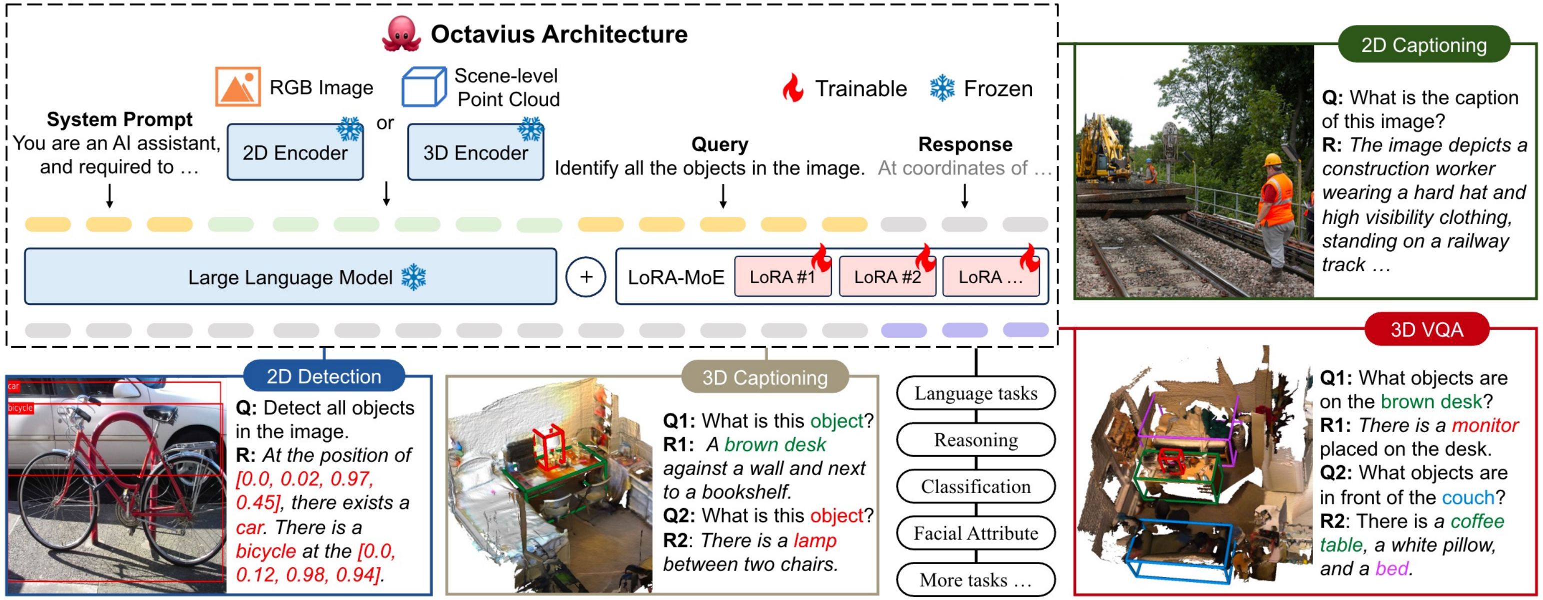
OCTAVIUS: MITIGATING TASK INTERFERENCE IN MLLMS VIA MOE

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], Yu Qiao[1], Jing Shao[1]
[1]Shanghai AI Laboratory, [2]School of Software, Beihang University
[3]Institute of Artificial Intelligence, Beihang University, [4]University of Sydney

Homepage is available at https://openlamm.github.io/paper_list/Octavius

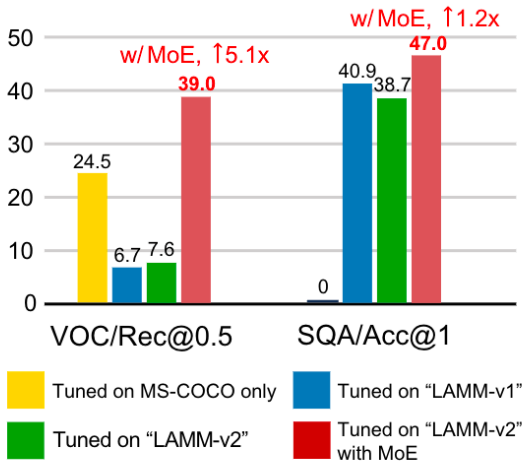
Overview

- Propose a method called **LoRA-MoE**, which combine Mixture-of-Experts(MoE) with LoRA in Multimodal Large Language Models(MLLMs).
- We designed a point cloud encoder called **Object-As-Scene** to provide language-aligned scene-level representations. Additionally, we constructed the **Scan2Inst** dataset for 3D instruction tuning using ScanNet.



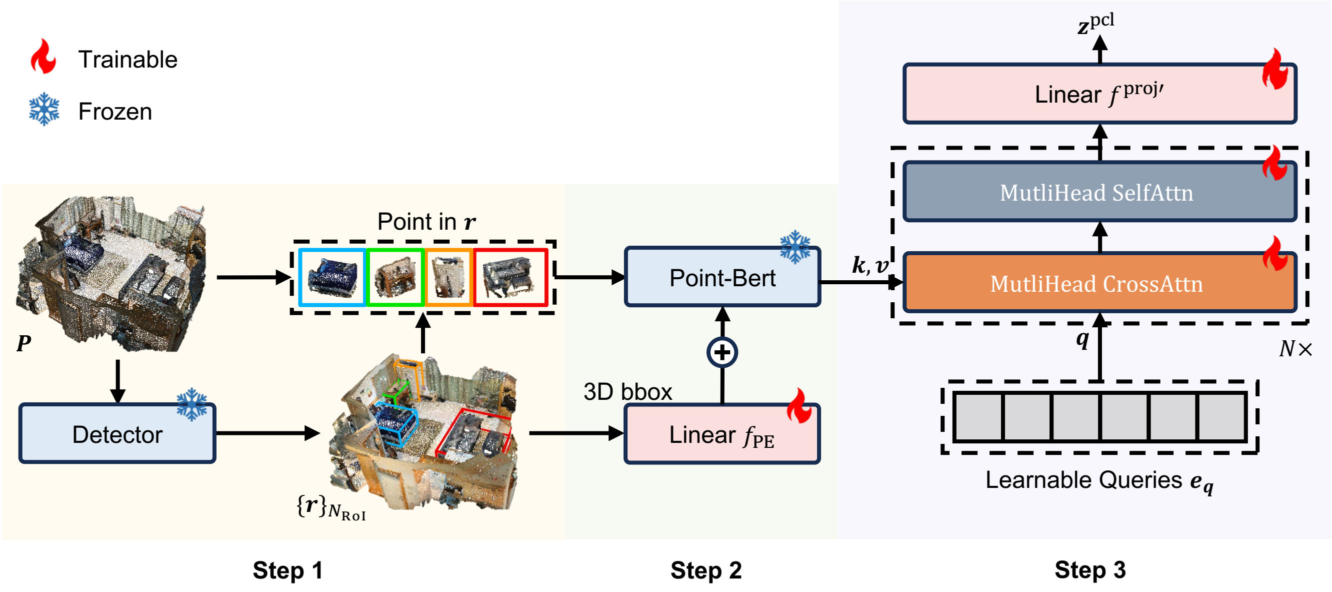
Introduction & Motivation

- For MLLMs with LoRA, simultaneously learning different tasks may cause conflicts, and ultimately compromise the performance of each downstream task.
  - E.g. For the detection task alone, the performance on VOC is 24.5. However, when combined with the VQA task, the performance drops to around 7.
- Objective: Design a versatile and scalable MLLM framework that can effectively address diverse multimodal tasks, even when instruction-tuning data is limited.
  - We propose LoRA-MoE, a combination of MoE and LoRA, which has led to improved performance on both VOC and SQA tasks.



Modality Encoder

- For image encoder, we use the pre-trained CLIP visual encoder ViT-L/14.
- For point cloud encoder, we propose a structure called Object-As-Scene.
  - Step 1: Locating regional RoIs as candidates.
  - Step 2: Extracting RoI features aligned with language and image.
  - Step 3: Aggregating RoI features as scene features.



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

$$h_q^{\text{pcl}} = \text{MHCA}(q = e_q, kv = \{h^{\text{pcl}} + f_{\text{PE}}(r)\}_{N_{\text{RoI}}})$$

$$z^{\text{pcl}} = f^{\text{proj}'}(h_q^{\text{pcl}})$$

3D Dataset Construction

- We employed GPT to enrich the question and answer in ScanQA.

Original question

What is over the chair beneath the blackboard?

Enriched questions

- Can you please describe the objects or items that are positioned on the chair beneath the blackboard?
- Are there any specific colors, patterns, or designs on the item placed over the chair beneath the blackboard?
- Could you provide information about the purpose or function of the object that is currently resting on the chair beneath the blackboard?

3D Dataset Construction

- We utilized annotations from ScanNet to construct data use templates.

Annotation

description: A black chair in front of the table.

Question templates

caption
1. What is the caption of the object?
2. What is the description of the object?
3. Can you provide a brief summary of the given text?
......

Caption task

Q: What is the caption of the object?
A: A black chair in front of the table.

Annotation

class: chair

Question templates

classification
1. What is the category of the object?
2. Which class does the item fall into?
3. What is the assigned label for the given item?
......

Answer templates

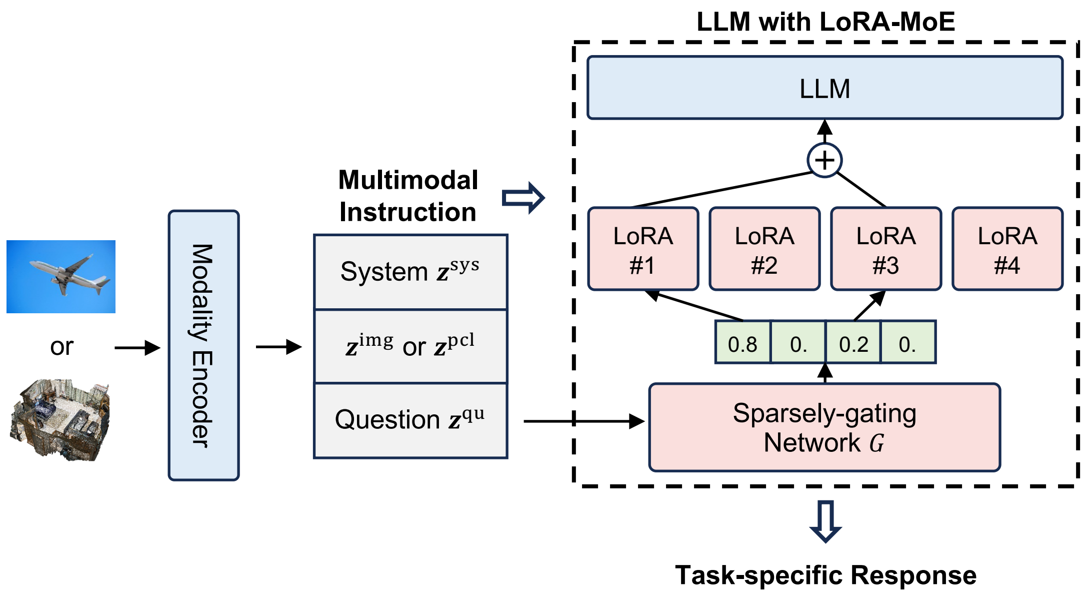
classification
1. The category of the object is {}.
2. The item falls into the class {}.
3. The entity fits in the category {}.
......

Classification task

Q: What is the category of the object?
A: The category of the object is chair.

Multimodal Decoder

- We propose a unified **LoRA-MoE** decoder based on an **instance-based gate** routing strategy.
  - Instance-based gate take the questions as input to predict routing scores for each expert.
  - We select sparsely-activated experts based on routing scores for each individual instance.



- For LoRA #k when generating i-th token:

$$Gate_k = G(\boldsymbol{z}^{\mathrm{qu}})_k$$

$$Emb_k = E_k^{\mathrm{LoRA}}(\mathtt{tok}_{0\dots i-1})$$

- Generate i-th token use sparsely-activated experts:

$$N = \{k; LoRA_k \text{ is activated}\}$$

$$\mathtt{tok}_i = f^{\mathrm{LLM}}(\mathtt{tok}_{0\dots i-1}) + \sum_{k}^{N} G(\boldsymbol{z}^{\mathrm{qu}})_k E_k^{\mathrm{LoRA}}(\mathtt{tok}_{0\dots i-1})$$

Experiments

- Comparison on image modalities

Models	MoE	FT. Dataset	Det. (IoU=0.5) Recall	Det. (IoU=0.5) Prec	VQA Acc@1	Cap. CIDEr	Cls. Acc@1	Facial Attr Hair Acc@1	Facial Attr Smile Acc@1	Avg.
LAMM		LAMM v2	7.61	5.95	40.31	0.21	**73.50**	58.04	50.15	–
LAMM	✓	LAMM v2	**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	–
LLaVA-LoRA	✓	LLaVA	–	–	**55.58**	23.08	**41.00**	3.93	**52.17**	18.36% ↑

- Comparison on point cloud modalities

Models	FT. Results Cap. (Scan2Cap) BLEU-1	Cap. (Scan2Cap) CIDEr	VQA (ScanQA) BLEU-1	VQA (ScanQA) CIDEr	Cls. (ScanNet) Acc@1	ZS. Results Cls. (ShapeNet) Acc@1	Cap. (Nr3d) BLEU-1	Cap. (Nr3d) CIDEr	ZS. Avg.
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% ↑

Experiments

- Comparison on image & point cloud modalities

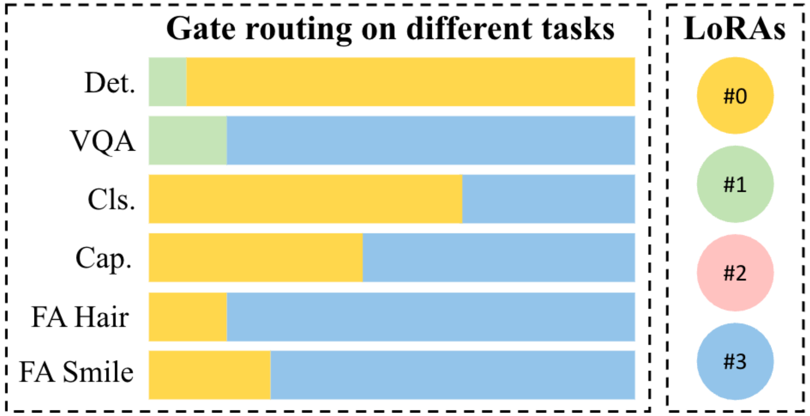
FT. Dataset	MoE	2D Results (ZS.) Det.						3D Results (FT.)			3D Results (ZS.)		Avg.
		Det.	VQA	Cap.	Cls.	Facial		Cap.	VQA	Cls.	Cls.	Cap.	
		Rec@0.5	Acc	CIDEr	Acc	Hair	Smile	CIDEr	CIDEr	Acc	Acc	CIDEr	
LAMM v2		7.61	40.31	13.28	73.50	58.04	50.15	–	–	–	–	–	–
	✓	39.04	46.95	26.71	65.40	60.93	59.82	–	–	–	–	–	–
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% ↑

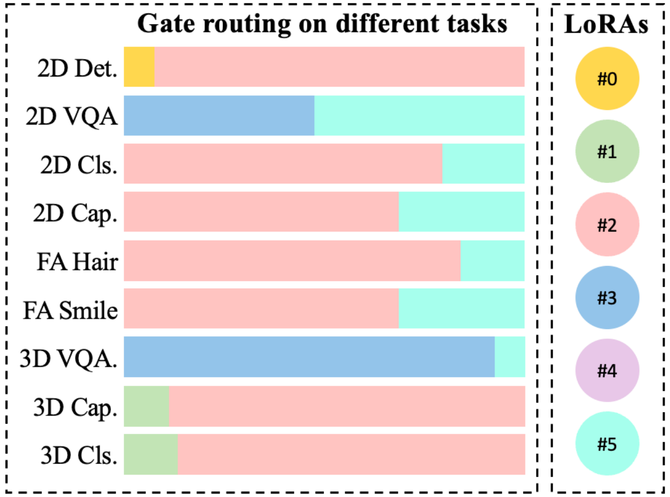
- Ablation studies of MoE on 2D tasks

Gate Type	Gate Input		#Experts	LoRA-Rank	Det. (VOC, IoU=0.5)		VQA	#Trainable Param.
	Question	System			Recall	Prec.	Acc@1	
– (Baseline)			1	32	7.61	5.95	40.31	0.4%
Sparse Top-2	✓		4	32	39.04	35.21	46.95	1.6%
Sparse Top-2	✓	✓	4	32	34.23	30.78	40.25	1.6%
Dense	✓		4	32	9.78	5.33	44.71	1.6%
Sparse Top-2	✓		4	16	32.81	24.46	39.11	0.8%
Sparse Top-2	✓		4	8	25.44	21.87	37.65	0.4%

Experiments

- Gate routing on different tasks



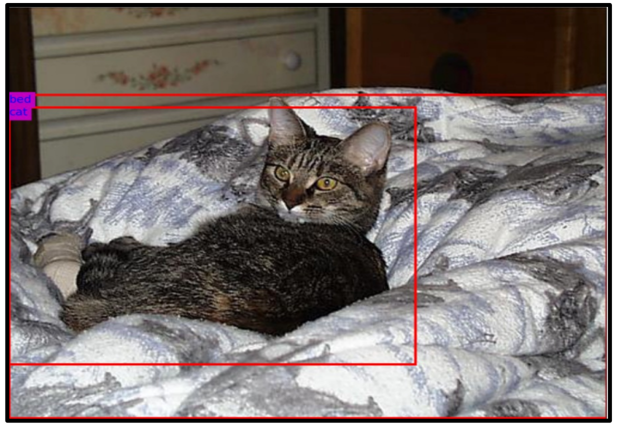


- Ablation studies on load balancing in MoE

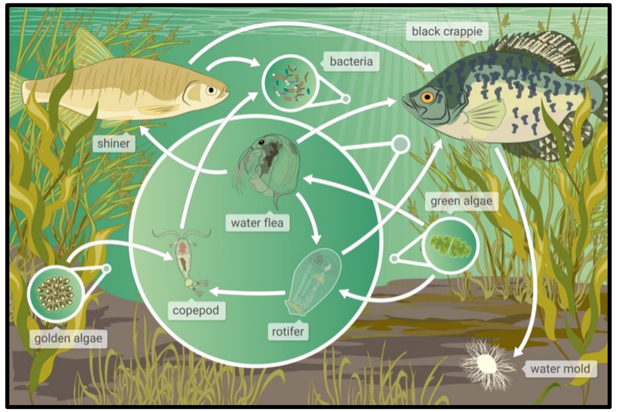
#Experts	LB	Token	Det.		VQA
			Recall	Prec.	Acc@1
4			39.04	35.21	46.95
4	✓		33.21	26.80	45.26
8			22.30	11.01	39.91
8	✓		21.52	12.10	37.53
4		✓	Fail		
4	✓	✓	Fail		

Visualization on 2D tasks

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

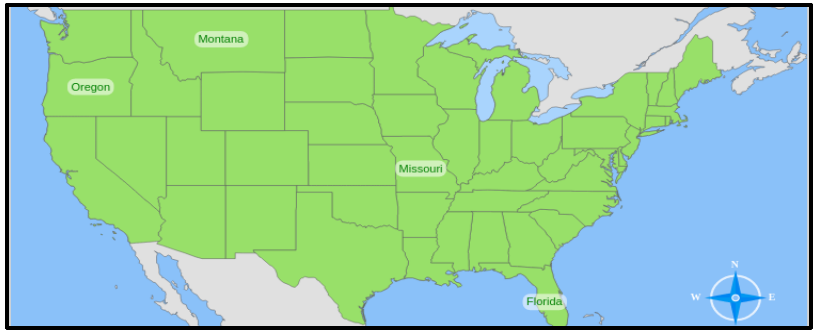


Question: Identify all objects in the image and provide their positions.
Answer: Positioned at [0.0, 0.24, 0.68, 0.86] within the image is an object that can be classified as belonging to the cat category, characterized by its unique texture and distinctive silhouette that differentiates it from other 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.



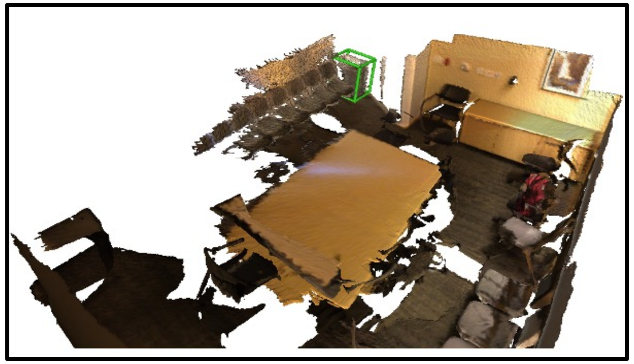
Question: Which of the following organisms is the primary consumer in this food web?
Context: Below is a food web from Little Rock Lake, a freshwater lake ecosystem in Wisconsin. A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter 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.

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.

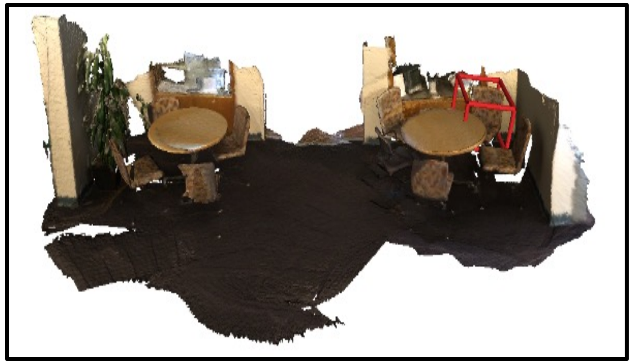


Question: Which of these states is farthest west?
Context: N/A
Choice: (A) Missouri (B) Florida (C) Montana (D) Oregon
Answer: After a thorough examination of the picture, it becomes apparent that the correct answer is (D) Oregon.

Visualization on 3D tasks



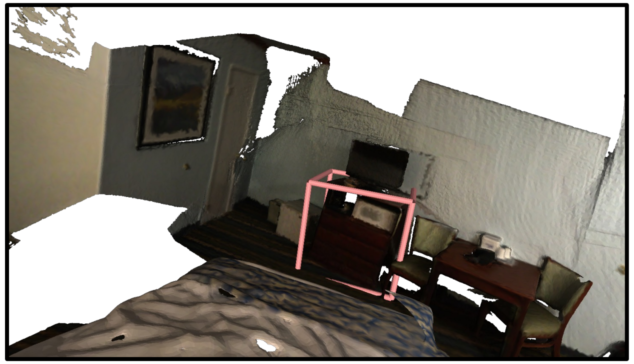
Question: What is the caption of the point cloud?
Answer: This is a trash can. it sets in the corner.



Question: What is located to the left of the display case?
Answer: A brown chair can be found in that position.

Question: What is the caption of the point cloud?
Answer: This is a black coffee table. it is in front of the couch.

Question: Which side of the monitor is the shelf positioned on?
Answer: The shelf is on the right side of the monitor.



Question: What is the shape of the object that is in contact with a wall surface?
Answer: The object in the scene is rectangular in shape.

Question: What is the color of the door located in the corner?
Answer: The door situated in the corner is brown in color.

Thanks for your watching!