TVDRNet: Text-driven Viewpoint Optimization via Differentiable Rendering for 3D Reasoning Segmentation

Zaiyang Yu[1], Changshuo Wang[2], Pinjie Xu[3], ZhangHuang[4], Yuan Shi[1,5], Linjun Sun[1,5], Weijun Li[1,5]

[1] Institute of Semiconductors, CAS, [2] UCL, [3] China University of Mining Technology - Beijing , [4] Beihang University, [5] UCAS

ICML International Conference On Machine Learning

1. Motivation

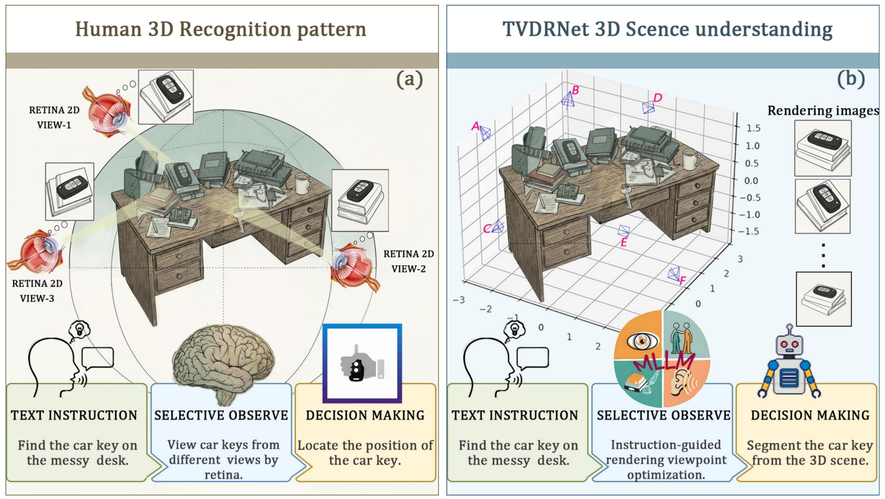

Figure 1. Humans have a visual pattern known as "**Active Vision**", in which select the optimal viewpoint to observe a target (**Left Figure**). Therefore, our original motivation for proposing TVDRNet was to explore whether the model could be guided to mimic this visual pattern—active vision—in order to focus on the visual regions most relevant to observation and reasoning tasks (**Right Figure**).

Human perception of the 3D world actively relies on selective observation. When asked to locate an item, humans do not arbitrarily scan the entire scene; instead, they selectively adjust their viewpoints based on the task instructions. Unfortunately, existing Multimodal Large Language Models (MLLMs) struggle with unstructured 3D point clouds, leading to severe erroneous localization and boundary ambiguity. To address these fundamental flaws, we draw inspiration from human active vision theory. By simulating this selective observation process computationally, an AI system can dynamically determine the most informative 2D virtual viewpoints. This mechanism ultimately establishes a complete and holistic 3D perception, significantly enhancing target localization and overcoming the limitations of static observation.

2. Introduction

Our task focuses on 3D reasoning segmentation, which aims to segment target objects based on natural language instructions and 3D spatial cues. The primary challenges are interpreting ambiguous text within complex scenes and handling the sparse, irregular geometry of 3D point clouds. To tackle this, we introduce TVDRNet, a framework employing a text-driven differentiable renderer. TVDRNet utilizes textual instructions to optimize rendering camera parameters, effectively teaching the model "where to look". This approach generates informative, task-relevant multi-view images. Consequently, it provides MLLMs with a holistic, unambiguous visual representation, directly mitigating the persistent challenges of erroneous target localization and ambiguous object boundaries in unstructured 3D environments.

3. Method

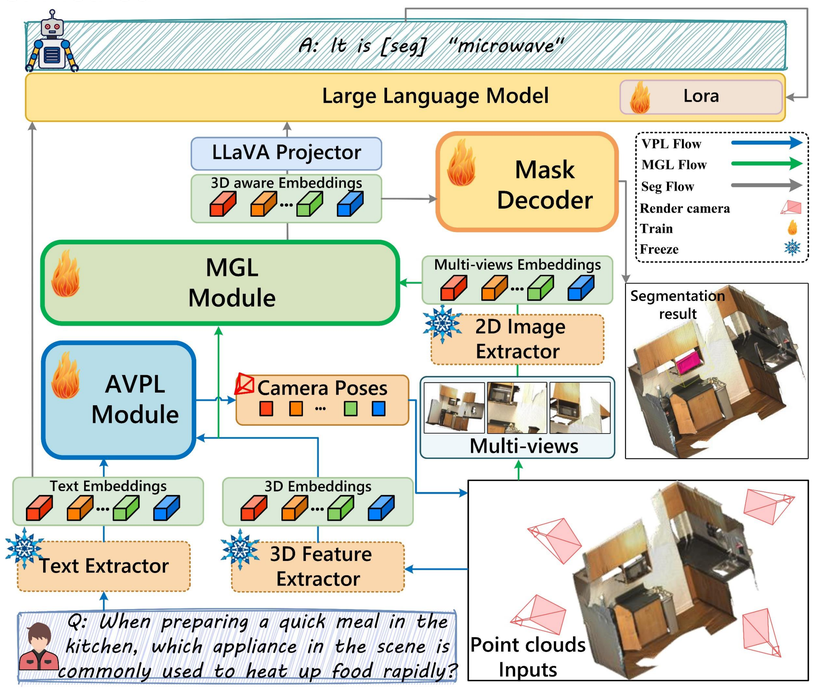

Figure 2.The design of TVDRNet.

To simulate active vision, TVDRNet relies on the Adaptive Viewpoint Position Learning (AVPL) module. The core insight is that optimal object viewing depends entirely on what the instruction requests. AVPL extracts features from input text and global 3D point clouds, feeding them into a multi-layer perceptron to map semantics directly to rendering camera parameters, such as focal length, azimuth, and elevation angles. These parameters guide a differentiable renderer to project the 3D point cloud into multi-view 2D images. Because the renderer is differentiable, gradients from the segmentation loss naturally flow back to adjust the camera parameters. This end-to-end optimization forces the network to spatially interpret text, iteratively discovering the absolute best geometric viewpoints.

Algorithm 1 The dataflow of TVDRNet

Input: Point Cloud $\mathcal{S} \in \mathbb{R}^{N\times 6}$; Text Instruction $\mathcal{T}_{ins}$
Output: Segmentation Mask $\mathcal{M} \in \{0,1\}^N$; Text Response $\mathcal{T}_{res}$
Stage 1: Adaptive Viewpoint Position Learning
Extract features: $\mathcal{T}_{emb} \leftarrow \mathcal{E}_T(\mathcal{T}_{ins})$, $\mathbf{F}_p \leftarrow \mathcal{E}_P(\mathcal{S})$
Predict rendering parameters $\mathbf{u} \leftarrow \mathcal{N}_{vp}(\mathcal{T}_{emb}, \mathbf{F}_p)$ (Eq. 1)
Decompose camera parameters (u_a, u_e, u_f) from $\mathbf{u}$ (Eq. 2)
Render multi-view images $\mathcal{M} \leftarrow \mathcal{R}(u_a, u_e, u_f, \mathcal{S})$ (Eq. 3)
Extract multi-view features $\mathbf{F}_v \leftarrow \mathcal{E}_I(\mathcal{M})$
Stage 2: Multi-modal Group Learning
Concatenate: $\mathbf{Z} \leftarrow [\mathbf{F}_p; \mathbf{F}_v] \in \mathbb{R}^{M\times D_p}$
Compute similarity matrix $\mathbf{S} \leftarrow \mathbf{Z}\mathbf{Z}^T$ (Eq. 4)
Compute and normalize coherence scores $\bar{s}_i$ (Eq. 5, 6)
Assign features to G groups based on score intervals
for each group $j = 1$ **to** G **do**
  Compute attention weights and aggregate prototype g_j (Eq. 7, 8)
end for
Compute global context c and inter-group attention β_j
Generate final feature $\mathbf{F}_{3Daware}$ (Eq. 9, 10)
Stage 3: Language Reasoning and Mask Prediction
Project features $\mathbf{F}_{3Daware\_proj} \leftarrow \text{Project}(\mathbf{F}_{3Daware})$
Generate response $\mathcal{T}_{res} \leftarrow \text{MLLM}(\mathcal{T}_{emb}, \mathbf{F}_{3Daware\_proj})$
Extract query h_{seg} and compute refined feature h'_{seg}
Generate mask $\mathcal{M} \leftarrow \sigma(\text{MLP}(h'_{seg}) \cdot \mathbf{F}^T_{3Daware})$
return $\mathcal{M}, \mathcal{T}_{res}$

Scan me

4. Experiments

In 3D reasoning segmentation, TVDRNet reached **43.92%** mIoU on Reason3D, surpassing the previous best model, OpenMaskDINO3D, by **4.11%**. On the Instruct3D benchmark, it similarly outperformed the leading MLLM-For3D with a **2.7%** mIoU gain. Furthermore, in the 3D visual grounding task on ScanRefer, TVDRNet achieved a dominant overall Acc@0.50 of **57.9%**, exceeding the Reason3D baseline by an impressive **16.0%** margin.

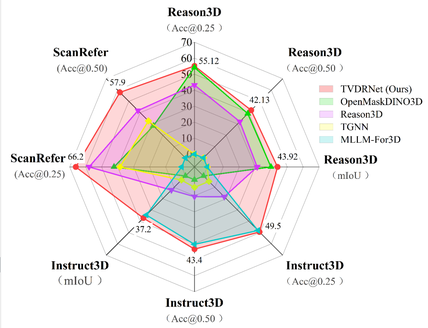

Figure 3.Comparision between recent efforts.

Table 1. 3D Reasoning Segmentation Results. The evaluation metric is accuracy at IoU 0.25, IoU 0.5 and mIoU.

Method	Venue	Input Modality	Reason3D			Instruct3D		
			Acc@0.25	Acc@0.50	mIoU	Acc@0.25	Acc@0.50	mIoU
OpenMaskDINO3D (Zhang, 2025)	Arxiv'25	RGB+3D+Text	54.21[#2]	39.14[#2]	39.81[#2]	-	-	-
MLLM-For3D (Huang et al., 2025a)	Arxiv'25	RGB+3D+Text	-	-	-	48.2[#2]	40.4[#2]	34.5[#2]
Reason3D (Huang et al., 2025b)	3DV'25	3D+Text	43.21[#3]	32.10[#3]	31.20[#3]	18.35	10.55	12.43
3D-STMN (Wu et al., 2024)	AAAI'24	3D+Text	25.43	17.78	18.23	-	-	-
OpenScene (Peng et al., 2023)	CVPR'23	3D+Text	24.68	7.14	15.03	-	-	-
OpenMask3D (Takmaz et al., 2023)	NeurIPS'23	3D+Text	20.78	6.82	13.38	-	-	-
LLM-Grounder (Yang et al., 2024)	ICRA'24	3D+Text	-	-	-	23.7[#3]	15.6[#3]	17.2[#3]
TGNN (Huang et al., 2021)	AAAI'21	3D+Text	-	-	-	4.76	4.76	3.51
†Avg	-	-	28.53	15.96	19.46	23.75	17.83	16.91
TVDRNet (Ours)	-	3D+Text	**55.12**[#1]	**42.13**[#1]	**43.92**[#1]	**49.5**[#1]	**43.4**[#1]	**37.2**[#1]

Table 2. 3D Visual Grounding Results on ScanRefer. The accuracy is evaluated by IoU 0.25 and IoU 0.5.

Method	Venue	Unique (∼19%)		Multiple (∼81%)		Overall	
		Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50
Task-specific reasoning							
UniVLG (Jain et al., 2025)	Arxiv'25	89.0[#2]	82.4[#3]	59.2[#2]	50.3[#2]	65.9[#2]	57.5[#2]
Chat-Scene (Huang et al., 2024a)	ECCV'24	88.9[#3]	80.1	54.2[#3]	48.6[#3]	62.0[#3]	55.7[#3]
AugRefer (Wang et al., 2025)	AAAI'25	86.2	70.8	50.0	39.1	55.7	44.0
X-RefSeg3D (Qian et al., 2024a)	AAAI'24	-	-	-	-	40.3	33.8
TGNN (Huang et al., 2021)	AAAI'21	69.3	57.8	31.2	26.6	38.6	32.7
ScanRefer (Chen et al., 2020)	ECCV'20	67.6	44.4	31.2	20.9	38.2	25.5
LMMs-based reasoning							
LLaVA-3D (Zhu et al., 2025a)	Arxiv'25	-	-	-	-	50.1	42.7
MCLN (Qian et al., 2024b)	ECCV'24	84.4	68.4	49.7	38.4	54.3	42.6
Reason3D (Huang et al., 2025b)	3DV'25	88.4	84.2[#2]	50.5	31.7	57.9	41.9
OpenMaskDINO3D (Zhang, 2025)	Arxiv'25	-	-	-	-	42.3	28.3
TVDRNet (Ours)	-	**89.2**[#1]	**84.7**[#1]	**60.3**[#1]	**51.2**[#1]	**66.2**[#1]	**57.9**[#1]

5. Conclusion

TVDRNet propose an active-vision paradigm for 3D reasoning segmentation. By integrating a differentiable rendering module, the model successfully learns "where to look" specifically based on what the textual instruction "asked to find". This dynamic viewpoint optimization process generates a semantically aligned visual representation, which, when effectively combined with raw point cloud features, delivers a powerful holistic, multimodal input to the reasoning model. TVDRNet sets a new standard on diverse benchmarks, proving highly robust at mitigating persistent issues like erroneous target localization and ambiguity in geometric boundaries.

6. Acknowledgement

The authors extend their profound gratitude to the anonymous reviewers for their highly constructive feedbacks.