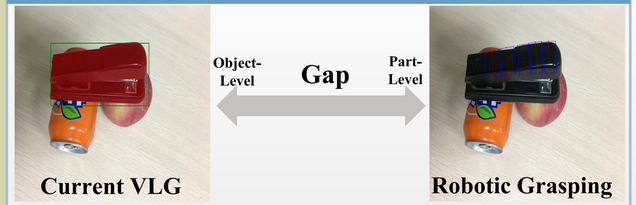
RealVLG-R1: A Large-Scale Real-World Visual-Language Grounding Benchmark for Robotic Perception and Manipulation

Linfei Li, Lin Zhang, Ying Shen
School of Computer Science and Technology, Tongji University

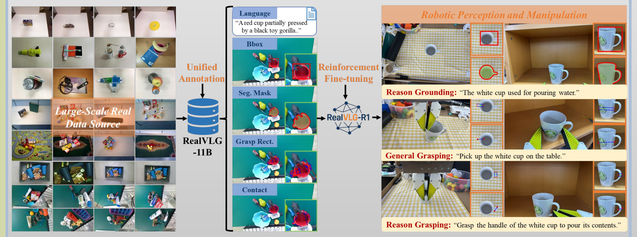
CVPR JUNE 3-7, 2026 DENVER COLORADO

1. Motivation & Problem

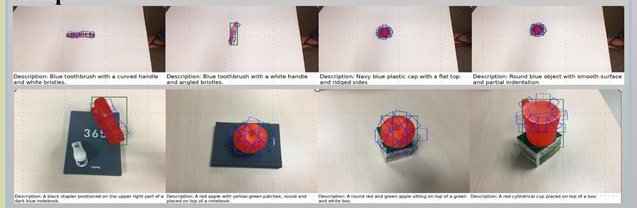


Gap: Coarse VLG vs. Geometric Grasping
Goal: Semantic-to-Manipulation Reasoning

Framework Overview



Samples:



RealVLG: Unified Vision-Language Grounding and Grasping

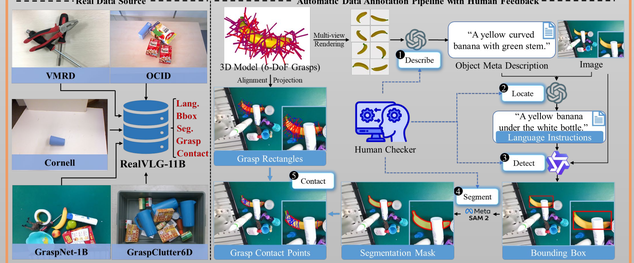
2. RealVLG-11B Dataset

DATASET STATS

165K Images	**1~20** Obj/Image	**15+** Grasp/Obj
1.3M Annotations	**800+** Instances	**11B** Grasps

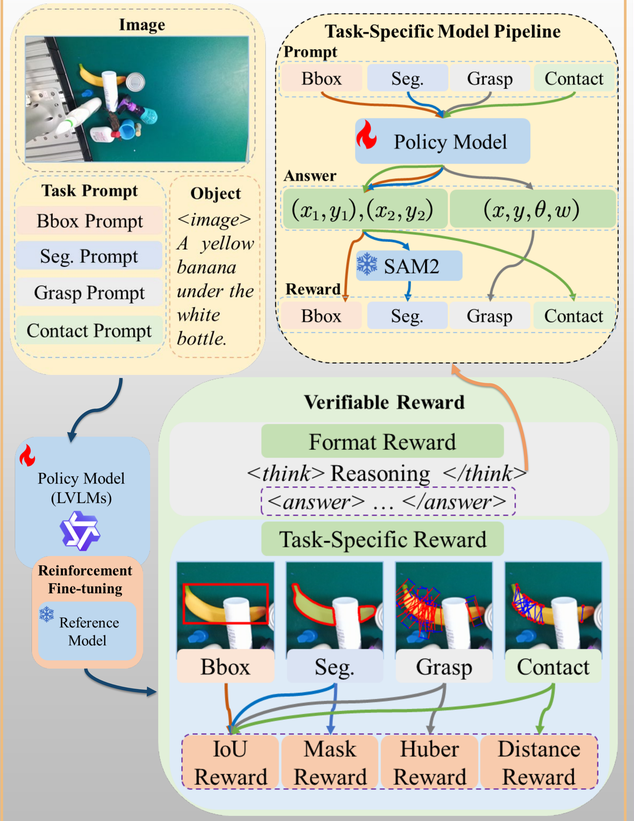
Type	Dataset	Grasp Label	Objects / Image	Scale			Grounding		Language Description			
				Images	Objects	Grasps	Seg.	Bbox	Object	Spatial	VLM-Check	Human-Check
Sim.	Dex-net [41]	Rect.	1	6.7M	1500	6.7M	✓	✗	✗	✗	✗	✗
	Jacquard [14]	Rect.	1	54K	11K	1.1 M	✓	✗	✗	✗	✗	✗
	VR-Grasping-101 [83]	6-DoF	1	10K	101	4.8M	✗	✗	✗	✗	✗	✗
	Kappler et al. [28]	6-DoF	1	700	80	300K	✗	✗	✗	✗	✗	✗
	6-DoF GraspNet [48]	6-DoF	1	206	206	7M	✗	✗	✗	✗	✗	✗
	Epper et al. [15]	6-DoF	1	21	21	1B	✗	✗	✗	✗	✗	✗
	EGAD! [47]	6-DoF	1	2231	2231	233K	✗	✗	✗	✗	✗	✗
	ACRONYM [16]	6-DoF	-	-	8872	17.7M	✗	✗	✗	✗	✗	✗
	MetaGraspNet [20]	6-DoF	~12	217K	82	-	✓	✗	✗	✗	✗	✗
Synth.	Grasp-Anything [76]	Rect.	-	1M	~3M	~600M	✓	✗	✗	✗	✗	✗
	Grasp-Anything++ [75]	Rect.	-	1M	~3M	~600M	✓	✗	✓	✗	✗	✗
Real	Cornell [26]	Rect.	1	1035	240	8019	✗	✗	✗	✗	✗	✗
	Pinto et al. [57]	Rect.	-	50K	150	50K	✗	✗	✗	✗	✗	✗
	Levine et al. [32]	Rect.	-	800K	-	800K	✗	✗	✗	✗	✗	✗
	VMRD [88]	Rect.	~5	4683	~200	100K	✗	✓	✗	✗	✗	✗
	Multi-Object [11]	Rect.	~3	96	-	2904	✗	✓	✗	✗	✗	✗
	GraspClutter6D [3]	6-DoF	~15	52K	200	9.3B	✓	✗	✗	✗	✗	✗
	OCID-Grasp [1]	Rect.	1~20	11K	~40	75K	✓	✗	✗	✗	✗	✗
	GraspNet [18]	6-DoF	~10	97K	88	~1.2B	✓	✗	✗	✗	✗	✗
	RealVLG-11B (Ours)	**Rect.**	**1~20**	**165K**	**~800**	**11B**	✓	✓	✓	✓	✓	✓

Annotation Pipeline with Human Verification



Closed-loop Verification Ensures Alignment of Language, Detection, Segmentation, and Grasping.

3. RealVLG-R1 Model



Adaptive Optimization in Multi-solution Grasp Scenarios.

4. RealVLG-Benchmark

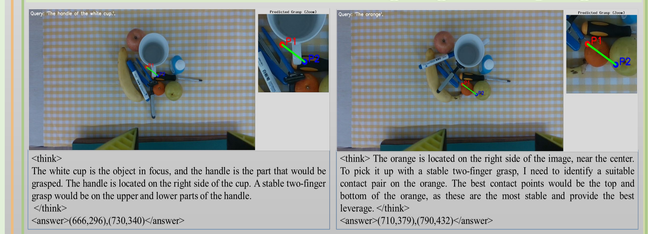
Benchmark Evaluation Summary

Benchmark

Method	Seen				Similar				Novel			
	Bbox gIoU/cIoU/R_c	Seg. F_1/S_α/R_c	Grasp mIoU/IgAcc/R_c	Contact mIoU/IgAcc/R_c	Bbox gIoU/cIoU/R_c	Seg. F_1/S_α/R_c	Grasp mIoU/IgAcc/R_c	Contact mIoU/IgAcc/R_c	Bbox gIoU/cIoU/R_c	Seg. F_1/S_α/R_c	Grasp mIoU/IgAcc/R_c	Contact mIoU/IgAcc/R_c
Qwen-VL-Max [13]	**92.3/93.2/63.3**	**87.8/44.8/71.5**	**16.0/16.7/14.2**	**34.7/25.6/**15.4	**89.4/90.8/64.3**	**86.8/43.5/66.4**	**14.9/7.1/**17.9	**32.0/7.0/**18.3	**88.4/88.8/67.7**	**47.3/30.6/64.0**	**8.1/5.4/**22.6	**18.8/20.0/**18.3
Gemini2.5-Flash [21]	60.1/61.5/**83.4**	3.1/4.1/**83.4**	3.2/1.2/**100**	5.2/3.6/**100**	59.2/60.5/**80.9**	3.8/13.1/**80.9**	2.5/2.2/**100**	6.3/4.7/**100**	62.6/63.7/**95.1**	3.9/14.9/**95.1**	3.4/1.9/**100**	4.0/4.3/**100**
Qwen2.5VL-3B [4]	50.0/51.9/89.2	75.1/43.5/99.2	4.3/2.9/94.1	25.2/17.7/87.0	51.2/52.4/94.5	75.1/42.4/94.5	4.0/1.4/92.3	27.8/22.9/85.5	52.9/53.7/96.9	45.9/38.7/97.0	3.1/1.3/95.7	14.5/8.6/98.8
Qwen2.5VL-3B + SFT	56.4/57.1/100	78.2/43.4/100	3.4/1.7/100	23.8/21.4/100	54.7/55.9/100	75.4/43.2/100	3.6/2.1/100	29.3/28.2/100	57.2/58.6/100	46.2/38.9/100	4.4/1.5/100	19.7/15.8/100
RealVLG-R1-3B (GRPO)	87.2/87.6/96.1	87.4/45.2/96.0	**34.7/40.3/100**	**51.9/33.2/100**	**86.5/87.1/97.9**	**85.6/43.9/97.9**	**28.2/31.9/100**	**47.8/31.9/100**	**78.5/79.8/98.2**	**58.3/39.4/98.2**	**16.3/17.1/100**	**26.9/20.2/100**
RealVLG-R1-3B (GSPO)	**87.7/88.4/100**	**88.7/45.3/100**	29.2/33.6/100	48.6/29.6/100	84.6/85.3/**100**	85.0/43.7/**100**	27.6/30.6/100	46.5/28.1/100	78.0/79.3/**100**	49.1/**39.6/100**	15.3/9.1/100	26.9/17.7/100
Qwen2.5VL-7B [4]	56.8/58.3/100	80.6/43.8/100	4.1/1.8/43.5	33.7/26.3/99.2	57.5/57.9/99.6	78.0/42.4/99.6	3.2/0.6/49.4	32.8/23.5/99.6	55.6/56.8/100	48.3/39.2/100	1.6/0.0/64.6	20.0/12.8/100
Qwen2.5VL-7B + SFT	64.8/63.5/100	82.2/44.1/100	5.1/3.1/80.6	38.7/31.2/100	66.7/62.7/100	79.6/42.6/100	4.6/2.7/85.4	37.4/29.6/100	64.9/65.2/100	48.5/39.4/100	3.5/1.4/78.1	26.4/16.4/100
RealVLG-R1-7B (GRPO)	88.0/88.5/100	**89.2/45.3**/100	22.2/**39.3**/99.6	53.3/33.1/99.2	83.4/84.2/100	85.0/43.8/100	**28.8/36.3**/99.6	50.5/34.9/97.4	84.6/85.4/100	**52.2/39.7/100**	**17.0/19.1/**99.8	**30.1/**20.7/100
RealVLG-R1-7B (GSPO)	**89.0/89.6/100**	88.9/**45.4**/100	20.6/32.8/100	**55.3/37.2/100**	**86.4/86.9/100**	**86.6/43.8/100**	27.9/39.2/**100**	**50.6/36.2/100**	**85.5/85.9/100**	52.0/39.7/100	16.5/18.3/**100**	28.9/**25.0/100**

Data Quality

Real-World Vision-Language Grasping



Conclusion & Future Work

- Unified Real-World VLG/Grasping framework
- Zero-shot deployment in clutter

→ **Future:** 3D VLG-to-Action, SmolVLM for realtime

Webpage **Code** **arXiv** **Video**

Contact
- Email: {cslinfeili, cslinzhang, yingshen}@tongji.edu.cn
- Corresponding author: Lin Zhang
- Group Website: https://cslinzhang.github.io/home/