CAPruner: Conceptual-Adjacent Scene Graph Pruner for Enhancing 3D Spatial Reasoning of Large Language Models

Shengli Zhou, Xiangchen Wang, Guanhua Chen✉, Feng Zheng✉
Southern University of Science and Technology

3D Vision-Language (3D VL) Tasks

- ▶ Input 3D point cloud + text and output textual response.
- ▶ Core: spatial reasoning, locate targets using spatial relations.

Examples:
[**3D Visual Grounding**] The bed next to the right of the nightstand and next to the chair. [Output] <obj002>
[**3D Visual Question-Answering**] What is the color of the chair next to the bed in the corner of the room? [Output] Blue.

Encoding Relations for LLMs

- ▶ Mainstream solution for 3D VL tasks:
  - ▪ Represent scene layout for LLM inference.
- ▶ Common data structure: scene graph.
  - ▪ Objects as nodes, relations as edges.
- ▶ **Challenge**: n objects yields $O(n^2)$ relations
  - ▪ Key spatial relation information is sparse.
  - ▪ $O(n^2)$ input tokens for the LLM, does not scale.

How to select O(n) key spatial relations?

Methodology

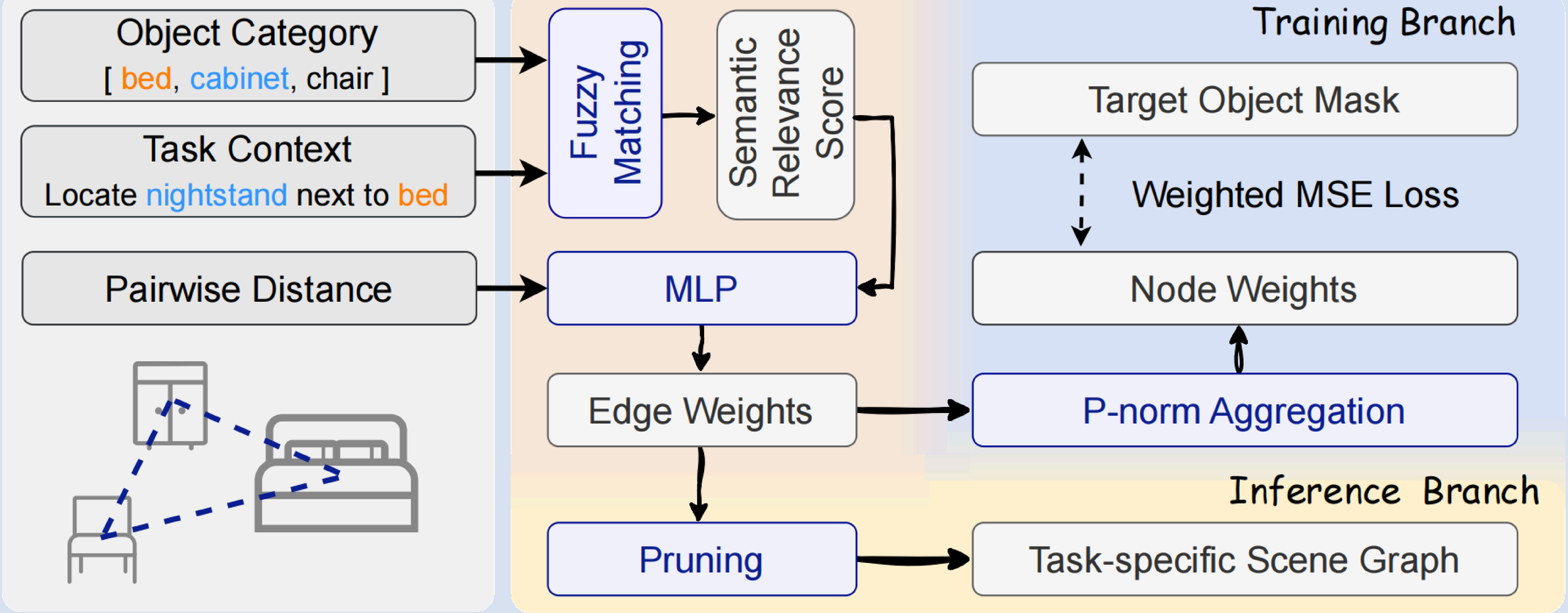
Core ideas: (1) **Semantic and spatial factors** should be considered for evaluating spatial relation importance.
(2) Instead of choosing an exact set of spatial relations used for inference, **choose its superset with O(n) size**.
→ **Fuzzy matching** enables approximate importance estimation with lower computation cost.

Semantic factor: let t_j be the j-th token in the input text, C_k be the k-th NYU40 category, the semantic relevance score s_i for object i with

$$s_i = \begin{cases} 1 & \exists_{j,k}(c_j \in C_k \ \wedge \ t_j \in C_k) \\ 0 & \text{otherwise} \end{cases}$$

Training Branch:

- ▶ Use p-norm to aggregate edge weights to nodes.
- ▶ Supervise target objects to have larger node weights.
- ▶ Aggregation: avoid $O(n^2)$ relation level annotation.



Spatial factor: proximity (according to Maxim of Relation).

Inference branch: retain edges with larger weights.

Comparative Experiment

Table 1. CAPruner provides key spatial relations to the LLM better, achieving consistent gains on mainstream 3D VL benchmarks.

Model	ScanRefer				ScanQA	SQA3D
	A.@0.25	A.@0.5	M.@0.25	M.@0.5	BLEU-4	EM@1
FFL-3DOG (Feng et al., 2021)	41.3	34.0	35.2	25.7	–	–
SeeGround (Li et al., 2025b)	44.1	39.4	34.0	30.0	–	–
CSVG (Yuan et al., 2025)	49.6	39.8	38.4	27.3	–	–
AugRefer (Wang et al., 2025)	55.7	44.0	50.0	39.1	–	–
MA2TransVG (Xu et al., 2024)	57.9	45.7	53.8	41.4	–	–
3D-VisTA (Zhu et al., 2023)	50.6	45.8	43.7	39.1	13.1	48.5
TSP3D (Guo et al., 2025)	56.5	46.7	–	–	–	–
Scene-LLM (Fu et al., 2025)	–	–	–	–	12.0	54.2
Chat-Scene-7B (Huang et al., 2024a)	55.5	50.2	–	–	14.3	54.6
PQ3D (Zhu et al., 2025)	–	51.2	–	46.2	–	47.1
QuatRoPE (Zhou et al., 2026)	58.2	52.5	54.3	49.2	–	55.2
3DGraphLLM-1B (Zemskova and Yudin, 2024)	52.5	47.5	45.0	40.5	12.2	52.6
CAPruner + Llama-3.2-1B (Ours)	55.0	49.6	48.0	42.8	13.0	52.8
3DGraphLLM-8B (Zemskova and Yudin, 2024)	60.2	54.6	*54.7*	*49.4*	12.5	55.2
CAPruner + Llama-3-8B (Ours)	61.7	56.0	55.3	49.9	13.2	56.3

Ablation Studies

Table 2. On downstream LLMs, CAPruner achieves better accuracies than proximity-based KNN.

Pruning Method	ScanRefer				ScanQA		SQA3D		
	A.@0.25	A.@0.5	M.@0.25	M.@0.5	B.-3	B.-4	EM@1	ROUGE	CIDEr
Proximity-based KNN	52.5	47.5	45.0	40.5	17.9	12.2	52.6	53.8	138.3
CAPruner (MST)	54.4	49.0	47.1	42.0	18.1	11.7	52.4	53.7	139.0
Gain	1.9	1.5	2.1	1.5	0.2	-0.5	-0.2	-0.1	0.7
CAPruner (KNN)	55.0	49.6	48.0	42.8	18.5	13.0	52.8	54.1	139.1
Gain	2.5	2.1	3.0	2.3	0.6	0.8	0.2	0.4	0.8

Project Page Paper Code