OCTAVIUS: MITIGATING TASK INTERFERENCE IN MLLMS VIA LORA-MOE

SUPPLEMENTARY MATERIALS

A ADDITIONAL IMPLEMENTATION DETAILS

Pre-training a language- and image-aligned Point-Bert following ULIP-like Pipeline. To improve the generalization performance of instance-level point cloud encoder, we select ScanNet (Dai et al., 2017) as our dataset due to its diverse object categories instead of the original ULIP dataset. Besides, we devised a memory bank in our pre-training framework for fast convergence and better representative capabilities.

For specific, given the instance-level point cloud $\boldsymbol{P}_i \in \mathbb{R}^{N\times 6}$ within each candidate RoIs $\boldsymbol{r}_i$, we first retrieve images $\boldsymbol{I}_i$ from related regions using its camera intrinsic matrix, and generate a simple prompt template $\boldsymbol{L}_i$, *e.g.*, "`a photo of {CLASS}`", obtaining a multimodal triplet $\langle \boldsymbol{P}_i, \boldsymbol{I}_i, \boldsymbol{L}_i \rangle$. We then extract corresponding features with a pre-trained CLIP (Gao et al., 2023) and a trainable Point-Bert encoder (Yu et al., 2022):

$$\boldsymbol{h}_i^{\text{pcl}} = f^{\text{Point-Bert}}(\boldsymbol{P}_i);\ \boldsymbol{h}_i^{\text{img}} = f^{\text{CLIP}}(\boldsymbol{I}_i);\ \boldsymbol{h}_i^{\text{lang}} = f^{\text{CLIP}}(\boldsymbol{L}_i), \tag{1}$$

We then contrast $\boldsymbol{h}_i^{\text{pcl}}$ with $\boldsymbol{h}_i^{\text{img}}$ and $\boldsymbol{h}_i^{\text{lang}}$, bringing 3D representation closer to semantic information of images and language:

$$\mathcal{L}_{\text{contrast}} = w_1 \mathcal{L}_{\langle \text{pcl,img} \rangle} + w_2 \mathcal{L}_{\langle \text{pcl,lang} \rangle}, \tag{2}$$

where $\mathcal{L}_{\langle \text{pcl,img} \rangle}, \mathcal{L}_{\langle \text{pcl,lang} \rangle}$ are respective contrastive loss and w_1, w_2 are corresponding loss weights. As mentioned before, the memory bank M is introduced to accommodate more negative samples for better feature alignment. For example, given i-th associated multimodal feature pair $\left\langle \boldsymbol{h}_i^{\text{pcl}}, \boldsymbol{h}_i^{\text{lang}} \right\rangle$, contrastive loss is given by

$$\mathcal{L}_{\langle \text{pcl,lang} \rangle} = -\sum_i \log \frac{\exp(\boldsymbol{h}_i^{\text{pcl}} \cdot \boldsymbol{h}_i^{\text{lang}} / \tau)}{\sum_{j \in \{i\} \bigcup M} \exp(\boldsymbol{h}_i^{\text{pcl}} \cdot \boldsymbol{h}_j^{\text{lang}} / \tau)}. \tag{3}$$

Eventually, the pre-trained Point-Bert is used for extracting instance-level 3D visual features that align with language and image in Object-As-Scene.

LLM Architecture and Training Scheme. We choose Vicuna-13B (Chiang et al.) as our LLM. Instructions are tokenized by SentencePiece (Kudo & Richardson, 2018). We apply LoRA-MoE on the language model for efficient fine-tuning and task-specific learning in all three setups. The number of experts in the above three setups is 4, 3, and 6, respectively. The rank of each LoRA expert is set to 32. During fine-tuning, we use an Adam (Kingma & Ba, 2014) optimizer with a total batch size of 64, a learning rate of 5×10^{-4}, and an epoch of 4 on all setups. All experiments are conducted on 4 NVIDIA A100 80GB GPUs.

Input images are all resized to 224×224 and split into 256 feature patches using CLIP ViT-L/14 (Radford et al., 2021). For point cloud data, we sample 1024 points from each RoI extracted by FCAF3D (Rukhovich et al., 2022) and generate corresponding features by pre-trained Point-Bert (Yu et al., 2022). Then we select N_{RoI} instances with bbox confidence larger than a threshold $\tau = 0.3$ for each scene. Next, we use 16 queries in the fusion module to obtain aligned 3D visual features. Furthermore, in the multimodal setup, we pad the output 3D visual features to 256 with masks for aligning with image patches.

Table 1: We conduct another pilot study to reveal the tug-of-war issues in multimodal learning.

FT. Dataset	MoE	Captioning			Avg.
		Flickr30k (2D ZS.)	Scan2Cap (3D FT.)	NR3D (3D ZS.)	
LAMM v2		0.21	-	-	-
Scan2Inst		-	35.10	16.19	-
LAMM v2+Scan2Inst		0.04	19.76	8.26	-
LAMM v2+Scan2Inst	✓	10.06	33.29	17.22	43.91% ↑

Table 2: **Ablation studies on point cloud encoder.** "PE" means positional embedding.

#Queries	PE	Fused Modality	Cap. (Scan2Cap)	VQA (ScanQA)	Cls. (ScanNet)	Avg.
			CIDEr	CIDEr	Acc	
16	Add	Lang.	35.11	168.21	47.40	83.57
16	Add	Lang. + Image	45.00	160.33	61.60	88.97
64	Add	Lang.	41.45	161.69	48.80	83.98
256	Add	Lang.	19.36	164.55	48.40	77.44
16	✗	Lang.	29.39	168.98	42.60	80.32
16	Concat	Lang.	26.65	174.86	47.40	82.97

Table 3: **Additional ablation studies on MoE architecture.** "Ques." and "Sys." refer to using question or system prompt in the instruction as gate input, respectively.

Gate Type	Gate Input		Det. (VOC, IoU=0.5)		VQA
	Ques.	Sys.	Recall	Prec.	Acc@1
– (Baseline)			7.61	5.95	40.31
Sparse Top-2	✓		39.04	35.21	46.95
Sparse Top-1	✓		22.42	21.23	36.88
Sparse Top-3	✓		38.57	36.02	43.89
Sparse Top-2	✓	✓	34.23	30.78	40.25

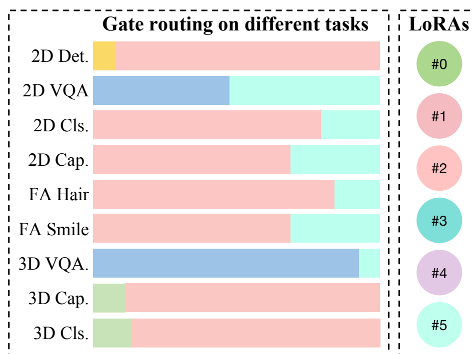


Figure 1: **Gate routing on 2D and 3D tasks.**

B ADDITIONAL EXPERIMENTS AND ABLATIONS

The Tug-of-War Issues in Multimodal Learning. We attempt to investigate the tug-of-war issues within the realm of multimodal learning. The results, delineated in Table 1, reveal that the tug-of-war issues not only prevail in multimodal learning, but also can be more severe. Here, we mainly focuses on 2D and 3D captioning tasks. When introducing more modalities during instruction tuning, a huge performance degradation can be observed, especially in 3D captioning tasks. After applying LoRA-MoE, the performance of 3D captioning tasks is enhanced, aligning with the levels achieved when fine-tuned on the single 3D modality. Meanwhile, the performance of 2D captioning is also greatly improved, underscoring the effectiveness of LoRA-MoE.

Point Cloud Encoder. As shown in Table 2, positional embedding (PE) improves the overall performance, since the position and scale of objects in PE can help the model better understand the semantic information of the scene and instances. We select "Add" in our model due to its more balanced downstream results. We also ablate different numbers of learnable queries. Considering that we extract about 50 RoIs in the scene, if we use far more queries than this number, the overall performance will decrease. 16∼64 is a reasonable range for the number of queries. Furthermore, we attempt to enhance the semantic information by introducing image features corresponding to 3D RoIs using the cross-attention mechanism. Despite the significant improvement in captioning and classification tasks, the additional cost of rendering images based on point cloud limits its practical usage, making it only a supplementary method.

Additional Ablations on LoRA-MoE. The results are shown in Table 3. By incorporating the "System Prompt" as an adjunctive input alongside the question, a performance decline can be found possibly due to the redundant global information embedded within the system prompt (*e.g.*, "You are an AI assistant, ..."), which introduces ambiguity and hampers the efficacy of the gate prediction network.

Furthermore, we also explore the impact of using gates of different top-k selection. Specifically, the top-1 gate exhibits poor performance due to its limited flexibility in selection, offering only four combinations of different experts, in contrast to the more versatile top-2 and top-3 gate. In scenarios that employing the top-3 gate, we find that the contribution of the third choice is relatively small across most tasks. For instance, in classification tasks, the distribution of routing weights often resembles "[0.1, 0, 0.4, 0.5]", which is closely similar to the scores in top-2 gate. Consequently, the top-3 gate yields comparable performance achieved by the top-2 gate.

More Analysis and Discussion on Gate Routing. Tasks such as classification, captioning and facial attribute recognition exhibit a tendency to engage similar experts with different routing scores. It can be attributed to underlying similarities among these tasks in training instructions and reponses, despite their varied downstream applications. For instance:

- **Captioning**

  Q: What is the caption of the image?

  A: The image depicts an older man and a young man in a close embrace. The older man, who is bald and wearing a red shirt, appears to be hugging the younger man. They are standing in front of a black wall, which enhances the contrast of their positions.

- **Classification**

  Q: What is the most accurate category label for this image?

  A: The image depicts a cute and adorable small white fluffy dog.

- **Facial Attribute**

  Q: What color is the person's hair in the image?

  A: The image shows a young woman with dark, long, and curly hair.

These examples indicate that in tasks like classification and facial attribute recognition, LLMs tend to offer comprehensive descriptions of the target object (including attributes like color, shape, and descriptive adjectives) rather than mere categorical labels, which are very similar in captioning tasks. Therefore it is reasonable that the gating network makes similar expert selections for these tasks. Conversely, in very different tasks like detection, gating networks generate distinct expert selections.

Moreover, we jointly fine-tune Octavius on both LAMM v2 and Scan2Inst datasets, supplementing our analysis with an additional illustration of gate routing on 2D and 3D tasks, as presented in Figure 1. Load balancing issues still occur within the distribution of gating scores, notably with experts #2 and #5. For the 3D modality, 3D captioning and 3D classification tasks mainly focus on instance-level perception, such as the caption or category of a specific object, while 3D VQA focuses more on inter-relations among multiple objects in the scene and the understanding of the entire scene. This divergence leads to two different pattern of routing weights between 3D captioning/classification and 3D VQA in Figure 1. Additionally, another interesting observation is the emergence of knowledge sharing across different modalities by certain experts (*e.g.*, expert #2), while others perfer for specilized modality (*e.g.*, experts #1 and #5).

Generalizability of Instance-based Gate Routing. To assess the generalizability of the proposed instance-based gate, we conduct several comprehensive ablation studies on the ScienceQA dataset Lu et al. (2022). The queries in ScicenceQA datasets comprise of distinct problem statements as well as contextual information. We employ GPT-3.5-turbo Brown et al. (2020) to enrich both the questions and the contexts separately, and ensure the enriched contents maintain consistency with the original semantics. Subsequently, we validate the proposed instance-based gate using these

Table 4: **Ablation studies on OOD generalization of query in VQA evaluation.** "Ctx." and "Ques." denotes context and question, respectively. We report top 1 accuarcy on ScienceQA Lu et al. (2022) test dataset.

Query Pattern		VQA
Enriched Ctx.	Enriched Ques.	
		46.95
✓		47.58
	✓	47.43
✓	✓	47.03

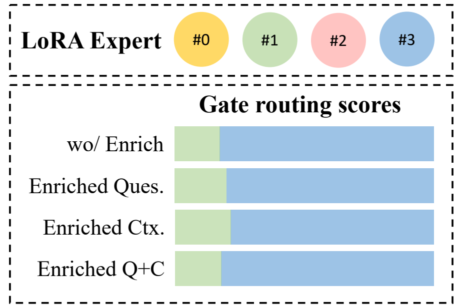


Figure 2: **Gate routing on different query pattern.** "Enriched Q+C" means using both enriched context and question as input.

enriched questions and contexts. As detailed in Table 4, Octavius achieve stable performance across all enriched data, highlighting the strong generalization capacity of instance-based gate in processing input queries of different patterns. We also present several examples in Figure 3. Furthermore, we provide a comparative analysis of the routing weights between the default and enriched queries in Figure 2. Remarkably, the model consistently selects similar gates with comparable weights, regardless of the modifications in the data. This consistency demonstrates the robustness of Octavius in effectively managing VQA tasks, proficiently navigating questions and contexts of diverse structures and complexities.

Complete Results on Downstream Tasks. We provide complete experimental results for detection, captioning, and VQA tasks in all setups, as shown in Table 5, 6 and 7. We report recall and precision at IoU thresholds of 0.5 and 0.25 in detection tasks, and "BLEU-1/2/3/4", "CIDEr", "METEOR" and "ROUGE-L" in both captioning and VQA tasks.

Table 5: **Complete results on 2D downstream tasks.** In the detection task, we also provide recall and precision of the predicted bounding box without categories.

MoE	Detection (PASCAL VOC)							
	w/ cls (IoU=0.5)		wo/ cls (IoU=0.5)		w/ cls (IoU=0.25)		wo/ cls (IoU=0.5)	
	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.
	7.61	5.95	10.1	7.91	20.96	16.41	27.14	21.24
✓	**39.04**	**35.21**	**44.16**	**39.63**	**51.38**	**46.12**	**59.19**	**53.13**

MoE	Captioning (Flickr30K)						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L
	13.283	7.328	3.733	1.883	0.21	12.482	17.707
✓	**26.705**	**15.163**	**8.296**	**4.566**	**5.66**	**16.979**	**26.849**

C ADDITIONAL VISUALIZATION

In this section, we provide several responses of Octavius in Figure 4, 5 and 6.

Table 6: **Complete results on 3D downstream tasks.** Here, † indicates the results of Scan2Cap is evaluated on a custom test set regenerated by 3D-LLM, which is different from ours.

Models	MoE	Captioning (Scan2Cap)						
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L
3D-LLM† (Flamingo)		36.10	24.50	18.70	15.60	–	17.60	35.80
Ours		34.16	20.92	12.45	7.56	**39.56**	13.03	**32.66**
Ours	✓	**35.93**	**21.66**	**12.79**	**7.75**	39.38	**13.34**	32.36

Models	MoE	VQA (ScanQA)						
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L
3D-LLM (Flamingo)		30.30	17.80	16.00	7.20	59.20	12.20	32.30
Ours		43.07	32.69	25.17	**19.26**	162.14	21.44	**45.08**
Ours	✓	**44.24**	**33.16**	**25.24**	19.16	**167.31**	**21.44**	44.87

Models	MoE	Captioning (Nr3d)						
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L
Ours		20.02	8.95	3.63	1.66	16.19	9.71	20.45
Ours	✓	**21.16**	**10.00**	**4.38**	**2.07**	**17.22**	**11.06**	**22.37**

Table 7: **Complete results on multimodal learning (2D & 3D).**

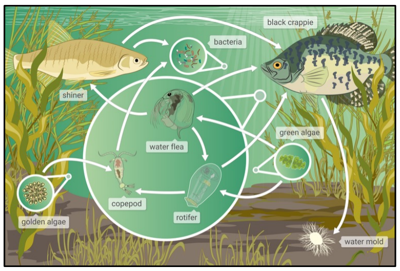
MoE	Detection (PASCAL VOC)							
	w/ cls (IoU=0.5)		wo/ cls (IoU=0.5)		w/ cls (IoU=0.25)		wo/ cls (IoU=0.5)	
	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.
	2.64	1.61	3.62	2.2	8.15	4.95	11.28	6.86
✓	**34.3**	**25.07**	**38.97**	**28.48**	**47.11**	**34.43**	**54.72**	**39.99**

MoE	Captioning (Flickr30K)						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L
	14.335	8.132	4.288	2.274	0.038	**13.673**	17.083
✓	**22.545**	**11.014**	**5.286**	**2.64**	**10.064**	11.6	**27.148**

MoE	Captioning (Scan2Cap)						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L
	26.16	14.79	7.91	4.36	**13.76**	29.26	19.76
✓	**36.62**	**21.91**	**12.56**	**7.29**	13.30	**31.69**	**33.29**

MoE	VQA (ScanQA)						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L
	45.63	**35.11**	**27.30**	21.01	**22.73**	**46.56**	**182.00**
✓	44.48	34.20	26.76	**21.04**	22.23	46.22	181.44

MoE	Captioning (Nr3d)						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L
	13.60	6.34	2.72	1.20	10.72	20.77	8.26
✓	**20.96**	**9.95**	**4.27**	**2.02**	**11.13**	**22.29**	**17.22**



Question: Which of the following organisms is the primary consumer in this food web?
Context: Below is a food web from Little Rock Lake, a freshwater lake ecosystem in Wisconsin. A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter moves between organisms in an ecosystem.
Choice: (A) copepod (B) black crappie (C) bacteria
Answer: After examining the picture closely, it can be concluded that copepod is the answer to this question.

Enriched Question: Who is the primary consumer among the organisms listed in this food web?
Context: Below is a food web from Little Rock Lake, a freshwater lake ecosystem in Wisconsin. A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter moves between organisms in an ecosystem.
Choice: (A) copepod (B) black crappie (C) bacteria
Answer: After analyzing the image, it can be concluded that copepod is the correct answer to this question.

Question: Which of the following organisms is the primary consumer in this food web?
Enriched Context: Here is a food web taken from Little Rock Lake, which is a freshwater lake ecosystem located in Wisconsin. A food web serves as a representation of how the matter consumed by organisms flows within an ecosystem. The arrows in a food web depict the movement of matter between organisms in the ecosystem.
Choice: (A) copepod (B) black crappie (C) bacteria
Answer: Upon careful observation of the picture, it becomes evident that copepod is the correct answer to this question

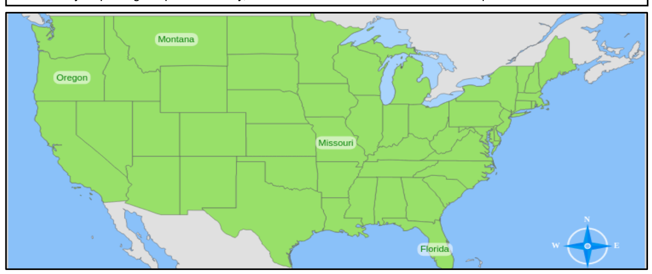
Enriched Question: Who is the primary consumer among the organisms listed in this food web?
Enriched Context: Here is a food web taken from Little Rock Lake, which is a freshwater lake ecosystem located in Wisconsin. A food web serves as a representation of how the matter consumed by organisms flows within an ecosystem. The arrows in a food web depict the movement of matter between organisms in the ecosystem.
Choice: (A) copepod (B) black crappie (C) bacteria
Answer: The answer to this question can be determined by closely observing the image, and it is copepod.

Question: Which property do these three objects have in common?
Context: Select the best answer.
Choice: (A) blue (B) hard (C) sticky
Answer: After examining the picture closely, it can be concluded that blue is the answer to this question.

Enriched Question: What is the common property among these three objects?
Context: Select the best answer.
Choice: (A) blue (B) hard (C) sticky
Answer: Upon careful observation of picture, it becomes evident that blue is the correct answer to this question.

Question: Which property do these three objects have in common?
Enriched Context: Please choose the most suitable answer.
Choice: (A) blue (B) hard (C) sticky
Answer: By observing the image closely, it becomes clear that the answer to this question is blue.

Enriched Question: What is the common property among these three objects?
Enriched Context: Please choose the most suitable answer.
Choice: (A) blue (B) hard (C) sticky
Answer: By inspecting the picture closely, one can conclude that the answer to this question is blue.



Question: Which of these states is farthest west?
Context: N/A
Choice: (A) Missouri (B) Florida (C) Montana (D) Oregon
Answer: After a thorough examination of the picture, it becomes apparent that the correct answer is (D) Oregon.

Enriched Question: Which state among these is the furthest to the west?
Context: N/A
Choice: (A) Missouri (B) Florida (C) Montana (D) Oregon
Answer: After a thorough examination of the picture, it becomes apparent that the correct answer is (D) Oregon.

Question: Which of these states is farthest west?
Enriched Context: N/A
Choice: (A) Missouri (B) Florida (C) Montana (D) Oregon
Answer: The correct answer to this question can be identified by closely examining the image, and it is Oregon.

Enriched Question: Which state among these is the furthest to the west?
Enriched Context: N/A
Choice: (A) Missouri (B) Florida (C) Montana (D) Oregon
Answer: After examining the picture closely, it can be concluded that Oregon is the answer to this question.

Figure 3: The response of Octavius given different query pattern in downstream VQA evaluation.

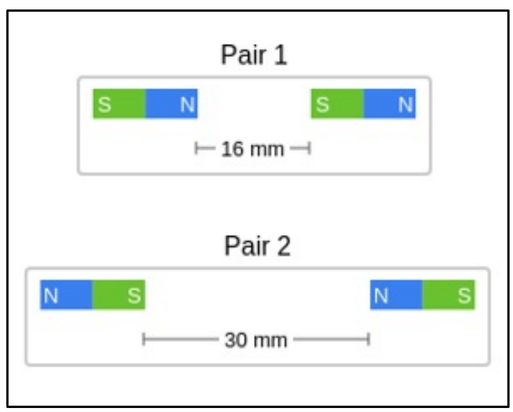
Question: What is the caption of the image ?
Answer: The image shows a man and woman holding hands in a river.

Question: What is the caption of the image ?
Answer: The image shows a man wearing a yellow safety vest, sitting on a chair with his arms extended.

Question: What is the caption of the image ?
Answer: The image depicts a scene where a group of people are gathered near a tower with a clock face.

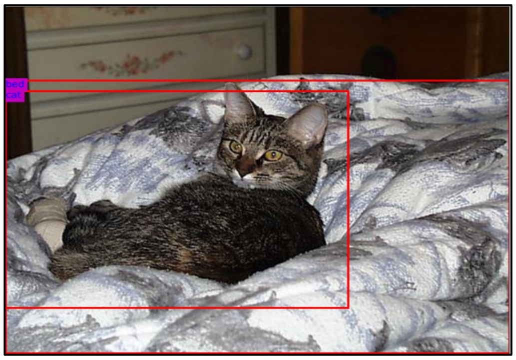
Question: What is the caption of the image ?
Answer: The image depicts a bustling street scene with numerous shops and vendors.

Question: What is the name of the colony shown?
Context: N/A
Choice: (A) New Hampshire (B) Vermont (C) New York (D) Delaware
Answer: Analysis of the image leads to the conclusion that New York is the correct response to this question.

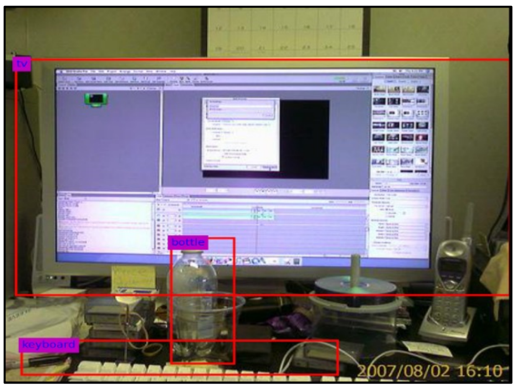


Question: Think about the magnetic force between the magnets in each pair. Which of the following statements is true?
Context: The images below show two pairs of magnets. The magnets in different pairs do not affect each other. All the magnets shown are made of the same material.
Choice: (A) The magnetic force is stronger in Pair 2. (B) The magnetic force is stronger in Pair 1. (C) The strength of the magnetic force is the same in both pairs.
Answer: By observing the image closely, it becomes clear that the answer to this question is B. The magnetic force is stronger in Pair 1.

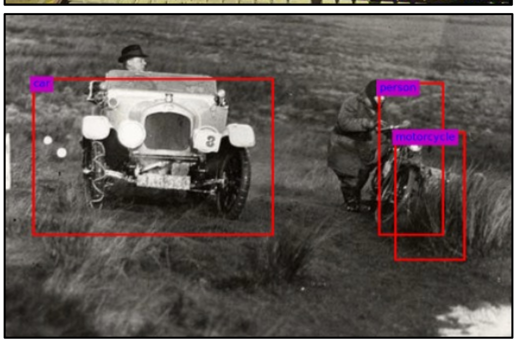
Figure 4: The response of Octavius on 2D captioning and VQA.



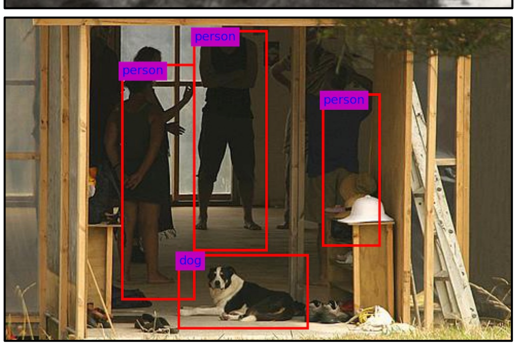
Question: Identify all objects in the image and provide their positions.
Answer: Positioned at [0.0, 0.24, 0.68, 0.86] within the image is an object that can be classified as belonging to the cat category, characterized by its unique texture and distinctive silhouette that differentiates it from other elements in the scene. This cat appears to be the focal point of the image, drawing the viewer's attention with its relaxed posture. At the [0.0, 0.21, 1.0, 0.99] position of the image, there is an object categorized as bed. The cat seems to be lying on the bed, adding a sense of tranquility and homeliness to the image.



Question: Identify all objects in the image and provide their positions.
Answer: Positioned at [0.15, 0.89, 0.7, 0.98] within the image is an object that can be classified as belonging to the keyboard category, and this object is also the only one in the image that has a specific type of texture and a distinctive shape that sets it apart from the other objects. At the coordinates of [0.41, 0.62, 0.52, 0.95] position of the image, there exists an object categorized as bottle, characterized by its unique shape. Upon close inspection of the image, it can be observed that there is an object positioned at [0.14, 0.15, 1.0, 0.77] that belongs to the tv category.



Question: Identify all objects in the image and provide their positions.
Answer: The object present at the [0.75, 0.42, 0.88, 0.73] coordinate in the image is classified as motorcycle. Positioned at [0.72, 0.3, 0.84, 0.67] within the image is an object that can be classified as belonging to the person category, and this object is also the only one in the image that has a specific type of texture and a distinctive shape that sets it apart from the other objects. At the coordinates of [0.07, 0.29, 0.52, 0.67] position of the image, there exists an object categorized as car.



Question: Identify all objects in the image and provide their positions.
Answer: The [0.62, 0.39, 0.73, 0.7] position of the image contains an object that can be classified as person. At the [0.23, 0.33, 0.37, 0.81] position of the image, there is an object that can be categorized as person. The [0.37, 0.26, 0.51, 0.71] position of the image contains an object that can be identified as belonging to the person category. At the exact coordinates of [0.34, 0.72, 0.59, 0.87] in the image, there is an object that can be identified as belonging to the dog category.

Figure 5: The response of Octavius on 2D detection.

Figure 6: The response of Octavius on 3D captioning and VQA.

REFERENCES

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.

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.

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.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pp. 1–15, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 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.

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.

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.

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.