Совместный семинар Российской ассоциации Искусственного интеллекта и Федерального исследовательского центра «Информатика и управление» РАН

28 февраля 2024

Нейросетевые методы построения мультимодальных карт и их использование для навигации и управления роботами

Дмитрий Александрович Юдин

к.т.н., заведующий лабораторией интеллектуального транспорта МФТИ - НКБ ВС,
Центр когнитивного моделирования МФТИ, старший научный сотрудник AIRI

AIRI

SBER Robotics Lab

MIPT MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

Center for Cognitive Modeling

ITL Intelligent Transport Lab

Cognitive Dynamic Systems

О чем поговорим

01 Мотивация

02 Методы построения мультимодальных карт

03 Плотные (Dense) методы

04 Разреженные (Sparse) методы

05 Базовые модели сегментации изображений с открытым словарем запросов (open vocabulary)

06 Наши исследования и приложения

07 Направления дальнейшего развития

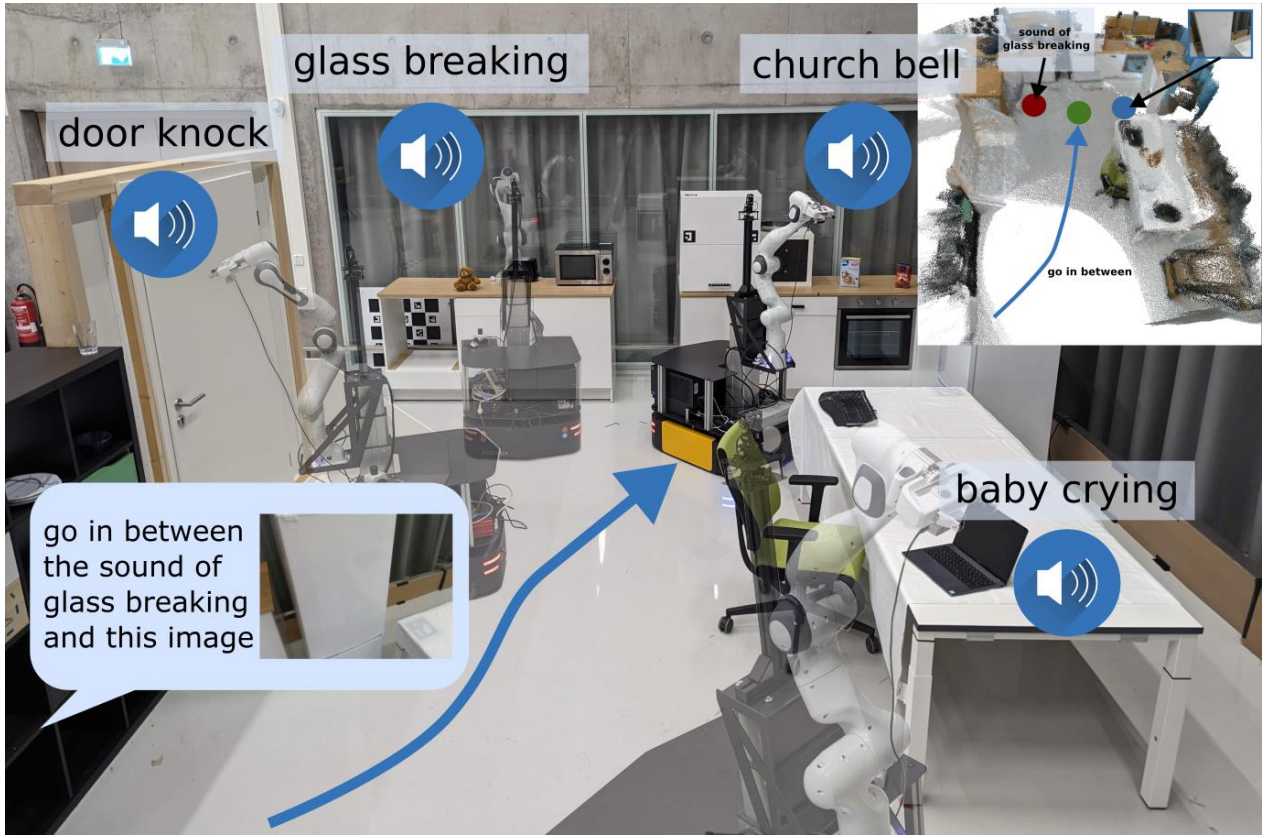
01

Мотивация

Будущее построения 3D-карт: пространственные запросы и ответы по трехмерным картам

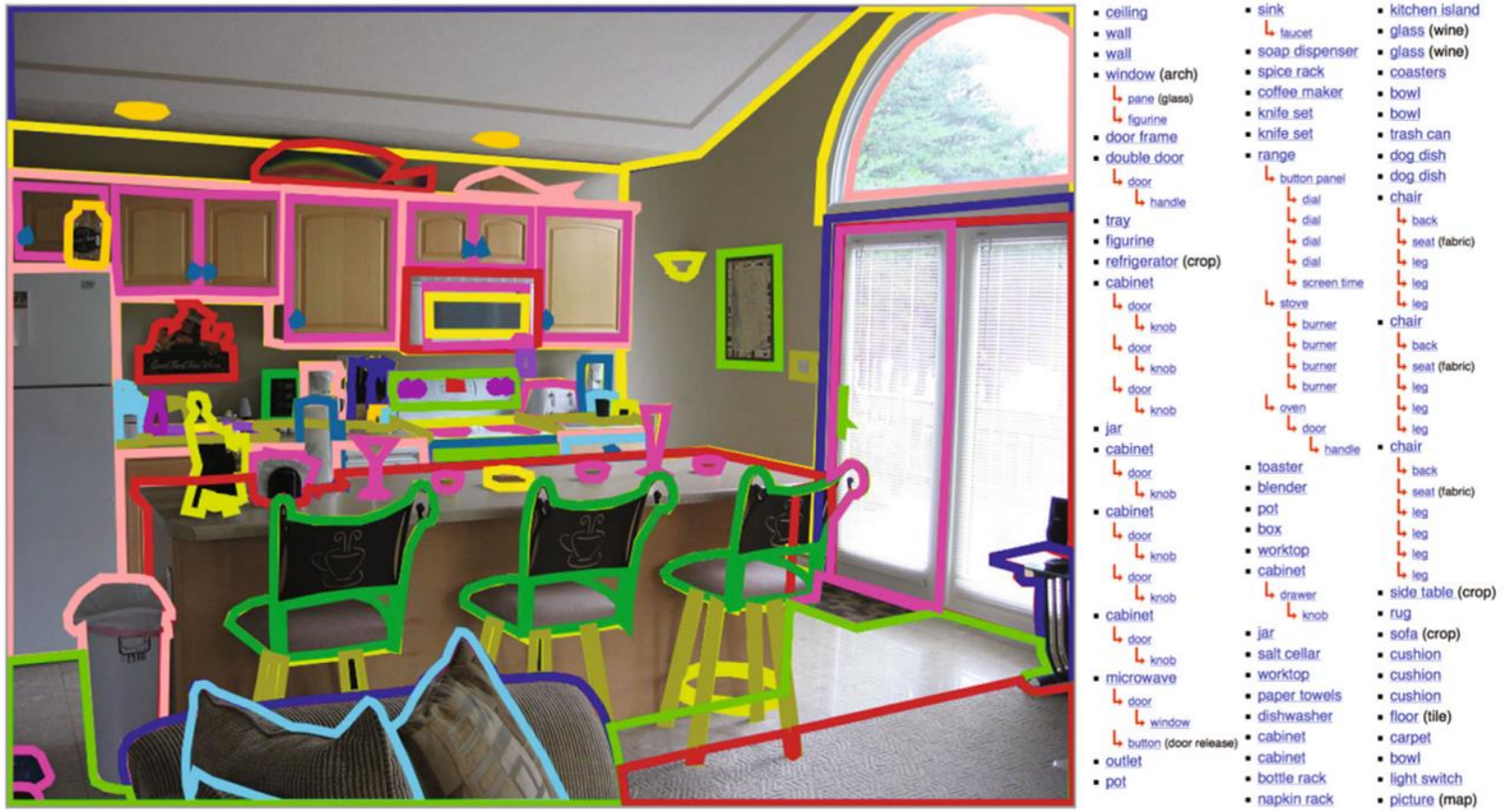
- **Текстовый запрос к 3D карте.**
  Пример: “Место для сидения (Sit)”

- **Мультимодальный запрос к 3D карте.**
  Пример: “Локация между местом, где был слышен звук разбитого стекла, и местом, показанном на фотографии”



Peng, Songyou, et al. Openscene: 3d scene understanding with open vocabularies. CVPR 2023. (Google Research, ETH Zurich, MPI for Intelligent Systems, Waymo LLC, Simon Fraser University)
Huang, Chenguang, et al. Audio visual language maps for robot navigation. arXiv preprint arXiv:2303.07522, 2023. (Freiburg University, Google Research, University of Technology Nuremberg)

Что делать с тысячами категорий объектов, которые надо сегментировать?



- Популярный набор данных ADE20K содержит 27,574 изображений внутри и вне помещений с 707,868 объектами, принадлежащими к 3,688 категориям(классам)
- Что делать, если мы хотим сегментировать объект, описанный на естественном языке?
- При подготовке собственных наборов данных мы хотим размечать объекты на изображениях в несколько кликов мыши

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. Semantic understanding of scenes through the ade20k dataset. ICCV, 127(3), 302-321. (2019)
(University of Hong Kong, Massachusetts Institute of Technology, Peking University, University of Toronto)

02

Методы построения мультимодальных карт

Какие методы бывают?

Методы построения мультимодальных карт

- **Плотные (Dense)**
  - **Точечные (Point-based)**
  - **Воксельные (Voxel-based)**
  - **Методы на основе NeRF (NeRF-based)**
  - **Методы на основе SDF (signed distance field) (SDF-based)**
- **Разреженные (Sparse)**
  - **Методы на основе объектов (Object-based)**
  - **Методы на основе иерархии объектов (Object-based with hierarchy)**
  - **Методы на основе графов объектов со связями (Object-based with relations, knowledge graphs)**

03

Плотные (Dense) методы

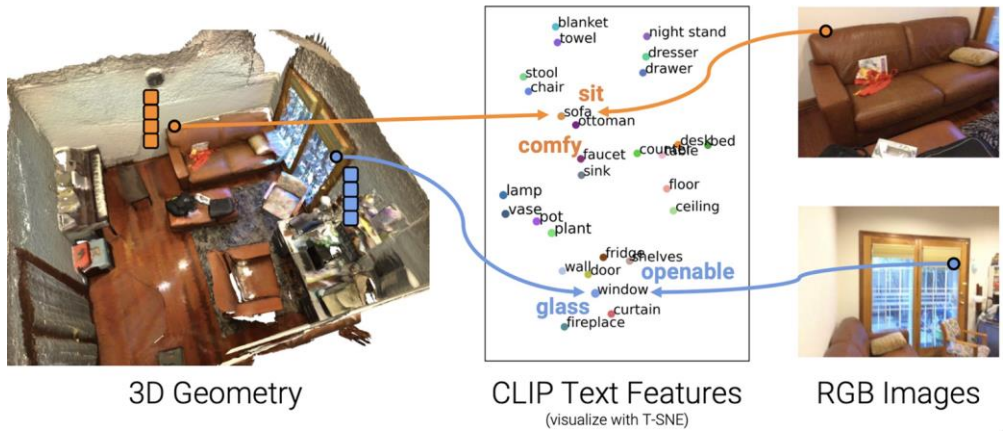
Постановка задачи

$$PCL_{Query} = \mathcal{A}((RGB_1 \dots RGB_N), [D_1 \dots D_N], (Pose_1 \dots Pose_N), Query), Query = (Text, [Image]).$$

$$I : Map = \mathcal{M}((RGB_1 \dots RGB_N), [D_1 \dots D_N], (Pose_1 \dots Pose_N)) = (Embedding_1 \dots Embedding_K),$$

$$where\ N - image\ sequence\ length,\ K - number\ of\ map\ points.$$

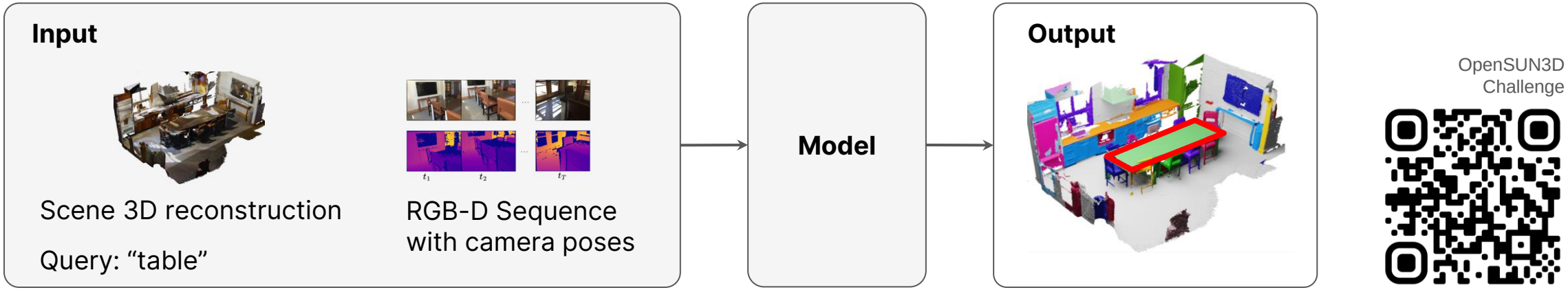
$$II : PCL_{Query} = \mathcal{Q}(Map, Query) = (X_1 \dots X_K, Y_1 \dots Y_K, Z_1 \dots Z_k, Score_1 \dots Score_K).$$



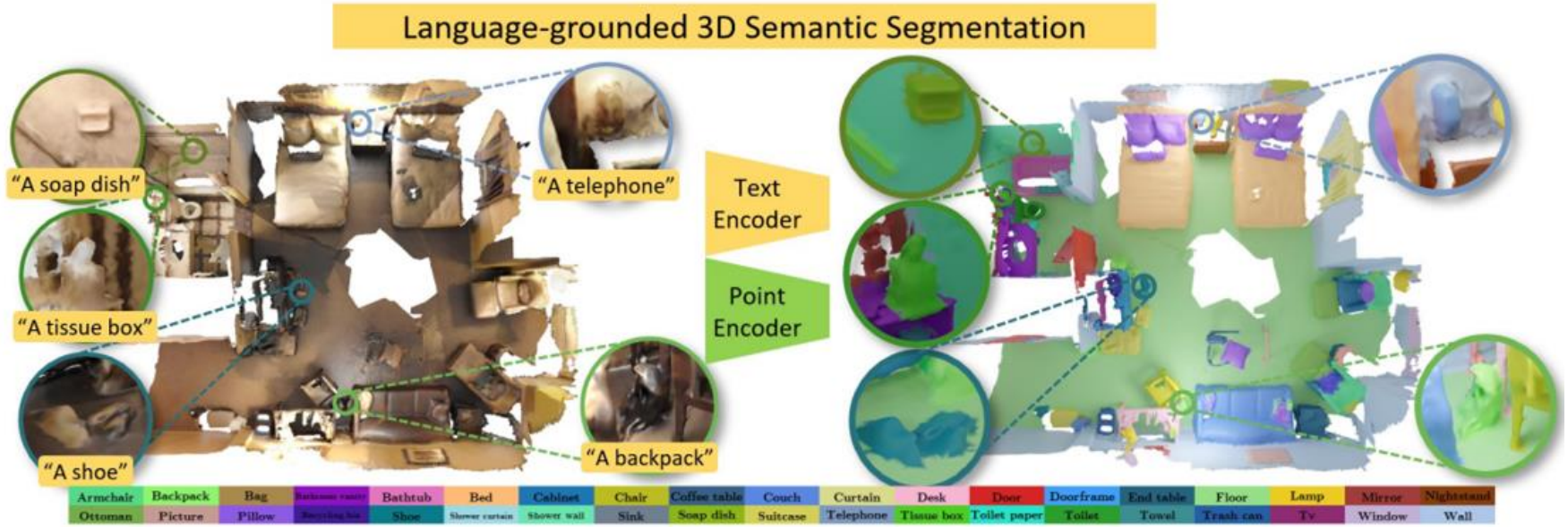
Peng, Songyou, et al. "Openscene: 3d scene understanding with open vocabularies." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. (Google Research, ETH Zurich, MPI for Intelligent Systems, Waymo LLC, Simon Fraser University)

OpenSUN3D Workshop Challenge on 3D Open-Vocabulary Scene Understanding (ICCV2023)

- TASK: Given an open-vocabulary, text-based query, the aim is to localize and segment the object instances that fit best with the given prompt, which might describe object properties such as semantics, material type, affordances and situational context.
- INPUT: An RGB-D sequence and the 3D reconstruction of a given scene, camera parameters, and a text-based input query.
- OUTPUT: Instance segmentation of the point cloud that corresponds to the vertices of the provided 3D mesh reconstruction, segmenting the objects that fit best with the given prompt.

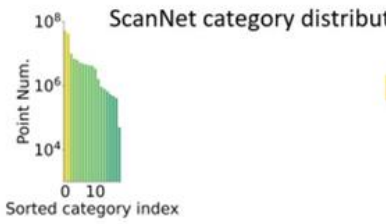


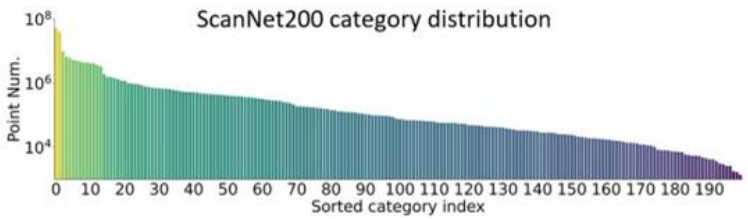
Language-Grounded Indoor 3D Semantic Segmentation in the Wild

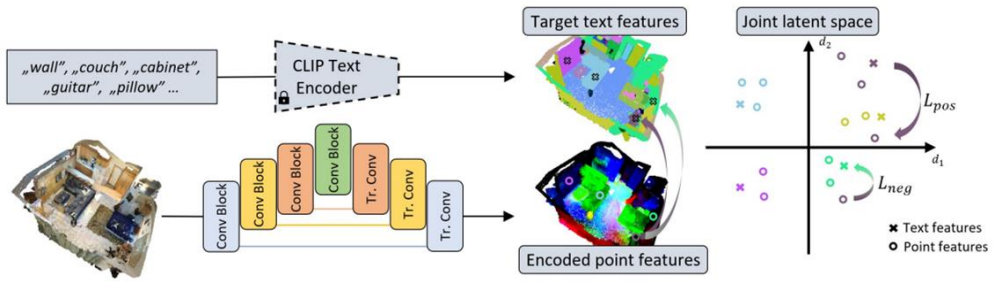


Project

Github

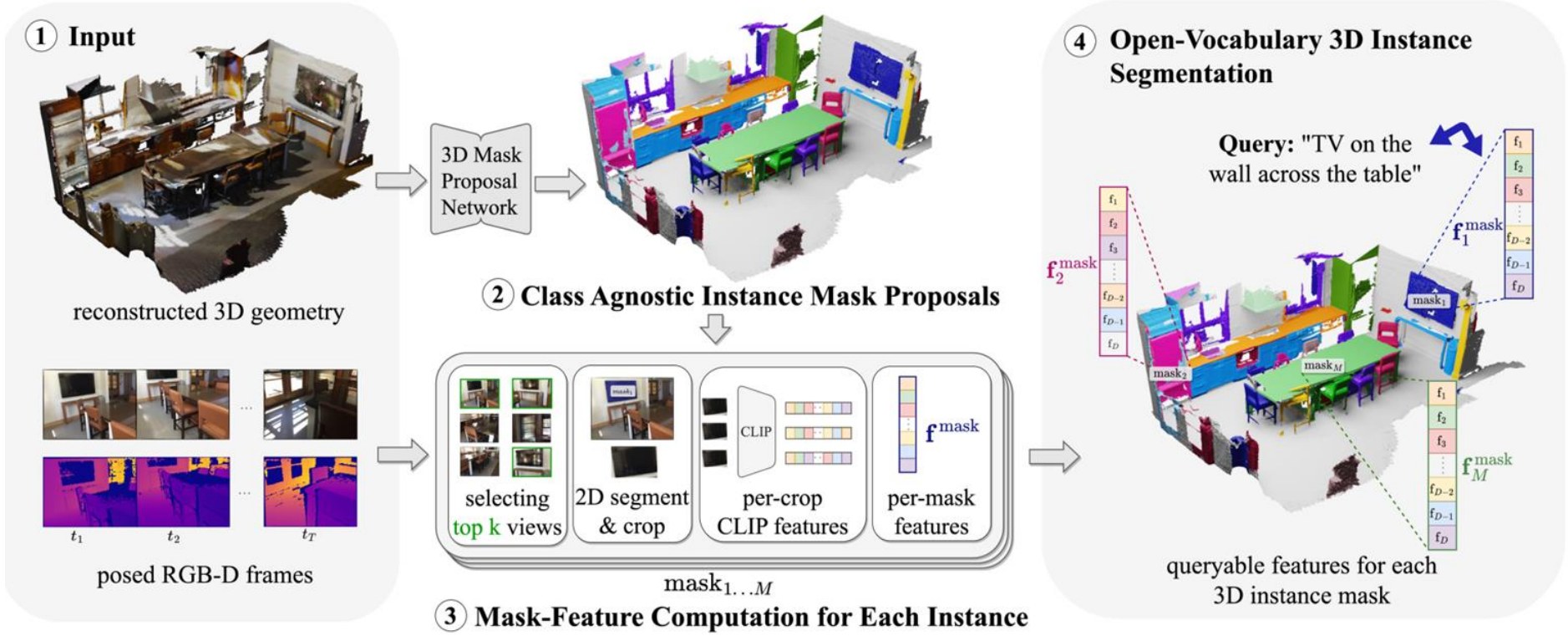






Rozenberszki, D., Litany, O., & Dai, A. (2022, October). Language-grounded indoor 3d semantic segmentation in the wild. In ECCV (pp. 125-141). (Technical University of Munich, NVIDIA)

OpenMask3D: Open-Vocabulary 3D Instance Segmentation

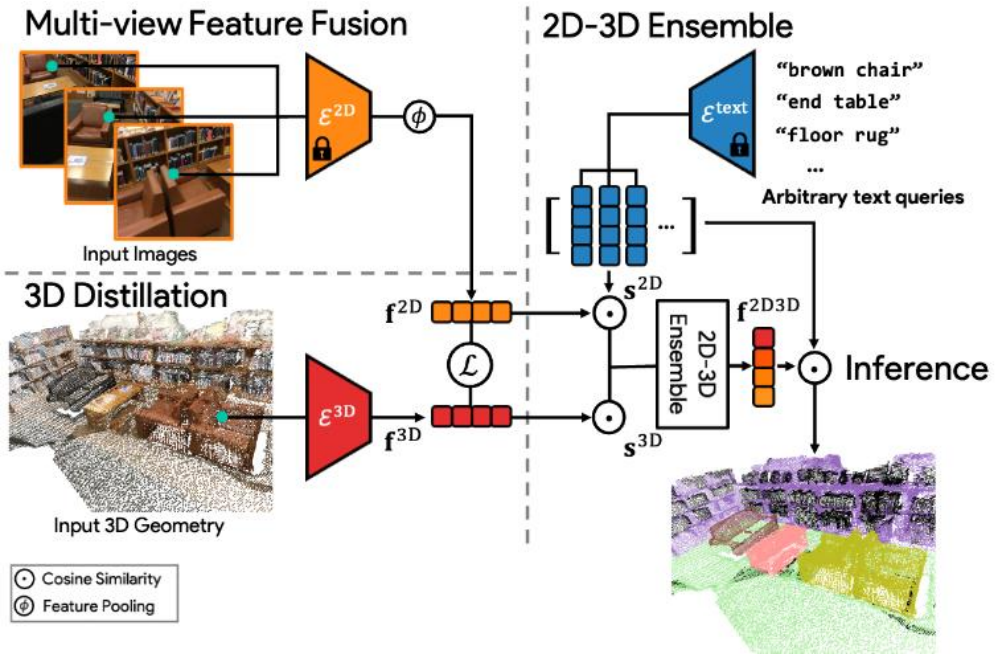


Project

Takmaz, A., Fedele, E., Sumner, R. W., Pollefeys, M., Tombari, F., & Engelmann, F. (2023). OpenMask3D: Open-Vocabulary 3D Instance Segmentation. arXiv preprint arXiv:2306.13631.
(ETH Zurich, ETH AI Center, Google Zurich)

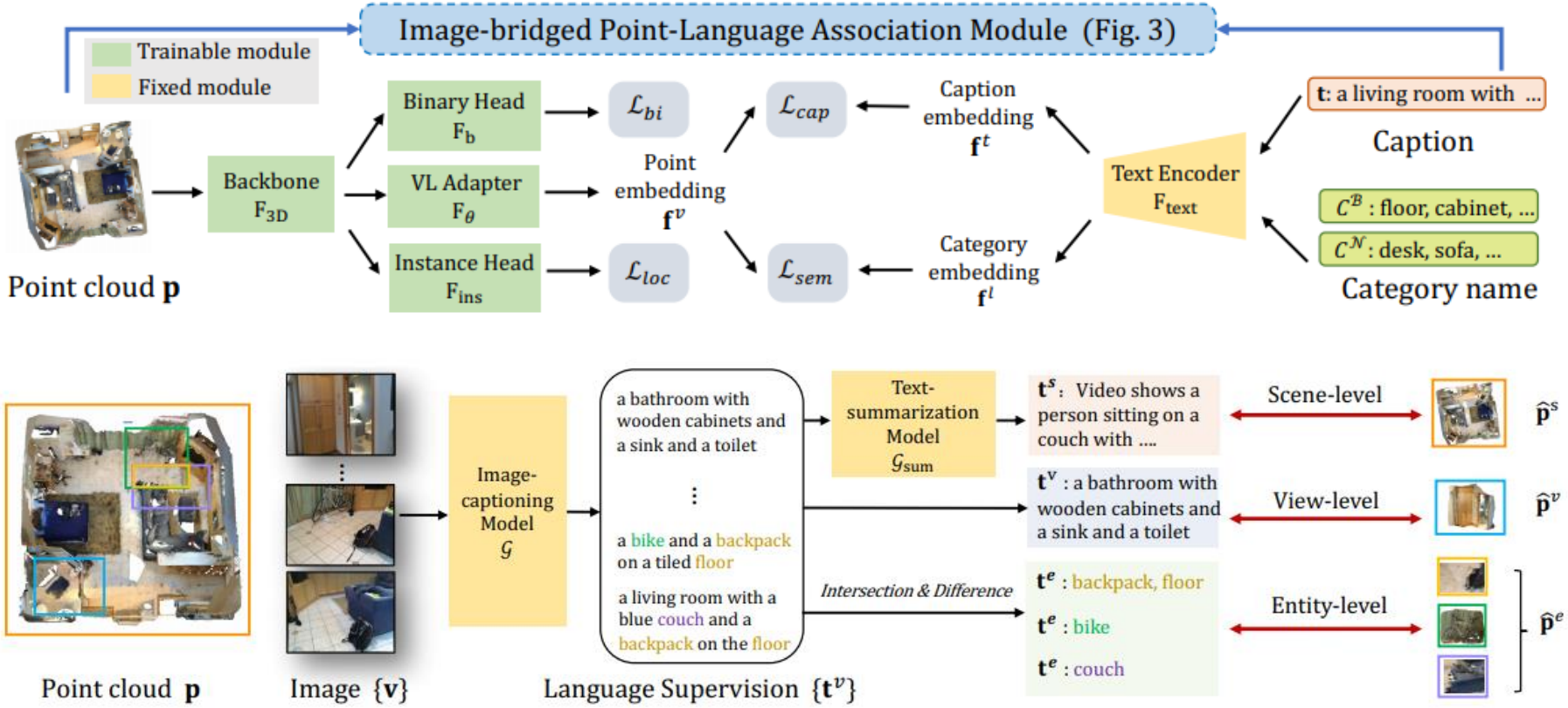
OpenScene: 3d scene understanding with open vocabularies

Produce text-image-3D co-embeddings



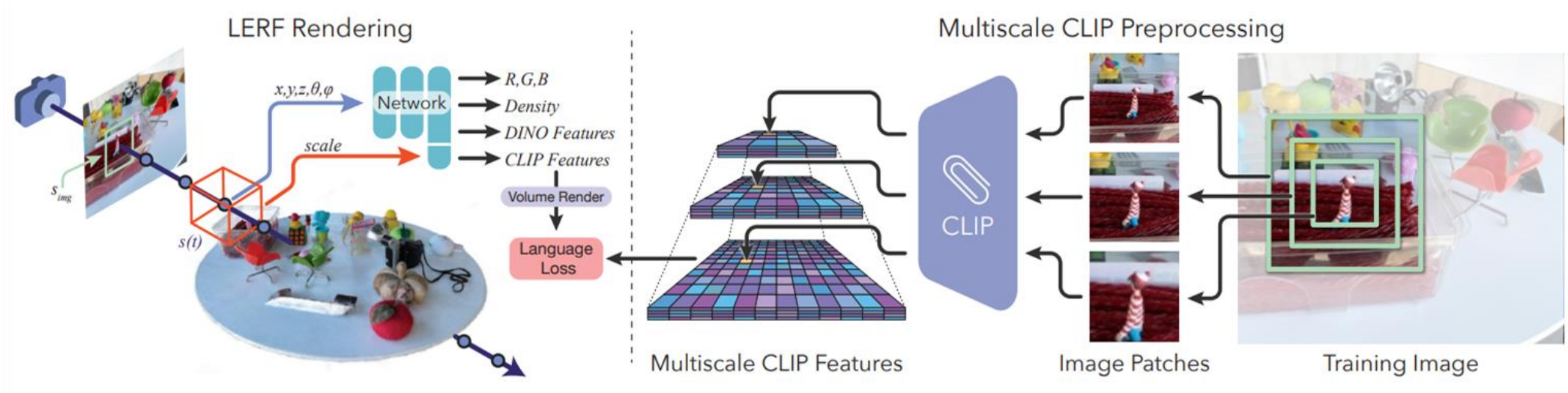
Peng, Songyou, et al. "Openscene: 3d scene understanding with open vocabularies." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. (Google Research, ETH Zurich, MPI for Intelligent Systems, Waymo LLC, Simon Fraser University)

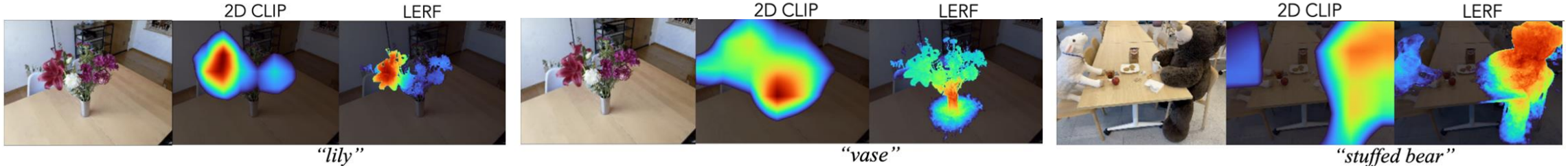
PLA: Language-Driven Open-Vocabulary 3D Scene Understanding



Ding, Runyu, et al. "PLA: Language-Driven Open-Vocabulary 3D Scene Understanding." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. (The University of Hong Kong, ByteDance)

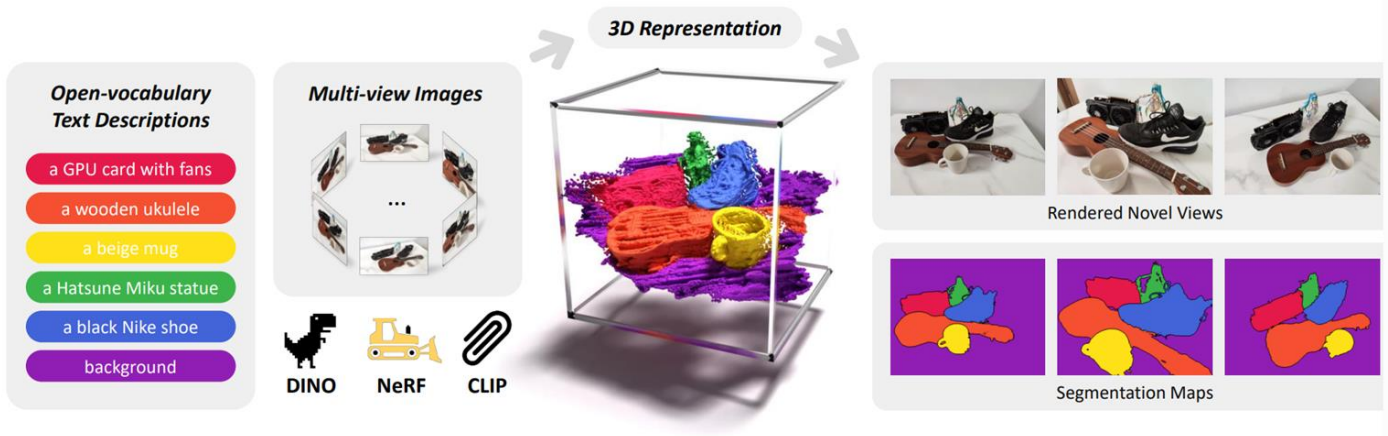
LERF: Language Embedded Radiance Fields



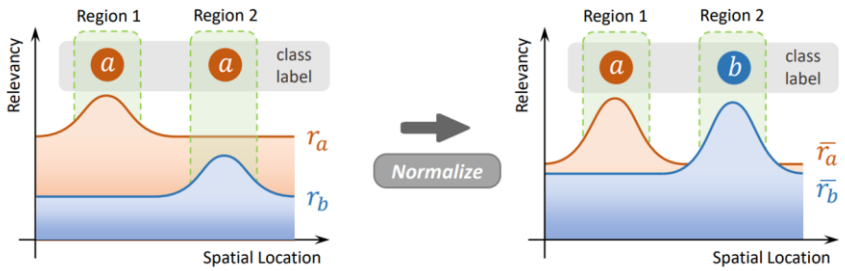


Kerr, J., Kim, C. M., Goldberg, K., Kanazawa, A., & Tancik, M. (2023). Lerf: Language embedded radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 19729-19739). (UC Berkeley)

3D-OVS: Weakly Supervised 3D Open-vocabulary Segmentation



Project
Github

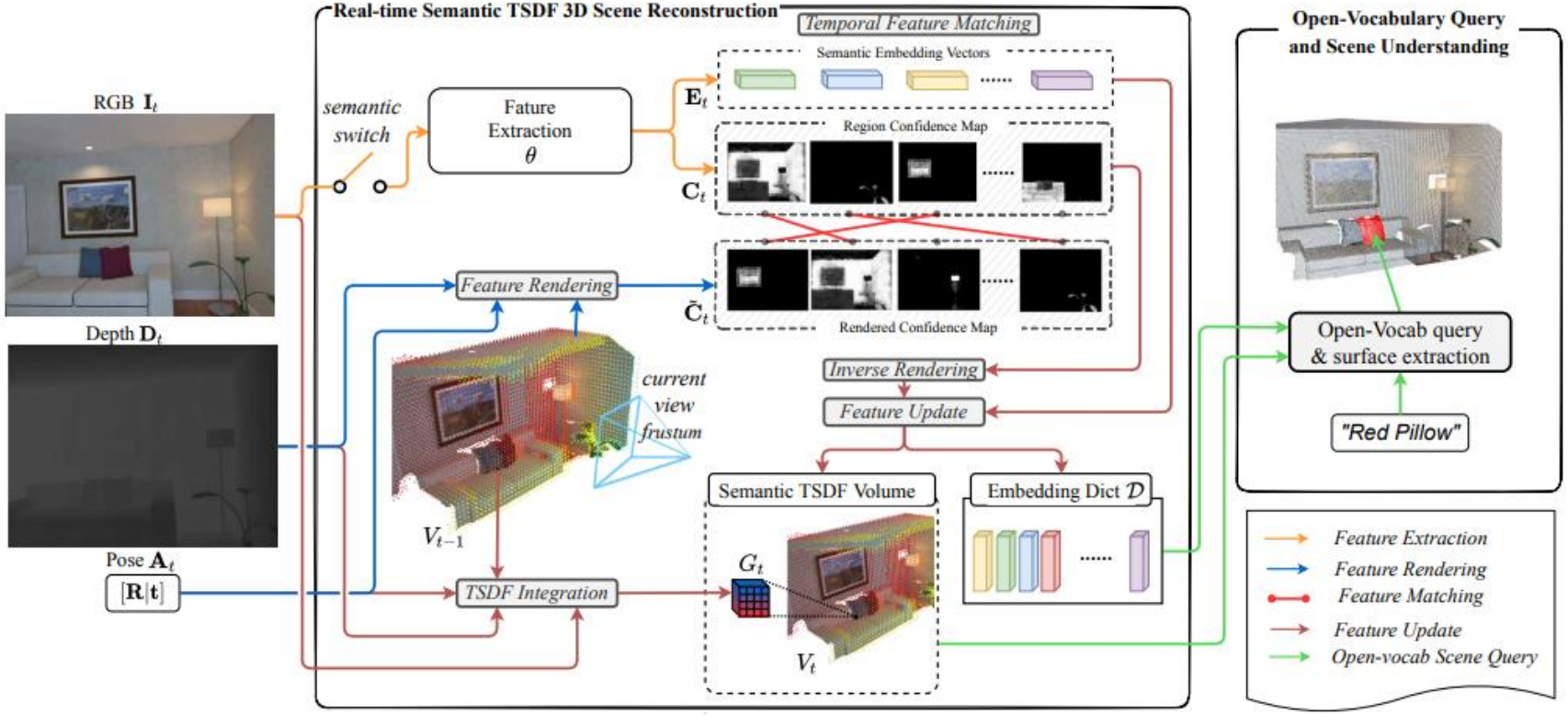


	Methods	bed		sofa		lawn		room		bench		table	
		mIoU	mAP	mIoU	mAP	mIoU	mAP	mIoU	mAP	mIoU	mAP	mIoU	mAP
2D	LSeg [8]	56.0	87.6	04.5	16.5	17.5	77.5	19.2	46.1	06.0	42.7	07.6	29.9
	ODISE [16]	52.6	86.5	48.3	35.4	39.8	82.5	52.5	59.7	24.1	39.0	39.7	34.5
	OV-Seg [19]	79.8	40.4	66.1	69.6	81.2	92.1	71.4	49.1	88.9	89.2	80.6	65.3
3D	FFD [4]	56.6	86.9	03.7	09.5	42.9	82.6	25.1	51.4	06.1	42.8	07.9	30.1
	Sem(ODISE) [45]	50.3	86.5	27.7	22.2	24.2	80.5	29.5	61.5	25.6	56.4	18.4	30.8
	Sem(OV-Seg) [45]	89.3	96.7	66.3	89.0	87.6	95.4	53.8	81.9	94.2	98.5	83.8	94.6
	LERF [2]	73.5	86.9	27.0	43.8	73.7	93.5	46.6	79.8	53.2	79.7	33.4	41.0
	3D-OVS	89.5	96.7	74.0	91.6	88.2	97.3	92.8	98.9	89.3	96.3	88.8	96.5

Mitigating CLIP features' ambiguities with normalized relevancy maps.

Liu, K., Zhan, F., Zhang, J., Xu, M., Yu, Y., Saddik, A. E., ... & Lu, S. (2023). Weakly Supervised 3D Open-vocabulary Segmentation. arXiv preprint arXiv:2305.14093. NeurIPS 2023

Open-Fusion: Real-time Open-Vocabulary 3D Mapping and Queryable Scene Representation



Map	Method	Representation	Foundation Model	Feature Level	Real-time [1]	Scene-specific	Sem-Query [2]
2D	CoW [8]	point	CLIP [3] + GradCAM [9]	point	-	✗	$O(P)$
	NLMap [10]	point	ViLD [11] + CLIP [3]	bbox	-	✗	$O(P)$
	VLMap [12]	point	LSeg [13]	point	✗	✗	$O(P)$
3D	CLIP-Fields [14]	NeRF	Detic [15] + CLIP [3]	bbox	✗	✓	-
	LERF [16]	NeRF	CLIP [3]	image patch	✗	✓	-
	SemAbs[17]	occupancy	CLIP [3] + GradCAM [9]	point	✗	✗	$O(P)$
	ConceptFusion [18]	point	SAM [19] + CLIP [3]	bbox	✗	✗	$O(P)$
	Open-Fusion	TSDF	SEEM [20]	region	✓	✗	$O(M)$

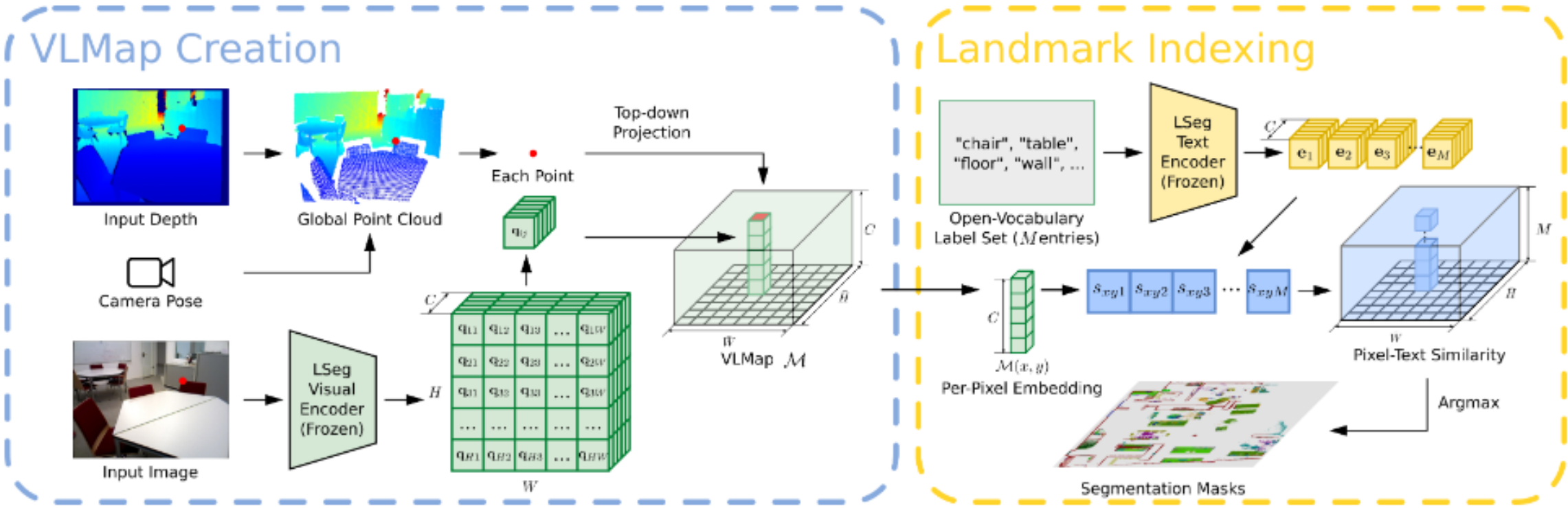
QUANTITATIVE COMPARISON OF **OPEN-SET SEMANTIC SEGMENTATION** AND **3D SCENE REPRESENTATION TIME** BETWEEN OPEN-FUSION AND EXISTING METHODS ON THE SCANNET DATASET.

	Method	Time (FPS)↑ 3D-Rec.[1]	Time (FPS)↑ Sem-3D-Rec[2]	Accuracy↑ mAcc	Accuracy↑ f-mIoU
Priv.[3]	LSeg	-	-	0.70	0.63
	OpenSeg	-	-	0.63	0.62
	CLIPSeg (rd64-uni)	-	-	0.41	0.34
	CLIPSeg (rd16-uni)	-	-	0.41	0.36
ZS.[4]	MaskCLIP	-	-	0.24	0.28
	ConceptFusion	1.5	0.15	**0.63**	0.58
	Open-Fusion	**50**	**4.5**	0.62	**0.59**

Project
Github

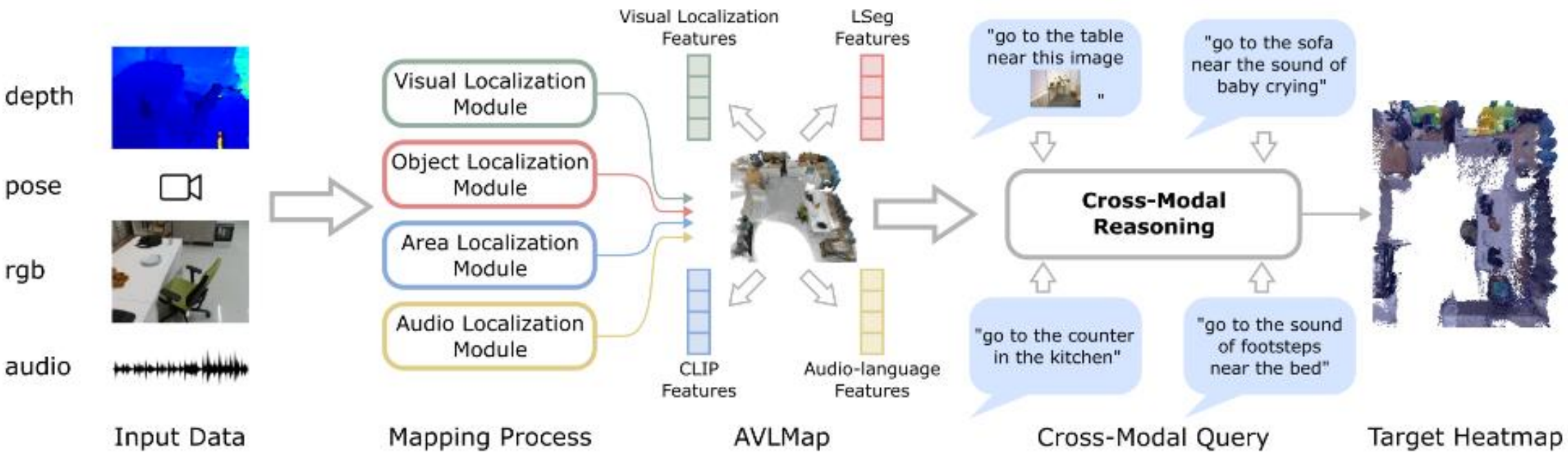
Yamazaki, Kashu, et al. "Open-Fusion: Real-time Open-Vocabulary 3D Mapping and Queryable Scene Representation." arXiv preprint arXiv:2310.03923 (2023).

VLMaps: Visual language maps for robot navigation



Huang, Chenguang, et al. "Visual language maps for robot navigation." 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023. (Freiburg University, Google Research, University of Technology Nuremberg)

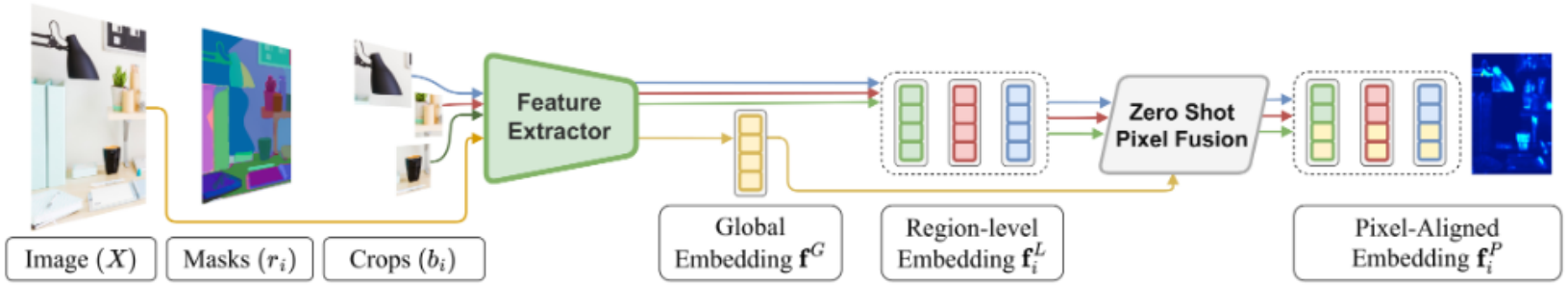
AVLMaps: Audio visual language maps for robot navigation



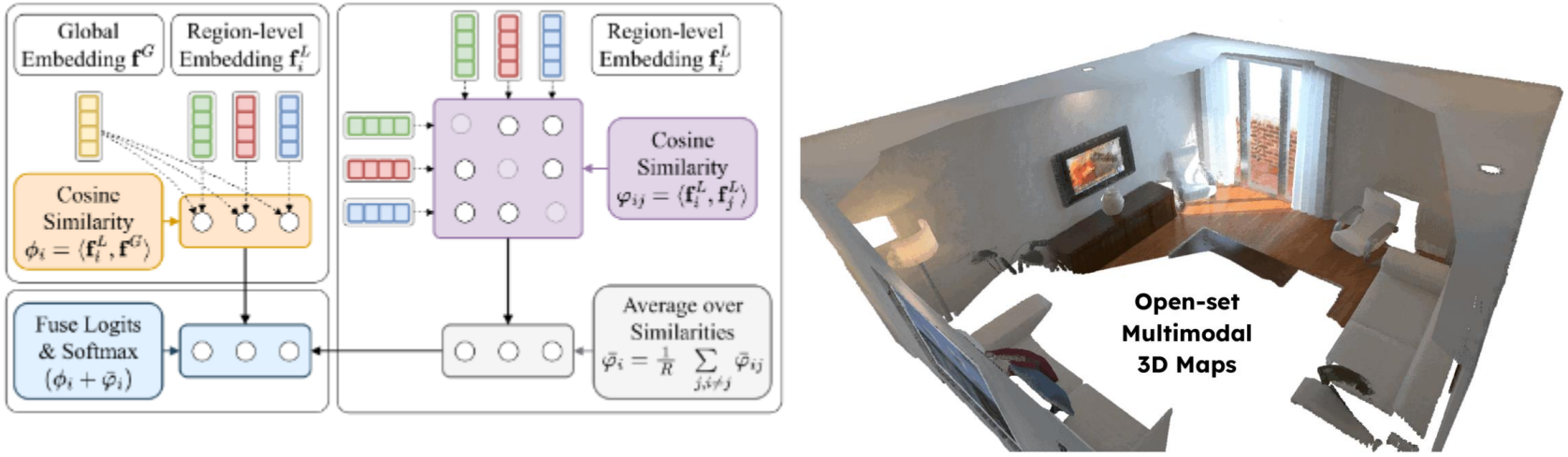
Huang, Chenguang, et al. "Audio visual language maps for robot navigation." arXiv preprint arXiv:2303.07522, 2023.
(Freiburg University, Google Research, University of Technology Nuremberg)

ConceptFusion: Open-set multimodal 3d mapping

Construct pixel-aligned features



Zero Shot Pixel Fusion



Jatavallabhula, Krishna Murthy, et al. "Conceptfusion: Open-set multimodal 3d mapping." arXiv preprint arXiv:2302.07241, 2023. (MIT, Universite de Montreal, University of Toronto, IIIT Hyderabad, CMU, Amazon, Matician, DEVCOM Army Research Laboratory)

04

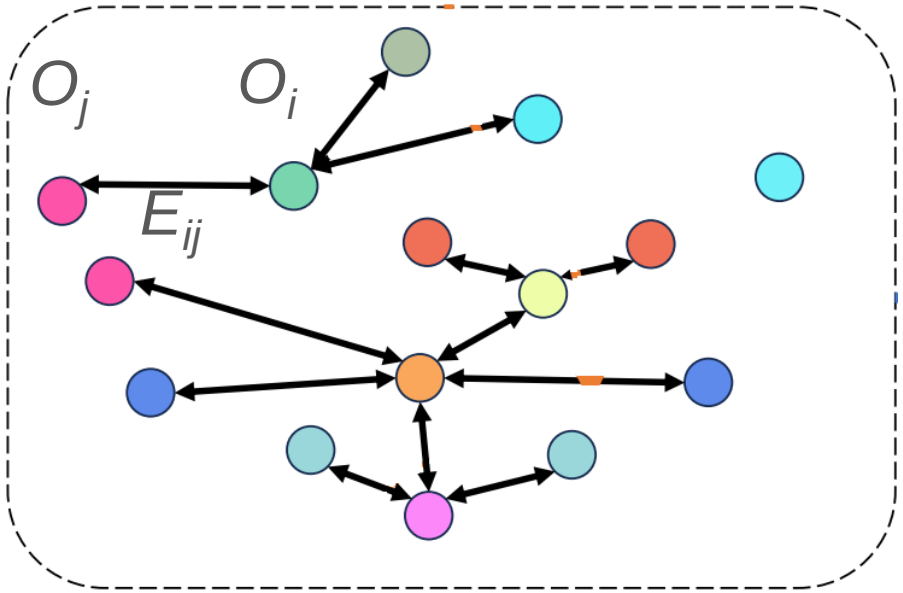
Разреженные (Sparse) методы

Постановка задачи

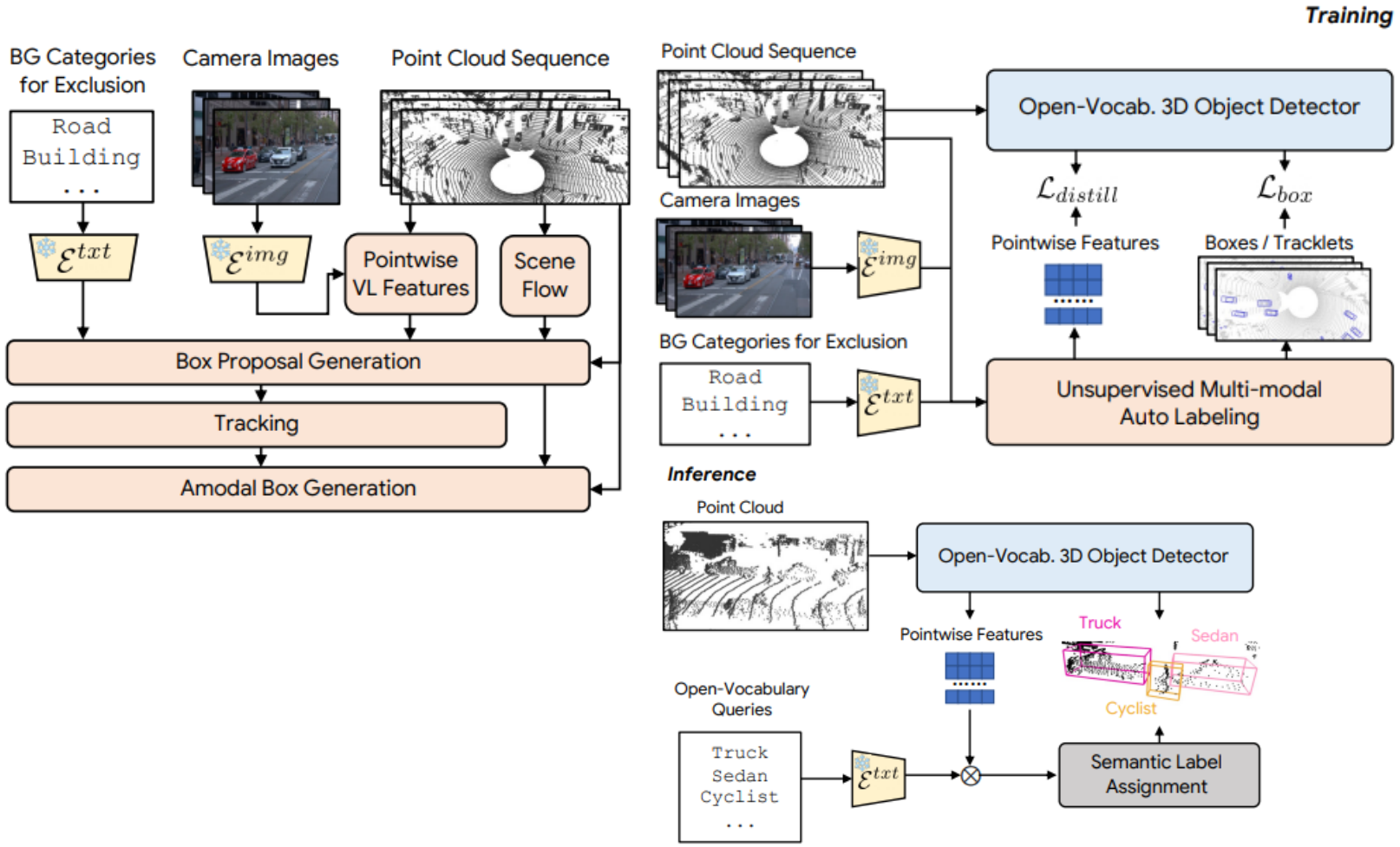
$$\text{I: } Map = M\big((\mathrm{RGB}_1 \dots \mathrm{RGB}_N), (\mathrm{D}_1 \dots \mathrm{D}_N), (\mathrm{Pose}_1 \dots \mathrm{Pose}_N)\big) = G(O, E),$$

$$\text{II: } \mathrm{O}_{\mathrm{Query}} = \mathrm{Q}(\mathrm{Map}, \mathrm{Query}).$$

$$\mathrm{O_i} = (\mathrm{X_{ic}}, \mathrm{Y_{ic}}, \mathrm{Y_{ic}}, \mathrm{Embedding_i}, \mathrm{PCL_i}, \mathrm{Caption_i}, \mathrm{Tag_i}).$$



Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving



Algorithm 1 Unsupervised multi-modal auto labeling.

Input: A sequence of images across T frames for each of the K cameras $\{\mathbf{I}_t^k\}$; a sequence of LiDAR point locations $\{\mathbf{P}_t\}$.

Requires: Cosine similarity threshold for background categories ϵ^{bg}; minimum scene flow magnitude ϵ^{sf}; maximum ratio of background points within a box r^{bg}; a set of a prior background categories $\mathbf{C}^{bg}$; a pre-trained open-vocabulary model with image encoder $\mathcal{E}^{img}$ and text encoder $\mathcal{E}^{txt}$.

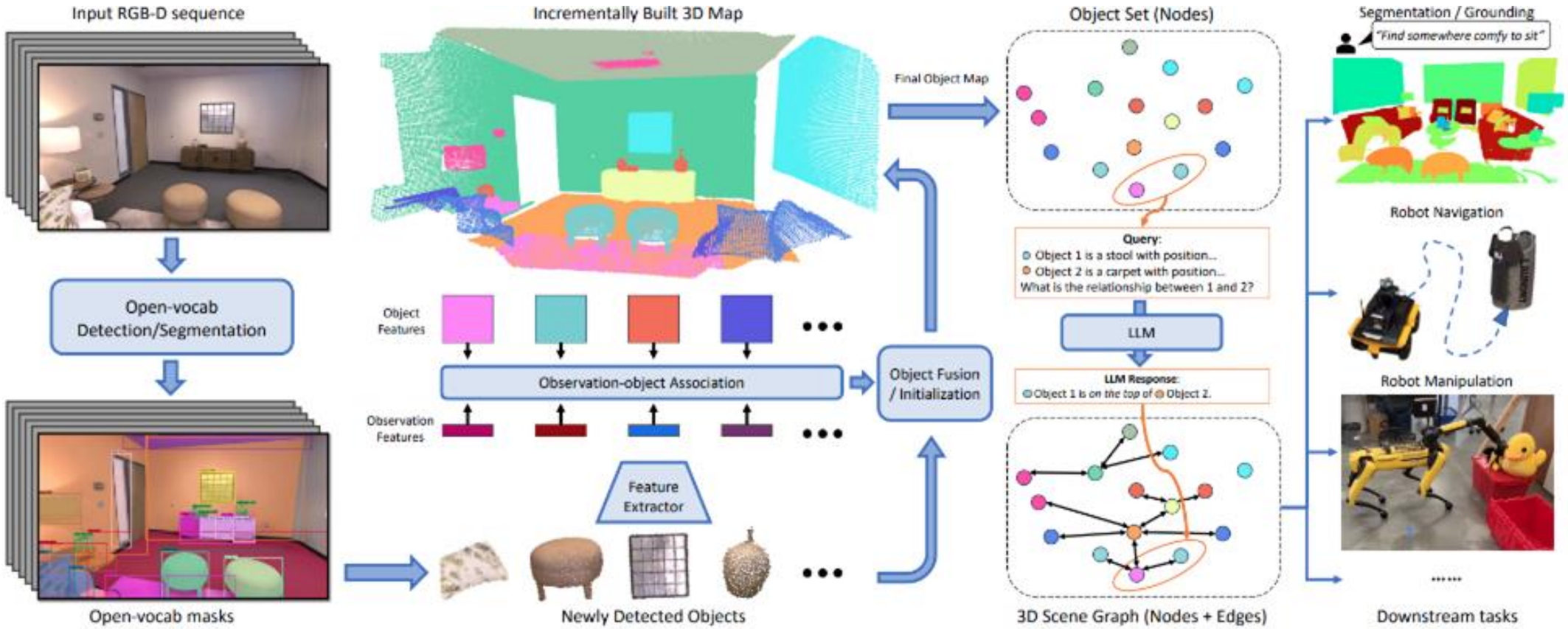
Output: Amodal 3D bounding boxes $\{\mathbf{B}_t\}$ and their track IDs $\{\mathbf{T}_t\}$; point-wise open-vocabulary features $\{\mathbf{F}_t^{vl}\}$.

Function:

for $t = 1$ to T **do**
  $\{\mathbf{V}_t^k\} \leftarrow \mathcal{E}^{img}(\{\mathbf{I}_t^k\})$ ▷ 2D VL features
  $\mathbf{F}_t^{vl} \leftarrow \text{Unprojection}(\{\mathbf{V}_t^k\}, \mathbf{P}_t)$ ▷ 3D VL features
  if $t \neq T$ **then**
    $\mathbf{F}_t^{sf} \leftarrow \text{NSFP++}(\mathbf{P}_t, \mathbf{P}_{t+1})$ ▷ Scene flow
  else
    $\mathbf{F}_t^{sf} \leftarrow -\text{NSFP++}(\mathbf{P}_t, \mathbf{P}_{t-1})$
  for $i = 1$ to N_t **do**
    $(\mathbf{M}_t^{sf})_i \leftarrow \mathbb{1}(\|(\mathbf{F}_t^{sf})_i\| \geq \epsilon^{sf})$
    $(\mathbf{M}_t^{bg})_i \leftarrow \mathbb{1}(\max_{c \in \mathbf{C}^{bg}} \frac{(\mathbf{F}_t^{vl})_i \cdot \mathcal{E}^{txt}(c)}{\|(\mathbf{F}_t^{vl})_i\| \|\mathcal{E}^{txt}(c)\|} \geq \epsilon^{bg})$
  $\widetilde{\mathbf{P}}_t, \widetilde{\mathbf{F}}_t^{sf} \leftarrow \mathbf{P}_t[\mathbf{M}_t^{sf}], \mathbf{F}_t^{sf}[\mathbf{M}_t^{sf}]$
  $\mathbf{B}_t^{vis} \leftarrow \text{InitialBoxProposal}(\widetilde{\mathbf{P}}_t, \widetilde{\mathbf{F}}_t^{sf}, \mathbf{M}_t^{bg}; r^{bg})$
$\{\mathbf{T}_t\} \leftarrow \text{Tracking}(\{\mathbf{B}_t^{vis}\})$
$\{\mathbf{B}_t\} \leftarrow \text{AmodalBoxGeneration}(\{\mathbf{B}_t^{vis}\}, \{\mathbf{T}_t\}, \{\mathbf{P}_t\})$
return $\{\mathbf{B}_t\}, \{\mathbf{T}_t\}, \{\mathbf{F}_t^{vl}\}$

Najibi, Mahyar, et al. "Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning



Project
Github

Gu, Qiao, et al. "ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning." arXiv preprint arXiv:2309.16650, 2023 (MIT, Universite de Montreal, University of Toronto, IIIT Hyderabad, JHU APL, JHU, UMass Amherst, DEVCOM Army Research Laboratory).

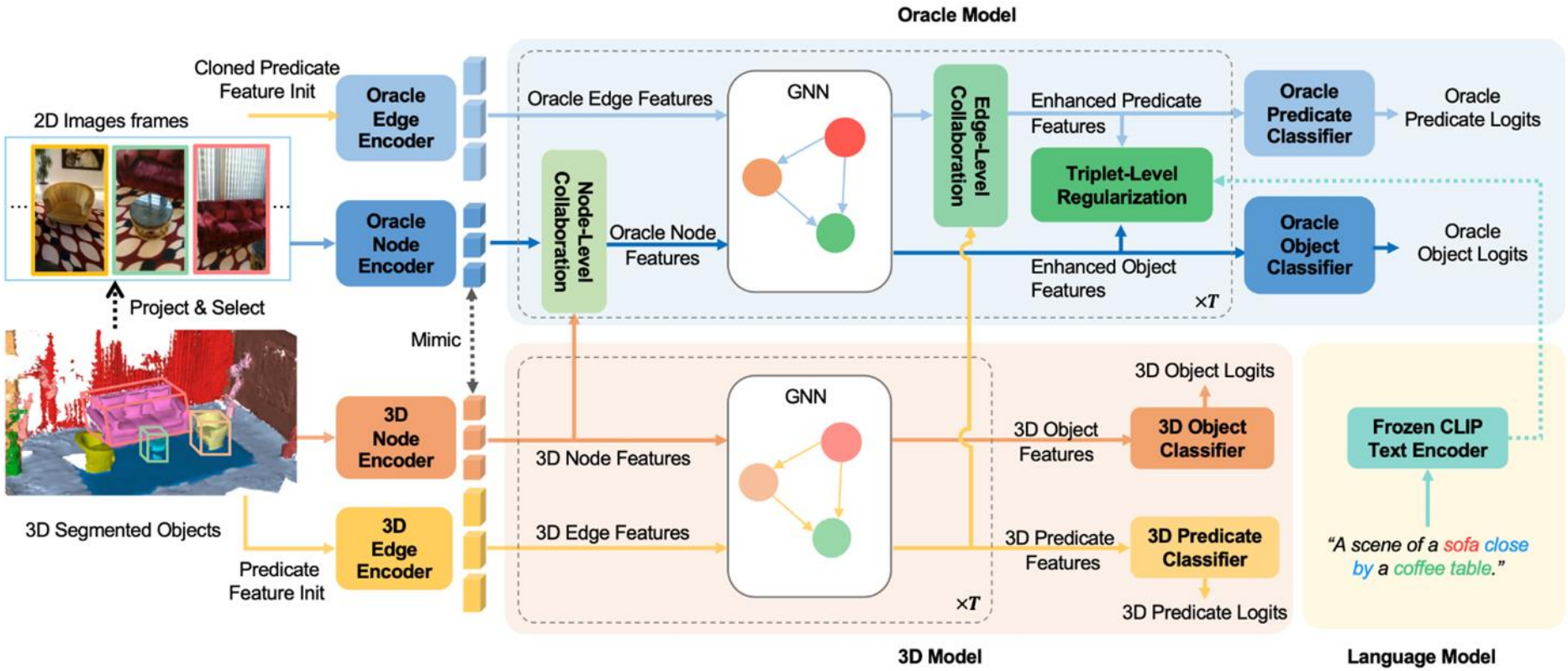
VL-SAT: Visual-Linguistic Semantics Assisted Training for 3D Semantic Scene Graph Prec

Dataset used: 3DSSG, (scenes from 3RSCAN)

- 1553 scenes
- 160 cat. objects
- 26 cat. predicates

Target:

- build an accurate 3D scene graph, i.e.:
- classify objects
- classify relationships between objects

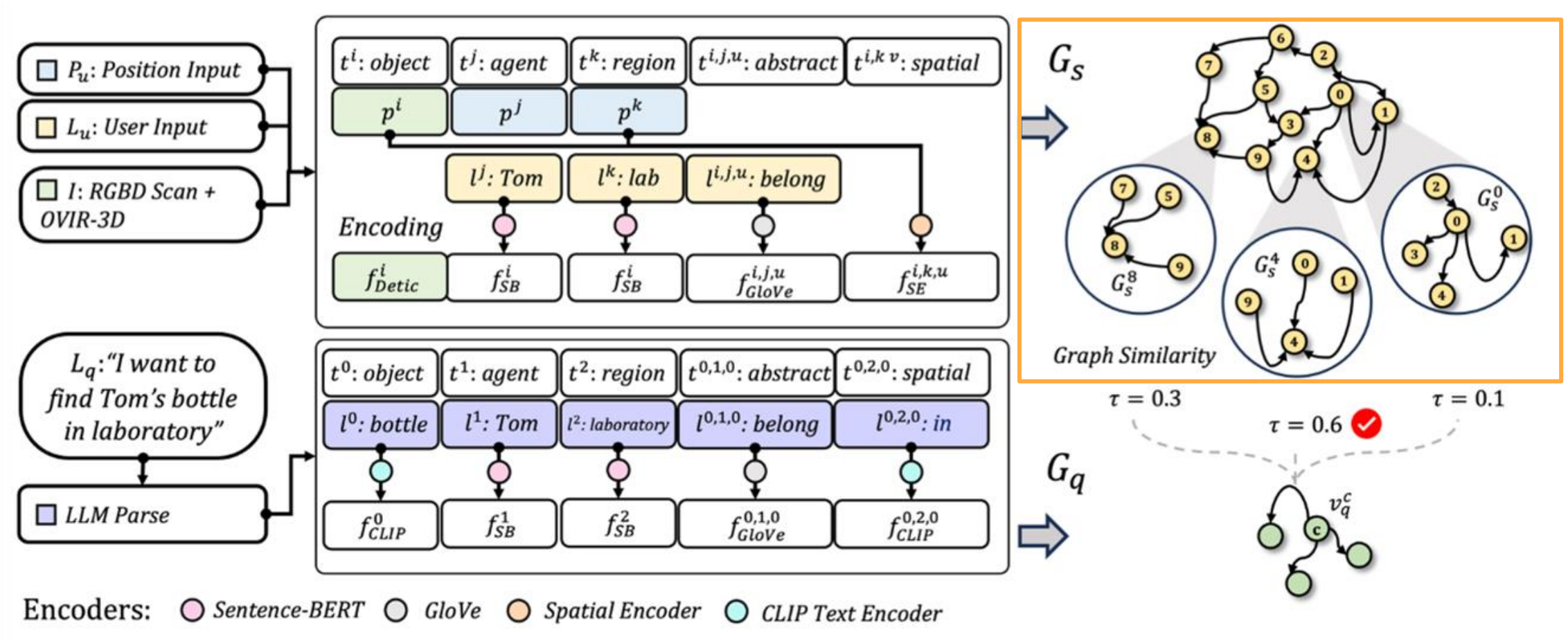


Wang, Z., Cheng, B., Zhao, L., Xu, D., Tang, Y., & Sheng, L. (2023). VL-SAT: Visual-Linguistic Semantics Assisted Training for 3D Semantic Scene Graph Prediction in Point Cloud. In Proceedings of the IEEE/CVF Conference on Computer

https://github.com/wz7in/CVPR2023-VLSAT

AIRI

Context-Aware Entity Grounding with Open-Vocabulary 3D Scene Graphs

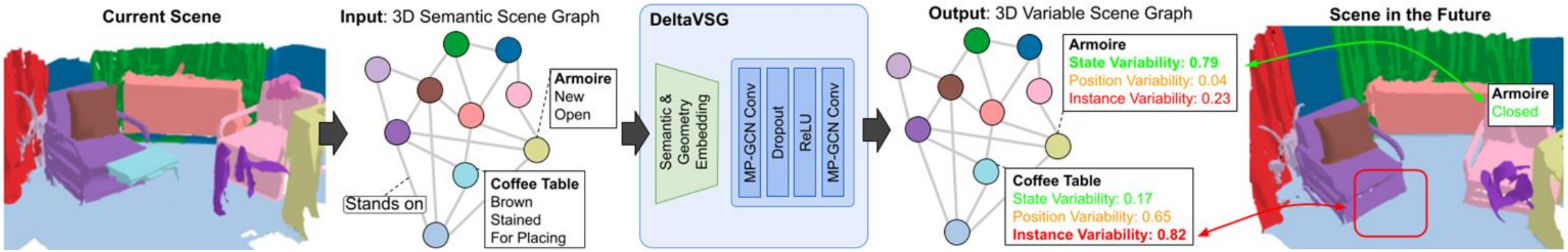


Chang, H., Boyalakuntla, K., Lu, S., Cai, S., Jing, E., Keskar, S., ... & Boularias, A. (2023). Context-aware entity grounding with open-vocabulary 3d scene graphs. *arXiv preprint arXiv:2309.15940*.

https://github.com/changhaonan/ovsg

AIRI

3D Variable Scene Graphs



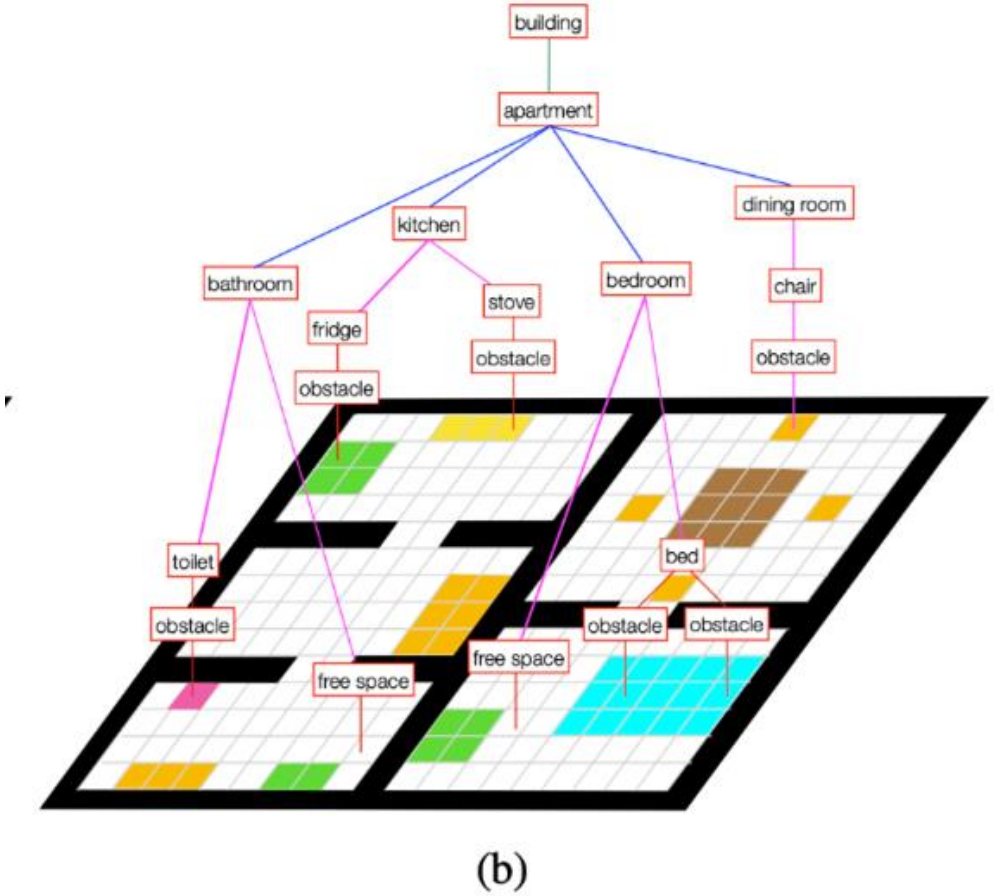
https://github.com/ethz-asl/3d_vsg

Looper, S., Rodriguez-Puigvert, J., Siegwart, R., Cadena, C., & Schmid, L. (2023, May). 3d vsg: Long-term semantic scene change prediction through 3d variable scene graphs. In *2023 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 8179-8186). IEEE.

Foundations of Spatial Perception for Robotics: Hierarchical Representations and Real-time Systems

(a) Original Image (b) OneFormer

(c) HRNet (d) Mobilenet



(b)

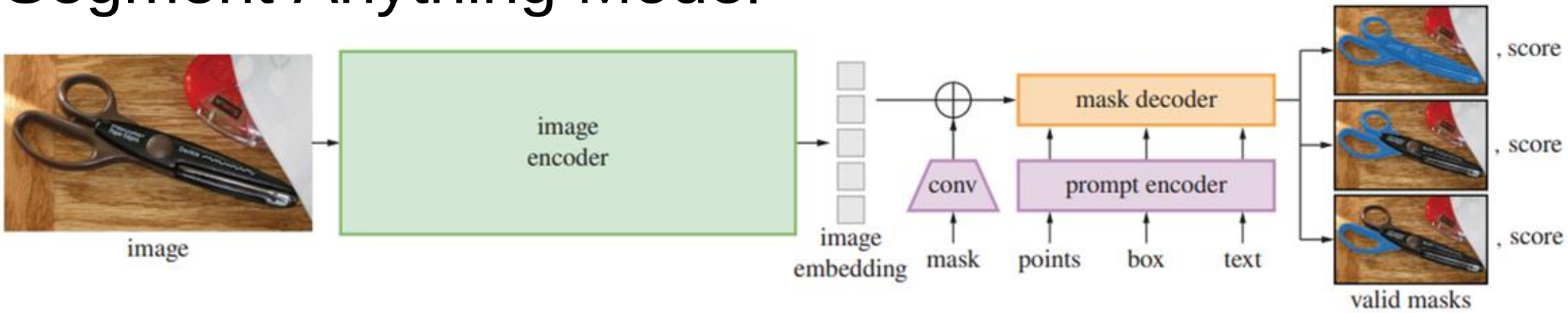
Hughes, N., Chang, Y., Hu, S., Talak, R., Abdulhai, R., Strader, J., & Carlone, L. (2023). Foundations of Spatial Perception for Robotics: Hierarchical Representations and Real-time Systems. *arXiv preprint arXiv:2305.07154*.

https://github.com/mit-spark/hydra

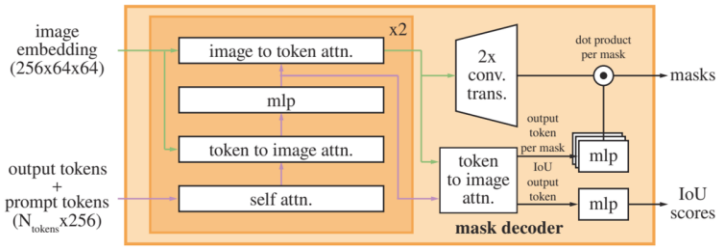
05

Базовые модели сегментации изображений с открытым словарем запросов (open vocabulary)

Segment Anything Model

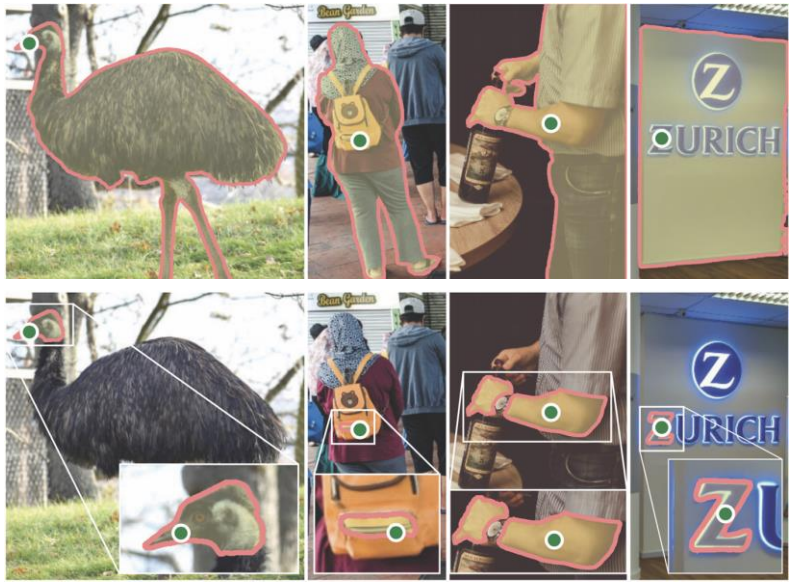


- Image encoder: Vision Transformer (ViT) ViT-H, ViT-L, ViT-B
- Prompt encoder: positional encoding for points and box corners, CLIP-based text encoder and convolutional mask encoder
- Mask decoder: lightweight attention-based model



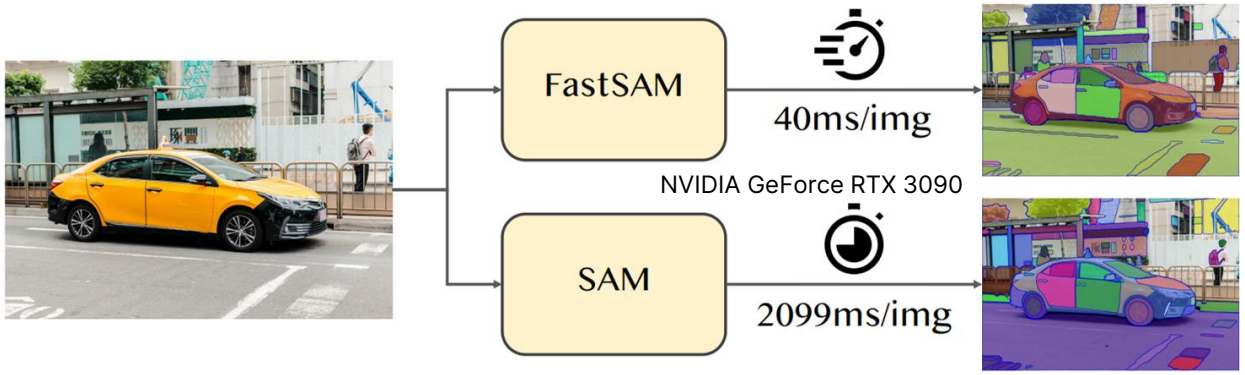
Project

Github

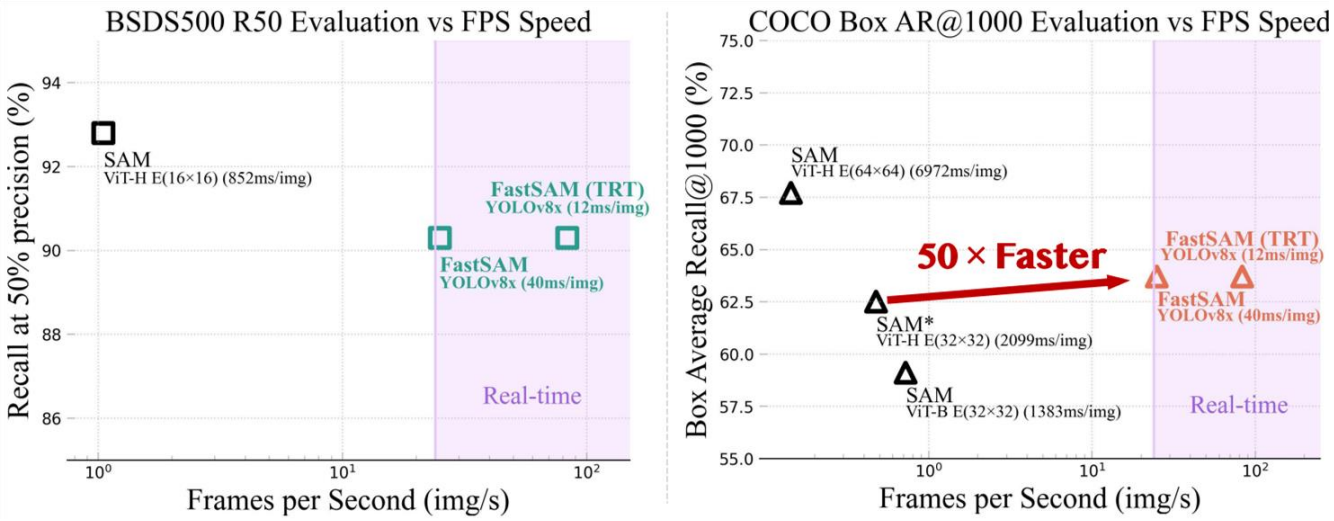


Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. arXiv preprint arXiv:2304.02643. ICCV2023 (Meta AI Research, FAIR)

Fast query-based segmentation: FastSAM



- Authors trained the existing real-time instance segmentation method YOLOv8-seg using only 1/50 of the SA-1B dataset published by SAM

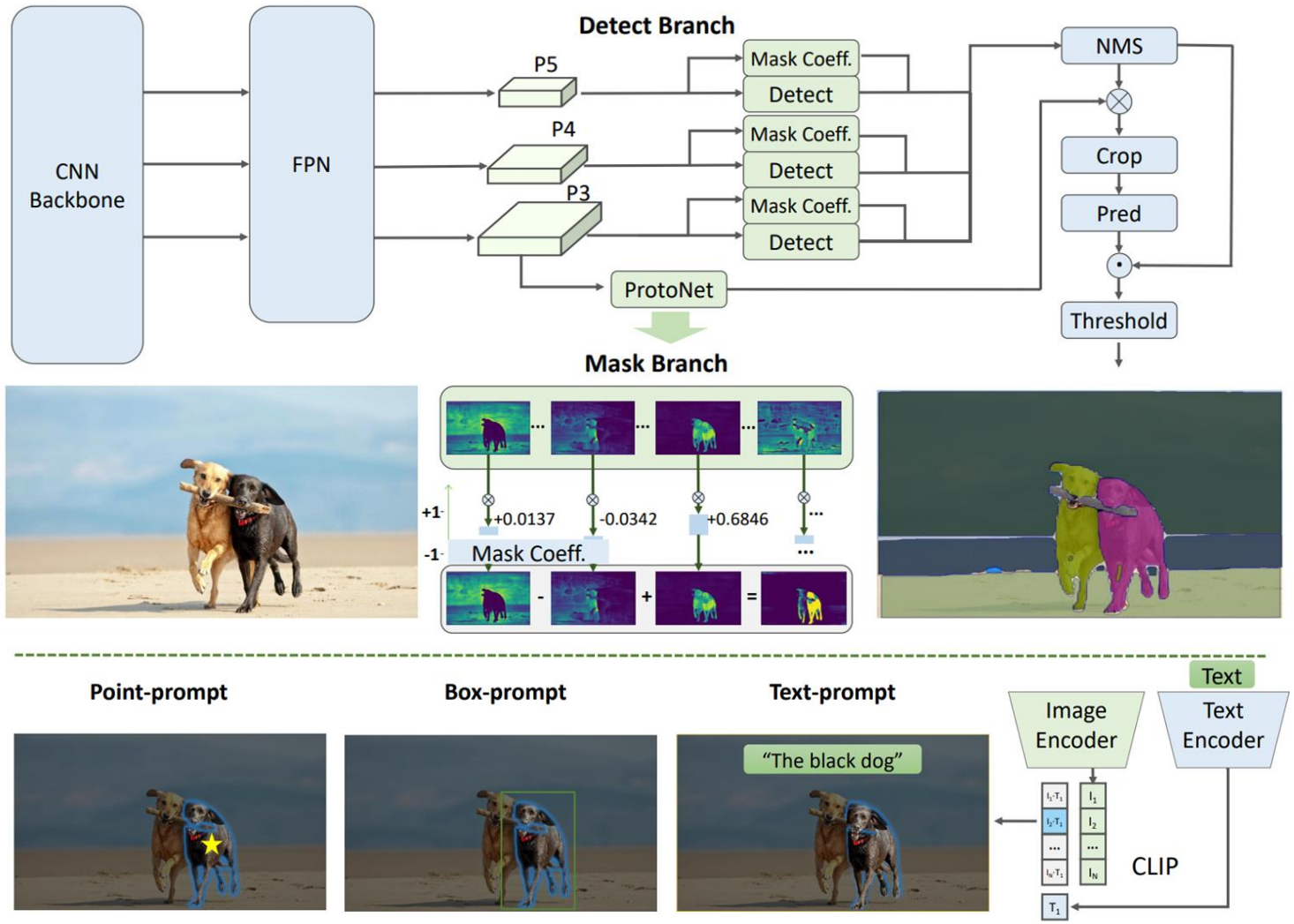


Project

Github

Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., ... & Wang, J. (2023). Fast Segment Anything. arXiv preprint arXiv:2306.12156.(Chinese Academy of Sciences, University of Chinese Academy of Sciences, Objecteye Inc., Wuhan AI Research)

Fast query-based segmentation: FastSAM

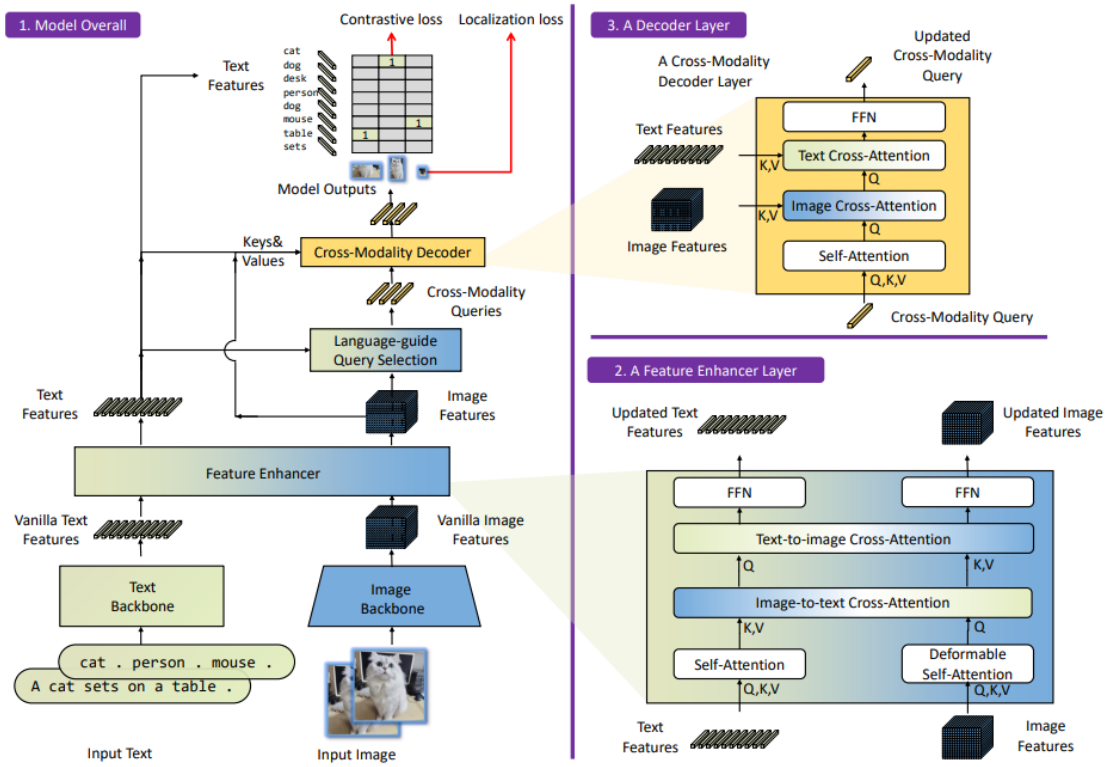


- The framework of FastSAM. It contains two stages: All-instance Segmentation (AIS) and Prompt-guided Selection (PGS).
- Authors use YOLOv8-seg to segment all objects or regions in an image.
- Then authors use various prompts to identify the specific object(s) of interest. It mainly involves the utilization of point prompts, box prompts, and text prompt. The text prompt is based on CLIP

Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., ... & Wang, J. (2023). Fast Segment Anything. arXiv preprint arXiv:2306.12156.(Chinese Academy of Sciences, University of Chinese Academy of Sciences, Objecteye Inc., Wuhan AI Research)

Fast query-based segmentation: Grounded Mobile SAM

Grounding DINO [1]



Mobile SAM [2]

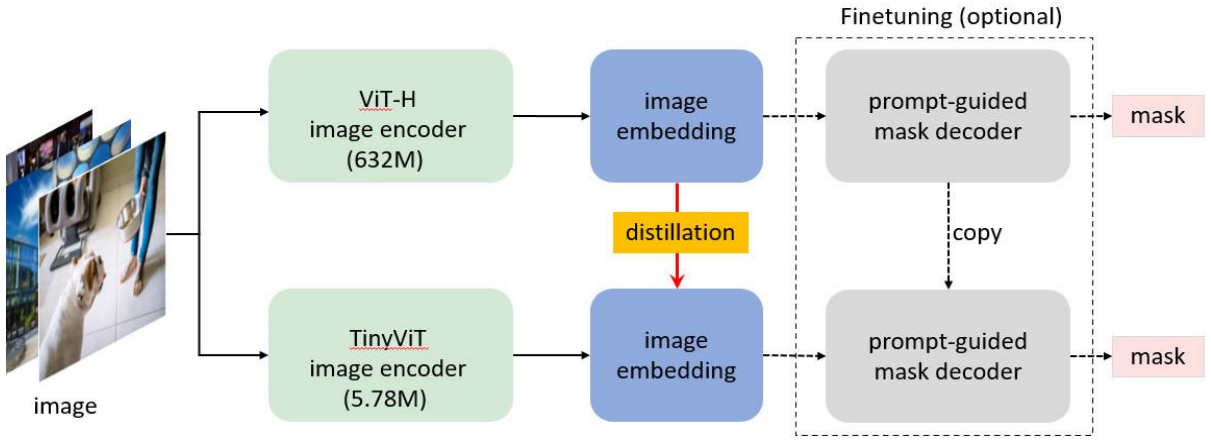


Table 6: Comparison between FastSAM and MobileSAM.

	FastSAM	MobileSAM	Ratio
Size	68M	9.66M	7
Speed	40ms	10ms	4

Table 7: mIoU comparison. With the assumption that the predicted mask from the original SAM is ground-truth, a higher mIoU indicates a better performance.

	100	200	300	400	500
FastSAM	0.27	0.33	0.37	0.41	0.41
MobileSAM	0.73	0.71	0.74	0.73	0.73

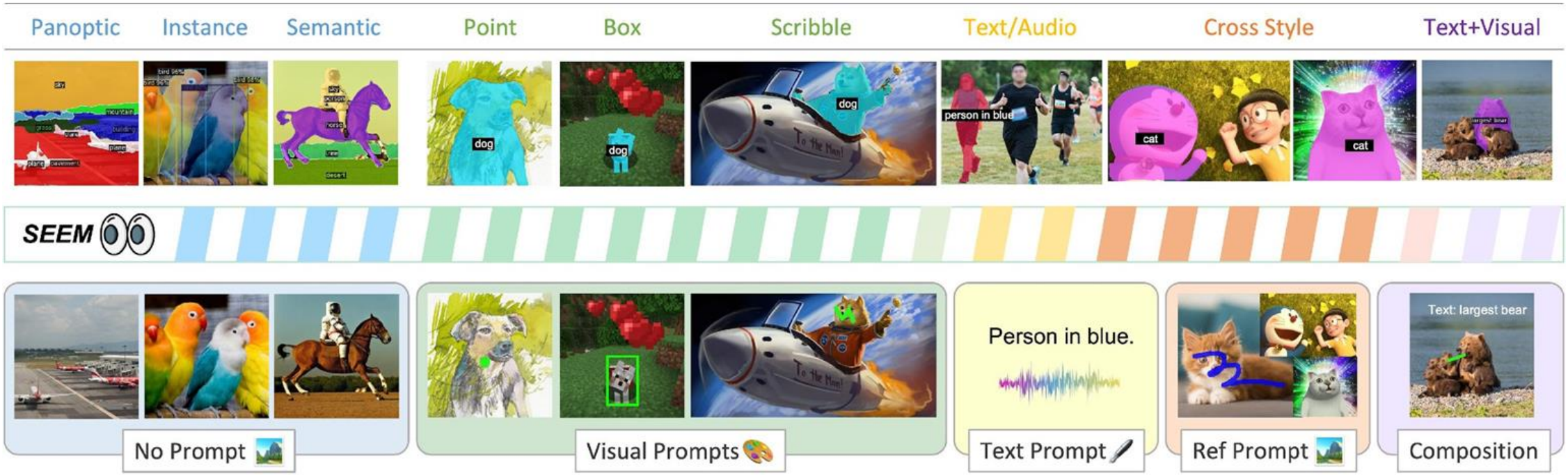
Project

Github

[1] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., ... & Zhang, L. (2023). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499. (Tsinghua University, IDEA, The Hong Kong University of Science and Technology, The Chinese University of Hong Kong (Shenzhen), Microsoft Research, Redmond)

[2] Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S. H., Lee, S., & Hong, C. S. (2023). Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. arXiv preprint arXiv:2306.14289. (Kyung Hee University)

SEEM: Segment Everything Everywhere All at Once



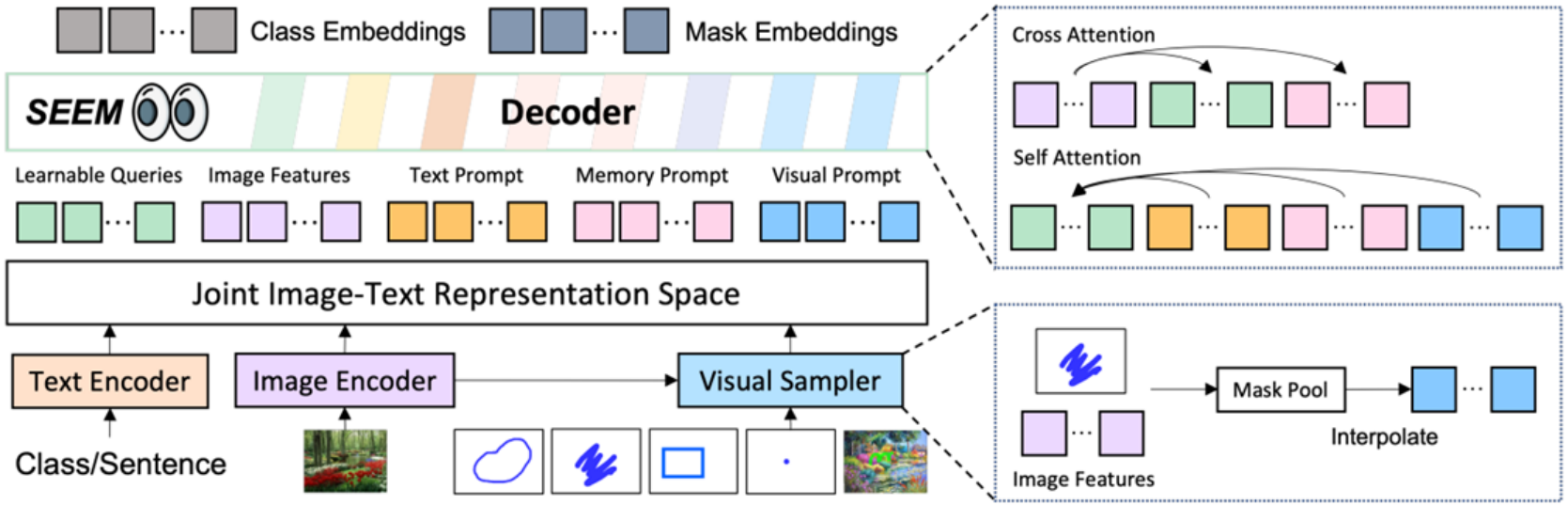
- Authors design a new prompting scheme that can encode various user intents into prompts in a joint visual-semantic space, enabling strong flexibility for various segmentation tasks and generalization capability to unseen prompts or their combinations.
- Authors build SEEM, a universal and interactive segmentation interface that integrates the newly designed prompting mechanism into a lightweight decoder for all segmentation tasks, leading to a model possessing properties of versatility, compositionality, interactivity, and semantic awareness

Project

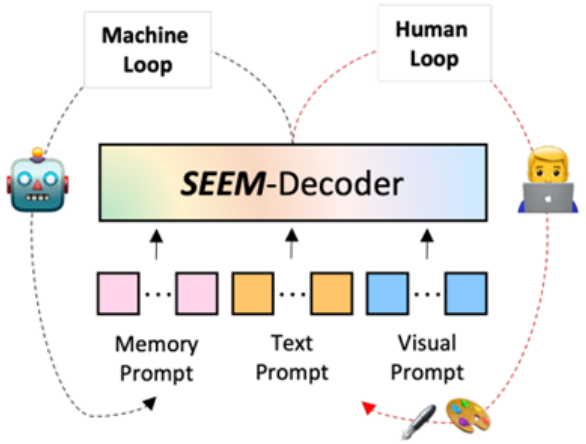
Github

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., & Lee, Y. J. (2023). Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718. NeurIPS2023 (University of Wisconsin-Madison, Microsoft Research, HKUST)

SEEM-Decoder



(a) Model Architecture



(b) Human-Model Interaction

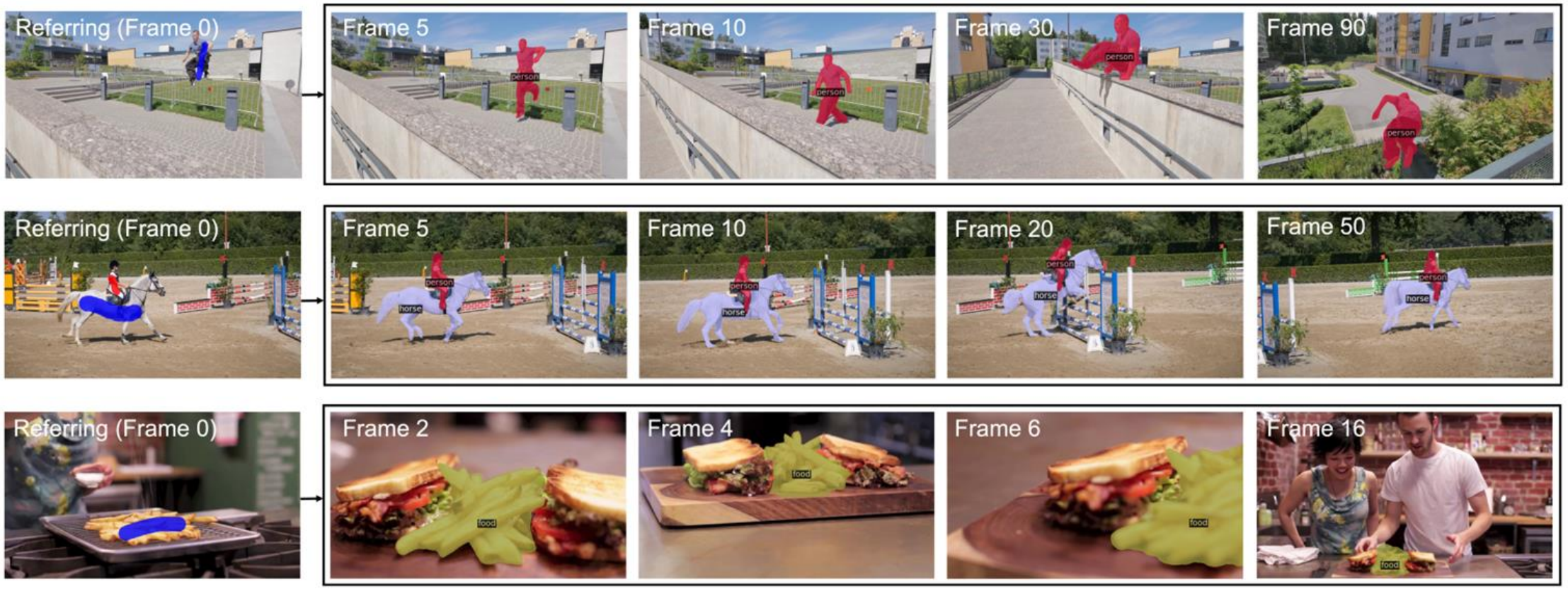
- SEEM encodes image, text, and human inputs into joint visual-semantic space as queries, features, and prompts, and then decodes queries to class and mask embeddings.
- With the benefit of SEEM decoder, the machine loop enables memorizing history mask information, and the human loop provides new corrections to the next round.
- For the vision backbone, authors use FocalT, DaViT-d3 (B), and DaViT-d5 (L). For the language encoder, they adopt a UniCL or Florence text encoder. For interactive segmentation, after one click on the image to generate the predicted mask, the next click is placed at the center of the area with the largest segmentation error

Project

Github

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., & Lee, Y. J. (2023). Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718. NeurIPS2023 (University of Wisconsin-Madison, Microsoft Research, HKUST)

SEEM: Segment Everything Everywhere All at



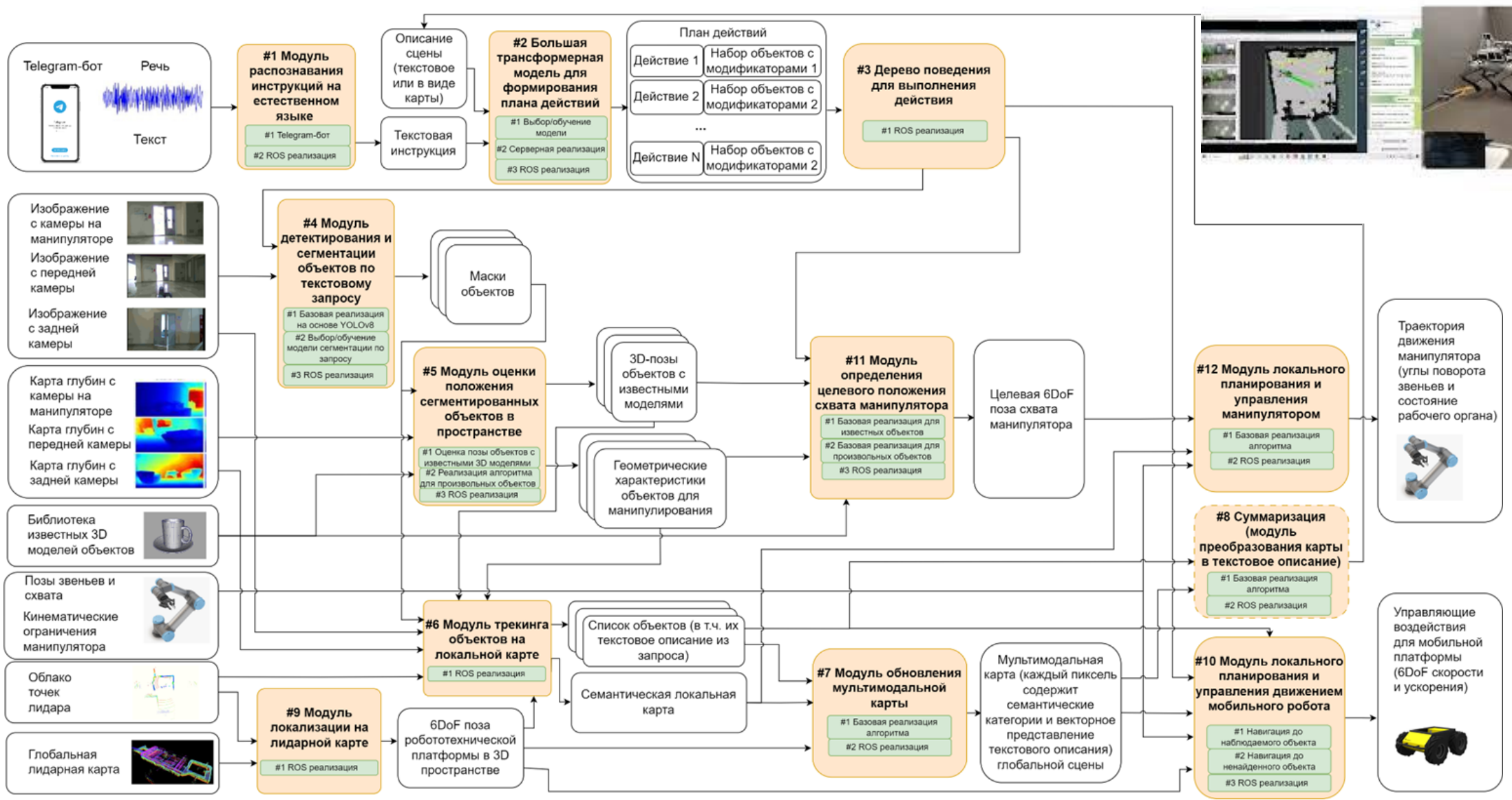
- Zero-shot video object segmentation using the first frame plus one stroke. From top to bottom, the videos are “parkour" and “horsejump-low” from DAVIS, and video 101 from YouCook2.
- SEEM precisely segments referred objects even with significant appearance changes caused by blurring or intensive deformations.

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., & Lee, Y. J. (2023). Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718. NeurIPS2023 (University of Wisconsin-Madison, Microsoft Research, HKUST)

06

Наши исследования и приложения

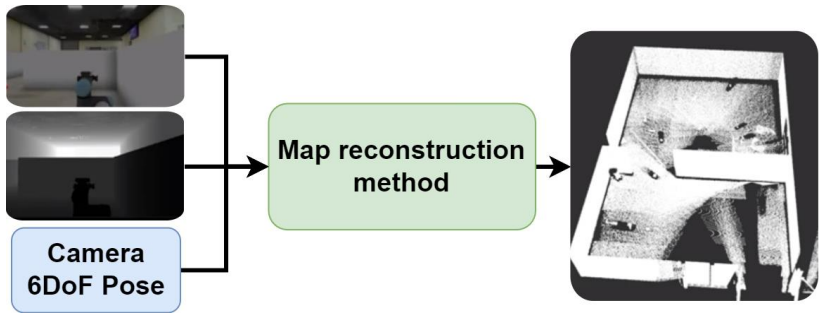
STRL-Robotics-LLM - Модульная система управления мобильным манипулятором



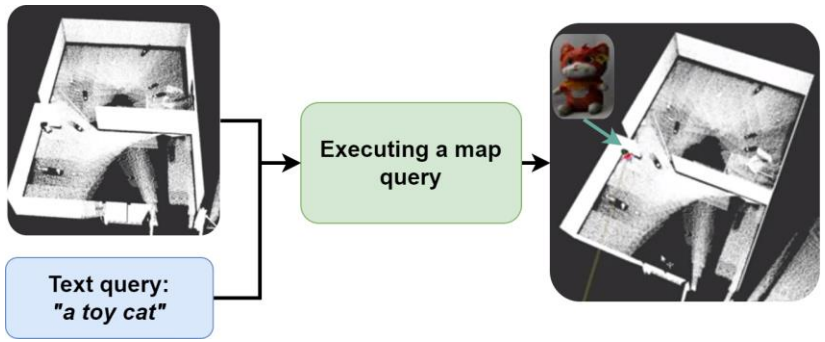
ROS1 реализация архитектуры управления

Integration with Robotic Operation System (ROS)

1st stage: map reconstruction



2nd stage: text query on the map



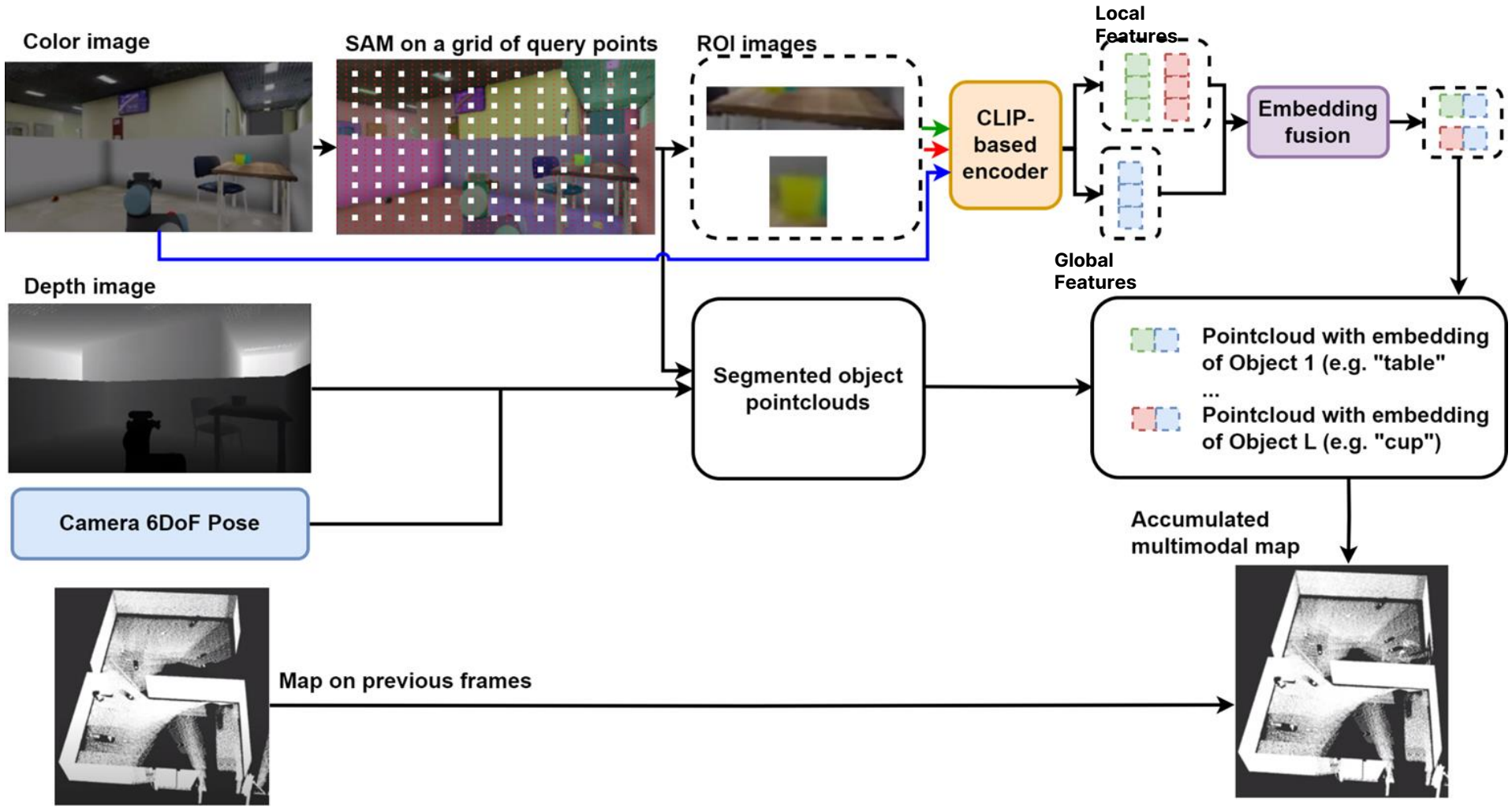
The proposed method has 2 stages:

- Create a map;
- Create a query on the map.

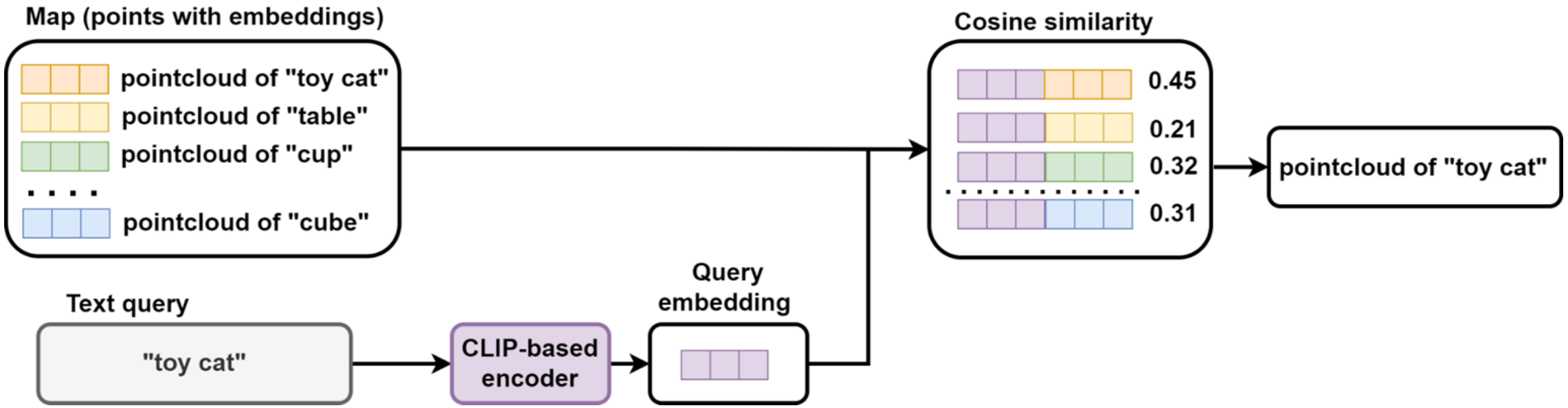
Nvidia Isaac Sim Scene

Real scene with Husky+UR5

Multimodal map reconstruction



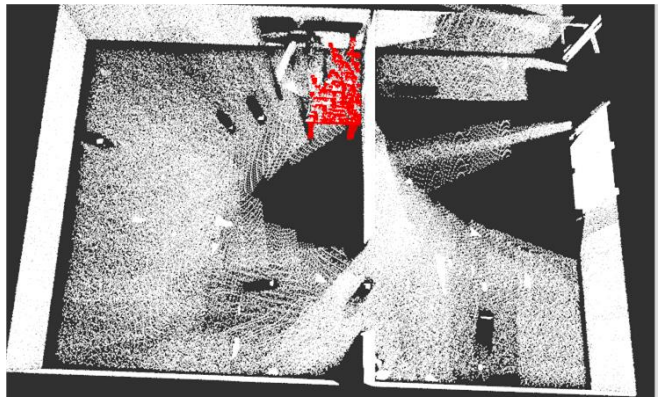
Executing a map query

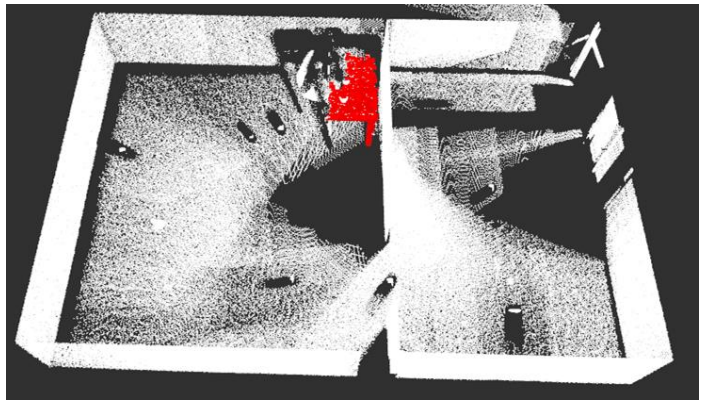


Validation of pre-trained models

Query: a table

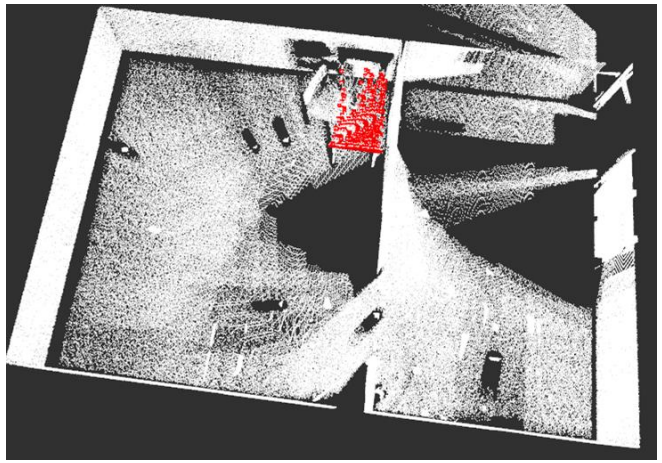
SAM
CLIP(ViT-H)

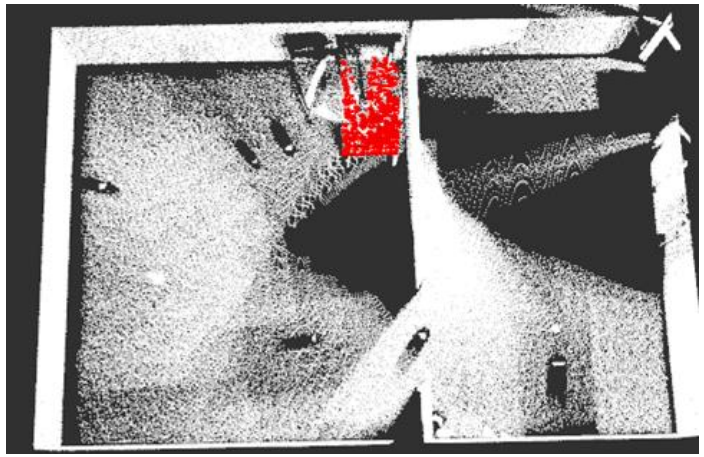




MobileSAM
CLIP(ViT-H)

SAM
CLIP(ConvNext)



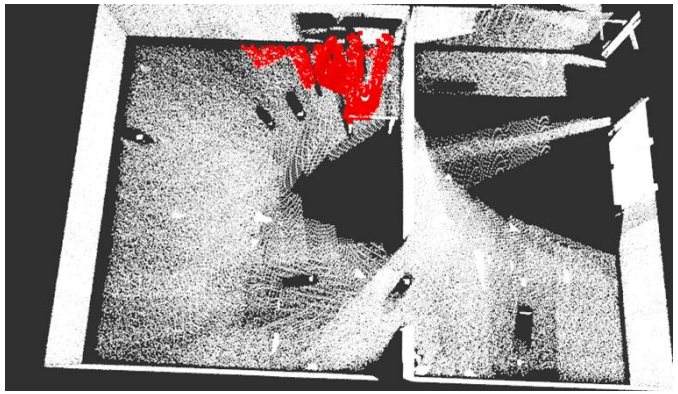


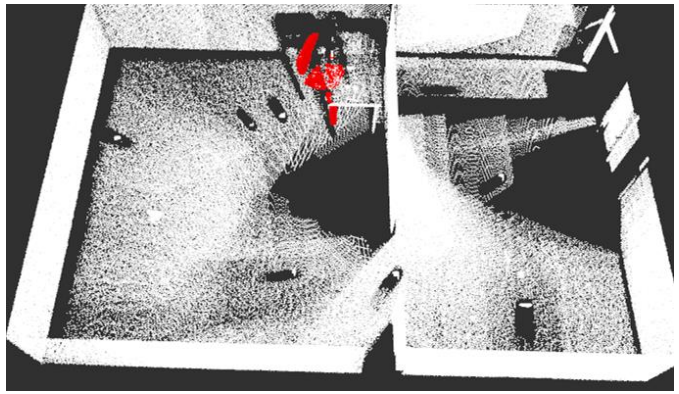
MobileSAM
CLIP(ConvNext)

Validation of pre-trained models

Query: a chair

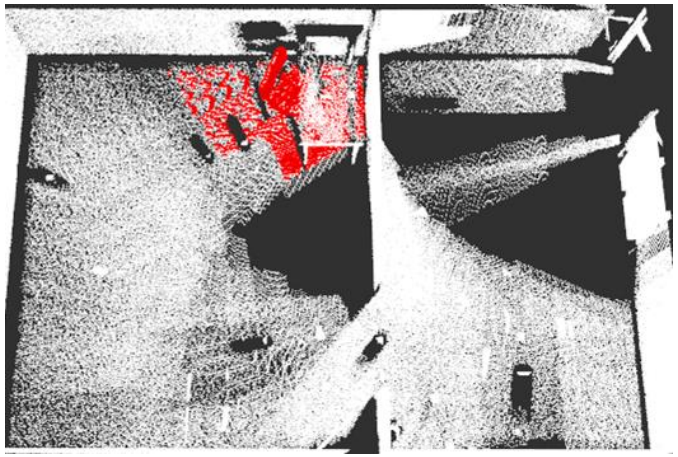
SAM
CLIP(ViT-H)

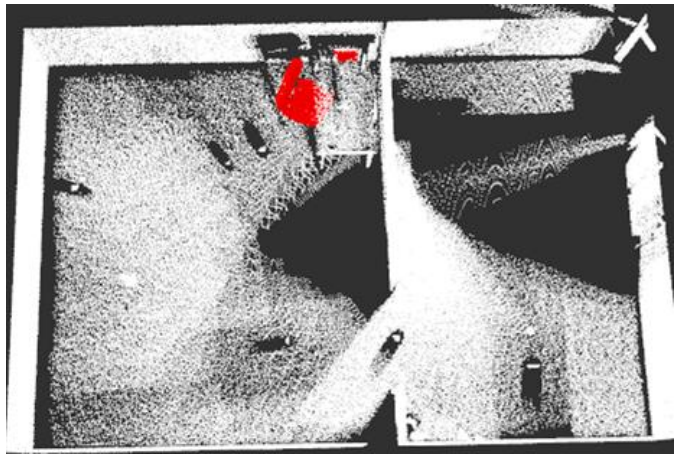




MobileSAM
CLIP(ViT-H)

SAM
CLIP(ConvNext)



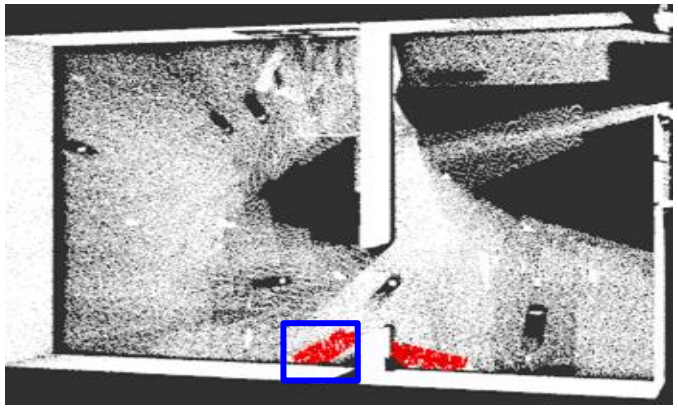


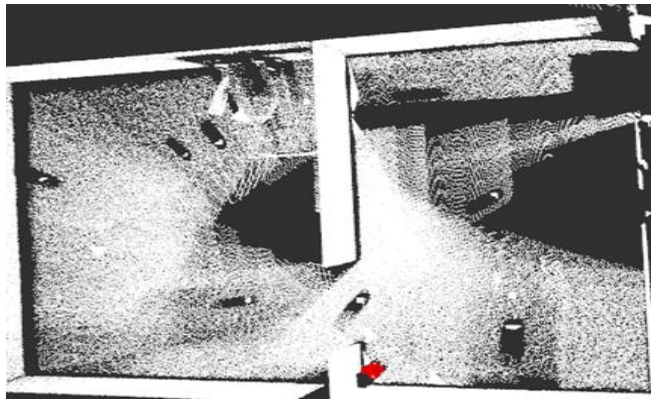
MobileSAM
CLIP(ConvNext)

Validation of pre-trained models

Query: a toy cat

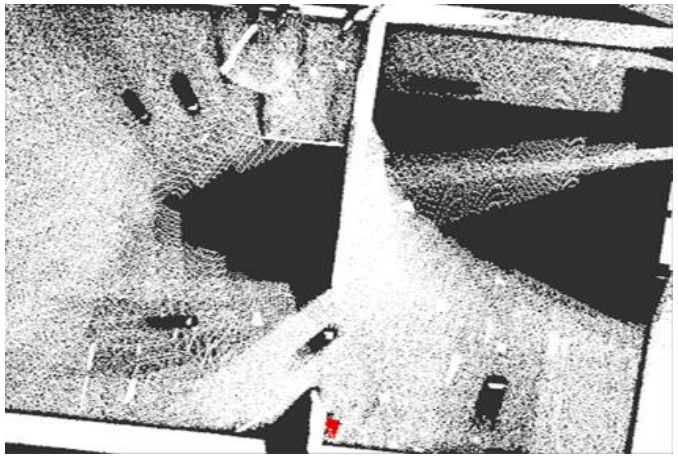
SAM
CLIP(ViT-H)

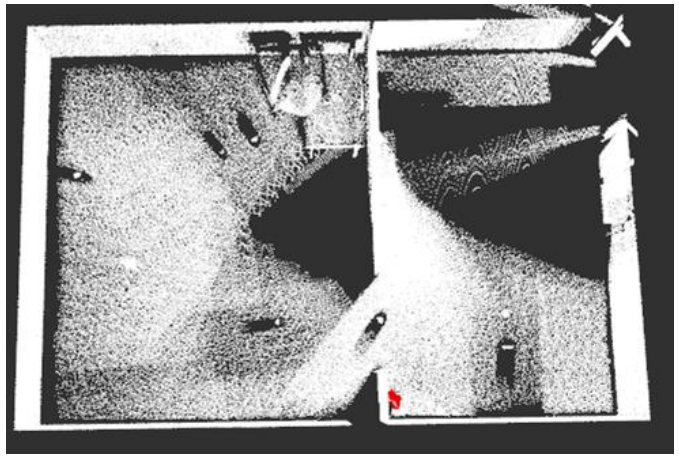




MobileSAM
CLIP(ViT-H)

SAM
CLIP(ConvNext)



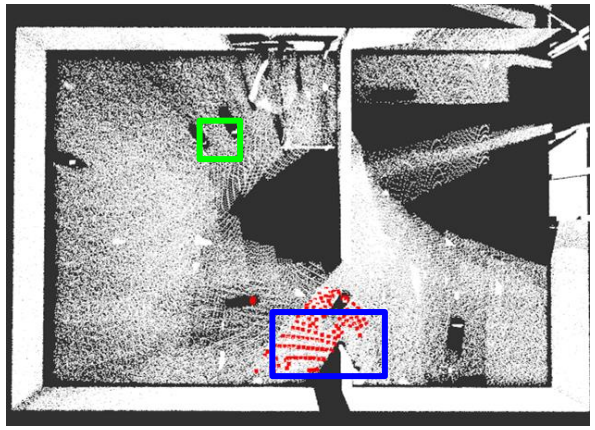


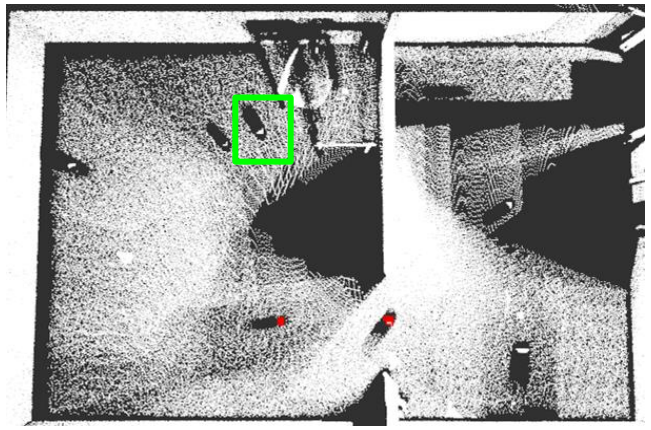
MobileSAM
CLIP(ConvNext)

Validation of pre-trained models

Query: a red cube

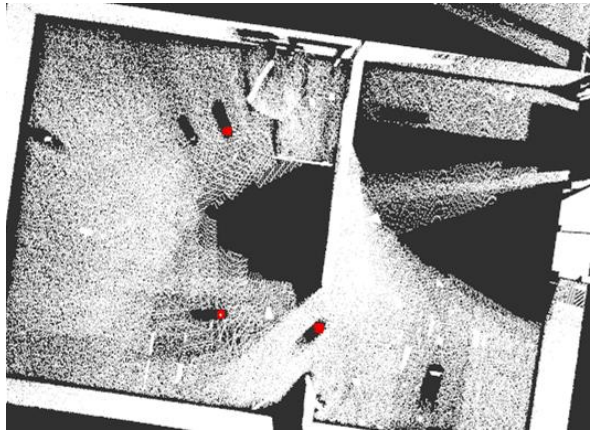
SAM
CLIP(ViT-H)

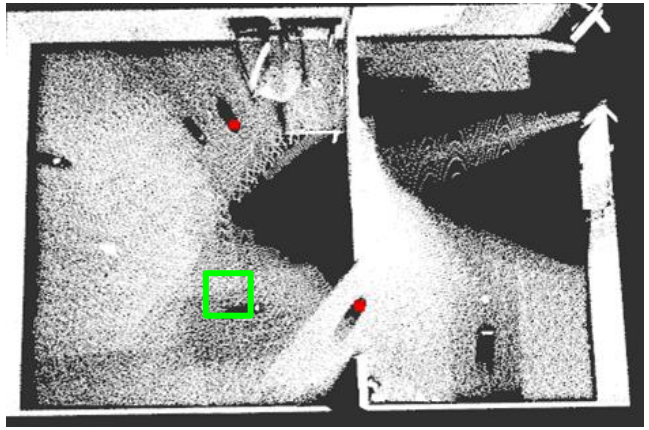




MobileSAM
CLIP(ViT-H)

SAM
CLIP(ConvNext)

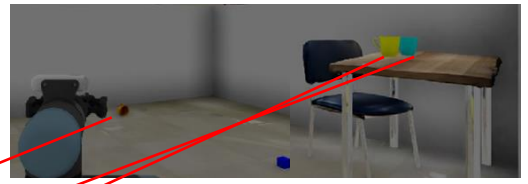




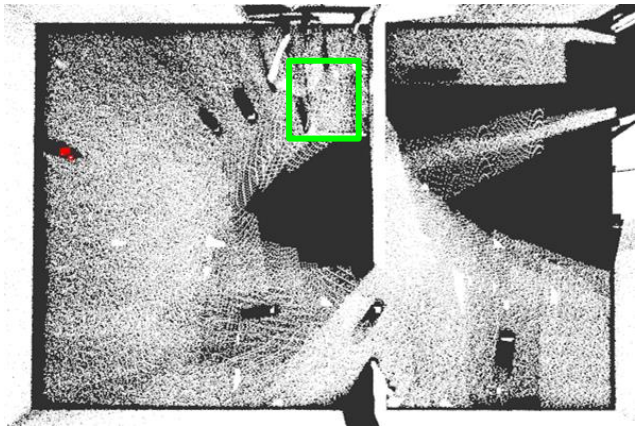
MobileSAM
CLIP(ConvNext)

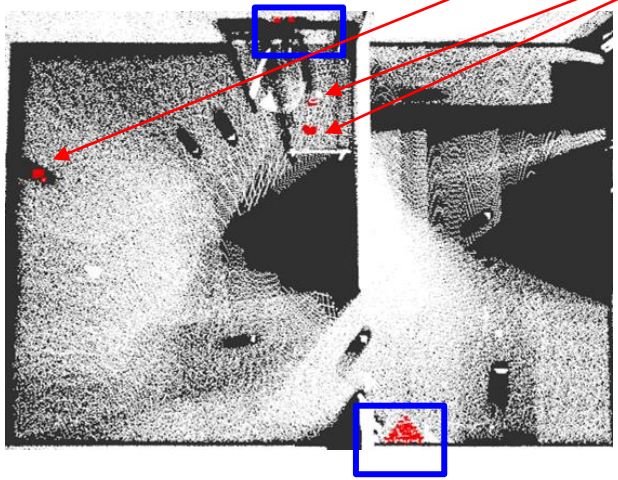
Validation of pre-trained models

Query: a cup



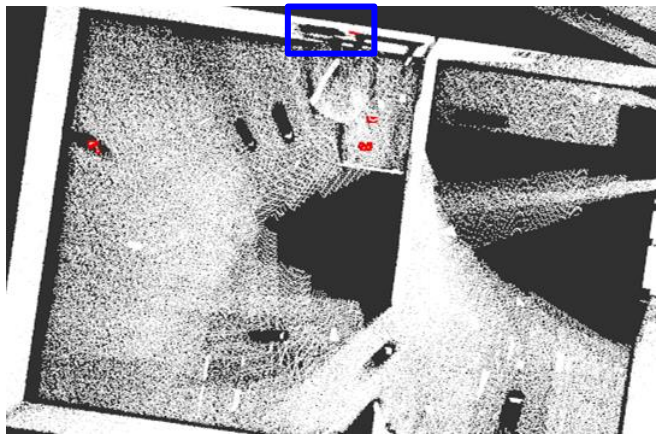
SAM
CLIP(ViT-H)

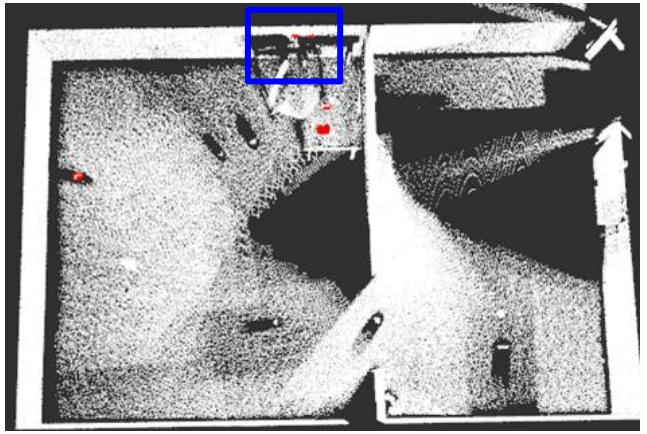




MobileSAM
CLIP(ViT-H)

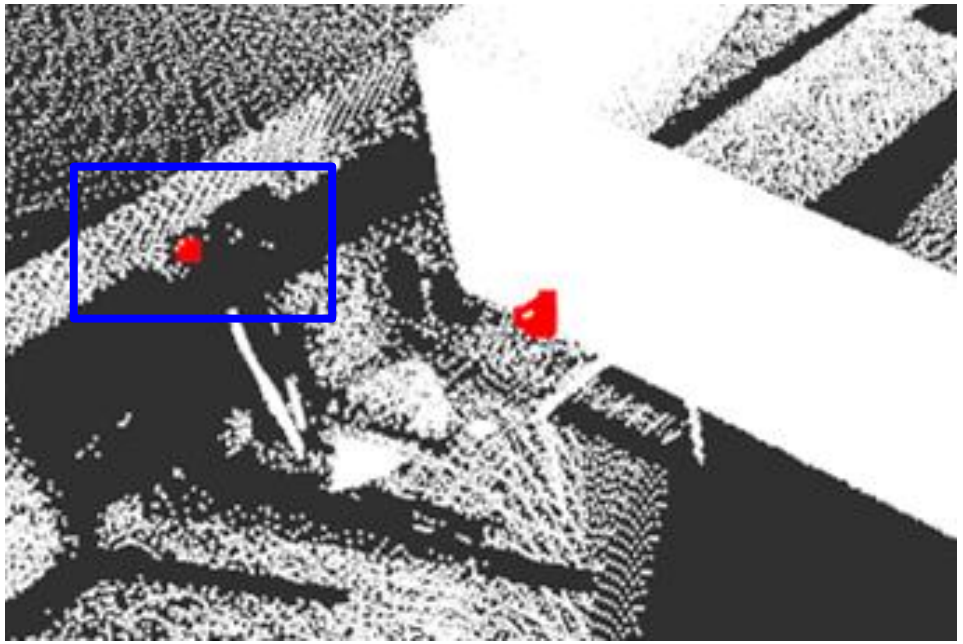
SAM
CLIP(ConvNext)



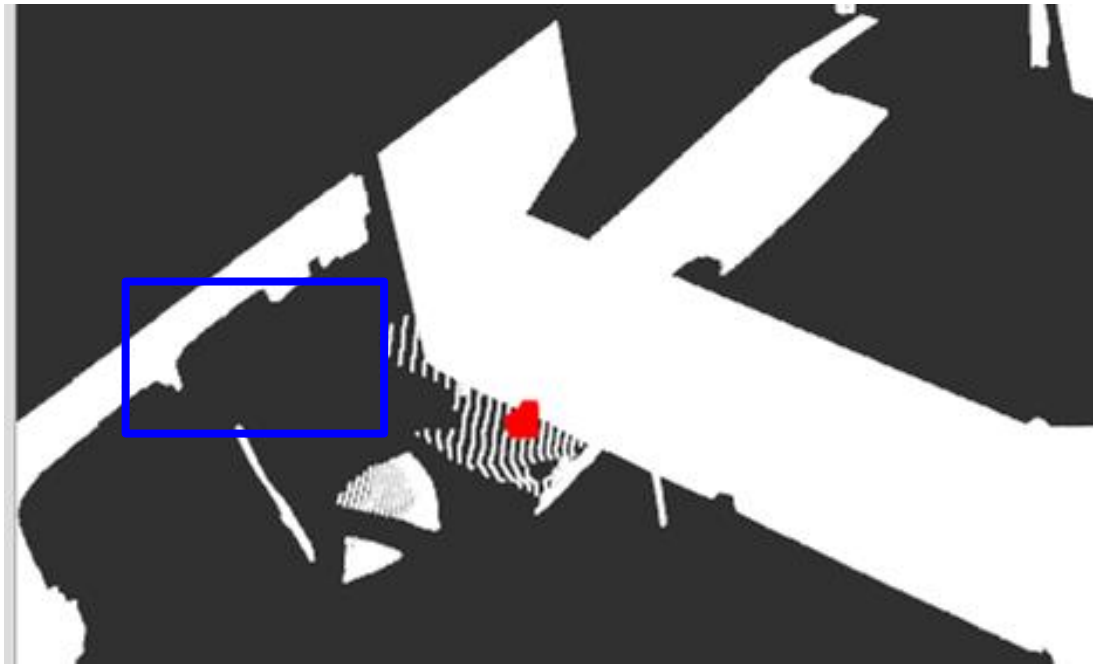


MobileSAM
CLIP(ConvNext)

Filtering the background on an object's point cloud



MobileSAM
CLIP(ConvNext) w/o filtering



MobileSAM
CLIP(ConvNext) + filtering

Performance of Multimodal Mapping Implementation for ROS

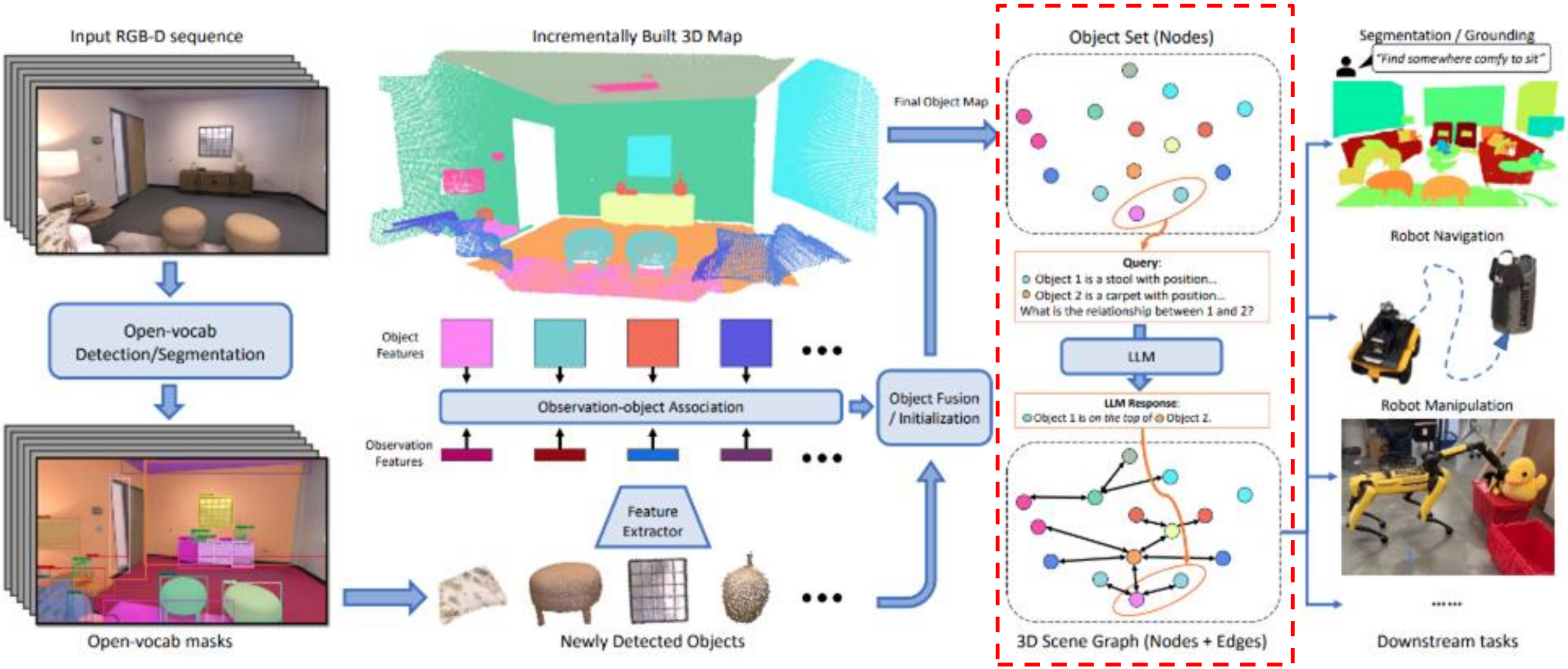
SAM Model***	CLIP Model	Time per frame in SAM, s **	Time per frame in CLIP on creating map, s **	Time in CLIP on query, s **	Memory (Reconstruct), MiB	Memory (Query), MiB
SAM ViT-H*	ViT-H-14 LAION 2B*	4.18	2.45	0.022	10’391	4’916
MobileSAM	ViT-H-14 LAION 2B	2.52 (-39.7%)	1.92(-21.6%)	0.022(-0%)	8’611(-17.1%)	4’916(-0%)
SAM ViT-H	ConvNext Base LAION 2B	4.18 (-0%)	0.73(-70.2%)	**0.013(-40.9%)**	7’226(-30.4%)	**1’694(-65.5%)**
MobileSAM	ConvNext Base LAION 2B	**2.51 (-39.9%)**	**0.59(-75.9%)**	**0.013(-40.9%)**	**5’461(-47.4%)**	**1’694(-65.5%)**

* This configuration is used in Conceptfusion
** On NVIDIA GeForce RTX 2080 Ti
*** We used SAM and MobileSAM with 32x32 seeding points

D.Yudin, S. Ladanova. Multimodal map reconstruction and application. 2023

Research of ConceptGraphs

GPT4 -> Mistral (with two stage query)



Gu, Qiao, et al. "ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning." arXiv preprint arXiv:2309.16650, 2023 (MIT, Universite de Montreal, University of Toronto, IIIT Hyderabad, JHU APL, JHU, UMass Amherst, DEVCOM Army Research Laboratory).

Quantitative assessment of LLM performance on a subsample of the sr3d dataset (253 requests)

LLM	Method for constructing graph nodes	Method for constructing graph edges	Recall @ 1	Recall @ 2	Recall @ 3	Number of valid answers (First sentence of LLM answer as top1 prediction)
Mistral-7B-Instruct-v0.1 16-bit quant	GT 3D bbox + GT object tag	-	0.19	0.21	0.21	190
Mistral-7B-Instruct-v0.1 16-bit quant	GT 3D bbox + GT object tag	list of GT objects relations	**0.34**	**0.35**	**0.36**	149
Mistral-7B-Instruct-v0.1 16-bit quant	GT 3D bbox + GT object tag	Objects caption without object ids	0.31	0.33	0.34	**233**

Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K. M., Sen, B., Agarwal, A., ... & Paull, L. (2023). Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*.

Quantitative assessment of LLM performance on a new sr3d subsample (different ways to construct graph with an LLM, 526 requests)

LLM	Method for constructing graph nodes	Method for constructing graph edges	Recall @ 1 from all answers	Recall @ 1 from valid answers	Invalid answers rate	Number of valid answers (from which id can be selected)
Mixtral 8x7b	two stages: first select semantic objects, second - use their GT bboxes	-	0.60	0.64	0.07	488
Mixtral 8x7b	two stages: first select semantic objects, second - use their GT bboxes	All relations	**0.66**	**0.69**	0.05	502
Mistral-7B-Instruct-v0.1 16-bit quant	two stages: first select semantic objects, second - use their GT bboxes	-	0.45	0.46	**0.03**	**513**

07

Направления дальнейшего развития

Что можно улучшить?

Перспективные направления:

- Построение мультимодальных карт в реальном времени
- Использование многоуровневой семантики для представления эмбеддингов точек (объектов) карты
- Создание большего количества датасетов и методов для мультимодальных запросов по мультимодальным картам.
- Учет шума в 6DoF позе камеры и картах глубин входной RGB-D последовательности;
- Эффективное распознавание ребер отношений в графе объектов карты;
- Использование одного или нескольких поясняющих примеров или дополнительной информации во время запросов к карте (one/few shot);
- Учет движения (динамики) объектов на карте;
- Временные запросы к динамическим мультимодальным картам (“Место, где люди часто ходят”, “Стул, который передвинули”, “Когда проезжали стол, какие объекты были на нем?”)

Distilled SAM: nanoSAM

Model	Jetson Orin Nano (ms)		Jetson AGX Orin (ms)		Accuracy (mIoU)			
	Image Encoder	Full Pipeline	Image Encoder	Full Pipeline	All	Small	Medium	Large
MobileSAM	TBD	146	35	39	0.728	0.658	0.759	0.804
NanoSAM (ResNet18)	TBD	27	4.2	8.1	0.706	0.624	0.738	0.7

- NanoSAM is a Segment Anything (SAM) model variant that is capable of running in real-time on NVIDIA Jetson Orin Platforms with NVIDIA TensorRT.
- NanoSAM is trained by distilling the MobileSAM image encoder on unlabeled images. For an introduction to knowledge distillation.
- Resnet18 image encoder, MobileSAM mask decoder.
- Training on COCO 2017 train.

Segment with keypoints (online using TRTPose detections)

Project

Github

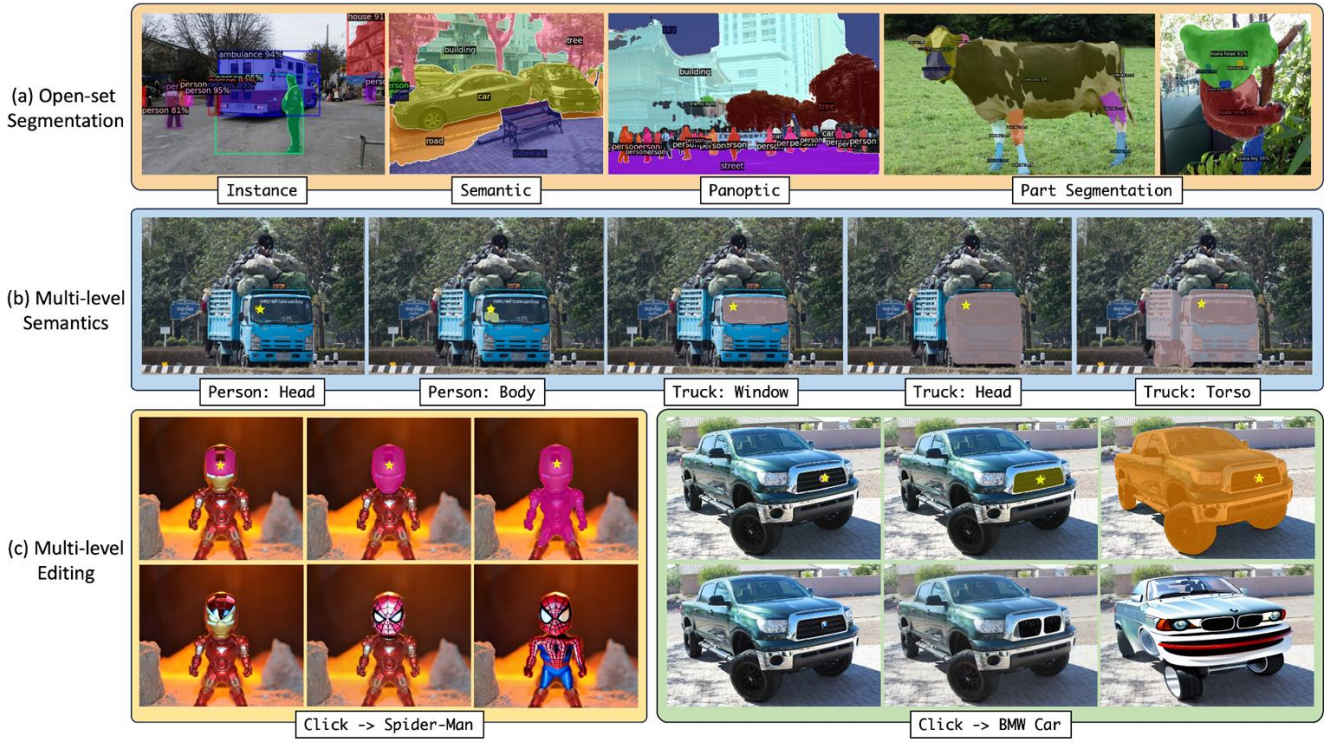
NanoSAM. A distilled Segment Anything (SAM) model capable of running real-time with NVIDIA TensorRT. 2023 Nvidia)

Что можно улучшить?

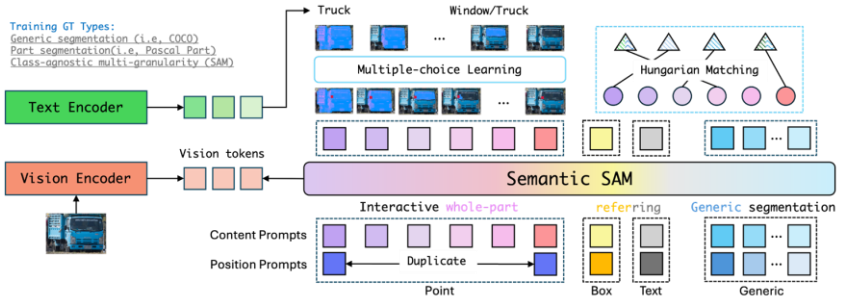
Перспективные направления:

- Построение мультимодальных карт в реальном времени
- Использование многоуровневой семантики для представления эмбеддингов точек (объектов) карты
- Создание большего количества датасетов и методов для мультимодальных запросов по мультимодальным картам.
- Учет шума в 6DoF позе камеры и картах глубин входной RGB-D последовательности;
- Эффективное распознавание ребер отношений в графе объектов карты;
- Использование одного или нескольких поясняющих примеров или дополнительной информации во время запросов к карте (one/few shot);
- Учет движения (динамики) объектов на карте;
- Временные запросы к динамическим мультимодальным картам (“Место, где люди часто ходят”, “Стул, который передвинули”, “Когда проезжали стол, какие объекты были на нем?”)

Query-based Segmentation with multi-level semantics: Semantic-SAM



- Semantic-SAM model supports Generic Segmentation, Part Segmentation, Multi-level Semantic Segmentation, Multi-Level Image Editing



Project

Github

Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., ... & Gao, J. (2023). Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767. (HKUST, Microsoft Research, IDEA, HKU, UW-Madison , Tsinghua)

Что можно улучшить?

Перспективные направления:

- Построение мультимодальных карт в реальном времени
- Использование многоуровневой семантики для представления эмбеддингов точек (объектов) карты
- Создание большего количества датасетов и методов для мультимодальных запросов по мультимодальным картам.
- Учет шума в 6DoF позе камеры и картах глубин входной RGB-D последовательности;
- Эффективное распознавание ребер отношений в графе объектов карты;
- Использование одного или нескольких поясняющих примеров или дополнительной информации во время запросов к карте (one/few shot);
- Учет движения (динамики) объектов на карте;
- Временные запросы к динамическим мультимодальным картам (“Место, где люди часто ходят”, “Стул, который передвинули”, “Когда проезжали стол, какие объекты были на нем?”)

Контакты

Дмитрий Александрович Юдин

к.т.н., заведующий лабораторией интеллектуального транспорта МФТИ - НКБ ВС, Центр когнитивного моделирования МФТИ, старший научный сотрудник AIRI

yudin@airi.net

@yuddim

cogmodel.mipt.ru