CuMo: Scaling Multimodal LLM with Co-Upcycled Mixture-of-Experts

Jiachen Li[1*] **Xinyao Wang**[2†‡] **Sijie Zhu**[2] **Chia-Wen Kuo**[2] **Lu Xu**[2]
Fan Chen[2] **Jitesh Jain**[1] **Humphrey Shi**[1‡] **Longyin Wen**[2]
[1]SHI Labs @ Georgia Tech & UIUC [2]ByteDance Inc., San Jose
https://github.com/SHI-Labs/CuMo

Abstract

Recent advancements in Multimodal Large Language Models (LLMs) have focused primarily on scaling by increasing text-image pair data and enhancing LLMs to improve performance on multimodal tasks. However, these scaling approaches are computationally expensive and overlook the significance of efficiently improving model capabilities from the vision side. Inspired by the successful applications of Mixture-of-Experts (MoE) in LLMs, which improves model scalability during training while keeping inference costs similar to those of smaller models, we propose **CuMo**, which incorporates **C**o-**u**pcycled Top-K sparsely-gated **M**ixture-**o**f-experts blocks into both the vision encoder and the MLP connector, thereby enhancing the multimodal LLMs with neglectable additional activated parameters during inference. CuMo first pre-trains the MLP blocks and then initializes each expert in the MoE block from the pre-trained MLP block during the visual instruction tuning stage, with auxiliary losses to ensure a balanced loading of experts. CuMo outperforms state-of-the-art multimodal LLMs across various VQA and visual-instruction-following benchmarks within each model size group, all while training exclusively on open-sourced datasets.

1 Introduction

The advent of GPT-4V [53] has sparked excitement within open-source communities to transform large language models (LLM) into multimodal LLMs. Recent multimodal LLMs [11, 46, 2] typically integrate pre-trained vision encoders with MLP connectors to LLMs, with visual instruction tuning data to fine-tune the pre-trained LLMs, enhancing their visual understanding capabilities. To further scale up multimodal LLMs, previous efforts [44, 45, 39, 51, 7, 42] primarily focus on training the model with a more extensive collection of text-image paired data and employing stronger LLMs, significantly increasing training efforts. On the vision side, recent work concentrates on leveraging multiple vision encoders [43, 18] to enrich visual content, employing larger vision encoders [9], and using advanced vision-language connectors [5] to improve performance on multimodal tasks. However, these techniques result in an increased number of additional parameters and generate extra visual tokens for LLMs to process, making it inefficient to scale.

In terms of efficiently scaling up models, Mixture-of-Experts (MoE) has become the de-facto framework in modern large-scale neural networks, particularly in natural language processing (NLP). Most large language models (LLM) are built upon the transformer [64] architecture, wherein sparse

[*] Work done during an internship at ByteDance San Jose, CA.
[†] Work done at ByteDance.
[‡] Corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

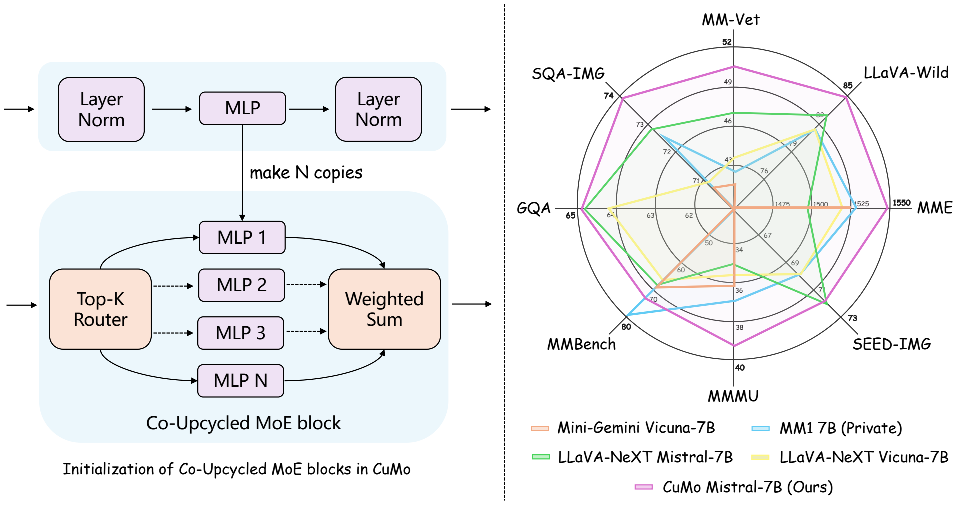


Figure 1: Left: Each MLP expert within the MoE block during the visual instruction tuning stage is initialized from the corresponding pre-trained MLP in CuMo. Right: CuMo outperforms strong open-sourced models such as Mini-Gemini and LLaVA-NeXT, as well as the private MM1 model.

MoE is used to replace the dense MLP block with the Top-K sparsely-gated MoE block [57]. Recent state-of-the-art open-sourced [28, 61] and private [55, 51] LLMs have predominantly adopted the sparse MoE architecture. These models are scaled up using the MoE design during training while maintaining relatively lower inference costs as only selected MLP experts are activated during the feed-forward process. Nevertheless, the development and optimization of MoE-based models have been largely tailored to LLMs, and the exploration of scaling multimodal LLMs with MoE, especially on the vision side, remains largely unexplored.

Motivated by these observations, we introduce CuMo, which integrates Top-K sparsely-gated MoE blocks into the vision encoder and the MLP connector of multimodal LLMs. We also explore the associated training recipe and methodology for CuMo. Firstly, we pre-train the MLP connector and perform pre-finetuning to warm up the whole model without introducing the MoE architecture, which stabilizes the following visual instruction tuning stage with newly incorporated sparse MoE blocks. Then, we replace each MLP block with the sparse MoE block in the MLP connector and the vision encoder through co-upcycling. Each expert within the sparse MoE block is initialized from the corresponding MLP block after the pre-training and the pre-finetuning stages, as shown in Figure 1. Additionally, each MoE block contains a Top-K router trained from scratch to select experts during the visual instruction tuning stage with auxiliary losses on the router to maintain a balanced loading of experts. We conduct further comparisons between co-upcycled LLMs and pre-trained MoE-based LLMs. The results show that the pre-trained MoE-based LLMs significantly outperform the co-upcycled LLMs. As a result, the co-upcycling of LLMs is not included in CuMo. Our models are trained fully on open-sourced datasets that are converted to visual instruction following formats. Experimental results demonstrate that CuMo outperforms other state-of-the-art multimodal LLMs on various VQA and multimodal instruction-following benchmarks within the same model size group, as illustrated in Figure 1.

Our contributions can be summarized as follows:

- We introduce CuMo, which integrates co-upcycled sparsely-gated MoE layers into both the MLP connector and the vision encoder, enhancing the multimodal LLM with slightly additional activated parameters from the vision side.
- We outline the training methodology for CuMo, including a three-stage training process with auxiliary losses to stabilize training and ensure a balanced loading of experts.
- We train CuMo exclusively on open-sourced datasets and pre-trained models. It outperforms state-of-the-art open-sourced and private multimodal LLMs across multiple competitive benchmarks within each model size group.

2 Related Works

2.1 Multimodal LLM

While the ultimate goal for mulitmodal models may be generative across various modalities [66, 3, 60], modern multimodal LLMs primarily focus on integrating additional modalities, such as vision, into LLMs. InstructBLIP [11] adopts Q-Former [37] to sample from visual tokens for LLM to feed-forward and follow the instructions. Flamingo [1] and IDEFICS [23, 32] use shared decoder for visual-language understanding. Qwen-VL [2] uses three-stage training to convert QwenLM to Qwen-VL. LLaVA series [46, 44, 45] adopt visual instruction tuning that uses instruction-following data to convert LLM into multimodal LLM. ShareGPT4V [7] collects detailed image caption data from GPT4V to augment the LLaVA models. HoneyBee [5] investigates different designs of the MLP connector for better alignment. VILA [42] unfreezes the LLM during pre-training with interleaved image-text data. MoE-LLaVA [41] adopts the MoE design in small LLMs and reaches comparable performance to LLaVA with large LLMs. VCoder [26] adopts various vision adapters to enhance visual perception abilities. SPHINX [43, 18] adopts multiple visual encoders to enrich the visual features with scaled data and models. InternLM-Xcomposer [69, 12] is trained with interleaved text-image composition data and achieves state-of-the-art performance. InternVL [9] scales up the vision encoder to a 6B ViT model. MM1 [51] summarizes the essential steps towards building a strong multimodal LLM from a pre-trained LLM. Mini-Gemini [39] further collects guided generation into the pipeline.

2.2 Mixture-of-Experts

Mixture-of-Experts [24] is proposed to utilize a set of expert networks to address specific tasks by employing a gating network to determine the selection of these experts. Recently, it has gained popularity in the design of large language models [15]. The mainstream practice [57] is to replace the dense MLP layers with Top-K sparsely-gated mixture-of-experts (MoE) layers in the transformer [64].
MoE in Language Subsequent works [33, 16] have further scaled up MoE-based large language models with improved stability and load balancing of experts. The design of gating networks often involves selecting the top-k experts for each token [57, 33]. Various routing strategies have been explored, such as choosing top-k tokens by experts [71], one-to-one matching between experts and tokens [34]. Besides routing strategies, maintaining the load balance of experts is crucial for training MoE models. ST-MoE [73] adopts loading balancing loss and router-z loss to ensure a balanced distribution of the experts. Upcycling [31] proposes training sparse experts from dense checkpoints to stabilize training and lower the cost. Recent large language models like Gemini-Pro [55] and DBRX [61] are also based on the MoE design.
MoE in Vision The success of MoE extends to the vision community, particularly following the popularity of vision transformers [13, 4, 72, 21, 20, 25, 36]. V-MoE [56] reaches comparable performance to dense ViT while only requiring half of the compute. LIMoE [52] replaces dense MLP layers with MoE layers in CLIP and observes improvements in zero-shot image classification. Residual MoE [65] corporates residual design into MoE transformer and saves over 30% training cost. AdaMV-MoE [8] proposes an adaptive MoE framework for multi-task learning.

3 Method

In this section, we first review the sparse MoE block structure and the upcycling strategy utilized in previous studies. Subsequently, we describe how these sparsely-gated MoE blocks are integrated into each module of multimodal LLMs using co-upcycling strategies. Then, we introduce the three-stage training process and auxiliary loss functions employed to stabilize training and balance the loads of experts.

3.1 Revisit Sparse MoE

Sparse MoE Structure Previous mainstream practice [57] is to replace the dense MLP blocks with sparsely-gated mixture-of-experts blocks. Given input $\mathbf{X} \in \mathbb{R}^{N \times C_{in}}$ and a MLP block,

$$X_{out} = \mathrm{MLP}(X) \in \mathbb{R}^{N \times C_{out}} \tag{1}$$

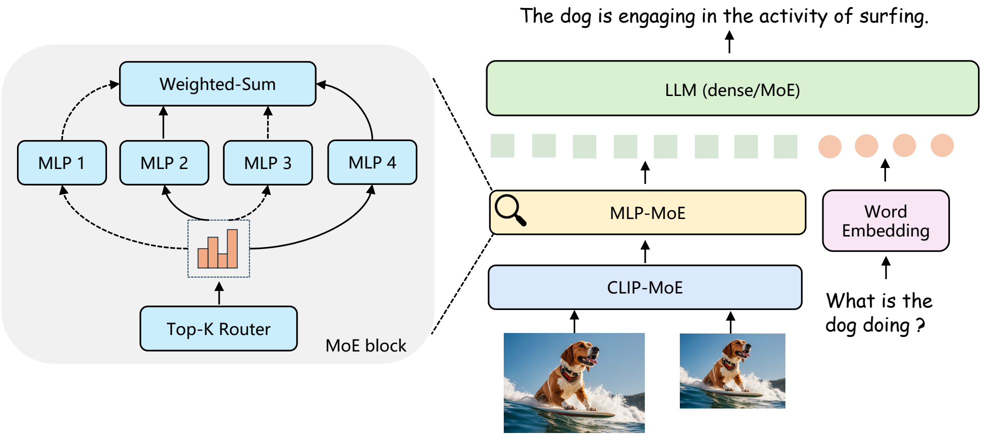


Figure 2: **Architecture of CuMo.** CuMo incorporates sparse Top-K MoE blocks into the CLIP vision encoder and vision-language MLP connector, thereby improving the multimodal LLM capabilities from the vision side. Skip connections are omitted for simplicity. Further implementation details are provided in Section 3.2.

To scale up the model with multiple MLP blocks in parallel, a sparse MoE block includes a router network to select Top-K experts out of S total experts. This router network has a linear layer to compute the normalized weight matrix based on the inputs $\mathbf{X}$ for voting, resulting in

$$W = \text{Softmax}(\text{Linear}(X)) \in \mathbb{R}^{N \times S} \tag{2}$$

The Top-K experts are selected for each token based on $\mathbf{W}$, and the re-normalized weights $\mathbf{W_K} \in \mathbb{R}^{N \times K}$ are computed using

$$W_K = \text{Softmax}(\text{TopK}(W)) \in \mathbb{R}^{N \times K} \tag{3}$$

Each selected expert is represented by an MLP block, and the final output is obtained through a re-weighted sum

$$X_{out} = \sum_{i}^{K} W_K^i \circ \text{MLP}_i(X) \in \mathbb{R}^{N \times C_{out}} \tag{4}$$

the output $\mathbf{X}_{out}$ maintains the same dimension as the output of a single dense MLP block.

Sparse Upcycling Training MoE-based designs from scratch can be unstable and costly. Sparse Upcycling [31] addresses this challenge by initializing the experts in each MoE block from the corresponding MLP block in pre-trained dense checkpoints. This initialization approach provides a better starting point for training MoE-based models and reduces training costs compared to training from scratch.

3.2 CuMo Architecture

Sparse MoE in MLP Connector The MLP connector converts visual tokens into word embedding space, aligning dimensions between visual and text tokens. An effective architecture for the vision-language connector is an MLP block [44] that contains two linear layers. We start from a single MLP block and replace it with a Top-K sparse MoE block, incorporating a Top-K router and a set of experts for projecting visual tokens into word embedding space.

Sparse MoE in Vision Encoder Vision encoders extract image features as sequences of visual tokens for reasoning in LLMs. CLIP [54] is one the most popular pre-trained vision encoders for multimodal LLM since it is pre-trained on large-scale image-text pairs, which makes it suitable for processing images for multimodal usage. The visual encoding part of CLIP is a ViT [13] model, which has consecutive MLP blocks in the transformer encoder. We substitute each MLP block with a Top-K sparse MoE block, retaining skip connections alongside MoE block outputs.

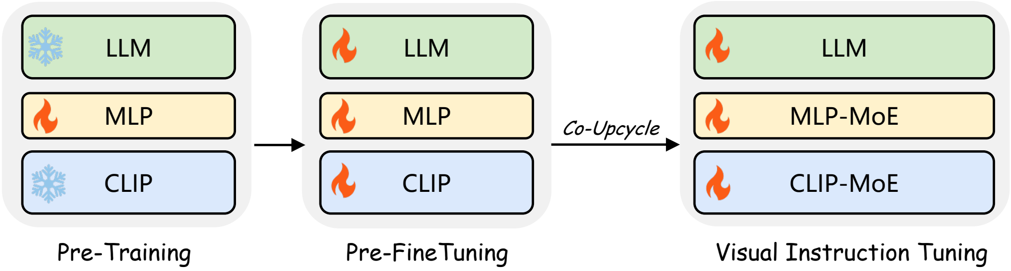


Figure 3: **Training Stages of CuMo.** The first stage involves pre-training the MLP for better alignment. Subsequently, the pre-finetuning stage trains all parameters as a warm-up before the next stage. Finally, the MLP experts within each MoE block are initialized from the weights of the corresponding MLP block, followed by training all parameters in the visual instruction tuning stage.

Sparse MoE in LLM In terms of using MoE in LLM, we compare the co-upcycled LLM with pre-trained MoE-based LLM. We start from Mistral-7B and the upcycled Mistral-7B-MoE slightly outperforms Mistral-7B on certain benchmarks. However, considering the constrained knowledge base of upcycled experts from Mistral-7B, we compare it with the pre-trained Mixtral 8x7B with pre-trained experts of a diverse knowledge base. Experimental results reveal that pre-trained Mixtral 8x7B significantly outperforms Mistral-7B-MoE. As a result, LLM is not co-upcycled with CLIP and MLP connectors since it brings marginal improvements with great additional parameters.

3.3 Training Recipe

Co-Upcycling MoE blocks We start with training the added MoE blocks from scratch while the model is struggling to converge. Attempts to address this issue with lower learning rates perform worse compared to the baseline. As a result, we adopt a co-upcycling approach, initializing each module that integrates sparsely-gated MoE blocks with pre-trained MLPs to replace corresponding MLP blocks, as shown in Figure 1. This strategy consistently improves training stability and model performance.

Three-Stage Training To further enhance training stability, we adopt a three-stage training strategy for CuMo models, as illustrated in Figure 3. In the first stage, we only pre-train the MLP connector, given that the vision encoder and LLM have already undergone pre-training on large-scale data. During the second pre-finetuning stage, we train all parameters using high-quality caption data to warm up the entire model before introducing MoE blocks in the subsequent stage. The third stage involves visual instruction finetuning, where the multimodal LLM is scaled up with upcycled MoE blocks and trained on visual instruction tuning data.

Loss Function To maintain a load balance between experts in each MoE block, we adopt auxiliary losses based on the language modeling cross-entropy loss. The auxiliary losses comprise loading balance loss and router z-loss [73]. Hence, the total loss is

$$L = L_{ce} + \alpha_b L_b + \alpha_z L_z \tag{5}$$

Here, L_{ce} represents the language modeling loss, which computes the cross-entropy of next-token predictions. α_b and α_z denote coefficients for loading balance loss L_b and router z-loss L_z, set to 0.1 and 0.01, respectively, across all experiments. These auxiliary losses, abbreviated as bzloss in Section 4, are individually applied to the MLP connector, vision encoder, and LLM for simplicity.

4 Experiments

We train the CuMo models on a mixture of open-sourced datasets, which are converted into the visual instruction tuning format. Then, we conduct comprehensive evaluations of the performance of CuMo models across various competitive VQA-based and instruction-following-based benchmarks. Additionally, we perform ablation studies on each module with upcycled MoE blocks with qualitative analysis of the results.

Method	LLM	Act. (B)	SQA IMG	Text VQA	GQA	POPE	MME	MMB EN	MMB CN	MM Vet	VQA v2	LLaVA Wild	SEED IMG	MMMU val
7B to 13B Models														
InstructBLIP [11]	Vicuna-7B	7.9	60.5	50.1	49.2	-	-	36.0	23.7	26.2	-	60.9	60.5	-
Qwen-VL-Chat [2]	Qwen-7B	-	68.2	61.5	57.5	-	1487.5	60.6	56.7	-	78.2	-	58.2	35.9
LLaVA-v1.5 [44]	Vicuna-7B	7.1	66.8	58.2	62.0	85.9	1510.7	64.3	58.3	30.5	78.5	63.4	66.1	-
VILA [42]	Vicuna-7B	7.1	68.2	64.4	62.3	85.5	1533.0	68.9	61.7	34.9	79.9	69.7	61.1	-
ShareGPT4V [7]	Vicuna-7B	7.1	68.4	-	-	-	1567.4	68.8	62.2	37.6	80.6	72.6	69.7	-
LLaMA-VID [38]	Vicuna-7B	-	68.3	-	64.3	86.0	1521.4	65.1	-	-	79.3	-	59.9	-
SPHINX-Intern2 [18]	InternLM2-7B	-	70.4	58.1	56.2	86.9	1260.4	57.9	-	36.5	75.5	57.6	68.8	-
LLaVA-NeXT [45]	Mistral-7B	7.6	72.8	65.7	64.8	86.7	1498	68.7	61.2	47.3	82.2	83.2	72.2	35.3
LLaVA-NeXT [45]	Vicuna-7B	7.1	70.1	64.9	64.2	86.5	1519	67.4	60.6	43.9	81.8	81.6	70.2	35.8
Mini-Gemini [39]	Vicuna-7B	7.3	65.2	-	-	-	1523	69.3	-	40.8	-	-	-	36.1
MM1 [51]	MM1-7B	-	72.6	72.8	-	86.6	1529.3	79.0	-	42.1	82.8	81.5	69.9	37.0
InstructBLIP [11]	Vicuna-13B	14.2	63.1	50.7	49.5	78.9	1212.8	-	-	25.6	-	58.2	63.1	-
LLaVA-v1.5 [44]	Vicuna-13B	13.4	71.6	61.3	63.3	85.9	1531.3	67.7	63.6	35.4	80.0	70.7	68.2	36.4
VILA [42]	Vicuna-13B	13.4	73.7	66.6	63.3	84.2	1570.1	70.3	64.3	38.8	80.8	73.0	62.8	-
InternVL-Chat [9]	Vicuna-13B	19	-	61.5	**66.6**	87.6	1586.4	-	-	-	81.2	-	-	-
LLaMA-VID [38]	Vicuna-13B	-	70.0	-	65.0	86.0	1542.3	66.6	-	-	80.0	-	62.3	-
SPHINX-Plus [18]	LLaMA2-13B	-	74.2	65.7	-	89.1	1457.7	71.0	-	47.9	-	71.7	74.8	-
Mini-Gemini[39]	Vicuna-13B	13.6	65.9	-	-	-	1565	68.5	-	46.0	-	-	-	38.1
LLaVA-NeXT [45]	Vicuna-13B	13.4	73.6	67.1	65.4	86.2	1575	70	64.4	48.4	82.8	**87.3**	71.9	36.2
CuMo	Mistral-7B	7.8	73.9	67.0	64.9	86.7	1548.6	73.0	66.6	**51.0**†	82.2	85.7†	72.1	39.1
7B MoE Models														
SPHINX-MoE [18]	Mixtral-8×7B	-	74.5	68.0	63.8	**89.6**	1485.3	71.3	-	40.9	81.1	70.2	73.0	31.1
MM1 [51]	MM1-7B-MoE	-	75.3	**72.8**	-	87.6	1629.0	**79.7**	-	47.0	**83.4**	82.0	70.4	40.9
Mini-Gemini [39]	Mixtral-8×7B	13.5	-	69.2	-	-	1639	75.6	-	45.8	-	-	-	41.8
CuMo	Mixtral-8×7B	13.5	**77.9**	66.0	63.8	85.7	**1639.5**	75.3	**68.0**	48.7†	81.8	84.7†	**73.2**	**45.0**

Table 1: Comparisons between CuMo and other state-of-the-art multimodal LLMs on competitive benchmarks. These models are grouped by the size of the base LLM and **bold** indicates the best performance on a certain benchmark. Act.: activated parameters during inference. Numbers with † are averaged by three inference runs of querying GPT API.

4.1 Implementation Details

Training Datasets During pre-training, we only utilize LLaVA-558K [46] to train the MLP connector for better alignment. In the subsequent pre-finetuning stage, detailed image caption data from ALLaVA [6] is employed to warm up all parameters of the multimodal LLM. For the final visual instruction tuning stage, a mixture of datasets including LLaVA-665K [44], ShareGPT4V [7], LAION-GPT-V [14], DocVQA [62], ChartQA [49], AI2D [29], InfoVQA [50], SynDog-EN [30], ALLaVA [6], and LIMA [70] is utilized to train the CuMo models with upcycled MoE blocks. The total data size for visual instruction tuning is approximately 1.65 million, and all training data are publicly accessible. The detailed breakdown of the training dataset is listed in Appendix A.

Evaluation Benchmarks Evaluation of CuMo models primarily focuses on academic VQA-based datasets such as VQAv2 [19], GQA [22], Science-QA [48], and TextVQA [59], as well as instruction-following-based LMM benchmarks including POPE [40], MME [17], MMBench [47], SEED-Bench [35], LLaVA-Wild [46], and MM-Vet [67]. Additionally, the challenging MMMU [68] is evaluated to assess the visual reasoning abilities of the multimodal LLMs.

Training Settings We employ the pre-trained CLIP ViT-L [54] as the vision encoder, a two-layer MLP as the vision-language connector, and Mistral-7B [27] as the LLM to establish the baseline model following LLaVA v1.5 [44]. We only use LLaVA-558K [44] as pre-training data and LLaVA-665K [44] as visual instruction tuning data to train the baseline model and make ablation studies for comparisons. The learning rate is set to 1e-3 for pre-training the MLP connector and reduced to 2e-5 for visual instruction tuning of both the MLP connector and CLIP. To further stabilize the visual instruction tuning process after scaling up with additional data, the learning rate is lowered to 2e-6 for all parameters of the CuMo models in the final results. More hyperparameters of the training process is listed in Appendix B.

Evaluation Settings During evaluation, we adhere to the settings outlined in the LLaVA series [44], employing a greedy decoding strategy for all benchmarks. The data and questions are converted into visual instructions to prompt the multimodal LLMs. For benchmarks that utilize GPT API for evaluation, we adopt gpt-4-0613 for LLaVA-Wild [46].

				SQA	Text				MMBench		MM	VQA	LLaVA	SEED
Method	LLM	PT	IT	IMG	VQA	GQA	POPE	MME	EN	CN	Vet	v2	Wild	IMG
InstructBLIP [11]	Vicuna-7B	129M	1.2M	60.5	50.1	49.2	-	-	36.0	23.7	26.2	-	60.9	60.5
InstructBLIP [11]	Vicuna-13B	129M	1.2M	63.1	50.7	49.5	78.9	1212.8	-	-	25.6	-	58.2	63.1
IDEFICS-9B [23]	LLaMA-7B	353M	1M	-	25.9	38.4	-	-	48.2	25.2	-	50.9	-	-
IDEFICS-80B [23]	LLaMA-65B	353M	1M	-	30.9	45.2	-	-	54.5	38.1	-	60.0	-	-
Qwen-VL [2]	Qwen-7B	1.4B	50M	67.1	**63.8**	59.3	-	-	38.2	7.4	-	78.8	-	56.3
Qwen-VL-Chat [2]	Qwen-7B	1.4B	50M	68.2	61.5	57.5	-	1487.5	60.6	56.7	-	78.2	-	58.2
LLaVA-v1.5 [44]	Vicuna-7B	558K	665K	66.8	58.2	62.0	85.9	**1510.7**	64.3	58.3	30.5	78.5	63.4	66.1
LLaVA-v1.5[†]	Mistral-7B	558K	665K	**72.8**	57.6	60.0	86.3	1414.9	66.5	60.1	32.1	78.2	**69.4**	66.4
CuMo	Mistral-7B	558K	665K	71.7	59.3	**63.2**	**87.1**	1428.6	**69.6**	**62.6**	**34.3**	**80.6**	68.8	**69.6**

Table 2: Comparisons between CuMo Mistral-7B and other multimodal LMM models with limited training data. The best performance are highlighted in **bold**. LLaVA-v1.5[†] with Mistral-7B is reproduced by us as a baseline model.

4.2 Main Results

Comparison with SoTA Multimodal LLMs In Table 1, we present a comparison of CuMo models with other state-of-the-art instruction-following-based multimodal LLMs. We categorize the models based on the size of the base LLMs, including 7B models, 13B models, and 7B MoE models. CuMo Mistral-7B outperforms other 7B-based state-of-the-art multimodal LLMs across multiple benchmarks. Moreover, the performance of the CuMo Mistral-7B model is comparable to many 13B-based multimodal LLMs. In the case of Mixtral-8×7B models, CuMo achieves results on par with SPHINX-MoE, MM1, and Mini-Gemini. LLaMA-based LLMs [10, 63] are not utilized in our experiments due to license constraints.

Comparison under limited training data To further evaluate the effectiveness of the co-upcycled MoE blocks, we train the vanilla CuMo mistral-7B under limited training data in Table 2. It shows that CuMo outperforms other 7B models and reaches comparable performance to LLaVA-v1.5 Vicuna-13B under the same training data.

4.3 Ablation Study

Upcycle MLP connector to MLP-MoE We initiate the ablation study by replacing the MLP connector with upcycled MLP-MoE, as depicted in Table 3(a). We start with a Top 2-in-4 router and train the MoE blocks from scratch, which leads to a clear performance drop on all benchmarks. Then, we adopt the upcycling strategy to initialize the MLP experts. We observe marginal improvements over the baseline, considering each expert comprises only two linear layers. Subsequently, the incorporation of bzloss to ensure a balanced loading of experts in the MLP-MoE yields noticeable enhancements on MMVet. However, employing a Top 2-in-8 router with upcycling and bzloss results in a slight performance decline, possibly due to the limited visual instruction tuning data to train robust and well-balanced eight experts.

Empower CLIP with CLIP-MoE In Table 3(b), initially unfreezing CLIP based on MLP-MoE leads to noticeable improvements on TextVQA and MMVet benchmarks. However, training with the added Top2-in-4 MoE blocks in CLIP from scratch proves unsuccessful, as the model fails to converge even with largely reduced learning rates. Consequently, adopting upcycled MoE blocks during the visual instruction tuning stage yields further enhancements on the TextVQA, MMVet, and SEED benchmarks, as well as a more stable training process.

Upcycle LLM vs Pre-trained LLM-MoE Upon replacing all MLP blocks with sparsely-gated MoE blocks in the visual part, we further investigate the utilization of the MoE architecture in the LLM. Starting from the Mistral-7B model, we first lower the learning rate to 2e-6 to set the baseline and the following experiments since a learning rate of 2e-5 induces training instabilities. Then, we upcycle each MLP block with a sparsely-gated MoE block, initializing the weight of each expert from the pre-trained MLP block. As demonstrated in Table 3(c), the upcycled Mistral-4×7B and 8×7B outperform the Mistral-7B model slightly except for TextVQA. However, considering that the upcycled experts significantly increase parameters without introducing new knowledge, we replace the upcycled Mistral 8×7B with Mixtral 8×7B [28]. In Mixtral 8×7B, all expert layers are pre-trained on large-scale language data, providing superior initialization and similar training stability compared to upcycling. The results indicate that CuMo Mixtral-8x7B outperforms its upcycled

Method	SQA	VQA^T	MMVet	SEED
Baseline on Mistral-7B	72.8	57.6	32.1	66.4
+ *Top 2-in-4 & Scratch*	68.1	55.6	29.3	65.1
$\rightleftharpoons$ *Top 2-in-4 & Upcycle*	73.7	57.2	32.3	67.1
+ *bzloss*	73.5	57.4	33.1	67.4
$\rightleftharpoons$ *Top 2-in-8 & Upcycle*	73.4	57.6	32.4	67.2

(a) MLP-MoE

Method	SQA	VQA^T	MMVet	SEED
MLP-MoE	73.5	57.4	33.1	67.4
+ *Unfreeze CLIP*	72.0	58.9	34.7	69.0
+ *Top 2-in-4 & bzloss*	72.8	59.7	35.4	69.8
$\rightleftharpoons$ *Top 2-in-8 & bzloss*	71.0	59.0	33.6	69.2

(b) CLIP-MoE

Method	SQA	VQA^T	MMVet	SEED
MLP-MoE & CLIP-MoE	72.8	59.7	35.4	69.8
+ *lower lr to 2e-6*	71.7	59.3	34.3	69.6
+ *Mistral 4×7B & Upcycle*	72.8	57.0	35.2	69.9
$\rightleftharpoons$ *Mistral 8×7B & Upcycle*	73.2	56.4	35.7	70.5
$\rightleftharpoons$ *Mixtral 8×7B*	74.2	60.6	40.0	72.6

(c) LLM-MoE

1×	2×	3×	SQA	VQA^T	MMVet	SEED
✓	-	-	71.7	59.3	34.3	69.6
✓	✓	-	71.7	60.6	35.0	69.7
✓	-	✓	72.9	61.0	37.0	69.7
✓	✓	✓	72.2	60.5	36.9	70.1

(d) Multi-resolution Feature

Method	SQA	VQA^T	MMVet	SEED
No PFT	71.7	59.3	34.3	69.6
+ *ShareGPT4V*	72.4	61.7	36.5	70.0
$\rightleftharpoons$ *ALLaVA*	73.0	62.8	37.2	70.9

(e) Pre-FineTuning Stage

Method	CLIP	MLP	LLM	Total
Mistral-7B	0.30	0.025	7.25	7.58
+ *MLP-MoE*	0.30	0.05	7.25	7.60
+ *CLIP-MoE*	0.50	0.05	7.25	7.80
$\rightleftharpoons$ *Mixtral-8x7B*	0.50	0.05	12.90	13.45

(f) Activated billions of parameters

Table 3: **Ablation Studies during building CuMo.** Each row represents a different configuration, with changes or additions marked using $\rightleftharpoons$ and $+$ symbols, respectively. Settings highlighted with a light blue background are those adapted for final model in Table 1. For (b): all MoE blocks in CLIP are initialized with upcycling.

counterparts significantly and is employed in the final models with bzloss to maintain a balanced loading of experts.

Multi-Resolution Visual Features Incorporating multi-resolution inputs is crucial for enhancing the understanding of image content in multimodal LLMs. Following the approach outlined in S^2[58], we introduce multi-resolution inputs to CLIP and concatenate the feature maps channel-wise to maintain the total number of visual tokens consistent with low-resolution inputs. As illustrated in Table 3(d), an empirical combination of 3× and 1× reaches the best performance and we adopt this configuration for the final CuMo models.

Pre-FineTuning Stage Previous ablation studies were conducted directly after the pre-training of the MLP connector, leading to observed training instabilities during visual instruction tuning. To address this, we introduce a pre-finetuning stage using high-quality image caption data, wherein all parameters are unfrozen. In Table 3(e), we leverage caption data from ALLaVA for this stage. Results indicate that ALLaVA data proves to be a superior option, providing fewer but higher-quality captions for training, ultimately leading to improved performance.

Added Parameters In Table 3(f), we keep track of the added activated parameters during inference of CuMo. It shows that adding MoE blocks in the vision side upon MLP connector and CLIP only brings 0.22B extra parameters compared to the baseline model on Mistral-7B. More details can be found in Appendix C.

4.4 More Analysis

Expert Distribution As shown in Figure 4, we visualize the expert distributions in the MoE block from selected layers in CLIP during inference. The dataset used for analyzation is the test set of the MME benchmark. The distribution indicates that the selected experts during infer-

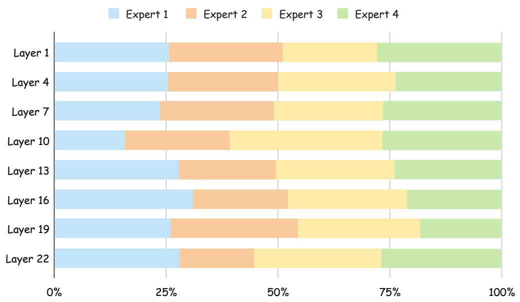
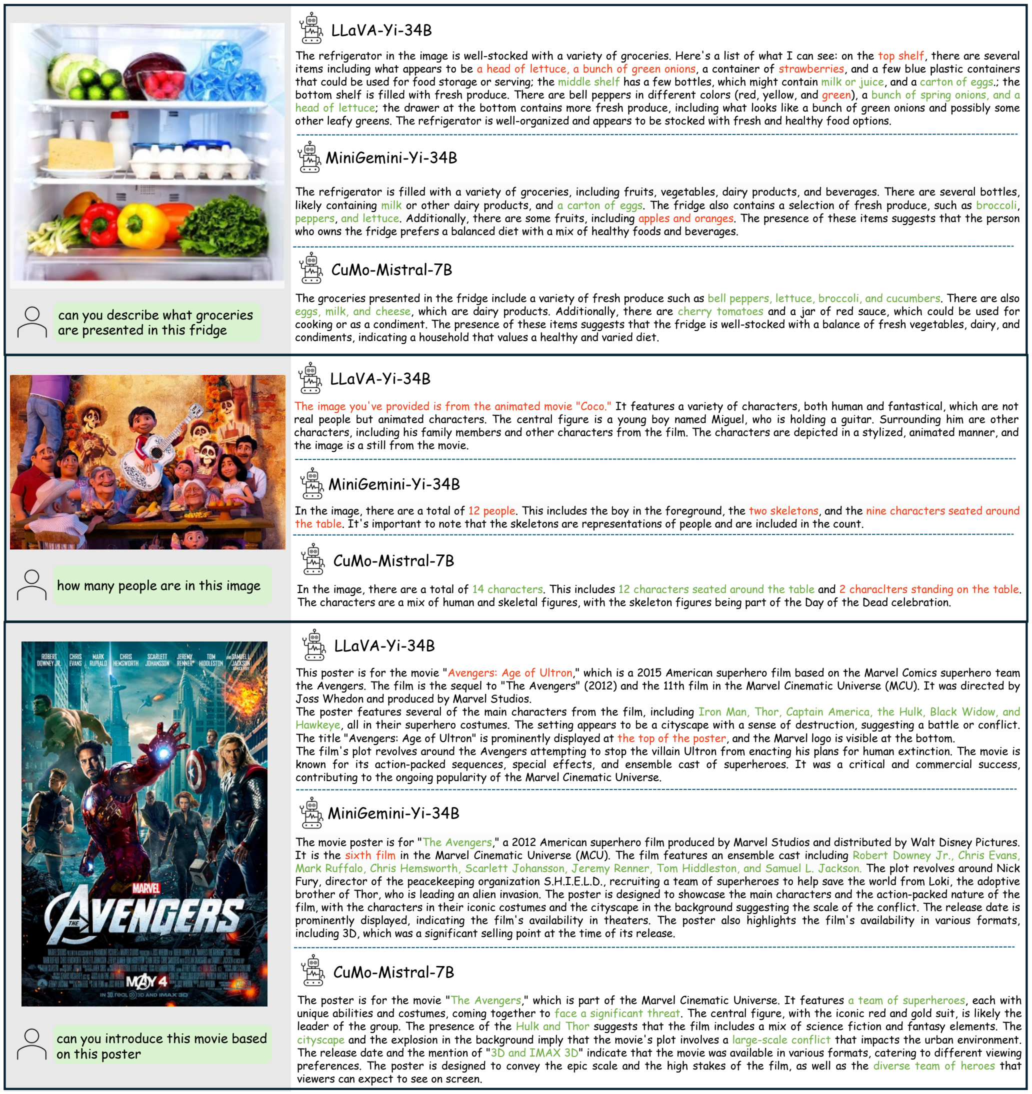


Figure 4: **Expert distributions of MoE blocks in CLIP.** We select layers from CLIP and summarize the activated experts during inference.



can you describe what groceries are presented in this fridge

LLaVA-Yi-34B

The refrigerator in the image is well-stocked with a variety of groceries. Here's a list of what I can see: on the top shelf, there are several items including what appears to be a head of lettuce, a bunch of green onions, a container of strawberries, and a few blue plastic containers that could be used for food storage or serving; the middle shelf has a few bottles, which might contain milk or juice, and a carton of eggs.; the bottom shelf is filled with fresh produce. There are bell peppers in different colors (red, yellow, and green), a bunch of spring onions, and a head of lettuce; the drawer at the bottom contains more fresh produce, including what looks like a bunch of green onions and possibly some other leafy greens. The refrigerator is well-organized and appears to be stocked with fresh and healthy food options.

MiniGemini-Yi-34B

The refrigerator is filled with a variety of groceries, including fruits, vegetables, dairy products, and beverages. There are several bottles, likely containing milk or other dairy products, and a carton of eggs. The fridge also contains a selection of fresh produce, such as broccoli, peppers, and lettuce. Additionally, there are some fruits, including apples and oranges. The presence of these items suggests that the person who owns the fridge prefers a balanced diet with a mix of healthy foods and beverages.

CuMo-Mistral-7B

The groceries presented in the fridge include a variety of fresh produce such as bell peppers, lettuce, broccoli, and cucumbers. There are also eggs, milk, and cheese, which are dairy products. Additionally, there are cherry tomatoes and a jar of red sauce, which could be used for cooking or as a condiment. The presence of these items suggests that the fridge is well-stocked with a balance of fresh vegetables, dairy, and condiments, indicating a household that values a healthy and varied diet.

how many people are in this image

LLaVA-Yi-34B

The image you've provided is from the animated movie "Coco." It features a variety of characters, both human and fantastical, which are not real people but animated characters. The central figure is a young boy named Miguel, who is holding a guitar. Surrounding him are other characters, including his family members and other characters from the film. The characters are depicted in a stylized, animated manner, and the image is a still from the movie.

MiniGemini-Yi-34B

In the image, there are a total of 12 people. This includes the boy in the foreground, the two skeletons, and the nine characters seated around the table. It's important to note that the skeletons are representations of people and are included in the count.

CuMo-Mistral-7B

In the image, there are a total of 14 characters. This includes 12 characters seated around the table and 2 characlters standing on the table. The characters are a mix of human and skeletal figures, with the skeleton figures being part of the Day of the Dead celebration.

can you introduce this movie based on this poster

LLaVA-Yi-34B

This poster is for the movie "Avengers: Age of Ultron," which is a 2015 American superhero film based on the Marvel Comics superhero team the Avengers. The film is the sequel to "The Avengers" (2012) and the 11th film in the Marvel Cinematic Universe (MCU). It was directed by Joss Whedon and produced by Marvel Studios.

The poster features several of the main characters from the film, including Iron Man, Thor, Captain America, the Hulk, Black Widow, and Hawkeye, all in their superhero costumes. The setting appears to be a cityscape with a sense of destruction, suggesting a battle or conflict.

The title "Avengers: Age of Ultron" is prominently displayed at the top of the poster, and the Marvel logo is visible at the bottom.

The film's plot revolves around the Avengers attempting to stop the villain Ultron from enacting his plans for human extinction. The movie is known for its action-packed sequences, special effects, and ensemble cast of superheroes. It was a critical and commercial success, contributing to the ongoing popularity of the Marvel Cinematic Universe.

MiniGemini-Yi-34B

The movie poster is for "The Avengers," a 2012 American superhero film produced by Marvel Studios and distributed by Walt Disney Pictures. It is the sixth film in the Marvel Cinematic Universe (MCU). The film features an ensemble cast including Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, Scarlett Johansson, Jeremy Renner, Tom Hiddleston, and Samuel L. Jackson. The plot revolves around Nick Fury, director of the peacekeeping organization S.H.I.E.L.D., recruiting a team of superheroes to help save the world from Loki, the adoptive brother of Thor, who is leading an alien invasion. The poster is designed to showcase the main characters and the action-packed nature of the film, with the characters in their iconic costumes and the cityscape in the background suggesting the scale of the conflict. The release date is prominently displayed, indicating the film's availability in theaters. The poster also highlights the film's availability in various formats, including 3D, which was a significant selling point at the time of its release.

CuMo-Mistral-7B

The poster is for the movie "The Avengers," which is part of the Marvel Cinematic Universe. It features a team of superheroes, each with unique abilities and costumes, coming together to face a significant threat. The central figure, with the iconic red and gold suit, is likely the leader of the group. The presence of the Hulk and Thor suggests that the film includes a mix of science fiction and fantasy elements. The cityscape and the explosion in the background imply that the movie's plot involves a large-scale conflict that impacts the urban environment. The release date and the mention of "3D and IMAX 3D" indicate that the movie was available in various formats, catering to different viewing preferences. The poster is designed to convey the epic scale and the high stakes of the film, as well as the diverse team of heroes that viewers can expect to see on screen.

Figure 5: **Dialogues between the user and multimodal LLMs on challenging images.** We highlight the correct answers and hallucinations from the responses of the multimodal LLMs.

ence are evenly spread across layers, providing further evidence of the effectiveness of the auxiliary losses in maintaining load balance.

Dialogue Comparisons Presented in Figure 5, we contrast the responses from CuMo-Mistral-7B, LLaVA-Yi-34B, and MiniGemini-Yi-34B under challenging content understanding cases. It demonstrates that CuMo-Mistral-7B can effectively follow instructions and provide mostly correct answers to challenging questions derived from complex scenes. However, CuMo also exhibits instances of hallucinations, such as responding with “2 characters standing on the table”, highlighting the need for further investigation to mitigate hallucinations and improve reliability of CuMo.

Limitations The main limitation of CuMo is that, similarly to other large language models, it can generate hallucinated responses. This may constrain its potentials in real-world multimodal applications like used as a chatbot. Future works, such as Reinforcement Learning with Human Feedback (RLHF) and Retrieval Augmented Generation (RAG), can be undertaken to mitigate these hallucinations and improve the model’s reliability.

5 Conclusion

In this study, we introduce the sparse mixture-of-experts design into multimodal LLMs from the vision side. Specifically, we replace each MLP block with a Top-K sparse MoE block in the MLP connector and the vision encoder. To enhance training stability, we employ a three-stage training approach, incorporating upcycled MoE blocks during the visual instruction tuning stage, along with auxiliary bzloss to maintain a balanced loading of experts. All CuMo models are trained and evaluated on fully open-sourced datasets and benchmarks. Through extensive experiments and ablation studies, we validate the effectiveness of the upcycled MoE blocks in each module. CuMo outperforms state-of-the-art models across multiple competitive benchmarks within the same group of model sizes.

Acknowledgments We extend our gratitude to Chunyuan Li, Lei Chen, and Haibin Lin for their insightful and valuable discussions throughout this project. Li, Jain, Shi are in part supported by National Science Foundation CAREER Award #2427478, and by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education under Award #2229873 - National AI Institute for Exceptional Education, Beckman Institute and ECE Department at UIUC, and Georgia Institute of Technology.

References

[1] 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. 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 3, 6, 7

[3] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning (ICML)*, pages 1692–1717. PMLR, 2023. 3

[4] 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*. Springer, 2020. 3

[5] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. *arXiv preprint arXiv:2312.06742*, 2023. 1, 3

[6] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv:2402.11684*, 2024. 6

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. 1, 3, 6

[8] Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17346–17357, October 2023. 3

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 1, 3, 6

[10] WL Chiang, Z Li, Z Lin, Y Sheng, Z Wu, H Zhang, L Zheng, S Zhuang, Y Zhuang, JE Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, mar. 2023. 7

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 1, 3, 6, 7

[12] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024. 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4

[14] LAION eV. Laion/gpt4v-dataset · datasets at hugging face. 6

[15] William Fedus, Jeff Dean, and Barret Zoph. A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*, 2022. 3

[16] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022. 3

[17] 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:2306.13394*, 2023. 6

[18] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, Wenqi Shao, Chao Xu, Conghui He, Junjun He, Hao Shao, Pan Lu, Hongsheng Li, and Yu Qiao. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *ArXiv*, abs/2402.05935, 2024. 1, 3, 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 *CVPR*, 2017. 6

[20] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv preprint arXiv:2209.15001*, 2022. 3

[21] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[22] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 6

[23] IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics, 2023. 3, 7

[24] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 3

[25] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 3

[26] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. *arXiv preprint arXiv:2312.14233*, 2023. 3

[27] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 6

[28] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv:2401.04088*, 2024. 2, 7

[29] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 6

[30] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 6

[31] Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints, 2023. 3, 4

[32] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 3

[33] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam M. Shazeer, and Z. Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *ArXiv*, abs/2006.16668, 2020. 3

[34] Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. Base layers: Simplifying training of large, sparse models, 2021. 3

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

[36] Jiachen Li, Vidit Goel, Marianna Ohanyan, Shant Navasardyan, Yunchao Wei, and Humphrey Shi. Vmformer: End-to-end video matting with transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6678–6687, 2024. 3

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[38] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv:2311.17043*, 2023. 6

[39] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models, 2024. 1, 3, 6

[40] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*, 2023. 6

[41] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 3

[42] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023. 1, 3, 6

[43] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models, 2023. 1, 3

[44] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 3, 4, 6, 7

[45] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 1, 3, 6

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 3, 6

[47] 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:2307.06281*, 2023. 6

[48] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022. 6

[49] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv:2203.10244*, 2022. 6

[50] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 6

[51] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis, insights from multimodal llm pre-training, 2024. 1, 2, 3, 6

[52] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts, 2022. 3

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

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4, 6

[55] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jeffrey Dean, and et al. Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 2, 3

[56] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts, 2021. 3

[57] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. 2, 3

[58] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? *arXiv preprint arXiv:2403.13043*, 2024. 8

[59] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 6

[60] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[61] The Mosaic Research Team. Introducing dbrx: A new state-of-the-art open llm, March 2024. 2, 3

[62] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In *ICDAR 2021*, 2021. 6

[63] Hugo Touvron and et al. Louis Martin. Llama 2: Open foundation and fine-tuned chat models, 2023. 7

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017. 1, 3

[65] Lemeng Wu, Mengchen Liu, Yinpeng Chen, Dongdong Chen, Xiyang Dai, and Lu Yuan. Residual mixture of experts, 2022. 3

[66] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 3

[67] 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:2308.02490*, 2023. 6

[68] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 6

[69] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 3

[70] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[71] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing, 2022. 3

[72] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. 3

[73] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models, 2022. 3, 5

Appendix

The supplementary material elaborates on further aspects of our work concerning the experimental setups and dataset usage. In Appendix A, we provide details on the datasets used for the visual instruction tuning stage and how we converted the mixture of datasets into the visual instruction following formats. In Appendix B, we present the hyperparameters used for the three-stage trainings. In Appendix E, we include additional examples of dialogues between the user and our CuMo models.

A Dataset Details

As outlined in Table 4, we provide detailed information on the datasets utilized for the three-stage training process mentioned in Section 3.3. All data are converted into the instruction-following format for training. For the Syndog-EN and DVQA datasets, we didn't use the entire training set as we observed that a large portion of synthetic data negatively impacts the zero-shot performance of the multimodal LLMs.

Dataset	Size
Pre-Training	
LCS-558K	558K
Pre-Finetuning	
ALLaVA-Caption	708K
Visual Instruction Tuning	
LLaVA-665K	665K
ShareGPT4V	102K
LAION-GPT-V	11K
DocVQA	10K
SynDog-EN	50K
ChartQA	4K
DVQA	50K
AI2D	2K
InfoVQA	4K
ALLaVA	708K
LIMA	1K
ALLaVA-Text	143K

Table 4: List of datasets used for three training stages.

Model	Vision Encoder	ImageNet Acc.	Res.	Params.	TextVQA	MMVet	SEED
LLaVA-v1.5	CLIP-ViT-L	76.6	336	0.30B	57.6	32.1	66.4
CuMo	CLIP-ViT-L	76.6	336	0.50B	59.3	34.3	69.6
LLaVA-v1.5	SigLIP-SO400M	83.2	384	0.43B	58.1	32.5	67.5
CuMo	SigLIP-SO400M	83.2	384	0.72B	59.4	34.1	69.8

Table 5: CuMo under different vision encoders.

B Experimental Setup Details

Table 6 provides an overview of the main hyperparameters used during the three-stage training process. For the final results presented in Table 1, the model was trained using 32 × A100 GPUs with a total batch size of 256 and a learning rate of 4e-6. All ablation studies were conducted with a total batch size of 128 and learning rates of 2e-5 and 2e-6, as detailed in Section 4.3.

Hyperparameter	PT	PFT	VIT
Learning rate	1e-3	2e-6	4e-6
LR schedule	Cosine	Cosine	Cosine
Batchsize per GPU	32	8	8
GPUs	8×A100	16×A100	32×A100
Zero	Zero2	Zero3	Zero3-offload
Optimizer	AdamW	AdamW	AdamW
MLP	Open	Open	Open
CLIP	Freeze	Open	Open
LLM	Freeze	Open	Open
MoE blocks	-	-	✓
Max token	2048	4096	4096

Table 6: Hyperparameters used in three-stage training on Mistral-7B. PT: Pre-Training stage. PFT: Pre-FineTuning stage. VIT: Visual Instruction tuning stage.

C More Vision Encoders

Table 5 shows CuMo with different vision encoders. Here we use the pre-trained SigLIP-SO400M as the vision encoder and add MoE to the vision encoder. SigLIP-SO400M has a much better performance on ImageNet zero-shot classification than CLIP-ViT-L (83.2 vs 76.6). The added MoE can still make improvements to this stronger vision encoder but the average improvement shrinks compared to CLIP-ViT-L. However, the training data here is limited to LLaVA-665K for quick verification, which may not show the full potential of the model if training with more data.

CuMo	CLIP	MLP	LLM	Total
Mistral-7B	0.30B	0.025B	7.25B	7.58B
$\rightleftharpoons$ Activation Params	0.30B	0.025B	7.25B	7.58B
+ Top 2-in-4 MLP-MoE	0.30B	0.10B	7.25B	7.65B
$\rightleftharpoons$ Activation Params	0.30B	0.05B	7.25B	7.60B
+ Top 2-in-4 CLIP-MoE	0.91B	0.10B	7.25B	8.26B
$\rightleftharpoons$ Activation Params	0.50B	0.05B	7.25B	7.80B
$\rightleftharpoons$ Mixtral-8x7B	0.91B	0.10B	46.70B	47.71B
$\rightleftharpoons$ Activation Params	0.50B	0.05B	12.90B	13.45B

Table 7: Change of model parameters of CuMo. The 7.80B and 13.45B activation parameters corresponds to Act. of CuMo in Table 1.

D Model Parameters

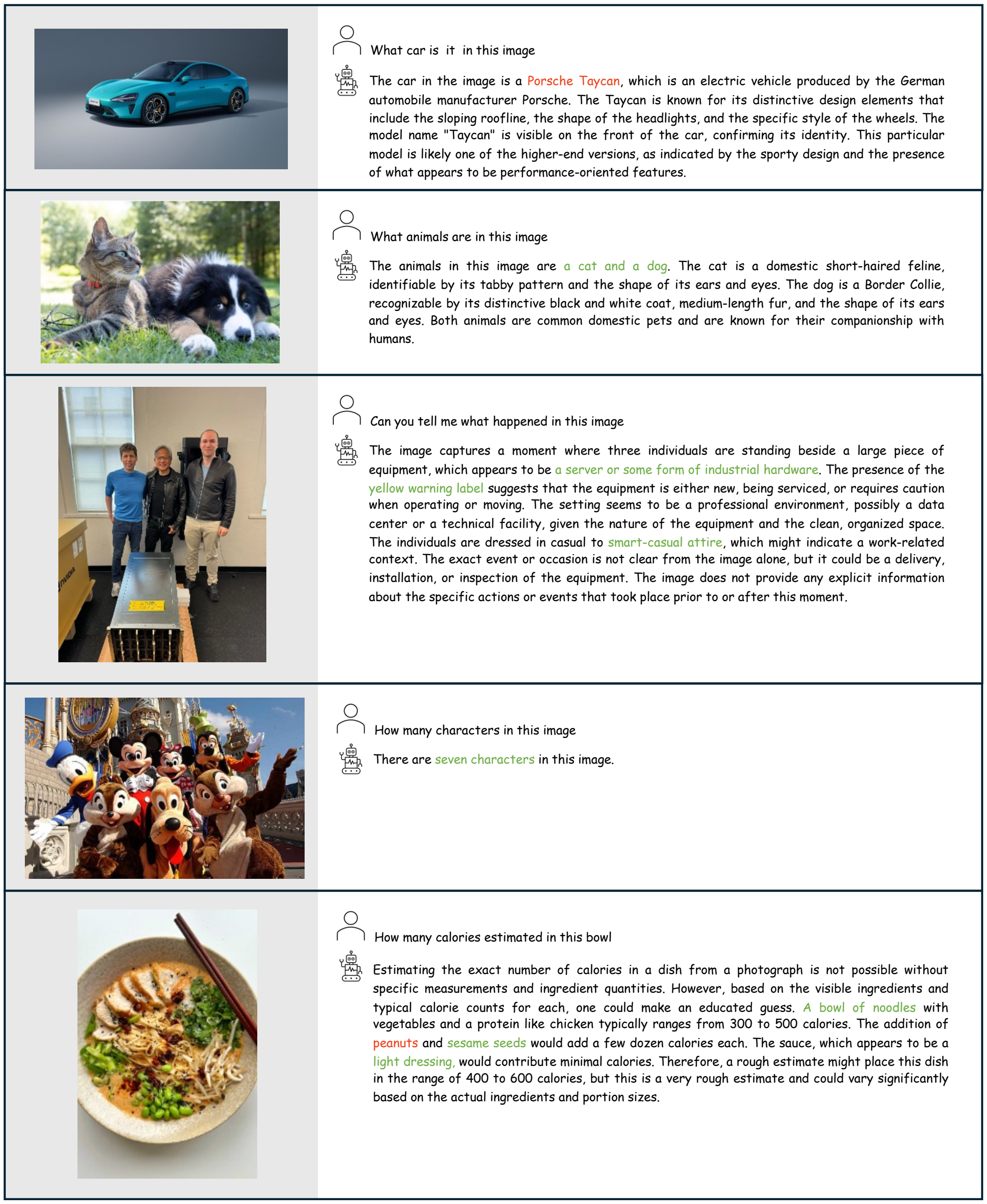
We include Table 7 to illustrate the evolution of parameters in the CuMo model throughout its construction process. The LLM constitutes a significant proportion of the total parameters, underscoring the potential for further scaling up the vision encoders to bolster the strength of multimodal LLMs.

E More Dialogues

We add more dialogues between the questions from the users and the response from CuMo-7B in Figure 6.

F Border Impact

The potential border impact of CuMo, as an open-sourced multimodal LLM, relies on two aspects: the development of artificial general intelligence (AGI) and the open-source community.



What car is it in this image

The car in the image is a Porsche Taycan, which is an electric vehicle produced by the German automobile manufacturer Porsche. The Taycan is known for its distinctive design elements that include the sloping roofline, the shape of the headlights, and the specific style of the wheels. The model name "Taycan" is visible on the front of the car, confirming its identity. This particular model is likely one of the higher-end versions, as indicated by the sporty design and the presence of what appears to be performance-oriented features.

What animals are in this image

The animals in this image are a cat and a dog. The cat is a domestic short-haired feline, identifiable by its tabby pattern and the shape of its ears and eyes. The dog is a Border Collie, recognizable by its distinctive black and white coat, medium-length fur, and the shape of its ears and eyes. Both animals are common domestic pets and are known for their companionship with humans.

Can you tell me what happened in this image

The image captures a moment where three individuals are standing beside a large piece of equipment, which appears to be a server or some form of industrial hardware. The presence of the yellow warning label suggests that the equipment is either new, being serviced, or requires caution when operating or moving. The setting seems to be a professional environment, possibly a data center or a technical facility, given the nature of the equipment and the clean, organized space. The individuals are dressed in casual to smart-casual attire, which might indicate a work-related context. The exact event or occasion is not clear from the image alone, but it could be a delivery, installation, or inspection of the equipment. The image does not provide any explicit information about the specific actions or events that took place prior to or after this moment.

How many characters in this image

There are seven characters in this image.

How many calories estimated in this bowl

Estimating the exact number of calories in a dish from a photograph is not possible without specific measurements and ingredient quantities. However, based on the visible ingredients and typical calorie counts for each, one could make an educated guess. A bowl of noodles with vegetables and a protein like chicken typically ranges from 300 to 500 calories. The addition of peanuts and sesame seeds would add a few dozen calories each. The sauce, which appears to be a light dressing, would contribute minimal calories. Therefore, a rough estimate might place this dish in the range of 400 to 600 calories, but this is a very rough estimate and could vary significantly based on the actual ingredients and portion sizes.

Figure 6: **More dialogues of CuMo-7B.** We highlight the correct answers and hallucinations from the responses of CuMo.

AGI: The development of AGI is a core topic, with the central component being the LLM. A multimodal LLM enhances its abilities to understand content such as images and videos, making it a more versatile agent and a positive contributor towards AGI.

Open-sourced MLLM: Open-sourced weights and code can accelerate the development of MLLMs. However, this also has potential negative impacts on society. Therefore, we have added a non-society license to constrain the usage of our model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize our contributions and scope in the abstract and introduction with the main experimental results to support our claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We add section 4.4 to discuss the limitations of our CuMo model.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper is application oriented and we conduct extensive experiments to validate our assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all experimental details including the datasets, hyperparameters, and training devices that we used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide all the open-sourced datasets and codes that we used for training and evaluation.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the training and inference details for reproducing and understanding our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For results with statistical variance like querying GPT API, we averaged the results by three times to reduce the variance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the details of the compute resources settings to reproduce our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly follow the code of ethics during the project.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We add the discussions of border impact in Appendix F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We provide usage guidelines for users to adhere when accessing to our model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We strictly follow the original licenses of existing assets including datasets and codes.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our models and codes are well documented for reproduction and protected by licenses.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not include crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not include crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.