GLOVER++: Unleashing the Potential of Affordance Learning from Human Behaviors for Robotic Manipulation

Teli Ma[1,†], **Jia Zheng**[1,†], **Zifan Wang**[1], **Ziyao Gao**[1], **Jiaming Zhou**[1], **Junwei Liang**[1,2,3,*]
[1] HKUST (GZ) [2] Spatialtemporal AI [3] HKUST
[†] Equal contribution. [*] Corresponding author.

Abstract: Learning manipulation skills from human demonstration videos offers a promising path toward generalizable and interpretable robotic intelligence—particularly through the lens of *actionable affordances*. However, transferring such knowledge remains challenging due to: 1) a lack of large-scale datasets with precise affordance annotations, and 2) insufficient exploration of affordances in diverse manipulation contexts. To address these gaps, we introduce **HOVA-500K**, a large-scale, affordance-annotated dataset comprising 500,000 images across 1,726 object categories and 675 actions. We also release a standardized benchmarking suite for multi-modal affordance reasoning. Built upon HOVA-500K, we present **GLOVER++**, a *global-to-local* affordance training framework that effectively transfers actionable affordance knowledge from human demonstrations to downstream open-vocabulary reasoning tasks. GLOVER++ achieves state-of-the-art results on the HOVA-500K benchmark and demonstrates strong generalization across diverse downstream robotic manipulation tasks. By explicitly modeling actionable affordances, GLOVER++ facilitates robust transfer across scenes, modalities, and tasks. We hope that HOVA-500K and the GLOVER++ framework will serve as valuable resources for bridging the gap between human demonstrations and robotic manipulation capabilities. Code, dataset and models are available at teleema.github.io/projects/GLOVER++.

Keywords: Actionable Affordance, Affordance Transfer, Vision-Language Model, Human Demonstrations, Robotic Manipulation

1 Introduction

Humans can naturally manipulate objects by following language instructions—distinguishing object types, locating them, and choosing affordable parts based on the task in a generalizable way. Hence, images and videos that depict human-object interaction and manipulation are prevalent [1, 2, 3], as such scenarios are highly common in our daily life and easily collectible. What can these data do for robotic manipulation? An intrinsic idea is absorbing the potential knowledge embodied in daily human behaviors and transferring it to facilitate robotic manipulation. However, how such knowledge can be learned and transferred remains unclear.

Some previous works [4, 5, 6, 7] focus on the policy of pretraining in human videos and finetuning in downstream robotic tasks, which reveals limited generalizability and lacks robustness to scene changes. Instead, recent works have paid attention to much more explicit and generalizable representations like **affordance** [8, 9, 10, 11], which refers to relational properties, or potentials for interaction between the environment and the animal as introduced by James J. Gibson [12]. The affordance embodies actionable human knowledge, reflecting the possibility of *where* and *how* to act. Previous affordance-based methods can be mainly categorized as 3D radiance field modeling [13, 14, 15, 16],

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

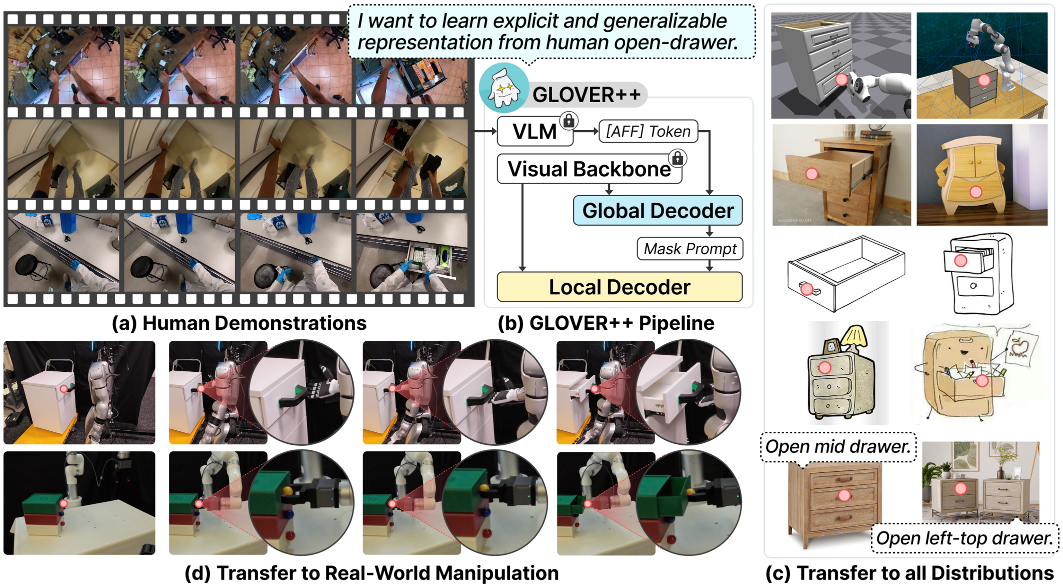


Figure 1: (**a**) GLOVER++ aims to learn generalizable affordance representation from human behaviors (*e.g. open drawer*). (**b**) The training pipeline of GLOVER++. We adopt a *global-to-local* decoding policy to balance global semantic decoding and local affordance decoding. (**c**) GLOVER++ is capable of transferring affordable knowledge to all kinds of distributions (*simulation, sketch, cartoon etc.*) in an open-vocabulary manner. It also presents strong spatial reasoning ability as shown in the bottom line. (**d**) By lifting inferred affordable points into 3D space, GLOVER++ provides perceptive awareness for real-world manipulation tasks. (Red dots represent affordable points.)

object retrieval [9, 8], and vision-language model (VLM) reasoning [10, 11]. However, existing methods have yet to adequately address how to distill actionable affordance knowledge from rich human videos, and how to demonstrate the effective transfer as an explicit representation for a variety of manipulation tasks.

To that end, we introduce **HOVA-500K**, a large-scale affordance-annotated dataset constructed from existing human videos and images. The HOVA-500K comprises 500,000 meticulously annotated images spanning 1,726 object categories and 675 action categories, creating a comprehensive taxonomy of human-object interactions. HOVA-500K offers three key advantages: 1) its unprecedented scale in terms of images and object/action categories enables large-scale affordance training, 2) the diversity of scenarios and views ensures broad coverage of real-world interaction contexts, 3) precise annotation of affordable points eliminates the ambiguity of the mask/region boundary and aligns better with robotic execution. Additionally, we provide a benchmarking evaluation set for standardized comparisons in multi-modal affordance reasoning.

Based on HOVA-500K, we present **GLOVER++**, an end-to-end framework that explores distilling actionable affordance knowledge from raw human demonstrations for multiple robotic tasks as Fig. 1 shows. To achieve the balance between global semantic perception and local affordance learning, a *global-to-local* affordance tuning policy is proposed to incorporate affordance reasoning capabilities while preserving the semantic understanding of VLM. We aim to unleash the potential of the affordance representation and push the boundaries of affordance transfer in diverse manipulation tasks. Extensive experiments in both simulation and the real world are conducted to demonstrate the effectiveness of GLOVER++, including functional zero-shot manipulation, multi-task imitation learning, and serving as an effective perception module for long-horizon and bimanual manipulation.

Our contributions are summarized as follows: **1**) We contribute a large-scale affordance-annotated dataset—HOVA-500K, that provides the necessary scale and diversity to learn generalizable affordance representations. **2**) We present GLOVER++, a global-to-local paradigm of affordance training policy based on HOVA-500K, showing fine-grained affordance representation and generalizable affordance reasoning capability. GLOVER++ achieves state-of-the-art performance in the HOVA-500K evaluation benchmark. **3**) Extensive applications in tasks like zero-shot manipulation, multi-task imitation learning, long-horizon and bimanual manipulation demonstrate the huge potential of HOVA-500K and GLOVER++.

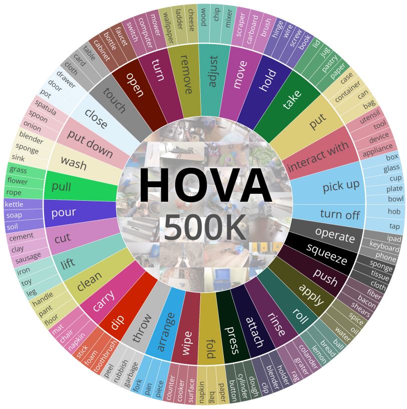


Figure 2: The distribution of primary action categories (>1,000 samples) and related objects in HOVA-500K.

Dataset	Img	Obj	Act	Format	Ego.	Exo.	Ann.
UMD [17]	30K	17	7	RGBD	✗	✓	Part
Sawatzky [18]	3K	17	7	RGBD	✗	✓	Part
AGD20k [19]	26K	50	36	RGB	✓	✓	Part
HANDAL [20]	200K	17	-	RGBD	✗	✓	Obj/Part
OPRA [21]	11K	-	7	RGB	✓	✓	Point
AED [22]	0.7K	13	8	RGB	✗	✓	Part
3DOI [23]	10K	-	-	RGB	✓	✓	Point
PAD [24]	4K	72	31	RGB	✗	✓	Obj
IIT-AFF [25]	8.8K	10	9	RGBD	✗	✓	Part
ADE-Aff [26]	10K	150	7	RGB	✓	✗	Scene
HOVA-500K	500K	1726	675	RGB	✓	✓	Point

Table 1: Comparisons between HOVA-500K and previous datasets. “**Format**”, “**Ann.**”, “**Ego.**” and “**Exo.**” refer to the image format, egocentric, exocentric, and annotation type, respectively. Our HOVA-500K annotates the action & object categories, and the affordance with more precise affordable points.

2 Related Works

Affordance Reasoning. Prior approaches to affordance inference can be categorized into three paradigms: (1) human-object interaction analysis [27, 28, 29], (2) scene understanding through geometric and semantic cues [14, 30, 16], and (3) 3D point cloud grounding [31, 32, 33, 34]. To achieve open-vocabulary affordance reasoning, recent advances have incorporated foundation models (LLMs/VLMs) to the model design [35, 10, 8, 36, 37, 38]. This integration approach with LLMs/VLMs comprises two primary types. The first [8, 9] constructs a memory of object affordances and reasons about the affordances of objects in novel scenes by retrieving from the affordance memory with the help of CLIP [39]. On the other hand, methods like AffordanceLLM [10] and GLOVER [11] fine-tune large VLMs [40] on affordance datasets, leveraging the world knowledge and reasoning capability of foundation models. However, these approaches are constrained by limited data availability and insufficient exploration of the VLM-based affordance fine-tuning mechanism. We contribute the HOVA-500K dataset to mitigate data scarcity, while providing an effective fine-tuning framework template, GLOVER++, to leverage such large-scale data.

Language-guided Zero-shot Manipulation. Integrating linguistic modalities into robotic manipulation tasks serves as a crucial approach for achieving zero-shot manipulation capabilities [15, 41, 42, 43, 14]. Currently, the related methods can be categorized into multiple types like visual-language-action (VLA) pretraining in robotic data [44, 45, 46, 47, 48], invoking LLMs/VLMs for planning or in-context learning [49, 50, 15, 41], and using VLMs for object/scene representations [16, 51, 30, 14, 52]. We follow the manner of leveraging VLMs for providing semantic scene understanding for the downstream manipulation tasks in this work. Different from the methods that distill features from 2D foundation models for building 3D feature fields [14, 52] via neural rendering [53, 54], we adopt a visual-linguistic affordance representation and project the inferred affordance into 3D space without requiring full reconstruction.

3 HOVA-500K Benchmark

Data Collection. HOVA-500K is primarily derived from three key sources: (1) human demonstration videos, which provide real-world interaction sequences for natural and diverse affordance learning (Ego4D [3], EPIC-KITCHEN [55]); (2) object-part segmentation masks, offering structural mask annotations to bridge semantic parts with actionable regions (HANDAL [20]); and (3) existing affordance datasets with labels of human-object affordable point (3DOI [23]). These datasets cover a broad range of scenarios, including both in-the-wild and household environments from ego/exocentric views. This intentional diversity ensures robust generalization across different spatial relationships and interaction modalities.

Affordable Points Annotation. Unlike previous affordance datasets that annotate object regions as segmentation masks [20], we reformulates affordance learning as a dense keypoint prediction task, where the model predicts a single maximum-probability interaction point per object based on the functionality. This shift from region-level to point-level representation offers two advantages: (1) it eliminates the ambiguity of mask boundaries in precision-sensitive tasks, and (2) it better aligns with robotic control, where end-effector contact requires millimeter-level accuracy.

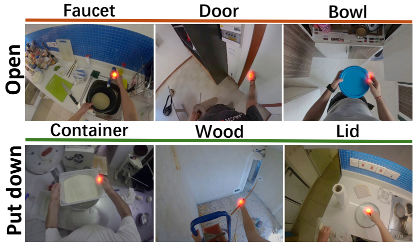


Figure 3: Some examples of HOVA-500K, showing action, object category, and Gaussian-distributed mask of affordable point.

Locating Affordable Points in Human Videos. Our approach builds upon the OCT model [56], applying skin segmentation [57] in the overlapping region between hand and object bounding boxes to obtain contact points. We then compute homography matrices between adjacent frames via RANSAC, enabling us to project contact points from the interaction frame back to the initial frame. This step effectively eliminates occlusion caused by hands. This semi-automatic pipeline enables efficient and scalable annotation of affordable points. The specific introduction and pipeline is shown in Sec. A.1 and Fig. 8.

Category & Action. For the existing uni-modal visual affordance datasets, we implement a semi-automated labeling pipeline using Qwen-2.5-VL-7B [58] VLM to generate object categories and actions. These automatically generated labels are subsequently verified by human annotators to eliminate clearly incorrect entries. Representative actions and object categories are visualized in Fig.2, and more details can be found in Fig. 9, 10, 11.

Benchmarking Testing Set. To comprehensively evaluate affordance prediction models, we construct a diverse test set by selecting 6,000 images from HOVA-500K. Our evaluation framework measures both prediction accuracy and functional realism. The metrics we used for evaluation include Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS). Moreover, we introduce SIM_{part}, a new metric that quantifies the practical plausibility of predicted affordance regions in real-world settings. Detailed metric descriptions appear in Sec. A.2.

4 Method

In this section, we elaborate on the training policy for distilling actionable affordance knowledge from human behaviors and aligning with human instructions. First, we briefly describe the task and preliminary (Sec. 4.1). Then, we introduce the global-to-local affordance fine-tuning of GLOVER++ in Sec. 4.2. The potential application of GLOVER++ is elaborated in Sec. 4.3.

4.1 Task Description and Preliminaries

GLOVER++ aims to predict executable affordable points $\mathcal{P}^{2D}$ in an open-vocabulary and end-to-end manner. Given the input $\mathcal{I} = (\mathcal{I}_R, \mathcal{T})$, where $\mathcal{I}_R, \mathcal{T}$ represent an RGB image and language instructions, we expect the model to generate $\mathcal{P}^{2D}$. To achieve this, we first convert ground-truth affordable points $\mathcal{P}^{2D}_{gt}$ into Gaussian-distributed heatmaps $\mathcal{M}^{2D}_{gt}$ that centered at $\mathcal{P}^{2D}_{gt}$ to ensure gradient continuity during training like annotations in Fig. 3. This translates discrete annotations into continuous optimization targets as shown in Fig. 3. The resulting $\mathcal{M}^{2D}_{gt}$ supervises the model to generate affordance mask $\mathcal{M}^{2D}_{\mathcal{A}}$, where each pixel value $\mathcal{A}^{2D}$ in $\mathcal{M}^{2D}_{\mathcal{A}}$ represents the affordable probability of the current position. We obtain $\mathcal{P}^{2D}$ by selecting the pixel with the highest probability:

$\mathcal{P}^{2D} = \underset{\mathcal{P}\in\mathcal{I}_R}{\arg\max}\mathcal{M}_{\mathcal{A}}^{2D}$. This point is then projected into 3D space via camera intrinsics to yield $\mathcal{P}^{3D}$ for robotic execution.

4.2 Global-to-Local Affordance Tuning

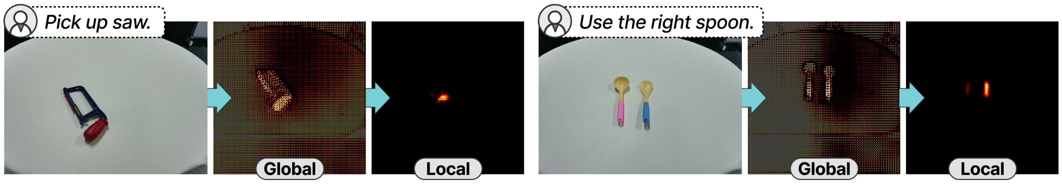


Figure 4: **Visualization of the decoded features by the global and local decoder** (the intensity of highlight scale with interest of regions). We can observe that the integration of the local decoder effectively eliminates the background noise from the global decoding.

We illustrate the global-to-local pipeline in Fig. 1 & 12. The following provides detailed descriptions of the components involved.

Multi-modal Encoding. To empower GLOVER++ with world knowledge and reasoning capability, we leverage a VLM to encode the multi-modal inputs $\mathcal{I}$ into a hidden latent token for affordance reasoning. Following the *Embedding-as-Mask* paradigm in LISA [59], we add a new affordance token `<AFF>` to encode combined visual and linguistic features with LLaVA-1.5 [40]. Given the language instruction $\mathcal{T}$ and RGB image $\mathcal{I}_R$, we feed them into the LLaVA model $\mathbf{F}_{LLaVA}$ to generate the responsive hidden latents $\boldsymbol{r}$, from where we detach the latent `<AFF>` token:

$$\texttt{<AFF>} \in \boldsymbol{r} = \mathbf{F}_{LLaVA}(\mathcal{I}_R, \mathcal{T}). \tag{1}$$

Global-to-Local Decoding. The `<AFF>` token aggregates **global** semantic context from $\mathcal{I}_{\mathcal{R}}$, while the affordance prediction requires **local** fine-grained reasoning. Hence, the core challenge lies in *balancing global semantic perception and precise local affordance representation learning*. To this end, we decompose the decoding process of `<AFF>` token into two stages: **global decoding** and **local decoding**. In the first stage, the `<AFF>` token guides global semantic decoding to generate a high-level semantic logits map that captures global contextual relationships. In the second stage, we refine the prediction through localized decoding: the semantic map $\mathcal{M}_{sem}^{2D}$ acts as a mask prompt to condition attention on relevant regions. This enables accurate region-specific affordance prediction:

$$\mathcal{M}_{sem}^{2D} = \mathbf{F}_{dec}^{glo}(\texttt{<AFF>}, \boldsymbol{v}), \quad \mathcal{M}_{\mathcal{A}}^{2D} = \mathbf{F}_{dec}^{loc}(\mathcal{M}_{sem}^{2D}, \boldsymbol{v}), \tag{2}$$

where $\boldsymbol{v}$ denotes the visual features from the vision backbone. The effectiveness of the global-to-local decoding mechanism is visually demonstrated in Fig. 4, and implementation details are specified in Sec. B.1.

Training Objective. Besides the sigmoid focal loss [60] used in GLOVER [11], we introduce an additional Kullback-Leibler Divergence (KLD) loss to constrain the predicted affordance distribution. This KLD term aligns the predicted heatmap $\mathcal{M}_{\mathcal{A}}^{2D}$ with the Gaussian-distributed ground truth $\mathcal{M}_{gt}^{2D}$, encouraging distributional consistency. The overall training objective becomes:

$$\mathcal{L} = \mathcal{L}_{FL}(\mathcal{M}_{\mathcal{A}}^{2D}, \mathcal{M}_{gt}^{2D}) + \mathcal{L}_{KL}(\mathcal{M}_{\mathcal{A}}^{2D}, \mathcal{M}_{gt}^{2D}), \tag{3}$$

More details about the training objective are specified in Sec. B.2.

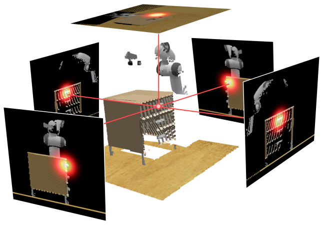


Figure 5: Explicit affordance representation for imitation learning in RLBench [61].

4.3 Unleash the Potential of Affordance Representation

Zero-shot Manipulation. GLOVER++ is capable of reasoning about affordance in an open-vocabulary way to acquire 3D graspable points, which inherently addresses zero-shot grasping challenges. The inferred affordable points $\mathcal{A}^{3D}$ can be combined with all kinds of pose estimators (*e.g.* GraspNet [62], AnyGrasp [63], FoundationPose [64]) or geometric-constraints estimation (*e.g.* superquadric recovery [65, 66, 11]) to generate the grasping pose $(\mathcal{A}^{3D}, \tau^{3D})$, where τ^{3D} is the rotation of the end-effector. For motion planning, we use Inverse Kinematics (IK) by default to reach $\mathcal{A}^{3D}$.

Imitation Learning. Instead of the previous *pretraining-finetuning* paradigm that transfers pretrained weights in human videos to downstream imitation-learning tasks [5, 6, 7], we adopt the reasoned affordance representation as an explicit knowledge prior to dynamically modulate attention weights, enabling the model to focus on task-relevant regions, as shown in Fig. 5. This representation serves as a structured guidance signal, making the learning process more interpretable and effective compared to implicit methods. The pipeline is shown in Fig. 14, and training details are specified in Sec. C.2.

5 Experiments

We evaluate the performance of GLOVER++ from four perspectives: vision-language affordance reasoning (Sec. 5.1), zero-shot manipulation (Sec. 5.2), imitation learning (Sec. 5.3), and extended capabilities (Sec. 5.4), including long-horizon manipulation with VLM planner and bimanual manipulation.

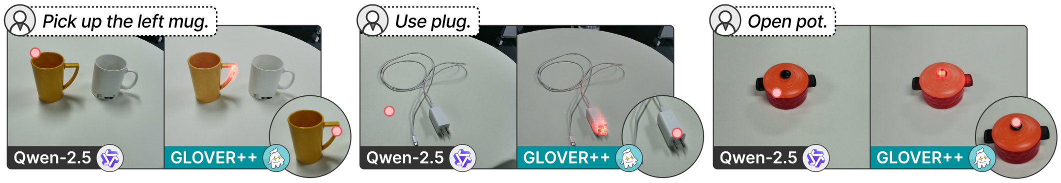


Figure 6: **Comparison with Qwen-2.5-VL**, GLOVER++ generates more physically plausible and functionally grounded prediction results, aligning better with real-world interaction constraints.

5.1 Vision-Language Affordance Reasoning

Training Details. GLOVER++ is trained on 8 NVIDIA A6000 GPUs for 10 epochs with a batch size of 32 per GPU. We employ AdamW [67] optimizer with a weight decay of 5e-4. The learning rates for the global decoder and local decoder are set to 5e-5 and 5e-4 to achieve a balance between preserving open-vocabulary knowledge and affordance learning. For more details, please refer to Sec. C.1.

Results. Table 2 shows the quantitative results of affordance reasoning on the HOVA-500K benchmark. The metrics specified in the benchmarking testing set (Sec. 3) are adopted to evaluate. We compare GLOVER++ with three methods, 3DOI [23], AffordanceLLM [10], and GLOVER [11]. The three methods are all based on affordance pretraining. Specifically, 3DOI relies solely on visual inputs. In contrast, both AffordanceLLM and GLVOER are pretrained models that integrate visual-language information for affordance learning. GLOVER++ significantly outperforms all baselines across all metrics, owing to its global-to-local decoding scheme and KLD-based optimization, which jointly enhance both affordance center prediction and distributional alignment.

Methods	KLD ↓	SIM ↑	SIM$_{part}$ ↑	NSS ↑
3DOI [23]	5.978	0.007	0.006	-0.311
AffordanceLLM [10]	5.041	0.018	0.161	1.665
GLOVER [11]	4.874	0.016	0.254	2.876
GLOVER++	**3.411**	**0.141**	**0.563**	**5.296**
		Ablations		
w/o global-to-local	3.465	0.101	0.483	4.925
w/o KLD	4.307	0.030	0.409	4.005
10→5 epoch	3.615	0.121	0.533	5.197

Table 2: Affordance reasoning results and ablations in the benchmarking dataset.

We also benchmark against Qwen-2.5-VL-7B [58], a powerful VLM with strong spatial understanding via Rotary Positional Embedding (RoPE). As shown in Fig. 6, while Qwen-2.5-VL exhibits decent localization, GLOVER++ provides affordance predictions that are more physically plausible and functionally grounded, despite using a comparable model size (∼7B).

Ablations. To assess the contribution of key components, we evaluate ablations by removing the global-to-local module, the KLD loss, and varying the training length (Table 2). Notably, removing global-to-local decoding or KLD optimization results in a 28.4% and 78.7% drop in the SIM metric, respectively, underscoring their importance. We also show the visualization of reasoned affordance in Fig. 16, where the effectiveness of model components is evident. Additional ablations are provided in Sec. C.3.

5.2 Zero-shot Manipulation

Setup. We perform extensive experiments in both the simulated and real-world settings. For simulation, we use IsaacGym [68] with GAPartNet [69] object sets, involving a 7-DoF Franka Panda arm for zero-shot manipulation. 50 objects across 5 categories (*Box, Pot, Drawer, TrashCan, Cabinet*) are used. Actions are listed in Table 3, with success defined as articulation joint or height exceeding a predefined threshold. Each task is tested 25 times with varied initial poses. For real-world experiments, we deploy a 7-DoF UFactory xArm, evaluated over five trials per task using RGB-D input from an Orbbec Femto Bolt (1280 × 960 resolution). Object categories and actions are provided in Table 4. Note that the "*Press Button*" task involves discriminating the correct color one and pressing it.

Object								**AVG**
Task	O	C	O	C	O	P	O	/
VRB [70]	4	60	4	56	20	24	16	26.3
Robo-ABC [8]	28	44	32	32	20	32	16	29.1
RAM [9]	36	**64**	40	60	24	36	32	41.7
GLOVER++	**40**	60	**40**	**68**	**32**	**44**	**44**	**46.9**

Table 3: Success rates of different methods in IsaacGym. O, C, P represent *Open, Close, Pickup*, respectively.

Baselines & Results. We compare GLOVER++ with three baselines. VRB [70] predicts contact points by learning from human demonstrations. Both Robo-ABC [8] and RAM [9] retrieve affordance from the pre-built memory and transfer it to new scenes via estimated similarity of CLIP [39]. For simulation and real-world experiments, we use the success rate (SR) as metric. Table 3 and 4 show the simulated and real-world results, respectively. We can see that GLOVER++ achieves favorable performance among the baseline methods in both the IsaacGym and real-world environments, yielding an average success of 46.9% and 73.3%, respectively. Compared to GLOVER++, RAM is limited in its ability to distinguish object properties (such as buttons of different colors) due to the retrieval mechanism based on object-level similarity. In contrast, GLOVER++ generalizes to novel objects and scenarios by directly grounding actionable affordance points from language instructions alone.

Object							**AVG**
Task	O	P	O	P	PR	P	/
RAM [9]	3/5	1/5	2/5	3/5	1/5	4/5	46.7
GLOVER [11]	3/5	2/5	**3/5**	3/5	3/5	4/5	60.0
GLOVER++	**4/5**	**4/5**	2/5	**4/5**	**4/5**	4/5	**73.3**

Table 4: Success rates of different methods in real-world experiments. O, P, PR means *Open, Pickup, Press*, respectively.

5.3 Imitation Learning

Setup. We validate affordance knowledge transfer in language-guided multi-task imitation learning using RLBench [61]. RLBench is a robot manipulation benchmark built on CoppelaSim [71] and PyRep [72]. We follow the protocols of PerAct [73] to test the model on 18 tasks in RLBench by controlling a Franka Panda robot with a parallel gripper. Each policy is trained on 100 demonstrations using RGB-D observations from four views (front, left shoulder, right shoulder, wrist) at 128 × 128 resolution. Tasks are tested 25 times per trial. Additional details are in Sec. C.2.

Baselines & Results. We compare with two imitation-learning baselines. RVT [74] utilizes a multi-view Transformer model to extract visual features from multi-view images rendered based

Models	put in drawer	drag stick	turn tap	slide block	open drawer	put in cupboard	sort shape	put in safe	push buttons	close jar	stack blocks	place wine	sweep to dustpan	meat off grill	screw bulb	place cups	insert peg	stack cups	Averaged Success Rate
RVT	92.0	100.0	**100.0**	76.0	76.0	52.0	40.0	84.0	96.0	92.0	24.0	88.0	64.0	88.0	48.0	0	0	0	62.2
RVT-AFF	92.0	100.0	96.0	64.0	76.0	**76.0**	40.0	**88.0**	96.0	**96.0**	**32.0**	**92.0**	**88.0**	**96.0**	**56.0**	**4.0**	**8.0**	**8.0**	67.1(+4.9)
RVT2	**96.0**	100.0	100.0	88.0	68.0	68.0	32.0	96.0	100.0	100.0	**80.0**	92.0	100.0	96.0	88.0	36.0	40.0	72.0	80.7
RVT2-AFF	92.0	100.0	100.0	**92.0**	**72.0**	68.0	**36.0**	**100.0**	100.0	100.0	76.0	**100.0**	100.0	96.0	**92.0**	**44.0**	**52.0**	**80.0**	83.3(+2.6)

Table 5: **Success rate of 18 tasks in RLBench**. With explicit affordance representation, both RVT [74] and RVT-2 [75] show improvements in the multiple imitation-learning tasks.

on point clouds, and predicting end-effector poses via deep-learning-based Q-function estimation. RVT-2 [75] learns more precise manipulations via zooming into the region of interest.

As shown in Table 5, GLOVER++ improves RVT and RVT-2 by +4.9% and +2.6%, respectively. Gains are particularly evident in tasks requiring precise control (*e.g.*, `insert peg`, `stack cups`, blue highlight in Table 5). The affordance representation constrains attention to task-relevant spatial and semantic features, improving policy effectiveness. The experiments demonstrate that the explicit representation like affordance is capable of transferring knowledge learned from human demonstrations to enhance robotic imitation learning performance.

5.4 Extended Capabilities

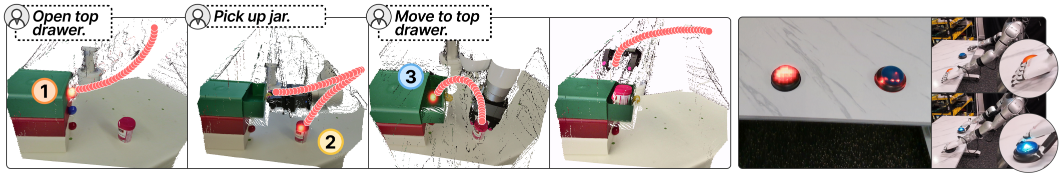


Figure 7: **Left:** GLOVER++ serves as a perceptual module for the VLM planner to complete long-horizon tasks. **Right:** GLOVER++ enables bimanual tasks by reasoning affordances for both left and right hands with spatial relationships.

Long-horizon Manipulation with VLM planner. GLOVER++ can serve as a perceptual backbone for a high-level VLM planner. We integrate it with Qwen-2.5-VL [58], which decomposes long-horizon instructions into subgoals. As shown in Fig.7-left, Qwen-2.5-VL split the task "*Put the jar into the top drawer*" into steps like "*Open top drawer*", "*Pick up jar*", "*Move to top drawer*" *etc.*, and invoking GLOVER++ when affordance grounding is required (①, ②, ③). This hybrid system combines semantic planning and precise affordance prediction, enabling robust multi-stage manipulation. Full flow is shown in Fig.22.

Bimanual Manipulation. Thanks to its spatial reasoning capabilities, GLOVER++ can interpret positional cues (*e.g.*, "left/right", "top/bottom") to enable dual-arm affordance reasoning. It generates graspable regions for both arms while maintaining spatial separation and feasibility (Fig.7, right). We execute dual-arm motions using obstacle-avoidance IK on the Unitree G1 humanoid robot (Fig.19).

6 Conclusion

In this work, we address the critical challenges of transferring actionable affordance knowledge from human demonstrations to robotic tasks by introducing HOVA-500K, a large-scale dataset with precise affordance annotations, and GLOVER++, a global-to-local framework for affordance reasoning. We demonstrate the potential of affordance representation and GLOVER++ in diverse tasks, including zero-shot manipulation, imitation learning, long-horizon and bimanual manipulation. We aim to foster future research in the explicit, interpretable and transferable representation learning for robotic manipulation from human behaviors.

7 Limitations

While HOVA-500K provides a large-scale affordance dataset, its annotations are primarily derived from static images, limiting coverage of dynamic interactions (e.g., tool-use trajectories or force-sensitive affordances). GLOVER++'s reliance on vision-language models may inherit biases from pre-trained VLMs, occasionally leading to over-generalized affordance predictions for novel object-action combinations. Additionally, GLOVER++ relies on imitation learning or an extra VLM planner to complete long-horizon manipulation tasks, lacking the ability to plan the grasping pose and trajectories by itself. Future work to address these issues includes: enlarging the HOVA-500K with annotated trajectories of human behaviors, empowering GLOVER++ with trajectory planning ability via finetuning, and training multiple versions of GLOVER++ based on different large VLMs. We also discuss the **failure cases** in Sec. C.7.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62306257) and the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2023A03J0008), Education Bureau of Guangzhou Municipality.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[3] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[4] X. Lin, J. So, S. Mahalingam, F. Liu, and P. Abbeel. Spawnnet: Learning generalizable visuomotor skills from pre-trained network. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4781–4787. IEEE, 2024.

[5] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[6] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[7] J. Zhou, T. Ma, K.-Y. Lin, Z. Wang, R. Qiu, and J. Liang. Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation. *arXiv preprint arXiv:2406.14235*, 2024.

[8] Y. Ju, K. Hu, G. Zhang, G. Zhang, M. Jiang, and H. Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. *arXiv preprint arXiv:2401.07487*, 2024.

[9] Y. Kuang, J. Ye, H. Geng, J. Mao, C. Deng, L. Guibas, H. Wang, and Y. Wang. Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. *arXiv preprint arXiv:2407.04689*, 2024.

[10] S. Qian, W. Chen, M. Bai, X. Zhou, Z. Tu, and L. E. Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7597, 2024.

[11] T. Ma, Z. Wang, J. Zhou, M. Wang, and J. Liang. Glover: Generalizable open-vocabulary affordance reasoning for task-oriented grasping. *arXiv preprint arXiv:2411.12286*, 2024.

[12] J. Gibson. The theory of affordances. *Perceiving, acting and knowing: Towards an ecological psychology/Erlbaum*, 1977.

[13] Y. Zheng, X. Chen, Y. Zheng, S. Gu, R. Yang, B. Jin, P. Li, C. Zhong, Z. Wang, L. Liu, et al. Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping. *arXiv preprint arXiv:2403.09637*, 2024.

[14] A. Rashid, S. Sharma, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *7th Annual Conference on Robot Learning*, 2023.

[15] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[16] R.-Z. Qiu, Y. Hu, G. Yang, Y. Song, Y. Fu, J. Ye, J. Mu, R. Yang, N. Atanasov, S. Scherer, et al. Learning generalizable feature fields for mobile manipulation. *arXiv preprint arXiv:2403.07563*, 2024.

[17] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015.

[18] J. Sawatzky, A. Srikantha, and J. Gall. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2795–2804, 2017.

[19] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022.

[20] A. Guo, B. Wen, J. Yuan, J. Tremblay, S. Tyree, J. Smith, and S. Birchfield. Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11428–11435. IEEE, 2023.

[21] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018.

[22] G. Li, N. Tsagkas, J. Song, R. Mon-Williams, S. Vijayakumar, K. Shao, and L. Sevilla-Lara. Learning precise affordances from egocentric videos for robotic manipulation. *arXiv preprint arXiv:2408.10123*, 2024.

[23] S. Qian and D. F. Fouhey. Understanding 3d object interaction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21753–21763, 2023.

[24] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao. One-shot affordance detection. *arXiv preprint arXiv:2106.14747*, 2021.

[25] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE, 2017.

[26] C.-Y. Chuang, J. Li, A. Torralba, and S. Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018.

[27] M. Hassan and A. Dharmaratne. Attribute based affordance detection from human-object interaction images. In *Image and Video Technology–PSIVT 2015 Workshops: RV 2015, GPID 2013, VG 2015, EO4AS 2015, MCBMIIA 2015, and VSWS 2015, Auckland, New Zealand, November 23-27, 2015. Revised Selected Papers 7*, pages 220–232. Springer, 2016.

[28] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021.

[29] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022.

[30] G. Lu, S. Zhang, Z. Wang, C. Liu, J. Lu, and Y. Tang. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation. *arXiv preprint arXiv:2403.08321*, 2024.

[31] Y. Geng, B. An, H. Geng, Y. Chen, Y. Yang, and H. Dong. Rlafford: End-to-end affordance learning for robotic manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5880–5886. IEEE, 2023.

[32] C. Ning, R. Wu, H. Lu, K. Mo, and H. Dong. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] R. Wu, K. Cheng, Y. Zhao, C. Ning, G. Zhan, and H. Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[35] D. Hadjivelichkov, S. Zwane, L. Agapito, M. P. Deisenroth, and D. Kanoulas. One-shot transfer of affordance regions? affcorrs! In *Conference on Robot Learning*, pages 550–560. PMLR, 2023.

[36] Y. Song, P. Sun, Y. Ren, Y. Zheng, and Y. Zhang. Learning 6-dof fine-grained grasp detection based on part affordance grounding. *arXiv preprint arXiv:2301.11564*, 2023.

[37] Y. Lu, Y. Fan, B. Deng, F. Liu, Y. Li, and S. Wang. Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 976–983. IEEE, 2023.

[38] K. Xu, S. Zhao, Z. Zhou, Z. Li, H. Pi, Y. Zhu, Y. Wang, and R. Xiong. A joint modeling of vision-language-action for target-oriented grasping in clutter. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11597–11604. IEEE, 2023.

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[40] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[41] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024.

[42] N. Di Palo and E. Johns. Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. *arXiv preprint arXiv:2402.13181*, 2024.

[43] P. Liu, Y. Orru, C. Paxton, N. M. M. Shafiullah, and L. Pinto. Ok-robot: What really matters in integrating open-knowledge models for robotics. *arXiv preprint arXiv:2401.12202*, 2024.

[44] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[45] Q. Vuong, S. Levine, H. R. Walke, K. Pertsch, A. Singh, R. Doshi, C. Xu, J. Luo, L. Tan, D. Shah, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.

[46] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[47] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[48] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, et al. Palm-e: An embodied multimodal language model. 2023.

[49] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[50] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[51] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023.

[52] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*, 2023.

[53] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[54] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[55] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

[56] S. Liu, S. Tripathi, S. Majumdar, and X. Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022.

[57] F. Saxen and A. Al-Hamadi. Color-based skin segmentation: An evaluation of the state of the art. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4467–4471. IEEE, 2014.

[58] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[59] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[60] T. Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[61] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

[62] H.-S. Fang, C. Wang, M. Gou, and C. Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.

[63] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 2023.

[64] B. Wen, W. Yang, J. Kautz, and S. Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024.

[65] A. Leonardis, A. Jaklic, and F. Solina. Superquadrics for segmenting and modeling range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11):1289–1295, 1997.

[66] D. Paschalidou, A. O. Ulusoy, and A. Geiger. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10344–10353, 2019.

[67] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[68] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[69] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, and H. Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023.

[70] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[71] E. Rohmer, S. P. Singh, and M. Freese. V-rep: A versatile and scalable robot simulation framework. In *2013 IEEE/RSJ international conference on intelligent robots and systems*, pages 1321–1326. IEEE, 2013.

[72] S. James, M. Freese, and A. J. Davison. Pyrep: Bringing v-rep to deep robot learning. *arXiv preprint arXiv:1906.11176*, 2019.

[73] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-Actor: A multi-task transformer for robotic manipulation. In *CoRL*, 2022.

[74] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023.

[75] A. Goyal, V. Blukis, J. Xu, Y. Guo, Y.-W. Chao, and D. Fox. Rvt-2: Learning precise manipulation from few demonstrations. *arXiv preprint arXiv:2406.08545*, 2024.

[76] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006.

[77] S. Yang, T. Qu, X. Lai, Z. Tian, B. Peng, S. Liu, and J. Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023.

[78] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[79] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.

A HOVA-500K

A.1 Locate Affordable Points in Human Videos.

Our approach builds upon the OCT model [56]. For a given contact frame $\mathcal{C}$, we initially apply skin segmentation [57] in the overlapping region between hand and object bounding boxes to obtain contact points $P = \{p_1, ..., p_2\}$. Subsequently, we sample 10 preceding frames as observation key frames $\mathcal{O} = \{o_1, ..., o_{10}\}$ (where o_{10} is temporally adjacent to $\mathcal{C}$). We aim to calculate a set of transformation matrices that link all observation frames to $\mathcal{C}$ by computing the homography between consecutive frames. To estimate those homographies, we mask out dynamic elements like detected hands and objects from each frame. Then, we establish feature correspondences in the unmasking regions using SURF descriptor [76]. The homography is computed by applying RANSAC algorithm to the established feature correspondences. Finally, we project the contact points back to the initial frame o_1 based on estimated tomography to acquire affordable points. The pipeline is shown in Fig. 8.

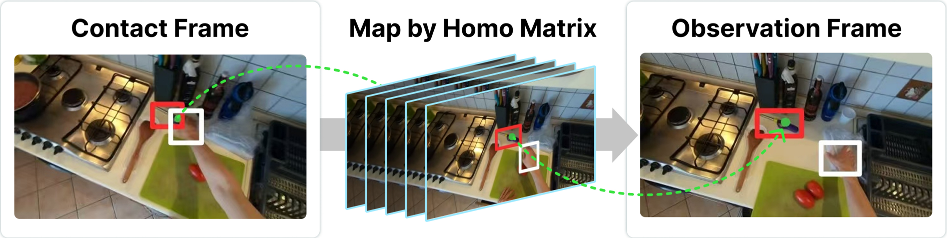


Figure 8: The pipeline of locating affordable points from the human videos following OCT [56]. We initially apply skin segmentation [57] in the overlapping region between hand and object bounding boxes to obtain contact points, and compute the homography between each pair of successive frames. We project the contact points back to the initial frame based on the homography.

A.2 Benchmark Evaluation Metrics

We utilize four metrics to evaluate the models' performance in the HOVA-500K benchmark, including KLD, SIM, NSS, and SIM_{part}. Specifically, these metrics can be formulated as:

$$KLD(\mathcal{M}^{2D}_{\mathcal{A}}, \mathcal{M}^{2D}_{gt}) = \sum_i {\mathcal{M}^{2D}_{gt}}_i \log{(\epsilon + \frac{{\mathcal{M}^{2D}_{gt}}_i}{\epsilon + {\mathcal{M}^{2D}_{\mathcal{A}}}_i})} \tag{4}$$

$$SIM(\mathcal{M}^{2D}_{\mathcal{A}}, \mathcal{M}^{2D}_{gt}) = \sum_i \min({\mathcal{M}^{2D}_{gt}}_i, {\mathcal{M}^{2D}_{\mathcal{A}}}_i) \tag{5}$$

$$NSS(\mathcal{M}_{\mathcal{A}}^{2D}, \mathcal{M}_{gt}^{2D}) = \frac{1}{N} \sum_i \hat{\mathcal{M}}_{\mathcal{A}}^{2D} \times \mathcal{M}_{gt}^{2D} \tag{6}$$

where $N = \sum_i \mathcal{M}_{gt\ i}^{2D}$, $\hat{\mathcal{M}}_{\mathcal{A}}^{2D} = \frac{\mathcal{M}_{\mathcal{A}}^{2D} - \mu(\mathcal{M}_{\mathcal{A}}^{2D})}{\sigma(\mathcal{M}_{\mathcal{A}}^{2D})}$, $\mu(\mathcal{M}_{\mathcal{A}}^{2D})$ and $\sigma(\mathcal{M}_{\mathcal{A}}^{2D})$ are the mean and standard deviation, respectively.

The formula for SIM_{part} is the same as that of SIM, except that the meaning of $\mathcal{M}_{gt}^{2D}$ differs. For an object with a handle, its graspable part is actually the entire handle. Thus, any point falling within the handle region after affordance argmax should be considered valid. Therefore, in SIM_{part}, we use the binary mask of the handle region as the ground truth for calculation.

A.3 Dataset Taxonomy

Our dataset consists of 1,726 object categories and 675 verb categories. The object categories can be referenced in the right panel of Figure 9, with examples including *hammer, screwdriver, adjustable wrenches, combinational wrenches, spatula, etc*. Some verb categories are shown in the left panel of Figure 9, with examples such as *pick up, interact with, put, take, hold, move, etc*. Figures 10 and 11 display the logarithmic frequency distribution histograms for object categories with more than 1,000 instances and verb categories with more than 100 instances, respectively.

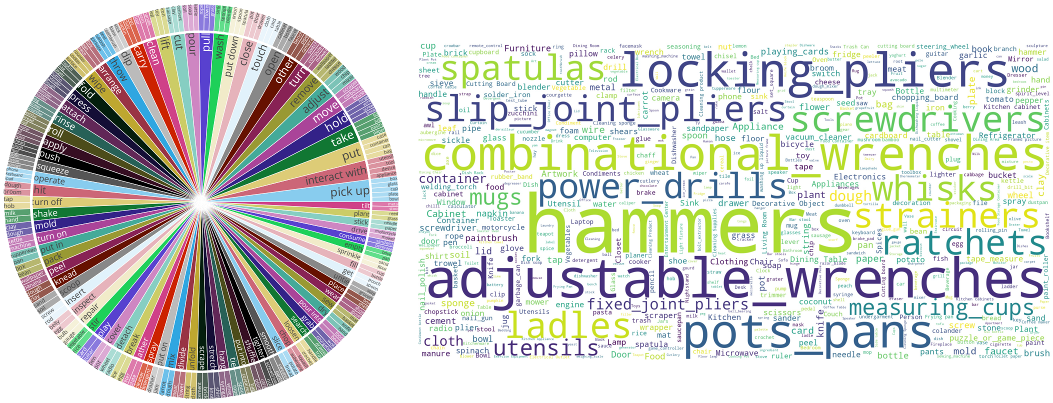


Figure 9: **Left:** The distribution of primary actions and related object categories in HOVA-500K (>100 data samples). **Right:** Word cloud of primary object categories in HOVA-500K.

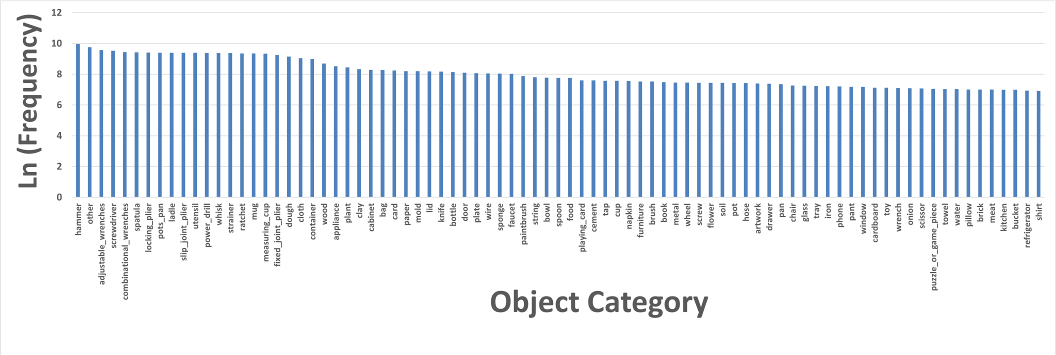


Figure 10: The distribution of primary object categories in HOVA-500K (>1000 data samples)

A.4 Manual Verification

For affordance point rationality, we randomly sampled 100 images and had three external individuals rate annotations on a 1-5 scale (1: highly unreasonable, 5: highly reasonable). **The average score was 4.897**, demonstrating high quality.

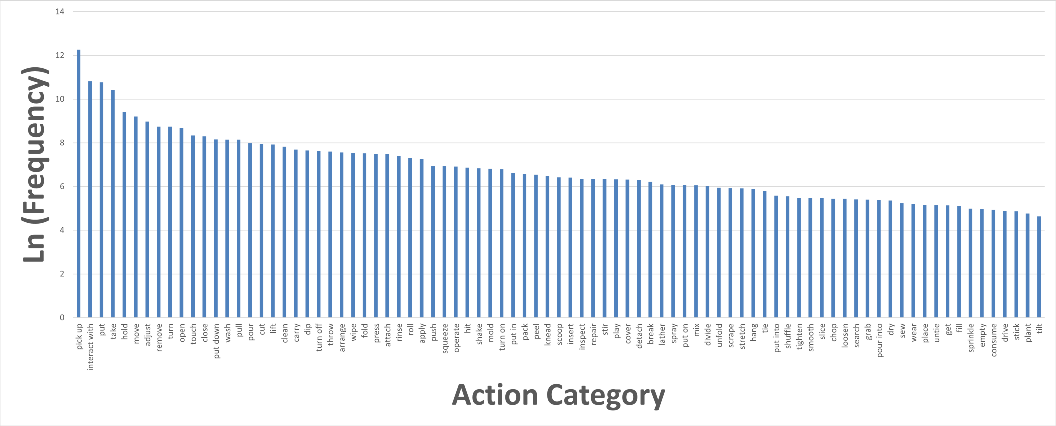


Figure 11: The distribution of primary action categories in HOVA-500K (>100 data samples)

B GLOVER++ Model

B.1 Model Pipeline

We show the detailed pipeline of GLOVER++ in Fig. 12. We utilize the pretrained vision backbone and LLaVA model from the LISA++ [77] to inherit the open-vocabulary reasoning knowledge from it. The vision backbone and LLaVA model are frozen during the training. We follow the SAM [78] to design the decoders, and both the global and local decoder comprises two layers of bi-directional Transformer layers.

The first global decoder processes the <AFF> token to generate a semantic-aware logits map, where the token aggregates vision-language features from the input to encode high-level image semantics. While this global understanding captures contextual relationships, it inevitably introduces background noise—irrelevant regions activated by broad semantic correlations (e.g., ”cut” may falsely highlight all sharp objects). Such noise conflicts with the localized nature of affordance learning, which demands precise spatial grounding of action-relevant object parts.

To resolve this discrepancy, we propose a cascaded local decoder that refines the global decoder's output. The local decoder treats the initial semantic logits map as a mask prompt, dynamically focusing on regions where language instructions align with local object geometries. This hierarchical design synergizes global semantic priors with local affordance specificity: the global decoder provides object-level awareness, while the local decoder resolves part-level actionable regions. We show the effectiveness of this global-to-local mechanism in Fig. 13. Also, in this process, we hope to preserve the world knowledge and open-vocabulary reasoning capability of the LLaVA in the first global decoding, thus distributing a relatively small learning rate to update the global decoder. For the local decoder, a large learning rate is set to ensure the thorough learning of the affordance representation.

Compared with GLOVER [11], GLOVER++ introduces negligible additional trainable parameters, as shown in Table 6, while achieving much better affordance reasoning performance.

Methods	**Trainable #param**	**Ratio**	**KLD** ↓	**SIM** ↑	**SIM**$_{part}$ ↑	**NSS** ↑
GLOVER [11]	4.1M	0.0526%	7.441	0.025	0.206	0.900
GLOVER++	8.1M	0.1050%	**3.411**	**0.120**	**0.506**	**5.151**

Table 6: Trainable parameters and their ratio to the total parameters. Our model surpasses GLOVER in performance with only a marginal increase in trainable parameters.

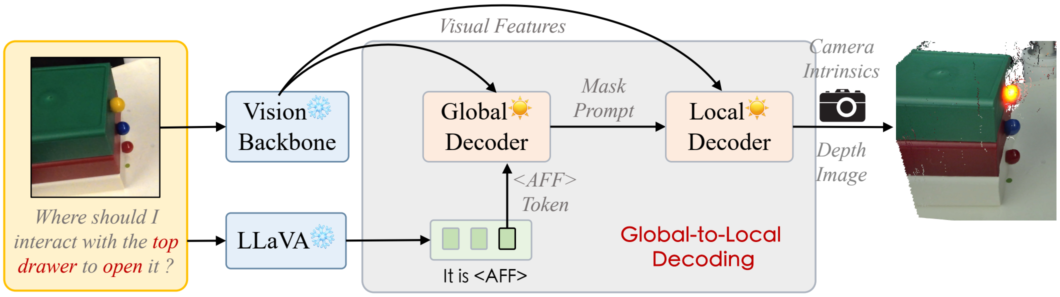


Figure 12: The pipeline of GLOVER++.

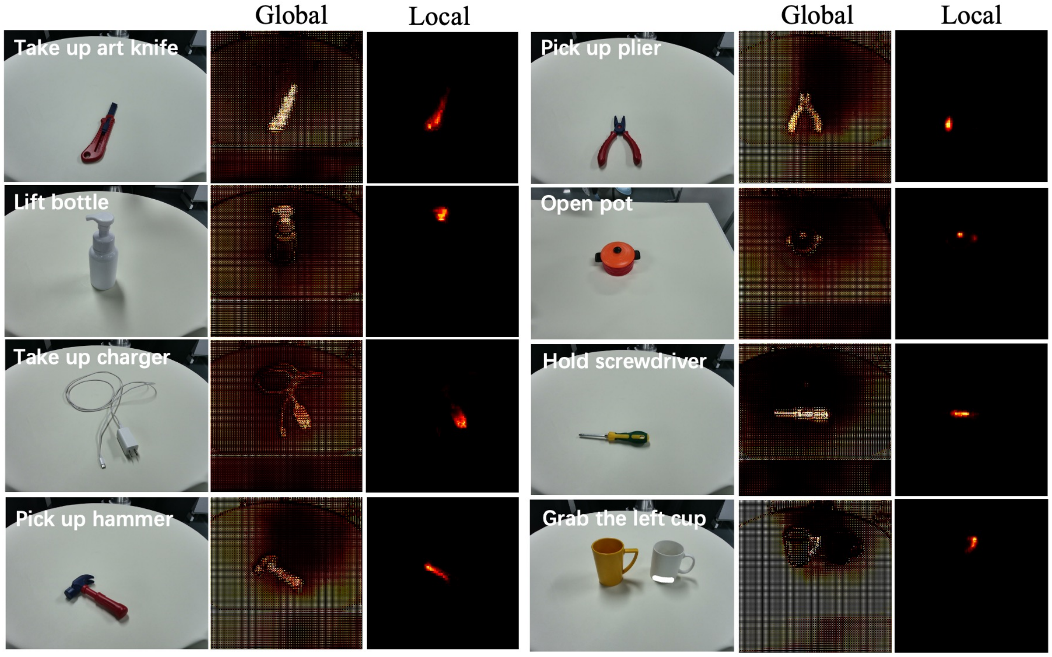


Figure 13: More visualization of the decoded features by the global and local decoder

B.2 Training Objective

The final training objective consists of two parts: the sigmoid focal loss and the Kullback-Leibler Divergence (KLD) loss. Here, α, γ denotes the focusing and balancing parameters in the focal loss, respectively. Additionally, α_t, p_t represent the soft versions of α, p in focal loss formulation. g_i, p_i correspond to the ground truth and predicted affordance value, respectively.

$$\mathcal{L} = -\lambda_{Sigmoid} \sum_i \alpha_t (1 - p_t)^{\gamma} \mathcal{L}_{CE} + \lambda_{KL} \sum_i g_i \log \frac{g_i}{p_i} \tag{7}$$

$$\mathcal{L}_{CE} = -g_i \log p_i \tag{8}$$

$$\alpha_t = \alpha g_i + (1 - \alpha)(1 - g_i) \tag{9}$$

$$p_t = p_i g_i + (1 - p_i)(1 - g_i) \tag{10}$$

C Experiments

C.1 GLOVER++ Training Details.

Our training is conducted on 8 NVIDIA 48G A6000 GPUs for 10 epochs. The training scripts are based on deepspeed engine[79]. We employ AdamW[67] optimizer($\beta_1 = 0.9, \beta_2 = 0.95$) with a weight decay of 0.0005. The learning rates for the mask decoder and affordance decoder are set to 5e-5 and 5e-4, respectively. We use WarmupDecayLR as the learning rate scheduler, with 188 warmup steps. The KL Loss and Focal Loss are both weighted at 0.1. Additionally, the batch size per GPU is configured as 32. The mask decoder is initialized with pretrained weights from LISA++[77], while the affordance decoder adopts weights from SAM [78].

C.2 Imitation Learning in RLBench

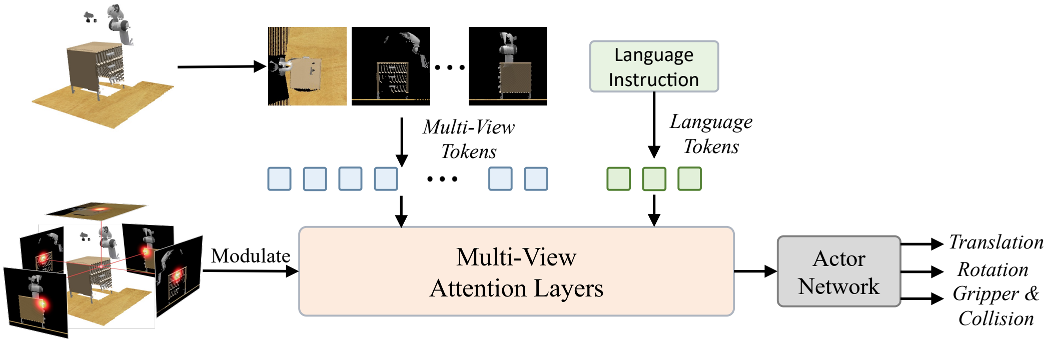


Figure 14: The pipeline of using actionable affordance as prior to guide the attention of multi-task language-guided manipulation.

We aim to demonstrate that explicit affordance representation can effectively guide imitation learning networks to focus on action-critical regions in visual inputs. We adopt RLBench [61] as a testbed to demonstrate. RLBench is a high-fidelity simulated environment built on PyRep (PyBullet wrapper) [72] with stable physics.

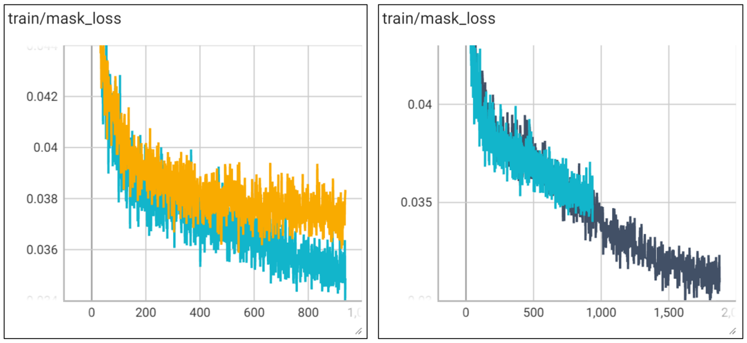
Our training closely follows RVT [74], utilizing cube-view re-rendered images generated from 3D point clouds. To ensure efficiency, the replay buffer of extracted keyframes is adopted to train the agent rather than all frames from episodes. For data augmentation, we adopt PerAct's [73] approach, applying random perturbation of translations within $\pm 0.125m$ and rotations along z-axis within $\pm 45^\circ$. We train the RVT and RVT-AFF for 5 epochs with a batch size of 24 and a learning rate of 1e-4. For the RVT2 and RVT2-AFF, we train the models for 100 epochs with a batch size of 24 and a learning rate of 1.25e-5.

Task	Language Template	# of Variations	Avg. Keyframes
open drawer	“open the __ drawer”	3	3.0
slide block	“slide the __ block to target”	4	4.7
sweep to dustpan	“sweep dirt to the __ dustpan”	2	4.6
meat off grill	“take the __ off the grill”	2	5.0
turn tap	“turn __ tap”	2	2.0
put in drawer	“put the item in the __ drawer”	3	12.0
close jar	“close the __ jar”	20	6.0
drag stick	“use the stick to drag the cube onto the __ target”	20	6.0
stack blocks	“stack __ __ blocks”	60	14.6
screw bulb	“screw in the __ light bulb”	20	7.0
put in safe	“put the money away in the safe on the __ shelf”	3	5.0
place wine	“stack the wine bottle to the __ of the rack”	3	5.0
put in cupboard	“put the __ in the cupboard”	9	5.0
sort shape	“put the __ in the shape sorter”	5	5.0
push buttons	“push the __ button, [then the __ button]”	50	3.8
insert peg	“put the __ peg in the spoke”	20	5.0
stack cups	“stack the other cups on top of the __ cup”	20	10.0
place cups	“place __ cups on the cup holder”	3	11.5

Table 7: Tasks we used in RLBench.

C.3 Ablations

We show more visualization of ablative studies to further demonstrate the effectiveness of the proposed components. The loss curves in Fig. 15 (a) show that global-to-local decoding leads to effective convergence of the total loss function, which conforms to better performance in Table 2. Also, the Fig. 15 (b) reveals the advantage of extending the training scheme in the GLOVER++'s fine-tuning.



(a) (b)

Figure 15: **The loss curves of training**. (**a**): The yellow and blue curve represents model w/o and w/ global-to-local decoding module, respectively. (**b**):The blue loss curve reflects a 5-epoch scheme of training, while the black one reflects a 10-epoch one.

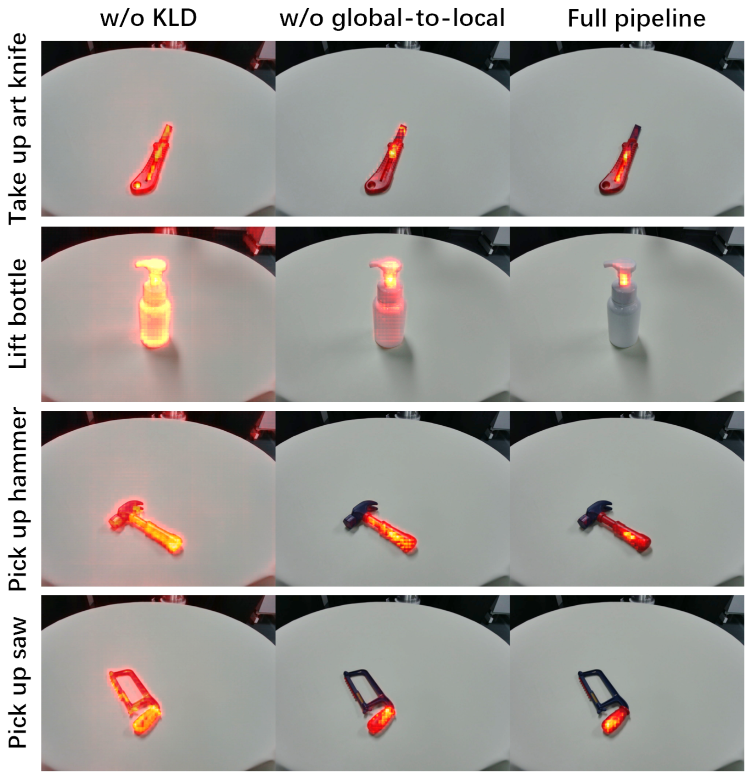


Figure 16: Visualization of the effectiveness of GLOVER++'s components in affordance reasoning.

The Fig. 16 illustrates the affordance reasoning visualization for ablative comparisons. Clearly, KLD loss and global-to-local decoding optimize local affordance representation learning for GLOVER++.

C.4 More Comparisons with Qwen-2.5

We show more comparisons between GLOVER++ and Qwen-2.5-VL-7B [58] model in Fig. 17. Compared to Qwen-2.5-VL, GLOVER++ demonstrates superior capability in generating task-compliant grasp points by explicitly modeling action-object semantics in language instructions. While Qwen-2.5-VL primarily relies on visual-language alignment for object localization, it often fails to disambiguate action-specific affordances to infer reasonable grasping regions.

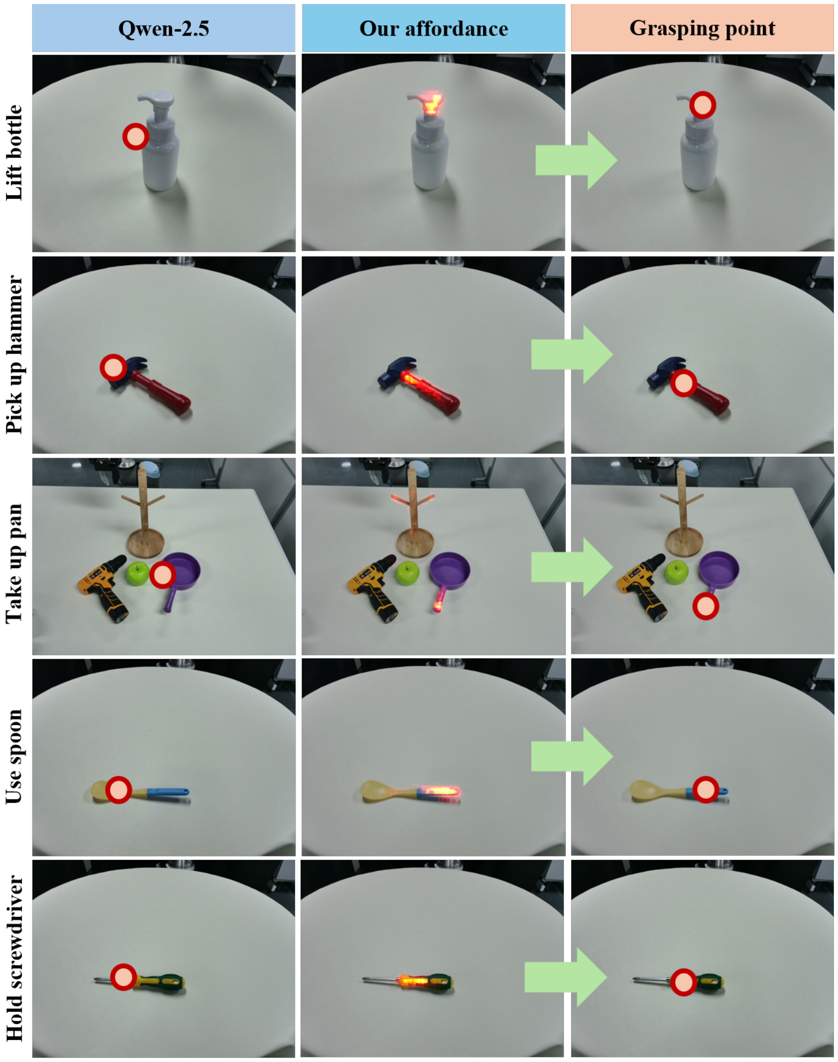


Figure 17: More visualization of comparison with Qwen-2.5-VL [58]. We show the inferred grasping point of GLOVER++ by argmaxing the affordance regions.

C.5 Real-world Experiments Setting

In the real-world experiments, we adopt two systems for manipulation tasks as Fig. 18 shows. The first system is based on a 7-DoF UFACTORY xAarm 7, equipped with DH-PGI gripper. We utilize an Orbbec Femto Bolt RGB-D camera for visual observations, with the image size of 1280 × 960 by default. The inverse kinematics (IK) to resolve the trajectory planning.

On the other hand, we leverage a Unitree G1 humanoid robot with two Inspire dexterous hands RH56DFX to construct a system for dexterous and bimanual grasping. We use the original head-mounted Intel394 RealSense D435i in Unitree G1 to capture RGB-D images with the size of 640 ×

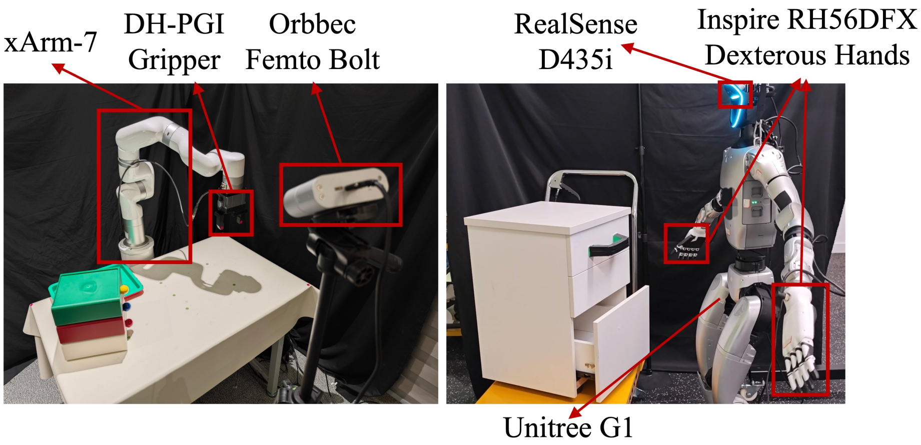


Figure 18: The real-world experiments settings.

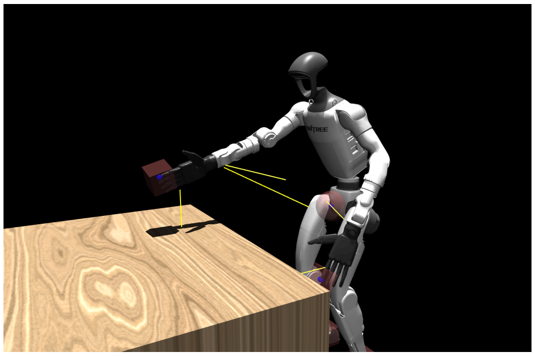


Figure 19: The illustration of the obstacle-avoidance IK we use to avoid self-collision.

480. For the motion planning, we use the obstacle-avoidance IK to avoid the self-collision as shown in Fig. 19.

C.6 Inference Speed Analysis

We have conducted a thorough inference time comparison on an Nvidia RTX4090, with an image size of 1280×960. As shown in Table 8, GLOVER++ significantly outperforms retrieval-based methods, achieving approximately 6× faster inference due to the elimination of the retrieval process. Compared to GLOVER, the additional inference time is negligible.

RAM [9]	GLOVER [11]	GLOVER++	LLaVA-7B in GLOVER++
6.78s	1.12s	1.21s	1.05s

Table 8: Comparison of inference speed between methods.

C.7 Failure Case

We illustrate the failure cases in the aspect of both affordance reasoning and real-world experiments.

Affordance Reasoning. The failure cases in the affordance reasoning includes two aspects as far as we know. First, when the viewpoint is excessively distant, GLOVER++ struggles to infer precise grasping regions and can only predict coarse object-level locations, as shown in Fig. 20 (a). Second, since the affordance reasoning is performed in 2D space, GLOVER++ sometimes struggles to dis-

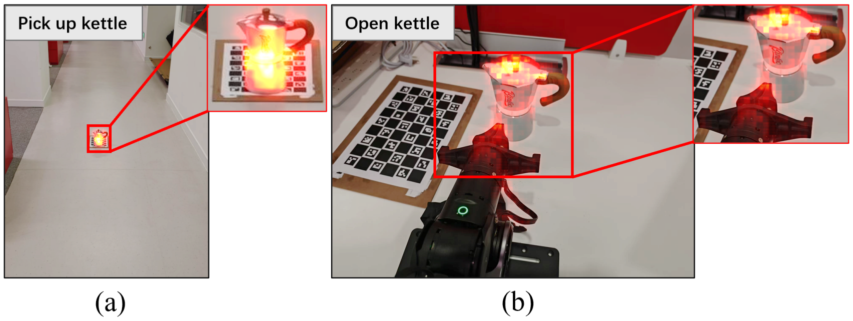


Figure 20: The failure cases of affordance reasoning, including: (a) distance viewpoint leads to inaccurate affordable regions, (b) failure to distinguish between overlapping objects sometimes, and results in background noise.

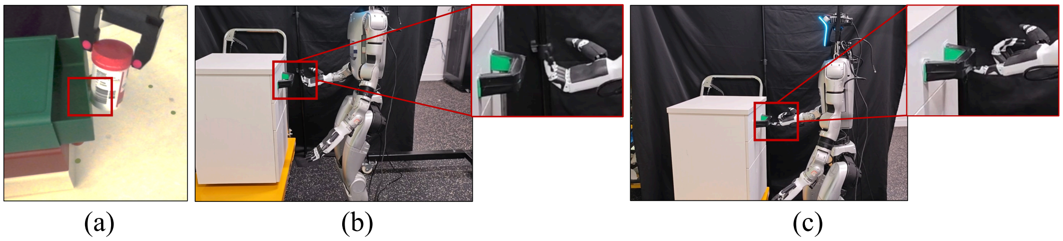
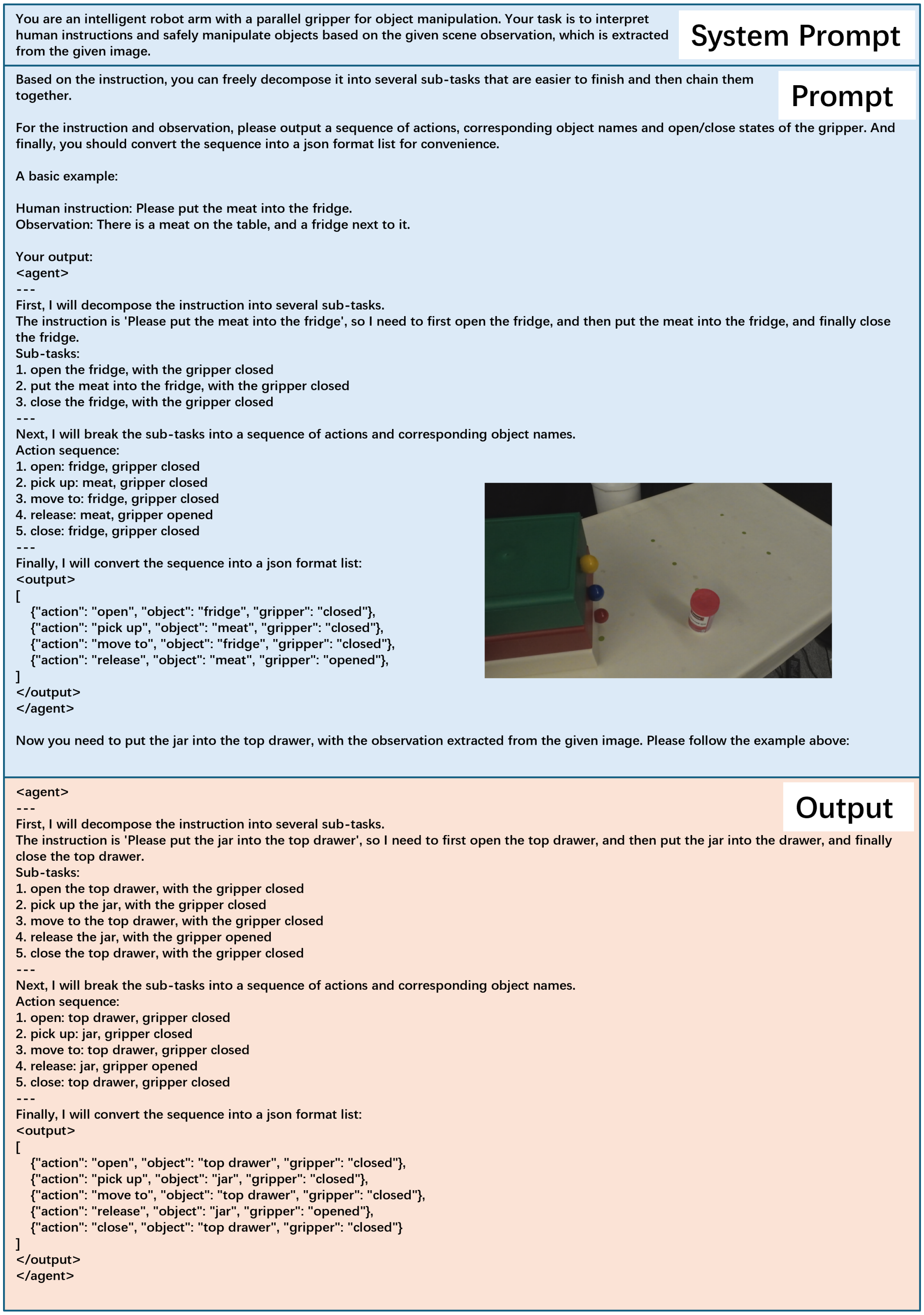


Figure 21: Failure cases in real-world experiments, including (a) collision problem, (b) z-axis inaccuracy, (c) imperfect grasping pose.

tinguish between overlapping objects at the pixel level, leading to noisy probability maps, although the highest-probability grasp points remain correct. The Fig. 20 (b) shows the circumstance.

Real-world Experiments. In the real-world experiments, the failure cases result from three primary reasons: (1) The collision caused by the imperfect rollout planning. (2) The projected affordable points may exhibit z-axis distance inaccuracies with one RGB-D camera, leading to the grasping failure. (3) The imperfect grasping pose of high-DoF dexterous hands leads to task failure. We show the above three cases in the Fig. 21.



```
You are an intelligent robot arm with a parallel gripper for object manipulation. Your task is to interpret
human instructions and safely manipulate objects based on the given scene observation, which is extracted
from the given image.
```

```
Based on the instruction, you can freely decompose it into several sub-tasks that are easier to finish and then chain them
together.

For the instruction and observation, please output a sequence of actions, corresponding object names and open/close states of the gripper. And
finally, you should convert the sequence into a json format list for convenience.

A basic example:

Human instruction: Please put the meat into the fridge.
Observation: There is a meat on the table, and a fridge next to it.

Your output:
<agent>
---
First, I will decompose the instruction into several sub-tasks.
The instruction is 'Please put the meat into the fridge', so I need to first open the fridge, and then put the meat into the fridge, and finally close
the fridge.
Sub-tasks:
1. open the fridge, with the gripper closed
2. put the meat into the fridge, with the gripper closed
3. close the fridge, with the gripper closed
---
Next, I will break the sub-tasks into a sequence of actions and corresponding object names.
Action sequence:
1. open: fridge, gripper closed
2. pick up: meat, gripper closed
3. move to: fridge, gripper closed
4. release: meat, gripper opened
5. close: fridge, gripper closed
---
Finally, I will convert the sequence into a json format list:
<output>
[
    {"action": "open", "object": "fridge", "gripper": "closed"},
    {"action": "pick up", "object": "meat", "gripper": "closed"},
    {"action": "move to", "object": "fridge", "gripper": "closed"},
    {"action": "release", "object": "meat", "gripper": "opened"},
]
</output>
</agent>

Now you need to put the jar into the top drawer, with the observation extracted from the given image. Please follow the example above:
```

```
<agent>
---
First, I will decompose the instruction into several sub-tasks.
The instruction is 'Please put the jar into the top drawer', so I need to first open the top drawer, and then put the jar into the drawer, and finally
close the top drawer.
Sub-tasks:
1. open the top drawer, with the gripper closed
2. pick up the jar, with the gripper closed
3. move to the top drawer, with the gripper closed
4. release the jar, with the gripper opened
5. close the top drawer, with the gripper closed
---
Next, I will break the sub-tasks into a sequence of actions and corresponding object names.
Action sequence:
1. open: top drawer, gripper closed
2. pick up: jar, gripper closed
3. move to: top drawer, gripper closed
4. release: jar, gripper opened
5. close: top drawer, gripper closed
---
Finally, I will convert the sequence into a json format list:
<output>
[
    {"action": "open", "object": "top drawer", "gripper": "closed"},
    {"action": "pick up", "object": "jar", "gripper": "closed"},
    {"action": "move to", "object": "top drawer", "gripper": "closed"},
    {"action": "release", "object": "jar", "gripper": "opened"},
    {"action": "close", "object": "top drawer", "gripper": "closed"}
]
</output>
</agent>
```

Figure 22: An example of prompting Qwen-2.5-VL-7B to decompose the long-horizon task for GLOVER++.