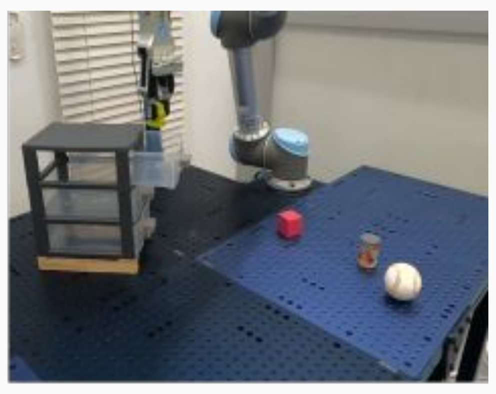
Towards Generalizable Vision-Language Robotic Manipulation: A Benchmark and LLM-guided 3D Policy

Ricardo Garcia-Pinel, Shizhe Chen*, Cordelia Schmid*

Inria Paris, École normale supérieure, PSL

Introduction



"put the frog toy in the top drawer"

Goal: Enhance the generalization capabilities of vision-language robotic manipulation policies.

Limitations of state-of-the-art methods:

- Train and test policies on the same task set
- Focus on a limited set of action skills (pick-and-place)

Our contributions:

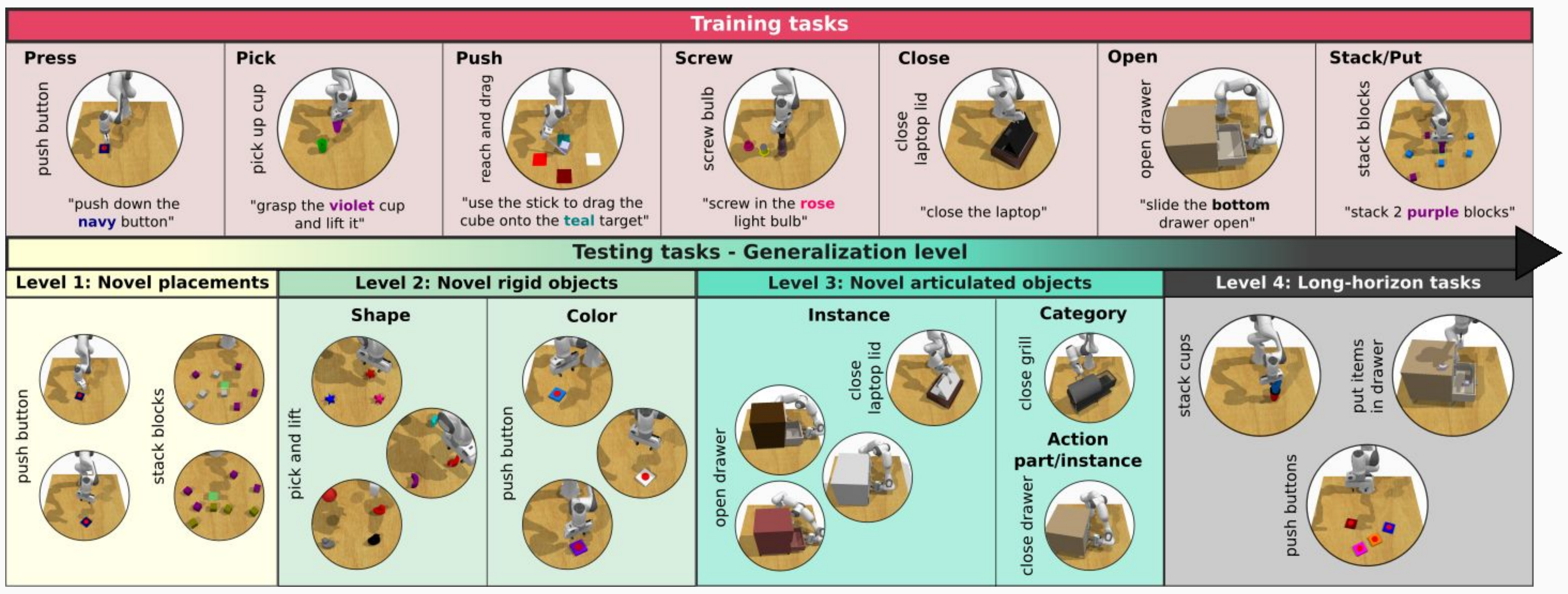
1. A **comprehensive benchmark:** covering 7 action skills and 4 generalization levels
2. A generalist **LLM-guided 3D policy**:

+ 3D-based robotic manipulation policy: more precise action prediction
+ Integration with LLMs and VLMs: improved generalization ability

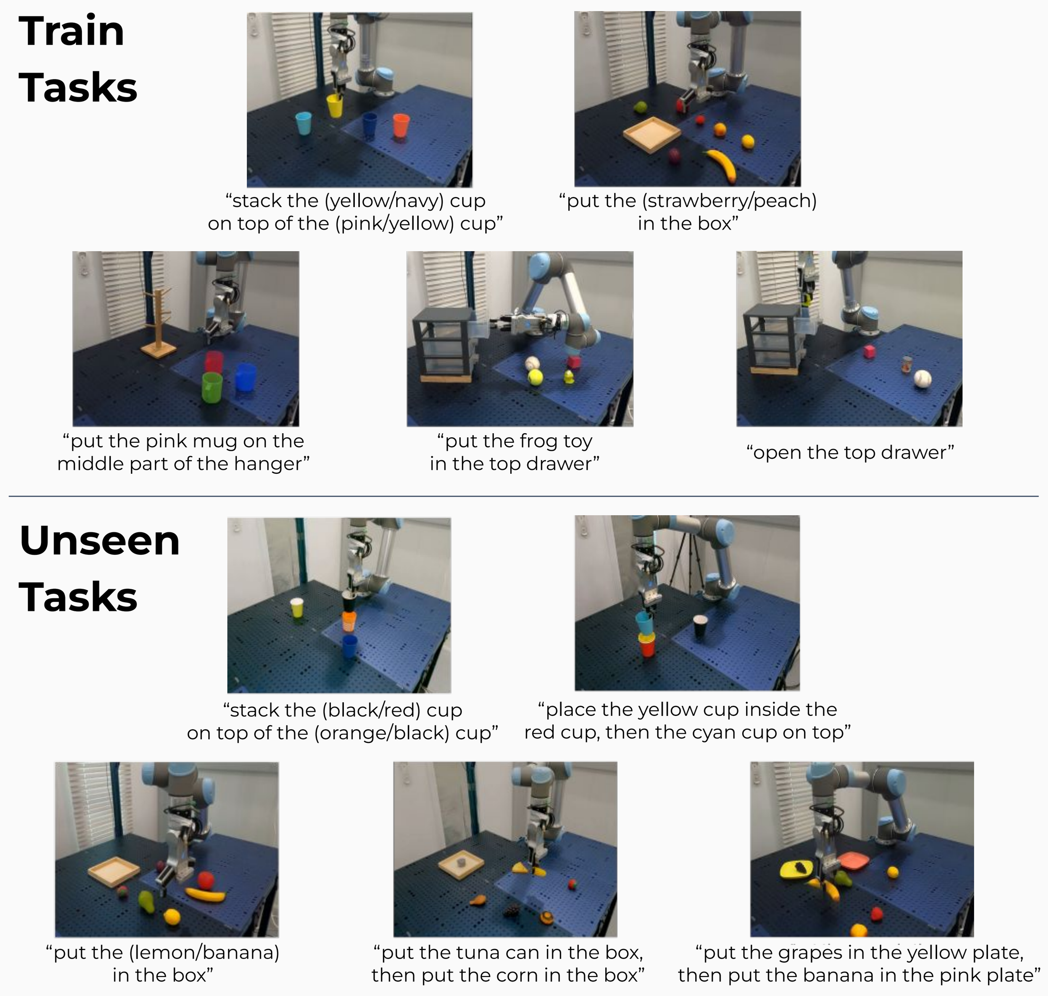
GEMBench: GEneralizable Vision-Language Robotic Manipulation Benchmark

Train set: 16 tasks (31 variations) / 7 action primitives.

Test set: 44 tasks (92 variations) / 4 levels of generalization.



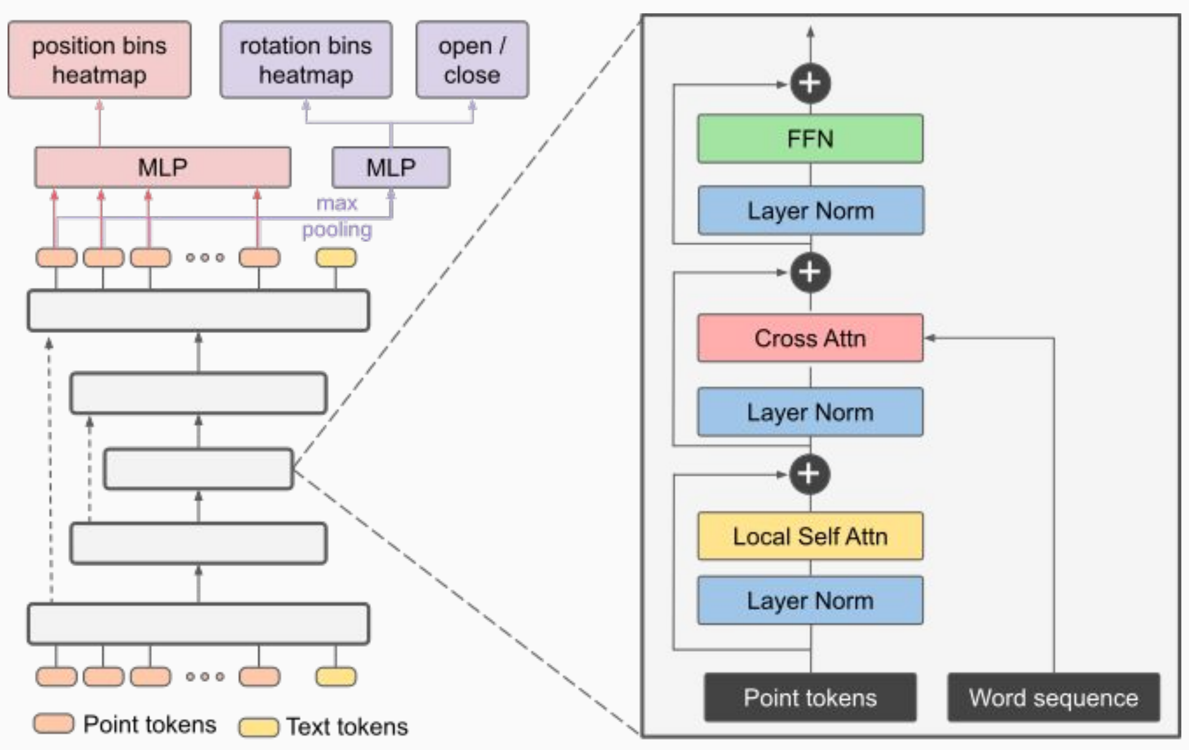
Real Robot Setups



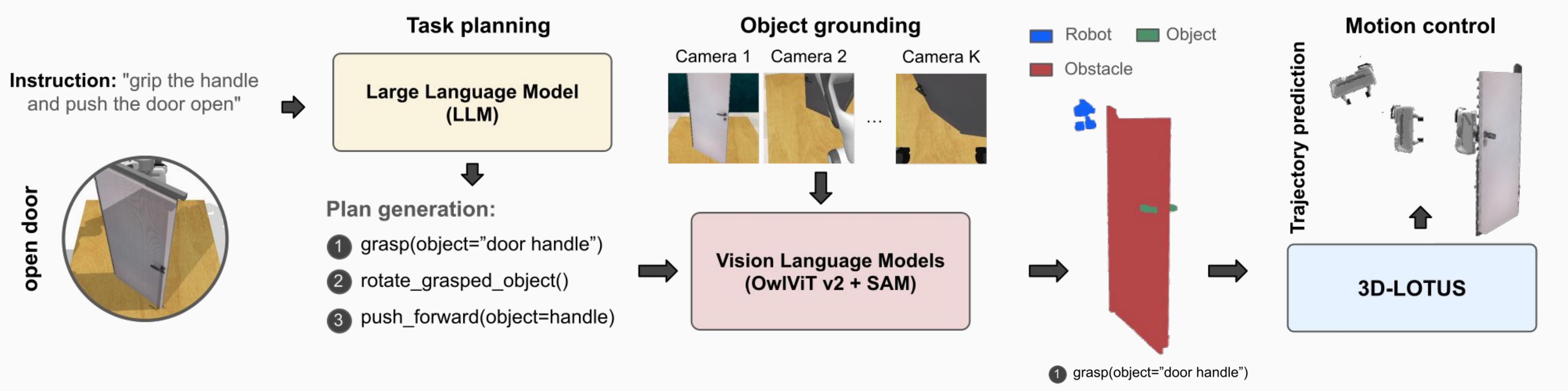
The Proposed Method

3D-LOTUS policy:

- Efficient Point Transformer v3 backbone
- Improved precision via point-wise classification



3D-LOTUS++ framework:



① LLaMA-8B performs task planning by splitting high-level goal instructions into a sequence of primitive actions.

② OWLv2 + SAM process multiple views and the target object name for visual grounding.

③ 3D-LOTUS visuomotor policy predicts precise robot trajectory for a given target object and primitive action.

Experimental Results

Evaluation on RLBench-18Task

Achieved SoTA performance and faster training speed.

	Avg. SR ↑	Avg. Rank ↓	Train time ↓
C2F-ARM-BC [38]	20.1	8.6	-
Hiveformer [17]	45.3	6.9	-
PolarNet [2]	46.4	6.4	8.9
PerAct [18]	49.4	6.2	128.0
RVT [34]	62.9	4.4	8.0
Act3D [4]	65.0	4.3	40.0
RVT2 [37]	81.4	2.4	6.6
3D diffuser actor [35]	81.3	2.3	67.6
3D-LOTUS	**83.1** $_{\pm 0.8}$	**2.2**	**2.2**[3]

Evaluation on GemBench

3D-LOTUS++ performs better on more challenging generalization levels.

Method	L1	L2	L3	L4
Hiveformer [17]	60.3 $_{\pm 1.5}$	26.1 $_{\pm 1.4}$	35.1 $_{\pm 1.7}$	0.0 $_{\pm 0.0}$
PolarNet [2]	77.7 $_{\pm 0.9}$	37.1 $_{\pm 1.4}$	38.5 $_{\pm 1.7}$	0.1 $_{\pm 0.2}$
3D diffuser actor [35]	91.9 $_{\pm 0.8}$	43.4 $_{\pm 2.8}$	37.0 $_{\pm 2.2}$	0.0 $_{\pm 0.0}$
RVT-2 [37]	89.1 $_{\pm 0.8}$	51.0 $_{\pm 2.3}$	36.0 $_{\pm 2.2}$	0.0 $_{\pm 0.0}$
3D-LOTUS	**94.3** $_{\pm 1.4}$	49.9 $_{\pm 2.2}$	38.1 $_{\pm 1.1}$	0.3 $_{\pm 0.3}$
3D-LOTUS++	68.7 $_{\pm 0.6}$	**64.5** $_{\pm 0.9}$	**41.5** $_{\pm 1.8}$	**17.4** $_{\pm 0.4}$

Ablation on GemBench

The motion policy and object grounding are the main bottlenecks for generalizable robotic manipulation.

Task Planning	Object Grounding	Avg.
GT	GT	63.0
GT	VLM	50.7
LLM	VLM	48.0

Real world results

Task	PolarNet	3D-LOTUS
Stack yellow cup in pink cup	**10/10**	9/10
Stack navy cup in yellow cup	9/10	**10/10**
Put strawberry in box	7/10	**10/10**
Put peach in box	**8/10**	**8/10**
Open drawer	6/10	**9/10**
Put item in drawer	1/10	**3/10**
Hang mug	6/10	**8/10**
Avg.	6.7/10	**8.1/10**

Seen Tasks

Task	3D-LOTUS	3D-LOTUS++
Stack red cup in yellow cup	0/10	**8/10**
Stack black cup in orange cup	0/10	**7/10**
Place the yellow cup inside the red cup, then the cyan cup on top	0/10	**7/10**
Put lemon in box	0/10	**9/10**
Put banana in box	0/10	**7/10**
Put tuna can in box, then corn in box	0/10	**8/10**
Put grapes in yellow plate, then banana in pink plate	0/10	**9/10**
Avg.	0/10	**7.9/10**

Unseen Tasks

Project Webpage

CVPR 2025 Challenge & Workshop