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

Shengli Zhou[1], Xiangchen Wang[1], Guanhua Chen[1]✉, Feng Zheng[1]✉
[1]Southern University of Science and Technology

{zhousl2022, wangxc2019}@mail.sustech.edu.cn, chengh3@sustech.edu.cn, f.zheng@ieee.org

3D Vision-Language (3D VL) Tasks

ACL 2026 SAN DIEGO

- Input: point cloud (3D scene) + text.
- Output: text (response).
- Core ability: spatial reasoning, locate target objects using spatial relations.

- Examples:

- [3D Visual Grounding] Locate 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?
- [Output] Blue.

Challenge: Scene Representation

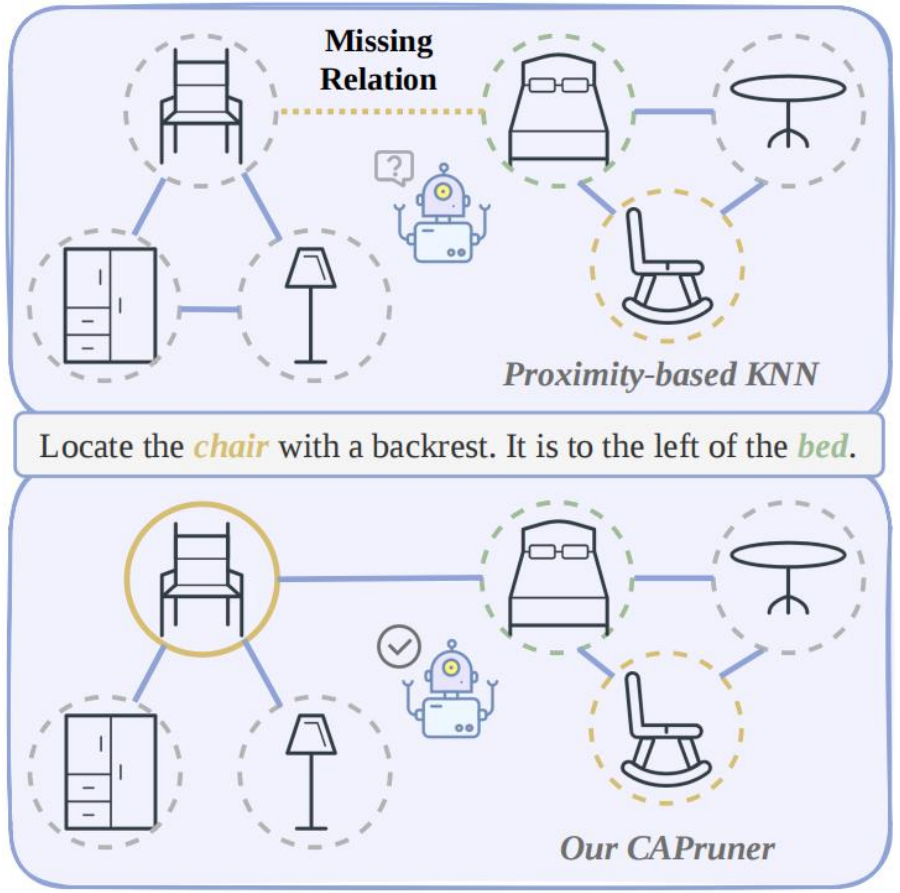
- Mainstream solution for 3D VL tasks: represent scene layout to LLMs for inference.
- Data structure: scene graph, objects as nodes, relations as edges.
- Challenge: encoding pairwise spatial relations yields $O(n^2)$ relations for n objects
  - Information for key spatial relations is sparse.
  - $O(n^2)$ input tokens for the LLM, does not scale.

How to select O(n) key spatial relations for the model?

Core Idea 1: Semantic + Spatial

How to select O(n) key spatial relations for the model?

- 3DGraphLLM [1]: Proximity-based KNN
  - Retain the relation between each object and its two nearest objects
  - But proximity != importance
- Each 3D VL query focuses on different relations
  - Context of the query should be considered
  - **Importance = Semantic + Spatial Relation**



[1] Tatiana Zemskova and Dmitry Yudin. 2024. 3dgraphllm: Combining semantic graphs and large language models for 3d scene understanding. *Preprint*, arXiv:2412.18450.

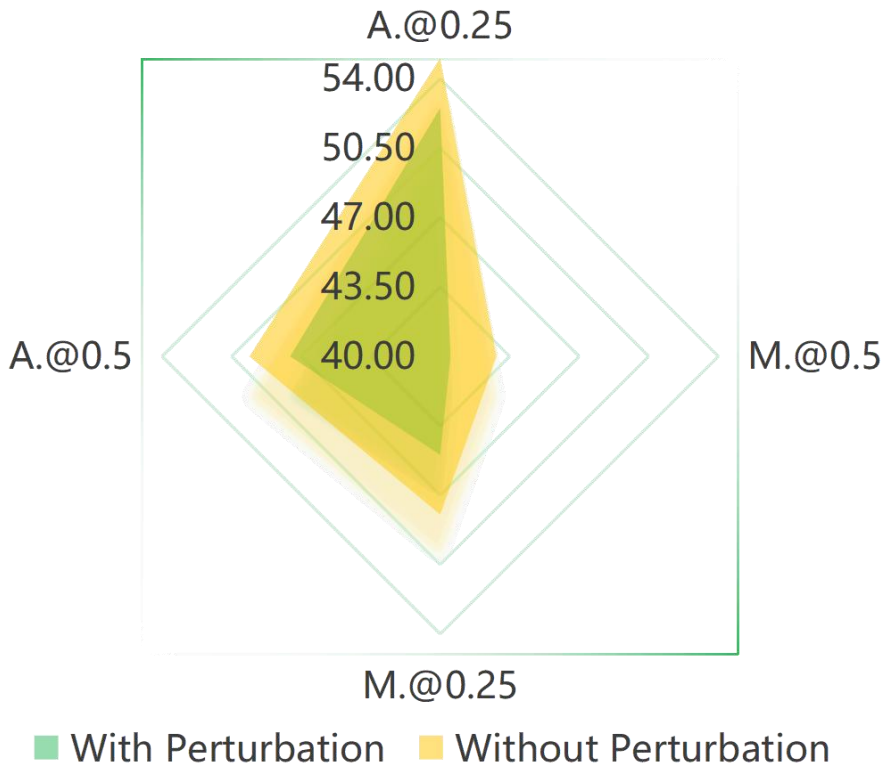
Core Idea 2: Select a Superset

How to select O(n) key spatial relations for the model?

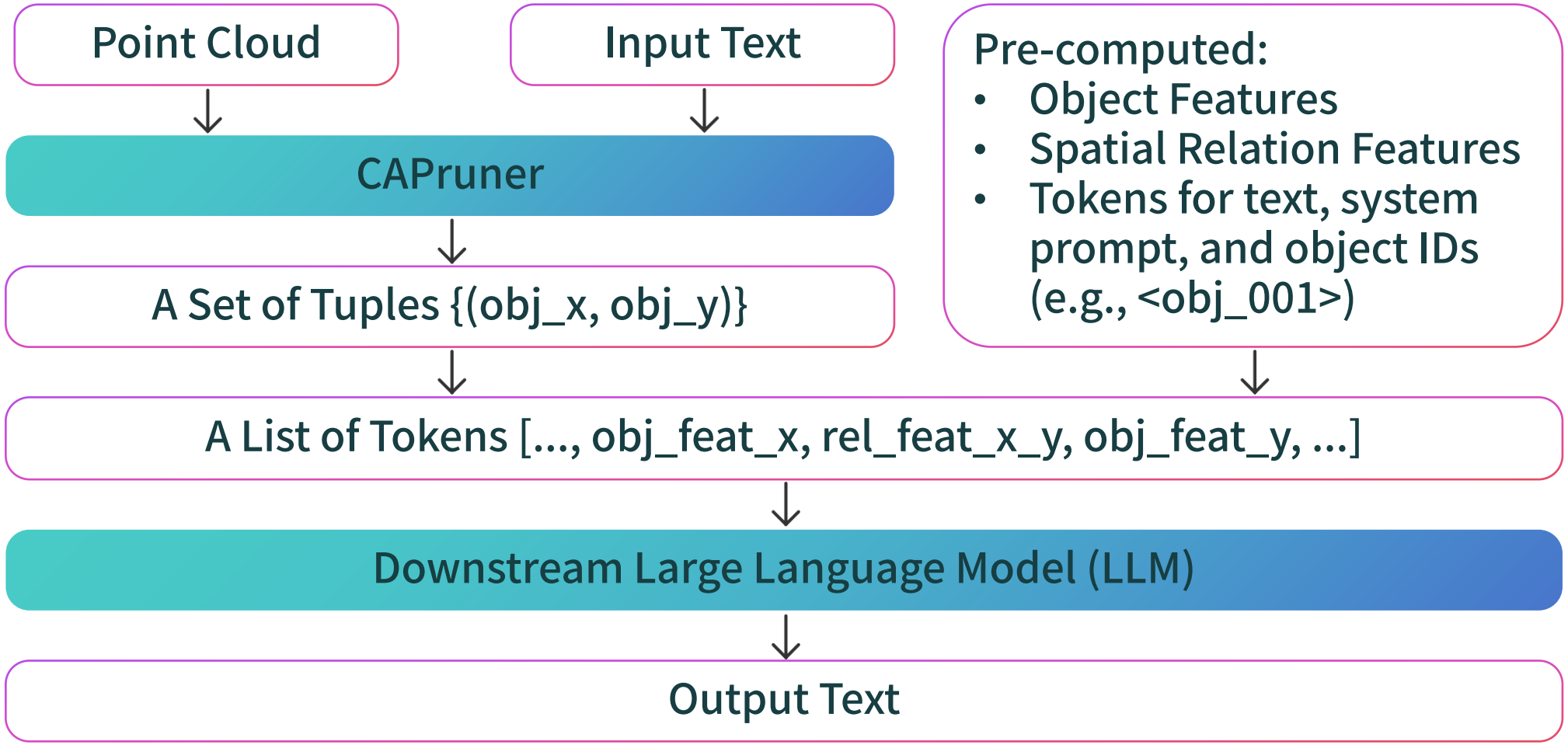
Experiment: replacing key spatial relations with irrelevant ones degrades accuracy

Key relations for inference should be retained

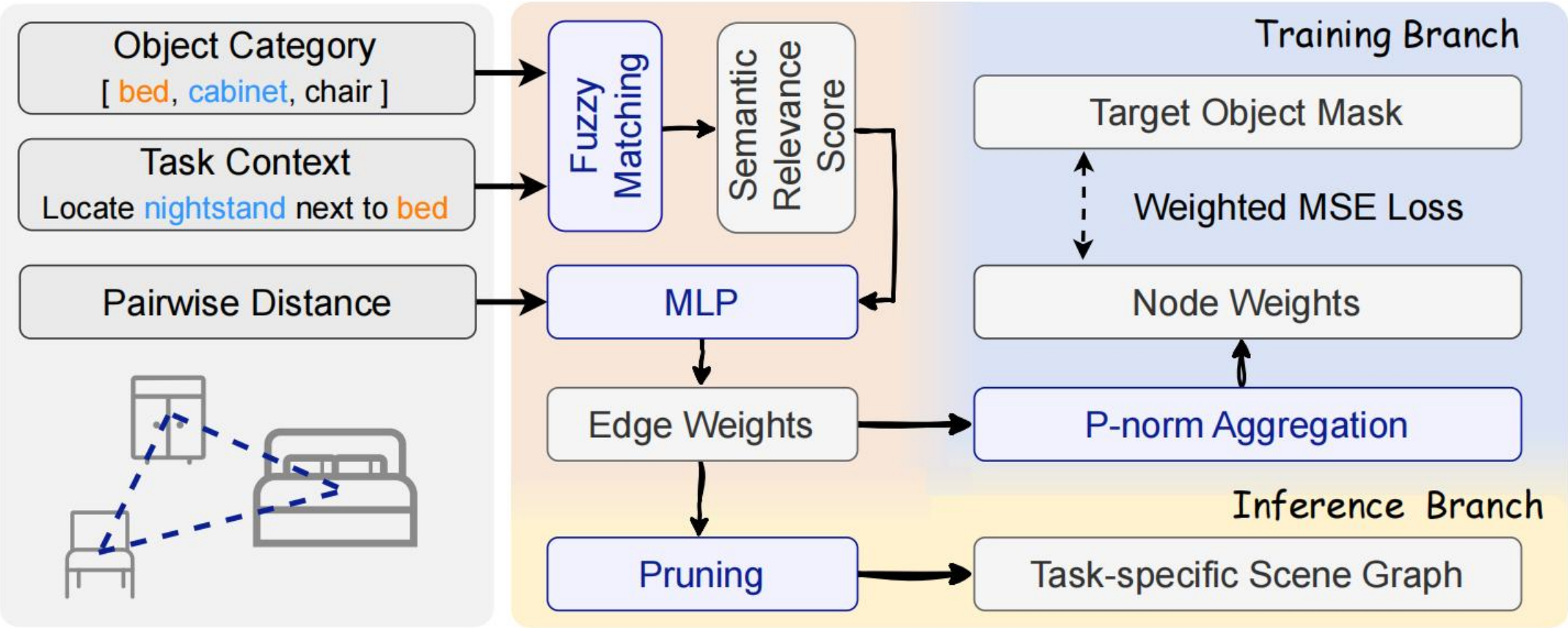
Goal: choose a set of relations with O(n) size that contains all key spatial relations



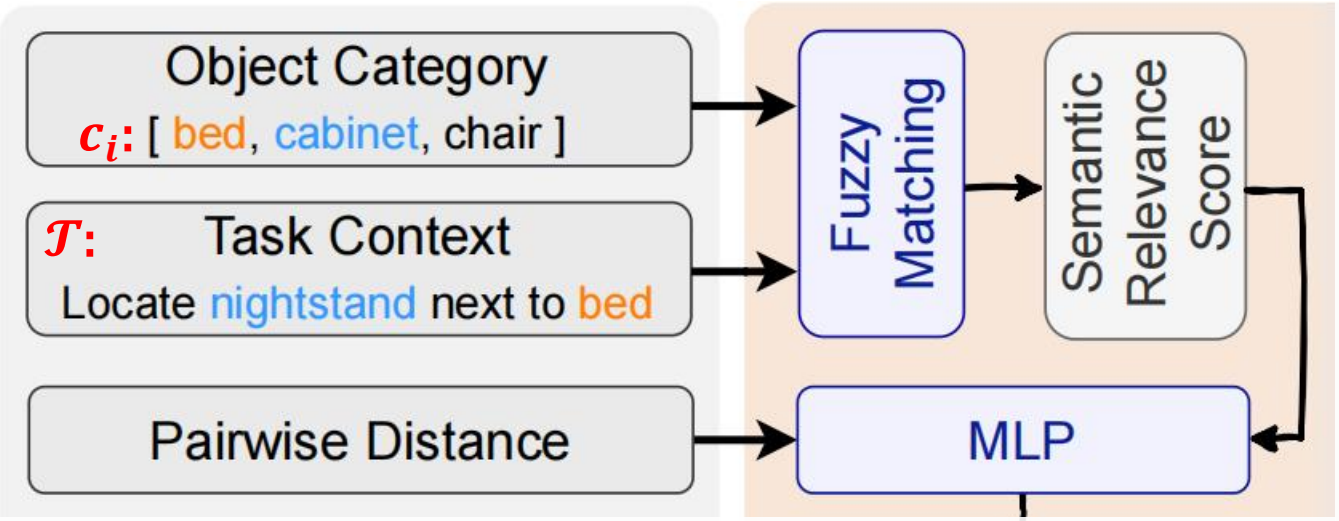
Pipeline Overview



Model Architecture



Model Architecture

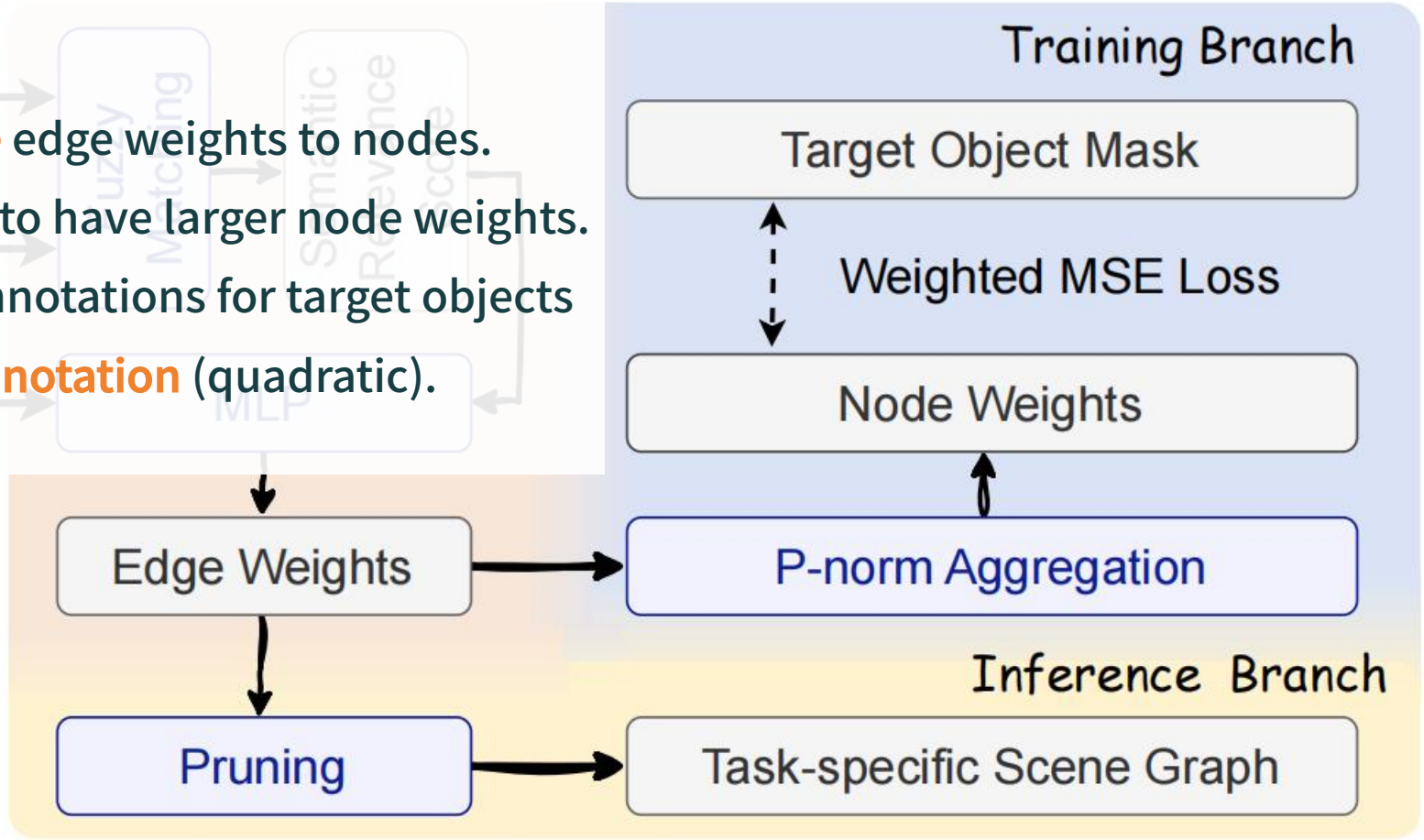


- Only need to select a superset → roughly estimate importance to lower overhead → fuzzy matching.
- **Semantic: fuzzy matching**
  - $s_i = \max_{t \in \mathcal{T}} \{\text{Similarity}(c_i, t)\}$

    where $\mathcal{T}$ is the input token set.
  - The similarity score for an object is 1 if there exists some word t in the input text that belongs to the same category c_i (e.g., bed, cabinet).
  - Otherwise the score is 0.
- **Spatial: Proximity**
  - Maxim of Relation: consider nearby objects first.
  - E.g., the window to the left of the door = the nearest window to the left of the door.

Model Architecture

- Training Branch:
  - Use p-norm to aggregate edge weights to nodes.
  - Supervise target objects to have larger node weights.
  - Why aggregation? Use annotations for target objects to avoid relation level annotation (quadratic).
- Inference Branch:
  - Pruning: retain edges with larger weights.
  - Output: a set of tuples.



Metrics for Experiments

- ScanRefer (3D Visual Grounding): Acc @ 0.25, Acc @ 0.5, Multi @ 0.25, Multi @ 0.5
  - The accuracy on the entire dataset / a subset with other objects of the same type as the target object (which requires spatial reasoning).
  - The output is regarded correct when its IoU with ground truth exceeds 0.25 / 0.5.
- ScanQA (3D Visual Question-Answering): BLEU-4
  - Measures sentence similarity by calculating overlapping 1-4 word phrases.
- SQA3D (3D Visual Question-Answering): EM @ 1
  - Accuracy, correct when top-confidence output exactly matches ground truth.

Comparative Experiment

ACL 2026 SAN DIEGO

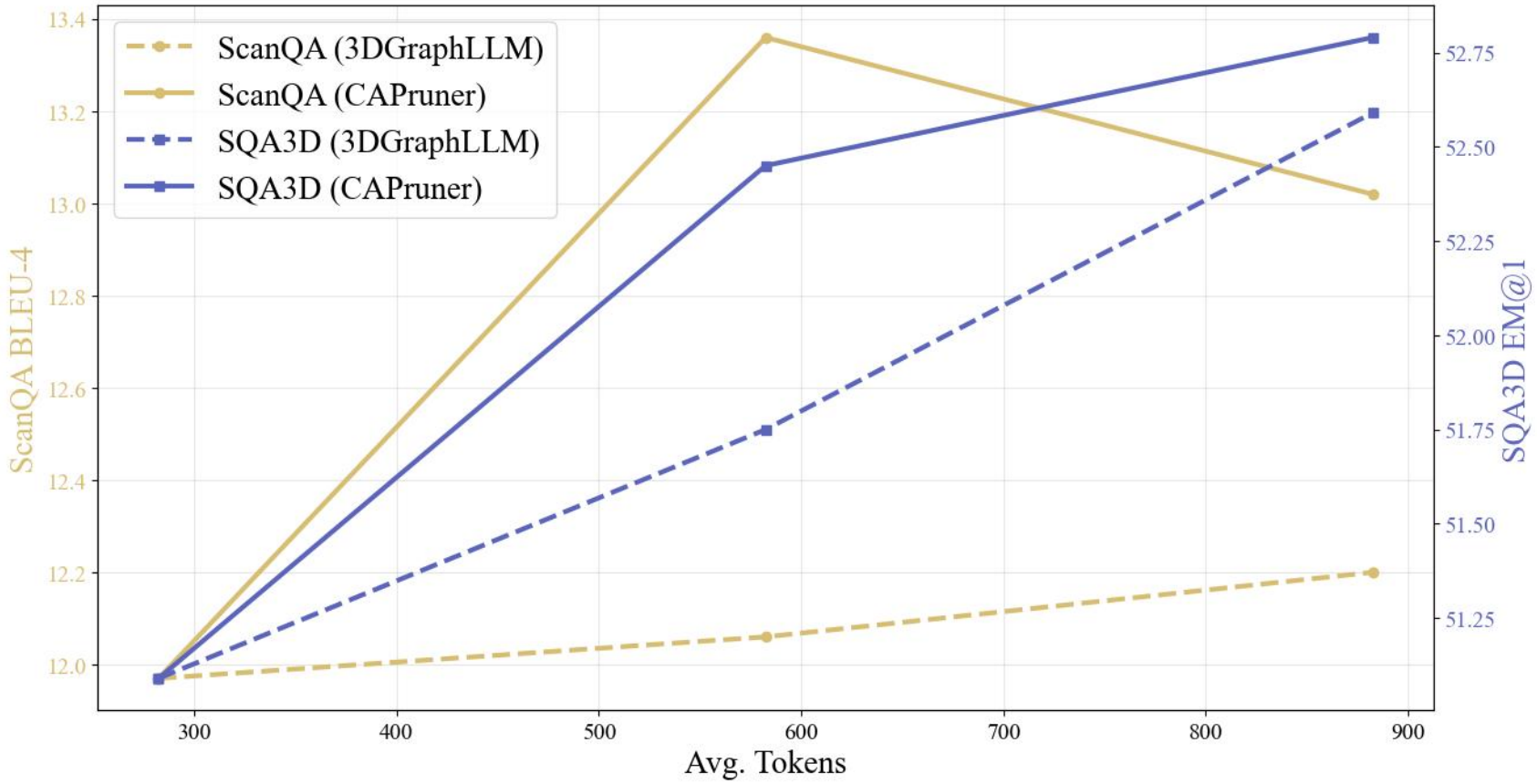
- CAPruner better provides key spatial relations to the LLM, achieving consistent gains.

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

Table 1: Comparison on ScanRefer, ScanQA, and SQA3D. With the same backbone LLM, CAPruner consistently outperforms base methods and achieves the strongest or highly competitive results. A. and M. denote accuracy on the overall and "multi" splits, respectively. Scores for 3DGraphLLM-1B and in italic are evaluated on our machine.

Saving Tokens

- Higher token efficiency: achieves **better accuracy with 34% fewer input tokens**.



Pruning Method Comparison

ACL 2026 SAN DIEGO

- CAPruner achieves better accuracies than proximity-based KNN on downstream LLMs.
  - Validating the importance of considering semantic information for pruning.

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

Table 2: Comparison of pruning policies after learning CAPruner edge weights. Applying KNN to the learned scores performs best. A. and M. denote accuracy on the overall and "multi" splits, respectively; B.-3 and B.-4 denote BLEU-3 and BLEU-4. All models are based on Llama-3.2-1B.

Conclusion

- We derive scene-graph pruning requirements when using LLMs for spatial reasoning.
- We propose CAPruner, a lightweight scene-graph pruning model.
  - Considers both semantic relevance and spatial proximity.
  - Aggregates edge weights to nodes to lower annotation cost.
- We validate the pruning rationale through extensive experiments.
  - CAPruner achieves consistent gains with fewer tokens.

Thank you!

Project Page

Paper

Code