# Robot Manipulation in Salient Vision through Referring Image Segmentation and Geometric Constraints

Chen Jiang<sup>†</sup>, Allie Luo<sup>†</sup> and Martin Jagersand<sup>†</sup>

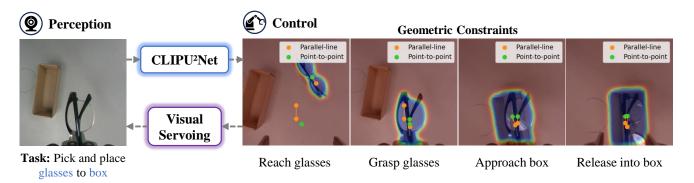


Fig. 1: To solve robot manipulation tasks in real-world environments, CLIPU<sup>2</sup>Net is first employed to segment regions most relevant to the target specified by referring language. Geometric constraints are then applied to the segmented region, generating context-relevant motions for uncalibrated image-based visual servoing (UIBVS) control.

Abstract-In this paper, we perform robot manipulation activities in real-world environments with language contexts by integrating a compact referring image segmentation model into the robot's perception module. First, we propose CLIPU<sup>2</sup>Net, a lightweight referring image segmentation model designed for fine-grain boundary and structure segmentation from language expressions. Then, we deploy the model in an eye-in-hand visual servoing system to enact robot control in the real world. The key to our system is the representation of salient visual information as geometric constraints, linking the robot's visual perception to actionable commands. Experimental results on 46 real-world robot manipulation tasks demonstrate that our method outperforms traditional visual servoing methods relying on labor-intensive feature annotations, excels in finegrain referring image segmentation with a compact decoder size of 6.6 MB, and supports robot control across diverse contexts.

## I. INTRODUCTION

An eye-in-hand visual servoing system [1], [2] depends on the flow of information from perception to control. In the perception phase, the system captures visual data with the eye-in-hand camera and processes this data to understand the manipulation context. The contextual data is then used in the control phase, where motor commands are generated to perform precise actions. For a robot to execute meaningful commands, it is necessary to interpret the evolving contexts of its workspace. However, the inherent movements of the eye-in-hand camera can cause visibility issues, where portions of objects may be obscured or fall outside the field of view, introducing significant challenges in perception.

Previous studies [3]–[7] have investigated the relationship between salient visual features and natural language in hand-eye cooperation, identifying regions that correlate with affordances. Recent advancements in large language models (LLMs) and vision-language models (VLMs) offer even more promising methods in generating point-based affordance representations from pixels for control [8]. However, the significant computational costs associated with LLMs pose difficulties for real-time robot control. Additionally, the use of points may overlook important details, such as fine boundaries and structures of the targets, making them less suitable for tasks requiring precise alignments. Consequently, exploring more compact models that balance between good affordance representations and computational efficiency becomes increasingly attractive. Can more compact models capture salient visual features from referring language, while enabling efficient robot manipulation?

To this end, extended from our previous work [9], we further explore how salient visual information interacts with robot control in real-world environments. We summarize our contributions as follows:

- We introduce CLIPU<sup>2</sup>Net, a new CLIP-driven model that delivers fine-grained referring image segmentation, with a compact decoder size of just 6.6MB.
- We approach robot manipulation in salient vision by framing it as a visual task specification problem, using CLIPU<sup>2</sup>Net and geometric constraints of points and lines to translate tasks into motions.
- Through experiments on 46 real-world tasks with varying appearances and contexts, we demonstrate the effectiveness of CLIPU<sup>2</sup>Net-inferred geometric constraints as universal visual representations for motions.

<sup>†</sup>Authors are with Department of Computing Science, University of Alberta, Edmonton AB., Canada, T6G 2E8. {cjiang2, luo3, mj7}@ualberta.ca

#### II. RELATED WORK

#### A. Salient Visual Features in Robot Control

Salient visual features have been extensively studied in robotics embodying cues. Particularly, classical methods like Vita [1], [2] relied on humans to annotate the cues as geometric constraints from parts of the objects. Jin et al [10] proposed to compose geometric constraints from dense visual descriptors into graph-based task functions. On the other hand, salient visual features can be learned unsupervised manifested as affordance signals, suggesting direct correlations between robotic actions and contexts. Visionbased approaches [4], [5], [7], [11], [12] involved using vision-language or contrastive models to infer cues from demonstration videos, while control-based approaches [13]-[15] learned to predict affordance scoring functions and actions. In particular, Bahl and Yoshida et al [7], [12] proposed to learn affordances from human videos, and deployed for robot control. Collin et al [15] learned to predict affordance heatmaps from joint vision, language and force inputs and enacted control. Still, how to effectively bridge from salient features to robot control remains challenging.

#### B. Vision-language Models in Robot Control

The uses of Vision-Language Models (VLMs) and Large Language Models (LLMs) have gained popularity due to their strong contextual reasoning capabilities. Methods such as Grounded Decoding [16], MOKA [8], AffordanceLLM [17], ManipLLM [18], and OVAL-Prompt [19] have demonstrated that LLMs can extract robust point-based affordance representations and be utilized as embodied agents. However, a significant drawback of LLM-based control systems is their high computational cost. For example, MOKA employed Grounded SAM [20] and Octo [21] to localize pixels and generate motion cues, while Grounded Decoding relied on CLIPort [14] for language-conditioned actions. These compositional strategies further amplify the computational demands of the control system.

In contrast, more compact approaches have been explored using smaller-scale models. Studies like Shridhar and Zhuo et al [14], [22] employed CLIP [23] to localize center attention or affordance cues for policy learning. However, these learning-based methods often require a large number of demonstrations per task for training. Jiang et al [9] utilized CLIPUNetr, a referring image segmentation model, to segment targets before applying UIBVS for control, though their work was limited to the reach-and-grasp context only.

# III. METHODOLOGY

We consider single-arm manipulation with an eye-in-hand camera. The goal is to segment regions related to the context specified by a referring expression in natural language, and compute geometric constraints using geometric points and lines for UIBVS control. Extended from [9], the overview of the system is presented in Figure 3. To enable the system, we first introduce CLIPU<sup>2</sup>Net, a compact model used to perform referring image segmentation. Then, we describe the process to compose geometric constraints in salient vision.

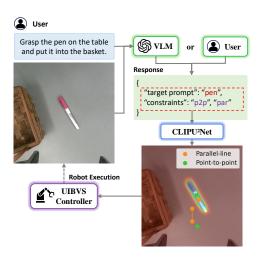


Fig. 3: Overview of the system to enact real-world robot control with CLIPU<sup>2</sup>Net and UIBVS control.

## A. Network Architecture

The architecture of CLIPU<sup>2</sup>Net, visualized in Figure 4, consists of three components: CLIP encoders to extract joint visual-text representations; a learnable masked multimodal fusion block that replaces Feature-wise Linear Modulation (FiLM); and a saliency module with U-squared decoder blocks for referring segmentation.

**CLIP Encoding** Given an image and text, the CLIP text and vision encoders compute the embedded text and image token features as  $X_l \in R^{D_l}$  and  $[X_v^{[CLS]}; X_v] \in R^{(L+1) \times D_v}$ , where  $X^{[CLS]} \in R^{D_v}$  is the image [CLS] token.

Masked Multimodal Fusion Previous studies [24], [25] have observed that Transformer-based models often have slow convergence due to processing context across tokens globally before slowly attending to local regions. We hypothesize that this issue also affects multimodal fusion between text and image features using Transformer-based models. To address this, we propose to use masked attention. We linearly project  $X_l$  and  $X_v$  into a joint embedding space and fuse them into a multimodal feature,  $X' \in R^{(L+1) \times D}$ :

$$X'_{l} = Proj_{l}(X_{l}), X'_{v} = Proj_{v}(X_{v})$$

$$X' = [X'_{l}; X'_{v}]$$
(1)

Replacing [CLS] token by the embedded token from text modality enforces multimodal encoder to pay more attention to the image tokens correlated to the context captured by text token. Furthermore, we composed of N Transformer layers with masked attention, where the masked self-attention mechanism is computed as:

$$X'_{n+1} = softmax((Q_n K_n^T + M_n)/\sqrt{D})V_n$$

$$M_n(i,j) = \begin{cases} 0, & \text{if } i = j \text{ or } j = 1, \\ -\infty, & \text{otherwise.} \end{cases}$$
 (2)

where n is the layer index,  $Q_n$ ,  $K_n$ ,  $V_n \in R^{(L+1)\times D}$  are the query, key, value features computed from the previous feature

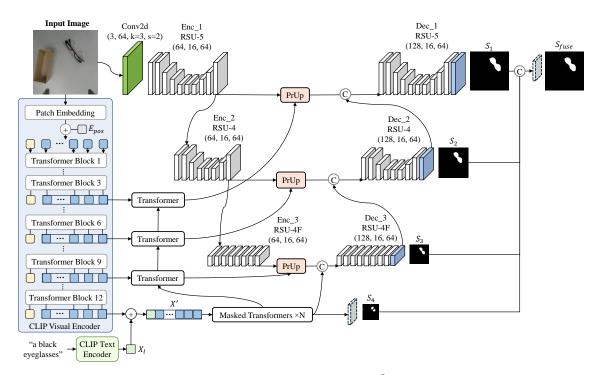


Fig. 4: The architecture of CLIPU<sup>2</sup>Net.

 $X_n^{'}$ . The attention mask  $M_n$  ensures that the self-attention mechanism ignores interactions among image tokens, computing attention only between text tokens and image tokens and the self-attention of each token. The final multimodal feature  $X_N^{'}$  is used to generate one side output  $S_4$ .

**Segmentation Decoder** The segmentation decoder integrates a miniature U²Net alongside the U-shaped transformer decoder to address the loss of multi-scale information caused by patch embedding in the CLIP visual encoder. The miniature U²Net comprises 6 blocks, each consisting of stacks of Residual U-blocks (RSU). The U²Net encoder processes the image input of size  $H \times W \times 3$  and generates three multiscale features  $F_e^1, F_e^2, F_e^3$  with three RSU encoding blocks, where each feature  $F_e^i \in \mathbb{R}^{H/2^i \times W/2^i \times D}$ .

Moreover, we integrate the U-shape transformer encoder from the original CLIPSeg to allow early visual information to flow into U2Net decoding process. Starting with the visual token feature  $X_v^9 \in \mathbb{R}^{L \times D_v}$  from the 9th CLIP visual encoding layer, it is processed by a transformer block, and reshaped into a 4D tensor  $X_v^{0}$  of D dimension. Then,  $X_v^{\prime 9}$  and  $F_e^3$  are taken by a Projection-Upsample (PrUp) block, where the low-resolution token feature is upsampled with bilinear interpolation and concatenated with the highresolution feature  $F_e^3$ , before being processed by a RSU block, generating the new multi-scale feature  $F_e^{'3}$ . Next,  $F_e^{'3}$ is concatenated with the final multimodal feature  $X'_L$  and processed by an RSU block, producing the decoding feature  $F_d^3$  and the third side output  $S_3$ . This process is repeated for the visual token features from the sixth and third layers of the CLIP visual encoder, generating additional side outputs  $S_2$ , and  $S_1$ . All side outputs are concatenated and refined by a  $1\times1$  convolution, producing the final probability map  $S_{fuse}$ .

**Training Loss** Similar to CLIPUNetr [9], the total loss is calculated as the sum of the losses from all outputs:

$$\mathcal{L} = \alpha_{fuse} \ell_{fuse} + \sum_{k=1}^{4} (\alpha_{side}^{k} \ell_{side}^{k})$$
 (3)

where  $\ell_{fuse}$  and  $\ell_{side}^k$  denote the loss for the fused and the side probability maps.  $\alpha_{fuse}$  and  $\alpha_{side}^k$  are the weights for each loss component, empirically chosen as 1. Each loss term  $\ell$  is computed as the sum of Focal and DICE loss.

## B. Geometric Constraints in Salient Vision

Geometric constraints [26] use points and lines to describe the alignment of a robot end effector to a target. Following [2], four base geometric constraints are used:

$$e_{pp}(\mathbf{f}) = f_2 - f_1$$

$$e_{pl}(\mathbf{f}) = f_1 \cdot f_{34}$$

$$e_{ll}(\mathbf{f}) = f_1 \cdot f_{34} + f_2 \cdot f_{34}$$

$$e_{par}(\mathbf{f}) = f_{12} \times f_{34}$$
(4)

where  $e_{pp}$ ,  $e_{pl}$ ,  $e_{ll}$ , and  $e_{par}$  are denoted as the error signals for point-to-point (p2p), point-to-line (p2l), line-to-line (l2l), and parallel-line (par) constraint, respectively. A line  $f_{ij}$  is computed as the cross product of two points,  $f_i$  and  $f_j$ . Given the error signals  $\dot{e}$  for the current visual observation, the UIBVS controller is governed by the visuomotor control law:

$$\dot{e} = J_u(q)\dot{q} \tag{5}$$

where  $\dot{q}$  is the control input of a robot with N degrees-of-freedom,  $J_u$  is the visuomotor Jacobian that maps the visual error signals to robot motions, and is updated online with Broyden's method [1]. From the definitions, two problems need to be addressed:

- What constraints should be used for the current task?
- Which points and lines should be used to dynamically form the error signals of the constraints?

To tackle these problems, we first establish a baseline method to automatically determine the appropriate geometric constraints by the initial task observations. Next, we define attention interactions to select the appropriate pairings of the points and lines in real time.

**Determining Geometric Constraints** Inspired by [8] that LLMs can generate robust point-based affordance representations to initiate control, we further investigate if LLMs are capable of inferring both point and line-based affordance representations encapsulated by the geometric constraints. Besides manually defining constraints by users through an HRI interface [2], [9], we now explore a baseline strategy to automatically determine these constraints using GPT-4o. Given only the initial visual observation of the workspace and a text describing the manipulation task goal, we prompt GPT-4o to infer the choice of geometric constraints  $E = e_i | e_i \in (e_{pp}, e_{pl}, e_{ll}, e_{par})$ , as well as the target prompt l to infer CLIP<sup>2</sup>Net.

Attention Interaction For the eye-in-hand camera configuration, attention interactions can be broadly categorized into two types: 1) object-gripper interaction, and 2) object-object interaction. For object-gripper interaction, the goal is to manipulate the target using the end effector. Affordances in this interaction arise from single-span attentions focused on the pixel locations of the target. Following the approach in [9], PCA is used to infer over the output saliency map  $S_{fuse}$ , extracting the target point  $f_2$  for the point-to-point constraint, or line  $f_{34}$  for point-to-line, line-to-line, or parallel-line constraints, respectively. A heuristic static point (e.g.  $f_1 = (W/2, 4H/5, 1)$ ) or a vertical line  $f_{12}$  passing through the mid-image center is used to complete the pairings.

For object-object interaction, where the object held by the end effector interacts with another target in the scene, affordances arise from dual attention spans: one over the object in the gripper and another over the target that completes the manipulation context. To generate geometric constraints, two prompts,  $l_1$  and  $l_2$ , are provided to describe the carried object and the target. CLIPU²Net is applied twice to generate two probability maps,  $S^1_{fuse}$  and  $S^2_{fuse}$ , on which PCA is applied to derive pairs of points and lines. Since the carried object appears static from the eye-in-hand camera's perspective, constraints are computed once for the carried object and dynamically updated for the target during control.

#### IV. EXPERIMENTS

# A. Experimental Settings

**Referring Image Segmentation** To evaluate the robustness of the proposed CLIPU<sup>2</sup>Net in pixel-wise localization as well as boundary prediction quality, we use three datasets:

- **PhraseCut** dataset [27] contains 340,000 referring phrases with associating regional segmentation masks. We follow its evaluation protocol and report the Mean intersection-over-union (mIoU) and cumulative intersection-over-union (cIoU) metrics.
- UMD+GT dataset [28] contains 30,000 RGBD images of 104 objects in multiple views and 6 affordance labels. Following [9], the labels are enriched with object-oriented and affordance-enriched prompts, and we report 4 metrics to measure prediction quality: Mean Absolute Error (MAE), structure measure  $(S_{\alpha})$ , weighted F-measure  $(wF_{\beta})$ , and max F-measure  $(F_{\beta}^{m})$ .
- **DIS5K** dataset [29] contains 5,470 images of 225 categories. Each image is manually annotated with highly accurate segmentation mask outlining structural complexities of the objects. For experiments, we use the category of the object as the prompts, and report the 6 metrics measuring the quality of the predicted object boundary and structure.

**Robot Control** We study the robot control system across four manipulation contexts: 1) Reach and grasp; 2) Pick and place; 3) Pull-open; and 4) Pour. 46 task targets with varying appearances are used. For each target, the robot is randomly positioned to view the target top-down or frontally. Two components are evaluated:

- Correctness of Constraints: For each manipulation target, a set of ground truth geometric constraints are selected by two skilled users. We assess the correctness of the baseline GPT-40 method to automatically determine the constraints, and report the accuracy.
- Real-time Control: Using the determined constraints, we enact real-world robot control with CLIPU<sup>2</sup>Net-augmented perception vs. classical perception, where visual tracking is used following [2]. Three attempts are allowed to complete a task with one target. One failed attempt results in a success rate of 50%, while two failed attempts result in a success rate of 0%. Success rates are averaged across all targets.

**Implementation Details** CLIPU<sup>2</sup>Net is implemented in PyTorch and trained on a single Nvidia Titan XP GPU with a batch size of 64, cosine annealing with an initial learning rate of 0.0005. The UIBVS control is implemented in ROS, utilizing Cartesian and joint spaces with support of velocity control for tabletop and front manipulation tasks. In comparison experiments, waypoint control is employed, moving the end effector incrementally until convergence.

# B. Results on Referring Image Segmentation

Quantitative Evaluation Table I and II present the quantitative results of referring image segmentation on the Phrase-Cut, UMD+GT, and DIS5K datasets, respectively. Our model achieves superior performance compared to MDETR [30], with only one-quarter of its size, demonstrating the benefits of using masked attention fusion for multimodal learning. Moreover, our model outperforms CLIPUNetr [9] in predicting fine boundaries and structures with fewer parameters, highlighting the effectiveness of the decoder design.

TABLE I: Quantitative results on PhraseCut and UMD+GT datasets.

	Phras	seCut	UMD+GT				
Model	mIoU	cIoU	MAE	$S_{\alpha}$	$wF_{\beta}$	$F_{\beta}^{m}$	
HulaNet [27]	0.413	0.508	-	-	-	-	
MDETR [30]	0.531	0.546	-	-	-	-	
CLIPSeg [31]	0.461	0.562	-	-	-	-	
CLIPSeg (PC+) [31]	0.434	0.547	0.134	0.663	0.384	0.566	
CLIPUNetr [9]	0.498	0.579	0.003	0.897	0.777	0.802	
AffKp [28]	-	-	0.004	0.876	0.675	0.803	
CLIPU <sup>2</sup> Net CLIPU <sup>2</sup> Net-Mix	0.542 <b>0.544</b>	0.589 <b>0.591</b>	0.001	0.932	0.843	0.866	

TABLE II: Quantitative results on DIS5K test set.

		DIS-TE (1-4)					
Model	$F_{\beta}^{m}$	$F_{\beta}^{w}$	MAE	$S_{\alpha}$	$E_{\phi}^{m}$	$HCE_{\gamma}$	
BASNet [32]	0.752	0.663	0.086	0.783	0.835	1313	
U <sup>2</sup> Net [33]	0.761	0.670	0.083	0.791	0.835	1333	
ISNet [29]	0.799	0.726	0.070	0.819	0.858	1016	
FP-DIS [34]	0.831	0.770	0.057	0.847	0.895	-	
UDUN [35]	0.831	0.772	0.057	0.844	0.892	977	
CLIPU <sup>2</sup> Net	0.846	0.779	0.054	0.858	0.899	1326	

Qualitative Evaluation Figure 5 shows the prediction results of CLIPU<sup>2</sup>Net compared to other methods. A common challenge in salient object segmentation is the uncertainty of what constitutes saliency, which should be adaptable based on a user's attention. Consequently, saliency models like DISNet [29] can struggle with localizing the full targets, while our model benefits from integrating CLIP for joint image-text representations to determine saliency and thus, achieving better localization capabilities. Our model also demonstrates comparable performance vs. other methods like CLIPSeg [31], avoiding checkerboard artifacts and having finer structures.

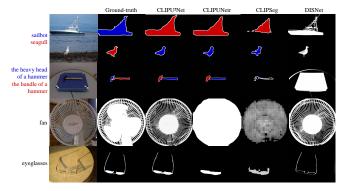


Fig. 5: Qualitative results for referring image segmentation.

**Ablation Study** We validate the effectiveness of each key component in CLIPU<sup>2</sup>Net by progressively removing the masked attention, U<sup>2</sup>Net decoder, or both. Additionally, we construct a simplified baseline model, referred to as CLIPU<sup>2</sup>Net-Bare. Similar to ViTSeg [31], the baseline model includes only a single convolution layer after the masked attention. The results are shown in Table III. Overall, both the

masked attention and U<sup>2</sup>Net decoder significantly contribute to learning fine-grain segmentation. The CLIPU<sup>2</sup>Net-Bare model also demonstrates considerably higher performance compared to ViTSeg, highlighting the effectiveness of the masked attention.

TABLE III: Ablation study with PhraseCut dataset.

Model	No Masked Attn	Coarse	Extreme	mIoU
CLIPU <sup>2</sup> Net-v1	✓	-	-	0.540
CLIPU <sup>2</sup> Net-v2	-	$\checkmark$	-	0.520
CLIPU <sup>2</sup> Net-v3	$\checkmark$	$\checkmark$	-	0.515
CLIPU <sup>2</sup> Net-Bare	-	$\checkmark$	$\checkmark$	0.449
ViTSeg [31]	✓	$\checkmark$	$\checkmark$	0.389

#### C. Results on Robot Control

Quantitative Evaluation We report the average success rate and comparison against classical Vita control [2] in Table IV, as well as the accuracy of the predicted geometric constraints. With the incorporation of CLIPU<sup>2</sup>Net, our robot control system achieves consistent performance in performing manipulation tasks with various targets. On the other hand, for contexts like pick-and-place, the classical interface struggles significantly. This is due to the fact that parts of the objects are obscured as the eye-in-hand camera moves, making the annotations and the tracking of the targets substantially harder. Moreover, the classical interface struggles to track transparent or more reflective objects like glass cups.

Additionally, GPT-40 shows some innate abilities in predicting line-based constraints from visual observations, especially for tasks involving part affordances like handles. This further validates the fact that manipulation contexts can arise from image geometry, which can be captured using point and line-based affordance representations. Still, the prediction quality varies across tasks, indicating some challenges in handling diverse contexts.

Qualitative Evaluation Figure 6 visualizes the geometric constraints for 8 out of the 46 assessed tasks. With the integration of CLIPU<sup>2</sup>Net, we can segment affordance regions, successfully completing tasks requiring fine-grain affordance information like reaching for the spoon handle in a bowl and placing a teabag into a mug, as seen in the first row. CLIPU<sup>2</sup>Net also handles diverse contexts effectively, from segmenting door handles for pulling tasks in the fourth row, to segmenting cans from top-down or frontal views for pickand-place tasks in the third row. Additionally, the usage of line-based constraints ensures stable motions, such as in the eyeglasses placement task (third row, fourth column), where the parallel-line constraint allow safe placement of the glasses into the tiny space of a box. In summary, geometric constraints are universal and effective visual representations of motion, adaptable to various contexts.

**Limitations** An issue with CLIPU<sup>2</sup>Net is its limitations in handling part segmentation, which affects its performance in tasks requiring detailed differentiation of object parts. We provide two failed tasks resulted from this limitation in Figure 7. In the first task, the model fails to segment the cap

TABLE IV: Results for robot control.

				Success Rate	
Context	Target	Constraint(s)	Acc	CLIPU <sup>2</sup> NETR	Classical [2]
Reach-and-Grasp	Apple. lemon, red pepper, carrot, banana, umbrella, tennis ball, 4 marker pen.	p2p, p2l, 12l, par	100%	100%	91%
	4 beverage can, 2 beverage bottles, 2 water bottles, med-bottle, 2 spray bottle.	p2l, p2p	72.7%	100%	91%
	Plastic cup, glass cup.	par, p21, p2p	50%	75%	38%
	Grasp-Fork-in-Plate. Grasp-Spoon-handle-in-Bowl.	par, p2p	100%	100%	50%
	The handle of a screwdriver, 2 hammers, 2 knife.	par, p2p	90%	60%	100%
Pick-and-Place	Strawberry, lemon $\rightarrow$ Bowl; Apple, carrot $\rightarrow$ Plate.	par, p2p; p2p	87.5%	100%	75%
	Coke can $\rightarrow$ Basket; Beverage can $\rightarrow$ Bucket.	p2p; p2p	100%	100%	75%
	Tea bag $\rightarrow$ Tea cup, 2 Mugs.	p2p; p2p	91.7%	83%	83%
	Eyeglasses $\rightarrow$ Eyeglass box; Marker pen $\rightarrow$ Rectangle basket.	par, p2p; par, p2p	50%	100%	67.5%
Pull-open	Left and right closet handles.	121, p2p	50%	100%	0%
	Drawer handle.	p2l, p2p	50%	100%	100%
Grasp-and-Pour	Pepper-in-bottle $\rightarrow$ Plate	p2l, p2p; p2p	100%	100%	0%



Fig. 6: Results of the predicted geometric constraints and motions for 8 of the 46 assessed tasks.

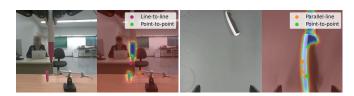


Fig. 7: Some failure cases.

of the marker pen when prompted with "pen cap". In the second task, the model fails to distinguish the blade from the handle of the knife. We hypothesize that this limitation arises from the annotations in the PhraseCut dataset, which lacks part annotations. We plan to evaluate this further in future work.

## V. CONCLUSIONS

In this paper, we introduce CLIPU<sup>2</sup>Net, a compact referring image segmentation model integrated into a robot's perception module, and use it to extract salient visual features as geometric constraints. Experimental results validate the effectiveness of our approach in enacting real-world robot control with eye-in-hand visual servoing. For future work, we aim to explore part segmentation from referring expressions to further refine contextual understanding and improve manipulation accuracy. Furthermore, we plan to investigate how geometric constraints can be leveraged to enhance the learning of referring image segmentation models themselves, optimizing their performance in real-world robot manipulation environments.

#### REFERENCES

- M. Jagersand, O. Fuentes, and R. Nelson, "Experimental evaluation of uncalibrated visual servoing for precision manipulation," in *Proceedings of International Conference on Robotics and Automation*, vol. 4. IEEE, 1997, pp. 2874–2880.
- [2] M. Gridseth, O. Ramirez, C. P. Quintero, and M. Jagersand, "Vita: Visual task specification interface for manipulation with uncalibrated visual servoing," in 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2016, pp. 3434–3440.
- [3] P. Abolghasemi, A. Mazaheri, M. Shah, and L. Boloni, "Pay attention!-robustifying a deep visuomotor policy through task-focused visual attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4254–4262.
- [4] C. Jiang, M. Dehghan, and M. Jagersand, "Understanding contexts inside robot and human manipulation tasks through vision-language model and ontology system in video streams," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 8366–8372.
- [5] Y. Zha, S. Bhambri, and L. Guan, "Contrastively learning visual attention as affordance cues from demonstrations for robotic grasping," in 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021, pp. 7835–7842.
- [6] Y. An, J. Yang, J. Li, B. He, Y. Guo, and G.-Z. Yang, "Skill learning in robot-assisted micro-manipulation through human demonstrations with attention guidance," in 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 15601–15607.
- [7] T. Yoshida, S. Kurita, T. Nishimura, and S. Mori, "Text-driven affordance learning from egocentric vision," arXiv preprint arXiv:2404.02523, 2024.
- [8] K. Fang, F. Liu, P. Abbeel, and S. Levine, "MOKA: Open-Set Robotic Manipulation through Mark-Based Visual Prompting," in *Proceedings* of Robotics: Science and Systems, Delft, Netherlands, July 2024.
- [9] C. Jiang, Y. Yang, and M. Jagersand, "Clipunetr: Assisting humanrobot interface for uncalibrated visual servoing control with clipdriven referring expression segmentation," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 6620– 6626.
- [10] J. Jin and M. Jagersand, "Generalizable task representation learning from human demonstration videos: a geometric approach," in 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022, pp. 2504–2510.
- [11] S. Yang, W. Zhang, W. Lu, H. Wang, and Y. Li, "Learning actions from human demonstration video for robotic manipulation," in 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019, pp. 1805–1811.
- [12] 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, 2023, pp. 13778–13790.
- [13] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2021, pp. 726–747.
- [14] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 894–906.
- [15] J. A. Collins, C. Houff, Y. L. Tan, and C. C. Kemp, "Forcesight: Text-guided mobile manipulation with visual-force goals," in 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 10874–10880.
- [16] W. Huang, F. Xia, D. Shah, D. Driess, A. Zeng, Y. Lu, P. Florence, I. Mordatch, S. Levine, K. Hausman et al., "Grounded decoding: Guiding text generation with grounded models for embodied agents," Advances in Neural Information Processing Systems, vol. 36, 2024.
- [17] 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, 2024, pp. 7587–7597.
- [18] X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, and H. Dong, "Manipllm: Embodied multimodal large language model for object-centric robotic manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18061–18070.

- [19] E. Tong, A. Opipari, S. Lewis, Z. Zeng, and O. C. Jenkins, "Oval-prompt: Open-vocabulary affordance localization for robot manipulation through llm affordance-grounding," 2024.
- [20] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan et al., "Grounded sam: Assembling open-world models for diverse visual tasks," arXiv preprint arXiv:2401.14159, 2024.
- [21] O. Mees, D. Ghosh, K. Pertsch, K. Black, H. R. Walke, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo et al., "Octo: An open-source generalist robot policy," in First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024.
- [22] F. Zhuo, Y. He, F. Yu, P. Li, Z. Zhao, and X. Sun, "Abm: Attention before manipulation," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, K. Larson, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2024, pp. 1816–1824, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2024/201
- [23] 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*. PMLR, 2021, pp. 8748–8763.
- [24] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision* and pattern recognition, 2022, pp. 1290–1299.
- [25] B. Lai, M. Liu, F. Ryan, and J. M. Rehg, "In the eye of transformer: Global-local correlation for egocentric gaze estimation and beyond," *International Journal of Computer Vision*, vol. 132, no. 3, pp. 854–871, 2024.
- [26] J. P. Hespanha, Z. Dodds, G. D. Hager, and A. S. Morse, "What tasks can be performed with an uncalibrated stereo vision system?" *International Journal of Computer Vision*, vol. 35, pp. 65–85, 1999.
- [27] C. Wu, Z. Lin, S. Cohen, T. Bui, and S. Maji, "Phrasecut: Language-based image segmentation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10216–10225.
- [28] R. Xu, F.-J. Chu, C. Tang, W. Liu, and P. A. Vela, "An affordance keypoint detection network for robot manipulation," *IEEE Robotics* and Automation Letters, vol. 6, no. 2, pp. 2870–2877, 2021.
- [29] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. Van Gool, "Highly accurate dichotomous image segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 38–56.
- [30] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1780–1790.
- [31] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7086–7096.
- [32] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7479–7489.
- [33] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, p. 107404, 2020
- [34] Y. Zhou, B. Dong, Y. Wu, W. Zhu, G. Chen, and Y. Zhang, "Dichoto-mous image segmentation with frequency priors." in *IJCAI*, vol. 1, no. 2, 2023, p. 3.
- [35] J. Pei, Z. Zhou, Y. Jin, H. Tang, and P.-A. Heng, "Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2139–2147.