NS3D: Neuro-Symbolic Grounding of 3D Objects and Relations

Joy Hsu, Jiayuan Mao, Jiajun Wu
joycj@stanford.edu, jiayuanm@mit.edu, jiajunw@stanford.edu

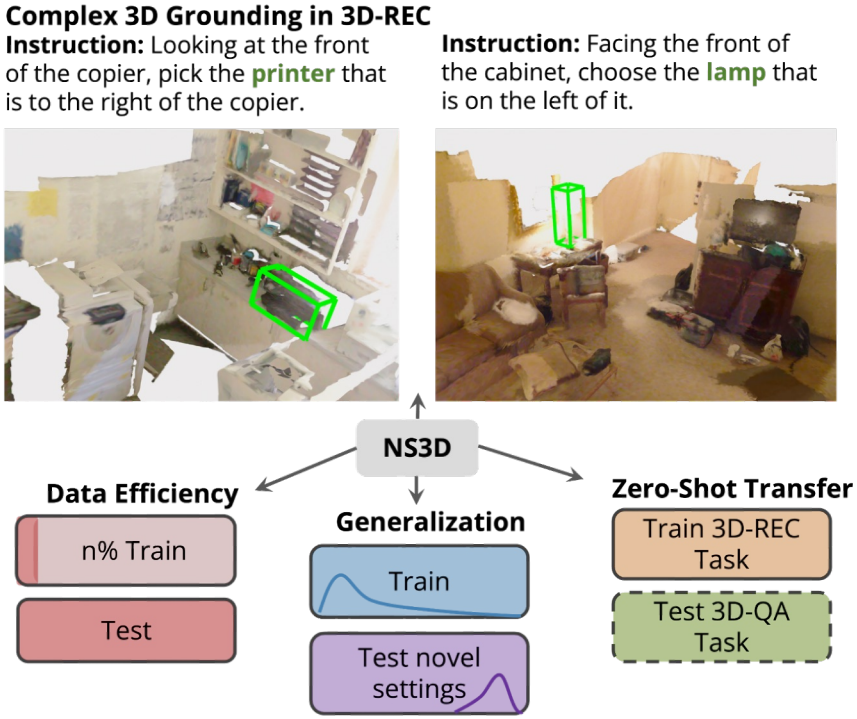
TUE-AM-249

Neuro-symbolic 3D grounding

Neuro-symbolic method to ground 3D objects and relations that integrates

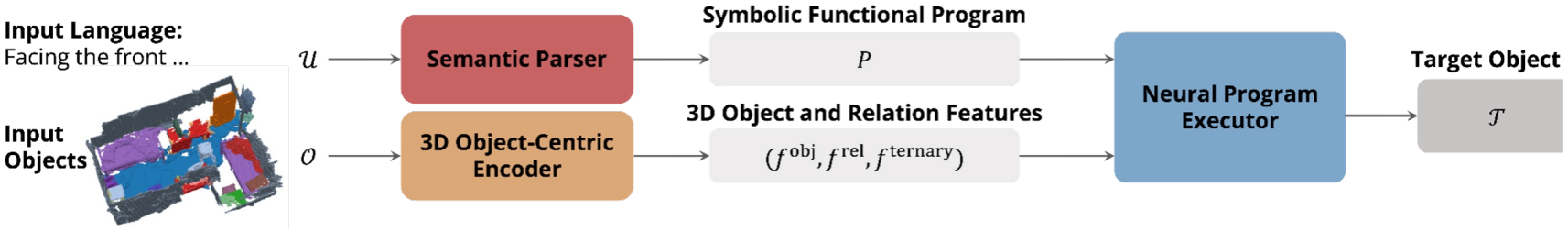
- **Large language models (LLMs)**
- **Modular neural networks**

Structured approach to reasoning that decomposes tasks into modules

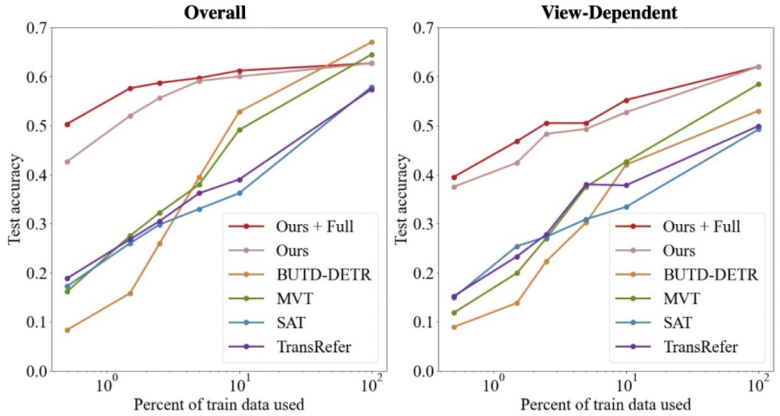


Joy Hsu joycj@stanford.edu

NS3D



NS3D outperforms prior end-to-end works



Project page:

Joy Hsu joycj@stanford.edu

NS3D

Neuro-symbolic method that integrates the power of **LLMs** and **modular neural networks**

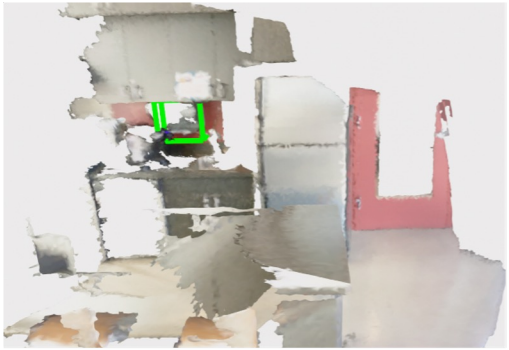
Neural program executor that reasons about **high-arity relations**

Strong performance in view-dependent grounding, **data-efficiency**, **generalization**, and ability to **zero-shot transfer**

Task

ReferIt3D input
- Object point clouds
- Referring expression

ReferIt3D output
- Target object

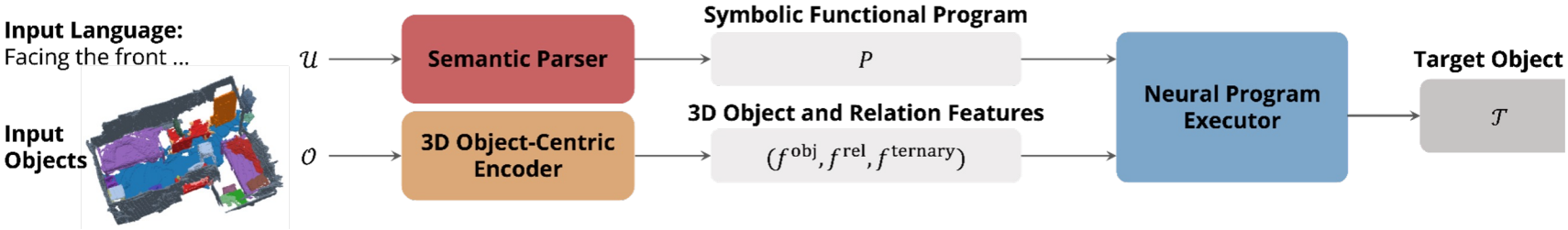


Instruction: Select the **microwave** that is close to the kitchen cabinets.

Instruction: Facing the front of the couch, choose the **towel** that is on the left side of it.

Instruction: Choose the **suitcase** that is in the middle of the bathtub and the cabinet.

NS3D



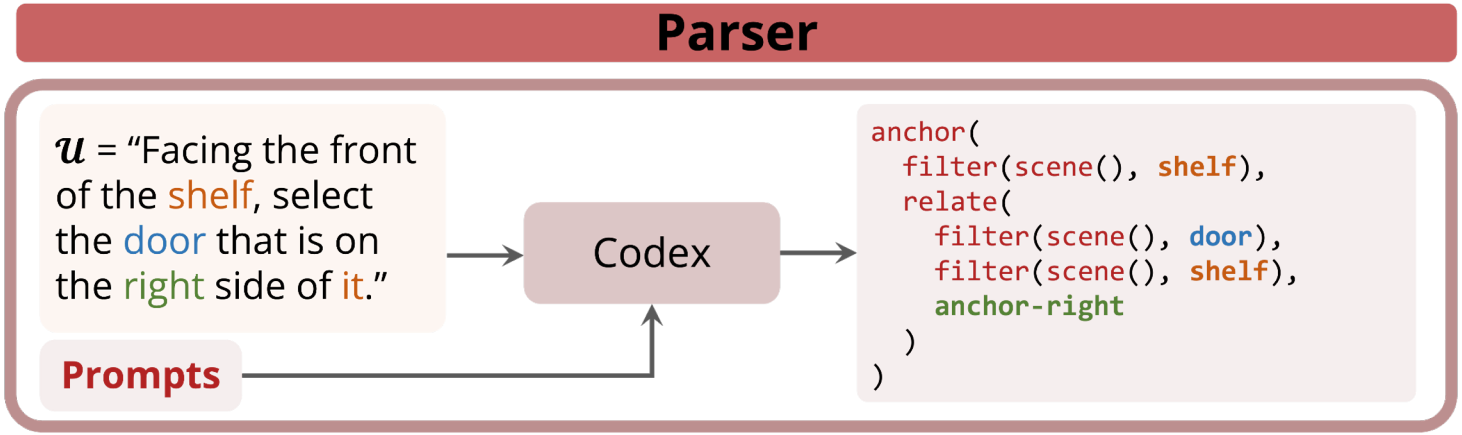
NS3D is composed of three main components:

1) A **semantic parser**

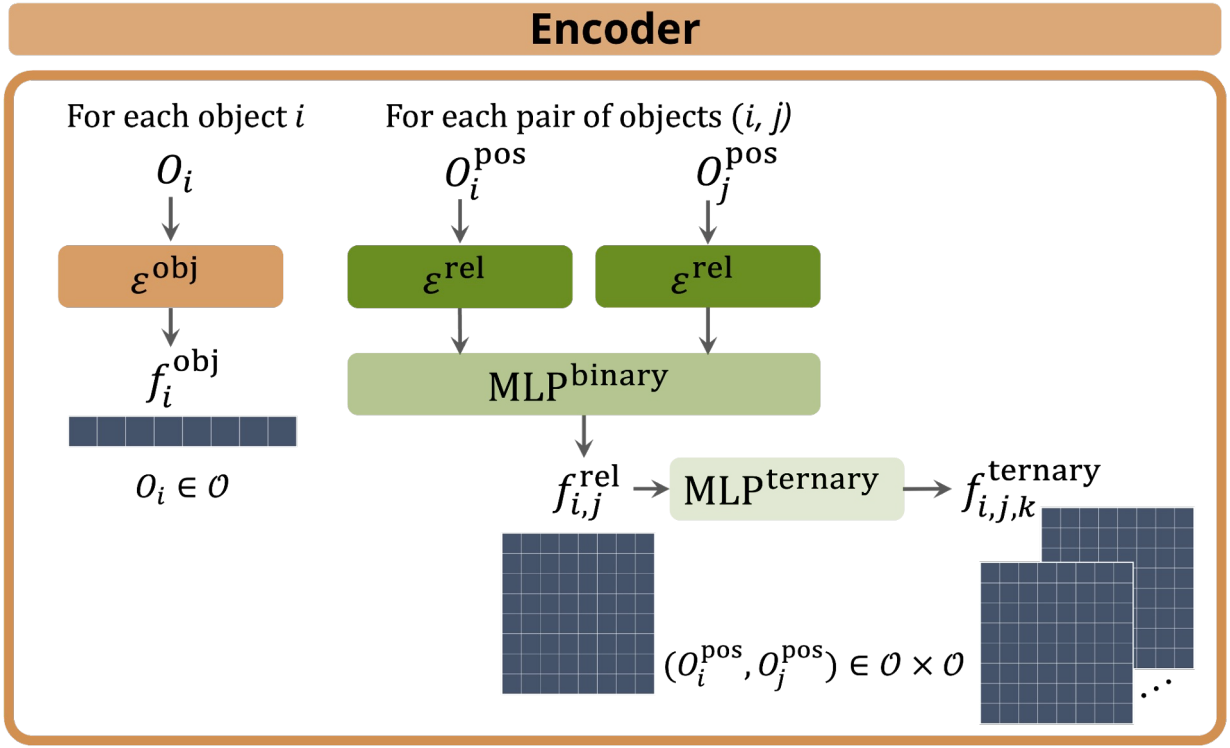
2) A **3D object-centric encoder**

3) A **neural program executor**

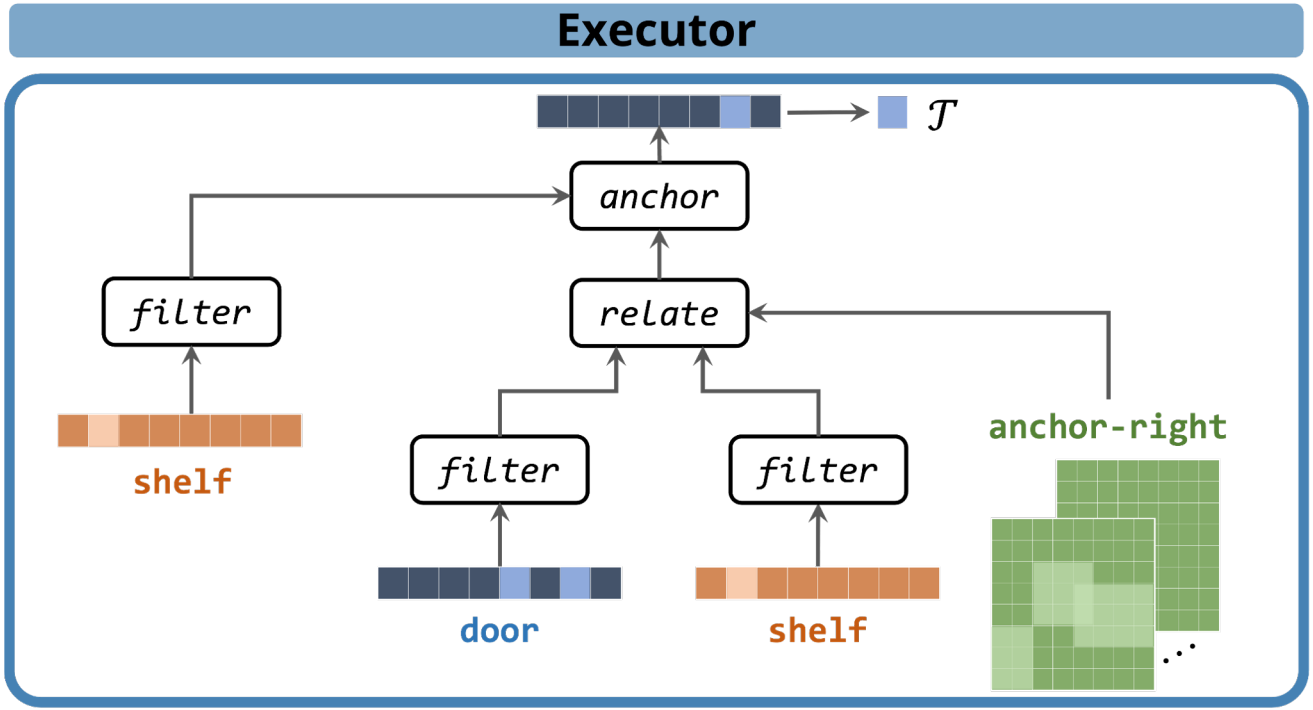
Semantic parser



3D object-centric encoder



Neural program executor



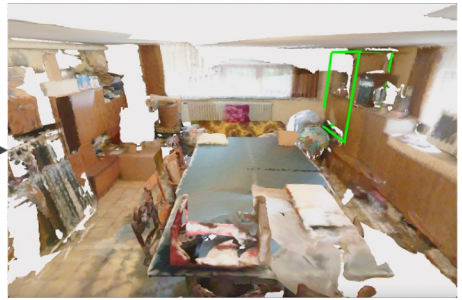
Execution trace

u = "Facing the couch from the side you sit on it, choose the cabinet that is to the right of this couch."

```
filter(scene(), couch)
```

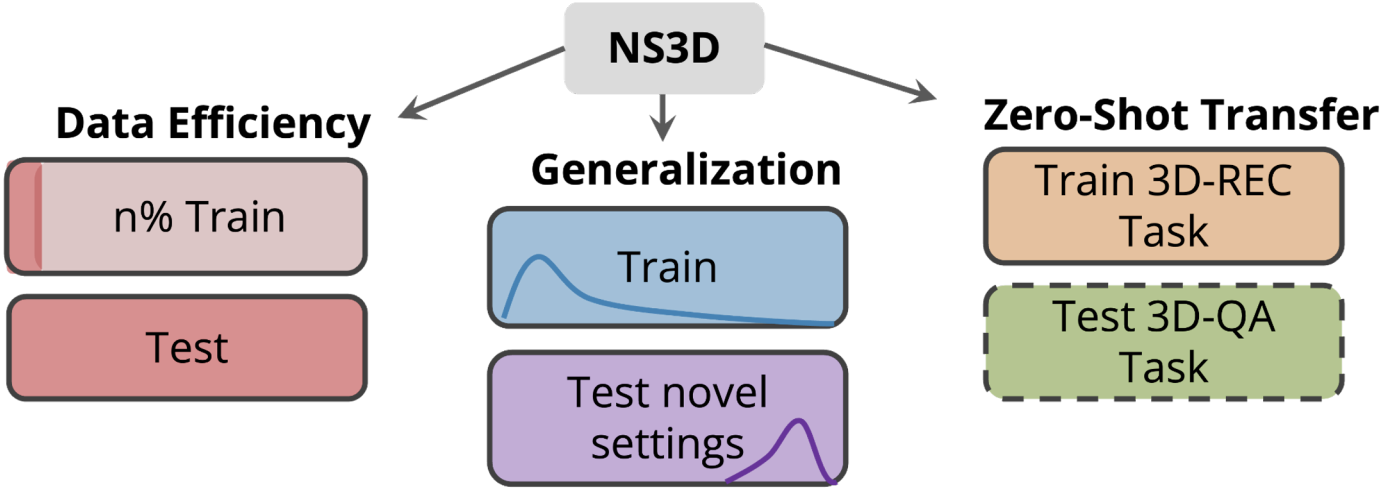
```
filter(scene(), cabinet)
```

```
anchor(
  filter(couch),
  relate(
    filter(cabinet),
    filter(couch),
    anchor-right
  )
)
```



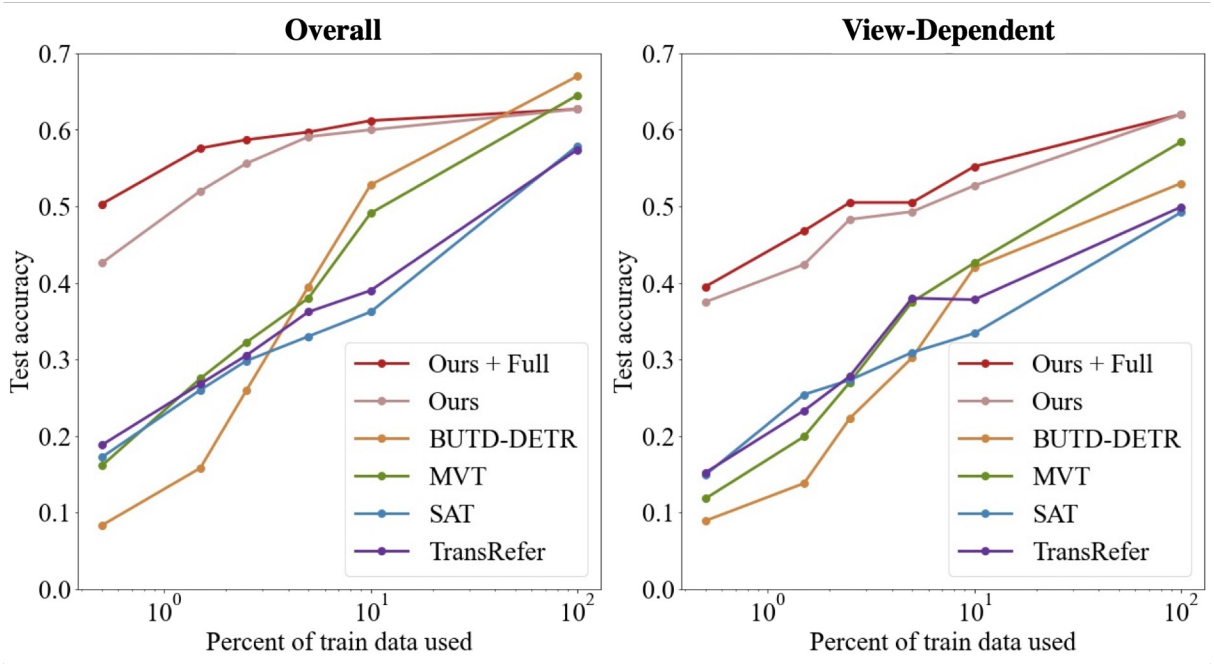
Joy Hsu joycj@stanford.edu

Results


NS3D
Data Efficiency
n% Train
Test
Generalization
Train
Test novel settings
Zero-Shot Transfer
Train 3D-REC Task
Test 3D-QA Task

Data efficiency

NS3D excels at data efficiency



Joy Hsu joycj@stanford.edu

Generalization

NS3D's modularity leads to strong generalization

	Pairs		Scene	
	All	V-dep.	All	V-dep
NS3D + full (Ours)	**0.612**	**0.635**	**0.563**	0.583
NS3D (Ours)	0.599	0.620	0.544	**0.611**
BUTD-DETR [20]	0.440	0.423	0.515	0.583
MVT [18]	0.420	0.353	0.502	0.500
SAT [37]	0.359	0.380	0.451	0.500
TransRefer [16]	0.322	0.344	0.384	0.361

Zero-shot transfer

NS3D can zero-shot transfer to novel, unseen tasks

	ALL	EXIST	COUNT	OBJ	REL
NS3D + CODEX	**0.68**	**0.80**	**0.67**	**0.60**	**0.60**
NS3D + T5	0.30	0.40	0.13	0.40	0.30
RANDOM	0.16	0.40	0.07	0.00	0.10

Exist

Q: Is there a **chair** between the **table** and the **window**?
A: Yes **NS3D:** Yes

Count

Q: How many **keyboards** are in the scene?
A: 6 **NS3D:** 6

Object

Q: What is the object under the **laptop**?
A: Table **NS3D:** Table

Relation

Q: Facing the **fridge**, what is the relation between the **stove** and the **fridge**?
A: Right **NS3D:** Right

Summary

NS3D integrates

- Large language models (LLMs)
- Modular neural networks

and enables

- Data efficiency
- Generalization
- Zero-shot transfer

Project page:
https://web.stanford.edu/~joycj/projects/ns3d_cvpr_2023