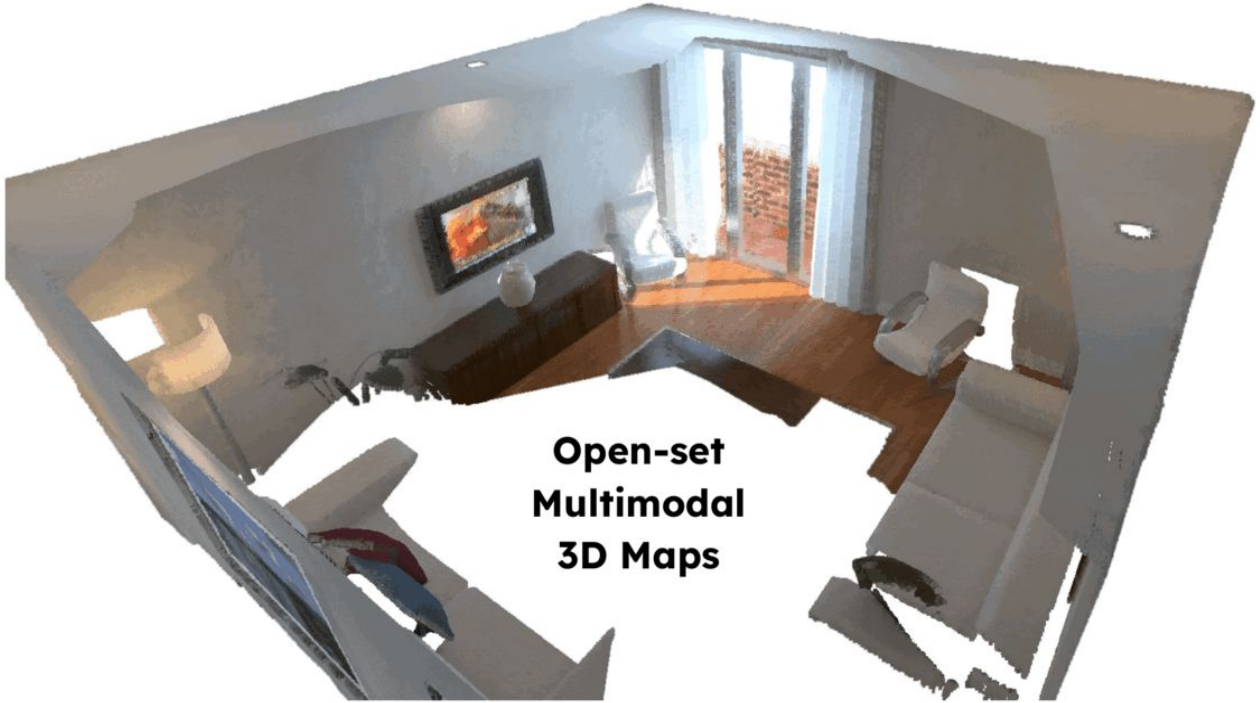
ConceptFusion: Open-set Multimodal 3D Mapping

Presenter: Dongmyeong Lee

Oct. 24, 2023

Motivation



https://concept-fusion.github.io/

Motivation

https://concept-fusion.github.io/

Motivation

https://concept-fusion.github.io/

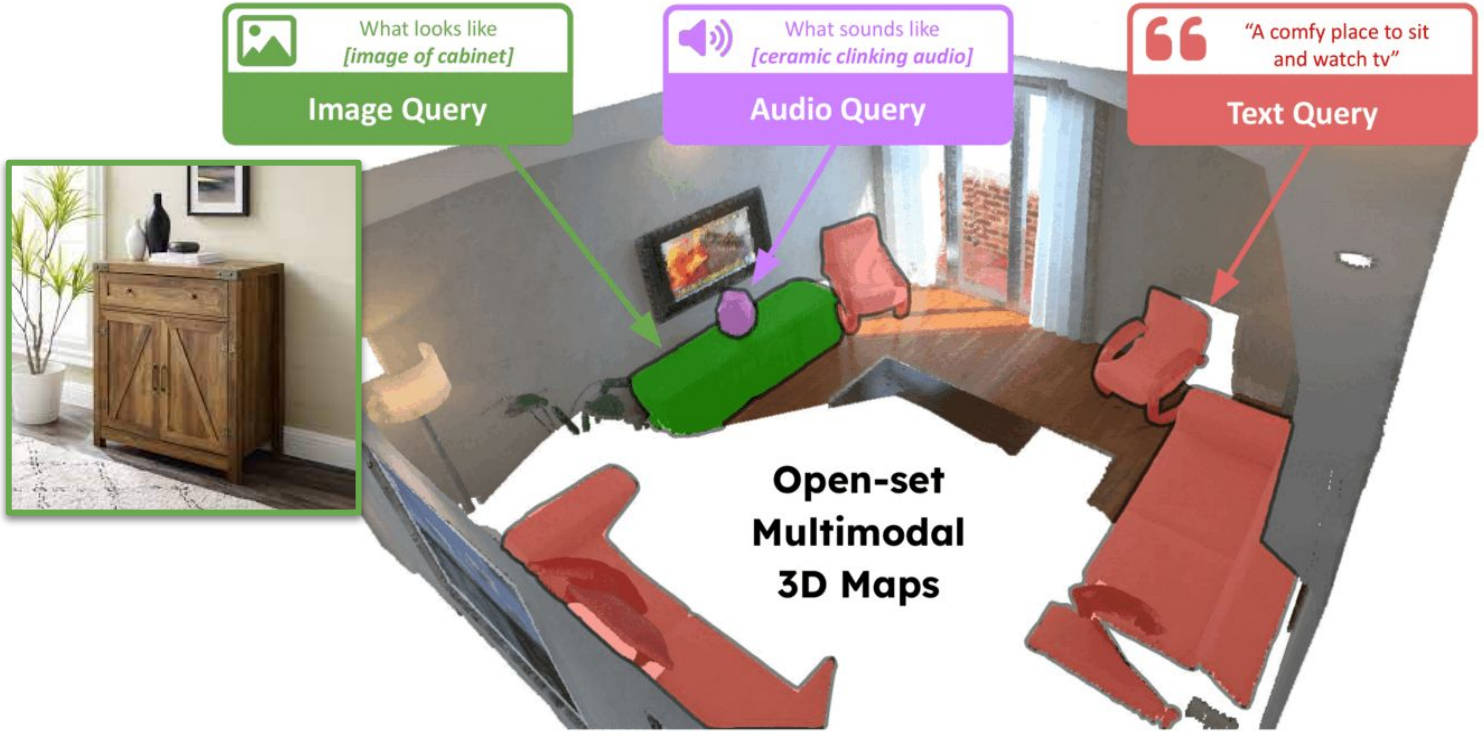
Motivation


What sounds like
[ceramic clinking audio]
Audio Query
"A comfy place to sit and watch tv"
Text Query
Open-set
Multimodal
3D Maps

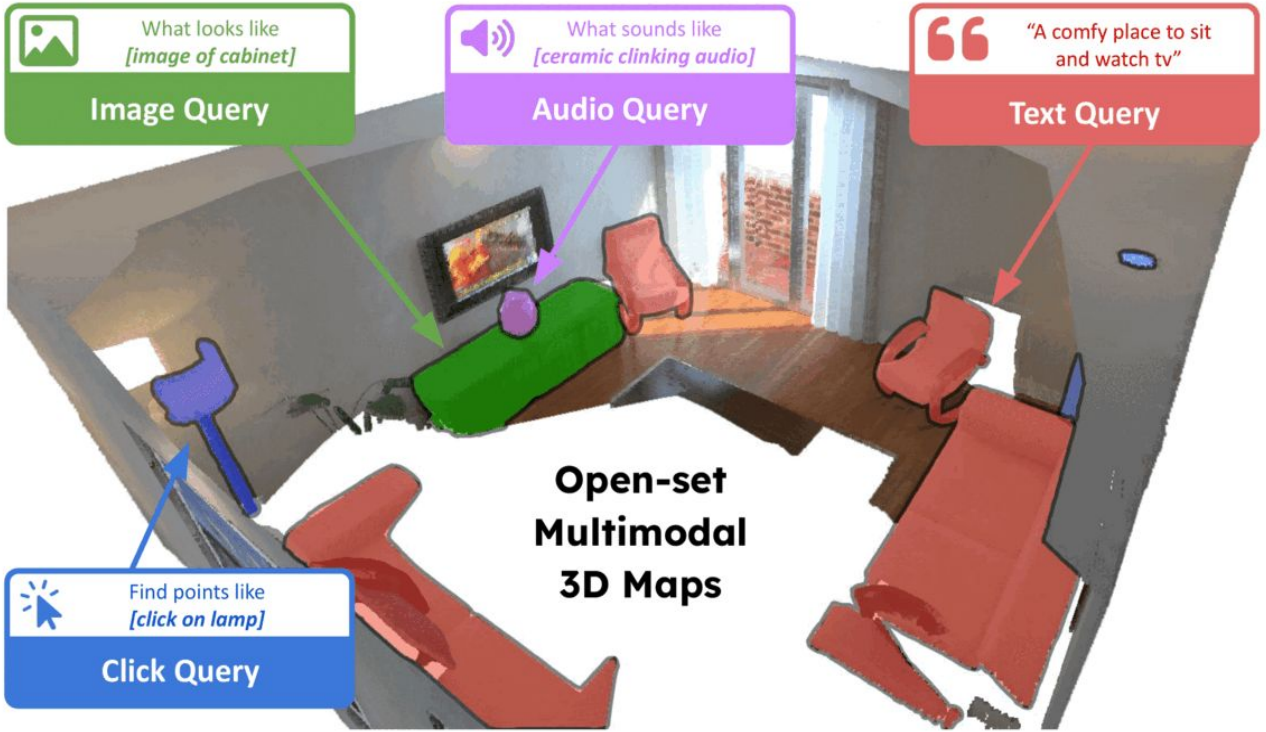
Motivation


What sounds like
[ceramic clinking audio]
Audio Query
"A comfy place to sit and watch tv"
Text Query
Open-set
Multimodal
3D Maps

Motivation


What looks like
[image of cabinet]
Image Query
What sounds like
[ceramic clinking audio]
Audio Query
"A comfy place to sit and watch tv"
Text Query
Open-set
Multimodal
3D Maps

Motivation


What looks like
[image of cabinet]
Image Query
What sounds like
[ceramic clinking audio]
Audio Query
"A comfy place to sit and watch tv"
Text Query
Open-set Multimodal 3D Maps
Find points like
[click on lamp]
Click Query

Main Problem

- 3D maps that have **2** capabilities

1. **Open-Set**

2. **Multimodal**

Main Problem

- 3D maps that have **2** capabilities

1. **Open-Set:** capture a large variety of concepts

   "can of soda" ≈ "something to drink" ≈ "Coke" ≈ "a refreshment"

2. **Multimodal**

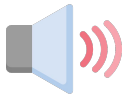
Main Problem

- 3D maps that have **2** capabilities

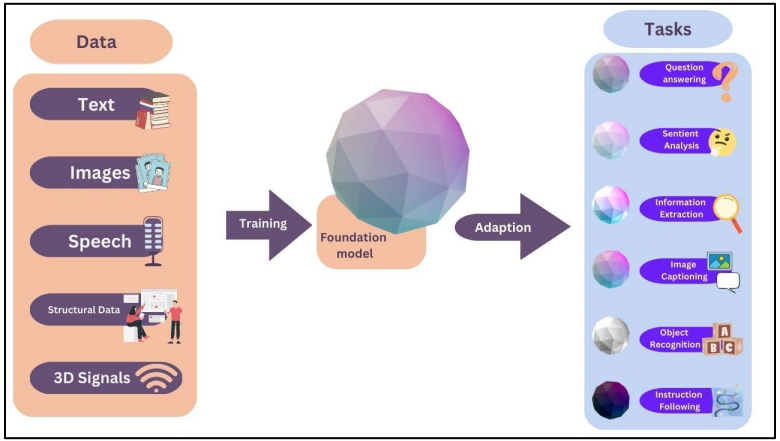
1. **Open-Set:** capture a large variety of concepts

   "can of soda" ≈ "something to drink" ≈ "Coke" ≈ "a refreshment"

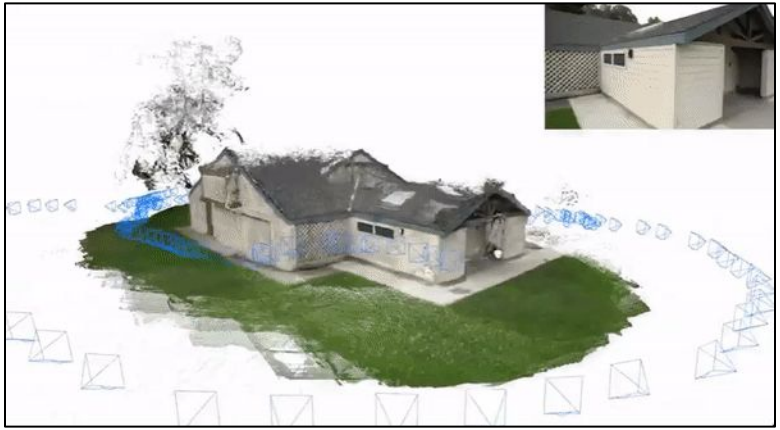
2. **Multimodal:** diverse range of possible queries

Key Ideas



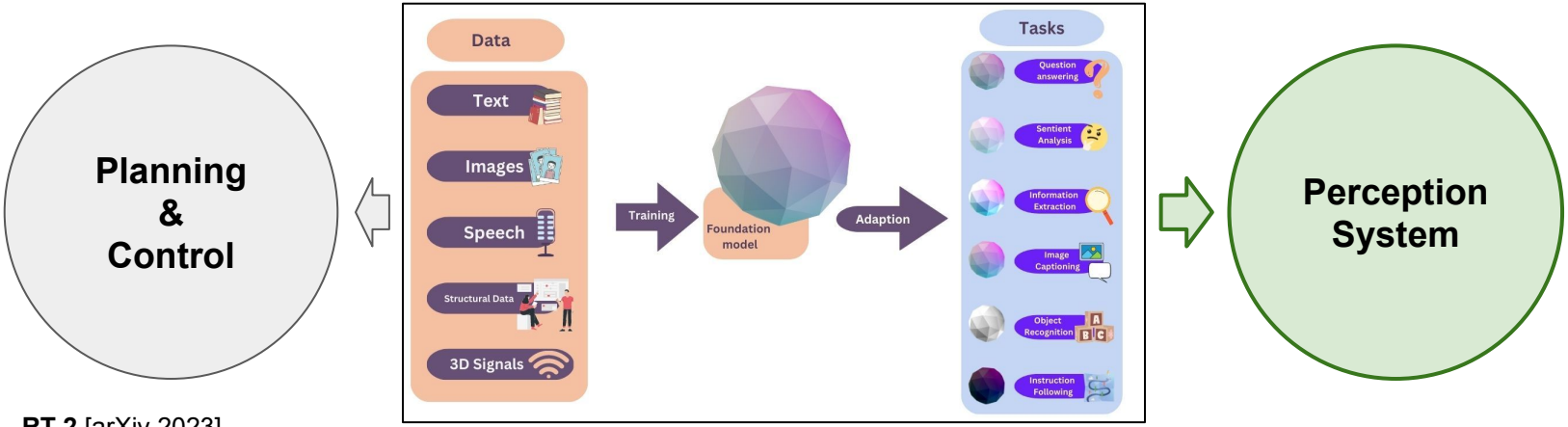
Foundation Model



SLAM, 3D reconstruction

Key Ideas

- **Foundation Models for Robotics**

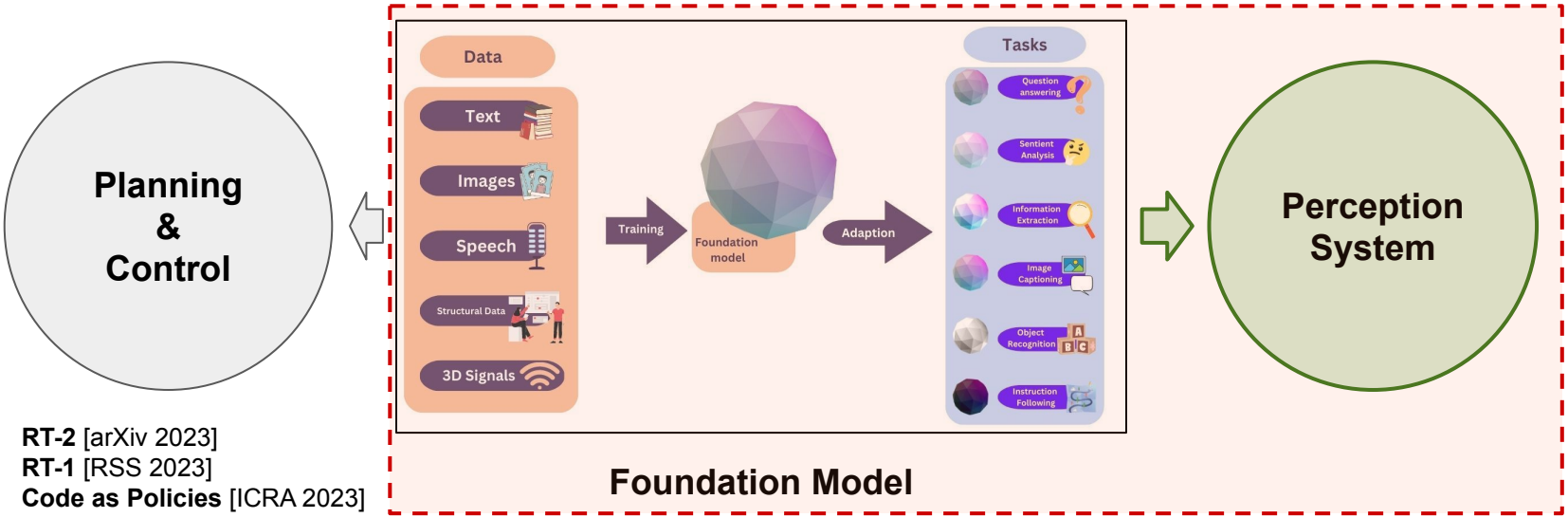


Foundation Model

RT-2 [arXiv 2023]
RT-1 [RSS 2023]
Code as Policies [ICRA 2023]
SayCan [CoRL 2022]
…

Key Ideas

- **Foundation Models for Robotics**



RT-2 [arXiv 2023]
RT-1 [RSS 2023]
Code as Policies [ICRA 2023]
SayCan [CoRL 2022]
…

"ConceptFusion"

Problem Setting

- **The open-set multimodal 3D mapping problem**

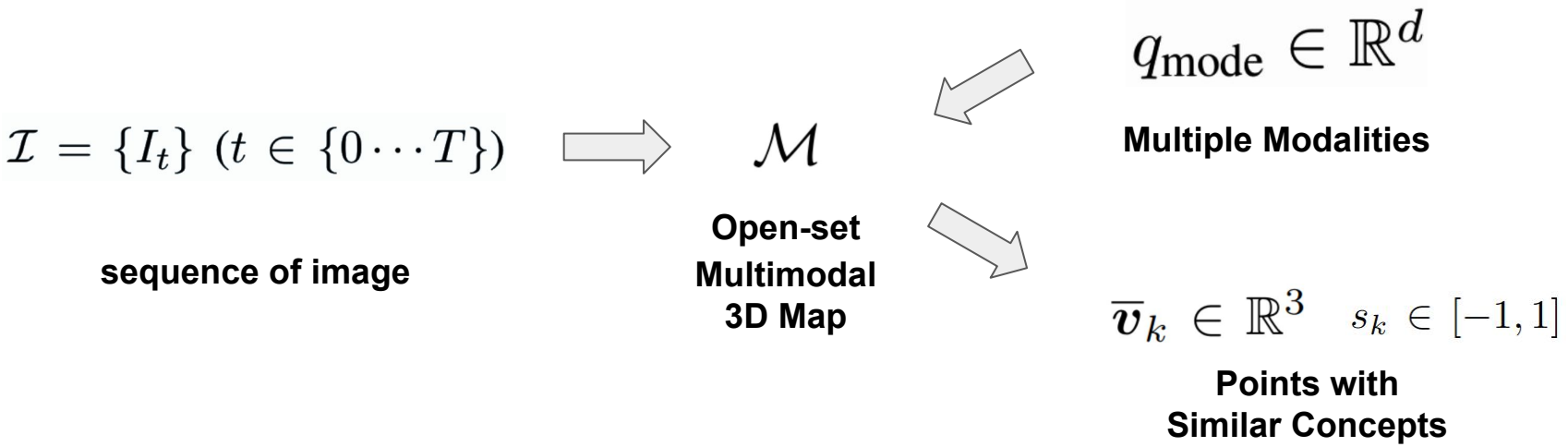
$$\mathcal{I} = \{I_t\}\ (t \in \{0 \cdots T\}) \quad \Longrightarrow \quad \mathcal{M}$$

sequence of image

Open-set Multimodal 3D Map

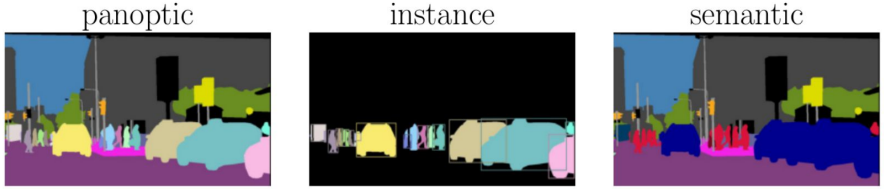
Problem Setting

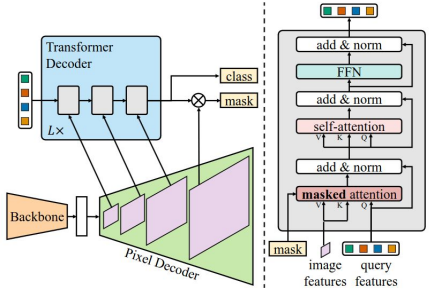
- **The open-set multimodal 3D mapping problem**



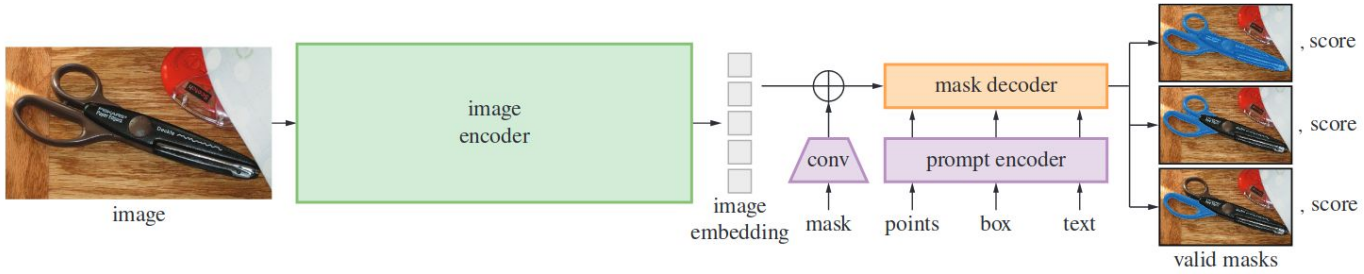
Related Work – **Instance Segmentation**

- **Mask2Former**: Masked-attention Mask Transformer for Universal Image Segmentation [Bowen Cheng et al., CVPR 2022]



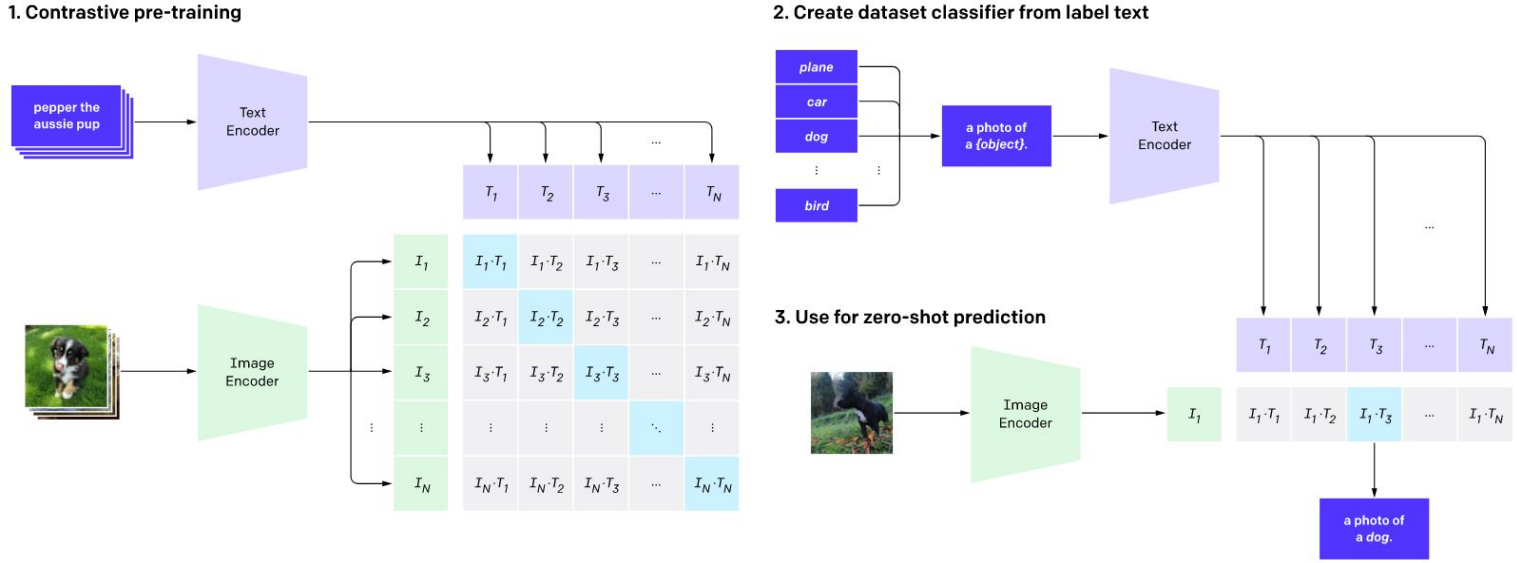


- **Segment Anything (SAM)** [Alexander Kirillov et al., ICCV 2023]



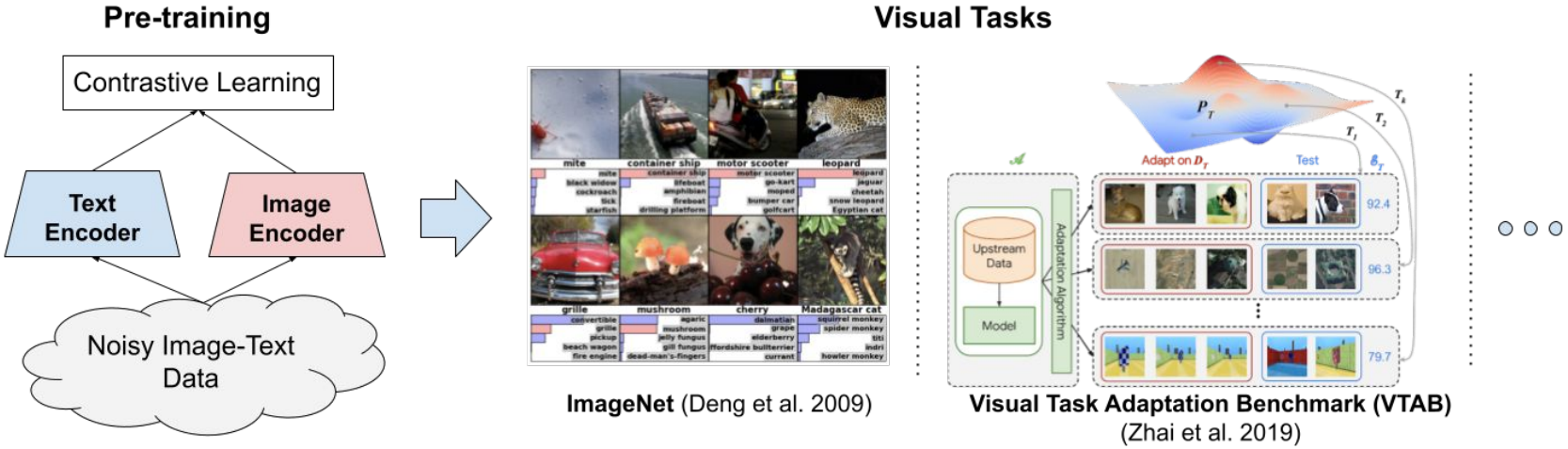
Related Work – **Foundation Model (Image-Language)**

- **CLIP**: Learning Transferable Visual Models From Natural Language Supervision [Alec Radford et al.,ICML 2021]



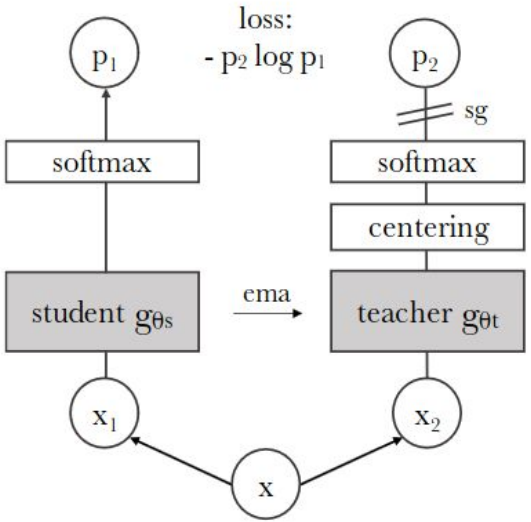
Related Work – **Foundation Model (Image-Language)**

- **ALIGN**: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision [Chao Jia et al., ICML 2021]



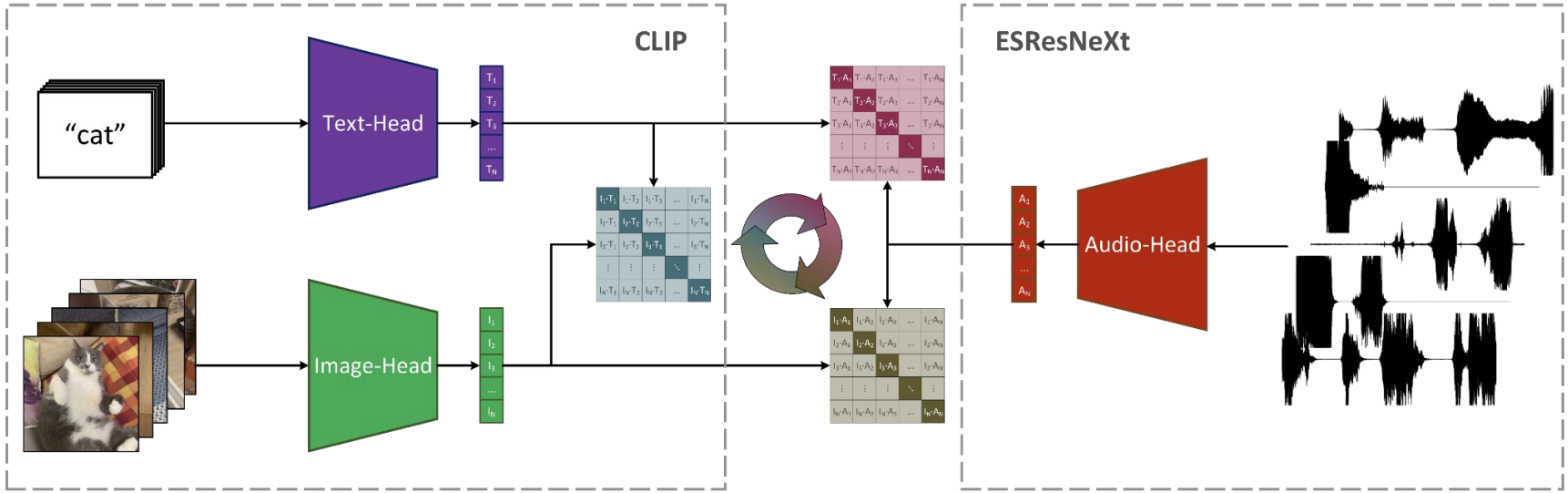
Related Work – **Foundation Model (Image-only)**

- **DINO**: Emerging Properties in Self-Supervised Vision Transformers
  [Caron et al., ICCV 2021]



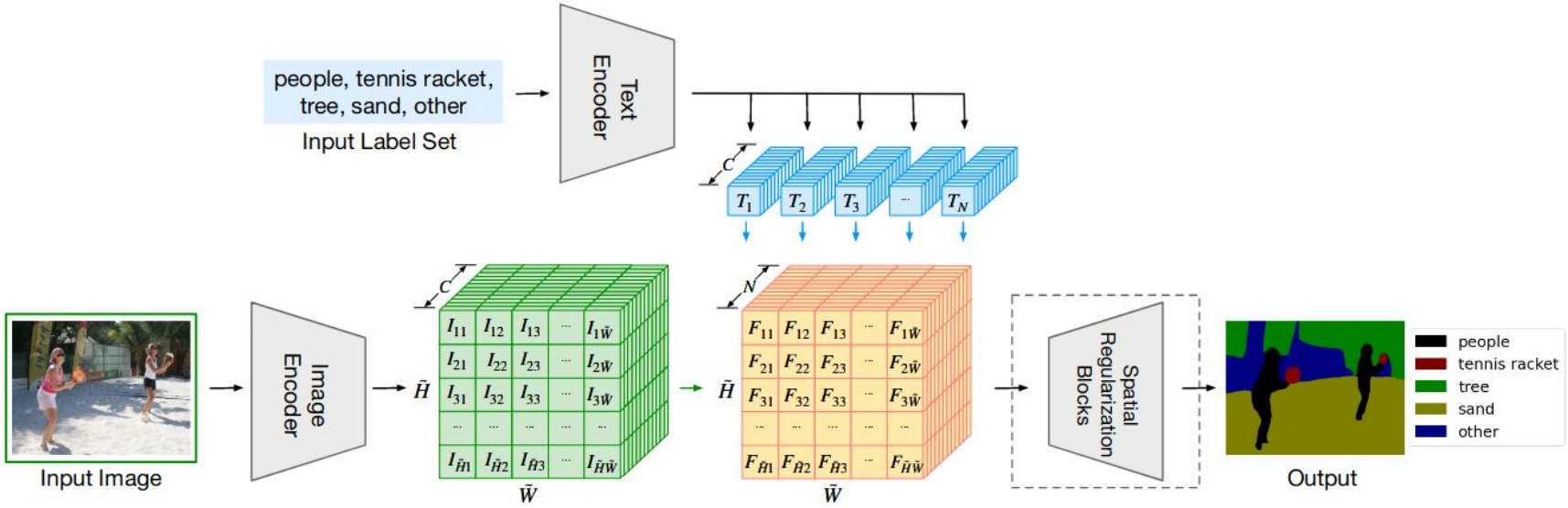
Related Work – **Foundation Model (Audio)**

- **AudioCLIP** [Andrey Guzhov et al., ICASSP 2022]



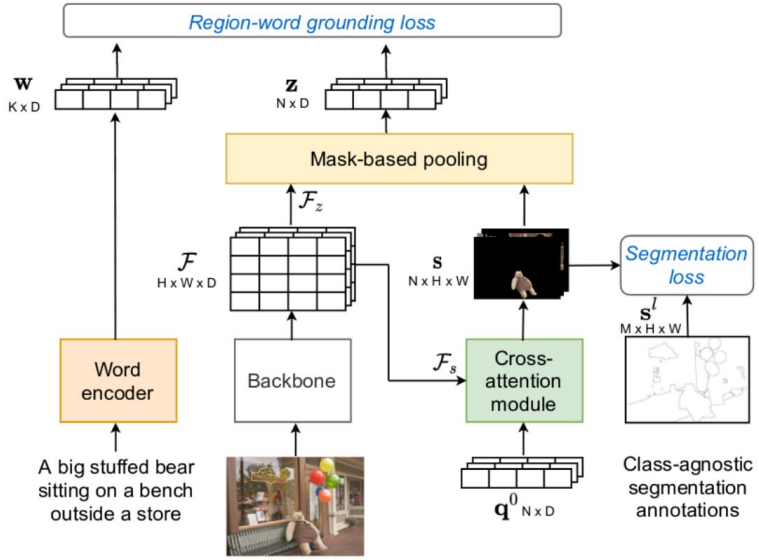
Related Work – **Pixel-aligned Foundation Features**

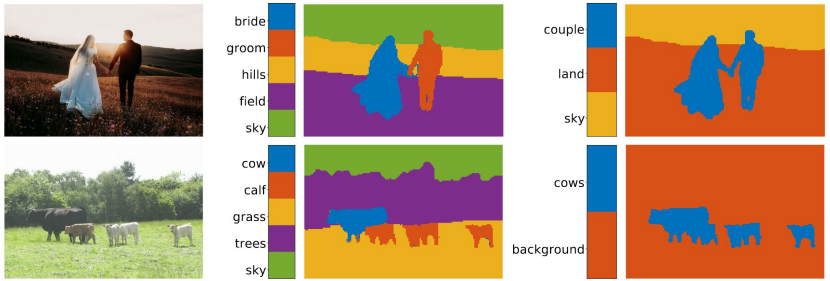
- **LSeg**: Language-driven Semantic Segmentation [Boyi Li et al., ICLR 2022]



Related Work – **Pixel-aligned Foundation Features**

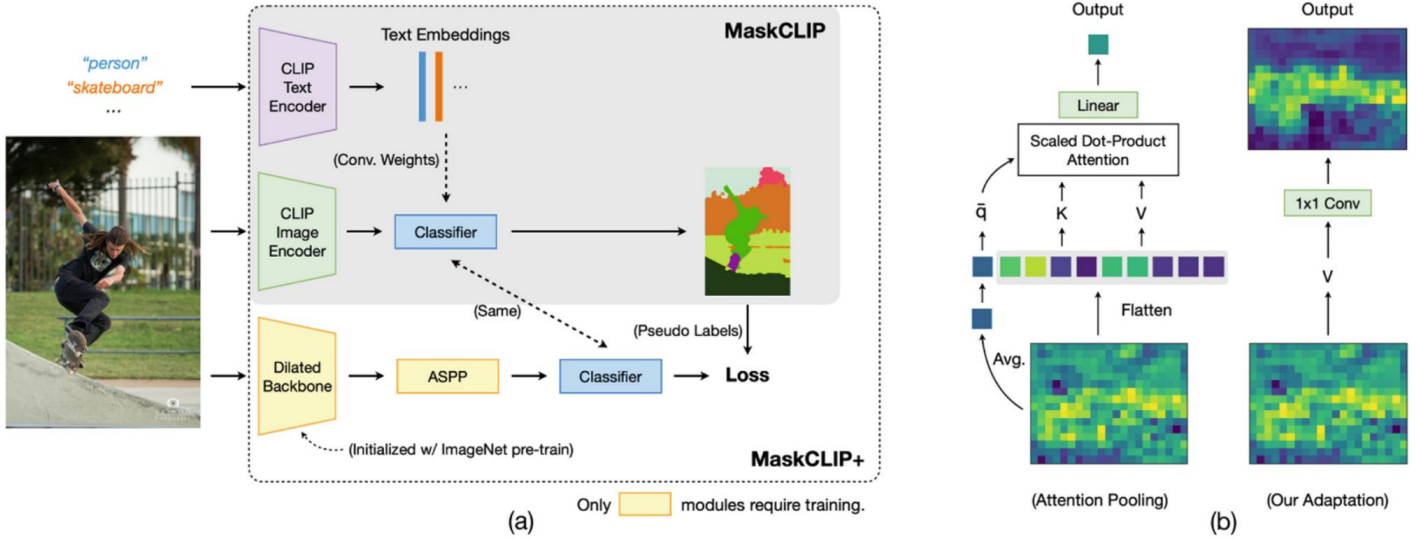
- **OpenSeg**: Scaling Open-Vocabulary Image Segmentation with Image-Level Labels [Golnaz Ghiasi et al., ECCV 2022]





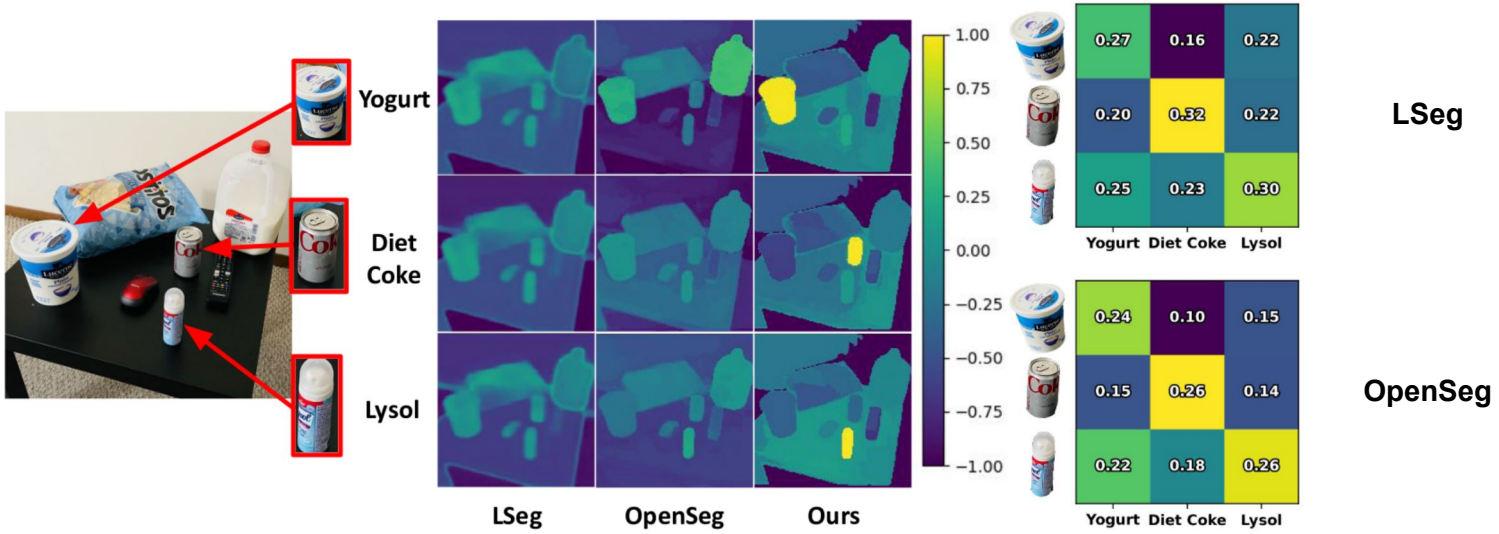
Related Work – **Pixel-aligned Foundation Features**

- **MaskCLIP:** Masked Self-Distillation Advances Contrastive Language-Image Pretraining [X Dong, J Bao et al., ECCV 2022]



Limitation of Prior Works – Pixel-aligned Features

- Struggle with delineating object **boundaries**
- Forget concepts **infrequent** in the label set (*long-tailed concept*)



ConceptFusion – Overview

Pixel-aligned
Foundation Features

CLIP

Sequence of Image

$\mathbf{f}_k^P$

Concept
Features

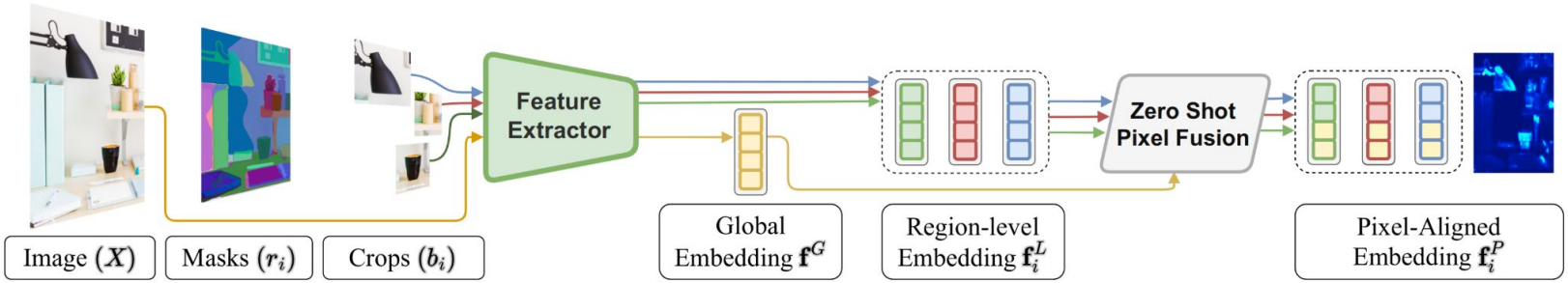
$\overline{\boldsymbol{v}}_k \in \mathbb{R}^3$
$\overline{\boldsymbol{n}}_k \in \mathbb{R}^3$
$\bar{c}_k \in \mathbb{R}$

gradslam

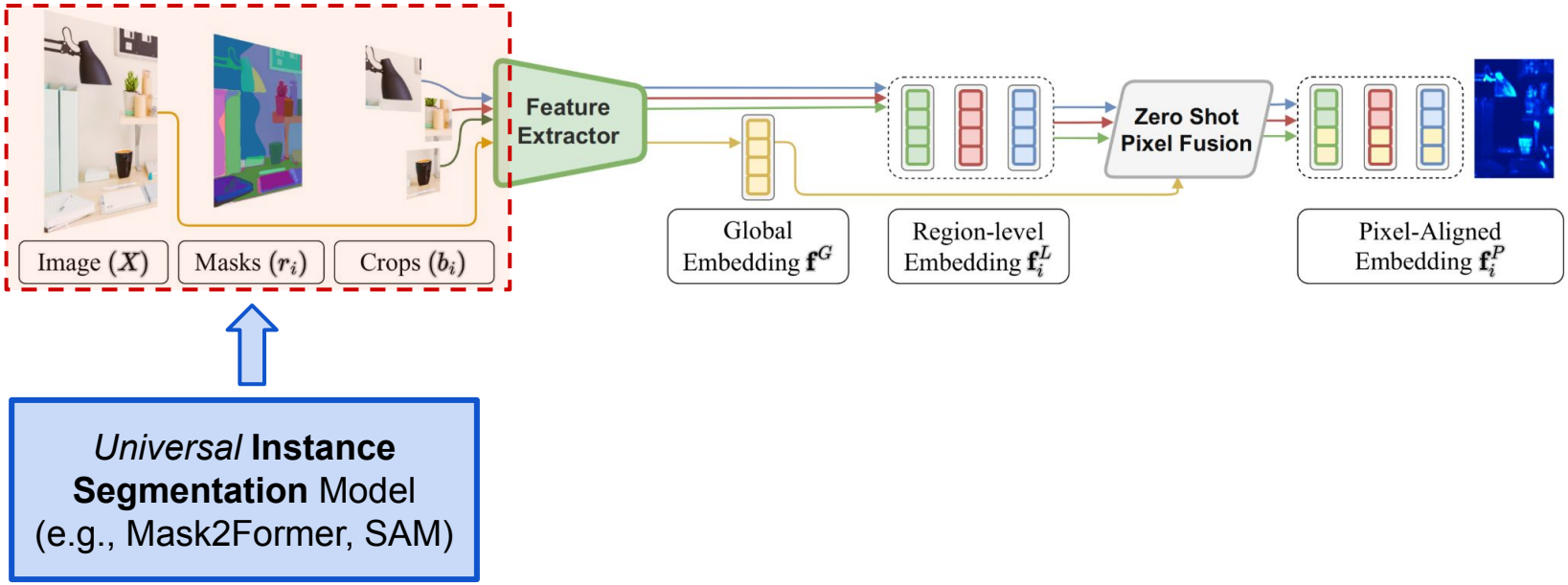
$\mathcal{M} := \{\text{point}_k\}$

Open-set Multimodal 3D Map

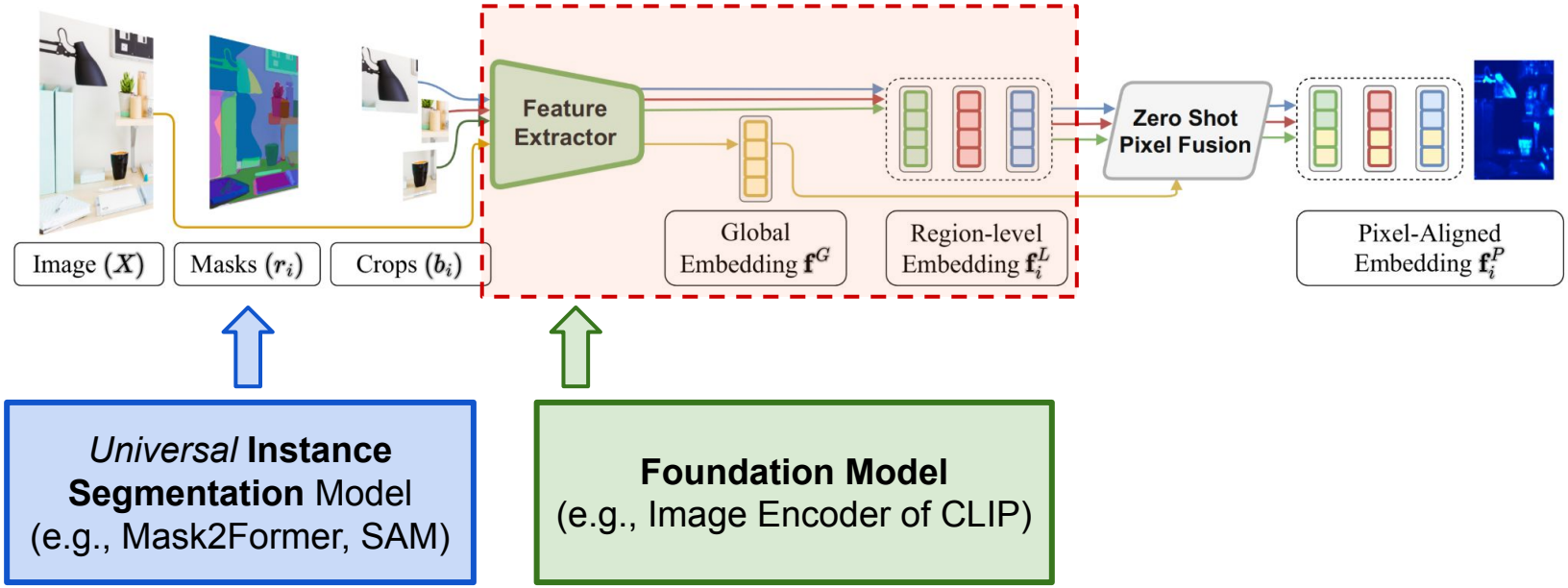
ConceptFusion – 1) Computing Pixel-aligned Features



ConceptFusion – 1) Computing Pixel-aligned Features

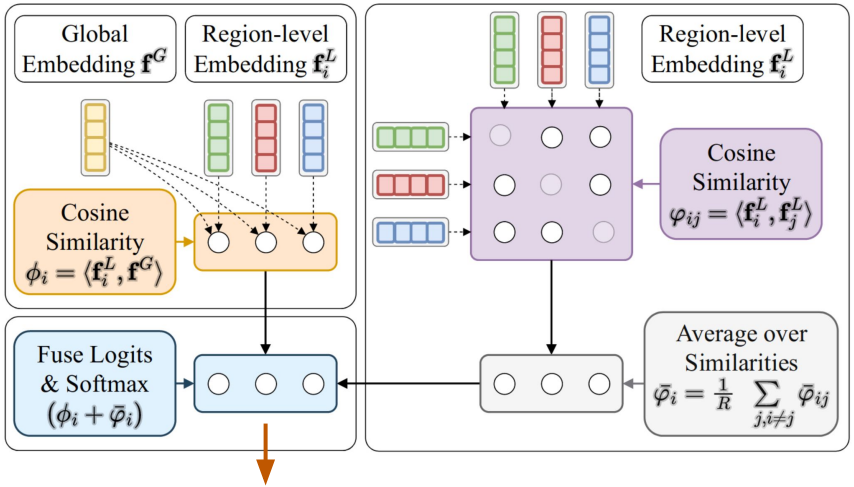


ConceptFusion – 1) Computing Pixel-aligned Features



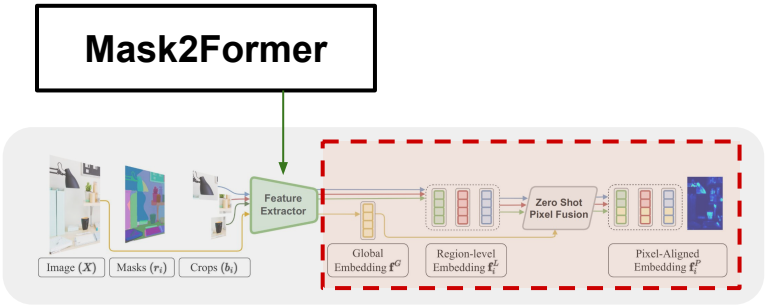
ConceptFusion – 1) Computing Pixel-aligned Features

- Fusing **Local** and **Global** Features



$$\mathbf{f}_i^P = w_i \mathbf{f}^G + (1 - w_i)\mathbf{f}^L$$

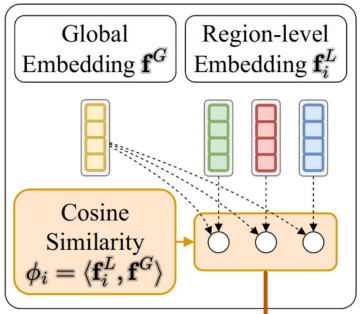
RSS 2023



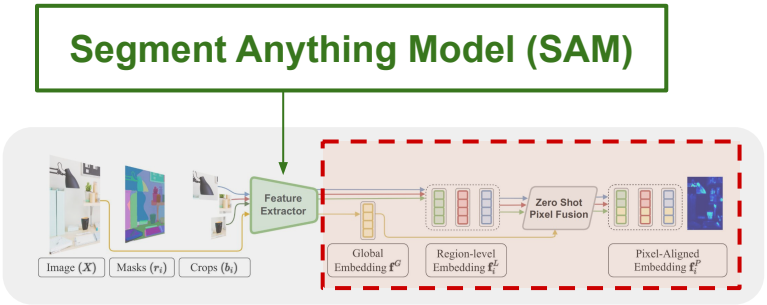
$$w_i = \frac{\exp\left(\frac{\phi_i + \bar{\varphi}_i}{\tau}\right)}{\sum_{i=1}^{R} \exp\left(\frac{\phi_i + \bar{\varphi}_i}{\tau}\right)}$$

ConceptFusion – 1) Computing Pixel-aligned Features

- Fusing **Local** and **Global** Features



arXiv, GitHub

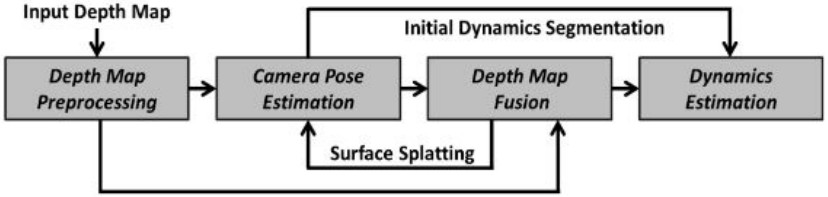


$$\mathbf{f}_i^P = w_i \mathbf{f}^G + (1 - w_i)\mathbf{f}^L$$

$$w_i = \frac{\exp\left(\frac{\phi_i}{\tau}\right)}{\sum_{i=1}^{R} \exp\left(\frac{\phi_i}{\tau}\right)}$$

ConceptFusion – 2) Fusing pixel-aligned features to 3D

- **PointFusion** (Indoor)
  [Maik Keller et al., 3DV 2013]



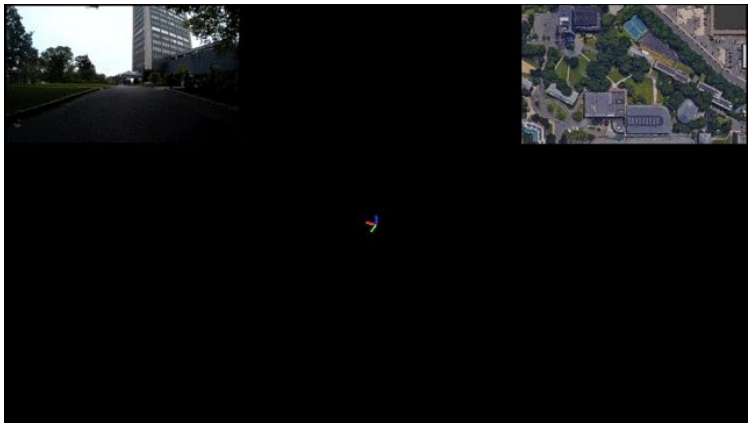
$$\mathbf{f}^P_{k,t} \leftarrow \frac{\bar{c}_k \mathbf{f}^P_{k,t-1} + \alpha \mathbf{f}^P_{u,v,t}}{\bar{c}_k + \alpha}$$

$$\bar{c}_k \leftarrow \bar{c}_k + \alpha$$

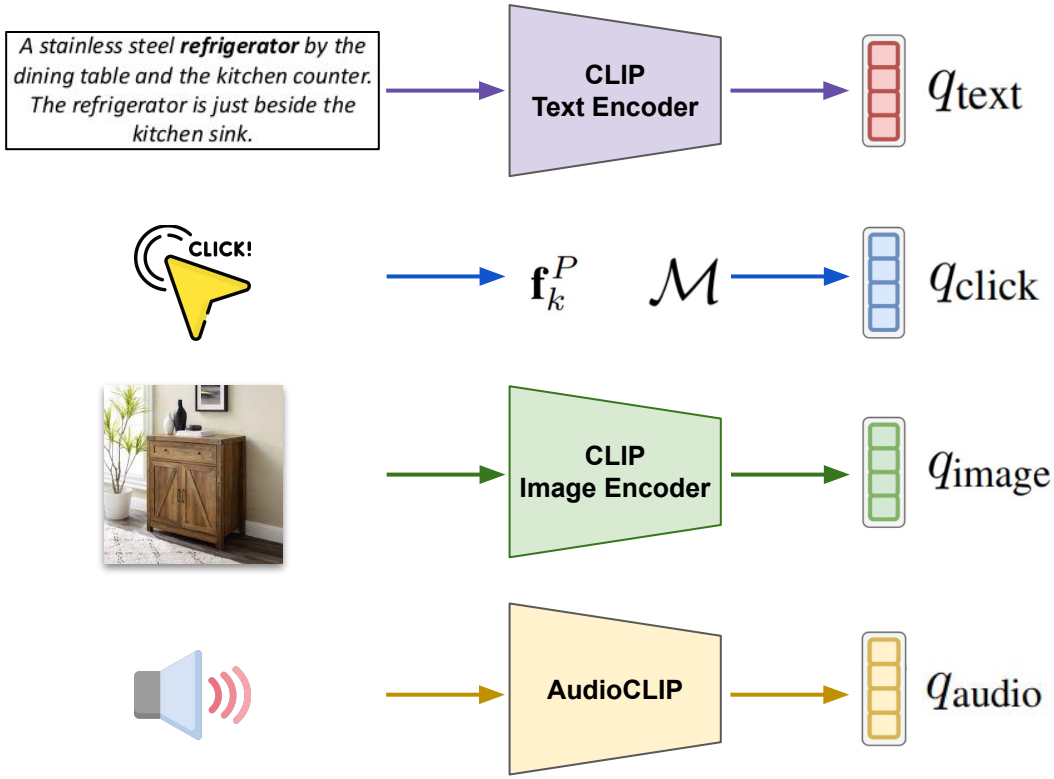
$$\alpha = e^{-\gamma^2/2\sigma^2}$$

Feature Fusion

- **LegoLOAM** (Outdoor)
  [T Shan, B Englot, IROS 2018]

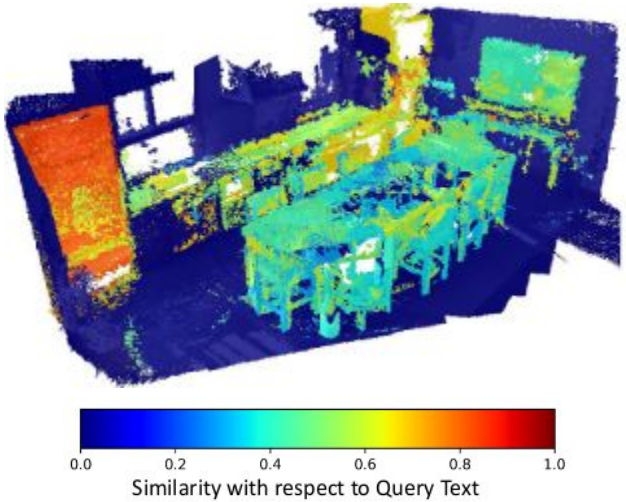


ConceptFusion – Multimodal Querying



$$s_k = \left\langle f_k, q_{\text{mode}} \right\rangle$$

Cosine Similarity



Experimental Setup

- **Benchmark**
  1) Indoor (apartment-scale, tabletop) scenes
  2) Outdoor (urban driving) scenes

- **Baselines**
  1) LSeg-3D
  2) OpenSeg-3D
  3) MaskCLIP-3D

UnCoCo Dataset
(tabletop scene)

Experimental Results – UnCoCo dataset

		3D mIoU	IoU >0.15	IoU >0.25	IoU >0.5
Supervised	LSeg-3D	0.128	25%	16.66%	9.72%
	OpenSeg-3D	0.289	43.05%	36.11%	27.78%
	MaskCLIP-3D	0.091	25.97%	9.09%	1.30%
	ConceptFusion	**0.446**	**77.78%**	**69.44%**	**45.83%**

Text Query - Structured

		3D mIoU	IoU >0.15	IoU >0.25	IoU >0.5
Supervised	LSeg-3D	0.122	31.45%	20.65%	5.65%
	OpenSeg-3D	0.153	27.26%	21.94%	11.29%
Zero-Shot	MaskCLIP-3D	0.092	20.63%	11.88%	3.06%
	ConceptFusion	**0.378**	**70.16%**	**59.52%**	**34.03%**

Text Query - Unstructured

Experimental Results – UnCoCo dataset

		3D mIoU	IoU >0.15	IoU >0.25	IoU >0.5
Supervised	LSeg-3D	0.122	31.45%	20.65%	5.65%
	OpenSeg-3D	0.153	27.26%	21.94%	11.29%
Zero-Shot	MaskCLIP-3D	0.092	20.63%	11.88%	3.06%
	ConceptFusion	**0.378**	**70.16%**	**59.52%**	**34.03%**

Image Query

		Accuracy (%)	IoU
source-ambiguous	Random	7.14%	N/A
	AudioCLIP [8]	23.81%	N/A
	ConceptFusion	**64.29%**	0.287
ecological	Random	5.56%	N/A
	AudioCLIP [8]	22.22%	N/A
	ConceptFusion	**66.67%**	0.301

Audio Query

Experimental Results – Other dataset

		ScanNet		Replica		Semantic KITTI	
		mAcc	f-mIoU	mAcc	f-mIoU	mAcc	f-mIoU
Priv.	LSeg [24]	0.70	0.63	0.52	0.33	0.84	082
	OpenSeg [18]	0.63	0.62	0.54	0.41	0.78	0.77
	CLIPSeg (rd64-uni) [55]	0.41	0.34	0.32	0.23	0.77	0.75
	CLIPSeg (rd16-uni) [55]	0.41	0.36	0.40	0.28	0.79	0.77
	CLIPSeg (rd64-uni-refined) [55]	0.23	0.24	0.13	0.13	0.28	0.26
ZS	MaskCLIP [27]	0.24	0.28	0.01	0.05	0.70	0.66
	Mask2former + Global CLIP feat	0.35	0.48	0.13	0.10	0.22	0.20
	ConceptFusion	**0.63**	**0.58**	**0.31**	**0.24**	**0.79**	**0.78**

Text Label Query

Experimental Results – on real robotic system

Autonomous Navigation

Experimental Results – on real robotic system

Zero-shot tabletop rearrangement

Discussion of Results

UnCoCo dataset

- Outperformed both **finetuned model** and **zero-shot model** for multimodal queries
  (LSeg, OpenSeg) (MaskCLIP)

Text label Query on other dataset (ScanNet, Replica, Semantic KITTI)

- Underperformed compared to finetuned model
- Outperformed zero-shot model

Real Robotic system

- Works well in localizing the novel objects

ConceptFusion – 3D Spatial Comparator (3DSC)

$$\text{RELATION}(\text{QUERY}_a, \text{QUERY}_b)$$

- Scalar (Distance)

  $\text{HOWFAR}(q_a, q_b)$

- Boolean (Spatial Relationship)

  $\text{ISTOTHELEFT}(q_a, q_b)$

  $\text{ONTOPOF}(q_a, q_b)$

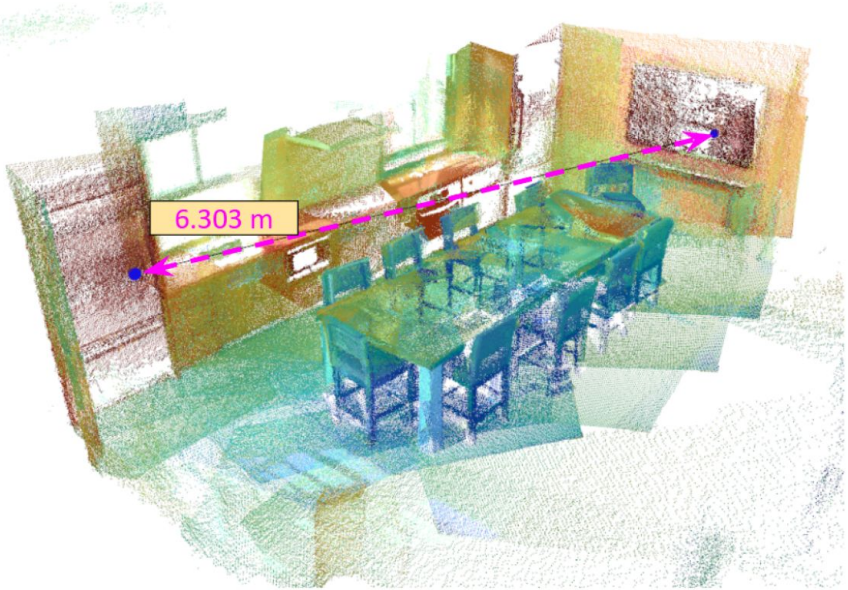
```
Here is a set of available functions:
1. howFar(object1, object2): returns the distance
    between object1 and object2
2. isToTheRight(object1, object2): returns true if
    object1 is to the right of object2
3. isToTheLeft(object1, object2): returns true if
    object1 is to the left of object2
4. isContained(object1, object2): returns true if
    object1 is contained in object2
5. onTopOf(object1, object2): returns true if
    object1 is on top of object 2
6. under(object1, object2): returns true if object1
    is underneath object 2
7. isBigger(object1, object2): returns true if
    object1 is bigger than object2
8. canFitInside(object1, object2): returns true if
    object1 can fit inside object2
Parse the provided queries into one of the above
    function formats.

Query: How close is the chair from the sofa?
Response: howFar(chair, sofa)
```

Prompt to LLM

Outlook

- 3D Spatial Reasoning Abilities



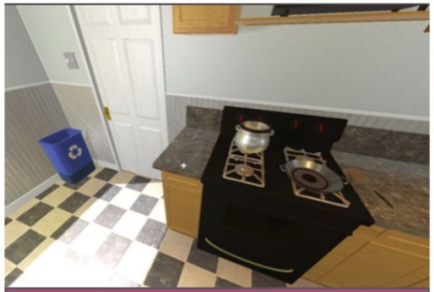
	Distance	Rel. position	Support	Containment
Random	24%	36%	52%	44%
2.5D (RGB-D image only)	32%	28%	76%	68%
ConceptFusion	**84%**	**76%**	**96%**	**72%**

Outlook

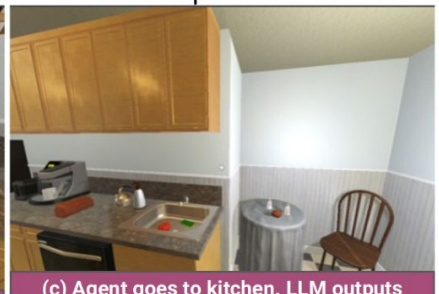
- Integration with Large Language Models (LLMs)

ConceptFusion-Agent	GenericLLM-Agent
User: I want to find some eggs. The rooms I see in the house are “bedroom”, “living room”, “kitchen”, “hallway”, and “yard”. Rank these in the decreasing order of likelihood for this task. **LLM**: 1. Kitchen 2. Hallway (if you have an egg storage container) 3. Yard (if you have chickens or hens) **User:** I am now in the kitchen. The objects I see in the room are “cabinets”, “sink”, “stove”, “trash can”, “microwave”, “fridge”, and “table”. Rank these in the decreasing order of likelihood of finding eggs. **LLM:** 1. Fridge 2. Trash can 3. Table 4. Cabinets 5. Sink 6. Stove 7. Microwave <SUCCESS>	**User:** I want to find some eggs. List some places where I can find them in the house. **LLM**: 1. Grocery store 2. Local farmers market 3. Health food store 4. Restaurant 5. Deli <FAIL>

(a) Agent spawns in house



(b) LLM outputs “kitchen”



(c) Agent goes to kitchen. LLM outputs “fridge”

(d) “Eggs” found

Outlook

- Multilingual Abilities (with Multilingual CLIP models)

Limitations

1. **Memory Intensity**
    - Millions of 3D points (apartment-scale scene)
    - High-dimensional concept embeddings for each point
2. **Computational Demand**
    - Requires processing similarity for every point for a single query
3. **Mapping Challenges in Dynamic Scene**
    - Offline feature extraction process (10~15 seconds / image)
    - Inherits challenges of SLAM

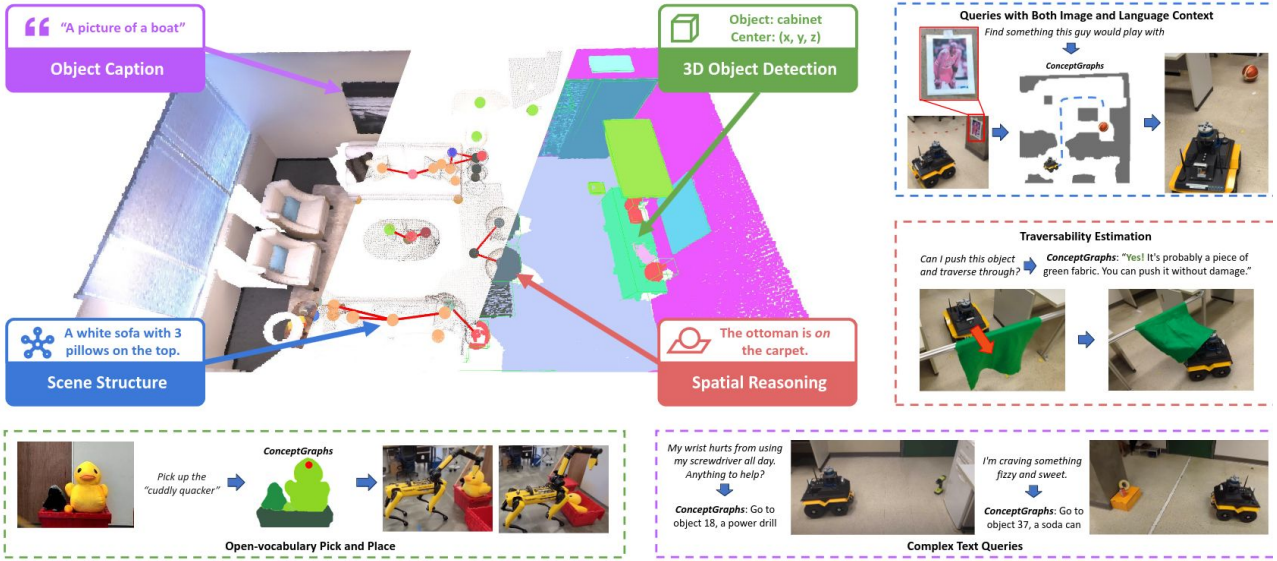
Future Work

1. **Scalability & Efficiency**
    - Filter essential features (e.g., Object-centric 3D representation)

2. **Dynamic updates**
    - Exclude irrelevant dynamic objects for map (e.g., Human)
    - Continual map updates for object movement

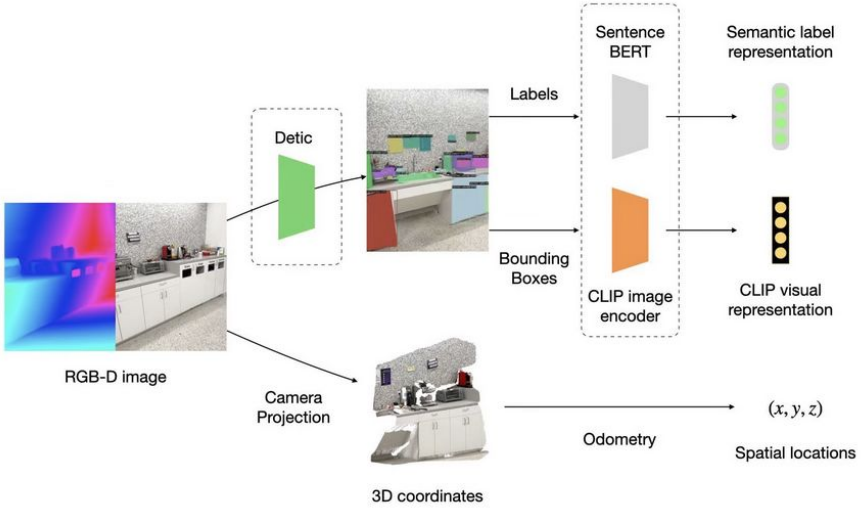
Follow-up Work

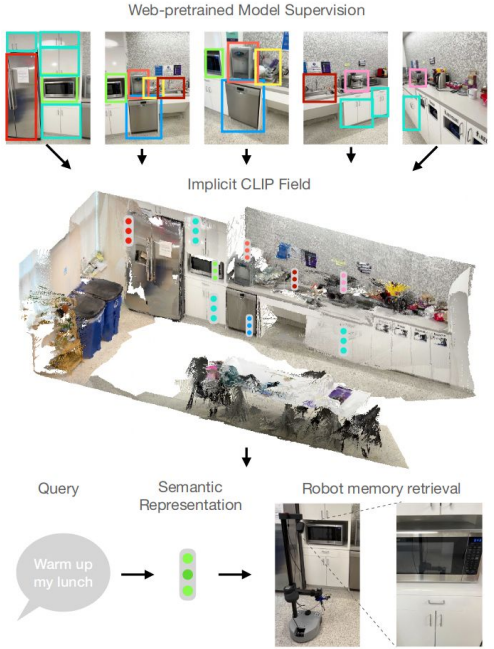
- **ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning**
  [Q Gu, A Kuwajerwala et al., arXiv. preprint arXiv:2309.16650, 2023]



Concurrent Work – 2D Foundation Features for 3D Scene

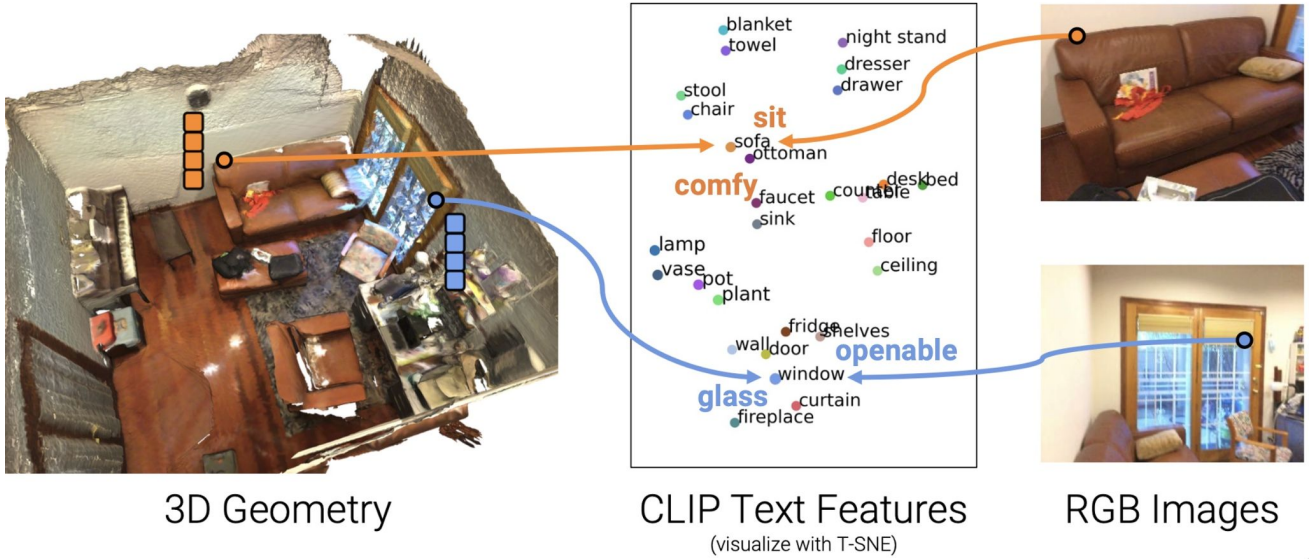
- **CLIP-Fields:** Weakly Supervised Semantic Fields for Robotic Memory
  [Mahi Shafiullah et al., RSS 2023]





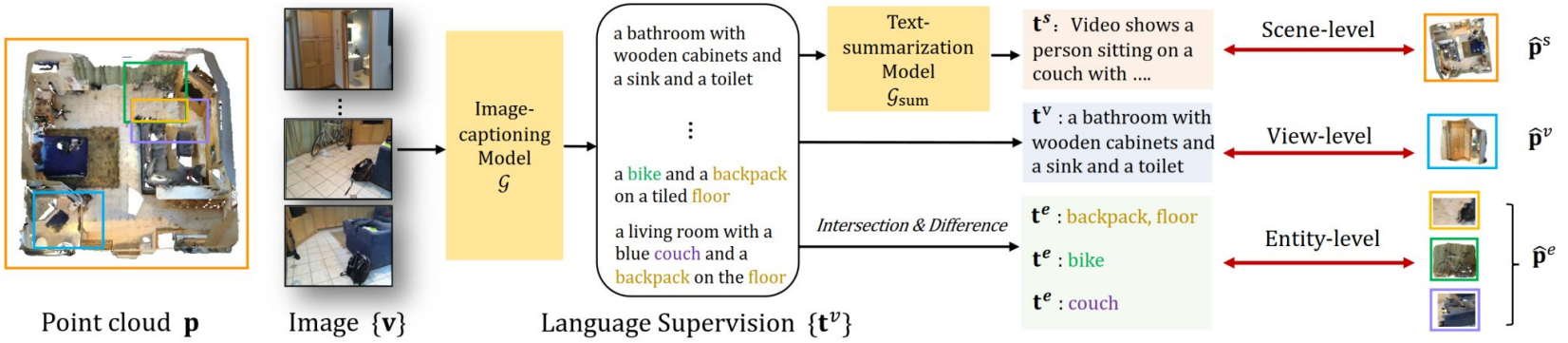
Concurrent Work – 2D Foundation Features for 3D Scene

- **OpenScene:** 3D Scene Understanding With Open Vocabularies
  [Songyou Peng et al., CVPR 2023]



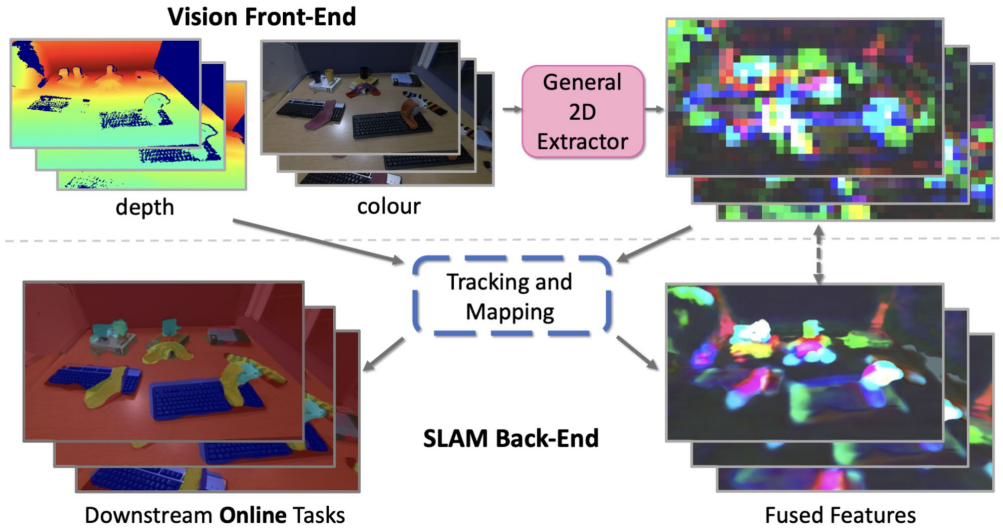
Concurrent Work – 2D Foundation Features for 3D Scene

- **PLA:** Language-Driven Open-Vocabulary 3D Scene Understanding
  [R Ding, J Yang et al., CVPR 2023]



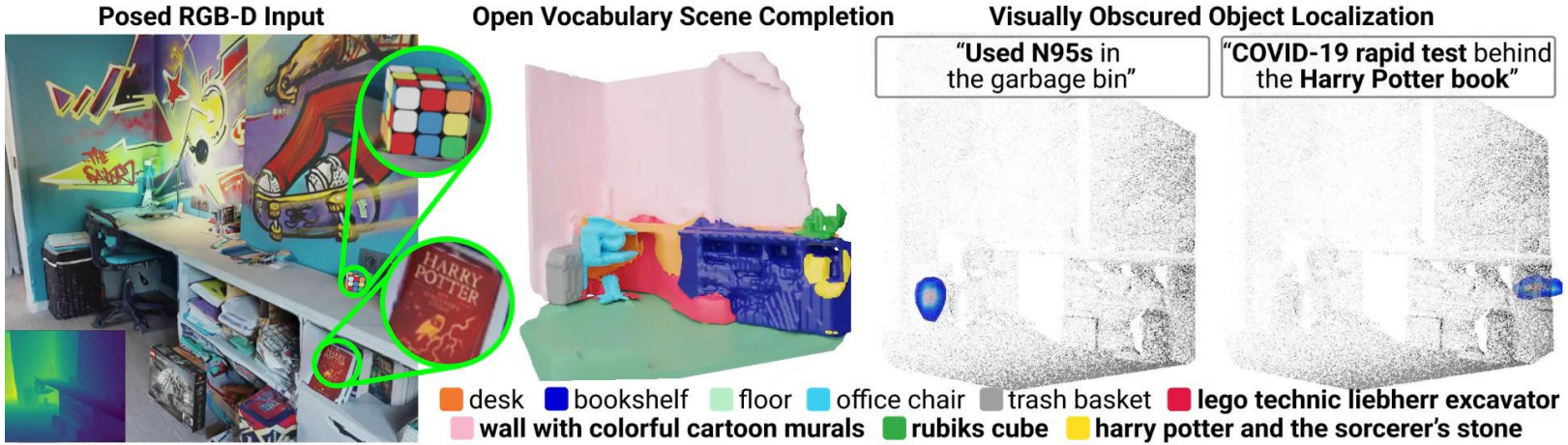
Concurrent Work – 2D Foundation Features for 3D Scene

- **Feature-Realistic Neural Fusion** for Real-Time, Open Set Scene Understanding
  [Kirill Mazur et al., ICRA 2023]



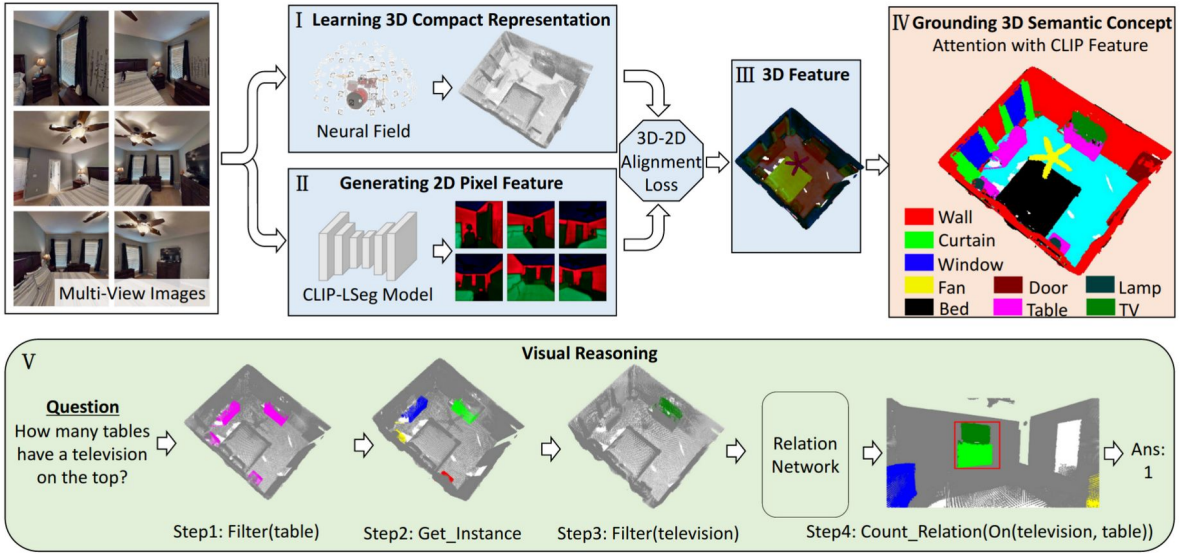
Concurrent Work – 2D Foundation Features for 3D Scene

- **Semantic Abstraction**: Open-World 3D Scene Understanding from 2D Vision-Language Models [Huy Ha & Shuran Song, CoRL 2022]



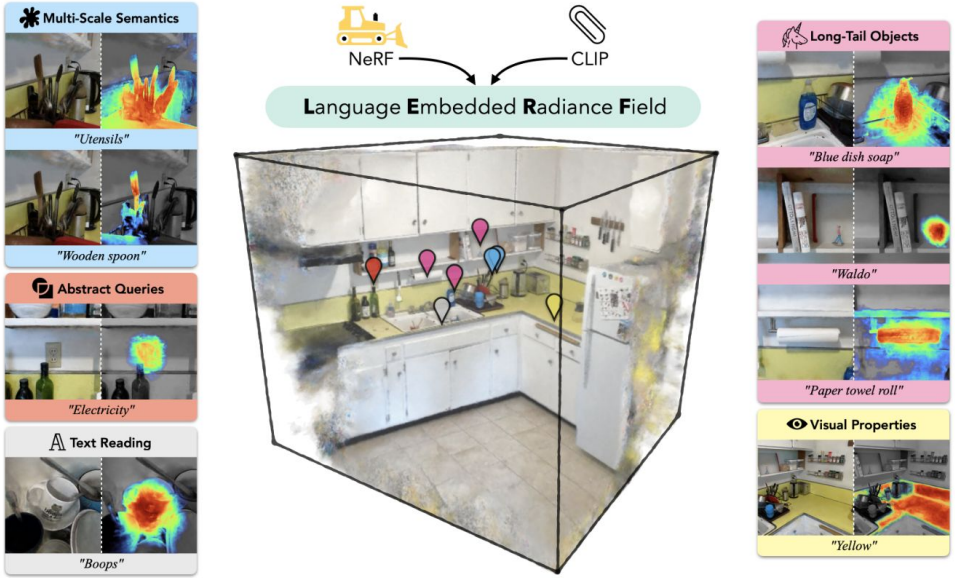
Concurrent Work – 2D Foundation Features for 3D Scene

- **3D-CLR**: 3D Concept Learning and Reasoning from Multi-View Images
  [Yining Hong et al., CVPR 2023]



Concurrent Work – 2D Foundation Features for 3D Scene

- **LERF**: Language Embedded Radiance Field [J Kerr, CM Kim et al., ICCV, 2023]



Extended Readings

- **Weakly Supervised 3D Open-vocabulary Segmentation**
  [Kunhao Liu et al., arXiv. preprint arXiv:2305.14093, 2023]

- **Audio Visual Language Maps for Robot Navigation**
  [Chenguang Huang et al., arXiv preprint arXiv:2303.07522, 2023]

- **RegionPLC: Regional Point-Language Contrastive Learning for Open-World 3D Scene Understanding** [J Yang, R Ding et al., arXiv preprint arXiv:2304.00962, 2023]

Summary

- ***ConceptFusion*** build **open-set multimodal queryable 3D map**

  → **Foundation models** + **Traditional Mapping System**

- Pixel-aligned features capturing **long-tailed** and **fine-grained** concepts

- No training required (Zero-Shot)