# Attribute-based Object Grounding and Robot Grasp Detection with Spatial Reasoning

Houjian Yu[1], Zheming Zhou[2], Min Sun[2,3], Omid Ghasemalizadeh[2], Yuyin Sun[2], Cheng-Hao Kuo[2], Arnie Sen[2], and Changhyun Choi[1]

***Abstract*— Enabling robots to grasp objects specified through natural language is essential for effective human–robot interaction, yet it remains a significant challenge. Existing approaches often struggle with open–form language expressions and typically assume unambiguous target objects without duplicates. Moreover, they frequently rely on costly, dense pixel–wise annotations for both object grounding and grasp configuration. We present Attribute–based Object Grounding and Robotic Grasping (OGRG), a novel framework that interprets open–form language expressions and performs spatial reasoning to ground target objects and predict planar grasp poses, even in scenes containing duplicated object instances. We investigate OGRG in two settings: (1) Referring Grasp Synthesis (RGS) under pixel–wise full supervision, and (2) Referring Grasp Affordance (RGA) using weakly supervised learning with only single–pixel grasp annotations. Key contributions include a bi-directional vision–language fusion module and the integration of depth information to enhance geometric reasoning, improving both grounding and grasping performance. Experiment results show that OGRG outperforms strong baselines in tabletop scenes with diverse spatial language instructions. In RGS, it operates at 17.59 FPS on a single NVIDIA RTX 2080 Ti GPU, enabling potential use in closed–loop or multi–object sequential grasping, while delivering superior grounding and grasp prediction accuracy compared to all the baselines considered. Under the weakly supervised RGA setting, OGRG also surpasses baseline grasp–success rates in both simulation and real–robot trials, underscoring the effectiveness of its spatial reasoning design. Project page: `https://z.umn.edu/ogrg`**

## I. Introduction

Target-oriented robot grasping is a fundamental task in robot manipulation, with wide-ranging applications in real-world scenarios. Compared to purely vision-driven robot grasping approaches [1], [2], [3], [4], language-driven robot grasping offers greater flexibility by leveraging object attributes (e.g., color, shape, category name, and spatial location) to differentiate the target object from others [5], [6], [7]. This capability reduces ambiguity in target identification. However, the reliance on detailed language descriptions introduces additional challenges for robust vision–language understanding.

Previous language-conditioned robot grasping approaches have been constrained to predefined vocabulary and simplistic attribute descriptions (e.g., "red apple"), while typically assuming the presence of distinct, non-duplicated objects in the scene [6], [7], [8]. Consequently, these methods cannot handle open-form language inputs or perform more challenging grounding tasks such as spatial reasoning. While recent advances in Multimodal Large Language Models (MLLMs), including Vision–Language–Model (VLM) and Vision–Language–Action (VLA) models, have demonstrated strong multimodal understanding for high-level planning and action generation [9], [10], [11], [12], [13], [14], their substantial computational demands for data generation, training, and inference often limit their deployment on resource-constrained robotic platforms. In addition, annotating large-scale robotic grasp datasets with detailed language descriptions remains labor-intensive for both human operators and autonomous systems.

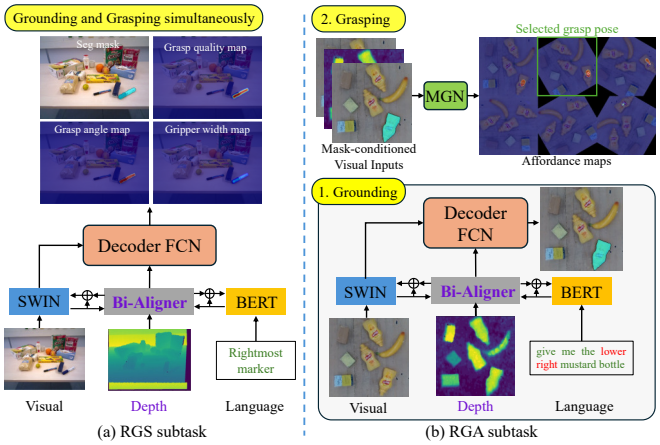



Fig. 1: **Object Grounding and Robot Grasping (OGRG) model with open-form expressions for spatial reasoning.** The model is designed to solve the attribute-based grounding and grasp detection task. The RGS subtask aims at predicting grasp rectangles with pixel-wise full supervision. The RGA subtask focuses on predicting grasp affordances with weak grasping supervision.

Despite the promising performance with MLLMs across various objects and environments, two critical challenges remain:

- **Grasp learning perspective**: *Can a compact and computationally efficient fusion module, serving as an alternative to MLLMs, be designed to effectively align vision and language features for real-world robot grasping?*
- **Data efficiency perspective**: *Can the grasping model be trained in a weakly supervised manner using sparse and imperfect labels?*

Without relying on pre-aligned vision–language mod-

*This work was supported by Amazon Lab126.

[1] The authors are with the Department of Electrical and Computer Engineering, Univ. of Minnesota, Minneapolis, USA {yu000487, cchoi}@umn.edu

[2] The authors are with Amazon Lab126, Sunnyvale, CA, USA {zhemiz, minnsun, ghasemal, yuyinsun, chkuo, senarnie}@amazon.com

[3] The authors is with National Tsing Hua University, Taiwan {sunmin}@ee.nthu.edu.tw

els [15], [16], we primarily investigate effective multimodal fusion and training with both dense and sparse labels for object grounding and grasp detection. In this paper, we propose a bi-directional multimodal fusion module to align vision, language, and depth features from different embedding spaces. The fused multimodal features are used for two grasp detection settings as shown in Fig .1: Referring Grasp Synthesis (RGS) [17], [18], [5] and Referring Grasp Affordance (RGA) [6], [17], [19] (see III-A for problem formulation details). Both tasks require pixel-level vision-language understanding for object grounding and planar grasp pose prediction. The most closely related work to ours is ETRG [17], which employs the CLIP model with a downsampling-then-upsampling strategy for parameter-efficient tuning, aiming to reduce the number of trainable parameters while maintaining multimodal alignment. However, the aggressive feature downsampling inevitably leads to information loss, resulting in suboptimal performance on object grounding. In contrast, our approach strikes a better balance between model compactness (approximately 240M total parameters) and task performance by introducing a novel fusion module that facilitates effective interaction between the vision and language backbones.

Our grounding and grasping system is capable of (1) accepting open-form object attribute descriptions, including colors, shapes, category names, and spatial reasoning, to predict target object masks and planar grasp poses in the format of grasp rectangles and grasp affordance maps, (2) effectively fusing multimodal features without relying on pre-aligned models and utilizing both dense and sparse robot grasping labels for predictions, and (3) grounding the target object with high accuracy while achieving a high success rate in grasping.

Our primary contributions are summarized as follows:

- An end-to-end Object Grounding and Robot Grasping (OGRG) model for RGS and RGA tasks, predicting object masks and 5-DoF grasp poses under dense supervision, and grasp affordances under weak supervision.
- A bi-directional multimodal fusion module is introduced to align vision, language, and OGRG downstream tasks, RGS and RGA.
- The approach is validated through comprehensive experiments in both simulation and real robot environments, demonstrating effectiveness across diverse objects and open-form language inputs compared to baseline methods.

## II. Related Work

### A. Language-guided Object Grounding

Language-guided object grounding focuses on localizing the target object referred to by language within an image. This problem can be categorized into two tasks based on the output type for localization: referring expression comprehension (REC), which predicts bounding boxes, and referring expression segmentation (RES), which generates binary segmentation masks. This paper concentrates on pixel-level multimodal alignment for predicting segmentation masks and grasping affordances, deriving from the RES task.

Early RES methods [20], [21], [22] employed fully convolutional networks for visual feature extraction and RNN/LSTM architectures for language embeddings. These approaches typically utilized simple multimodal feature concatenation or multiplication, followed by convolutional layers and upsampling, to decode target masks. More recent works [17], [23], [24], [25] leverage well-aligned vision-language models, such as the CLIP model [15], pretrained on large-scale datasets, to enhance vision-language fusion. Another line of research, including LAVT [26], CGFormer [27], and DMMI [28], incorporates Swin Transformer [29] as the visual feature extractor while actively exploring cross-modal attention mechanisms to embed vision-language features into a shared space. The method LAVT [26] is closely related to ours, which introduces early-stage uni-directional fusion modules for visual and linguistic feature interactions. In contrast, this work proposes bi-directional multimodal fusion modules that further integrate depth features, extending the RES task to the robotics domain for object grounding and grasp pose prediction.

### B. Language-guided Robot Grasping

Recent advancements in vision-language models [15], [30], [31] have significantly advanced the field of language-guided robot grasping, enabling robots to identify and manipulate objects based on user-provided natural language instructions [7], [17], [18], [9], [5], [19], [32], [10]. These models leverage pre-trained embeddings, such as CLIP and similar architectures, to align visual and textual modalities effectively, allowing robots to process diverse and open-vocabulary instructions. Early approaches primarily focused on tasks like object detection and segmentation, often relying on handcrafted features and task-specific training data. Recent methodologies extend these capabilities by integrating multimodal transformers and attention mechanisms to enhance contextual understanding and reasoning [9], [5]. These models excel at handling ambiguous or complex instructions, such as spatial references or multi-object contexts, by generating grasp affordance maps and candidate grasp poses with high precision. Such advancements pave the way for more robust and flexible applications in unstructured and dynamic environments, addressing key challenges in open-world robot manipulation. Our work builds on previous vision and language models while targeting object grounding and robot grasping tasks with strong and weak supervision.

## III. Method

In this section, we propose the OGRG model for attribute-based language-guided object grounding and robot grasping. The OGRG method is designed to address two primary grasp detection tasks: Referring Grasp Synthesis (RGS) and Referring Grasp Affordance (RGA).

### A. Problem Formulation

The detailed formulations for the RGS and RGA subtasks are first introduced, as shown in Fig. 1. These tasks involve

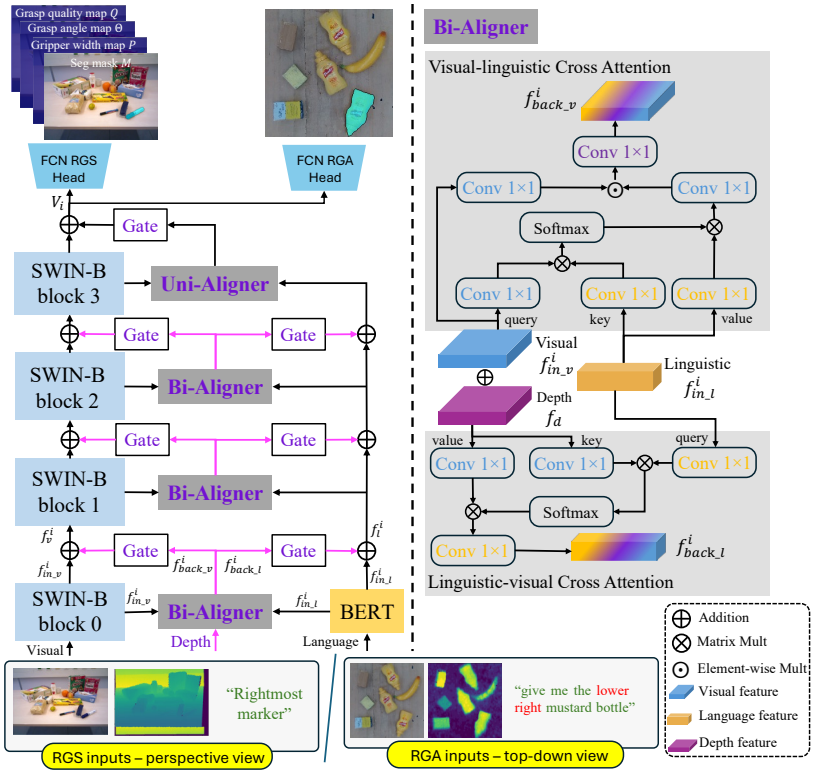



Fig. 2: **OGRG Architecture.** OGRG processes open-form language expressions, visual images, and depth maps as inputs to generate task predictions. The bidirectional aligner (Bi-Aligner) fuses the multimodal features extracted from Swin Transformer [29] at different stages and the BERT language model [33]. The updated multimodal features $f^i_{back\_v}$ and $f^i_{back\_l}$ are fed back into their corresponding visual and linguistic branches via feature gates. Finally, the light-weighted fully convolutional network (FCN) heads processes the updated visual features at different stages to produce the task-specific outputs.

processing multimodal inputs, including an RGB image denoted as $I \in \mathbb{R}^{H\times W\times 3}$, where $(H, W)$ represent the image dimensions; an attribute-based language description with token length $L$ represented as $T \in \mathbb{R}^L$; and a depth image denoted as $D \in \mathbb{R}^{H\times W}$. The proposed OGRG model leverages these visual observations and open-form language expressions to generate predictions for object grounding binary masks $M \in \mathbb{R}^{H\times W}$ and robot grasp pose $G$. We use a unified representation for grasp detection here: $G = \{x, y, z, \theta, l\}$ is parametrized by: $(x, y)$, the gripper center location in the image coordinate; $z$, the depth value; $\theta$, the gripper rotation angle in the camera frame; and $l$, the open width of the gripper.

**RGS Subtask:** This task aims to predict a segmentation mask and planar 5-DoF grasp poses in the format of grasp rectangles, given an RGB image $I$, a depth map $D$ from the camera perspective view, and a language expression $T$. Following [18], the OGRG model predicts three grasp-related maps to reconstruct the gripper pose. The ground-truth grounding and grasping maps are from the OCID-VLG dataset [18]. The system is formulated as $\Phi(I, D, T) = \{M, Q, \Theta, P\}$ to predict the object grounding mask $M$, the grasp quality map $Q$, the grasp angle map $\Theta$, and the gripper open width map $P$. Specifically, $(x, y)$ is determined by the pixel coordinate of the maximum value in $Q$. The rotation angle $\theta$ is derived as $\theta = \Theta(x, y)$ from $\Theta$. $z$ can be derived from depth map $D(x, y)$. Finally, the gripper open width $l$ is obtained from $P$ as $l = P(x, y)$.

**RGA Subtask:** Unlike RGS, RGA will predict grasp affordance maps $A \in \mathbb{R}^{H\times W\times N}$ with $N$ discretized rotation angles [6] (Fig. 1b). We adopt a segment-then-grasp pipeline with OGRG and a Mask-Conditioned Grasping Network (see section III-F). The 5-DoF grasp pose is derived from the affordance maps, where $(x^*, y^*, \theta^*) = \text{argmax}_{(x,y,\theta)} A(x, y, \theta)$ and $z = D(x^*, y^*)$. Here, $(x^*, y^*)$ corresponds to the pixel coordinate with the maximum affordance value, $\theta^*$ represents the optimal rotation angle, and $z$ provides the depth value for the grasp motion. We use a predefined gripper open width $l^*$ for all grasp attempts in RGA to facilitate the data collection process.

### *B. OGRG Multimodal Feature Fusion*

Fig. 2 illustrates the details of the proposed OGRG model. Depending on the different settings for RGS and RGA subtasks, the OGRG model will provide 4 different grounding and grasping maps $\{M, Q, \Theta, P\}$ simultaneously for the RGS subtask after passing the task-specific FCN head. On the other hand, the OGRG model will only predict the object grounding map for RGA subtask.

The OGRG model employs Swin Transformer [29] as the visual backbone and BERT Transformer [33] as the language feature extractor. For the depth branch, a ResNet-18 [34] model is utilized to extract depth features $f_d$. To enable efficient vision-language alignment, the model incorporates

a four-stage hierarchical multimodal fusion process with multiple aligners. At each stage, the visual features $f^i_{in\_v} \in \mathbb{R}^{C_i \times H_i \times W_i}$ and linguistic features $f^i_{in\_l} \in \mathbb{R}^{L \times C_t}$ interact through two distinct cross-attention mechanisms, resulting in fused multimodal features $f^i_{back\_v}$ and $f^i_{back\_l}$ for their respective branches. Here, $C_i$, $H_i$, and $W_i$ represent the number of channels, height, and width of the $i$-th stage ($i \in \{1, 2, 3, 4\}$), while $L$ and $C_t$ denote the language token length and token dimension, respectively.

The fused features are passed through learnable feature gates $g_i$ and added element-wise to $f^i_{in\_v}$ and $f^i_{in\_l}$, generating enhanced visual and linguistic features $f^i_v$ and $f^i_l$. Finally, the light-weighted Fully Convolutional Network (FCN) head processes the four-stage intermediate multimodal visual feature maps to produce the final task-specific outputs.

### C. Bidirectional Aligner

Inspired by the unidirectional fusion module from LAVT [26], a bidirectional aligner is proposed to update the two branches simultaneously. The bidirectional aligner (Bi-Aligner in Fig. 2) consists of visual-linguistic and linguistic-visual cross-attention mechanisms for multimodal fusion. The visual and depth features are first fused via element-wise addition, $f_{in\_vd} = f_{in\_v} + f_d$, where the depth feature $f_d$ is used only at the first stage. Given the flattened visual-depth features $f^i_{in\_vd} \in \mathbb{R}^{C_i \times D}$, where $D = H_i \times W_i$, and the linguistic features $f^i_{in\_l} \in \mathbb{R}^{L \times C_t}$ from the model backbone, cross-attention features are computed using the transformer attention formulation:

$$f^i_{cross\_v} = \text{softmax}\left(\frac{(W^V_q f^i_{in\_vd})^T (W^V_k f^i_{in\_l})}{\sqrt{C_i}}\right)(W^V_v f^i_{in\_l})^T, \tag{1}$$

$$f^i_{cross\_l} = \text{softmax}\left(\frac{(W^L_q f^i_{in\_l})^T (W^L_k f^i_{in\_vd})}{\sqrt{C_t}}\right)(W^L_v f^i_{in\_vd})^T, \tag{2}$$

where $W^V_q, W^V_k, W^V_v, W^L_q, W^L_k, W^L_v$ are projection matrices that unify the visual-depth and linguistic feature dimensions. The resulting cross-modal features are reshaped into $f^i_{cross\_v} \in \mathbb{R}^{C_i \times H_i \times W_i}$ and $f^i_{cross\_l} \in \mathbb{R}^{L \times C_t}$. These features are further processed with $1 \times 1$ convolutions and ReLU activations to produce the fused visual and linguistic features, $f^i_{back\_v}$ and $f^i_{back\_l}$.

Following the language pathway design from LAVT [26], learnable gates $g_i$ are applied to enhance the features, yielding the final visual and linguistic outputs:

$$f^i_v = f^i_{in\_v} + g_i(f^i_{back\_v}), \tag{3}$$

$$f^i_l = f^i_{in\_l} + g_i(f^i_{back\_l}). \tag{4}$$

### D. Task Specific FCN Head

The annotation $V_i$, $i \in \{1, 2, 3, 4\}$ is used to represent the intermediate visual features as inputs to the FCN head. From an empirical result, we use $V_i = \{f^i_v\}$ in RGS, and $V_i = \{f^i_{back\_v}\}$ in RGA for best performance. The decoding process is formulated as:

$$\begin{cases} Y_4 &= V_4, \\ Y_i &= \text{Conv}([\text{Up}(Y_{i+1}); V_i]), \quad i = 3,\ 2,\ 1, \end{cases} \tag{5}$$

where $\text{Conv}(\cdot)$ denotes a $3 \times 3$ convolution layer followed by batch normalization and a ReLU activation function, and $\text{Up}(\cdot)$ indicates bilinear interpolation upsampling. The decoded $Y_1$ serves as the final model prediction, outputting $\{M, Q, \Theta, P\}$ for RGS (conducting grounding and grasping simultaneously) and $M$ for RGA (only for object grounding).

To train the OGRG-RGS model, the loss function combines the cross-entropy loss for object grounding mask prediction ($M$) with smooth L1 losses for the grasp quality map ($Q$), grasp angle map ($\Theta$), and gripper open width map ($P$). For the OGRG-RGA model, dice loss and focal loss are applied for object grounding mask prediction.

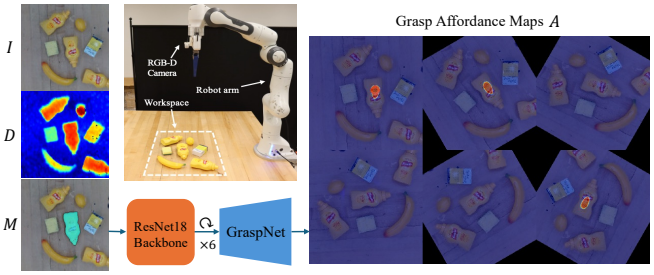



Fig. 3: **Mask-conditioned Grasping Network (MGN).** Conditioned on the object grounding mask $M$ predicted from OGRG, the MGN network uses a fully convolutional encoder-decoder architecture for pixel-level grasp affordance prediction with different rotation angles.

### E. Mask-conditioned Grasping Network for RGA

As shown in Fig. 3, Mask-conditioned Grasping Network (MGN) predicts pixel-level affordance maps $A$ for $N = 6$ discrete rotation angles, each a multiple of $\theta = 30^\circ$ based on the RGB image $I$, the depth map $D$, and the OGRG grounding mask $M$. Specifically, the inputs are concatenated and passed through a ResNet-18 backbone. To handle challenging grasping rotations, tensor transformations are applied, and the processed features are fed into an FCN-based GraspNet, which consists of standard convolutional layers, batch normalization, and ReLU activations. A Sigmoid layer serves as the final output layer to produce the grasp affordances.

To minimize human annotation effort, this problem is formulated in a weak-supervision manner, where only a binary $\{0, 1\}$ ground-truth label is provided for the sampled grasp location—a single pixel among the $N$ rotation maps—while other pixels remain unlabeled. The training process employs a motion loss $\mathcal{L}_{grasp}$ from Attribute-Grasp [6]

### F. Dataset Collection for RGA

We collect training and testing data for the RGA task in the CoppeliaSim simulator [35]. From a pool of 32 objects, 7 were randomly selected—primarily from the YCB dataset [36]—to construct grasping scenes, as illustrated in Fig. 4a. To improve robustness to environmental variations, we applied domain randomization to the background textures. Language instructions were generated from multiple

templates incorporating object color, shape, category name, and spatial location. Spatial relationships were expressed in two forms: *Absolute* (relative to the workspace) and *Relative* (relative to other reference objects), increasing the variety of spatial reasoning cases. In total, we collected over 16,000 visual–language–grasp triplets for model training.

## IV. Experiments

In this section, attribute-based object grounding and robot grasping experiments are conducted to evaluate the proposed method. The objectives of the experiments are to verify the following: 1) The Bi-Aligner with depth fusion effectively fuses multimodal features without relying on pre-aligned vision-language backbones. 2) The OGRG model demonstrates the ability to understand object-attribute descriptions and achieves a high grasping success rate, even with complex spatial relationship language inputs. 3) The proposed OGRG-RGA, combined with the MGN, successfully addresses the weakly supervised RGA problem and efficiently adapts from simulation to real robot experiments.

### A. OGRG-RGS with Depth Fusion

**Implementation Details:** The OGRG-RGS model is trained on the OCID-VLG dataset for 26 epochs with a batch size of 4 per GPU, using a total of 8 NVIDIA V100 GPUs. The training process employs the AdamW optimizer with an initial learning rate of $\lambda = 0.00005$ and a polynomial learning rate decay. For fair comparisons, input images are resized to a resolution of $416 \times 416$, and the maximum sentence length is capped at 20 tokens for all baselines.

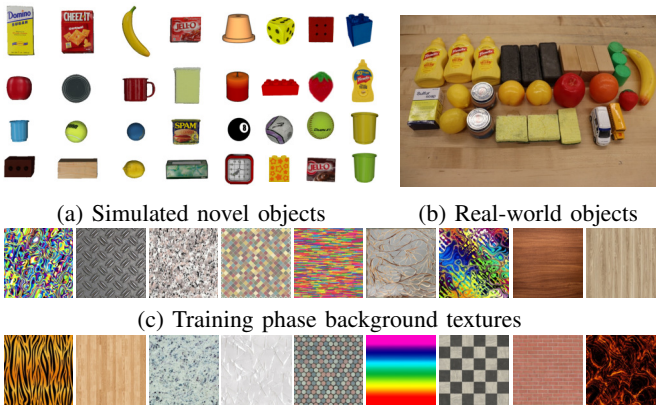
**RGS Evaluation Dataset:** The OGRG-RGS model and corresponding baselines are evaluated on the OCID-VLG dataset [18]. This dataset is designed for target object grounding and grasp pose prediction based on open-form language descriptions. It includes 58 unique object candidates, over $89.6k$ referring language expressions describing a wide range of object attributes, and more than $75k$ hand-annotated grasp rectangles.

**Evaluation Metrics:** The image segmentation results for the language-referred object grounding task are evaluated using the mean intersection over union (mIoU). For robot grasping, the Jaccard Index $J@N$ metric, as described in [18], is employed. This metric measures the top-$N$ grasp rectangles that achieve an IoU greater than 0.25 and have rotation angle differences of less than $30^\circ$ compared to the ground-truth grasp rectangles.

**Baselines:** The RGS evaluation results are reported with the following baselines: 1) CROG, proposed by [18], extends the referring expression segmentation model CRIS [25] for grasp map prediction. This approach involves full model fine-tuning, including the pre-trained CLIP model [15]. 2) ETRG [17] is a CLIP-based method that employs a parameter-efficient tuning framework with depth fusion branches. Instead of fine-tuning the full CLIP model, it uses a bidirectional adapter optimized for multiple tasks. 3) HiFi-CS [37] applies hierarchical FiLM [38] fusion for multimodal alignment and serves as another CLIP-based object grounding method. 4) LAVT [26] adopts unidirectional aligners for multimodal fusion. 5) OGRG-nodepth applies our bi-directional aligner but without depth inputs.

TABLE I: **OGRG-RGS ablation study and baseline comparison on the OCID-VLG dataset.** Our proposed method improves both object grounding performance and the accuracy of grasp rectangle predictions.

| Baselines | Grounding | Grasping | |
|---|---|---|---|
| | mIoU | J@1 | J@Any |
| CROG [18] | 81.10 | 77.20 | 87.70 |
| ETRG [17] | 80.11 | 89.38 | 93.49 |
| HiFi-CS [37] | 88.26 | - | - |
| LAVT [26] | 92.52 | 87.55 | 91.77 |
| OGRG-nodepth | 94.87 | 88.49 | 93.70 |
| **OGRG (Ours)** | **95.60** | **90.81** | **94.70** |



(a) Simulated novel objects (b) Real-world objects

(c) Training phase background textures

(d) Testing novel background textures from VIMA [39]

Fig. 4: Target objects used in both simulation and the real world include 32 instances for simulation and 15 for real-robot experiments. We apply domain randomization to the robot workspace during training for robust performances and use the textures from VIMA [39] in testing.

**RGS Results.** Table I presents the language-guided object grounding and robot grasping performance comparison with the selected baselines on the OCID-VLG dataset [18]. The proposed OGRG-RGS consistently outperforms all baselines across different backbone architectures. For the object grounding task, the method achieves a significant improvement of $+14.5\%$ mIoU compared to CROG. Additionally, the Bi-Aligner with depth fusion enhances grounding performance by $+3.08\%$ and $+0.73\%$ mIoU compared to LAVT and OGRG-nodepth. In terms of robot grasping performance, OGRG-RGS demonstrates a substantial improvement, achieving $+13.61\%$ $J@1$ over CROG. The ablation study further highlights that both the Bi-Aligner and depth fusion contribute significantly to the overall accuracy of grasp rectangle predictions. During model inference, OGRG is able to run on a single RTX 2080Ti GPU with an inference speed of 17.59 FPS.

### B. Referring Grasp Affordance with Weak Supervision

**Implementation Details:** Following the collection of the RGA dataset, both the OGRG-RGA grounding model and the MGN affordance prediction model are trained for 50 epochs, with batch sizes of 12 and 32, respectively. The AdamW optimizer is employed for the OGRG-RGA model, with an

TABLE II: **OGRG-RGA object grounding performance (oIoU) in simulation.**

| Baselines | Abs | Rel | Attr-cls | Attr-base | Avg |
|---|---|---|---|---|---|
| ETRG [17] | 93.67 | 84.35 | 92.85 | 91.28 | 90.54 |
| LAVT [26] | 95.62 | 85.83 | 95.34 | 94.24 | 92.76 |
| OGRG-nodepth | 95.88 | 84.59 | 95.55 | 94.91 | 92.73 |
| OGRG-db | 96.49 | 85.88 | 96.51 | **95.69** | 93.64 |
| **OGRG (Ours)** | **97.00** | **87.05** | **96.55** | 95.49 | **94.02** |

initial learning rate of $5 \times 10^{-5}$ and a polynomial learning rate decay. The maximum sentence length is set to 25 tokens. All RGA-related models are trained and tested on a single NVIDIA RTX 2080 Ti GPU.

**Evaluation Metrics:** The object grounding task is evaluated using the overall Intersection over Union (oIoU), similar to the metric used in RGS. For object grasping, the evaluation metric is defined as the object instance grasp success rate: $\frac{\text{\# of successful grasps on the correct target}}{\text{\# of total grasps}}$. In each test case, a single grasp attempt is executed. A grasp is considered successful only if the correct target object is grasped.

**OGRG-RGA Object Grounding in Simulation:** Test scenes were collected with varying numbers of objects, ranging from 1 to 7 per scene. For spatial attribute grounding experiments, the *Abs* and *Rel* settings correspond to absolute and relative spatial reasoning, respectively. A total of 1000 scenes were formulated, with and without object repetition (e.g., five objects consisting of three apples, one banana, and one tissue box). For general attribute grounding experiments, two language templates were used: *Attr-cls*, which includes color, shape, and category names, and *Attr-base*, which includes only color and shape attributes. These experiments were conducted on an additional 777 scenes.

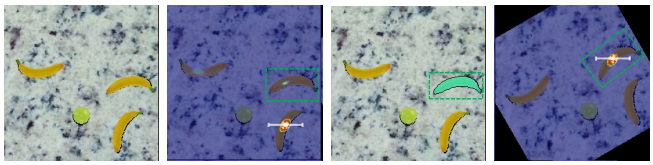
The object grounding results are presented in Table II. ETRG [17] is used as a baseline, modified to predict grounding masks with minor adjustments to its architecture. LAVT [26] uses a unidirectional aligner and no depth fusion; Variants of the OGRG-RGA model were also evaluated, OGRG-nodepth, which incorporates a bidirectional aligner without depth fusion; and OGRG-db, where the multimodal fused features $f^i_{back\_v}$ and $f^i_{back\_l}$ are passed directly to the next following aligner, similar to the depth branch in ETRG. The proposed OGRG-RGA model outperforms all baselines, achieving an average improvement of $+3.48\%$ in oIoU compared to ETRG.

**OGRG-RGA Robot Grasping in Simulation:** For testing, 1,600 test cases were created across 32 objects with pre-selected query language descriptions based on object attributes. In each test case, three identical object instances were randomly dropped into the workspace, and spatial language expressions were used to specify the target. For fair comparisons, all baselines were tested under identical scenes and language expressions.

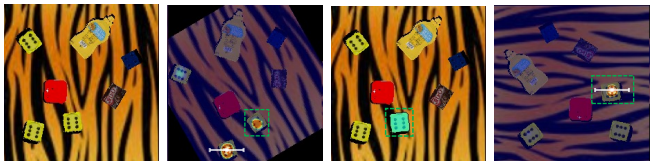
The robot grasping performance results are presented in Table III. The OGRG-RGA method significantly outperforms the state-of-the-art ETRG [17] grasp affordance method, achieving a $+4.86\%$ improvement in overall grasping performance on the spatial reasoning task. Detailed qualitative visualizations are shown in Fig. 5. As demonstrated, the ETRG method in Fig. 5a and Fig. 5b fails to localize the target and provide successful grasp poses, while the proposed OGRG-RGA accurately segments the language-referred target and predicts feasible grasp poses. Furthermore, the grasp affordance maps generated by ETRG (Fig. 5c) exhibit redundant high values on incorrect object candidates, whereas the proposed approach focuses directly on the correct target object.

TABLE III: **Grasp-success rates of OGRG-RGA in simulation**, reported for seen and unseen background (BG) conditions to demonstrate generalization capability.

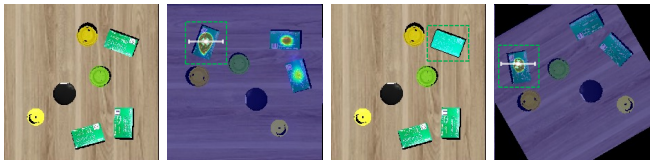
| BG | Baselines | Absolute | | Relative | AVG |
|---|---|---|---|---|---|
| | | 4-obj | 7-obj | 7-obj | |
| Seen | ETRG [17] | 90.75 | 88.38 | 86.56 | 88.56 |
| | OGRG-nodepth | 96.31 | 95.75 | 80.38 | 90.81 |
| | **OGRG (Ours)** | **96.50** | **96.88** | **86.88** | **93.42** |
| Unseen | ETRG | 92.06 | 91.00 | **88.56** | 90.54 |
| | **OGRG (Ours)** | **97.06** | **95.86** | 82.75 | **91.90** |



(a) Language input: pass me the banana that is to the middle right of the workspace



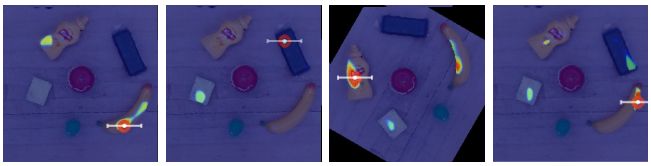
(b) Language input: grasp me the bottom center dice



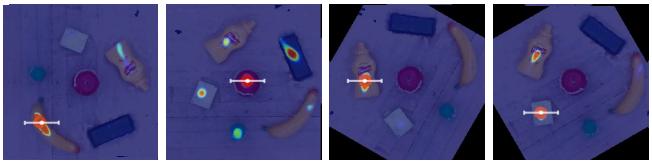
(c) Language input: grasp the tissue box that is to the upper right of the green cylinder green cup

Fig. 5: **Grounding masks and grasp affordances with spatial reasoning in simulation.** The green bounding boxes highlight the correct language-referred target object. The first column shows the input scene. The second column shows the affordance predictions from ETRG [17]. The third and fourth columns denote the grounding mask and grasp affordances from OGRG pipeline.

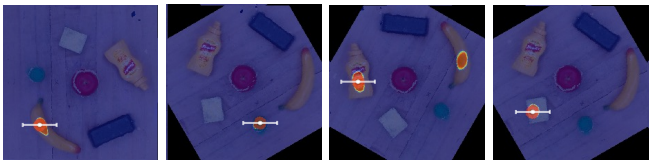
**OGRG-RGA Real Robot Grasping:** Real robot experiments were conducted using a Franka Emika Panda robot equipped with a FESTO DHAS soft gripper. As shown in Fig. 4b, 15 household objects were collected, and testing scenes were created by randomly sampling 6 target objects. A total of 100 visual-language-grasp triplets were manually collected in the real robot setup, and 246 grasping data points were generated after applying data augmentation [6]. For each baseline, all models were fine-tuned using the same augmented dataset. Comparing with ETRG [17] test scene setup, we parsed object candidates that have similar object attributes formulating more challenging real-robot scenes for



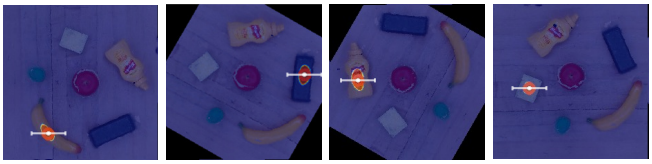
(a) Grasp affordances *Attribute-Grasp* [6]



(b) Grasp affordances *ETRG* [17]

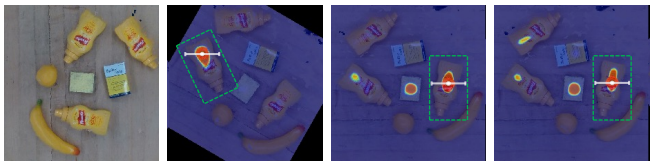


(c) Grasp affordances *OGRG-nodepth*

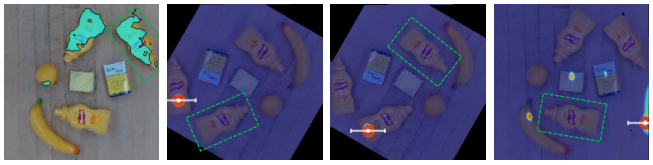


(d) Grasp affordances *OGRG* ***(Ours)***

Fig. 6: **Visualization of grasping affordances for real robot experiments with distinct challenging objects.** (a) to (d) show the grasp affordances maps from the RGA models: *Attribute-Grasp* [6], *ETRG* [17], *OGRG-nodepth*, and *OGRG (Ours)*, respectively. The language inputs from left to right columns are: yellow banana, black cuboid, mustard bottle, and yellow cuboid sponge.



(a) *ETRG* grasping affordances



(b) *OGRG-nodepth* grounding and grasping maps

(c) *OGRG* ***(Ours)*** grounding and grasping maps

Fig. 7: **Visualization of grounding masks and grasp affordance for challenging real robot spatial reasoning.** The green bounding boxes highlight the correct language-referred target object. The visualization in the first column shows the original scene arrangement and the grounding masks, while the rest of the three columns indicate the grasp affordances corresponding to spatial language descriptions on the same scene. The language expressions from left to right are: (1) grasp the upper right mustard bottle; (2) the mustard bottle that is to the top center of the workspace; (3) pass the mustard bottle that is to the lower right of the sponge; and (4) pass the mustard bottle below the sponge. Our proposed method demonstrates strong grounding capability, rapid adaptation, and accurate grasp pose predictions.

open-form language comprehension.

Fig. 6 illustrates the qualitative results on general object attribute reasoning with distinct objects. Compared to the state-of-the-art Attribute-Grasp method [6], the proposed OGRG-RGA approach accurately localizes the target and generates precise grasp poses with distinct rotation angles (Fig. 6d). In contrast, Attribute-Grasp (Fig. 6a) struggles when object candidates have similar colors or shapes as the target. Fig. 7 presents visualizations from challenging real robot spatial reasoning experiments. The OGRG-RGA method (Fig. 7c) produces clean and correctly focused grounding and grasping affordance maps. While ETRG (Fig. 7a) successfully localizes language-referred objects, it predicts high affordance values on incorrect objects. Additionally, OGRG-nodepth fails in both grounding and grasping tasks under these conditions.

Across 24 grasp attempts for each baseline, the qualitative evaluation of real robot grasping success rates is presented in Table IV. Due to the challenge of obtaining ground-truth target object masks in the real robot setup, a grounding accuracy metric is introduced to evaluate whether the grasping motion in each scene aligns with the target object. The proposed OGRG-RGA model achieves the highest grounding and grasping success rates, validating the effectiveness of its design, including the Bi-Aligner and depth fusion components. Despite the small dataset size used for fine-tuning, the model efficiently adapts to new scenes with novel objects.

TABLE IV: **OGRG-RGA real robot experiment results.** We evaluate the grounding accuracy and the grasping success rate under different scenes with challenging objects.

| Methods | Grounding (%) | Grasp Succ. (%) |
|---|---|---|
| ETRG [17] | 75.0 | 62.5 |
| OGRG-nodepth | 75.0 | 33.3 |
| **OGRG (Ours)** | **87.5** | **70.8** |

## V. Conclusion

We introduced OGRG, a framework that aligns visual and language features through a bidirectional aligner without relying on pre-aligned vision–language models. By incorporating depth fusion, OGRG supports open-form, attribute-based grounding and demonstrates rapid adaptability to new scenes and novel objects. On both RGS and RGA benchmarks, it achieves competitive performance in language-guided grounding and grasping. Although our experiments focus on planar grasps using a parallel-jaw gripper, the proposed approach is embodiment-agnostic and can be transferred to humanoid tabletop manipulation tasks. Furthermore, OGRG facilitates cross-embodiment grasping data collection by reusing language-grounded supervision across different end-effectors.

While OGRG achieves strong results on RGS and RGA tasks, our current evaluation is limited to common household objects with regular geometries in planar grasping settings. The experiments assume a fixed camera viewpoint, a parallel-jaw gripper, and relatively uncluttered tabletop environments. Addressing scenarios involving complex object geometries, severe occlusions, and diverse gripper configurations remains an important direction for future work.

## REFERENCES


[1] X. Lou, H. Yu, R. Worobel, Y. Yang, and C. Choi, "Adversarial object rearrangement in constrained environments with heterogeneous graph neural networks," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1008–1015.

[2] H. Yu and C. Choi, "Self-supervised interactive object segmentation through a singulation-and-grasping approach," in *European Conference on Computer Vision*. Springer, 2022, pp. 621–637.

[3] H. Yu, X. Lou, Y. Yang, and C. Choi, "Iosg: Image-driven object searching and grasping," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 3145–3152.

[4] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9626–9633.

[5] A. D. Vuong, M. N. Vu, B. Huang, N. Nguyen, H. Le, T. Vo, and A. Nguyen, "Language-driven grasp detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 902–17 912.

[6] Y. Yang, H. Yu, X. Lou, Y. Liu, and C. Choi, "Attribute-based robotic grasping with data-efficient adaptation," *IEEE Transactions on Robotics*, 2024.

[7] 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)*. IEEE, 2023, pp. 11 597–11 604.

[8] H. Ahn, O. Kwon, K. Kim, J. Jeong, H. Jun, H. Lee, D. Lee, and S. Oh, "Visually grounding language instruction for history-dependent manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 675–682.

[9] S. Jin, J. Xu, Y. Lei, and L. Zhang, "Reasoning grasping via multimodal large language model," *arXiv preprint arXiv:2402.06798*, 2024.

[10] 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.

[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] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," *arXiv preprint arXiv:2407.08693*, 2024.

[13] G. Tziafas and H. Kasaei, "Towards open-world grasping with large vision-language models," *8th Conference on Robot Learning (CoRL 2024)*, 2024.

[14] Y. Qian, X. Zhu, O. Biza, S. Jiang, L. Zhao, H. Huang, Y. Qi, and R. Platt, "Thinkgrasp: A vision-language system for strategic part grasping in clutter," *arXiv preprint arXiv:2407.11298*, 2024.

[15] 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.

[16] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[17] H. Yu, M. Li, A. Rezazadeh, Y. Yang, and C. Choi, "A parameter-efficient tuning framework for language-guided object grounding and robot grasping," *arXiv preprint arXiv:2409.19457*, 2024.

[18] G. Tziafas, X. Yucheng, A. Goel, M. Kasaei, Z. Li, and H. Kasaei, "Language-guided robot grasping: Clip-based referring grasp synthesis in clutter," in *7th Annual Conference on Robot Learning*, 2023.

[19] O. Mees, J. Borja-Diaz, and W. Burgard, "Grounding language with visual affordances over unstructured data," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 576–11 582.

[20] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 108–124.

[21] R. Li, K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia, "Referring image segmentation via recurrent refinement networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5745–5753.

[22] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, "Recurrent multimodal interaction for referring image segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1271–1280.

[23] Z. Xu, Z. Chen, Y. Zhang, Y. Song, X. Wan, and G. Li, "Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 503–17 512.

[24] Y. Wang, J. Li, X. Zhang, B. Shi, C. Li, W. Dai, H. Xiong, and Q. Tian, "Barleria: An efficient tuning framework for referring image segmentation," in *The Twelfth International Conference on Learning Representations*.

[25] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu, "Cris: Clip-driven referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 686–11 695.

[26] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 155–18 165.

[27] J. Tang, G. Zheng, C. Shi, and S. Yang, "Contrastive grouping with transformer for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 570–23 580.

[28] Y. Hu, Q. Wang, W. Shao, E. Xie, Z. Li, J. Han, and P. Luo, "Beyond one-to-one: Rethinking the referring image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4067–4077.

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[30] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[31] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[32] 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.

[33] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[35] E. Rohmer, S. P. N. Singh, and M. Freese, "V-rep: A versatile and scalable robot simulation framework," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 1321–1326.

[36] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*. IEEE, 2015, pp. 510–517.

[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)*. IEEE, 2023, pp. 976–983.

[38] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[39] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: General robot manipulation with multimodal prompts," in *Fortieth International Conference on Machine Learning*, 2023.