Language and affordance guided grasp pose generation

BY JOHANNES KOLHOFF

Outline

Motivation/reminder

Task definition

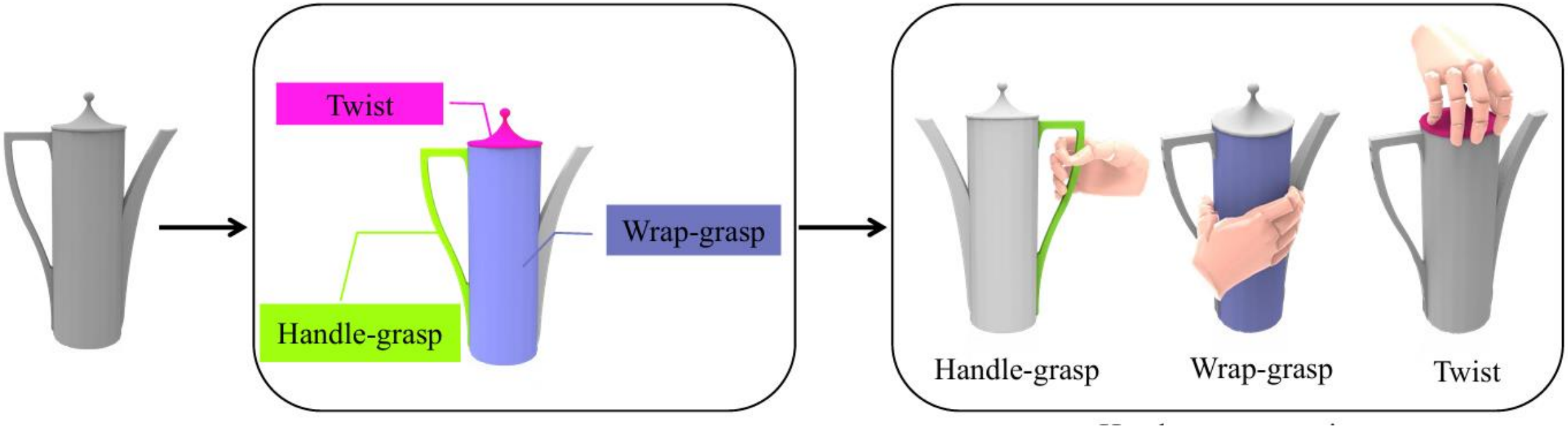
Current models:

- Language-Conditioned Affordance-Pose Detection
- Language Guided Affordance (AffordDexGrasp)

Summary

Motivation

Where would you grasp this teapot?



[2]

What is Affordance

„Affordance“ depends on:

Object shape

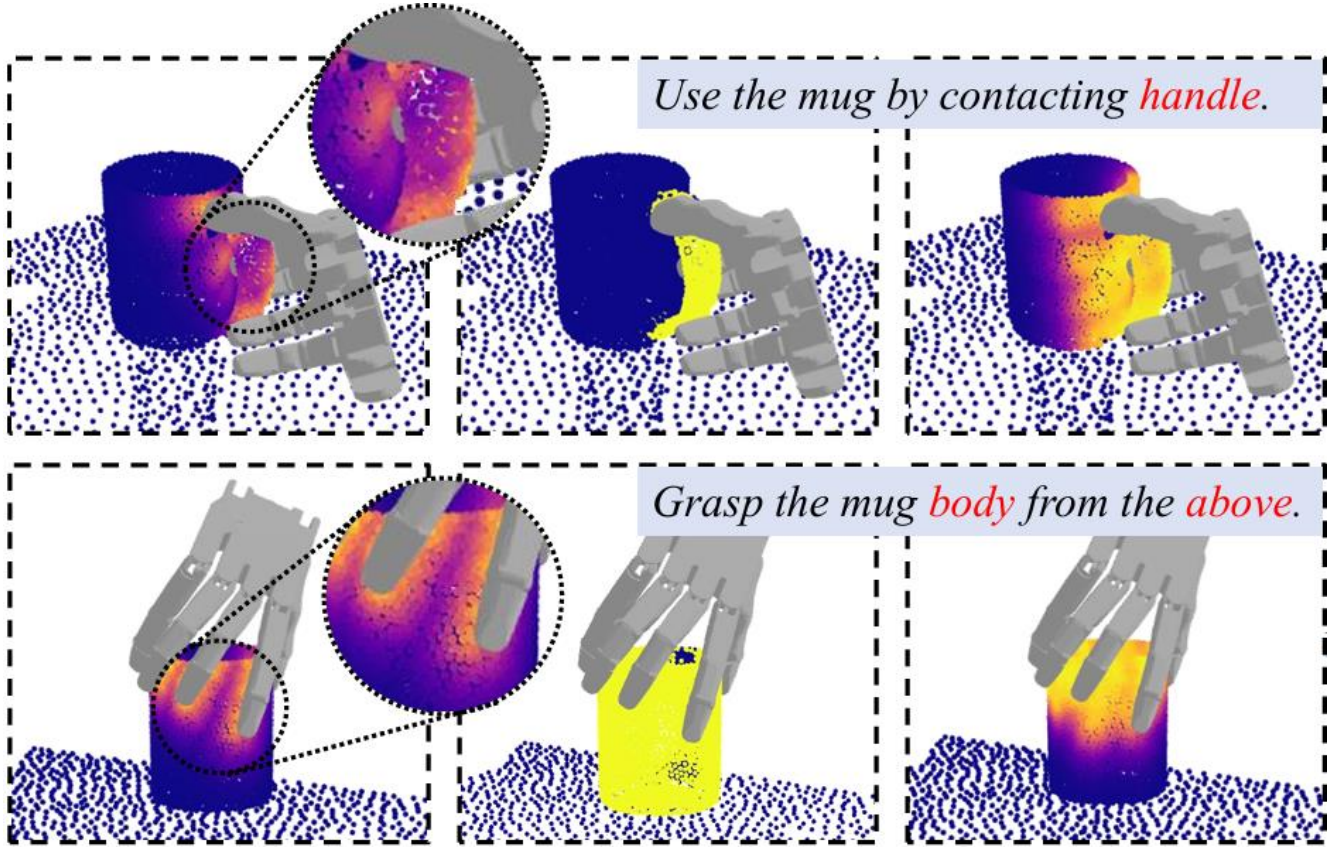
Object type (what is it)

Intention (what do you want to do?)

Also, your manipulator

=> Object in its context!

How an object can be used



[5]

Challenges:

- Understanding the objects
- Unknown objects
- Many affordance category
- Bridge to 6DoF grasp pose
- Bridge from language guided
- Cluttered environments (pile)
- Datasets

Generalizability

Task description

Input

- 3D point cloud
- Natural language

Affordance: predict Object context

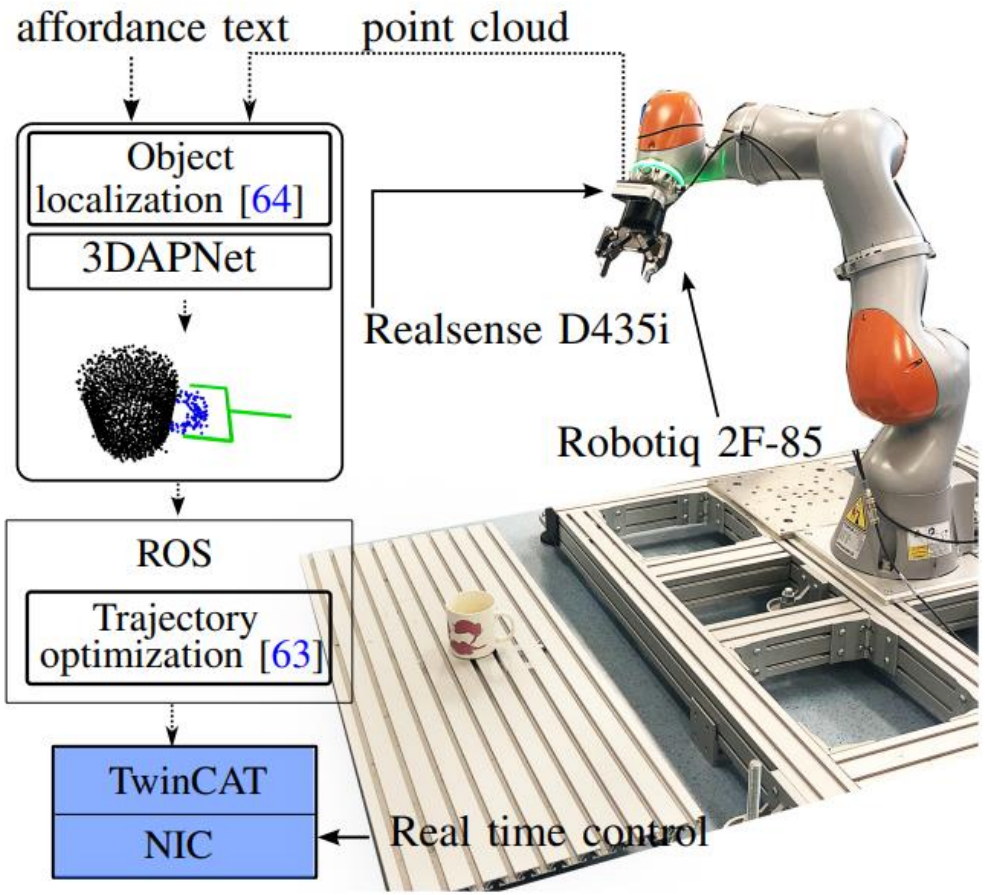
Generate according grasp pose

Output: Grasp pose

[2]

Excluded from Task

- Identifying object
- Obtaining point cloud
- [System]
- Kinematics
- Path planning
- Etc.



[2]

Approaches

Encoding point cloud and text as embeddings

Processing point cloud to explicit affordance map (point-based label)

Implicit affordance map

Pose generation

Language-Conditioned Affordance-Pose Detection - Dataset

New dataset based on 3D AffordanceNet and 6-DoF GraspNet

28K gripper poses in total

Manual annotation of affordance category's

Converted to point cloud (set to 2,048 points)

Dataset is triplet of point could, affordance category and pose

No explicit affordance in data set! (No affordance ground truth)

[4]

Language-Conditioned Affordance-Pose Detection

Input: point cloud and text (affordance)

Point cloud and text are encoded separately

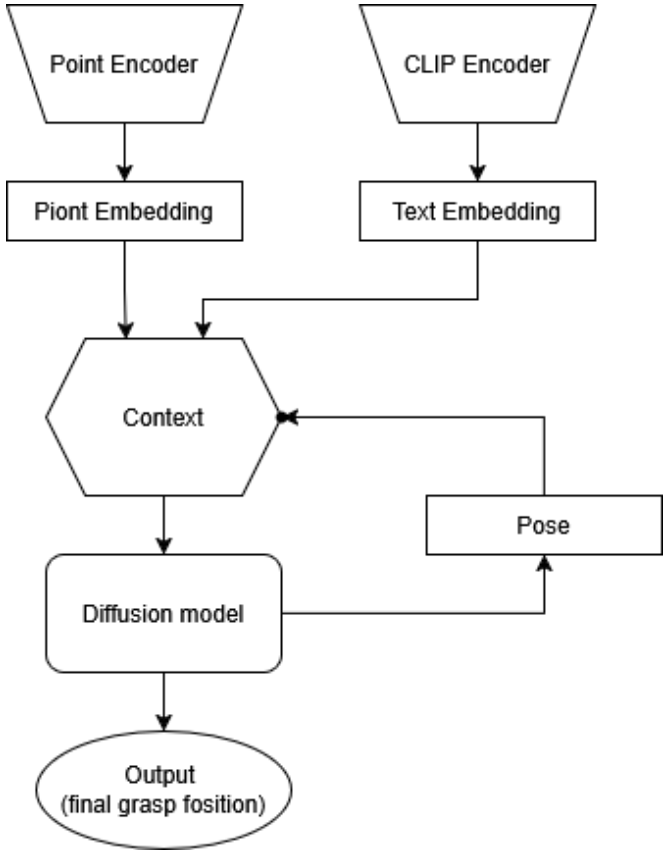
Grasp estimation is diffusion based

Context for diffusion is made up of embeddings of:

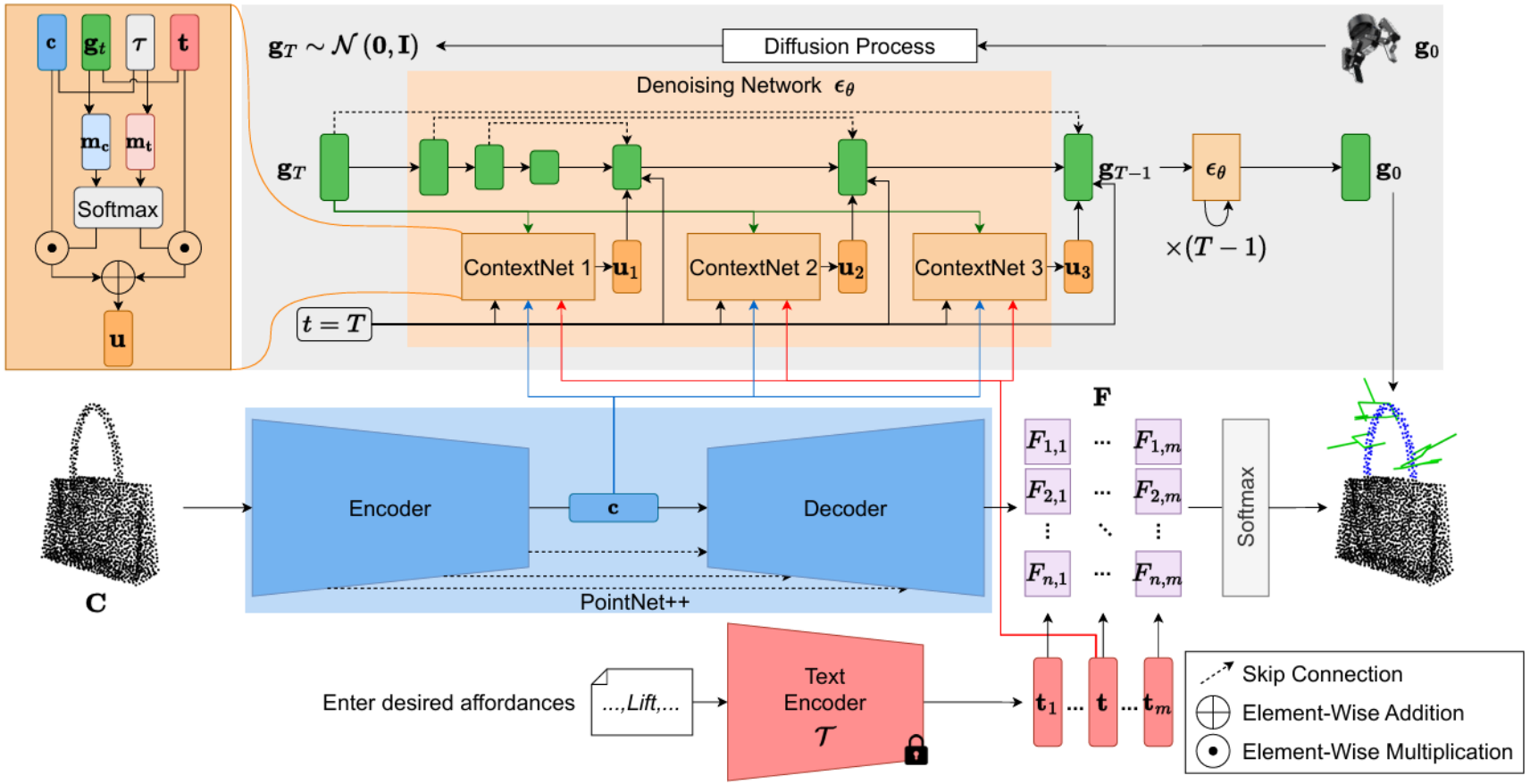
- Point cloud, text, robot state and timestep

One network (end to end)

Language-Conditioned Affordance-Pose Detection



Language-Conditioned Affordance-Pose Detection



[4]

Language-Conditioned Affordance-Pose Detection – Training

Text encoder is frozen

Point encoder is trained to correspond to text embeddings

- Affordance loss

Diffusion model it trained on gaussian noise on pose T = 1000

- Pose loss

Whole network is trained

- Total loss = Affordance loss + Pose loss

Trained as single network

Language-Conditioned Affordance-Pose Detection – Results

Affordance metrics:

- mIoU (mean Intersection over union)
- Acc (overall accuracy of all points)
- mAcc (mean accuracy over all affordances)

Pose generation metrics:

- mESM (mean evaluated similarity metric)
- mCR (mean coverage rate)

Confirmed the two loss functions correlate

Training Context net is worth it

TABLE II

SINGLE-BRANCH VS. JOINTLY LEARNED NETWORK

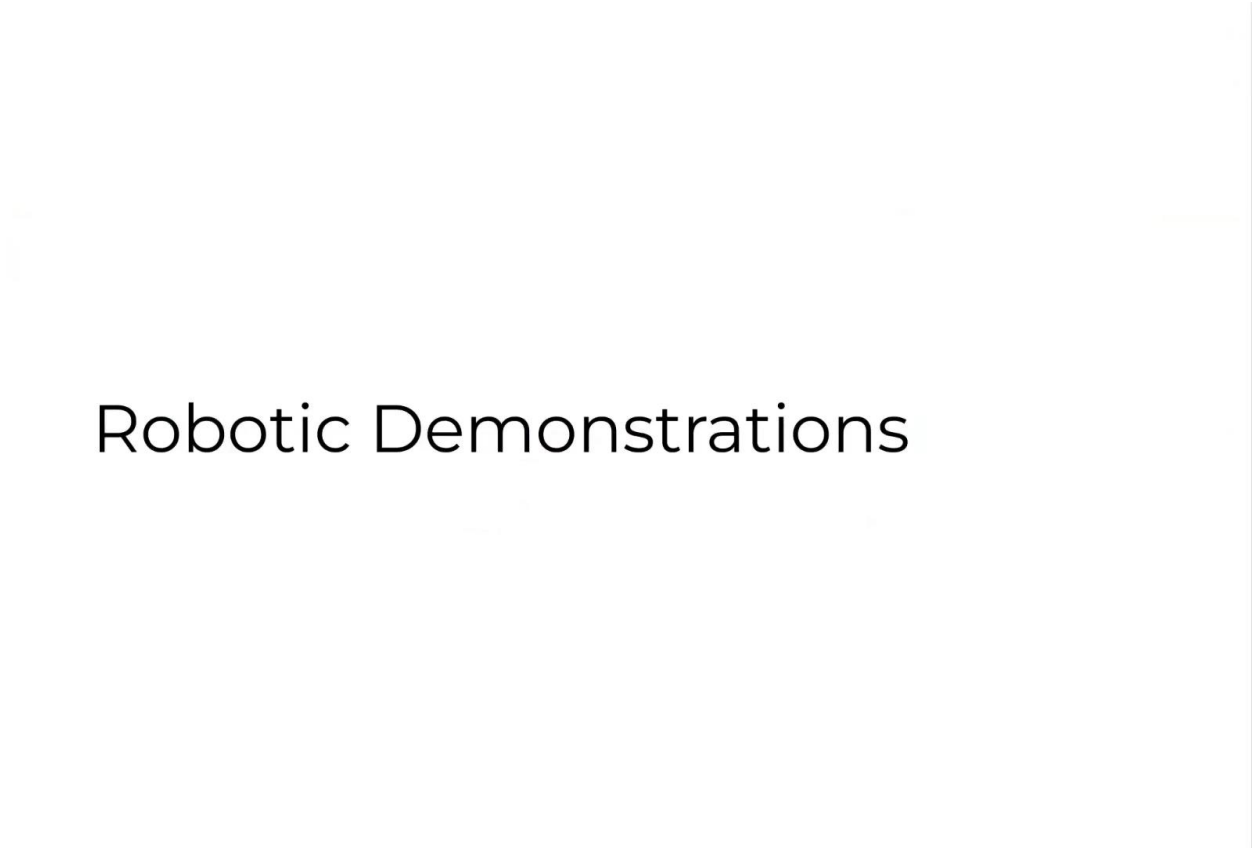
Method	Affordance Detection			Pose Estimation	
	mIoU↑	Acc↑	mAcc↑	mESM↓	mCR↑
Affordance Only	55.65	60.73	58.80	–	–
Pose Only	–	–	–	0.147	41.29
Both	**56.18**	**61.77**	**59.26**	**0.120**	**44.63**

TABLE III

THE EFFECTIVENESS OF CONTEXTNET

ContextNet	Affordance Detection			Pose Estimation	
	mIoU↑	Acc↑	mAcc↑	mESM↓	mCR↑
✗	53.97	60.20	58.49	0.433	12.61
✓	**56.18**	**61.77**	**59.26**	**0.120**	**44.63**

Demonstration



Language-Conditioned Affordance-Pose Detection

ADVANTAGE

Open vocabulary

Expansion to new affordances and objects possible

Diffusion (few samples)

DISADVANTAGE

Always requires affordance instruction

Expansion to new affordances limited

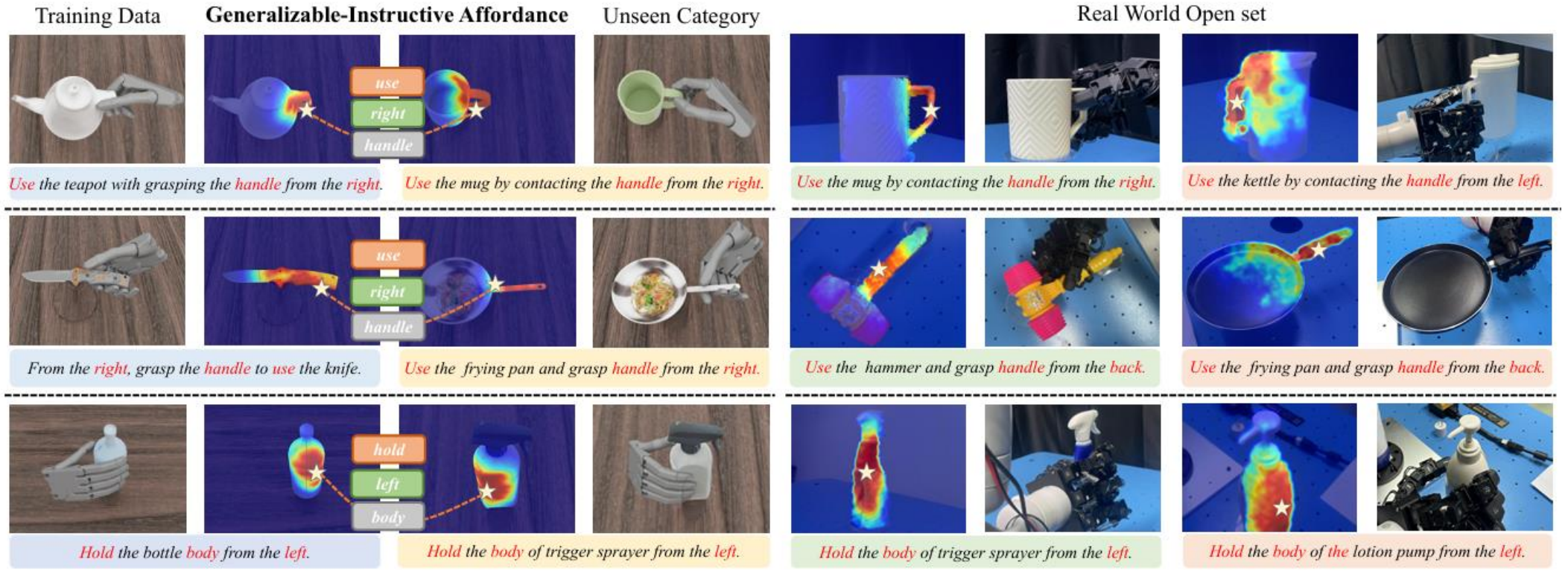
Word choice matters

Single object only

No explicit affordance

AffordDexGrasp: Dataset

New open set dataset based on language-guided dexterous grasp dataset
tabletop environment and 33 categories



[5]

AffordDexGrasp:

Input:

- Natural language user input
- RGB image
- Point cloud

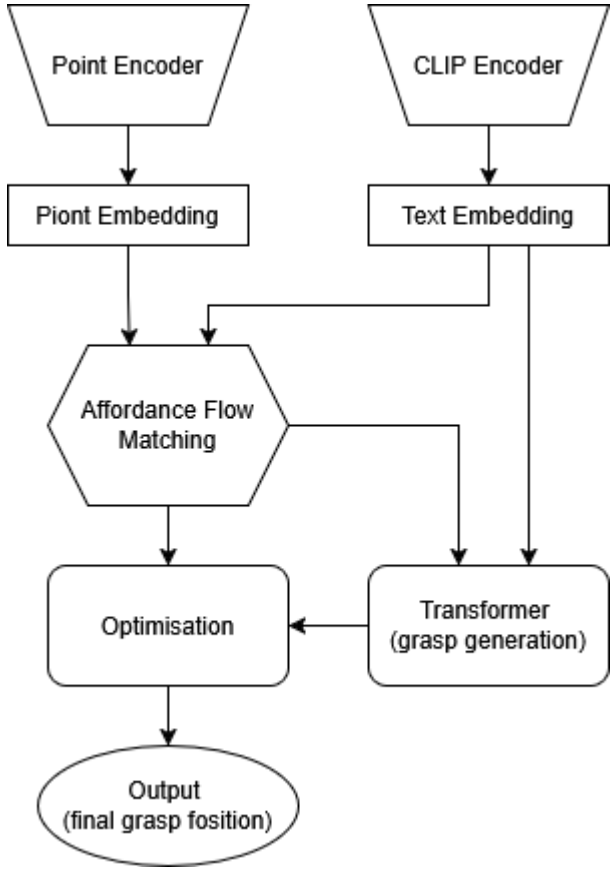
Utilizes MLLM to intuit and guide affordance

Closer to sequential processing:

- First create an affordance map
- Then determine gasp from affordance map + affordance encoding

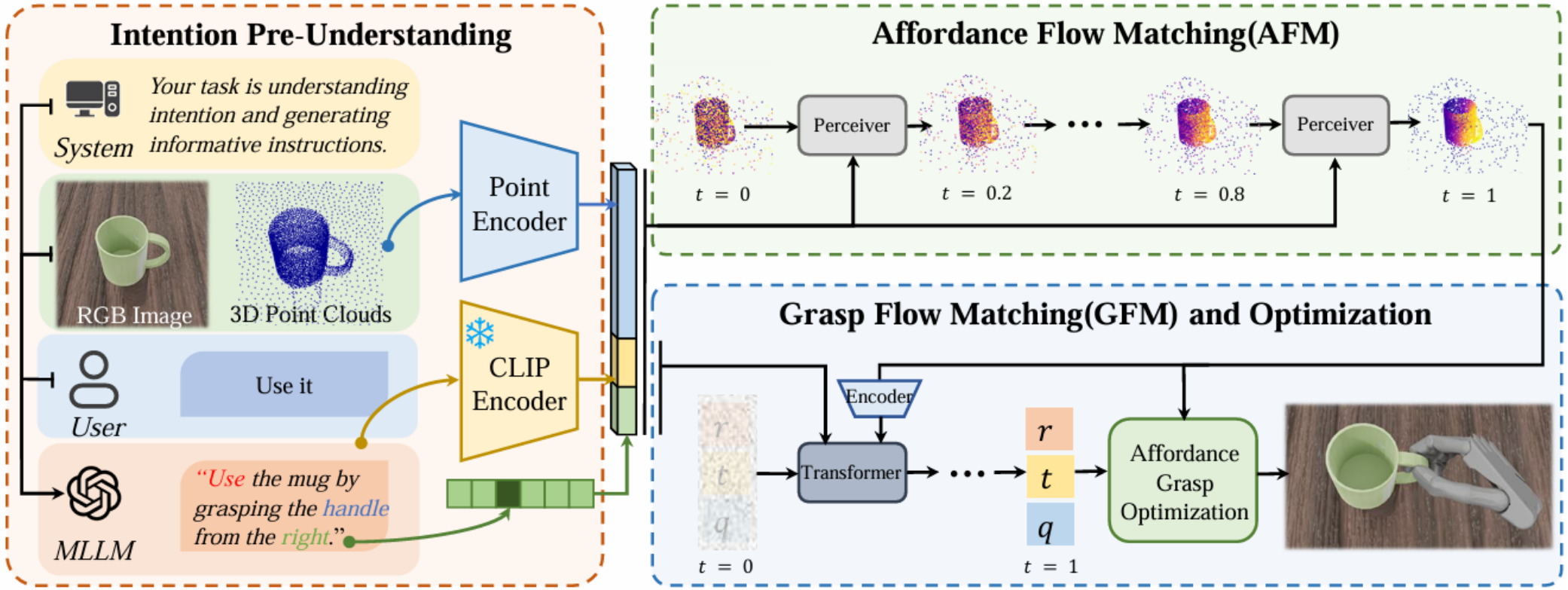
[5]

AffordDexGrasp:

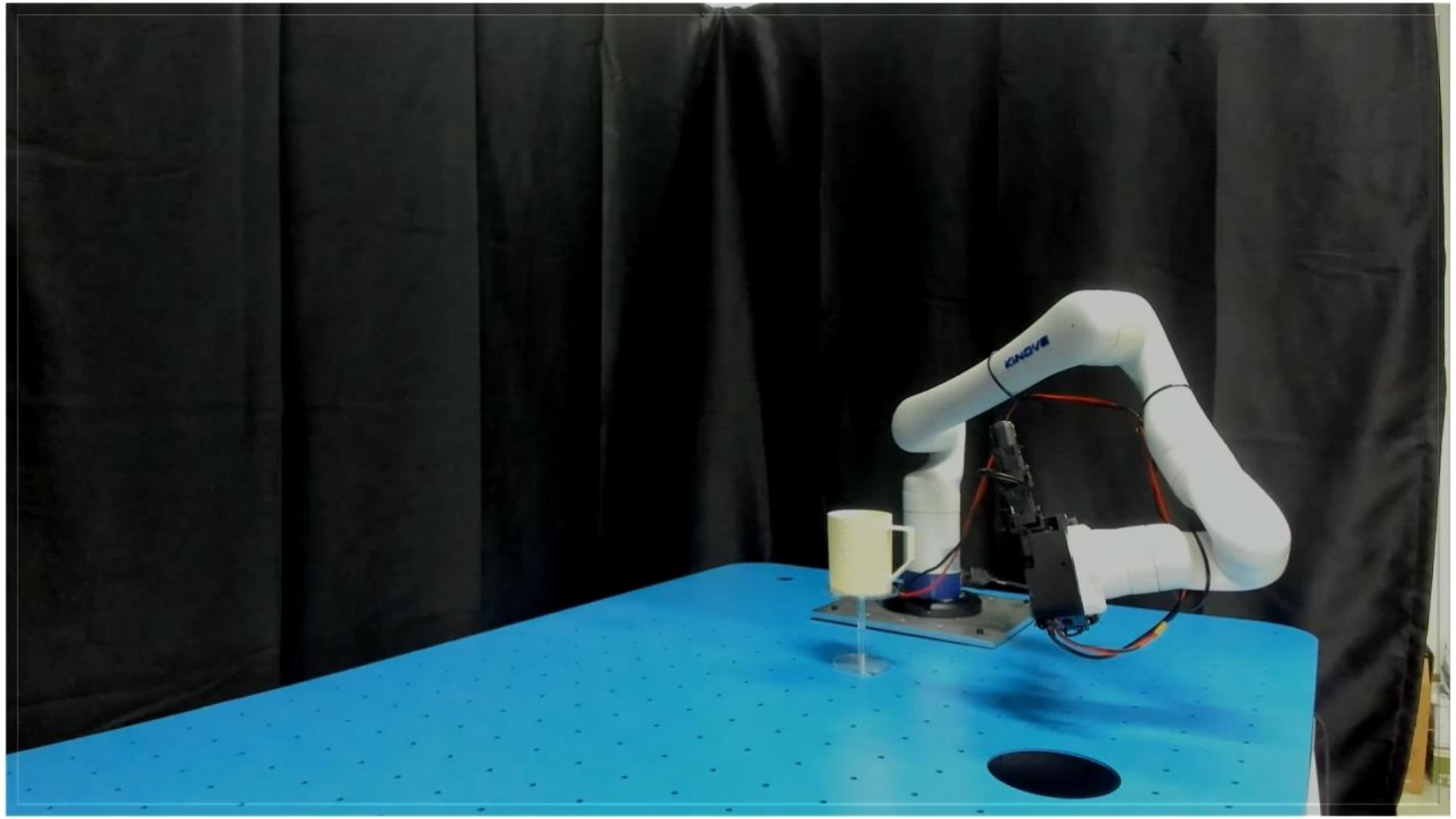


[5]

AffordDexGrasp:



Demonstration



[5]

AffordDexGrasp:

ADVANTAGE

Open to natural language instructions
- But not necessary!

Use of RGB image

Robust to different formulations

Actual affordance map

DISADVANTAGE

Input needs RBG image, point cloud and instruction for best performance

Still has real world gap

Affordance map is based on gaussian distribution

Summary

Utilize 3D/2.5D vision

Gather context information (image + text/situation)

Integration of affordance and grasp pose estimation

Challenges:

- Pressley defining the problem (+dataset)
- Measuring success
- Applying to real world (multiple objects, point cloud render etc.)

Sources

[1] “Visual Affordance and Function Understanding: A Survey“ 2018 Mohammed Hassanin, Salman Khan, Murat Tahtali

[2] “AffordPose: A Large-scale Dataset of Hand-Object Interactions with Affordance-driven Hand Pose” 2023 Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, Jian Liu

[3] “Learning 6-DoF Task-oriented Grasp Detection via Implicit Estimation and Visual Affordance” 2022 Wenkai Chen, Hongzhuo Liang, Zhaopeng Chen, Fuchun Sun and Jianwei Zhang

[4] “Language-Conditioned Affordance-Pose Detection in 3D Point Clouds” 2024 Toan Nguyen, Minh Nhat Vu, Baoru Huang, Tuan Van Vo, Vy Truong, Ngan Le Thieu Vo, Bac Le, Anh Nguyen

[5] “AffordDexGrasp: Open-set Language-guided Dexterous Grasp with Generalizable-Instructive Affordance” 2025 Yi-Lin Wei, Mu Lin, Yuhao Lin, Jian-Jian Jiang, Xiao-Ming Wu, Ling-An Zeng, Wei-Shi Zheng