



Object Navigation based on multimodal scene representation for mobile robots

Tatiana Zemskova

Researcher, Cognitive AI Systems Lab, AIRI
PhD student, MIPT

Agenda

- 01 Task description
- 02 Segmentation as auxiliary task for navigation
- 03 Active perception for navigation
- 04 Scene Graphs for 3D Referred Object Grounding
- 05 Conclusion

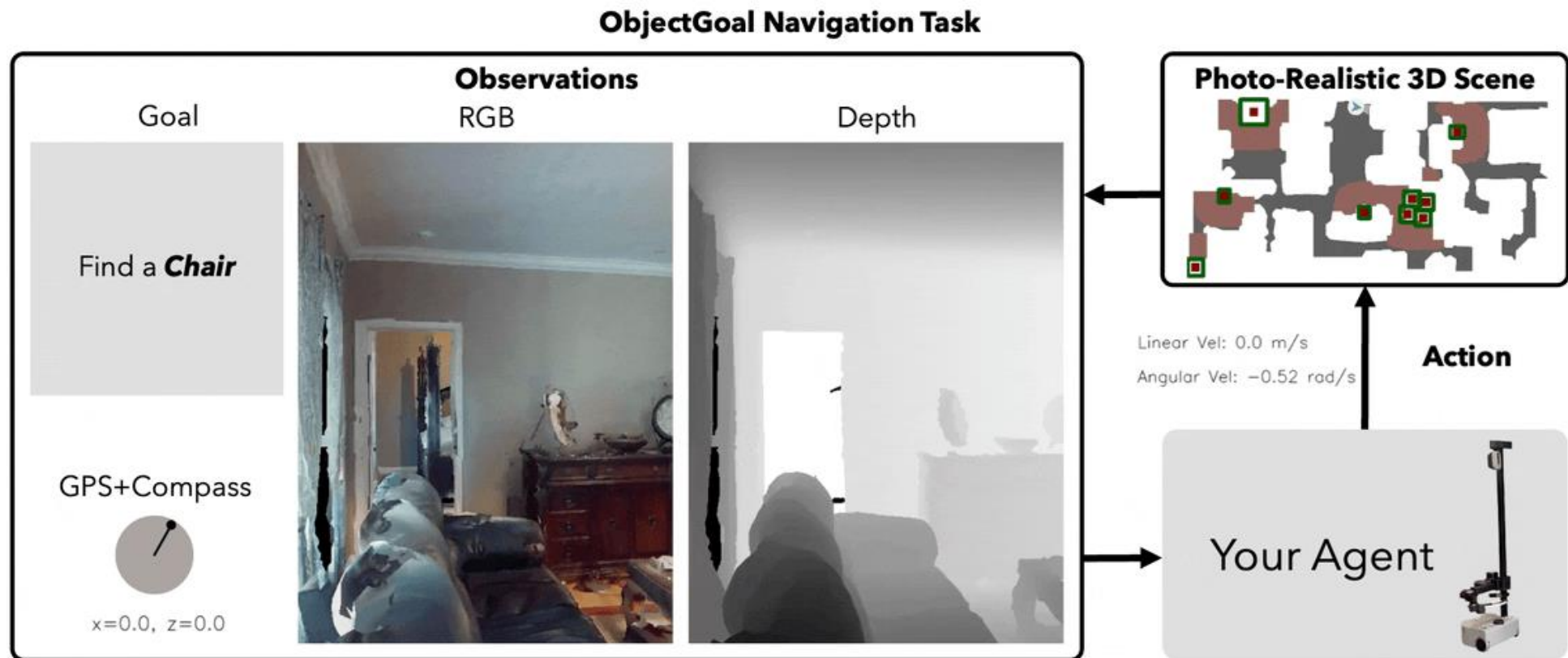
01



Task description

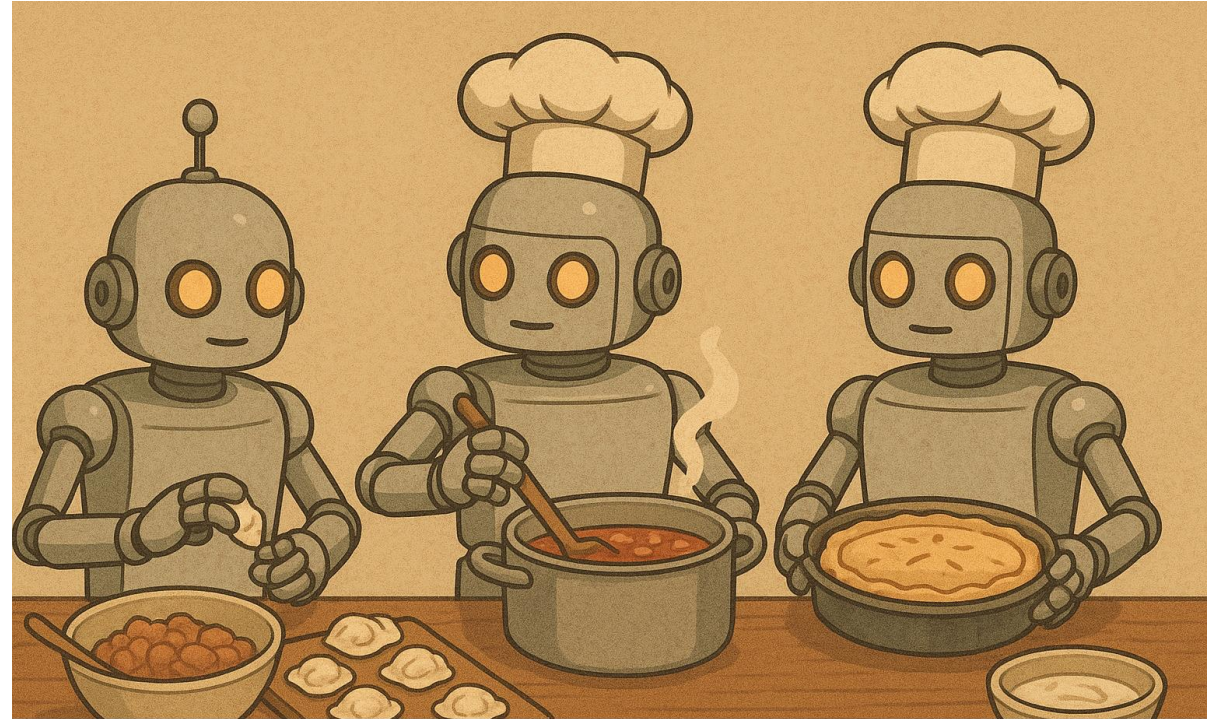
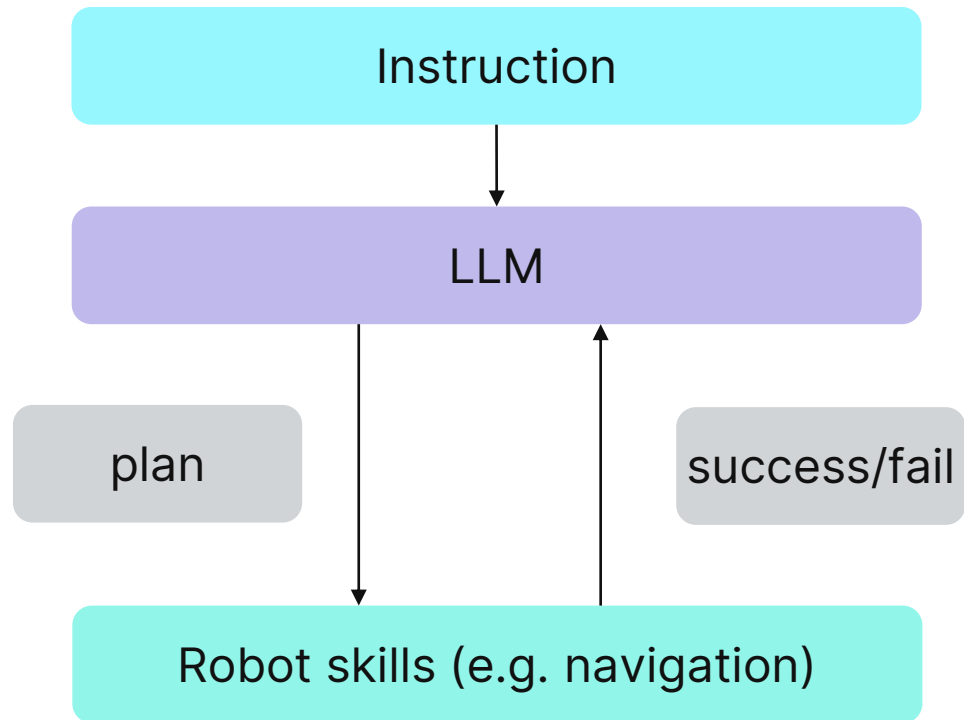
Task description (ObjectGoal Navigation)

Task: Navigate within unfamiliar environment to the target object described by text.



The navigation task as a part of complex robotics tasks

The navigation task is a subtask of mobile manipulation.

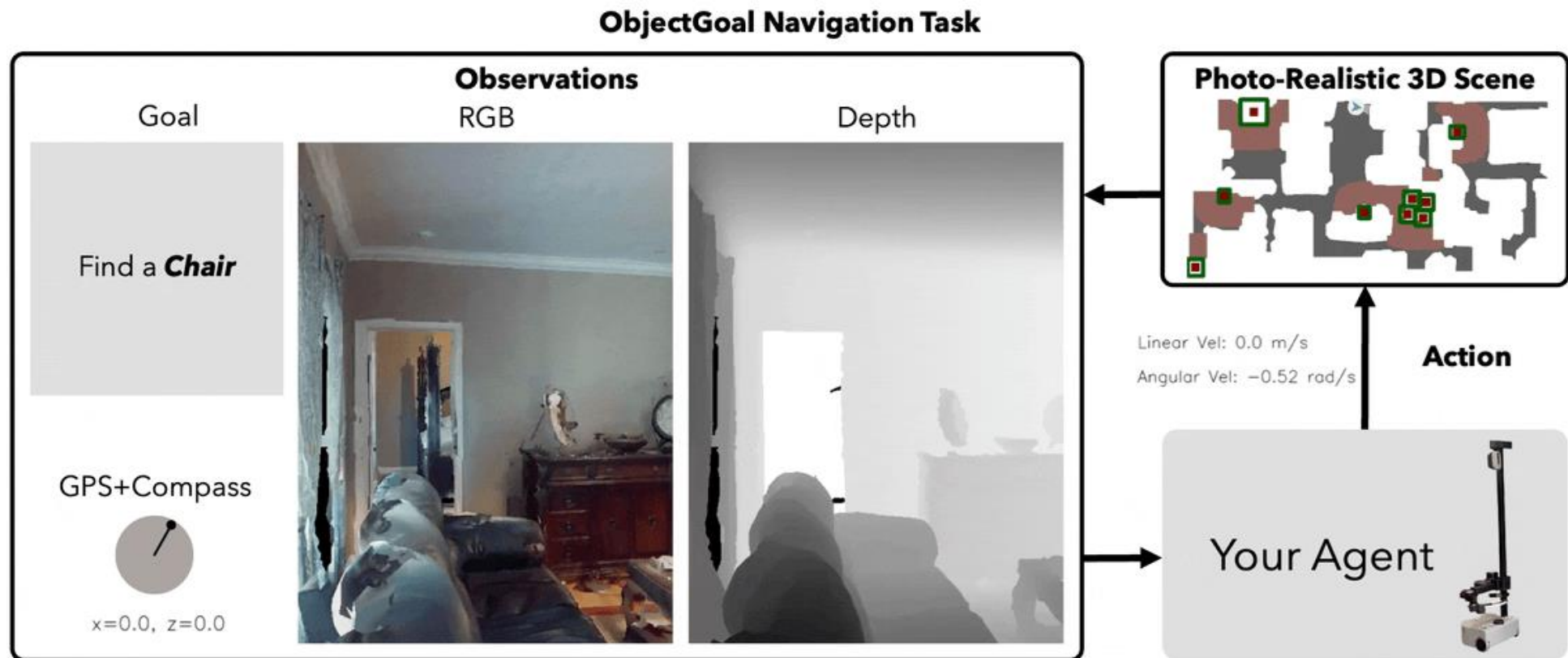


02

Segmentation as auxiliary task
for navigation

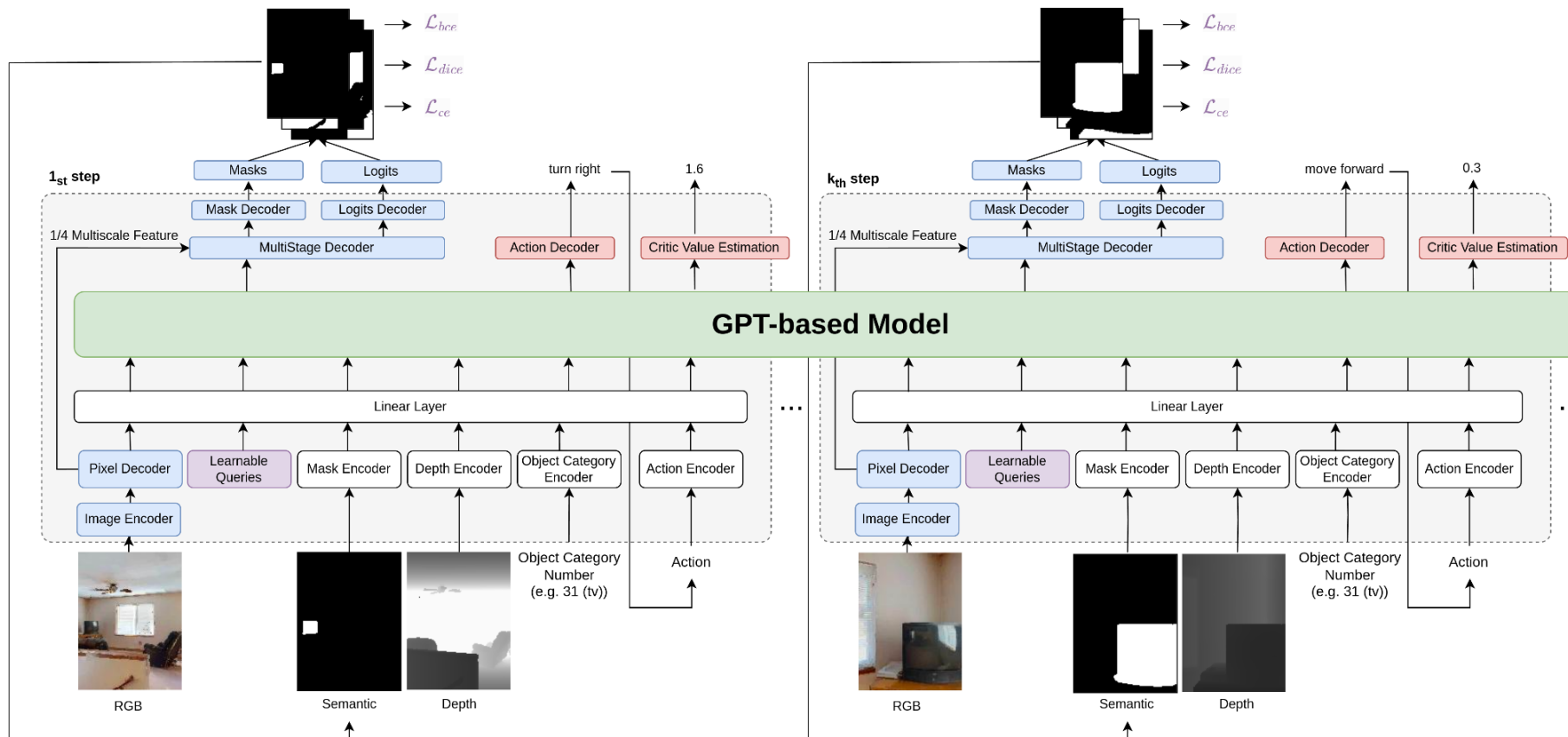
Task description (ObjectGoal Navigation)

Task: Navigate within unfamiliar environment to the target object described by text.



Segmenting Decision Transformer (SegDT)

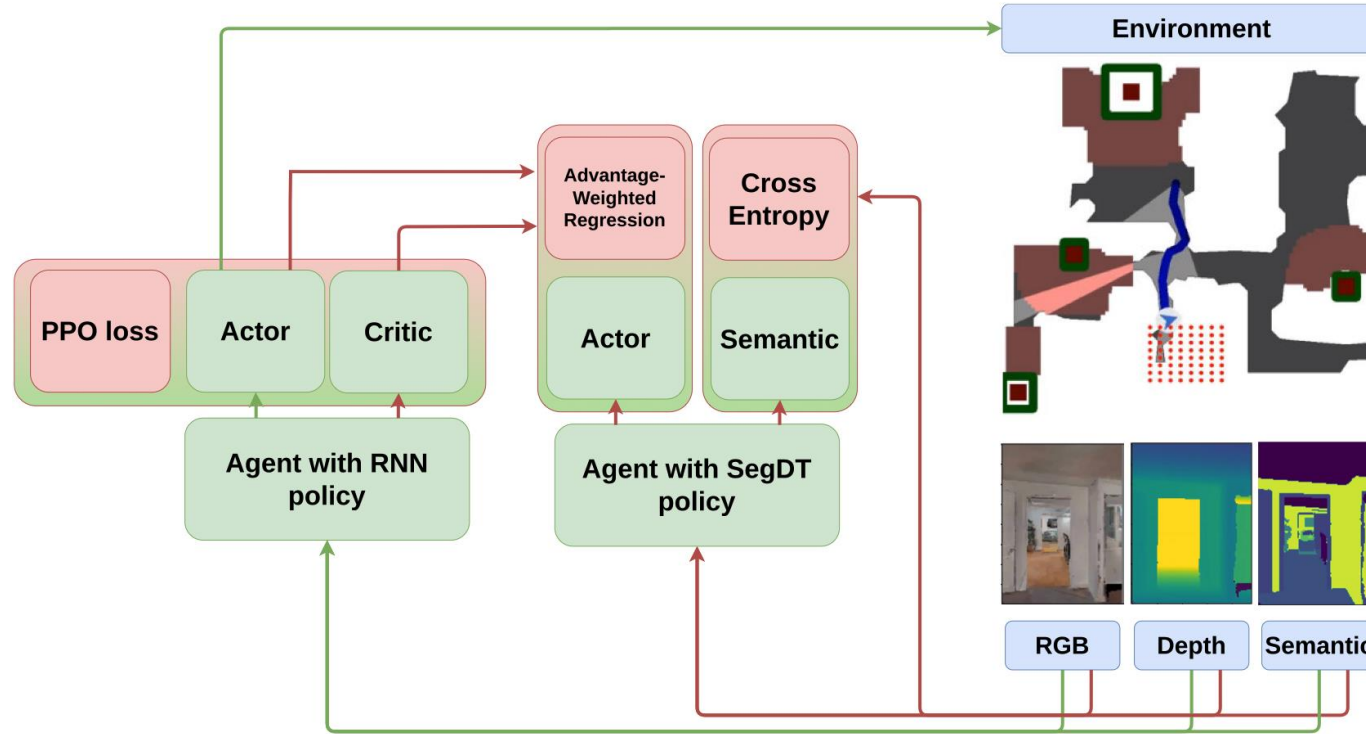
Architecture for jointly solving semantic segmentation and navigation tasks (reaching the target object).



Code SegDT
Github 

Online fine-tuning of SegDT

- An RNN-based strategy using GT segmentation generates expert trajectories that guide SegDT via an advantage-weighted regression loss.
- This knowledge transfer mechanism enables efficient learning of both segmentation and navigation while maintaining computational efficiency.



Quantitative Results

- SegDT achieves state-of-the-art results in the goal reaching task, demonstrating strong performance under a limited camera field of view and without using localization for navigation.

Integration of SegDT into the pipeline for solving the full target-object navigation task

Exploration	GoalReacher	SR	SPL	SoftSPL
PIRLNav	PIRLNav	61.9	26.0	28.3
PIRLNav	SegDT	63.6	30.3	32.6

Method	Sensors	FOV	Turn Angle	SR	SPL	SoftSPL
DD-PPO (500 steps)	RGB-D, GPS+Compass	79°	30°	10.2	2.1	14.6
OnavRIM	RGB-D, GPS+Compass	79°	30°	5.6	2.2	40.1
OnavRIM (500 steps)	RGB-D, GPS+Compass	79°	30°	56.3	16.2	17.3
PIRLNav	RGB-D, GPS+Compass	79°	30°	51.7	41.5	49.2
PIRLNav (500 steps)	RGB-D, GPS+Compass	79°	30°	68.7	43.3	44.6
SegDT with predicted segmentation	RGB-D	79°	30°	53.4	<u>50.1</u>	<u>56.7</u>
SegDT with predicted segmentation (500 steps)	RGB-D	79°	30°	<u>69.4</u>	58.8	58.9
RL with RNN and GT segmentation	RGB-D	42°	15°	49.1	36.4	58.5
SegDT with GT segmentation	RGB-D	42°	15°	47.3	44.7	56.3
RL with RNN and predicted segmentation	RGB-D	42°	15°	31.2	28.2	46.2
SegDT with predicted segmentation	RGB-D	42°	15°	40.2	38.3	51.5

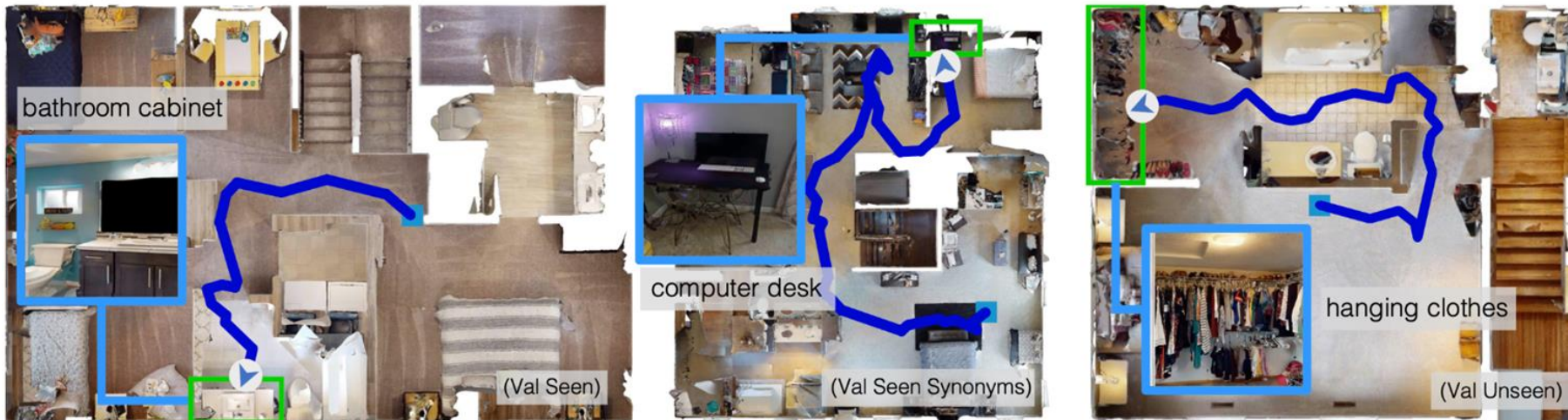
Task description (Open-Vocabulary ObjectGoal Navigation)

Navigation within unfamiliar environment to an object of an arbitrary category.

Category split	Train	Val seen	Val seen (synonyms)	Val unseen
Categories number	280	79	50	49

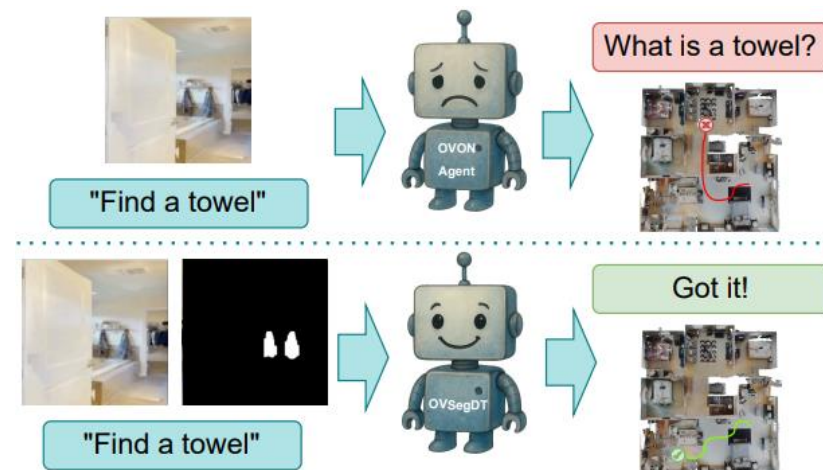
Method	VAL SEEN		VAL SEEN SYNONYMS		VAL UNSEEN	
	SR(↑)	SPL(↑)	SR(↑)	SPL(↑)	SR(↑)	SPL(↑)
1) BC	11.1 $\pm$ 0.1	4.5 $\pm$ 0.1	9.9 $\pm$ 0.4	3.8 $\pm$ 0.1	5.4 $\pm$ 0.1	1.9 $\pm$ 0.2
2) DAgger	18.1 $\pm$ 0.4	9.4 $\pm$ 0.3	15.0 $\pm$ 0.4	7.4 $\pm$ 0.3	10.2 $\pm$ 0.5	4.7 $\pm$ 0.3
3) RL	39.2 $\pm$ 0.4	18.7 $\pm$ 0.2	27.8 $\pm$ 0.1	11.7 $\pm$ 0.2	18.6 $\pm$ 0.3	7.5 $\pm$ 0.2
4) BCRL	20.2 $\pm$ 0.6	8.2 $\pm$ 0.4	15.2 $\pm$ 0.1	5.3 $\pm$ 0.1	8.0 $\pm$ 0.2	2.8 $\pm$ 0.1
5) DAgRL	41.3$\pm$0.3	21.2$\pm$0.3	29.4 $\pm$ 0.3	14.4 $\pm$ 0.1	18.3 $\pm$ 0.3	7.9 $\pm$ 0.1
6) *VLFM [12]	35.2	18.6	32.4	17.3	35.2	19.6
7) DAgRL+OD	38.5 $\pm$ 0.4	21.1 $\pm$ 0.4	39.0$\pm$0.7	21.4$\pm$0.5	37.1$\pm$0.2	19.9$\pm$0.3

*deterministic method



OVSegDT

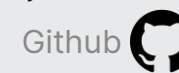
- OVSegDT is a **lightweight (130M) transformer-based architecture for open-vocabulary navigation**, that incorporates a semantic encoder for binary target masks into the policy's observation space.
- Goal mask encoder enables **robust navigation to both seen and unseen object categories**, even with noisy predicted segmentation.
- On HM3D-OVON, our model achieves performance on unseen categories comparable to that on seen ones and **establishes state-of-the-art results** (44.7% SR, 20.6% SPL on val unseen) **without using depth, odometry, or large vision-language models**.



Method	Depth	Odometry	Val Seen		Val Seen Synonyms		Val Unseen	
			SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
BC	✗	✗	11.1 ± 0.1	4.5 ± 0.1	9.9 ± 0.4	3.8 ± 0.1	5.4 ± 0.1	1.9 ± 0.2
DAgger	✗	✗	18.1 ± 0.4	9.4 ± 0.3	15.0 ± 0.4	7.4 ± 0.3	10.2 ± 0.5	4.7 ± 0.3
RL	✗	✗	39.2 ± 0.4	18.7 ± 0.2	27.8 ± 0.1	11.7 ± 0.2	18.6 ± 0.3	7.5 ± 0.2
BCRL	✗	✗	20.2 ± 0.6	8.2 ± 0.4	15.2 ± 0.1	5.3 ± 0.1	8.0 ± 0.2	2.8 ± 0.1
DAgRL	✗	✗	41.3 ± 0.3	21.2 ± 0.3	29.4 ± 0.3	14.4 ± 0.1	18.3 ± 0.3	7.9 ± 0.1
Uni-NaVid (LLM-based)	✗	✗	41.3	21.1	43.9	21.8	39.5	19.8
VLFM	✓	✓	35.2	18.6	32.4	17.3	35.2	19.6
DAgRL+OD	✓	✓	38.5 ± 0.4	21.1 ± 0.4	39.0 ± 0.7	21.4 ± 0.5	37.1 ± 0.2	19.8 ± 0.3
TANGO (LLM-based)	✓	✓	-	-	-	-	35.5 ± 0.3	19.5 ± 0.3
MTU3D	✓	✓	55.0	23.6	45.0	14.7	40.8	12.1
OVSegDT (ours)	✗	✗	<u>43.6 ± 0.4</u>	20.1 ± 0.2	40.1 ± 0.4	17.9 ± 0.1	44.7 ± 0.4	20.6 ± 0.2



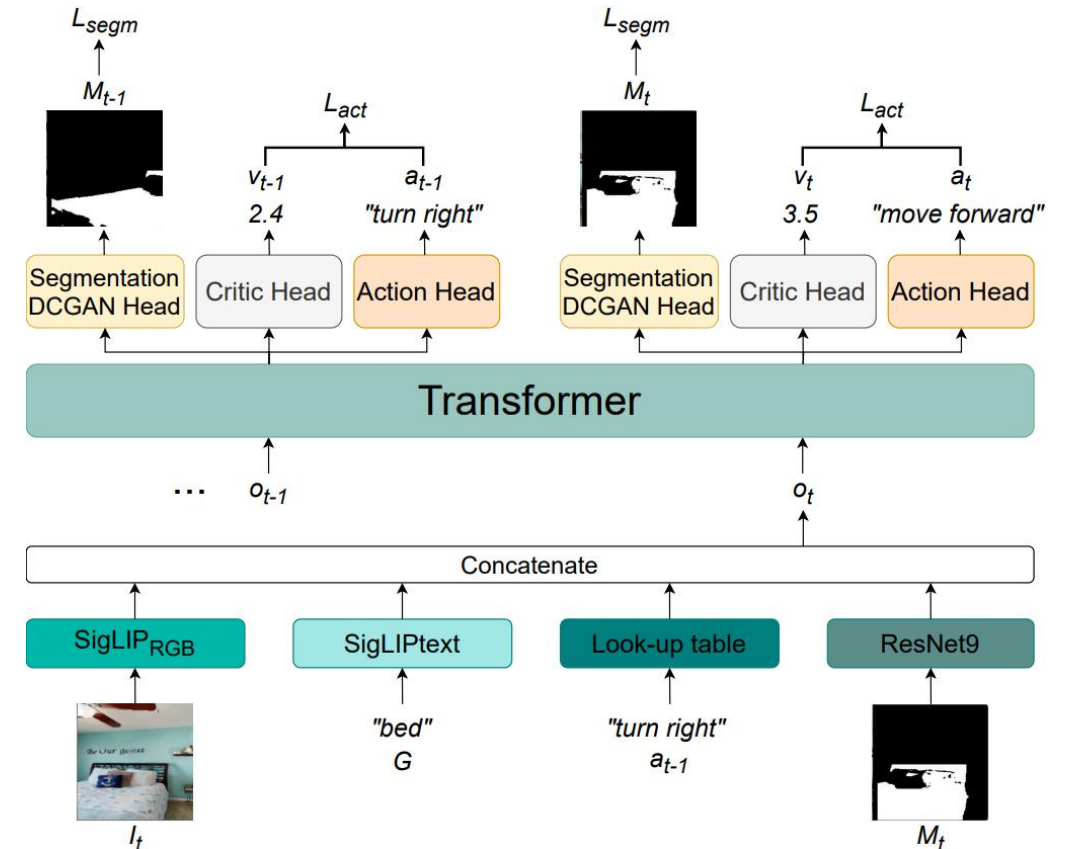
Project OVSegDT



Key components of OVSegDT (1) – architecture

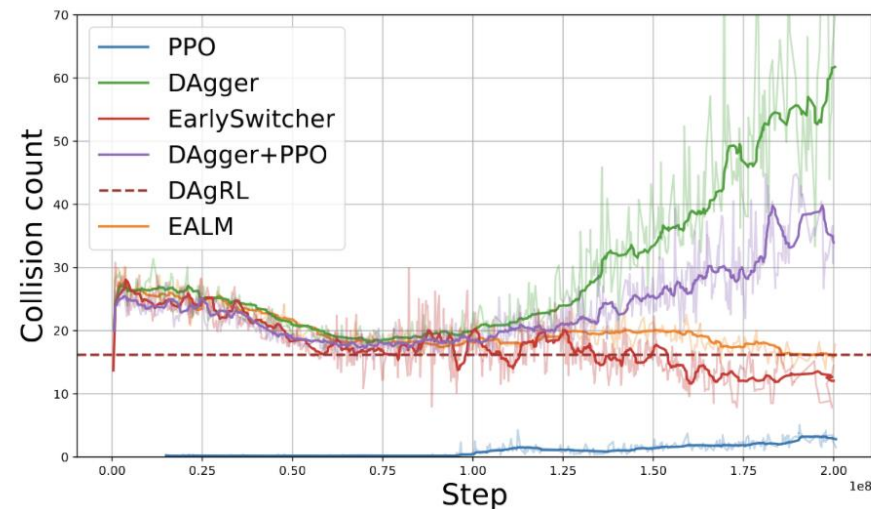
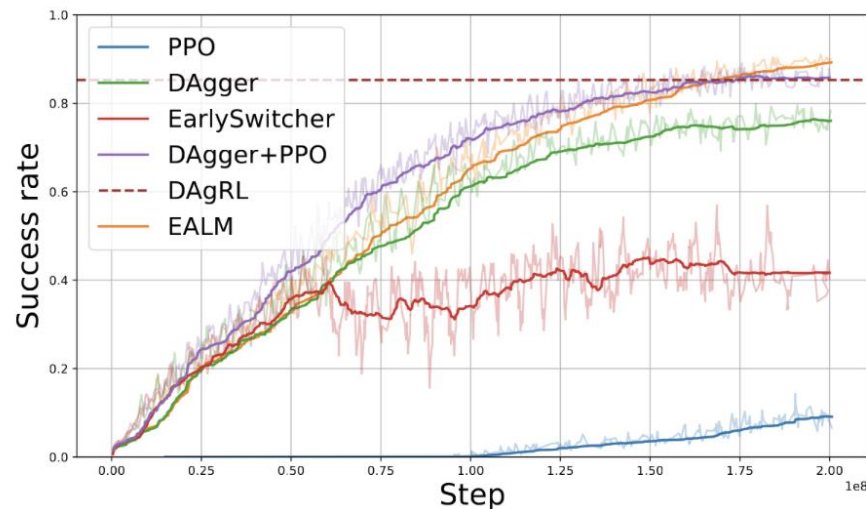
- The synergy between **navigation** and **semantic scene understanding** is achieved through two components: the architecture and the training objectives.
- **Architecture:** the binary goal segmentation mask is encoded as a part of observation.
- **Training objectives:**
 - **Auxiliary segmentation loss function** $\mathcal{L}_{seg}$ is computed between the reconstructed binary target mask and the ground-truth target mask for the current frame.
 - **Semantic reward** r^{sem} is given to the agent for increasing the area of the observed ground-truth target mask and for moving closer to the target.

Method	Goal mask	r^{sem}	$\mathcal{L}_{seg}$	SR $\uparrow$	SPL $\uparrow$	Collisions $\downarrow$
DagRL	✗	✗	✗	18.3 $\pm$ 0.3	7.9 $\pm$ 0.1	50.2
OVSegDT	✓	✗	✓	26.7 $\pm$ 0.5	13.2 $\pm$ 0.2	29.4 $\pm$ 3.1
OVSegDT	✓	✓	✗	<u>36.9 $\pm$ 0.7</u>	<u>16.5 $\pm$ 0.1</u>	49.3 $\pm$ 0.4
OVSegDT	✓	✓	✓	44.7 $\pm$ 0.4	20.6 $\pm$ 0.2	<u>45.4 $\pm$ 0.6</u>



Key components of OVSegDT (2) - EALM

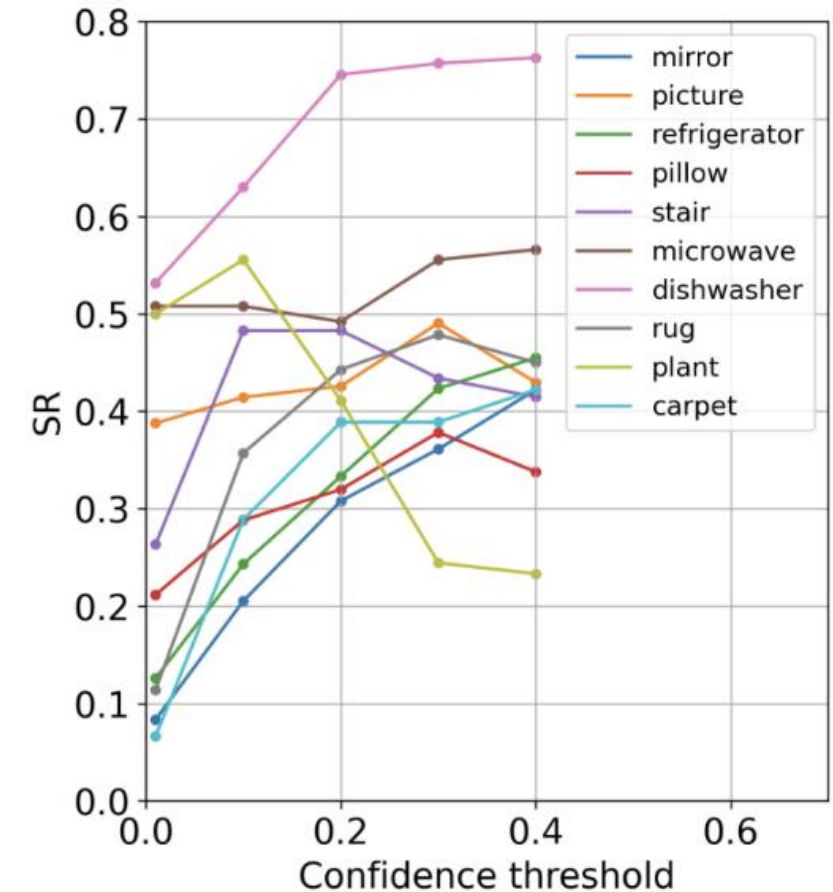
- Open-vocabulary navigation requires a policy to decide how strongly it should rely on **imitation signals versus exploration-driven reinforcement learning**.
- **To eliminate brittle manual phase switches** between these signals we propose a **per-sample scheduler EALM** (Entropy-Adaptive Loss Modulation).
- EALM solves this by **continuously blending** Behavior Cloning (BC) and Proximal Policy Optimization (PPO) objectives according **to the policy's uncertainty, quantified by its action entropy**.
- The gradual transition from BC to PPO helps **reduce collisions** and **increase the success rate**, while **improving sample efficiency** by 33% (200M steps vs. 300M steps).



Key components of OVSegDT (3) — navigation with predicted segmentation

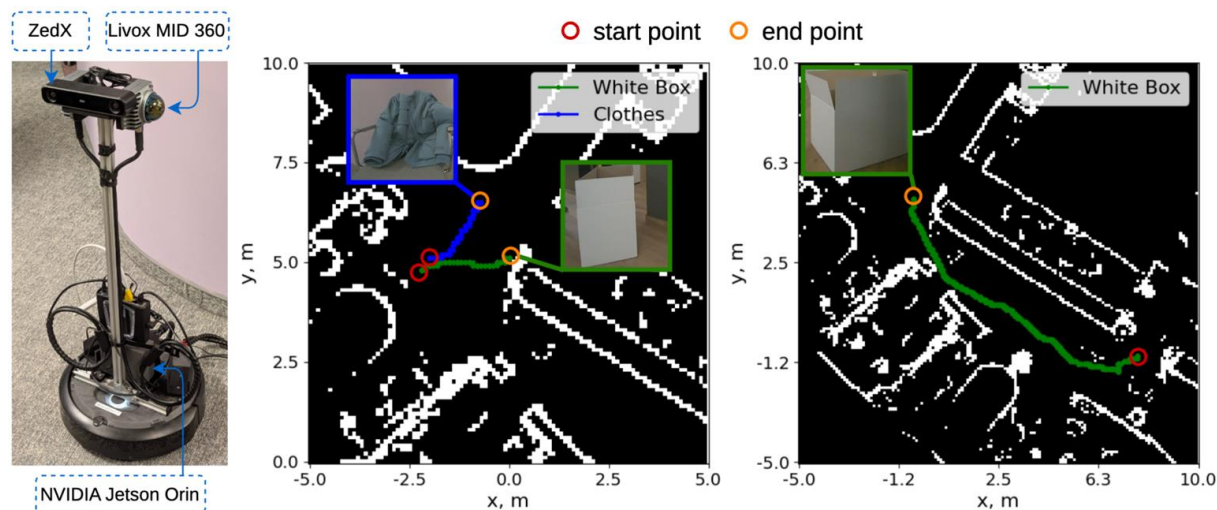
- The **navigation performance** using the **YOLOE-predicted** binary target mask **drops significantly** compared to **navigation using the ground-truth segmentation**.
- To **improve navigation performance** when using predicted segmentation, **we propose two methods**:
 - **YOLOE per-category confidence calibration**.
 - **Fine-tuning the model for navigation with predicted segmentation**:
The reward for increasing the area of the ground-truth segmentation mask serves as **an active perception signal**, leading to a **reduction in false-negative target detections** through the agent's action selection.

Predicted finetune	mask	Segmentation source	SR $\uparrow$	SPL $\uparrow$	Collisions $\downarrow$
$\times$		GT	69.3 $\pm$ 0.2	39.1 $\pm$ 0.2	23.3 $\pm$ 0.3
$\times$		YOLOE	28.8 $\pm$ 0.4	14.0 $\pm$ 0.2	19.6 $\pm$ 0.2
$\times$		YOLOE+calib.	35.6 $\pm$ 0.4	16.2 $\pm$ 0.3	20.2 $\pm$ 0.8
$\checkmark$		YOLOE+calib.	44.7 $\pm$ 0.4	20.6 $\pm$ 0.2	45.4 $\pm$ 0.6



Real-world Robot Experiments

- We conduct **demonstration experiments** on a real **iRobot Create 3** robot equipped with a **ZED X camera**, Livox MID 360 LiDAR, and an **Nvidia Jetson Orin PC**.
- The robot successfully performs **target-oriented navigation**, correctly identifying and reaching **both visible** and **initially hidden targets**.
- The navigation system demonstrates robust **long-horizon behavior**, including **effective obstacle avoidance** during movement through complex environments.

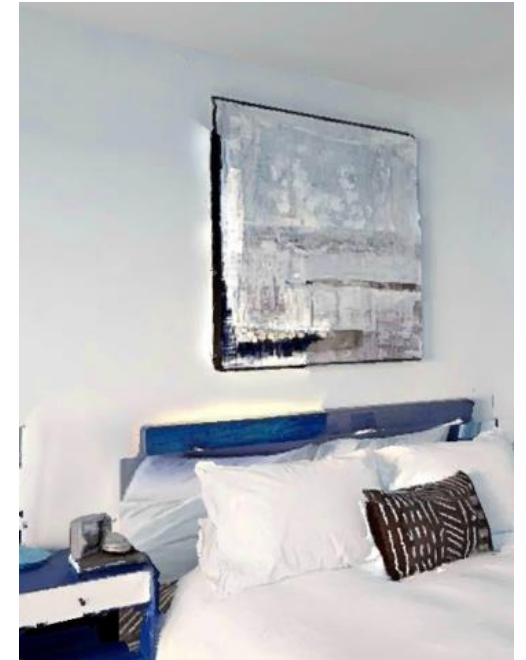
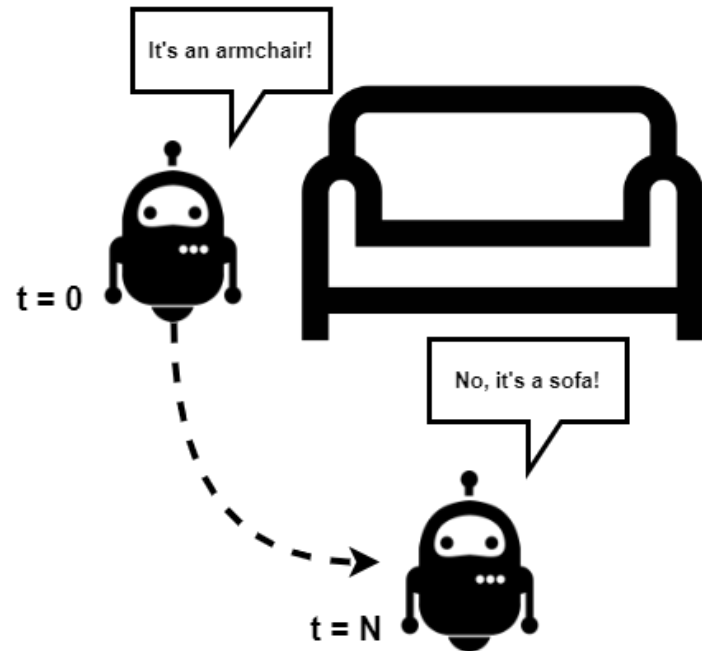


03

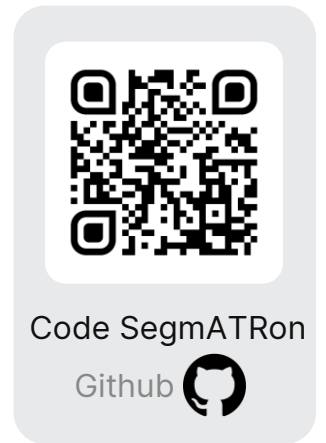
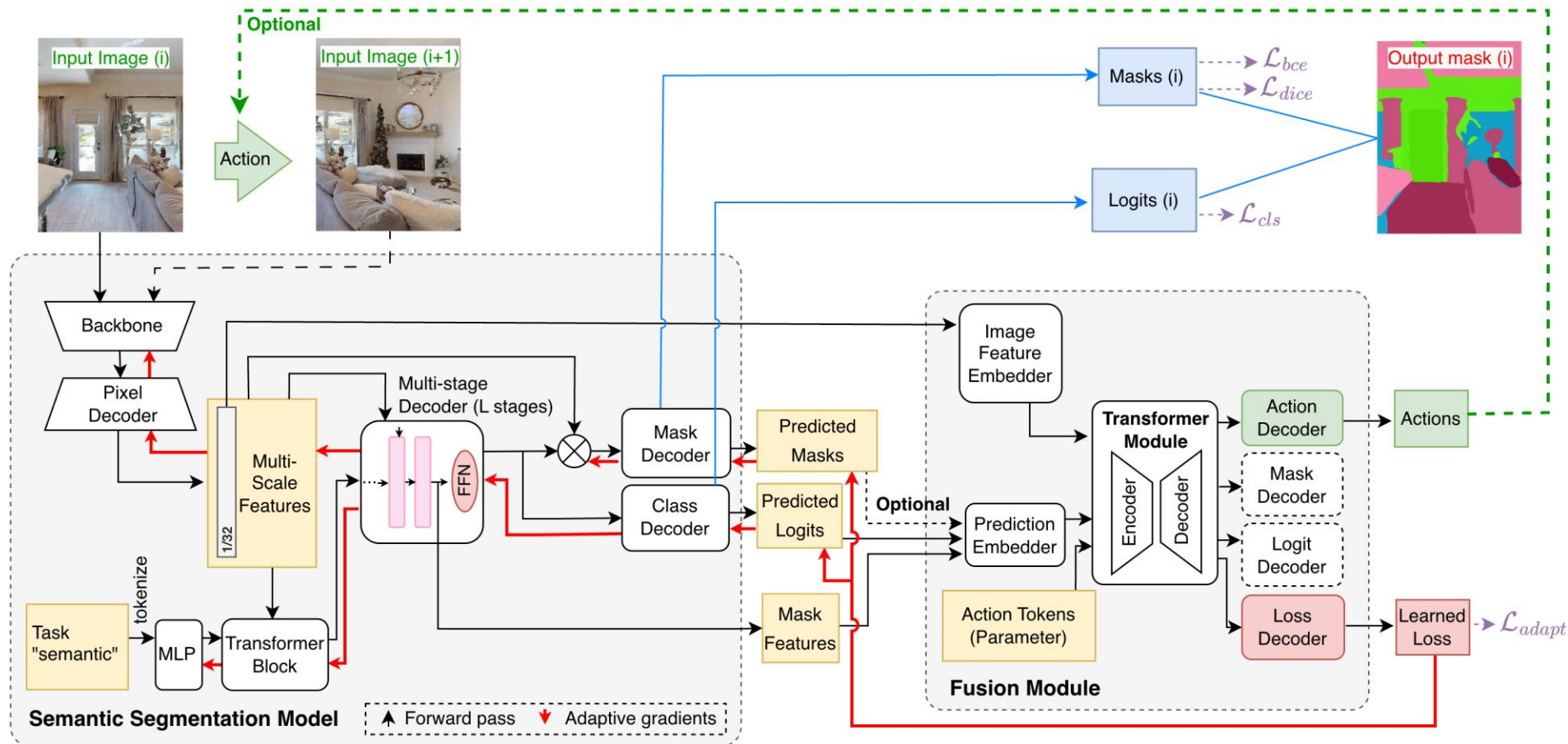
Active perception for navigation

Task setup (Active Perception)

Task: Predict actions in order to improve semantic perception of the environment.



SegmATRon architecture



Improving segmentation on the current frame using active segmentation

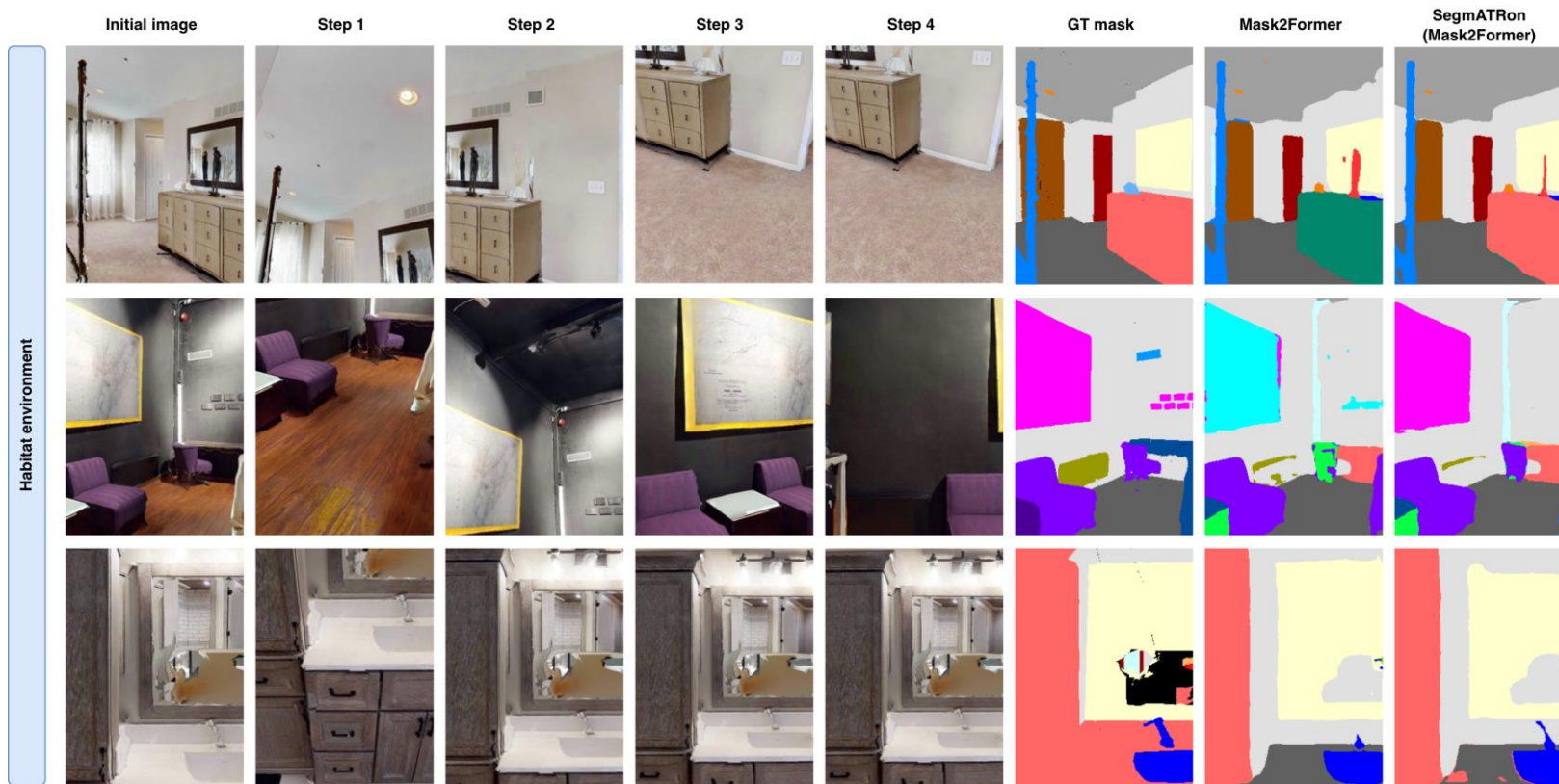
→ **SegmATRon with a random action strategy** significantly outperforms the baseline **Mask2Former (M2F)** approach on the **validation dataset in the Habitat environment** in terms of **segmentation quality**.

→ Using a **learned strategy** based on a **weighted best-value loss function** enables an even **greater improvement in segmentation quality** for SegmATRon (Mask2Former).

Method	Adaptation on Inference	Action Policy	mIoU, %	fwIoU, %	mACC, %	pACC, %
Mask2Former	no	Single Frame	21.9	65.1	31.3	76.6
SegmATRon (M2F) 4 Steps	no	Random	20.5 (−6.4%)	62.5 (−4.0%)	27.5 (−12.1%)	75.9 (−0.9%)
SegmATRon (M2F) 4 Steps	yes	Random (mean)	22.6 (+3.2%)	66.5 (+2.2%)	31.6 (+1.0%)	78.0 (+1.8%)
SegmATRon (M2F) 4 Steps	yes	Random (max)	23.7 (+8.2%)	67.0 (+2.9%)	32.7 (+4.5%)	78.5 (+2.5%)
SegmATRon (M2F) 4 Steps	yes	Best loss	<u>23.3</u> (+6.4%)	<u>66.9</u> (+2.8%)	32.7 (+2.6%)	<u>78.2</u> (+2.1%)

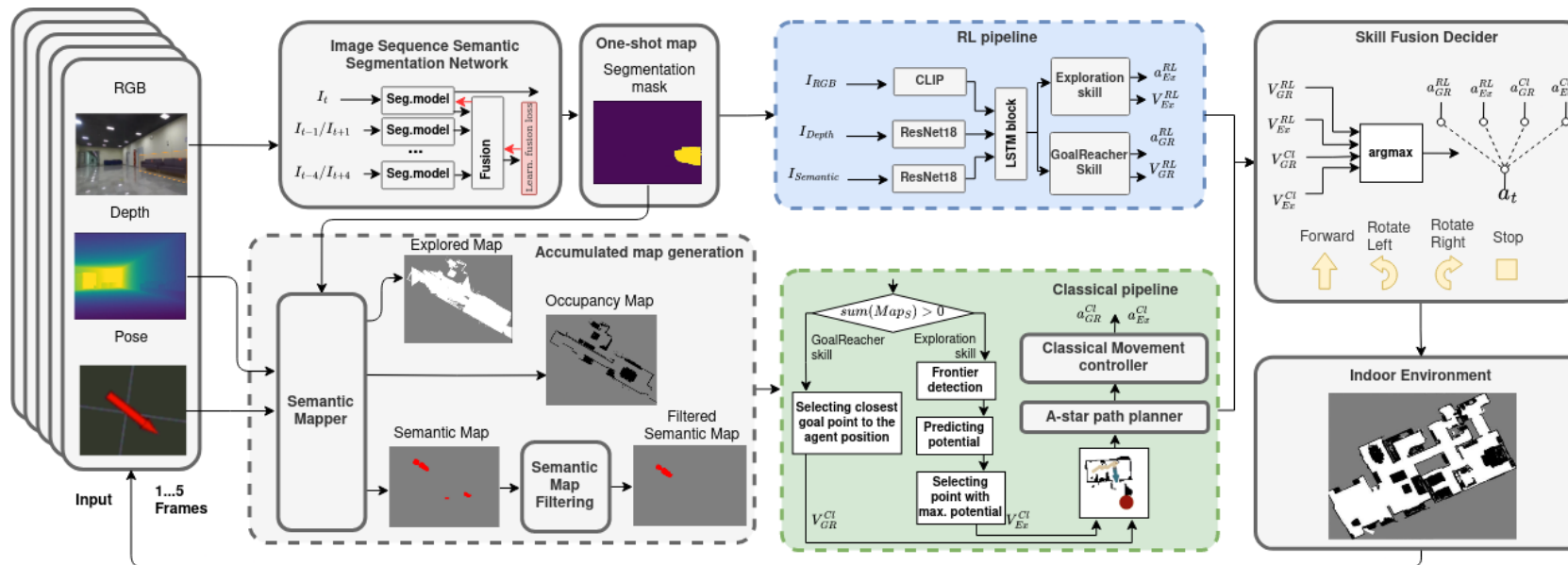
Examples of active segmentation

Mask2Former-based SegmATRon uses additional frames to segment objects more accurately, correcting object classification errors and improving mask quality.



Active perception and semantic segmentation enhancement for environment mapping (SkillTron)

- **Two-level representation of a semantic map** that aims to reduce semantic noise in two stages.
- A **one-shot map** in the form of a **segmentation mask** is constructed **during the interaction** of the embodied agent with the environment.
- At this stage, **additional observations** are used to refine the **2D segmentation mask**.
- The accumulated map is formed using a **unique filtering algorithm** for the semantic **bird's-eye-view map**.



Code SkillTron
Github

SkillTron performance on the HM3DSem v0.2 dataset

- **SkillTron surpasses the state-of-the-art approach SkillFusion** by a considerable margin of 4% of success rate, 2% of SPL and 2% of SoftSPL metrics.
- The **SkillTron significantly outperforms** previous **state-of-the-art methods** for the indoor Object Goal Navigation task such as DD-PPO and Aux-RL.

Method	Semantic method	Number of steps	Turn Angle	Success	SPL	SoftSPL
SkillTron	OneFormer	-	30°	0.57	0.28	0.35
<i>SkillTron</i>	<i>SegmATRon(B)</i>	<u>1</u>	<u>30°</u>	<u>0.58</u>	0.29	<u>0.35</u>
SkillTron	SegmATRon(B)	2	30°	0.59	<u>0.28</u>	0.36
SkillTron	SegmATRon (B)	4	30°	0.57	0.27	<u>0.35</u>
SkillTron	SegmATRon (B)	1	15°	0.51	0.25	0.33
SkillTron	SegmATRon (B)	2	15°	0.49	0.24	0.31
SkillTron	SegmATRon (B)	4	15°	0.50	0.25	0.32

Method	Exploration	Goalreacher	Semantic method	Success	SPL	SoftSPL
DD-PPO	-	-	-	0.07	0.04	0.28
Aux-RL	-	-	-	0.18	0.10	0.31
Host_74441_Team	-	-	-	0.12	0.05	0.27
ICanFly	-	-	-	0.43	<u>0.26</u>	0.37
SkillFusion	FBE+RL	Classic+RL	OneFormer	<u>0.55</u>	<u>0.26</u>	0.34
SkillTron	PONI + RL	Classic + RL	SegmATRon(B)(2Steps)	0.59	0.28	<u>0.36</u>

04

Scene Graphs for 3D Referred Object Grounding

Task setup (3D Referred Object Grounding)

Task: Predict the coordinates of the 3D bounding boxes of the object corresponding to the description.



Examples of ScanRefer and Multi3DRefer queries (ScanNet scenes)

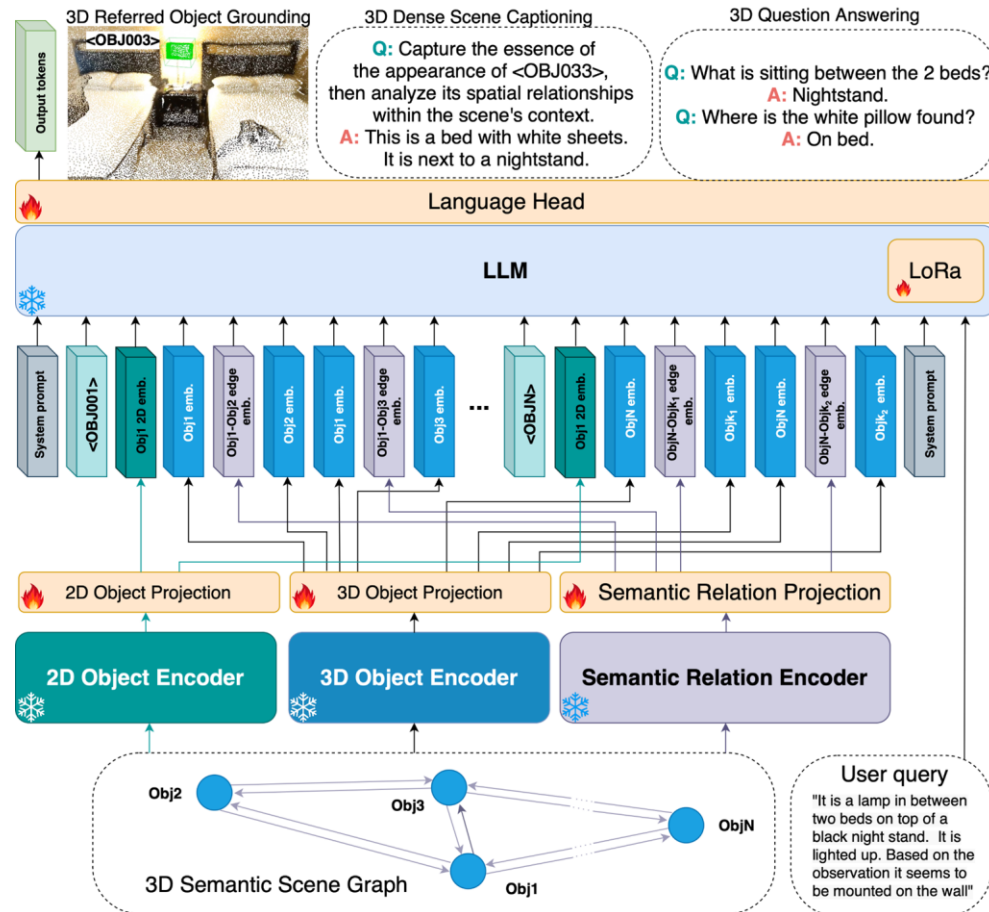
Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5828-5839).

Chen, D. Z., Chang, A. X., & Nießner, M. (2020, August). Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision* (pp. 202-221). Cham: Springer International Publishing.

Zhang, Y., Gong, Z., & Chang, A. X. (2023). Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 15225-15236).

3DGraphLLM: Learnable representation of a 3D scene graph for LLM

- 3DGraphLLM is a method for creating a learnable representation of a 3D scene graph, which improves the quality user text queries for an LLM.
- Each object is represented by its subgraph, consisting of k nearest neighbors.
- The connection with each of the k nearest neighbors is described using a triplet: (object1, VL-SAT connection descriptor, object2).
- Additionally, we utilize learnable object identifier embeddings and 2D object features to improve 3DGraphLLM performance on 3D vision-language tasks.



Project
3DGraphLLM
(ICCV 2025)



3DGraphLLM Results (Mask3D instance segmentation)

- Neural network instance segmentation Mask3D for extracting objects to create graph vertices.
- 3DGraphLLM provides more **accurate 3D scene understanding** compared to its closest baseline: the Chat-Scene method.
- Implicit scene graph representation allows for **increased inference speed** in solving 3D vision-language tasks compared to an explicit text or video representation of the scene.

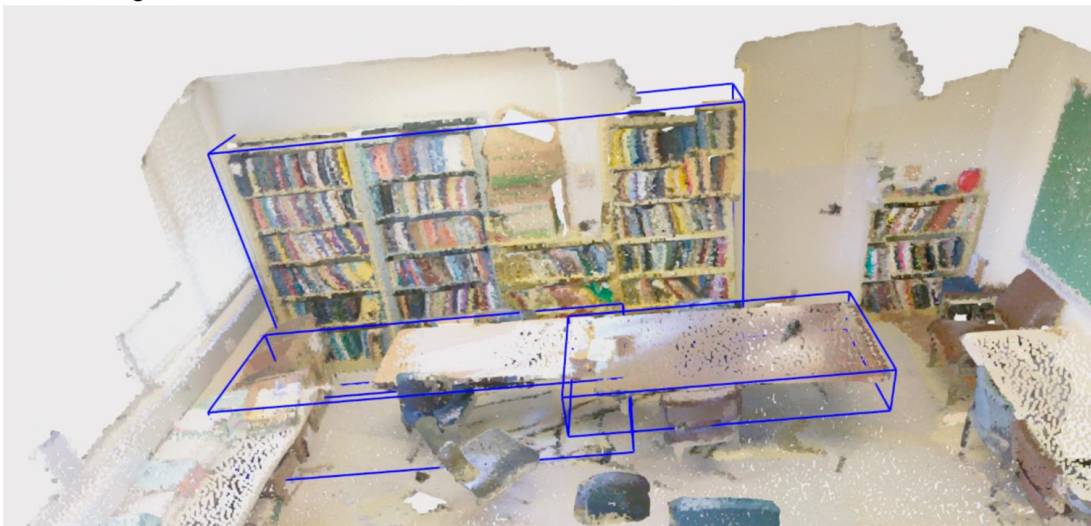
	Dataset	Method	
		3DGraphLLM	GPT4Scene
Input token number per scene		800	10400
Inference speed, sec	ScanRefer	0.4	1.9
	Multi3DRefer	0.5	2.0
	Scan2Cap	0.9	2.2
	ScanQA	0.4	1.9
	SQA3D	0.4	1.7

	Methods	2D features	3D features	LLM	ScanRefer		Multi3DRefer		Scan2Cap		ScanQA		Sqa3D
					A@0.25↑	A@0.5↑	F1@0.25↑	F1@0.5↑	C@0.5↑	B-4@0.5↑	C↑	B-4↑	EM↑
Expert models	ScanRefer [5]	✓	✓	✗	37.3	24.3	-	-	-	-	-	-	-
	MVT [27]	✓	✓	✗	40.8	33.3	-	-	-	-	-	-	-
	3DVG-Trans [61]	✓	✓	✗	45.9	34.5	-	-	-	-	-	-	-
	ViL3DRel [6]	✗	✓	✗	47.9	37.7	-	-	-	-	-	-	-
	M3DRef-CLIP [60]	✓	✓	✗	51.9	44.7	42.8	38.4	-	-	-	-	-
	Scan2Cap [9]	✓	✓	✗	-	-	-	-	35.2	22.4	-	-	-
	ScanQA [3]	✓	✓	✗	-	-	-	-	-	-	64.9	10.1	-
	Sqa3D [35]	✗	✓	✗	-	-	-	-	-	-	-	-	47.2
	3D-VisTA [65]	✗	✓	✗	50.6	45.8	-	-	66.9	34.0	72.9	13.1	48.5
	BUTD-DETR [28]	✗	✓	✗	52.2	39.8	-	-	-	-	-	-	-
LLM-based models	PQ3D [66]	✓	✓	✗	-	51.2	-	50.1	80.3	36.0	87.8	-	47.1
	ZSVG3D [58]	✓	✓	GPT4	36.4	32.7	-	-	-	-	-	-	-
	3D-LLM [21]	✓	✓	Flamingo	21.2	-	-	-	-	-	59.2	7.2	-
	3D-LLM [21]	✗	✓	BLIP2-flant5	30.3	-	-	-	-	-	69.4	12.0	-
	Chat-3D v2 [24]	✗	✓	Vicuna-7B-v0	35.9	30.4	-	-	-	-	77.1	7.3	-
	Scene-LLM [15]	✓	✓	Llama-2-7B	-	-	-	-	-	-	80.0	12.0	54.2
	LEO [26]	✗	✓	Vicuna-7B-v1.1	-	-	-	-	72.4	38.2	101.4	13.2	50.0
	LL3DA [7]	✗	✓	OPT-1.3B	-	-	-	-	65.2	36.8	76.8	13.5	-
	Grounded 3D-LLM [8]	✗	✓	Tiny-Vicuna-1B	47.9	44.1	45.2	40.6	70.6	35.5	72.7	13.4	-
	Robin3D [31]	✓	✓	Vicuna-7B-v1.5	60.8	55.1	64.9	59.7	87.2	38.4	-	-	56.0
	GPT4Scene-HD [43]	✓	✓	Qwen2-VL-7B	50.9	46.4	53.7	50.0	74.4	37.9	89.9	15.9	57.2
	GPT4Scene-HDM [43]	✓	✓	Qwen2-VL-7B	62.6	57.0	64.5	59.8	86.3	40.6	96.3	15.5	59.4
	Chat-Scene [25] (baseline)	✓	✓	Vicuna-7B-v1.5	55.5	50.2	57.1	52.4	77.1	36.3	87.7	14.3	54.6
	3DGraphLLM (ours)	✓	✓	Vicuna-7B-v1.5	58.6	53.0	61.9	57.3	79.2	34.7	91.2	13.7	55.1
	3DGraphLLM (ours)	✓	✓	LLAMA3-8B-Instruct	<u>62.4</u>	<u>56.6</u>	<u>64.7</u>	59.9	81.0	36.5	88.8	15.9	55.9

Qualitative example of 3DGraphLLM performance

Q: I am organizing the bookshelf under the boxes. Is the table behind me round or rectangle?

A: Rectangle.



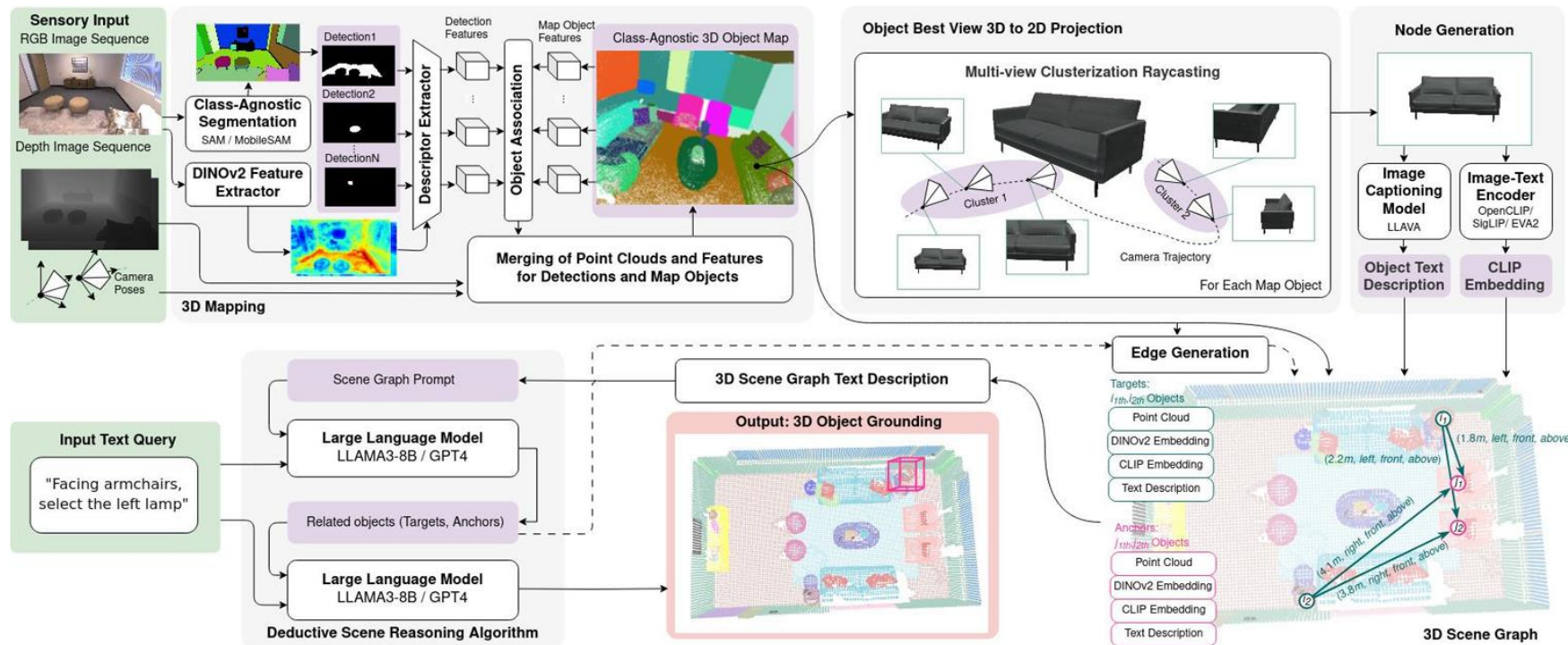
Q: It is a black suitcase on the floor. It is sitting beside the mini fridge.

A: <OBJ014>



Beyond Bare Queries: Open-Vocabulary Object Grounding with 3D Scene Graph

- DINO features for object aggregation
- Deductive Scene Reasoning Algorithm
- Best view selection for image captioning



Project BBQ
(ICRA 2025)

Github

Quantitative results on public datasets

The results on the Sr3D+, Nr3D, and ScanRefer datasets demonstrate the effectiveness of the BBQ modular method.

Sr3D+										
Methods	Overall		Easy		Hard		View Dep.		View Indep.	
	A@0.1	A@0.25	A@0.1	A@0.25	A@0.1	A@0.25	A@0.1	A@0.25	A@0.1	A@0.25
OpenFusion	12.6	2.4	14.0	2.4	1.3	1.3	3.8	2.5	13.7	2.4
BBQ-CLIP	14.4	8.8	15.4	9.0	6.7	6.7	11.4	5.1	14.4	8.8
ConceptGraphs	13.3	6.2	13.0	6.8	16.0	1.3	15.2	5.1	13.1	6.4
BBQ	34.2	22.7	34.3	22.7	33.3	22.7	32.9	20.3	34.4	23.0

Nr3D										
Methods	Overall		Easy		Hard		View Dep.		View Indep.	
	A@0.1	A@0.25	A@0.1	A@0.25	A@0.1	A@0.25	A@0.1	A@0.25	A@0.1	A@0.25
OpenFusion	10.7	1.4	12.9	1.4	5.1	1.5	8.5	0.0	11.4	1.9
BBQ-CLIP	15.3	9.4	18.1	11.0	8.1	5.6	8.1	6.1	17.2	10.5
ConceptGraphs	16.0	7.2	18.7	9.2	9.1	2.0	12.7	4.2	17.0	8.1
BBQ	28.3	19.0	30.5	21.3	22.8	13.2	23.6	18.2	29.8	19.3

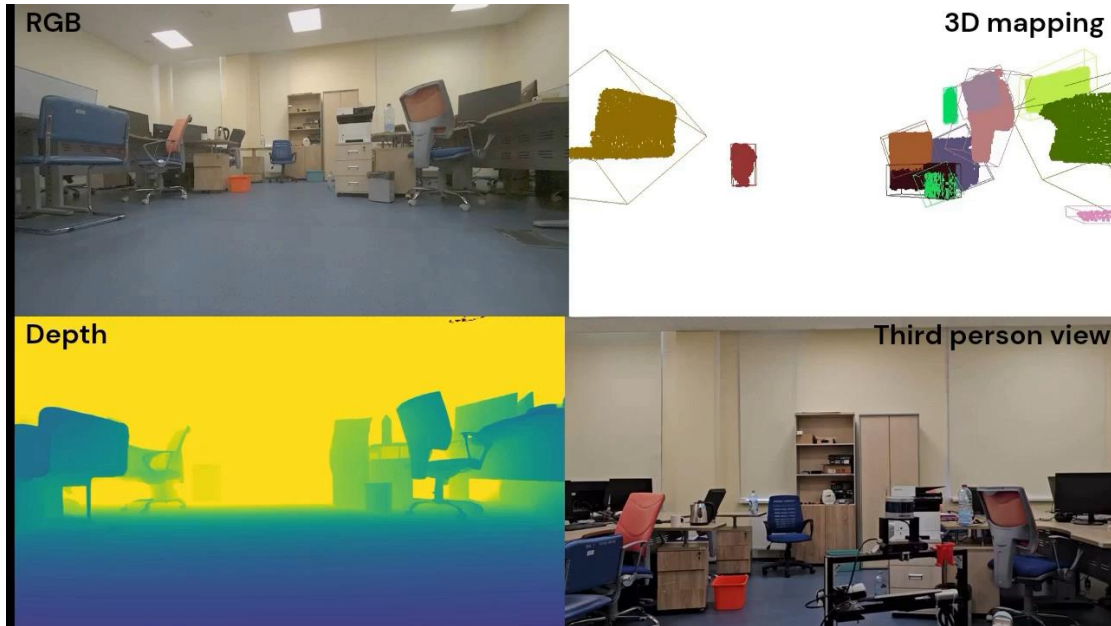
Grounding accuracy on Sr3D+/Nr3D

Methods	A@0.25	A@0.5
LERF	4.4	0.3
OpenScene	13.0	5.1
LLM-Grounder	17.1	5.3
BBQ	19.4	11.6

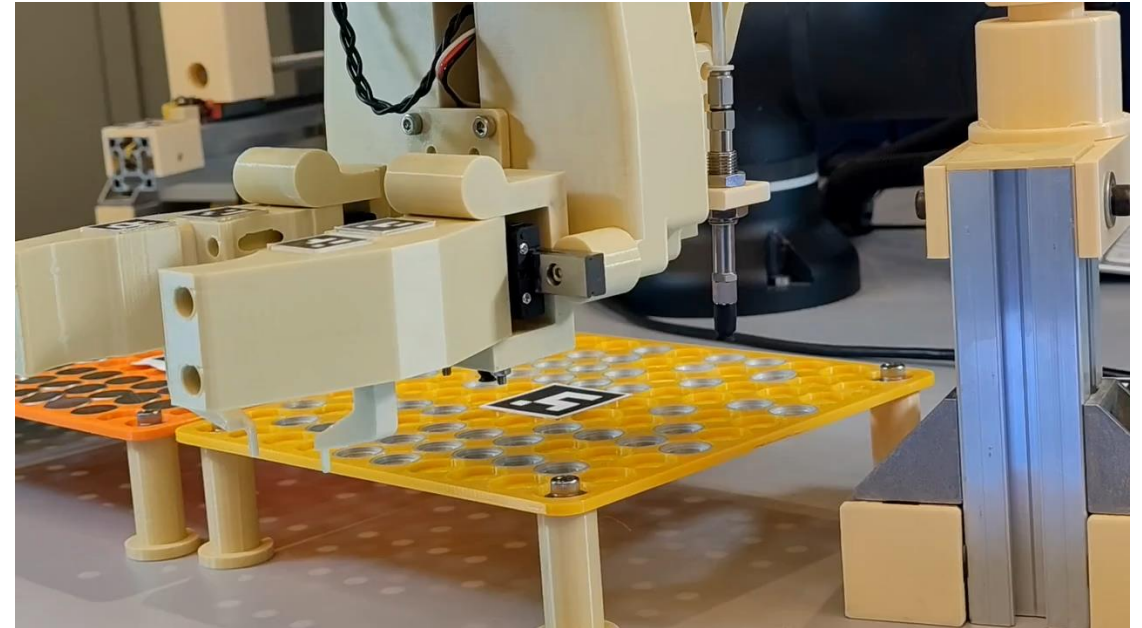
Grounding accuracy on ScanRefer

BBQ Applications

- Example of map construction and query execution in an office environment using a Husky Clearpath mobile robot.



- BBQ is used to determine the locations of target objects for manipulation by a robotic chemist during the automated assembly of CR032 batteries.



05



Conclusion

Conclusion

- In open-vocabulary navigation, a major challenge is the accurate recognition of the target by the agent.
 - A binary target mask and training objectives for target mask segmentation improve navigation performance for unseen categories.
 - Additional frames and learnable policy for their collection can be used to improve semantic segmentation and therefore object goal navigation quality.
 - A multimodal knowledge graph provides a compact and structured representation of a scene. The graph-based representation improves the accuracy of object retrieval in a 3D scene compared to a flat list of objects.
- Open challenges include developing embodied open-vocabulary navigation methods capable of handling dynamic environments and a large number of goal categories, including those that require contextual understanding for recognition (e.g., “a mug on the kitchen table”).

Contacts



Tatiana Zemskova

Research Scientist, AIRI
PhD Student, MIPT

✉ zemskova@airi.net





airi.net



[airi_research_institute](https://t.me/airi_research_institute)



[AIRI Institute](https://vk.com/AIRI_Institute)



Telegram

AIRI