Unifying 3D Vision-Language Understanding via Promptable Queries

Ziyu Zhu[1,2], Zhuofan Zhang [1,2], Xiaojian Ma[2], Xuesong Niu[2], Yixin Chen[2], Baoxiong Jia [2], Zhidong Deng[1] ✉, Siyuan Huang[2] ✉, Qing Li[2] ✉
[1]Tsinghua University [2]National Key Laboratory of General Artificial Intelligence, BIGAI, China

北京通用人工智能研究院 Beijing Institute for General Artificial Intelligence

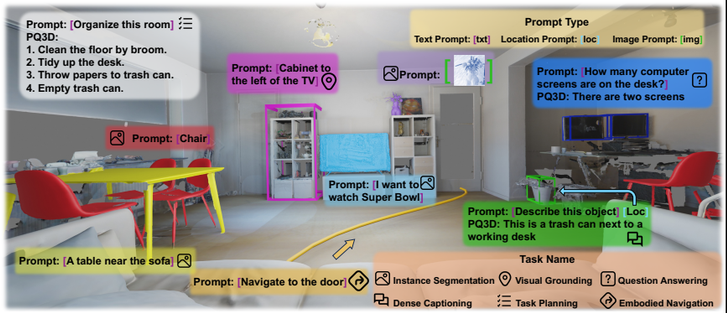
Project Page

Contribution

a) We introduce PQ3D, a **unified model** adept at solving a broad spectrum of 3D-VL tasks with **promptable queries**.

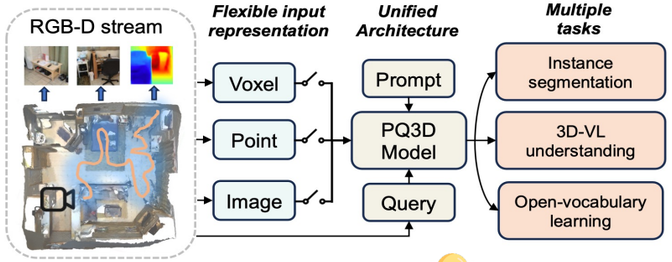
b) PQ3D aligns *voxels*, *point clouds*, and *multi-view images* into a **shared 3D space** for joint training.

c) PQ3D not only achieves **competitive results** but also sets new records across various 3D-VL tasks.



Highlight

Unifying multiple representations from RGB-D stream, including **Voxel**, **Point**, and **Multi-view images**.



Dataset	Task	Prompt	Heads	Size
ScanNet200 [52]	instance segmentation	category	mask,grounding	240K
ScanRefer [6]	visual grounding	sentence	grounding	37K
Nr3D [2]	visual grounding	sentence	grounding	119K
Sr3D [2]	visual grounding	sentence	grounding	66K
Multi3DRefer [66]	visual grounding	sentence	grounding	44K
ScanQA [3]	question answering	question	grounding,generation	30K
SQA3D [41]	question answering	question	generation	89K
Scan2Cap [9]	dense captioning	3D box	generation	37K
Total	-	-	-	662K

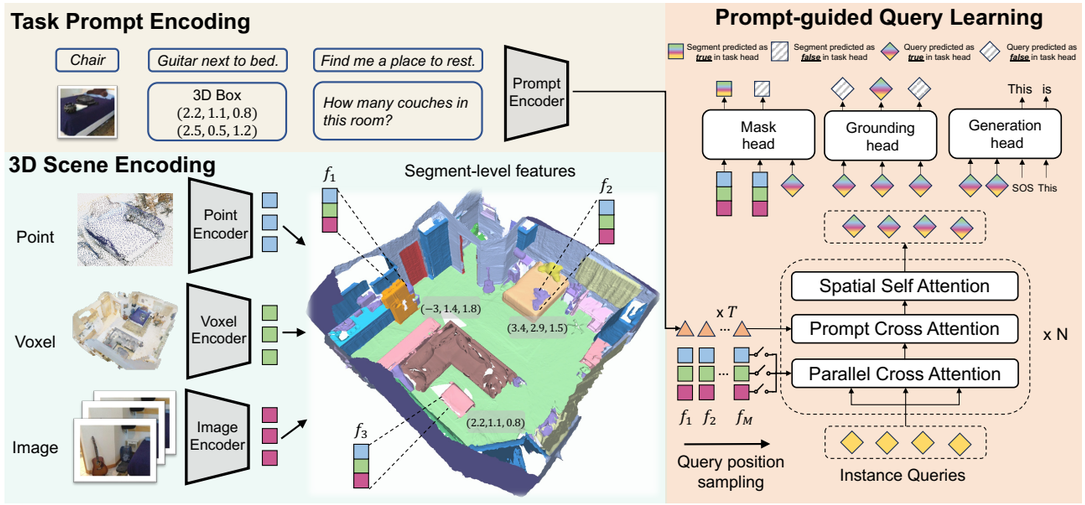
Unified training on **broad range** of tasks

Segmentation ✅ Refer ✅
Captioning ✅ QA ✅

PQ3D Model Design

a) Diverse prompt and output format for multi-task learning: Diverse prompts including text, image, and location are projected to a shared feature space. Instance queries can retrieve task-relevant information from prompts and be fed to different task heads for unified training.

b) Leverage multiple representations in one model: Point cloud, voxel grids, and multi-view images are aligned to segment level with shared coordinate space.



a) Task Prompt Encoding

CLIP text, image encoder for VL prompts

MLP for encoding location prompt

b) 3D Scene Encoding

Point cloud: PointNet++ feature for each point

Voxel: Apply SparseConvUNet to voxel grid.

Multi-view image: OpenSeg pixel feature

c) Prompt-guided Query Learning

Parallel cross attention between queries and features.

Cross attention between queries and prompts.

Different task heads for unified learning.

Mask $p_{\text{mask}} = \sigma(f_s(\mathbf{V}+\mathbf{I}+\mathbf{P}) \cdot f_q(\mathbf{Q})^T)$

Grounding $p_{\text{grd}} = \sigma(f_g(\mathbf{Q}))$

Generation T5 small autoregressive

$$\mathbf{Q}_l' = \text{FFN}(\text{Norm}(\mathbf{Q}_l + \sum_{\mathbf{F}\in\{\mathbf{V},\mathbf{I},\mathbf{P}\}} \text{MaskedCrossAttn}(\mathbf{Q}_l, \mathbf{F}))),$$

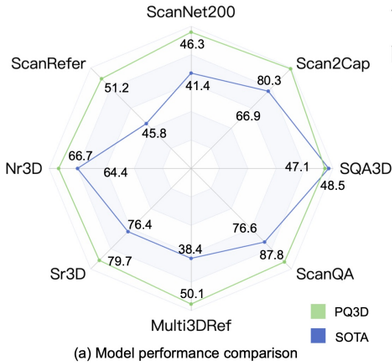
$$\mathbf{Q}_l'' = \text{FFN}(\text{Norm}(\mathbf{Q}_l' + \text{CrossAttn}(\mathbf{Q}_l', \mathbf{t}))),$$

$$\mathbf{Q}_{l+1} = \text{FFN}(\text{Norm}(\text{SpatialSelfAttn}(\mathbf{Q}_l''))).$$

$$\text{Attn}(Q,K,V,S) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_h}} + \log\sigma(Sw)\right)V,$$

Results and Insights

a) PQ3D demonstrates superior performance on most tasks, from low-level segmentation to high-level reasoning.



(a) Model performance comparison

Instance segmentation

Model	ScanNet200 AP	AP_{50}	AP_{25}	head	common	tail
Closed-vocabulary						
Mask3D [62]	26.9	36.2	41.4	**39.8**	21.7	17.9
PQ3D (*w/cls*)	**27.0**	**38.9**	**46.3**	35.8	**24.2**	**20.0**

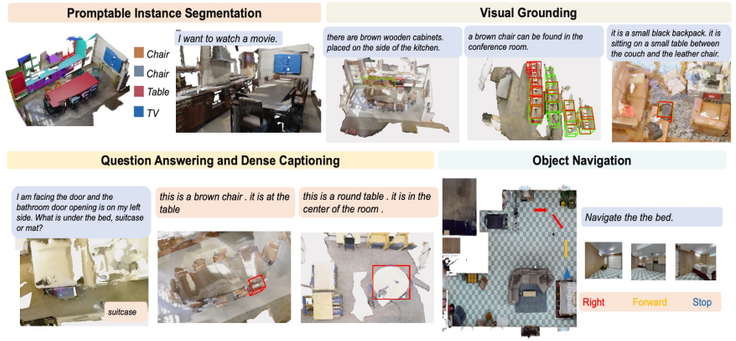
Visual grounding

Method	ScanRefer Unique	Multiple	Avg.	Nr3D Easy	Hard	Avg.
UniT3D [13]	73.1	31.1	39.1	-	-	-
M3DRef-CLIP [74]	77.2	36.8	44.7	55.6	43.4	49.4
3D-VisTA [79]	75.1	39.1	45.8	72.1	56.7	64.2
PQ3D (*sg.*)	76.6	42.0	47.4	73.3	56.7	64.9
PQ3D	**78.2**	**46.2**	**51.2**	**75.0**	**58.7**	**66.7**

Object navigation

Model	Success ↑	SPL ↑	Soft-SPL ↑
VC-1 (ViT-B) [50]	57.1	0.31	0.41
PQ3D	**80.0**	**0.50**	**0.60**
PQ3D w/o GPS*	75.0	0.45	0.50

b) PQ3D is a versatile model can take various prompts and solve a broad range of tasks in a flexible way.



c) Unifying multiple representations and tasks benefits 3D vision language understanding.

V	P	I	Refer	QA	Caption
✓			46.1 / 47.1	43.7 / 44.2	67.8 / 68.1
✓	✓		49.2 / 49.4	45.4 / 45.8	74.6 / 74.7
✓	✓	✓	51.2	47.1	80.3

Task	Data	Refer	QA	Caption
+Refer	-	1.8 ↑	2.2 ↑	
+QA	0.0 ↑	-	1.2 ↑	
+Caption	0.6 ↑	0.7 ↑	-	