MoVA: Adapting Mixture of Vision Experts to Multimodal Context

Zhuofan Zong[1,2,*] **Bingqi Ma**[2,*] **Dazhong Shen**[3] **Guanglu Song**[2]

Hao Shao[1] **Dongzhi Jiang**[1] **Hongsheng Li**[1,3,4,†] **Yu Liu**[2,†]

[1]CUHK MMLab   [2]SenseTime Research   [3]Shanghai AI Laboratory   [4]CPII under InnoHK

Abstract

As the key component in multimodal large language models (MLLMs), the ability of the visual encoder greatly affects MLLM's understanding on diverse image content. Although some large-scale pretrained vision encoders such as vision encoders in CLIP and DINOv2 have brought promising performance, we found that there is still no single vision encoder that can dominate various image content understanding, e.g., the CLIP vision encoder leads to outstanding results on general image understanding but poor performance on document or chart content. To alleviate the bias of CLIP vision encoder, we first delve into the inherent behavior of different pre-trained vision encoders and then propose the MoVA, a powerful and novel MLLM, adaptively routing and fusing task-specific vision experts with a coarse-to-fine mechanism. In the coarse-grained stage, we design a context-aware expert routing strategy to dynamically select the most suitable vision experts according to the user instruction, input image, and expertise of vision experts. This benefits from the powerful model function understanding ability of the large language model (LLM). In the fine-grained stage, we elaborately conduct the mixture-of-vision-expert adapter (MoV-Adapter) to extract and fuse task-specific knowledge from various experts. This coarse-to-fine paradigm effectively leverages representations from experts based on multimodal context and model expertise, further enhancing the generalization ability. We conduct extensive experiments to evaluate the effectiveness of the proposed approach. Without any bells and whistles, MoVA can achieve significant performance gains over current state-of-the-art methods in a wide range of challenging multimodal benchmarks. Codes and models are available at `https://github.com/TempleX98/MoVA`.

1 Introduction

Significant achievements in multimodal large language models (MLLMs) [1, 2, 3, 4, 5, 6, 7] have been witnessed due to their remarkable proficiency in solving open-world tasks. MLLMs acquire visual perception capacity while inheriting sophisticated reasoning abilities and knowledge from large language models (LLMs) [8, 9, 10]. The core idea behind MLLMs is projecting the vision representation into an LLM through a projector, facilitating a general-purpose multimodal understanding.

General multimodal understanding requires comprehending complex image contexts across various tasks and scenarios. The CLIP [11] vision encoder, pre-trained on large-scale image-text pairs with a contrastive loss, is widely considered as a flexible and popular choice among the latest leading MLLMs. However, training data and optimization target of the vision encoder determine its

---

*Equal contribution.

†Corresponding authors.

Table 1: **Comparison of CLIP *vs.* state-of-the-art task-specific vision encoders.** Our evaluation criteria encompass a variety of dimensions: comprehensive benchmarks [16], text-oriented Visual Question Answering (VQA) [17, 18], general VQA [19], object hallucination [20], Referring Expression Comprehension (REC) [21], Referring Expression Segmentation (RES) [21], and medical VQA benchmark SLAKE [22]. We use the same data for each model.

Vision Encoder	Task	MMB	DocVQA	ChartQA	GQA	POPE	REC	RES	SLAKE
CLIP [11]	Image-text Contrastive	**64.9**	35.6	35.3	62.5	85.7	81.5	43.3	63.7
DINOv2 [15]	Visual Grounding	57.5	14.7	15.9	**63.9**	86.7	**86.1**	47.5	59.4
Co-DETR [23]	Object Detection	48.4	14.2	14.8	58.6	**88.0**	82.1	48.6	55.3
SAM [24]	Image Segmentation	40.7	13.9	15.0	54.0	82.0	79.2	**49.3**	57.7
Pix2Struct [25]	Text Recognition	41.9	**57.3**	53.4	51.0	78.1	59.2	32.2	44.0
Deplot [26]	Chart Understanding	36.2	40.2	**55.8**	48.1	75.6	51.1	27.0	44.5
Vary [12]	Document Chart Parsing	28.1	47.8	41.8	42.6	69.1	21.6	16.0	40.9
BiomedCLIP [27]	Biomedical Contrastive	40.0	15.3	16.8	50.8	76.9	57.8	27.4	**65.1**
Plain fusion	-	63.4	46.5	48.9	63.0	86.4	85.7	45.3	64.7
MoVA	-	**65.9**	**59.0**	**56.8**	**64.1**	**88.5**	**86.4**	**49.8**	**66.3**

inconsistent performance across tasks and scenarios, which will bias the generalization of multimodal large language models. For instance, MLLMs with a single CLIP vision encoder usually perform poorly on fine-grained tasks such as grounding and optical character recognition (OCR) [12]. Several works have attempted to incorporate extra state-of-the-art vision encoder experts to cope with the challenge. For example, both SPHINX [13] and MoF [14] integrate vision self-supervised learning features of DINOv2 [15] with MLLMs to enhance their visual grounding capabilities. Vary [12] introduces a new vision encoder expert for improved fine-grained document and chart parsing ability. Intuitively, it is necessary to explore the utilization of more task-specific vision encoder experts in MLLMs to promote model generalization across various domains.

We aim to start the exploration through empirical analysis of readily available vision experts. In particular, we focus on the multimodal capabilities of seven distinct state-of-the-art vision encoders based on LLaVA-1.5-7B [28]. The results in Table 1 reveal that MLLMs with these task-specific vision encoders achieve optimal performance in their respective area. Concurrently, we note that the plain fusion (concatenation) of vision encoder experts adopted in previous works [13] would not bring consistent improvement compared with the single task-specific vision expert in its proficient task. The inherent bias of each expert introduces biased information and leads to performance degradation in the plain fusion paradigm. For example, DINOv2 serves as an expert in visual grounding but performs poorly at text-oriented tasks. Representation of DINOv2 would be regarded as biased information in text-related scenarios so incorporating DINOv2 for these tasks would inevitably cause performance decrease. Consequently, a flexible method of vision encoder ensemble that dynamically activates and weights context-relevant task-specific vision experts can fully unleash the capacity of these models while avoiding model bias.

In this paper, we propose MoVA, a powerful MLLM, adaptively routing and fusing task-specific vision experts with a coarse-to-fine mechanism. Inspired by the powerful tool-use capabilities of LLM [29], the coarse-grained context-aware expert routing aims to employ LLM to select vision experts with strong relevance to the user's image and instruction from the expert model pool. Thanks to the strong generalization ability of LLM, we also can perform model routing for vision experts in open scenarios. The fine-grained expert fusion facilitates better extraction and integration of expert representations based on multimodal context. Specifically, the expert knowledge extractor in the mixture-of-vision-expert adapter (MoV-Adapter) will extract diverse task-specific knowledge from various vision experts through mixture-of-expert (MoE) cross-attention layers. The dynamic gating network can allocate precise expert-wise soft weights for the integration of extracted task-specific knowledge. Under the coarse-to-fine paradigm, we provide a flexible and effective manner of leveraging representation from experts based on multimodal context and model expertise, further enhancing the model generalization ability. As presented in Table 1, MoVA can preserve the optimal performance of a single relevant vision encoder by ignoring non-relevant experts on the GQA, POPE, and REC task. Besides, MoVA can further boost performances via the fine-grained fusion of multiple relevant vision experts on other tasks.

We conduct comprehensive experiments on various benchmarks to evaluate the effectiveness of MoVA, including MLLM benchmarks, visual question answering (VQA), visual grounding, and biomedical understanding. Without any bells and whistles, MoVA can achieve significant performance gains over current state-of-the-art methods.

The **contributions** of this work are three-fold: **(i)** By analyzing the performance of individual vision encoders versus the plain fusion of multiple encoders across various tasks, we reveal that the inherent bias of each vision encoder can diminish its generalization ability across other irrelevant domains. **(ii)** We propose MoVA, a powerful MLLM composed of coarse-grained context-aware expert routing and fine-grained expert fusion with MoV-Adapter. Based on multimodal context and model expertise, MoVA fully leverages representation from multiple context-relevant vision experts flexibly while avoiding biased information of irrelevant experts. **(iii)** We demonstrate the effectiveness of each component in MoVA by elaborate ablation studies. MoVA can achieve significant performance gains over state-of-the-art methods in a wide range of challenging benchmarks.

2 Related Work

Multimodal architectures [1, 3, 6, 30, 31, 32, 33, 34], optimization paradigm [35, 36], applications [37, 38, 39, 40, 41], and benchmarks [42, 43, 44, 45, 46, 47, 48] have recently achieved remarkable progress and garnered unprecedented attention within the academic community. Multimodal large language models (MLLMs) usually leverage the alignment from visual features to the linguistic feature space to achieve superior vision-language understanding capabilities based on off-the-shelf LLMs and vision encoders. CLIP vision encoder [11], which is trained in contrastive learning from billions of diverse image-text pairs [49, 50], is widely used among these works. For example, LLaVA [3] adopts an MLP projector to align visual tokens from the frozen CLIP vision encoder to the embedding layer of LLM. However, The representation from CLIP exhibits strong discriminative abilities in classification and recognition but only has limited performance on downstream tasks like location and relation understanding [51]. To break through this bottleneck, some works [4, 52] turn to unlock the CLIP vision encoder and further fine-tune the parameter with training data for downstream tasks. For instance, Qwen-VL [6] collected massive training data for grounding and OCR to jointly optimize the CLIP vision encoder and LLM. Recent works propose to involve an extra frozen vision encoder to enhance the performance of MLLMs. SPHINX [13] is one of the pioneers, where grounding capabilities have been significantly improved with the assistance of the DINOv2 [15]. Vary [12] introduces an extra encoder training on large-scale charts and document data to improve the performance on related downstream tasks.

3 MoVA Methodology

3.1 Overview

MoVA comprises five key components: **(i)** a pre-trained large language model (LLM) that generates accurate responses given the image tokens and instructions; **(ii)** a base vision encoder; **(iii)** vision experts that generate task-specific vision latent features; **(iv)** mixture-of-vision-expert adapter (MoV-Adapter) that performs fine-grained expert fusion based on the multimodal context.

As illustrated in Figure 1, MoVA consists of two stages: coarse-grained context-ware expert routing and fine-grained expert fusion with MoV-Adapter. First, our coarse-grained context-ware expert routing leverages the tool-use capabilities of LLM, routing the most appropriate experts from N expert candidates via LLM to help the model answer the user's question. In the second stage, we turn to enhance the visual representation with a novel MoV-Adapter module in a fine-grained manner. More specifically, we leverage the mixture-of-expert (MoE) cross-attention layers to extract the task-specific knowledge of representations from chosen experts. Meanwhile, the dynamic gating network in MoV-Adapter can allocate soft weights to the extracted knowledge of each expert according to the input image and instruction. Then the extracted knowledge can be effectively integrated into the foundational representation of the base vision encoder. Finally, the enhanced visual representation with instruction tokens is fed to the LLM to generate an accurate response. In Section 3.2 and Section 3.3, we will focus on our core contributions, the context-aware expert routing strategy, and the expert fusion with MoV-Adapter. In Section 3.4, we will introduce the training process.
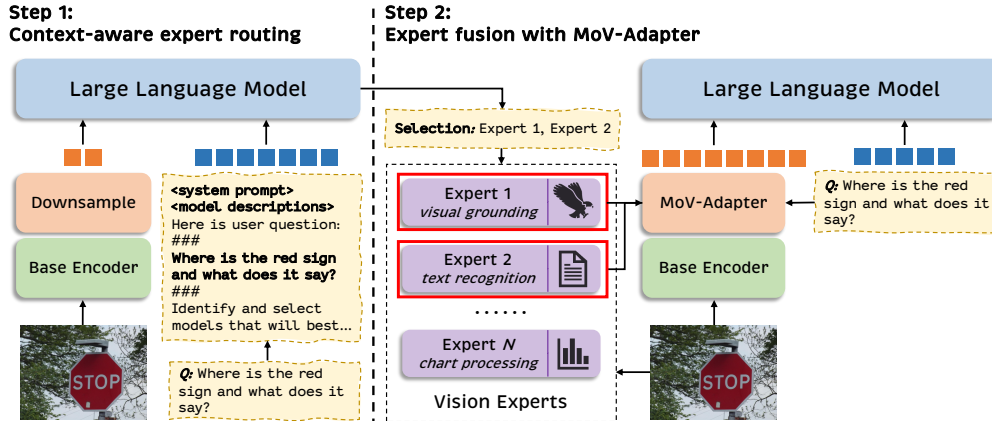
Figure 1: **The pipeline of MoVA.** MoVA performs coarse-to-fine routing to solve a given question. The coarse context-aware expert routing is performed in the first stage to select context-relevant experts. Next, we adopt the MoV-Adapter to extract and fuse the task-specific knowledge from these selected experts in a fine-grained manner.

Pretrained Vision Encoders and LLM. The vision encoders in MoVA consist of a base encoder and multiple task-specific vision encoder experts. We choose the pre-trained CLIP ViT-L-336px as the base encoder. Our vision experts include several state-of-the-art task-specific encoders: DINOv2, Co-DETR, SAM, Pix2Struct, Deplot, Vary, and BiomedCLIP. The corresponding expertise is presented in Table 1. For example, both Pix2Struct and Vary will be used when the user asks the MLLM to scan the document image. MoVA is flexible and easy to generalize to all decoder-only LLMs. We mainly consider Vicuna-7B [8], Llama3-8B [3], and Yi-34B [53] as our language models in this work.

3.2 Coarse-grained Context-aware Expert Routing

Pipeline of Context-aware Routing. The context-aware expert routing strategy aims to employ the impressive tool-use capacity of LLM to select vision experts with strong relevance to the user's image and instruction from a model pool. Specifically, we perform the context-aware expert routing in three steps during inference. First, the input image, user questions, and descriptions of expert models are converted into appropriate instructions that prompt the MLLM to perform expert selection. An example of the prompt instruction input and selection output is shown in Table 2. Such a routing task does not require image details and high-resolution input images, hence we directly downsample the base encoder's visual feature to obtain a coarse image embedding (*e.g.*, 144 image tokens). The downsampled image tokens and instruction tokens are then fed to the LLM as inputs. Finally, the LLM generates the output text and we parse it to determine which vision expert should be selected for fine-grained knowledge extraction in the second stage. For instance, as depicted in Table 2, the LLM directly outputs the option's letter of DINOv2 and Pix2Struct, thus we only utilize them for the subsequent extraction. During training, we do not perform context-aware expert routing and replace the routing outputs with our routing annotations to improve efficiency.

Routing Data Construction. Compared with other MLLMs, MoVA requires additional routing annotations. We first introduce the formal definition of the data structure for an unambiguous understanding of the routing data. The data structure for expert routing introduces additional routing annotation $\mathcal{R}$ to the conventional multimodal data $(\mathcal{I}, \mathcal{Q}, \mathcal{A})$. Here, $\mathcal{I}$ represents the image, $\mathcal{Q}$ and $\mathcal{A}$ refer to the question-answer pair, and $\mathcal{R}$ refers to the expert set which contains the most appropriate ones to solve this question. Then the construction process for routing data can be formulated as $(\mathcal{I}, \mathcal{Q}, \mathcal{A}) \rightarrow \mathcal{R}$, with the primary objective being to derive vision experts that optimally align with the sample $(\mathcal{I}, \mathcal{Q}, \mathcal{A})$. Intuitively, the language modeling loss can serve as an effective metric for evaluating how a data sample aligns with the vision expert. Specifically, we can reuse the LLaVA-1.5-7B models with various vision encoders presented in Section 1 to perform loss computation. Here, we denote the model with the base encoder as $\mathcal{M}_0$ and the model with j-th expert among N experts as

---

[3]https://github.com/meta-llama/llama3

Table 2: One example of the instruction-following data for context-aware expert routing. We present the multimodal inputs in the top block and the language response in the bottom block. The detailed model descriptions are released in the Appendix.

Routing Prompt Input

You are a helpful assistant router. Based on the visual content, questions, and model pool the user provides, you need to consider the expertise of these models to select the most 3 suitable models to help you answer the questions. Answer with the model's letter from the given choices directly. If no models are selected, just answer 'none'.

Model pool:
A. <DINOv2 model description>
B. <Co-DETR model description>
C. <SAM model description>
D. <Pix2Struct model description>
E. <Deplot model description>
F. <Vary model description>
G. <BiomedCLIP model description>
Question:
Where is the red sign and what does it say?

Routing Prompt Output

A, D

$\mathcal{M}_j$. For the i-th sample $(\mathcal{I}_i, \mathcal{Q}_i, \mathcal{A}_i)$, we send it to models $\{\mathcal{M}_j | j \in \{0, 1, \ldots, N\}\}$ and calculate the language modeling loss $\{\mathcal{L}_{i,j} | j \in \{0, 1, \ldots, N\}\}$. The j-th expert is regarded as a useful expert for the i-th sample only if $\mathcal{L}_{i,j} < \mathcal{L}_{i,0}$ and will be added to the routing set $\mathcal{R}_i$. Note that we only keep up to 3 vision experts to avoid computation costs brought by too many additional experts. All the routing annotations of our training data are generated offline. We can directly parse and input these offline results to the subsequent expert fusion component during training.

Routing Data Augmentation. To preserve the expert routing robustness and generalization ability in open scenarios, we only randomly select 2K samples for training, remove the model name in model description, and rewrite the model descriptions using ChatGPT [54] for each expert. We also shuffle the model pool and randomly truncate the model pool during training.

3.3 Fine-grained Expert Fusion with MoV-Adapter

We propose the MoV-Adapter to facilitate fine-grained expert representation extraction and integration based on multimodal context. As shown in Figure 2, the MoV-Adapter consists of L adapter blocks and a text encoder. Each block contains an expert knowledge extractor, a dynamic gating network, and a transformer block. For the i-th block, the input feature is denoted as $\mathbf{X}^i \in \mathbb{R}^{C \times H \times W}$ and we take the CLIP base encoder feature $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ as the input feature $\mathbf{X}^1$ of the first block. We use $\mathbf{G}$ to indicate the indices of chosen K experts. The expert feature set is $\{\mathcal{F}_j | j \in \mathbf{G}\}$. The final output feature of L adapter blocks is $\mathbf{X}^{L+1}$. Additionally, we apply two residual blocks [55] with an average pooling to $\mathbf{X}^{L+1}$ to obtain a coarser image feature $\mathbf{X}^{L+1}_{out} \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$, which is further connected to the LLM text embedding space by an MLP layer.

Text Encoder. We introduce a pre-trained BERT as the text encoder to extract language context information from the user's instruction. We take the [CLS] token from the output of the text encoder as the text token $\mathbf{X}_T \in \mathbb{R}^{C_T}$. It is worth noting that all the adapter blocks share the same text token.

Expert Knowledge Extractor. We adopt N cross-attention layers as the expert knowledge extractor to achieve efficient knowledge extraction. Note that only the expert features $\{\mathcal{F}_j | j \in \mathbf{G}\}$ and their corresponding cross-attention layers are involved in the extraction. For each selected expert feature $\mathcal{F}_j \in \mathbb{R}^{C_j \times H_j \times W_j}$, we first align its resolution to $\mathbf{X}^i$ with bilinear interpolation:

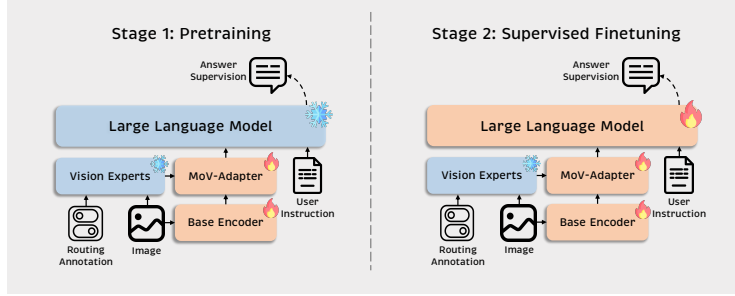$$\hat{\mathcal{F}}_j = \text{Interpolate}(\mathcal{F}_j, H, W). \tag{1}$$

Figure 3: **The training strategy of MoVA.** We enhance the task-specific knowledge extraction capacity in the first stage. Then, we excite model multimodal capacities in the last stage.

For the i-th MoV-Adapter block and the j-th cross-attention layer, we take input feature $\mathbf{X}^i$ as query, and the aligned expert feature $\hat{\mathcal{F}}_j$ as the key and value:

$$\mathbf{Y}_j^i = \mathbf{X}^i + \text{Attention}(\mathbf{X}^i, \hat{\mathcal{F}}_j). \tag{2}$$

Dynamic Gating Network. We employ a dynamic gating network to contribute to a fine-grained knowledge integration process for the conditional representation $\{\mathbf{Y}_j^i | j \in \mathbf{G}\}$. It is implemented with the softmax over the logits of an MLP layer, processing multimodal representation to generate expert-wise soft weight $\mathbf{P}^i \in \mathbb{R}^K$ for the output of each cross-attention layer in the extractor. Specifically, the input to the gating network is the concatenated vector of a visual token $\mathbf{X}_V^i \in \mathbb{R}^C$ and the text token $\mathbf{X}_T \in \mathbb{R}^{C_T}$. We obtain $\mathbf{X}_V^i$ with a global average pooling operation to $\mathbf{X}^i$. Then we concatenate them to compute the gating weights and the expert-wise outputs by computing the weighted sum:

$$\hat{\mathbf{X}}^i = \sum_{j \in \mathbf{G}} \mathbf{Y}_j^i \cdot \mathbf{P}_j^i, \tag{3}$$

Figure 2: MoV-Adapter architecture.

where $\mathbf{P}_j^i \in (0, 1)$ is the soft weight for the j-th expert in the i-th block.

Transformer Block. The transformer block in the adapter block follows the vanilla design, consisting of a self-attention layer and an FFN layer. Taking the fused visual representation $\hat{\mathbf{X}}^i$, its output will serve as the input feature $\mathbf{X}^{i+1}$ for the next adapter block.

3.4 Training Paradigm

As shown in Figure 3, the training process of MoVA consists of pretraining and supervised finetuning.

Pretraining. To improve multimodal generalization, we first construct 15M visual instruction samples across diverse domains as the training data: **(i)** Image caption data that covers 4M randomly selected samples from DataComp-1B [56], ShareGPT4V-PT [52], and ALLaVA-4V [57]. **(ii)** Visual grounding and localization dataset that encompasses Objects365 [58], RefCOCO [21], VisualGenome [59], PointQA [60], and Flickr30K [61]. **(iii)** Chart understanding data that includes MMC-Instruction [62], Chart2Text [63], DVQA [64], and SciGraphQA [65]. **(iv)** Text recognition and document parsing data that covers LLaVAR-PT [66] and 3M English document images from Common Crawl [4]. **(v)** LLaVA-Med [67] for biomedical image understanding. During the pretraining phase, we only optimize the MoV-Adapter along with the base vision encoder while preserving the capabilities of the initial large language model. Meanwhile, we leverage the routing annotations generated via the method proposed in Section 3.2 to choose experts and ignore representations from irrelevant ones during training.

Supervised Finetuning. We utilize high-quality visual instruction tuning data that build upon LLaVA-665K [28] for finetuning. Additionally, we integrate several visual question answering
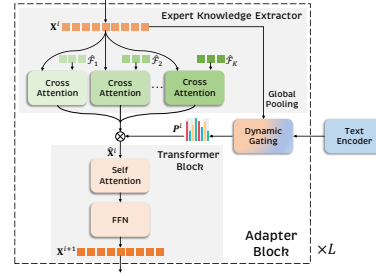
---

[4]https://commoncrawl.org

Table 3: **Performance comparison with current state-of-the-art frameworks on popular MLLM benchmarks.** PT and SFT indicate the number of multimodal training samples in pretraining and finetuning stage. #IMG means the number of image tokens processed by LLM.

Model	LLM	PT	SFT	#IMG	MME	MMB	MMB^{CN}	QBench	MathVista	MathVerse	POPE
					Proprietary MLLMs						
Qwen-VL-Plus [6]	–	–	–	–	–	66.2	68.0	66.0	43.3	11.8	–
Qwen-VL-Max [6]	–	–	–	–	–	77.6	75.1	73.6	51.0	24.8	–
Gemini-Pro [79]	–	–	–	–	–	73.6	74.3	68.2	45.2	22.3	–
GPT-4V [54]	–	–	–	–	–	75.8	73.9	74.5	49.9	38.3	–
					Open-source MLLMs						
Qwen-VL [6]	Qwen-7B	1.4B	50M	256	–	38.2	7.4	59.4	–	–	–
Qwen-VL-Chat [6]	Qwen-7B	1.4B	50M	256	1488/361	60.6	56.7	–	–	–	–
LLaVA-1.5 [28]	Vicuna-7B	558K	665K	576	1511/316	64.3	58.3	58.7	–	14.3	85.9
LLaVA-1.5 [28]	Vicuna-13B	558K	665K	576	1531/295	67.7	63.6	62.1	27.6	17.0	85.9
mPLUG-Owl2 [80]	LLaMA2-7B	348M	1.2M	64	1450/–	64.5	–	62.9	–	4.6	85.8
SPHINX-2k [13]	Vicuna-13B	115M	–	2880	1471/327	65.9	57.9	–	27.8	–	87.2
LLaVA-NeXT [81]	Vicuna-7B	558K	760K	2880	1519/332	67.4	60.6	–	34.6	–	86.5
LLaVA-NeXT [81]	Hermes-Yi-34B	558K	760K	2880	**1631**/397	79.3	79.0	–	**46.5**	23.4	87.7
MoVA	Vicuna-7B	15M	1.6M	576	1562/371	70.4	63.7	69.3	37.6	19.7	88.6
MoVA	Llama3-8B	15M	1.6M	576	1596/348	75.3	67.7	**70.8**	37.7	21.4	**89.3**
MoVA	Hermes-Yi-34B	15M	1.6M	576	1603/**455**	**81.3**	**79.0**	70.7	44.3	**23.7**	88.3

datasets across various domains, such as DocVQA [17], ChartQA [18], InfographicVQA [68], AI2D [69], ST-VQA [70], TextVQA [71], SynthDoG-en [72], Geometry3K [73], PGPS9K [74], Geo170K [75], RefCOCO, LLaVA-Med, VQA-RAD [76], and SLAKE [22]. We also encompass equivalent comprehensive captions [52, 57, 77, 78] generated by the advanced GPT4-V [54] for improved world knowledge. Apart from the above instruction tuning data, we convert the selected 2K routing annotations to instructions and incorporate them into the training data. In the supervised fine-tuning stage, only task-specific vision experts are frozen and we jointly optimize other components. The objective of supervised fine-tuning is to align the visual representation and the embedding of LLM, boosting its visual instruction-following capabilities.

4 Experiments

4.1 Implementation Details

As mentioned in Section 3.4, our training pipeline consists of two stages. In the pretraining stage, we use the AdamW optimizer with an initial learning rate of 2×10^{-4}, a batch size of 1024, and train the model for 1 epoch. We jointly finetune the weights of all components except additional vision experts with a batch size of 128 and an initial learning rate of 2×10^{-5} during supervised finetuning. We use 3 transformer blocks ($L=3$) in the MoV-Adapter and its hidden dimension is 1024, which is consistent with the base vision encoder CLIP. The input resolution of the base vision encoder is 672×672. Two residual blocks with an average pooling are employed in the MoV-Adapter to reduce the number of output image tokens from 2304 to 576. For the experiment performed in Table 1, we follow the default setting of LLaVA-1.5 but incorporate several additional datasets, including DocVQA [17], ChartQA [18], RefCOCO [21], LLaVA-Med [67], VQA-RAD [76], and SLAKE [22]. More details about vision experts, ablations, and analysis are available in Appendix A.1 and A.3.

4.2 MLLM Benchmarks

We empirically analyze the multimodal capacity and generalization ability of MoVA on a wide range of challenging MLLM benchmarks in Table 3. This comprehensive assessment is conducted on MME [82], MMBench [16], QBench [83], MathVista [84], MathVerse [85], and POPE [20]. Compared to other open-source MLLMs with similar model complexity, MoVA with Vicuna-7B achieves the best performance across 7 MLLM benchmarks while offering a more favorable balance between training efficiency and performance. For instance, MoVA-7B surpasses the recent state-of-the-art LLaVA-NeXT-7B [81] with a dynamic high resolution design, processing only 20% image tokens. Furthermore, we adopt Hermes-Yi-34B [86] as the LLM to validate the scaling property of MoVA. As depicted in Table 3, the performance of MoVA-34B is on par with popular proprietary

Table 4: **Performance comparison on VQA benchmarks**. We present the number of model parameters of each MLLM for a clear complexity comparison. * denotes zero-shot evaluation.

| Model | LLM | Params | General VQA | | | Text-oriented VQA | | | |
			VQA^{v2}	GQA	SQA^I	TextVQA	ChartQA	DocVQA	AI2D
Generalist models									
Qwen-VL [6]	Qwen-7B	10B	79.5	59.3	67.1*	63.8	65.7	65.1	62.3
Qwen-VL-Chat [6]	Qwen-7B	10B	78.2	57.5	68.2*	61.5	66.3	62.6	57.7
LLaVA-1.5 [28]	Vicuna-7B	7B	78.5	62.0	66.8*	58.2*	–	–	–
LLaVA-1.5 [28]	Vicuna-13B	7B	80.0	63.3	71.6*	61.3*	–	–	–
SPHINX-2k [13]	Vicuna-13B	16B	80.7	63.1	70.6*	61.2	–	–	65.1
Vary-base [12]	Qwen-7B	7B	–	–	–	–	65.3	76.3	–
CogAgent [87]	Vicuna-7B	18B	83.7	–	–	76.1	68.4	81.6	–
Specialist models									
Pix2Struct-Large [25]	–	1.3B	–	–	–	–	58.6	76.6	42.1
PALI-X-55B [88]	–	55B	**86.0**	–	–	71.4	70.9	80.0	81.2
MoVA	Vicuna-7B	10B	83.5	64.8	74.4*	76.4	68.3	81.3	74.9
MoVA	Llama3-8B	11B	83.5	65.2	74.7*	77.1	70.5	83.4	77.0
MoVA	Hermes-Yi-34B	38B	82.3	63.9	**79.0***	**77.8**	**73.8**	**84.2**	**83.0**

Table 5: **Performance comparison (Acc@0.5) on RefCOCO REC task.** Specialists are specifically designed for the grounding task or finetuned on RefCOCO data.

| Type | Model | RefCOCO | | | RefCOCO+ | | | RefCOCOg | |
		val	test-A	test-B	val	test-A	test-B	val	test
Generalist	Shikra-13B [5]	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16
	Ferret-13B [91]	89.48	92.41	84.36	82.81	88.14	75.17	85.83	86.34
	Qwen-VL [6]	89.36	92.26	85.34	83.12	88.25	77.21	85.58	85.48
	SPHINX-2k [13]	91.10	92.88	87.07	85.51	90.62	80.45	88.07	88.65
	MoVA-7B	92.55	94.50	88.81	87.70	92.05	82.94	89.28	89.70
	MoVA-8B	92.18	**94.75**	88.24	88.45	92.21	82.82	90.05	90.23
	MoVA-34B	**93.38**	94.66	**90.58**	**89.64**	**92.53**	**84.03**	**91.09**	**90.78**
Specialist	G-DINO-L [92]	90.56	93.19	88.24	82.75	88.95	75.92	86.13	87.02
	UNINEXT-H [93]	92.64	94.33	91.46	85.24	89.63	79.79	88.73	89.37

MLLMs (*e.g.*, Gemini-Pro [79]) and outperforms Qwen-VL-Plus [6] on 5 MLLM benchmarks. For example, MoVA establishes new records on MMBench and MMBench-CN, even surpassing the GPT-4V [54]. These results suggest that the ensemble of vision experts with adaptive expert routing can serve as an effective dimension for MLLM model scaling.

4.3 Visual Question Answering

The evaluation results on VQA benchmarks are outlined in Table 4. We divide these benchmarks into general VQA benchmarks [89, 19, 90] and text-oriented VQA benchmarks [71, 18, 17, 69]. Thanks to the dynamic and efficient task-specific knowledge extraction, MoVA achieves state-of-the-art performances across diverse VQA benchmarks. For general VQA benchmarks, MoVA-7B outperforms SPHINX-2k [4] equipped with Vicuna-13B on VQAv2 [89] and GQA by 4.2% and 1.9%, respectively. Besides, MoVA shows its proficiency in text recognition in various scenarios, encompassing scene text, chart, document, and diagram. For instance, MoVA-7B catches up to the current state-of-the-art generalist CogAgent [87] with 18 billion parameters on these text-oriented benchmarks with smaller model size. The MoVA model with 38B parameters even surpasses the well-established specialist model PALI-X-55B [88] by clear margins. These outstanding performances demonstrate MoVA's robust generalization capabilities across diverse domains.

4.4 Visual Grounding

We conduct experiments on Referring Expression Comprehension (REC) benchmarks [21] to evaluate the visual grounding ability of MoVA. The results are presented in Table5. The performance of

Table 6: Comparisons on the biomedical VQA datasets.

Model	VQA-RAD		SLAKE	
	Open	Close	Open	Close
LLaVA-Med	28.6	56.3	70.6	54.6
LLaVA-1.5	35.3	68.9	73.1	63.7
MoVA	**38.3**	**68.9**	**78.2**	**68.8**
LLaVA-Med (ft)	61.5	84.2	83.1	85.3

Table 7: Results of component-wise ablation studies.

Design	GQA	ChartQA	DocVQA
MoVA	**64.8**	**68.3**	**81.3**
Random routing	63.1	60.4	71.6
w/o routing	63.4	62.5	73.7
w/o MoV-Adapter	62.7	65.2	77.1

Table 8: Results of K varying from 1 to 3.

K	GQA	ChartQA
Dynamic	**64.8**	**68.3**
1	64.0	64.9
2	63.5	66.7
3	63.2	67.4

Table 9: Comparisons of expert routing criteria.

Design	#Models	POPE	GQA	ChartQA
Separate	4	88.6	**64.8**	68.3
Combination	14	88.9	64.6	68.7

Table 10: Open-world expert routing results.

Design	#Samples	Accuracy
MoVA	2K	**92.5%**
MoVA	5K	12.5%
w/o Augmentation	2K	0%
MLP classifier	2K	0%

Table 11: Performance of various MoV-Adapter variants.

Design	MME^P	MMB	POPE	GQA
MoVA	1562	**70.4**	**88.6**	**64.8**
2 blocks	1526	70.1	87.9	63.9
4 blocks	**1578**	69.4	88.3	64.5
Uniform gating	1521	69.1	87.5	64.1

MoVA-7B is on par with the state-of-the-art specialist models that are elaborately designed for grounding tasks. For example, MoVA-7B achieves a score of 90.22% on RefCOCO+ val, which is 2.46% higher than the score of UNINEXT-H [93]. Our largest model MoVA-34B further pushes the performance bound of visual grounding on these benchmarks. These impressive results demonstrate MoVA's remarkable visual grounding capacity.

4.5 Medical Visual Question Answering

This experiment is conducted on popular medical VQA benchmarks VQA-RAD and SLAKE. We directly leverage the medical VQA evaluation metric adopted by LLaVA-Med. Each sample of VQA-RAD and SLAKE is observed only once during the training process of MoVA and LLaVA-1.5. For a fair comparison, we compare MoVA with the LLaVA-Med variant that is finetuned with only 1 epoch on the benchmark. The performance of the LLaVA-Med specialist that is fully finetuned on downstream tasks is also reported. As presented in Table 6, MoVA-7B consistently yields higher scores than LLaVA-Med and LLaVA-1.5, exhibiting its medical visual chat ability.

4.6 Ablation Study

Component-wise analysis. As presented in Table 7, we perform an ablation to thoroughly delve into the effect of each component. First, we try to replace the context-aware routing with random routing. Without task-relevant vision experts, the performance drops by a large margin, especially on the text-oriented VQA benchmarks. Removing context-aware routing to leverage all vision experts also leads to similar results. It proves that both these modifications introduce biased information from irrelevant vision experts due to the removal of context-aware routing. Then, we ablate the effectiveness of the MoV-Adapter by replacing it with simple linear layers. The removal of fine-grained expert feature fusion downgrades performance across all datasets. These results delineate that each component in MoVA can consistently yield significant gains.

Number of activated experts. In the context-aware routing phase, the number of activated experts K is dynamic. We compare such a data-dependent design with other variations of constant K in this experiment. As presented in Table 8, the overall performance of dynamic K consistently outperforms other models with constant K. This reveals this dynamic implementation can fully exploit the task-specific knowledge of relevant experts while avoiding the incorporation of biased information.

Criteria for choosing better experts. To reduce the costs, our method only adopts $\binom{N}{1}$ models with N various encoders to identify the better vision expert *separately*. However, we do not explicitly consider the combination of the chosen vision experts. In this experiment, we compare our method with another strategy that considers vision encoders combination and enumerates $\sum_{i=1}^{3} \binom{N}{i}$ models for routing data construction. Specifically, we set $N = 4$ and employ a smaller model pool of DINOv2, Co-DETR, Pix2Struct, and Deplot to reduce training costs. As shown in Table 9, our method achieves comparable performance while requiring much less models for data construction.

9

Expert routing in open scenarios. We develop 105 human-verified testing samples that should be answered using novel experts for the expert routing task. These novel experts encompass 7 vision models [94, 72, 95, 92, 96, 55, 97] on various computer vision tasks and each expert corresponds to 15 evaluation samples. We manually check the correctness of the expert routing result. As presented in Table 10, a lightweight network, such as a MLP classifier fails to generalize to this open-world setting. Besides, increasing the routing training samples and removing data augmentations also lead to severe performance degradation. The results demonstrate our coarse-grained context-aware routing preserves the generalization ability for expert routing in open scenarios.

Adapter Design. In this section, we conduct ablation studies on the design of the MoV-Adapter. As presented in Table 11, we compared the impact of using 2, 3, and 4 adapter blocks on the model's performance. We observed that the baseline with 3 blocks can achieve better performance than other settings. Then, we substituted our multimodal gating for uniform gating to investigate its effectiveness. Each of the experts is assigned the same soft weight in the uniform gating. We find uniform gating brings consistent performance drops in the test benchmarks. It indicates that the lack of the dynamic soft-weight harms the overall performance since it fails to perform precise knowledge extraction.

Inference Analysis. As illustrated in Figure 1, MoVA consists of two stages: coarse-grained context-ware expert routing and fine-grained expert fusion with MoV-Adapter. This two-stage inference pipeline can be further broken down into five steps: **(i)** Data preprocessing. We first process the input image with image processors and convert the input text into a token sequence with the LLM tokenizer. **(ii)** Base encoder forward. We extract the base image feature using the base CLIP encoder. Note that we only run the base encoder once since its output feature can be preserved and reused in the fourth step. **(iii)** LLM routing generation. We compress the base image features into 144 image tokens. The LLM generates a concise routing answer based on the compressed image feature and routing instruction. Vision experts and MoV-Adapter forward. **(iv)** According to the multimodal context and routing results generated in the previous step, we fuse vision features of the base encoder and activated experts in a coarse-to-fine manner. **(v)** LLM response generation. The LLM generates the final response given the fused vision features and user instructions. To investigate the inference efficiency of each step, we randomly select 200 images from the COCO val2017 dataset and adopt the common image caption instruction: *Describe this image.* The temperature for generation is 0. The latency is measured using bfloat16 and flash-attention 2 on an A100 80G GPU. We present the average inference latency of each step and show the average sequence length of the routing output and final response. The inference latencies for each step are 0.19s, 0.05s, 0.14s, 0.07s, and 10.24s, respectively. The average length of the routing output is 3.24 tokens, while the average length of the final response is 405.06 tokens. Compared to the LLM response generation (Step 5), the LLM expert routing (Step 3) generates much fewer output tokens and its latency is negligible (0.14s *v.s.* 10.24s). Therefore, our method does not bring significant inference costs.

5   Conclusion

In this paper, we reveal that the inherent bias of each vision encoder can diminish its generalization ability across other irrelevant domains by analyzing the performance of individual vision encoders versus the plain fusion of multiple encoders across various tasks. To deal with the problem, we propose MoVA, a powerful MLLM composed of coarse-grained context-aware expert routing and fine-grained expert fusion with MoV-Adapter. Based on multimodal context and model expertise, MoVA fully leverages representation from multiple context-relevant vision encoder experts flexibly and effectively while avoiding biased information brought by irrelevant experts. MoVA can achieve significant performance gains over current state-of-the-art methods in a wide range of benchmarks.

Limitations. We acknowledge some limitations in our paper that require attention. One limitation is the hallucination, which refers to the generation of text that appears plausible or coherent but is factually incorrect and misleading. This issue potentially presents in all powerful MLLMs. Additionally, the performance may be affected by failure cases of the context-relevant vision experts, leading to potential degradation. We plan to explore solutions for these limitations in future works.

Acknowledgments and Disclosure of Funding

References

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

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

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

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

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

[6] 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.

[7] Wenliang Dai, Junnan Li, Dongxu Li, AnthonyMeng Huat, 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.

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[9] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[10] Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, et al. Map-neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv:2405.19327*, 2024.

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

[12] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023.

[13] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

[14] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024.

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

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

[18] 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 preprint arXiv:2203.10244*, 2022.

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

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

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

[22] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[23] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023.

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[25] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023.

[26] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*, 2022.

[27] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[29] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

[30] Hao Shao, Yuxuan Hu, Letian Wang, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. *arXiv preprint arXiv:2312.07488*, 2023.

[31] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

[32] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Lumen: Unleashing versatile vision-centric capabilities of large multimodal models. *arXiv preprint arXiv:2403.07304*, 2024.

[33] Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models, 2024.

[34] Hongcheng Guo, Jian Yang, Jiaheng Liu, Liqun Yang, Linzheng Chai, Jiaqi Bai, Junran Peng, Xiaorong Hu, Chao Chen, Dongfeng Zhang, et al. Owl: A large language model for it operations. *arXiv preprint arXiv:2309.09298*, 2023.

[35] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

[36] Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024.

[37] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *arXiv preprint arXiv:2404.03653*, 2024.

[38] Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. *arXiv preprint arXiv:2412.09618*, 2024.

[39] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. *arXiv preprint arXiv:2407.05131*, 2024.

[40] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024.

[41] Zaiquan Yang, Yuhao Liu, Jiaying Lin, Gerhard Hancke, and Rynson WH Lau. Boosting weakly-supervised referring image segmentation via progressive comprehension. *arXiv preprint arXiv:2410.01544*, 2024.

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

[43] Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, and Yu-Gang Jiang. Eyes can deceive: Benchmarking counterfactual reasoning abilities of multi-modal large language models. *arXiv preprint arXiv:2404.12966*, 2024.

[44] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024.

[45] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, et al. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024.

[46] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*, 2024.

[47] Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*, 2024.

[48] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023.

[49] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

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

[51] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[52] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[53] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

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

[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[56] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[57] 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 preprint arXiv:2402.11684*, 2024.

[58] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.

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

[60] Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. Point and ask: Incorporating pointing into visual question answering. *arXiv preprint arXiv:2011.13681*, 2020.

[61] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[62] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023.

[63] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, 2022.

[64] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.

[65] Shengzhi Li and Nima Tajbakhsh. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*, 2023.

[66] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

[67] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[68] 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.

[69] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.

[70] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019.

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

[72] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 7:15, 2021.

[73] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

[74] Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. A multi-modal neural geometric solver with textual clauses parsed from diagram. *arXiv preprint arXiv:2302.11097*, 2023.

[75] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

[76] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[77] LAION. Gpt-4v dataset. `https://huggingface.co/datasets/laion/gpt4v-dataset`, 2023.

[78] Carter Jimmy. Textocr-gpt4v. `https://huggingface.co/datasets/jimmycarter/textocr-gpt4v`, 2024.

[79] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

[81] 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.

[82] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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

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

[85] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024.

[86] 01-AI. Yi. `https://huggingface.co/01-ai`, 2023.

[87] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023.

[88] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023.

[89] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

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

[91] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

[92] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[93] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023.

[94] Zhuofan Zong, Dongzhi Jiang, Guanglu Song, Zeyue Xue, Jingyong Su, Hongsheng Li, and Yu Liu. Temporal enhanced training of multi-view 3d object detector via historical object prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3781–3790, 2023.

[95] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[96] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.

[97] Zhuofan Zong, Kunchang Li, Guanglu Song, Yali Wang, Yu Qiao, Biao Leng, and Yu Liu. Self-slimmed vision transformer. In *European Conference on Computer Vision*, pages 432–448. Springer, 2022.

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

[99] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

A Appendix / supplemental material

Table 12: **Vision expert model configurations of vision experts in MoVA.** Methods with * use a convolution layer to compress the output feature.

Model	Params	Resolution	Width	Depth	Output shape
DINOv2-giant [15]	1.1B	518×518	1536	40	1536×37×37
Co-DETR-large* [23]	304M	1280×1280	1024	24	256×80×80
SAM-huge* [24]	632M	1024×1024	1280	32	256×64×64
Pix2Struct-large [25]	513M	720×720	1536	18	1536×45×45
Deplot-base [26]	92M	720×720	768	12	768×45×45
Vary-base* [15]	86M	1024×1024	768	12	512×32×32
BiomedCLIP-base [15]	86M	224×224	768	12	768×16×16

Table 13: **Introduction of datasets used in the MoV-Adapter pretraining stage.** The <class> placeholder represents the object category in the object detection task. The <expr> placeholder represents the expression in the REC task. The placeholder denotes the bounding box coordinates. The <point> placeholder denotes the coordinate of a point. We directly use the original question as the instruction for MMC-Instruction and ScigraphQA.

Task	Dataset	Task template
Image Caption	Datacomp [56]	Please describe this image. Provide a one-sentence caption for the provided image.
	ShareGPT4V-PT [52]	Can you elaborate on the elements of the picture provided? Write a detailed description of the given image.
	ALLaVA-4V [57]	Can you elaborate on the elements of the picture provided? Write a detailed description of the given image.
Grounding and Localization	Objects365 [58]	Detect all objects among <class> in the image. Perform object detection given the image within <class>.
	RefCOCO [21]	Locate the region this sentence describes: <expr>. Please provide the bounding box coordinates. Please generate a short and spotlighted mention of the part seen in the photo.
	Visual Genome [59]	Locate the region this sentence describes: <expr>. Please provide the bounding box coordinates. Please generate a short and spotlighted mention of the part seen in the photo.
	PointQA [60]	How many of these objects in picture? How many of these objects <point> in picture?
	Flickr30K [61]	Take a look at the image and give me the location details for any mentioned items. Unravel the aspects of the image and give the bounding box for the mentioned items.
Chart Understanding	MMC-Instruction [62]	-
	Chart2Text [63]	What significant details and conclusions can be drawn from this chart? Can you extract the data points in this image?
	ScigraphQA [65]	-
Document Parsing	LLaVAR-PT [66]	Report on any text that can be clearly read in the image. Identify any text visible in the image provided.
	English Documents	Extract every piece of text from this image. I request you to apply optical character recognition to this image.
Biomedical Understanding	LLaVA-Med [67]	Write a terse but informative summary of the picture. Share a comprehensive rundown of the presented image.

A.1 Vision Experts

Model Configuration. We present the detailed model configurations of our task-specific vision experts in 12. We adopt the official checkpoint weights that are publicly available.

Model Description. The model descriptions used in the routing prompt are released in Table 17.

A.2 Training Data Details

The training process of MoVA consists of two stages. In the Appendix, we present the training datasets with corresponding task templates of the first stage in Table 13. For the training data of the second stage, we follow the prompt format of LLaVA-1.5 [28]. The MoVA models with Vicuna-7B and LLama3-8B are pretrained using 64 A100 80G GPUs for 2 days, and finetuned using 32 A100 80G GPUs for 1 day. The MoVA with 34B LLM is pretrained using 128 A100 80G GPUs for 5 days and finetuned using 64 A100 80G GPUs for 2 days.

Table 14: Performance of various routing component.

Design	#Samples	MME	MMB
LLM	2K	**1562/371**	**70.4**
BERT	**1.6M**	1520/326	68.8
MLP	**1.6M**	1483/305	68.1

Table 15: Results of various vision encoder combination for routing.

Design	MMB	GQA	DocVQA
CLIP	**70.4**	64.8	**81.3**
+DINOv2	70.1	**65.1**	80.5
+DINOv2+Pix2Struct	69.5	64.4	80.9

Table 16: Effects of routing image tokens.

#IMG	MMB	ChartQA
144	70.4	68.3
256	69.8	68.4
576	**70.6**	**68.7**

A.3 More Experiments

Image segmentation. In this experiment, we aim to investigate if task-specific knowledge can improve MoVA on the segmentation task. Therefore, we introduce a simple design to extend MoVA to segmentation tasks. Unlike segmentation generalists [98] that adopt an additional pixel decoder with high-resolution images for high-quality mask generation, we just formulate the referring segmentation task as sequential polygon generation [99]. We finetune MoVA and the baseline with a SAM-Huge [24] backbone on the RefCOCO referring segmentation datasets. MoVA achieves 57.1% gIoU on the testA benchmark, which is 2.6% higher than the 54.5% of baseline. This result indicates that MoVA is capable of exploiting task-specific knowledge to solve segmentation tasks.

Effect of LLM for expert routing. In this experiment, we investigate the effect of the LLM in our coarse-grained expert routing. As presented in Table 14, expert routing with LLM achieves the best performance. When the LLM is substituted for a lightweight MLP classifier and a BERT encoder, we need to increase the number of routing training samples from 2K to 1.6M to preserve model performance. Besides, both MLP classifier and BERT encoder fail to perform expert routing in open scenarios as stated in Table 10. Therefore, the strong tool-use capacity and generalization ability of LLM is critical to our flexible and effective expert routing.

Vision encoder for expert routing. In the coarse-grained expert routing, we only adopt the image feature of base vision encoder CLIP. As presented in Table 15, the method with CLIP achieves slightly better performance than other methods since such a routing task does not require elaborate expert knowledge. Besides, incorporating other vision experts with plain fusion also brings biased information and increases cost. To achieve a better trade-off between efficiency and performance, we only use CLIP for coarse-grained expert routing.

Number of image tokens in expert routing. We ablate the number of image tokens used in the expert routing stage. As shown in Table 16, routing with 144 tokens can achieve comparable performance to methods with more tokens. Considering the additional cost brought by processing more image tokens, we only use 144 tokens for routing.

Data construction. In this experiment, we analyze the effectiveness of our routing data construction method. The test split is constructed by randomly selecting 500 samples from the routing data. We first prompt GPT4-V [54] to determine whether the expert routing result is reasonable or not. These evaluation results are then manually checked and we finally compute the routing accuracy using the human-verified evaluation results. Our loss-driven data construction approach achieves 94.6% accuracy (473 samples are regarded as "reasonable" by GPT4-V), demonstrating its effectiveness.

A.4 Qualitative Results

We apply MoVA to a wide range of real-world understanding and reasoning tasks to investigate its multimodal comprehension capacity. As presented in Figure 4 and Figure 5, MoVA successfully solves these diverse and complex cases thanks to the context-relevant expert knowledge.

A.5 Potential Societal Impacts

Positive impacts. The positive impact is to use MoVA to solve various open-world tasks in real-world scenarios. For example, MoVA can serve as a useful assistant to help people improve work efficiency and answer their questions.

Negative impacts. The potential negative social impact is to use MoVA model to generate misleading and false contents. This issue potentially presents in all multimodal large language models. We will try to improve its safety in future works.

Multimodal Commonsense

User:
What is the person with the highest detection confidence score doing?

MoVA:
The person with the highest detection confidence score, **labeled "person 0.89", is eating ice cream**. She is holding a **white container**, and she appears to be **taking a bite of the ice cream**. She is wearing a light pink shirt with **a red logo on the left chest area**, black shorts with a white stripe on the side, and white sneakers with black accents. Her hair is styled in a bob cut, and she is wearing **a pair of white earphones around her neck**. The person is standing with one hand slightly raised, possibly to hold the container steady while eating. The image captures her in a candid moment, enjoying her snack while walking or standing outdoors.

Scene Text Recognition

User:
What are all the scene text in the image?

MoVA:
The text on the road sign reads:
– **"CLEMENCEAU AVE"** at the **top** of the sign.
– **"River Valley Rd"** with an arrow **pointing right**.
– **"Havelock Rd"** with an arrow **pointing left**.
– **"Chin Swee Rd"** with an arrow **pointing left**.

Joke and Meme

User:
Can you explain this meme?

MoVA:
The image is a humorous meme that compares **the experiences of different parties involved in online learning during the COVID-19 pandemic**. The meme is divided into four panels, each representing a different entity: **Zoom, the Wifi company, the teacher, and the students.**
In the first panel, a man is shown throwing money into the air, **symbolizing the financial success of Zoom, a popular video conferencing platform that has seen a surge in usage during the pandemic**. The text "Zoom" is written above the image.
The second panel shows a man lying on a couch, holding a laptop, with a relaxed and content expression on his face. This represents **the teachers who have been able to work from home and enjoy a more flexible lifestyle**. The text "Teacher" is written above the image.
In the third panel, a man is shown throwing money into the air, similar to the first panel, but this time the text "Wifi company" is written above the image. This suggests that **the companies providing internet services have also benefited from the increased demand for online services during the pandemic**.
The fourth and final panel shows a clown with a sad expression, standing in front of a mirror with the words **"PUT ON A HAPPY FACE"** written on it. This represents **the students who have had to adapt to online learning and may be feeling overwhelmed or stressed**. The text "Students" is written above the image.
Overall, **the meme humorously highlights the different experiences and impacts of online learning during the pandemic**, with Zoom and the Wifi company profiting, teachers enjoying a more relaxed lifestyle, and students feeling the pressure to adapt to a new learning environment.

UI Notification Understanding

User:
Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.
Below is what you see on the computer screen. **Describe the image and predict your next move**. If the action involves moving the mouse, describe the location and the icon as detailed as possible.

MoVA:
This is a screenshot of a computer desktop with various icons and a window open. The desktop displays an image of a bridge over a body of water, possibly a sea or ocean. The bridge is a long, straight structure connecting two land masses. The water is a deep blue, and the land is rocky with some greenery. **The window also shows some text at the bottom that reads "Kevin Lin is calling you,"** suggesting that someone named Kevin Lin is trying to reach the user. The desktop has several icons, including one for the browser, a folder, and some other applications. The taskbar at the bottom of the screen shows **the time as 11:18 AM and the date as 7/21/2023**. There are also icons for various applications, including a search bar and a folder.

As a robot, I would not have personal preferences or needs, so my next move would depend on the instructions given to me. If I were asked to open a new tab in the browser, I would click on the browser icon in the taskbar. **If I were asked to answer the call from Kevin Lin, I would click on the phone icon in the notification area of the taskbar**. If I were asked to open a specific application, I would click on the corresponding icon on the desktop. If I were asked to search for something, I would click on the search bar icon in the taskbar.

Figure 4: Qualitative multimodal understanding results of MoVA.

Visual Referring Prompting

User:
What does this part of the poster highlights?

MoVA:
This part of the poster highlights the "ReCo: Region-Controlled Text-to-Image Generation" framework, which is a system designed to generate images from text descriptions. The poster outlines the framework's components and benefits, such as the ability to generate images with precise region control, improved semantic correctness, and better image quality. It also emphasizes the flexibility of the box tokens and the ease of use, suggesting that the system can create images that are both visually appealing and semantically accurate.

User:
In the DocVQA task, how many points is the model with the highest score higher than the model with the lowest score?

MoVA:
The model with the highest score in the DocVQA task is "Pix2Struct large" with a score of 76.6, while the model with the lowest score is "Dessurt" with a score of 63.2. The difference in scores is 76.6 − 63.2 = 13.4 points.

Table Understanding and Reasoning

Model	Train Param	TS FT	Doc VQA	Info VQA	Deep Form	KLC	WTQ	TabFact	ChartQA	TextVQA	TextCaps	Visual MRC
Dessurt	127M	✓	63.2	-	-	-	-	-	-	-	-	-
Donut	176M	✓	67.5	11.6	61.6	30.0	18.8	54.6	41.8	43.5	74.4	93.91
Pix2Struct_base	282M	✓	72.1	38.2	-	-	-	-	56.0	-	88.0	-
Pix2Struct_large	1.3B	✓	76.6	40.0	-	-	-	-	58.6	-	95.5	-
UReader	86M	✗	65.4	42.2	49.5	32.8	29.4	67.6	59.3	57.6	118.4	221.7

Flow Chart Understanding

User:
Can you translate the flowchart to a C++ code?

MoVA:
Sure, here's a possible C++ code that implements the logic of the flowchart:
```
#include <iostream>

int main() {
    int num1, num2;
    std::cout << "Input num1: ";
    std::cin >> num1;
    std::cout << "Input num2: ";
    std::cin >> num2;

    if (num1 > num2) {
        std::cout << "The larger number is: " << num1 << std::endl;
    } else {
        std::cout << "The larger number is: " << num2 << std::endl;
    }

    return 0;
}
```
This code prompts the user to input two numbers, num1 and num2, and then compares them using an if-else statement. If num1 is greater than num2, the code will display num1; otherwise, it will display num2.

Dense Text Recognition

Abstract

As the key component in multimodal large language models (MLLMs), the ability of the visual encoder greatly affects MLLM's understanding on diverse image content. Although some large-scale pretrained vision encoders such as vision encoders in CLIP and DINOv2 have brought promising performance, we found that there is still no single vision encoder that can dominate various image content understanding, e.g., the CLIP vision encoder leads to outstanding results on general image understanding but poor performance on document or chart content. To alleviate the bias of CLIP vision encoder, we first delve into the inherent behavior of different pre-trained vision encoders and then propose the MoVA, a powerful and novel MLLM, adaptively routing and fusing task-specific vision experts with a coarse-to-fine mechanism. In the coarse-grained stage, we design a context-aware expert routing strategy to dynamically select the most suitable vision experts according to the user instruction, input image, and expertise of vision experts. This benefits from the powerful model function understanding ability of the large language model (LLM). In the fine-grained stage, we elaborately conduct the mixture-of-vision-expert adapter (MoV-Adapter) to extract and fuse task-specific knowledge from various experts. This coarse-to-fine paradigm effectively leverages representations from experts based on multimodal context and model expertise, further enhancing the generalization ability. We conduct extensive experiments to evaluate the effectiveness of the proposed approach. Without any bells and whistles, MoVA can achieve significant performance gains over current state-of-the-art methods in a wide range of challenging multimodal benchmarks.

User:
Please recognize the text in the image.

MoVA:
Abstract
As the key component in multimodal large language models (MLLMs), the ability of the visual encoder greatly affects MLLMs' understanding on diverse image content. Although some large-scale pretrained vision encoders such as vision encoders in CLIP and DINOv2 have brought promising performance, we found that there is still no single vision encoder that can dominate various image content understanding, e.g., the CLIP vision encoder leads to outstanding results on general image understanding but poor performance on document or chart content. To alleviate the bias of CLIP vision encoder, we first delve into the inherent behavior of different pre-trained vision encoders and then propose the MoVA, a powerful and novel MLLM, adaptively routing and fusing task-specific vision experts with a coarse-to-fine mechanism. In the coarse-grained stage, we design a context-aware expert routing strategy to dynamically select the most suitable vision experts according to the user instruction, input image, and expertise of vision experts. This benefits from the powerful model function understanding ability of the large language model (LLM). In the fine-grained stage, we elaborately conduct the mixture-of-vision-expert adapter (MoV-Adapter) to extract and fuse task-specific knowledge from various experts. This coarse-to-fine paradigm effectively leverages representations from experts based on multimodal context and model expertise, further enhancing the generalization ability. We conduct extensive experiments to evaluate the effectiveness of the proposed approach. Without bells and whistles, MoVA can achieve significant performance gains over current state-of-the-art methods in a wide range of challenging multimodal benchmarks.

Figure 5: Qualitative multimodal understanding results of MoVA.

Table 17: **Model descriptions used in the context-aware expert routing.** We describe the pros and cons of each expert model in the routing prompt. We only present 3 captions for each expert here.

DINOv2 description
(1) This model demonstrates exceptional prediction capabilities across a range of image-related tasks, including image classification, object detection, segmentation, and image retrieval. The model leverages advanced self-supervised learning techniques to achieve high performance without relying heavily on labeled data.
(2) This model shows very strong prediction capabilities on tasks such as image classification, detection, segmentation, and image retrieval. However, it encounters challenges in accurately reading text within images.
(3) This model can effectively extract the accurate spatial and semantic information from natural images.

Co-DETR description
(1) This model is a state-of-the-art object detector pretrained on natural images. It can enable models to solve object-centric problems. Nonetheless, this model struggles with processing background elements in natural scenes.
(2) This model is a cutting-edge object detection model that can accurately detect objects in images. However, it struggles with identifying text in images.
(3) This model is a state-of-the-art object detector that can identify objects in images.

SAM description
(1) This model is an image segmentation model. This model can segment the precise location of either specific objects in an image or every object in an image.
(2) This model is a leading image segmentation framework and achieves strong zero-shot segmentation performance.
(3) This model is a promotable segmentation system with zero-shot generalization to unfamiliar objects and images.

Pix2Struct description
(1) This model excels in text recognition, achieving state-of-the-art text analysis results across distinct domains: documents, illustrations, user interfaces, natural images containing text, and images of charts.
(2) This model demonstrates exceptional proficiency in text recognition, delivering cutting-edge text analysis performance across various domains.
(3) This model can automate the extraction of information from scanned documents, making it easier to digitize and manage large volumes of paperwork.

Deplot description
(1) This model is a specialized model designed to achieve state-of-the-art plot and chart understanding performance.
(2) This model is a fine-tuned version of an existing text recognition model. It has been specifically trained to achieve superior performance in plot and chart understanding tasks.
(3) This model can help detect the text within the input document, diagram, and chart images.

Vary description
(1) This model can achieve more fine-grained vision perception for images with text, such as document-level Chinese/English OCR, book image to markdown or LATEX, Chinese/English chart understanding.
(2) This model can handle images with text effectively and accurately, enabling advanced tasks such as document OCR and chart understanding.
(3) This model can accurately process images with text, enabling tasks such as OCR. However, it cannot process natural images without text.

BiomedCLIP description
(1) This model is a foundation model designed for biomedical vision-language processing.
(2) This model is capable of biomedical images, such as chest X-ray and radiology images.
(3) This model is a state-of-the-art biomedical vision-language model. It has been shown to achieve significant improvements in biomedical image-text tasks.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

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 clarify our main claims and contribution in the abstract and introduction section.

   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 discuss the limitations of MoVA in the conclusion section.

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: The paper does not include theoretical results.

   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 fully release the implementation details and experiment settings in the Section Experiments and Section Appendix.

   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: [No]

   Justification: The codes and models will be released when the paper is accepted.

   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 present the experiment details in Section Experiments and Section Appendix.

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: [No]

Justification: We conduct experiments only once following the previous baseline LLaVA [28]. It would be too computationally expensive to conduct the pretraining and finetuning multiple times.

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 release the experiments compute resources in the Appendix.

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: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

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 discuss the potential societal impacts of our work in the Appendix.

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: [NA]

Justification: The paper poses no such risks.

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: Please see Section Experiments.

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: Please see Section Experiment and supplementary material.

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: The paper does not involve crowdsourcing nor research with 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: The paper does not involve crowdsourcing nor research with 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.