instead, it accepts the gate results (G_{global} and G_{experts}) computed in the MoE from another layer for routing. We will intersperse the use of these two types of MoE architectures throughout the model, as detailed in Chapter 4.1.

3.2 Stable Routing Strategy

In this work, our MoE design in Awaker2.5-VL has two main differences from traditional MoE structures in LLMs: (1) Our MoE operates at the instance-level rather than the token-level (used by traditional MoE), meaning that all tokens within each instance will activate the same expert. (2) In traditional MoE structures, the gate network of each MoE module receives the output from the previous transformer layer. However, in our practical implementation, we found that this routing strategy can lead to training instability. In this work, we thus simplify the routing strategy of MoE as shown in Figure 2.

Specifically, for a input instance, the gate layer at each transformer layer receives the output from the embedding layer of LLM as input, which keeps the same across all gate layers of Awaker 2.5-VL. Moreover, to reduce the gap between training and inference, we use only the embedding of the instruction part as input to the gate network during training. The label part of the training data does not participate in the routing decision. That is, for multi-modal data, this includes the embedding of both the images and the question. For pure-text data, it is just the embedding of the question text.

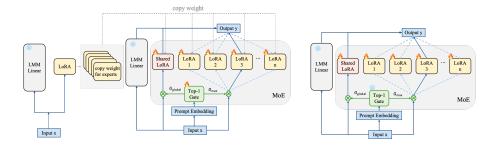


Fig. 3. The Traing Pipeline of Awaker 2.5-VL. From Left to Right: Stage I, Stage II, and Stage III.

3.3 Training Process

We train Awaker2.5-VL through three stages as shown in Figure 3.

Stage I: Initialization Training. In the first stage, we add a LoRA module to the base model for training. During this stage, we freeze the entire base model and only train the LoRA parameters.

Stage II: MoE Training. We further replace the LoRA module from the first stage with an MoE module in Awaker2.5-VL. Each expert in the MoE module is initialized with the LoRA parameters trained in the first stage. During this stage, we freeze the base model and only train the MoE module (including the gate layer and all the experts).

Stage III: Instruction Fine-Tuning. In the third stage, we freeze the gate layer of the MoE module and only train the experts.

4 Experiments

4.1 Implementation Details

Model Details. Our Awaker2.5-VL is based on the Qwen2-VL-7B-Instruct model, which is a multimodal large language model built on Qwen2. We integrate our MoE architecture into the Attention and MLP modules of Qwen2. However, in this work, we utilize different types of adapters (LoRA or MoE) based on the different layers in the model. Concretely, in the Attention module, we receptively inject a single LoRA to the q_proj, k_proj, and v_proj layers, while adding an MoE module to the o_proj layer. In the MLP module, we receptively add MoE modules to the gate_proj and Simplified MoE modules to up_proj and down_proj layers. This means that the gate results from the gate_proj layer are shared with the up_proj and down_proj layers.

Hyperparameters. We set n=4 experts and 1 general expert in our MoE architecture, where each expert takes the hyperparameters $r=256, \alpha=512$. During the training process, due to cost constraints, we set the maximum image resolution to 1,103,872 (i.e., max_pixels=1,103,872 in Qwen2-VL-7B-Instruct)

Table 1. Evaluation Results on MME-Realworld-CN Benchmark.

Model	Parameters	Institutions	Chinese		
			Overall	Perception	Reasoning
Awaker2.5-VL (ours)	10.8B	Metabrain AGI	62.7	67.71	52.07
Qwen2-VL	8B	Alibaba	55.5	59.80	46.46
InternVL-2	7B	Shanghai AI Lab	54.3	57.79	46.65
InternVL-Chat-V1.5	20B	Shanghai AI Lab	47.9	49.90	43.74
Claude3.5 Sonnet	-	Anthropic	47.0	48.25	44.31
Yi-VL-34B	34B	01.AI	42.0	42.45	41.16
CogVLM2-Llama3-Chat	8B	THU & Zhipu AI	39.8	38.57	42.25
GPT-40	-	OpenAI	38.8	43.44	29.05
Mini-Gemini-34B-HD	34B	CUHK	38.5	38.31	38.75
Cambrian-1-8B	8B	NYU	33.6	32.44	35.97
LLaVA-NeXT-Qwen-72B	72B	Bytedance	30.6	30.02	31.67
Gemini-1.5-Pro	-	Google	28.1	36.10	11.14
DeepSeek-VL	7B	DeepSeek AI	27.6	27.63	27.63
GPT-4o-mini	-	OpenAI	25.9	26.32	25.16

Table 2. Evaluation Results on MME-Realworld Benchmark.

Model	Parameters	Institutions	English		
			Overall	Perception	Reasoning
Awaker2.5-VL (ours)	10.8B	Metabrain AGI	60.8	63.14	43.74
LLaVA-OneVision	8B	Bytedance	57.4	59.59	41.17
Qwen2-VL	8B	Alibaba	56.5	58.96	40.39
InternVL-2	7B	Shanghai AI Lab	53.5	55.82	38.74
Claude3.5 Sonnet	-	Anthropic	51.6	52.90	44.12
InternVL-Chat-V1.5	20B	Shanghai AI Lab	49.4	51.36	36.48
Mini-Gemini-34B-HD	34B	CUHK	45.9	48.05	31.73
GPT-4o	-	OpenAI	45.2	46.43	37.61
CogVLM2-Llama3-Chat	8B	THU & Zhipu AI	44.6	45.84	37.25
Cambrian-1-8B	8B	NYU	42.7	43.82	36.16
Gemini-1.5-Pro	-	Google	38.2	39.63	29.19
GPT-4o-mini	-	OpenAI	36.4	37.12	32.48
DeepSeek-VL	7B	DeepSeek AI	32.4	33.14	27.98
Yi-VL-34B	34B	01.AI	31.0	30.97	32.45
LLaVA-NeXT-Qwen-72B	72B	Bytedance	28.7	29.01	27.86

and batch_size=4. We adopt the cosine lr_scheduler, and set the learning rating lr=1e-5 in Stage I, lr=1e-5 Stage II, and lr=5e-6 in Stage III.

4.2 Training Data

For model training, we construct a dataset of approximately 12 million pieces of data, which includes 7 million pieces of English data and 5 million pieces of Chinese data. The English data mainly comes from open-source datasets and has been filtered and selected. This primarily includes Cambrain (2M), LLaVA-OneVision (4M), Infinity-MM (800K), MathV360k (360K). The Chinese data mainly comes from our self-built dataset. In the past two years, we have carefully built a Chinese instruction dataset of about 5 million pieces of data, which includes abundant instruction data from various multimodal tasks such as Image Caption, VQA, Object Detection, and OCR.

Chinese Model Parameters Institutions Overall MMBench_v1.1 MMBench Qwen2-VL-72B 73.4BAlibaba 86.3 85.8 86.7 InternVL2-40B 40B Shanghai AI Lab 85.784.986.4InternVL2-Llama-76B 76B Shanghai AI Lab 85.5 85.5 Taiyi Megvii 85.2 85.085.4 JT-VL-Chat-V3.0 China Mobile 84.7 83.5 85.8 73B LLaVA-OneVision-72B ByteDance 84.6 83.9 85.3 Step-1.5V StepFun 84.0 83.5 84.5 Claude3.5-Sonnet-20241022 Anthropic 83.0 82.5 83.5 Awaker2.5-VL (ours) 10.8BMetabrain AGI 82.681.883.4GPT-4o (0513, detail-low) 82.3 82.5 82.1 OpenAI LLaVA-OneVision-7B 8BByteDance 81.8 80.9 82.7 GPT-4o (0513, detail-high) OpenAI 81.5 82.1 81.8 InternVL2-26B 26B Shanghai AI Lab 81.5 80.9 82.1 CongROng CloudWalk 81.2 80.4 81.9 MMAlava2 26B DataCanvas 80.9 79.7 82.1 Ovis1.6-Gemma2-9B 10.2BAlibaba 80.879.582.0 Qwen2-VL-7B 8BAlibaba 80.5 80.3 80.6 LLaVA-OneVision-72B (SI) 73B ByteDance 80.0 81.9 78.0 InternVL-Chat-V1.5 26B Shanghai AI Lab 79.1 80.7 79.9 InternLM-XComposer2.5 8BShanghai AI Lab 79.9 78.8 80.9 GPT-4o (0806, detail-high) OpenAI 79.8 79.280.3 GPT-4V (0409, detail-high) OpenAI 79.2 78.2 80.2

Table 3. Evaluation Results on MMBench-CN Benchmark.

4.3 Main Results

We conduct evaluation on the latest two multimodal large model benchmarks: (1) MME-RealWorld [15]: this benchmark considers images from domains such as autonomous driving, remote sensing, video surveillance, newspapers, street views, and financial charts. It contains 29,429 annotations, covering 43 sub-tasks, with each task having at least 100 questions. (2) MMBench [12]: this benchmark is a visual-language model evaluation benchmark developed by the OpenCompass research team. It enables a granular assessment of capabilities ranging from perception to cognition, covering 20 fine-grained evaluation dimensions including object detection, text recognition, action recognition, image understanding, and relational reasoning.

Results on MME-Realworld. We make a comprehensive performance evaluation of our Awaker2.5-VL on the MME-Realworld benchmark and its Chinese version (MME-Realworld-CN). Table 1 and Table 2 show the evaluation results in terms of perception, reasoning, and overall scores on both MME-Realworld and MME-Realworld-CN, respectively. All compared models are ranked by the average/overall scores. The results of the competitors are directly cited from https://mme-realworld.github.io/home_page.html#leaderboard. From the two tables, we have the following observations:

1) Awaker 2.5-VL ranks the first in overall score, perception score, and reasoning score on MME-RealWorld-CN, outperforming all other models. It is even the only one that takes the overall score over 60 on MME-Realworld-CN.

GPT-4V (0409, detail-high)

English Model Parameters Institutions Overall MMBench_v1.1 MMBench Qwen2-VL-72B 73.4BAlibaba 86.5 86.1 InternVL2-40B40BShanghai AI Lab 86.085.186.885.7 84.7 86.7 Taivi Megvii InternVL2-Llama-76B 76B Shanghai AI Lab 85.5 85.5LLaVA-OneVision-72B 73B ByteDance 85.485.0 85.8 JT-VL-Chat-V3.0 China Mobile 84.5 83.685.4 Awaker2.5-VL (ours) 10.8BMetabrain AGI 83.7 82.5 84.9 GPT-4o (0513, detail-high) OpenAI 83.2 83.0 83.4 GPT-4o (0513, detail-low) OpenAI 83.283.183.3Step-1.5V StepFun 82.9 80.4 85.3 InternVL2-26B 26B Shanghai AI Lab 82.5 81.5 83.4 10.2BOvis1.6-Gemma2-9B Alibaba 82.5 81.5 83.4 RBDash-v1.2-72B 79BDLUT 82.5 81.7 83.2 Qwen2-VL-7B 8BAlibaba 82.4 81.8 83.0 LLaVA-OneVision-7B 8BByteDance 82.1 80.9 83.2 GPT-4o (0806, detail-high) OpenAI 82.0 81.8 82.1LLaVA-OneVision-72B (SI) 73B ByteDance 81.9 83.3 80.5 Qwen-VL-Plus-0809 Alibaba 81.9 81.1 82.7 CongROng CloudWalk 80.9 82.8 81.9 Claude 3.5 - Sonnet - 2024 1022Anthropic 81.8 80.9 82.6 26B MMAlava2 DataCanvas 81.6 80.6 82.5 InternVL-Chat-V1.5 26B Shanghai AI Lab 81.3 80.3 82.3 InternLM-XComposer2.5 8BShanghai AI Lab 81.180.182.0

Table 4. Evaluation Results on MMBench Benchmark.

2) Awaker 2.5-VL still holds the top-1 position in the perception and overall scores on MME-Realworld, even though there is a slight decrease in the reasoning score when compared to the state-of-the-art.

OpenAI

80.0

80.5

81.0

3) Awaker2.5-VL demonstrates exceptional performance in Chinese scenarios (see Table 1), with an overall score improvement of 5 points compared with the base model Qwen2-VL-7B-Instruct, a 6-point increase in perception tasks, and a 3-point increase in reasoning tasks.

Results on MMBench. We compare Awaker 2.5-VL with the latest competitors on four MMBench series benchmarks: MMBench, MMBench_v1.1, MMBench_CN, and MMBench_v1.1_CN. We separately present the results on the Chinese benchmarks (MMBench_CN and MMBench_1.1_CN) and the English benchmarks (MMBench and MMBench_v1.1) in Table 3 and Table 4, respectively. All compared models are ranked by the average/overall scores. The results of the competitors are directly cited from https://mmbench.opencompass.org.cn/leaderboard. From the two tables, it can be observed that:

- 1) Awaker2.5-VL ranks 7th on MMBench and 9th on MMBench-CN. The performance of Awaker2.5-VL on both benchmarks exceeds that of other models with similar parameter sizes.
- 2) Compared to the base model Qwen2-VL-7B-Instruct, Awaker2.5-VL shows significant improvements on all four benchmarks.

5 Conclusion and Future Work

We release Awaker2.5-VL, a large multimodal Mixture of Experts (MoE) model. Our Awaker2.5-VL mitigates the "multi-task conflict" issue in MLLM through the MoE architecture and has outperformed the latest competitors on multiple benchmarks. Furthermore, we hope to enhance the capabilities of Awaker2.5-VL in the following areas in our ongoing research:

- (1) The current prompt embeddings used for routing are derived from the embedding layers of ViT and LLM, respectively. We believe that these shallow embeddings have limited capability of representation, especially for text prompts. Therefore, in our future work, we will explore more suitable methods for representing prompts to improve routing performance.
- (2) The MoE model in Awaker 2.5-VL is currently applied only to the LLM side of the multimodal model. We plan to conduct further research on applying the MoE model to the ViT as well.

References

- AI, ., :, Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., Huang, W., Hu, X., Ren, X., Niu, X., Nie, P., Xu, Y., Liu, Y., Wang, Y., Cai, Y., Gu, Z., Liu, Z., Dai, Z.: Yi: Open foundation models by 01.ai (2024), https://arxiv.org/abs/2403.04652
- Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
- 3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023), https://arxiv.org/abs/2308.12966
- 4. Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., Huang, J.: A survey on mixture of experts (2024). https://arxiv.org/abs/2407.06204
- Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks (2024), https://arxiv.org/abs/2312.14238
- 6. Gou, Y., Liu, Z., Chen, K., Hong, L., Xu, H., Li, A., Yeung, D.Y., Kwok, J.T., Zhang, Y.: Mixture of cluster-conditional lora experts for vision-language instruction tuning (2024), https://arxiv.org/abs/2312.12379
- Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T.L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mixtral of experts (2024), https://arxiv.org/abs/2401.04088
- 8. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models (2020), https://arxiv.org/abs/2001.08361
- Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models (2023), https://arxiv.org/abs/2301.12597

- 10. Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., Liu, S., Jia, J.: Minigemini: Mining the potential of multi-modality vision language models (2024), https://arxiv.org/abs/2403.18814
- 11. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), https://arxiv.org/abs/2304.08485
- 12. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? (2024), https://arxiv.org/abs/2307.06281
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
- 14. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
- Zhang, Y.F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., Wang, L., Jin, R., Tan, T.: Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? (2024), https://arxiv.org/abs/2408.13257
- Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models (2023), https://arxiv.org/abs/2304.10592