OCTAVIUS: MITIGATING TASK INTERFERENCE IN MLLMS VIA LORA-MOE

Zeren Chen[1, 2*], **Ziqin Wang**[1, 3*], **Zhen Wang**[2], **Huayang Liu**[2]
Zhenfei Yin[1,4†], **Si Liu**[3], **Lu Sheng**[2‡], **Wanli Ouyang**[1,4], **Jing Shao**[1‡]
[1] Shanghai AI Laboratory, [2] School of Software, Beihang University,
[3] Institute of Artificial Intelligence, Beihang University, [4] University of Sydney
{czr1604,wzqin,wz1234,hy43,liusi,lsheng}@buaa.edu.cn,
{yinzhenfei,shaojing}@pjlab.org.cn, wanli.ouyang@sydney.org.cn.

ABSTRACT

Recent studies have demonstrated Large Language Models (LLMs) can extend their zero-shot generalization capabilities to multimodal learning through instruction tuning. As more modalities and downstream tasks are introduced, negative conflicts and interference may have a worse impact on performance. While this phenomenon has been overlooked in previous work, we propose a novel and extensible framework, called Octavius, for comprehensive studies and experimentation on multimodal learning with Multimodal Large Language Models (MLLMs). Specifically, we combine the well-known Mixture-of-Experts (MoE) and one of the representative PEFT techniques, *i.e.,* LoRA, designing a novel LLM-based decoder, called LoRA-MoE, for multimodal learning. To the best of our knowledge, we are one of the pioneering efforts to introduce MoE into MLLMs to address this problem. The experimental results (about 20% improvement) have shown the effectiveness and versatility of our design in various 2D and 3D downstream tasks. Code and datasets are available at https://openlamm.github.io/paper_list/Octavius.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Huang et al., 2023; Liu et al., 2023; Li et al., 2023a; Zhu et al., 2023) have been considered as promising general-purpose interfaces that can perform various multimodal tasks under few-/zero-shot settings. Apart from leveraging the powerful Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023a) as the universal interfaces that unify the responses to different types of tasks as task-specified textual sequences, the keys to the success of MLLMs are to reliably perceive more modalities and be efficiently fine-tuned to adapt more downstream tasks.

To achieve this goal, MLLMs rely on the instruction-tuning scheme (Ouyang et al., 2022) where the model is fine-tuned based on multimodal instruction-following dialogues orchestrated from various multimodal tasks. Moreover, thanks to the Parameter-Efficient Fine-Tuning (PEFT) techniques (*e.g.*, LoRA (Hu et al., 2021) and Adapter (Houlsby et al., 2019)) where only small trainable components are injected in the model and updated during fine-tuning, recent MLLMs (Zhang et al., 2023; Yin et al., 2023; Ye et al., 2023) can efficiently learn to solve downstream tasks with a small scale of annotated data, while preserve the language proficiency and generalizability to novel situations. Remarkably, these models achieve comparable performance at low costs in comparison to LLaVA (Liu et al., 2023), KOSMOS series (Huang et al., 2023; Peng et al., 2023) and Shikra (Chen et al., 2023), which are learned by full model fine-tuning with a large amount of multimodal data.

However, PEFT has to address the crucial tug-of-war problem (Hadsell et al., 2020), where simultaneously learning different tasks may cancel each task-specific optimization out, and ultimately com-

*Equal contribution.
†Project leader.
‡Corresponding author.

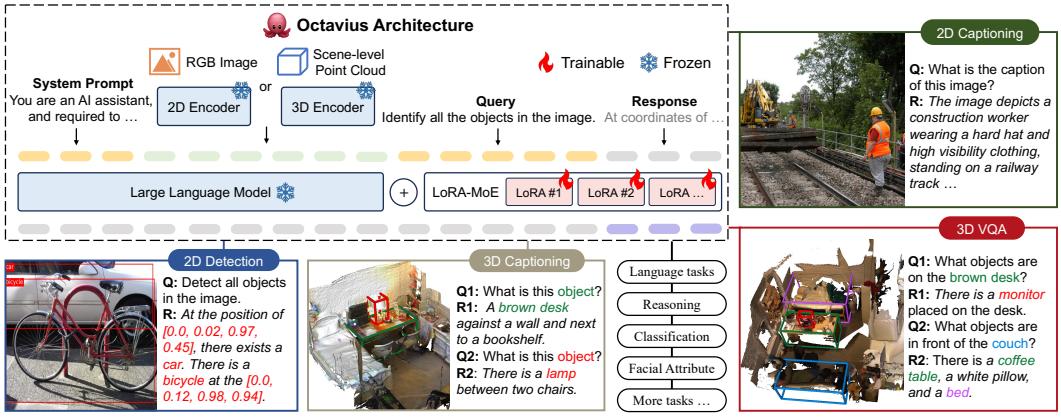

Figure 1: **Octavius** is a unified, multimodal large language model with a novel capability to comprehend various tasks across different modalities, including but not limited to 2D captioning, 2D detection, 3D VQA, and 3D dense captioning.

promise the performance of each downstream task. This problem is much more severe in MLLMs, especially when more modalities and tasks are involved, but only a few well-annotated data are available. First, the features from new modalities are not easy to be aligned with each other, not to mention compatible with the LLM-based language decoders. Second, simultaneously learning to acquire knowledge at distinct granularities, such as the instance-level perception (*e.g.*, object detection) and logical reasoning (*e.g.*, VQA), may lead to significant interference. Third, it is more complicated by using an LLM-based decoder to generate textual responses that meet the special requirements of tasks in modalities other than natural language, such as the bounding box coordinates in detection tasks, or action sequences for robotics.

To resolve this issue, we propose **LoRA-MoE**, which combines the well-known Mixture-of-Experts (MoE) (Jacobs et al., 1991; Jordan & Jacobs, 1994) and one of the representative PEFT techniques, *i.e.*, LoRA (Hu et al., 2021). Based on LoRA-MoE, an LLM-based decoder can efficiently be involved in more downstream tasks and more modalities by learning more LoRA modules. Different from conventional MoE models (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Du et al., 2022), we adopt a simple yet effective instance-based gate routing scheme, sparsely activating independent LoRA experts with instance-level instructions and further acquiring task-specific knowledge for better aligning different tasks. Notably, to the best of our knowledge, we are one of the pioneering efforts to introduce MoE into MLLMs to address the tug-of-war problem.

To validate the effectiveness of LoRA-MoE, in this work, we investigate a more complicated scenario, where the MLLMs should simultaneously learn downstreaming tasks from more additional modalities, such as 2D images and 3D point clouds. This scenario is especially useful for embodied agents (Duan et al., 2022; Mu et al., 2023; Driess et al., 2023). Specifically, in addition to the off-the-shelf image encoder, we design a point cloud encoder called **Object-As-Scene**, which provides language-aligned scene-level point cloud representations. This encoder at first gathers language-aligned point cloud features of each instance (Xue et al., 2023) in a scene, which are then aggregated into a scene-level feature based on the attention operation guided by the input instructions.

Based on the aforementioned contributions, we introduce a novel and extensive framework called **Octavius**, which learns the MLLMs upon the instruction-following datasets adapted from LAMM (Yin et al., 2023) and ScanNet (Dai et al., 2017). As shown in Figure 1, Octavius can successfully address various 2D/3D vision and language tasks, including but not limited to 2D detection, 2D captioning, 3D VQA, and 3D dense captioning. We conduct various experiments to validate the effectiveness and versatility of our design, improving multiple downstream tasks by about 20% while increasing only a few trainable parameters.

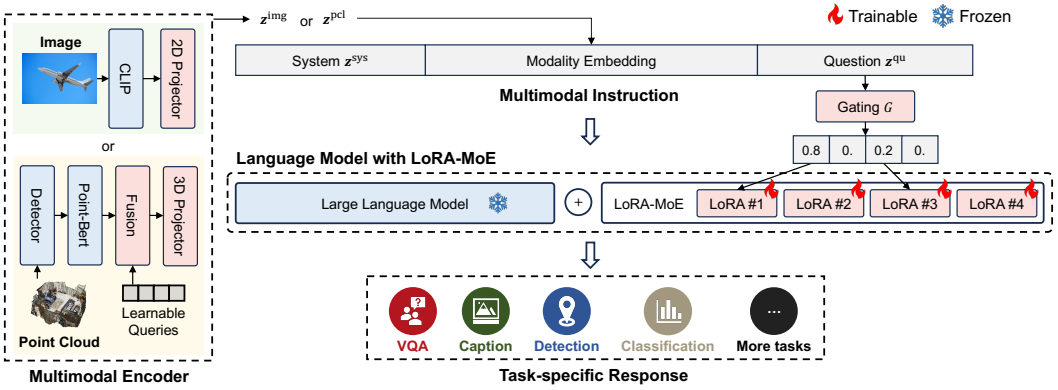

Figure 2: **Overall pipeline of Octavius**. We design corresponding encoders for different modalities, with the primary objective of empowering the LLMs to gain a deeper understanding of visual features. Additionally, we propose a dynamic gating network that selects distinct LoRA experts based on input instructions, thereby proficiently mitigating interference arising from multimodal learning.

2 Related Works

Large Language Models (LLMs) & PEFT. Recently, Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Chiang et al.; Touvron et al., 2023b) have gained significant attention due to their impressive capabilities in language generation (Zhang et al., 2022a), in-context learning (Wei et al., 2022), and reasoning (Touvron et al., 2023a). For both data- and compute-efficient adaptation on certain downstream tasks, several PEFT (Parameter-Efficient Fine-Tuning) (Li & Liang, 2021; Houlsby et al., 2019; Karimi Mahabadi et al., 2021; Hu et al., 2021) are proposed. For instance, LoRA (Hu et al., 2021) represents weight updates using two smaller matrices through low-rank decomposition, where original weights are kept frozen while the new update matrices are trained. In this work, we adopt LoRA for efficient MLLMs fine-tuning.

Multimodal Large Language Models (MLLMs). Several recent studies have attempted to extend the capability of LLMs to multimodal tasks. Alayrac et al. (2022); Li et al. (2023b); Liu et al. (2023); Zhang et al. (2023); Yin et al. (2023); Chen et al. (2023); Peng et al. (2023) introduce image modality in LLMs for comprehending 2D visual content. Hong et al. (2023) combines LLMs with 3D modality by rendering point clouds into 2D images and utilizing them to represent 3D visual features. Driess et al. (2023); Mu et al. (2023); Brohan et al. (2023) establish connections between visual inputs and embodied controls for robotic tasks. Despite its wide range of multimodal applications, the performance degradation caused by interference between tasks and modalities during fine-tuning in MLLMs receives inadequate attention.

Mixture-of-Experts (MoE). Deep MoE models are proposed to increase the number of model parameters without adding computational overhead in the field of computer vision (Riquelme et al., 2021; Mustafa et al., 2022; Shen et al., 2023) and natural language processing (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022). Different from these approaches, we aim to address conflicts between tasks with MoE. Adamix (Wang et al., 2022a), an approach related to but distinct from ours, randomly selects experts during training and uses the average weights of experts in inference, which may be analogous to dropout (Srivastava et al., 2014) in certain cases. In this paper, we desire for a dynamic gate routing strategy to automatically calculate the weights of each LoRA expert according to the input instructions, adapting MLLMs for broader multimodal applications.

3 Methodology

As illustrated in Figure 2, we propose an extensible framework called **Octavius** for multimodal instruction tuning. In Section 3.1, we first elaborate on the tug-of-war problem and propose a unified LoRA-MoE decoder to break through the bottleneck caused by interference between different tasks and modalities. We then verify our design using both image and point cloud modality in this work and describe corresponding encoders in Section 3.2.

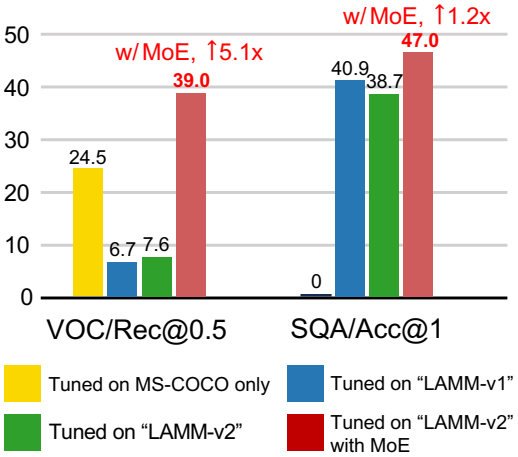

Figure 3: We conduct a simple pilot study on PASCAL VOC and ScienceQA to demonstrate the tug-of-war problem and the effectiveness of our proposed LoRA-MoE. Recall@0.5 denotes recall at an IoU threshold of 0.5, respectively.

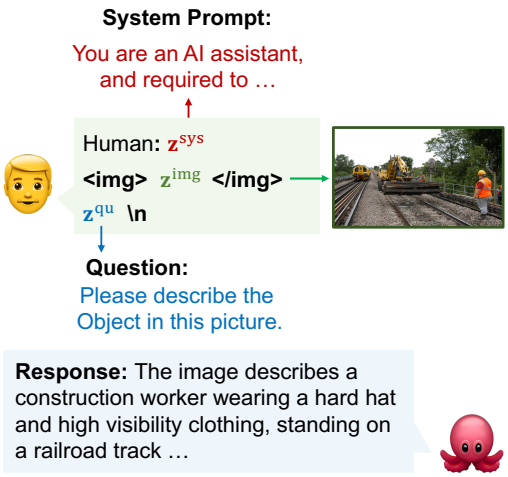

Figure 4: We follow previous works, *e.g.*, LAMM (Yin et al., 2023), LLaVA (Liu et al., 2023), to apply an instruction-following training pipeline.

3.1 Multimodal Decoder

3.1.1 The Tug-of-war Problem

Interference among different modalities and tasks is a common and critical issue (Zhao et al., 2018; Vandenhende et al., 2021) in multimodal and multitask learning. While MLLMs can alleviate this problem by adopting the same learning objective, *i.e.*, next-token prediction loss, for all tasks, there still exists task-specific divergences that limit their potential in various downstream tasks. Since previous works have yet to delve into the tug-of-war phenomenon in MLLMs, we conduct a simple pilot study on image modality to reveal this problem.

The pipeline of image encoders in MLLMs are simple and similar in previous works. Here, we select LAMM (Yin et al., 2023) as our model due to its rich benchmarks on downstream tasks. We fine-tune LoRA and projector in LAMM following (Yin et al., 2023) and validate zero-shot performance on PASCAL VOC (Everingham et al., 2010) and ScienceQA (Lu et al., 2022) datasets. We report the recall and precision at an Intersection over Union (IoU) threshold of 0.5 on PASCAL VOC and the accuracy of multiple-choice questions on ScienceQA.

The results are shown in Figure 3. Although the original dataset of LAMM, referred to as "LAMM v1", contains numerous images from MS-COCO (Lin et al., 2014), the lack of sufficient detection instructions results in poor performance on PASCAL VOC. To overcome this problem, we leverage the entire COCO detection annotations and GPT-API (OpenAI, 2023) to generate additional detection instructions as supplementation, constructing a new dataset called "LAMM v2" for better generalization of detection tasks. After verifying the detection ability of LAMM by using COCO detection instructions alone, we find using a mixed dataset does not lead to a huge improvement in detection performance. Also, there is a decline in the VQA tasks. Moreover, we can achieve the same results on another dataset used in LLaVA (Liu et al., 2023). It can be concluded that MLLMs suffer from a severe tug-of-war problem on image modality, not to mention incorporating more modalities for training simultaneously.

3.1.2 LoRA-MoE

Some prior works (Kendall et al., 2018; Chen et al., 2018) have attempted to balance the magnitudes of losses or gradients across different tasks to address the tug-of-war issue. However, considering that different objectives are defined for each task in multitask learning, it is challenging to directly extend existing methods to MLLMs that adopt a unified optimization objective for all tasks. In this section, we introduce the concept of Mixture-of-Experts (Jacobs et al., 1991; Jordan & Jacobs,

1994; Shazeer et al., 2017), proposing a unified LoRA-MoE decoder based on a instance-based gate routing strategy.

Revisiting MoE Models. A typical MoE Model injects multiple MoE layers into LLM to accommodate a greater number of parameters. The MoE layer consists of a group of N expert networks $E_1, E_2, ..., E_N$ and a gating network G, taking the previous tokens as input and producing the probability of the next token:

$$\texttt{tok}_i = \sum_k^N G(\texttt{tok}_{0...i-1})_k E_k(\texttt{tok}_{0...i-1}), \tag{1}$$

where $\texttt{tok}_i$ denotes i-th token. We refer to this kind of gating network, which based on token-level input, as **token-based gate**. Furthermore, to prevent G from consistently producing imbalanced weights that favor only a few experts, an auxiliary loss $\mathcal{L}_{\text{balance}}$ (Shazeer et al., 2017; Zhou et al., 2022; Wu et al., 2022) is introduced to balance gating routing. For example, Shazeer et al. (2017) minimize the coefficient of variation of the gate values for each token, encouraging all experts to have equal importance:

$$\mathcal{L}_{\text{balance}} = \alpha \, \texttt{CV}\Big[\sum_i G(\texttt{tok}_{0...i-1})\Big]^2, \tag{2}$$

where α is a hyper-parameter and $\texttt{CV}(\cdot)$ is the coefficient of variation, which is the ratio of the standard deviation to the mean.

LoRA-MoE and Instance-based Gate Routing. Different from token-based gate in LLM, we design a simple but effective routing strategy for MLLMs, assigning downstream tasks to independent experts for specific knowledge based on individual instances, called **instance-based gate**. It is motivated that the input questions $\boldsymbol{z}^{\text{qu}}$ applied in multimodal instructions will substantially affect the responses generated by MLLMs, we take the questions as input to predict routing scores for each expert. Then, we select sparsely-activated experts based on routing scores for each individual instance to generate the entire sentence. In this work, the LoRA module is treated as an expert in MLLMs, combining instance-based gate with it to alleviate interference arise from multimodal learning, named **LoRA-MoE**. By replacing LoRA in each projection layer of language model f^{LLM} with a group of independent LoRA experts $\{E^{\text{LoRA}}\}_N$, we can predict the i-th token value as follow:

$$\texttt{tok}_i = f^{\text{LLM}}(\texttt{tok}_{0...i-1}) + \sum_k^N G(\boldsymbol{z}^{\text{qu}})_k E_k^{\text{LoRA}}(\texttt{tok}_{0...i-1}). \tag{3}$$

Additionally, we find that $\mathcal{L}_{\text{balance}}$ is incompatible and infeasible with LoRA-MoE in an instance-based gate scenario. For example, it is more reasonable to assign detection samples to a LoRA expert proficient in localizing than the other experts for the purpose of balancing. Therefore, we can observe some imbalance phenomenon in experiments (see Section 4.3 for details), unless the amount of data for each task in the whole dataset is balanced.

Compared with previous MoE models, LoRA-MoE allows for efficient fine-tuning on small datasets and faster convergence with instance-based gate routing. During the inference phase, if the downstream tasks and input questions are specified, LoRA-MoE can also merge parameter weights with language model like vanilla LoRA to reduce storage requirements and inference costs.

Instruction Tuning with LoRA-MoE. Given the target modal features $\boldsymbol{z}^{\text{img}}$ or $\boldsymbol{z}^{\text{pcl}}$, we construct image-text conversation pairs in an instruction-following format based on previous works (Zhang et al., 2023; Yin et al., 2023; Peng et al., 2023), as shown in Figure 4. The language model with LoRA-MoE is then trained to predict corresponding responses based on the system prompts, target modal features and questions.

3.2 Modality Encoder

3.2.1 Image Encoder

Benefit from the pioneer vision-language model (Radford et al., 2021) that bridges the gap between the image and language modality, Li et al. (2023a); Liu et al. (2023); Chen et al. (2023); Zhang et al.

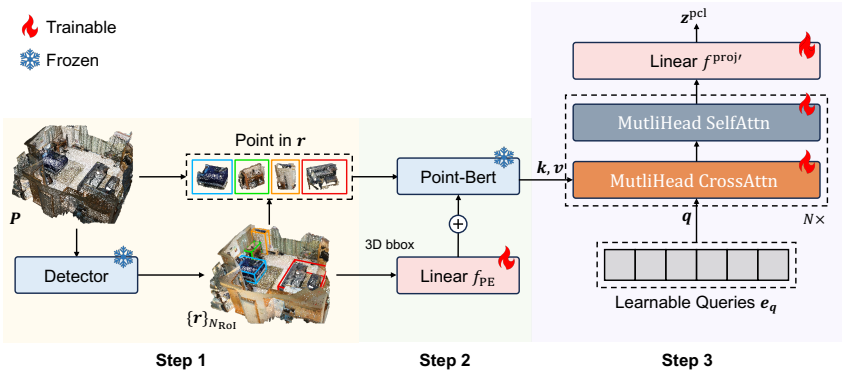

Figure 5: **Structure of Object-As-Scene**. To acquire scene-level features, we follow a three-step process. Firstly, we obtain RoIs from a given point cloud using a pre-trained detector. Next, we pre-train a Point-Bert model following a ULIP-like pipeline and employ it to extract instance-level 3D features. Finally, by aggregating features from visual embedding, we derive the final scene-level feature.

(2023); Yin et al. (2023) achieve impressive results. We follow their pipeline to extract features for image modality. Specifically, for an image input $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, we use the pre-trained CLIP visual encoder ViT-L/14 f^{CLIP} (Radford et al., 2021) to extract the language-aligned visual feature $\boldsymbol{h}^{\text{img}}$, following by a trainable linear layer f^{proj} to match the dimension of $\boldsymbol{h}^{\text{img}}$ with the word embedding space in language model:

$$\boldsymbol{h}^{\text{img}} = f^{\text{CLIP}}(\boldsymbol{I});\ \boldsymbol{z}^{\text{img}} = f^{\text{proj}}(\boldsymbol{h}^{\text{img}}), \tag{4}$$

where $\boldsymbol{z}^{\text{img}}$ is the output features of image modality for further instruction tuning.

3.2.2 Point Cloud Encoder

Conventional 3D methods typically apply 3D CNNs (Yan et al., 2018; Shi et al., 2020; Qi et al., 2017) or Transformers (Zhao et al., 2021; Fan et al., 2022) as feature extractors to process sparse point cloud data. However, they still retain numerous background points with low-density information, which may confuse the subsequent language models in MLLMs, ignoring the pivotal elements in the scene. Besides, the unavailability of encoders capable of aligning scene-level 3D features with language may pose significant challenges for LLMs in comprehending the semantic information of the entire scene. To address these issues, we propose Object-As-Scene as our point cloud encoder dedicated to language-aligned scene-level 3D representation generation, as illustrated in Figure 5.

Step 1: Locating Regional RoIs as Candidates. An intuitive way to avoid excessive background points is to identify specific regions in the scene that may contain instances or relevant semantic information and encode entire scene with these regions. Specifically, given a 3D point cloud scene, we employ a pre-trained object detector, *i.e.*, FCAF3D (Rukhovich et al., 2022), to locate candidate RoIs (Region-of-Interest) $\{\boldsymbol{r}\}_{N_{\text{RoI}}}$. Note that N_{RoI} denotes the number of RoIs. Besides, for tasks such as captioning or classification that primarily focus on instances mentioned in the conversations, we directly use regional features associated with these instances as input.

Step 2: Extracting RoI Features Aligned with Language and Image. Inspired by a recent work (Xue et al., 2023), we pre-train a Point-Bert (Yu et al., 2022) encoder $f^{\text{Point-Bert}}$ aligned with both language and image modalities following a ULIP-like pre-training pipeline, allowing us to extract instance-level 3D visual features from points $\boldsymbol{P} \in \mathbb{R}^{N\times 6}$ in candidate RoIs $\{\boldsymbol{r}\}_{N_{\text{RoI}}}$ during instruction tuning:

$$\{\boldsymbol{h}^{\text{pcl}}\}_{N_{\text{RoI}}} = f^{\text{Point-Bert}}(\boldsymbol{P}, \{\boldsymbol{r}\}_{N_{\text{RoI}}}). \tag{5}$$

More details about improved ULIP-like pre-training pipeline can be found in the Appendix.

Step 3: Aggregating RoI Features as Scene. Next, we adopt a fusion module with two stacked transformer layers to fuse scene-level features with RoI features. Specifically, we utilize multi-head cross-attention mechanism (Vaswani et al., 2017) (denoted as `MHCA`) to attend a group of trainable

Table 1: **Comparisons on image modality.** We investigate the zero-shot performance of our method on multimodal (2D) scenarios. We use the following abbreviations in this table and subsequent experiments for ease: "Det." for detection tasks, "Cap." for captioning task, and "Cls." for classification tasks.

Models	MoE	FT. Dataset	Det. (IoU=0.5)		VQA	Cap.	Cls.	Facial Attr		Avg.
			Recall	Prec	Acc@1	CIDEr	Acc@1	Hair Acc@1	Smile Acc@1	
LAMM		LAMM v2	7.61	5.95	40.31	0.21	**73.50**	58.04	50.15	–
	✓		**39.04**	**35.21**	**46.95**	**5.66**	65.40	**60.93**	**59.82**	20.89% ↑
LLaVA-LoRA		LLaVA	–	–	52.35	**30.75**	2.89	**12.50**	50.23	–
	✓		–	–	**55.58**	23.08	**41.00**	3.93	**52.17**	18.36% ↑

queries $\boldsymbol{e}_q$ with RoI features $\{\boldsymbol{h}^{\text{pcl}}\}_{N_{\text{RoI}}}$:

$$\boldsymbol{h}_q^{\text{pcl}} = \texttt{MHCA}(\boldsymbol{q} = \boldsymbol{e}_q, \boldsymbol{kv} = \{\boldsymbol{h}^{\text{pcl}} + f_{\text{PE}}(\boldsymbol{r})\}_{N_{\text{RoI}}}). \tag{6}$$

f_{PE} is used to transform 3D bbox coordinates into positional embedding to enrich spatial information. And $\boldsymbol{q}, \boldsymbol{k}$ and $\boldsymbol{v}$ denote query, key, and value in attention. As in image encoder (Section 3.2.1), a trainable linear layer is also applied for final 3D features $\boldsymbol{z}^{\text{pcl}}$:

$$\boldsymbol{z}^{\text{pcl}} = f^{\text{proj}'}(\boldsymbol{h}_q^{\text{pcl}}) \tag{7}$$

3D Instruction Data. We construct a 3D instruction tuning dataset called **Scan2Inst** using ScanNet (Dai et al., 2017) as our 3D instruction tuning dataset due to its diverse tasks and annotated categories. Following Wang et al. (2022b), we use GPT-API to generate a total of 80k data pairs comprising instructions and responses based on original dataset.

4 Experiment

4.1 Experiment Setup

To explore the effectiveness of our framework in multimodal learning, we fine-tune Octavius in three modality setups: **i.)** image modality, **ii.)** point cloud modality and **iii.)** both image and point cloud modalities. We then evaluate the zero-shot performance using these three fine-tuned models on various downstream tasks. More details about architecture and training scheme are provided in Appendix.

Instruction Datasets. For image modality, we follow Yin et al. (2023) to construct "LAMM v2", an instruction dataset consisting of MS-COCO (Lin et al., 2014) and Bamboo (Zhang et al., 2022b), which includes object detection, classification, captioning, and other common 2D tasks. For point cloud modality, we utilize ScanNet (Dai et al., 2017) to generate an instruction dataset called "Scan2Inst" which contains VQA, captioning, and classification tasks. In the multimodal learning (2D&3D) setup, we merge the image and point cloud instruction dataset to fine-tune the entire framework simultaneously.

Quantitative Zero-shot Evaluation. We perform zero-shot evaluation on various downstream tasks for both image and point cloud modalities. For image modality, we perform Visual Question Answering (VQA) on ScienceQA (Lu et al., 2022), classification on CIFAR-10 (Krizhevsky et al., 2009), captioning on Flickr30K (Young et al., 2014) and facial attribute recognition on CelebA (Liu et al., 2018). Note that we evaluate performance in VQA tasks through multiple-choice selection. For point cloud modality, we perform classification on ShapeNet (Chang et al., 2015) (55 classes) and captioning on NR3D (Achlioptas et al., 2020). We also evaluate the performance of classification, captioning, and QA on the test split of ScanNet to verify the proposed Object-As-Scene encoder.

4.2 Quantitative Results.

All results are provided in Table 1, 2, 3 for different modality setups. Severe interference can be found in the experimental results, especially between the localizing tasks and semantic understanding tasks like VQA and captioning. After equipping with LoRA-MoE, we can observe a remarkable

Table 2: **Comparisons on point cloud modality.** We report both Fine-Tuning (FT.) Results and Zero-shot (ZS.) Results on the 3D tasks. We also compare our model with 3D-LLM (Hong et al., 2023). Here, † indicates the results of Scan2Cap are evaluated on a custom test set regenerated by 3D-LLM, which is different from ours.

Models	FT. Results					ZS. Results			ZS. Avg.
	Cap. (Scan2Cap)		VQA (ScanQA)		Cls. (ScanNet)	Cls. (ShapeNet)	Cap. (Nr3d)		
	BLEU-1	CIDEr	BLEU-1	CIDEr	Acc@1	Acc@1	BLEU-1	CIDEr	
3D-LLM (Flamingo)	36.10†	–	30.30	59.20	–	–	–	–	–
Ours	33.58	35.11	43.21	**168.21**	47.40	19.75	20.02	16.19	–
Ours w/ MoE	**35.94**	**39.38**	**44.24**	167.31	**48.80**	**24.85**	**21.16**	**17.22**	17.06% ↑

Table 3: **Comparison on image & point cloud modalities.** While there are some performance gaps compared to the model fine-tuned on a single modality, our model with the inclusion of MoE exhibits a superior performance (∼20%) compared to its counterpart.

FT. Dataset	MoE	2D Results (ZS.)						3D Results (FT.)			3D Results (ZS.)		Avg.
		Det.	VQA	Cap.	Cls.	Facial		Cap.	VQA	Cls.	Cls.	Cap.	
		Rec@0.5	Acc	CIDEr	Acc	Hair	Smile	CIDEr	CIDEr	Acc	Acc	CIDEr	
LAMM v2		7.61	40.31	13.28	73.50	58.04	50.15	–	–	–	–	–	–
	✓	39.04	46.95	26.71	65.40	60.93	59.82	–	–	–	–	–	–
Scan2Inst		–	–	–	–	–	–	39.56	162.14	47.60	19.75	16.19	–
	✓	–	–	–	–	–	–	39.38	167.31	43.40	24.85	17.22	–
LAMM v2+Scan2Inst		2.64	**39.71**	0.04	**71.66**	42.47	50.66	19.76	**182.00**	38.80	14.85	8.26	–
	✓	**34.30**	35.80	**10.06**	56.86	**51.52**	**54.22**	**33.29**	181.44	**47.20**	**21.10**	**17.22**	21.40% ↑

improvement of approximately 20% in all setups, demonstrating the effectiveness of our design in resolving the tug-of-war issue. Additionally, we compare our proposed point cloud encoder Object-As-Scene with a recent work, 3D-LLM (Hong et al., 2023) in Table 2. We achieve a comparable performance on Scan2Cap, and outperform 3D-LLM on ScanQA by a significant margin, suggesting a better scene-level understanding capability of Object-As-Scene. Besides, as shown in Table 3, as the complexity of interference among tasks increases, especially when tasks of different modalities are introduced, we can observe a huge performance drop when training the model with different modalities like image and point cloud simultaneously compared to separate training. Our proposed MoE-based decoder partially alleviates the degradation and even achieves comparable performance with separate training in some tasks.

4.3 Ablation and Analysis

LoRA-MoE. The results are shown in Table 4. We first ablate on MoE architecture by individually employing dedicated LoRAs for each tasks (denoted as "Individual" in the table). While the individual gate exhibits performance merits in specific tasks, its primary challenge is the difficulty in assigning suitable experts for tasks that are not encountered in the instruction dataset, thereby compromising the model's generalizability. Another observation is the superior efficacy of the sparse gate relative to the dense gate (denoted as "Weighted Sum" in the table), which is intuitive when considering that the dense gate can essentially be regarded as a singluar LoRA with additional parameters. Furthermore, we compare the performance of the sparse gate against the baseline model (single LoRA) under conditions of parameter-consistency during inference (*i.e.*, sparse top-2 gate only uses half of rank in LoRA compared with baseline model). The enhanced performance of sparse gate demonstrate that **MoE transcends a mere aggregation of parameters**.

Gate Routing in MoE and Load Balancing. As shown in Figure 6, there is a huge discrepancy in expert selection between detection and VQA tasks, which demonstrates the tug-of-war phenomenon, and explains why using a single LoRA yields poor performance on both tasks simultaneously. Additionally, it is found that the routing weights of experts assigned by gate network tend to concentrate on a subset of specific experts. In particular, in a 4-expert model, despite the superior performance compared to a 3-expert model, the final converged model ends up utilizing only 3 of 4 available experts. To further explore this imbalance issue, we conduct several experiments with load balancing loss (Equation 2) in Table 5. As a result, no improvements and better routing results are

Table 4: **Ablation studies on MoE architecture on 2D tasks.** "Sparse Top-2" gate picks out top-2 ranked experts based on routing scores. "Weighted Sum" gate uses the weighted sum of all experts as output. "Individual" gate employs different experts for each 2D tasks individually. "#Trainable Param." denotes the proportion of trainable parameters to total parameters.

Gate Type	LoRA-Rank	Det. (VOC, IoU=0.5)		VQA	#Trainable Param.
		Recall	Prec.	Acc@1	
– (Baseline)	32	7.61	5.95	40.31	0.4%
Sparse Top-2	32	39.04	35.21	46.95	1.6%
Weighted Sum	32	9.78	5.33	44.71	1.6%
Individual	32	28.38	25.64	48.54	2.4%
Sparse Top-2	16	32.81	24.46	39.11	0.8%
Sparse Top-2	8	25.44	21.87	37.65	0.4%

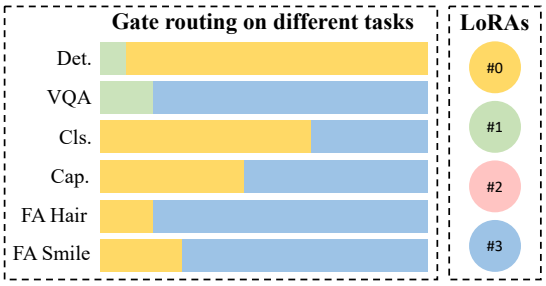

Figure 6: **Gate routing on different 2D tasks.** We use different color to represent i-th LoRA. The proportion represents the score of each expert.

Table 5: **Ablation studies on load balancing in MoE.** "LB" means apply load balance loss. **Token** means using token-based gate.

#Experts	LB	Token	Det.		VQA
			Recall	Prec.	Acc@1
4			39.04	35.21	46.95
4	✓		33.21	26.80	45.26
8			22.30	11.01	39.91
8	✓		21.52	12.10	37.53
4		✓		Fail	
4	✓	✓		Fail	

observed, which is intuitive because directly employing load balancing strategies is incompatible and infeasible in an instance-based gate scenario (see Section 3.1.2). We also attempt to replace the instance-based gate with token-based gate used in conventional MoE methods (Shazeer et al., 2017; Lepikhin et al., 2020). However, since only the LoRA modules are trained in our approach, the number of trainable parameters differs from conventional MoE models where all parameters of the entire foundation model are trained, resulting in poor convergence.

Besides, we also provide more ablations and analysis on point cloud encoder, LoRA-MoE, gate routing and qualitative results in Appendix.

5 Conclusion and Limitations

In this paper, we propose Octavius, a unified multimodal framework, to effectively address the critical challenge of task interference in complex learning scenarios. By integrating the Mixture-of-Experts (MoE) with LoRA, we present LoRA-MoE decoder, which delivers specialized learning paths for different tasks and modalities. After the validation across multiple modalities and tasks, Octavius alleviates the severe tug-of-war issue and achieves a significant performance boost in both 2D and 3D tasks.

Limitations. Compared to separate training on a single modality, introducing simultaneously multiple modalities for joint training may result in performance degradation, posing a challenge for future research. The combination of MLLMs and MoE still has great potential in addressing this problem, especially for a more complicated real-world system like embodied AI scenarios that require more modalities as input. Besides, we will further explore the token-based gate with load balancing strategies, especially when the number of downstream tasks increases.

Acknowledgement. This work is supported in part by National Key Research and Development Program of China (NO. 2021YFB1714300) , and National Natural Science Foundation of China (62132001).

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 422–440. Springer, 2020.

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.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

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.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pp. 5547–5569. PMLR, 2022.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *ijcv*, 88(2):303–338, 2010.

Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8458–8468, 2022.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1): 5232–5270, 2022.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *arXiv preprint arXiv:2307.12981*, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

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

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

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

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

OpenAI. Gpt-4 technical report, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

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*, pp. 8748–8763. PMLR, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In *European Conference on Computer Vision*, pp. 477–493. Springer, 2022.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*, 2023.

Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10529–10538, 2020.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3614–3633, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. *arXiv preprint arXiv:2210.17451*, 2022a.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Lemeng Wu, Mengchen Liu, Yinpeng Chen, Dongdong Chen, Xiyang Dai, and Lu Yuan. Residual mixture of experts. *arXiv preprint arXiv:2204.09636*, 2022.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1179–1189, 2023.

Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19313–19322, 2022.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022a.

Yuanhan Zhang, Qinghong Sun, Yichun Zhou, Zexin He, Zhenfei Yin, Kun Wang, Lu Sheng, Yu Qiao, Jing Shao, and Ziwei Liu. Bamboo: Building mega-scale vision dataset continually with human-machine synergy. *arXiv preprint arXiv:2203.07845*, 2022b.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268, 2021.

Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu. A modulation module for multi-task learning with applications in image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 401–416, 2018.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.