

RESEARCH ARTICLE

Multi-modal interaction with transformers: bridging robots and human with natural language

Shaochen Wang¹ , Zhangli Zhou¹, Bin Li², Zhijun Li^{3,4} and Zhen Kan¹

¹Department of Automation, University of Science and Technology of China, Hefei, 230026, China, ²Department of Electronic and Information Science, University of Science and Technology of China, Hefei, 230026, China, ³Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, 230031, China, and ⁴School of Mechanical Engineering, Tongji University, Shanghai, 200092, China

Corresponding author: Zhen Kan; Email: zkan@ustc.edu.cn

Received: 24 May 2023; **Revised:** 25 September 2023; **Accepted:** 12 October 2023;

First published online: 13 November 2023

Keywords: multi-modal robot perception; robotic grasping; human-robot interaction

Abstract

The language-guided visual robotic grasping task focuses on enabling robots to grasp objects based on human language instructions. However, real-world human-robot collaboration tasks often involve situations with ambiguous language instructions and complex scenarios. These challenges arise in the understanding of linguistic queries, discrimination of key concepts in visual and language information, and generation of executable grasping configurations for the robot's end-effector. To overcome these challenges, we propose a novel multi-modal transformer-based framework in this study, which assists robots in localizing spatial interactions of objects using text queries and visual sensing. This framework facilitates object grasping in accordance with human instructions. Our developed framework consists of two main components. First, a visual-linguistic transformer encoder is employed to model multi-modal interactions for objects referred to in the text. Second, the framework performs joint spatial localization and grasping. Extensive ablation studies have been conducted on multiple datasets to evaluate the advantages of each component in our model. Additionally, physical experiments have been performed with natural language-driven human-robot interactions on a physical robot to validate the practicality of our approach.

1. Introduction

In today's rapidly evolving world, robots [1, 2] are progressively integrating into our daily lives, excelling in tasks ranging from household assistance [3] to industrial operations [4]. As robots diversify their applications, the need for more intuitive and natural human-robot communication becomes paramount. Traditional human-robot interfaces [5], however, impose significant constraints, often requiring specialized training and proficiency in complex programming skills. These limitations hinder not only the widespread adoption of robots but also efficient collaboration between humans and their mechanical counterparts. To address this, there is a growing demand for robots to engage in natural communication with humans, with natural language emerging as an attractive option due to its convenience, directness, and user-friendly nature compared to traditional interfaces like keyboards, mice, and programming languages.

This work emphasizes the potential of harnessing natural language to facilitate seamless interactions between humans and robots. By leveraging the power of language, our aim is to bridge the communication gap and enhance the collaborative capabilities of robots across a variety of real-world scenarios. It is worth noting that humans rely on vision for processing and perceiving over 70% of information [6]. Therefore, relying solely on language may prove insufficient for robots attempting to perceive and understand their surroundings. Consequently, the integration of visual and language modalities is emerging as a promising research area in the field of human-robot interaction. Just like humans, language helps

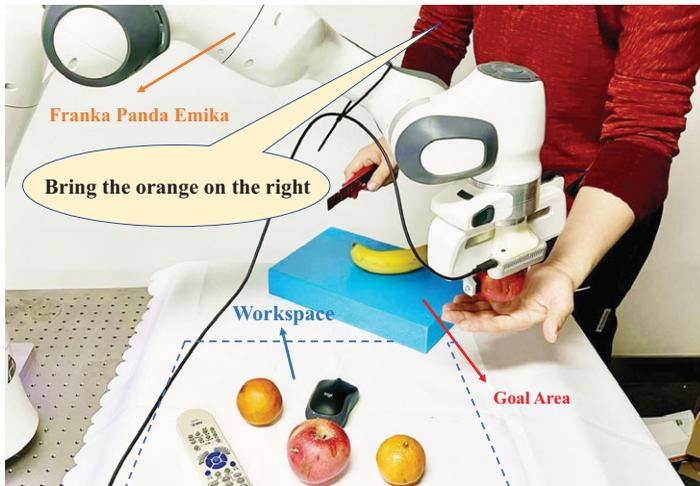


Figure 1. The robot assists in fruit grasping based on user instructions.

robots capture higher-level context and task intent, while vision provides detailed information about the environment and objects. Combining these modalities empowers robots to comprehend both the “what” (object recognition) and the “why” (user intent) of a task. This integration of language and vision enhances natural and intuitive communication between humans and robots. Furthermore, by incorporating both language and vision, robots become more versatile in comprehending and adapting to diverse tasks. Language enables users to convey complex and dynamic instructions, while vision aids in adapting to changing environments and identifying objects not explicitly mentioned in the commands. Different modalities can complement each other, compensating for limitations in one modality with information from the other. For example, if an object is occluded from the robot’s view, language input may help the robot infer missing information.

However, the inherent ambiguity and complexity of human language, coupled with the captured visual information, poses a significant challenge for robots to achieve effective joint reasoning between these two data streams. For instance, consider a language query instructing the robot to grasp the orange on the right, as depicted in Fig. 1. This task becomes challenging when the desk is cluttered with unrelated objects like a mouse, an apple, and a remote, alongside multiple oranges. The model must not only accurately identify and locate these objects but also comprehend their spatial relationships. Furthermore, depending on the region of interest (ROI) specified by the robot, it must generate executable manipulations for the end-effector to grasp objects that align with the desired shape, size, and color specified by the user’s instructions.

In the context of multi-modal robot learning, previous research [7] has focused on parsing textual and perceptual information using manually designed rules. However, recent studies have witnessed a gradual transition from traditional language representation approaches to the adoption of deep learning methodologies. For example, Chen et al. [8] employed ResNet for visual information processing and long short-term memory (LSTM) for textual input handling. Nonetheless, their approach is limited in handling only simple instructions due to constrained inner feature alignment. Another approach by Lin et al. [9] utilized the powerful language model BERT [10] for manipulation tasks; however, it still struggles with understanding spatial relationships between objects. On the other hand, significant advancements have been made in vision-language understanding tasks, such as image captioning and visual grounding. Regrettably, this line of research lacks a coherent mapping from language command comprehension to physical robot actions.

The goal of this work is to bridge the gap between current vision-language models and human-robot interaction models. Although natural language is the most natural form of communication with humans, existing language interaction-based robot models still struggle with understanding spatial relationships among objects or rely on a limited set of instructions [11]. To tackle these challenges, our framework

incorporates the following three key ideas. *Semantic Parsing*: The visual-linguistic encoder extracts meaningful information from the visual scene and natural language instructions, where the visual branch performs visual scene perception and the linguistic branch breaks down the given instructions into actionable commands, taking into account linguistic cues that imply spatial relationships. *Cross-Modal Embeddings*: Our model highlights the significance of cross-modal embeddings, where linguistic cues are aligned with visual cues. This enables the robot to associate words with corresponding visual features, aiding in the recognition of spatial arrangements. *Contextual Understanding*: Our model integrates the cross-attention mechanism to reinforce contextual understanding, thereby enabling the model to inherently capture and interpret spatial relationships.

In this paper, we present a unified transformer framework designed to effectively map natural language commands and raw visual observations to executable actions performed by a robotic manipulator. Our framework incorporates a vision and language transformer to process image and linguistic instructions. Through multi-stage contextual information reasoning, we fuse aggregated visual and linguistic token representations in the encoder stage to identify discriminative features. Specifically, the model utilizes textual embeddings to query visual representations using the cross-attention mechanism, enabling focused attention on regions relevant to the given language expressions. Moreover, a grasping decoder leverages progressive attention layer propagation to determine the grasping configuration parameters of the robotic gripper. Experimental results demonstrate the significant improvements achieved by our model on mainstream benchmarks for both visual-linguistic understanding and visual grasping. Furthermore, we validate the effectiveness of our approach in real-world applications by conducting physical experiments using a Franka Panda robot, in addition to numerical simulations.

The highlights of using a unified transformer for visual-linguistic robotic understanding can be summarized as follows: (1) Token-based representations facilitate a seamless alignment of visual and linguistic features within the robotic semantic space. (2) The attention-based context mechanism empowers robots with a comprehensive perception of the scene, enabling them to achieve more efficient and reliable grasping in unstructured and cluttered environments. (3) Adopting a unified modeling perspective enables a more effective mapping between the perception space of robots and their corresponding physical action space.

In a nutshell, the contributions of this paper can be summarized in three folds:

- This work presents a novel approach to multi-modal human-robot interface tailored to the domain of robotic grasping. To the best of our knowledge, our work represents one of the pioneering attempts to leverage multi-modal transformers, enabling the integration of vision-language understanding with physical robotic manipulations.
- We propose an elegant framework for seamlessly combining visual and linguistic information in multi-modal human-robot interaction tasks. Remarkably, our model facilitates robots in comprehending the semantics of complex natural language instructions, including spatial relationships between objects and color information.
- Extensive experiments have been conducted to validate the effectiveness of our approach, demonstrating significant improvements across several widely used benchmarks.

2. Related work

2.1. Visual grasping and human-robot interaction

Reliable grasping plays a vital role in numerous applications [12, 13]. Recent studies [14–15] have focused on leveraging deep learning techniques for visual grasping. Notably, Lenz et al. [16] were the pioneers in neural networks as classifiers for grasp prediction. Building upon this, Guo et al. [17] extended the model by integrating tactile information to enhance grasping performance. Our studies encompass a range of tasks including transformer-based visual grasping [18], grasp planning [19], and unsupervised grasp detection [20]. These research endeavors have significantly advanced our

comprehension of how robots perceive and interact with their surroundings. Furthermore, our research endeavors have delved into transformer-based models, including their applications in visual grasping [18] and depth prediction [21]. Our profound familiarity with these models has paved the way for our exploration of their adaptability to vision-language models tailored for robotic applications. As robotics continues to permeate everyday life, it becomes increasingly important to develop natural and user-friendly ways of interacting with robots. Traditional control methods, such as mouse, keyboard, and touchscreen interfaces, often require users to issue complex commands that can be challenging for non-experts in robotics. Wang et al. [22] have proposed a sight-based robotic arm assistance system that enables users to operate robots even when their upper limbs are injured or occupied. In contrast, natural language-based interfaces offer a user-friendly approach with robots, and reduced effort required to educate and familiarize the personnel or users who will be interacting with the robot. However, existing methods, like the one presented by Chen et al. [8], employed separate model architectures, utilizing a convolutional neural network (CNN) for visual feature extraction and an LSTM to capture textual information. Unfortunately, these distinct modalities struggle to align their semantic meaning accurately within different model architectures.

2.2. Multi-modal vision and language understanding

When humans perceive the world, they naturally integrate and align information from various sources, such as visual and audio cues. In the field of artificial intelligence (AI), there is a growing interest in enabling AI systems [23, 24] to learn from multi-modal data. Recently, there has been growing interest in perceptual models that simultaneously incorporate visual and linguistic signals [25–26]. These systems manipulate images to generate corresponding text, as seen in image caption tasks, or engage in vision-language interaction tasks like visual question answering. Among these tasks, visual grounding, which involves localizing regions based on verbal descriptions, is particularly relevant to our objective. Existing approaches to tackle this task typically follow either a two-phase or one-phase pipeline. Two-phase approaches [27–28] initially detect a set of region proposals, which are then matched against linguistic queries to identify the top-ranked proposals. Within the two-stage methods, several appealing techniques have been developed to enhance the modeling of multi-modal relationships, such as modular attention networks [29] and scene graphs [23]. On the other hand, one-stage approaches [30–31] integrate text information with image features to directly produce dense predictions. For instance, Fast and Accurate One-stage Approach (FAOA) [32] incorporates linguistic features into each spatial position of visual feature maps by concatenating them. Distinguishing itself from prior works, our study tailors the vision-language model to suit robotic tasks in the physical world.

2.3. Transformer

The attention architecture [33] has demonstrated significant potential in sequence modeling and machine translation tasks. As an attention-based model, transformers [10] have emerged as the predominant approach in natural language processing. Furthermore, transformers have also made notable contributions to computer vision tasks. An exemplary work, DETection TRansformer (DETR) [34], formulates object detection as a set prediction problem by employing learnable queries and a contextual attention mechanism to capture object relationships. In a related study, Wang et al. [18] adopt a pure transformer architecture to capture global relationships within a grasping scene using attention mechanisms, resulting in improved grasp detection.

3. Method

3.1. Problem formulation

Our objective is to develop a flexible linguistic-based human-robot interface that enables seamless understanding of human intentions by robots. Our system finds applications in scenarios such as industrial

assembly lines, where robots follow human instructions to perform assembly tasks, or situations where individuals with impairments employ language to control a robotic arm for handling inconvenient tasks. Let I represent the captured image and L denote the user's instruction. The ultimate goal is to learn a function f that maps the image and user commands to a grasp representation, satisfying the user's specifications:

$$g = f(I, L). \quad (1)$$

A grasp configuration, denoted as g , can be represented in the image space as a 5-dimensional tuple: $g = \{x, y, \theta, w, h\}$. Here, (x, y) represents the grasping center, θ denotes the orientation angle of the gripper, and w and h indicate the width and height of the parallel gripper when opened, respectively. In the context of robotic visual grounding tasks, there are additional constraints compared to common visual grounding tasks. For instance, due to the limitations of the physical gripper, the angle of the grasping rectangle must fall within the range of $[-\frac{\pi}{2}, \frac{\pi}{2}]$, and the opening width of the gripper cannot exceed its predefined limit.

3.2. Architecture overview

In Fig. 1, our framework comprises four essential components: (1) **Linguistic Branch**, (2) **Visual Branch**, (3) **Language-Conditioned Visual Fusion Module**, and (4) **Fine-Grained Grasping Module**. Before introducing our framework, we address a fundamental challenge in grounded language-based robot understanding: establishing a connection between human language and the visual perceptual world in which the robot operates. To tackle this challenge, we break it down into two sub-problems. Firstly, we aim to construct a unified semantic space that bridges the gap between visual and linguistic concepts. Secondly, we endeavor to establish a mapping between these semantic concepts and the corresponding robotic actions in the physical world. Our approach employs a cascaded encoder-decoder architecture. Initially, the RGB image and the human language query are individually processed through the visual and text branches, respectively. The tokenized features extracted from both branches are then fused using the language-conditioned fusion module to derive the corresponding bounding region. Subsequently, this bounding region is integrated with the scene depth image and provided as input to the grasping decoder, which generates the optimal grasp configuration.

In summary, our visual-linguistic robot grasping system offers the following key features:

Alignment: Leveraging transformer-based models to seamlessly integrate visual and linguistic modalities, facilitating a unified approach to robotic tasks.

Representation: Utilizing homogeneous token-based representations to enhance interaction between visual concepts and textual entities within the system.

Co-learning: Incorporating synchronized learning of visual-verbal understanding and robotic grasping, enabling the establishment of a mapping from the human-interactive world to corresponding robot actions.

The general framework is broadly presented in the following three parts.

3.2.1. Text and vision encoder

As shown in Fig. 2, the input consists of images and language instructions, which are processed separately in two branches. In the visual branch, a convolutional neural network (CNN) backbone is utilized to extract visual features. To facilitate the model's understanding of the relationship between text and images, we propose using a unified model, such as a transformer, to process both modalities. In this unified model, image information is treated as word-tokens within the transformer model. Following the CNN backbone, we employ a stack of attention layers to handle the visual features. As an illustration, let's consider the process of passing an image $I_0 \in \mathbb{R}^{3 \times H \times W}$ through a backbone network, resulting in a 2-dimensional feature map $z \in \mathbb{R}^{C \times H_z \times W_z}$. Here, C represents the output channel of the feature map, while H_z and W_z denote the dimensions of the feature map. To ensure compatibility with the following

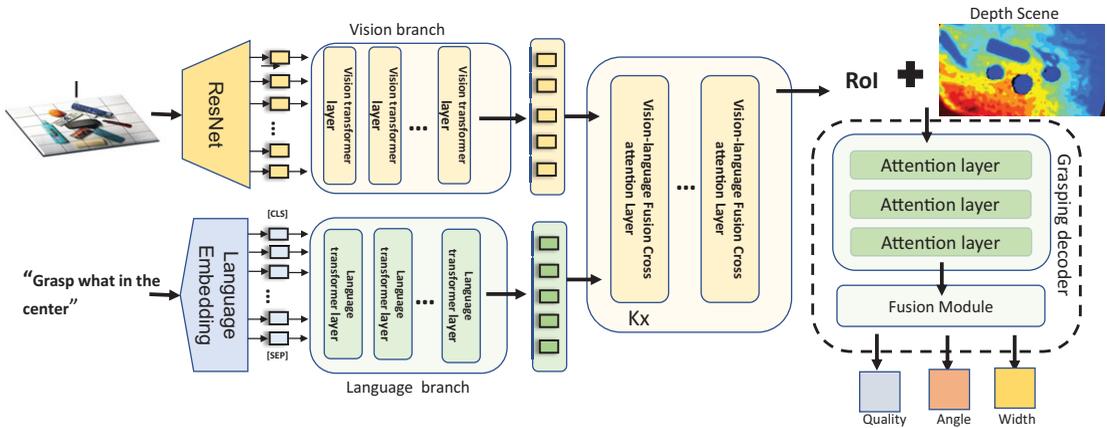


Figure 2. The overview of the framework. The language and visual encoders are served as prefixes to the fusion decoder, and the resulting grounding information is then combined with the scene depth image as input to the grasping model to produce the grasping configuration according to language instructions.

attention layer, the internal variable is flattened into $z \in \mathbb{R}^{C \times L}$, where $L = H_z \times W_z$ represents the size of the input tokens. By employing visual tokens in the vision transformer branch, global features are captured by focusing on multiple areas across the entire image.

In parallel to the vision branch, the language branch operates as a sibling branch that receives natural language instructions as input. The linguistic branch comprises a token embedding layer and a series of stacked transformer layers. In order to leverage the benefits of pre-trained BERT model [10], the design is inspired by the BERT architecture, with each linguistic transformer layer having an output channel dimension of 768. In terms of linguistic expressions, each word in a sentence is transformed into a one-hot vector, which is then passed through an embedding layer to obtain the corresponding token-based representation. Furthermore, our approach aligns with machine translation techniques, wherein [CLS] and [SEP] tokens are inserted at the beginning and end of each sentence. The resulting latent embeddings are subsequently inputted into linguistic transformer layers, generating the corresponding features denoted as $z_l \in \mathbb{R}^{C \times N_l}$. Notably, the image features are also treated as word-tokens to establish a homogeneous token-based representation shared between linguistic and visual features, allowing different modal information to be embedded into a joint semantic space.

3.2.2. Language-conditioned fusion

The language-conditioned fusion module refines visual features through textual embedding using the latent features obtained from the previous two branches. This refinement process helps identify regions that are relevant to the referring instructions. Instead of directly feeding visual and textual information into the grasping decoder, we propose a multi-stage language-guided fusion approach that progressively gathers contextual information from verbal and visual features to facilitate the localization of the referred object. Below we briefly recall the attention mechanism. An input sequence $X \in \mathbb{R}^{n \times d}$ is linearly transformed to obtain three successive vectors namely (query Q , key K , and value V) in shaping the interaction between the input data and the attention mechanism, where n and d is the length and dimension of the input x . The vectors are computed via

$$Q = XW_Q, K = XW_K, V = XW_V, \tag{2}$$

where $W_Q \in \mathbb{R}^{d \times d_q}$, $W_K \in \mathbb{R}^{d \times d_k}$, $W_V \in \mathbb{R}^{d \times d_v}$ are three linear projection matrices, which refer to learnable weight matrices used to transform the input data (queries, keys, and values) before applying the attention mechanism. In this work, we have $d_q = d_k = d_v = d$. The attention is then calculated using the formula:

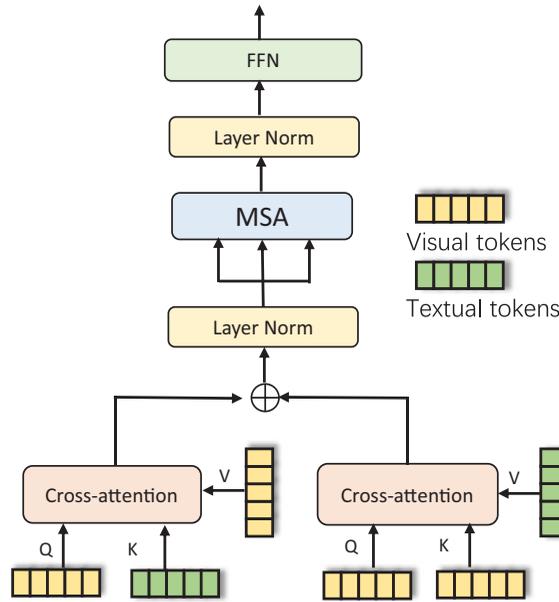


Figure 3. An illustration of the vision-language fusion cross-attention module.

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V, \tag{3}$$

where the attention is obtained by taking the dot-product between query Q and key K , and \sqrt{d} serves as a scaling factor. Finally, a softmax operator is applied to obtain a normalized distribution that is assigned to value V .

As illustrated in Fig. 3, the input visual tokens serve as the query and value, while the input textual tokens act as the key in the cross-attention mechanism. In the attention layers, the data is processed through a multi-head self-attention (MSA). This involves splitting the input embeddings into multiple “heads” or parallel subspaces. Each head computes attention scores between different elements of the input sequence, capturing dependencies and relationships. Following the MSA step, the data is passed through a feed-forward network. This process gathers visual information relevant to the text. Similarly, in parallel, the visual tokens serve as the key and query, while the textual tokens are used as the value in another cross-attention operation. This dual cross-attention mechanism enhances the textual-related visual features by computing their correlation, effectively suppressing irrelevant visual and textual information. The features obtained from the dual cross-attention are then aggregated and passed through a normalization layer to ensure internal alignment. In the subsequent fusion stage, the output features serve as the input for cross-model reasoning. Through our cross-model attention fusion approach, the model progressively eliminates redundant information in visual features using linguistic queries, enabling the model to focus more on the regions crucial for human grasping.

3.2.3. Grasping decoder

The grasping decoder takes a depth image and a coarse location region specified by a language expression to accurately estimate the precise grasping pose. To achieve this, we convert the task of planar grasp detection into pixel-level prediction tasks. We employ a series of attention layers that capture the interrelationship between different parts of the object, thereby identifying regions suitable for grasping. These attention layers enable us to retrieve relevant contextual information. Subsequently, a feature pyramid

fusion module is employed to incorporate the contextual features gathered from the previous attention layers. We refer to the set of grasps in the image space as the grasp map, which is denoted as

$$\mathbf{G} = (Q, \Theta, W) \in \mathbb{R}^{3 \times H \times W} \quad (4)$$

The grasp map \mathbf{G} estimates the parameters of a set of grasps, executed at the Cartesian point \mathbf{p} , corresponding to each pixel. We represent the grasp map \mathbf{G} as a set of images:

- Q is an image describing the quality of a grasp executed at each point (x, y) . The value is a scalar within the range $[0, 1]$ where a value closer to 1 indicates higher grasp quality, i.e. higher chance of successful grasp.
- Θ represents the grasp angle that should be executed at each point. Given the symmetry of the antipodal grasp around $\pm \frac{\pi}{2}$ radians, the angles are provided in the range of $[-\frac{\pi}{2}, \frac{\pi}{2}]$.
- W indicates the width of the gripper to be employed at each point. To achieve depth invariance, we set the range of the variable w to $[0, 150]$ pixels, which can be converted to a physical measurement utilizing the depth camera parameters and the measured depth.

Finally, three grasping heads are connected in parallel to the top of the fusion layer to estimate the grasping score estimation head Q , the gripper angle estimation head Θ , and the gripper width head W . Each head generates a heatmap of the same dimensions as the input depth image. The feature maps obtained from each stage of the fusion module are resized to match the resolution of the final output. These feature maps are then passed through convolutional layers with 1×1 kernels to generate the three grasp heatmaps. Each position in the grasping score heatmap Q produces a value between 0 and 1, indicating the likelihood of successful grasping at that particular pixel. Similarly, the width and angle heads provide information about the gripper's width and rotation angle during the grasping process. Instead of sampling the input image to create grasp candidates, the grasp point \mathbf{g}' is determined by the highest confidence score in the grasping quality heatmap Q . Mathematically, $\mathcal{G}_{\text{pos}} = \arg \max Q$, signifying the pixel location where successful grasping is most likely to occur. The predicted angle θ and the corresponding orientation width w are obtained from the angle Θ and width W heatmaps.

3.3. Training optimization

The optimization of the entire architecture consists of two stages. Firstly, the visual-linguistic transformer encoder is responsible for the rough localization of the referred object by capturing the relationships between visual and language tokens. Secondly, the grasping decoder generates precise grasping parameters based on the position and shape of the referred object.

During the visual grounding procedure, the model directly calculates the regression loss between the predicted bounding box \hat{b} and the ground truth bounding box b . The training objective is defined as:

$$\mathcal{L}_{\text{vg}} = \lambda_1 L_{l1} + \lambda_2 L_{\text{GIoU}}(b), \quad (5)$$

where the loss is composed of two terms: L_{l1} and the GIoU (Generalized Intersection over Union) loss denoted by L_{giou} . The hyperparameters λ_1 and λ_2 control the relative contribution of each loss term. b refers to the ground truth bounding box, which represents the actual location and size of an object in the image. \hat{b} represents the predicted bounding box, which is the model's estimated location and size for the object.

During the optimization process of grasping, we approach the problem of estimating the grasping pose as a regression problem. To achieve this, we utilize a grasping decoder, which establishes a mapping from the designated region of interest to the corresponding executable grasping configuration parameters. The objective is to minimize the L_2 distance between the grasping heatmaps and the ground truth by defining a loss function as follows:

$$\mathcal{L}_{\text{grasp}} = L_{\text{grasp-score}} + L_{\text{angle}} + L_{\text{width}}. \quad (6)$$

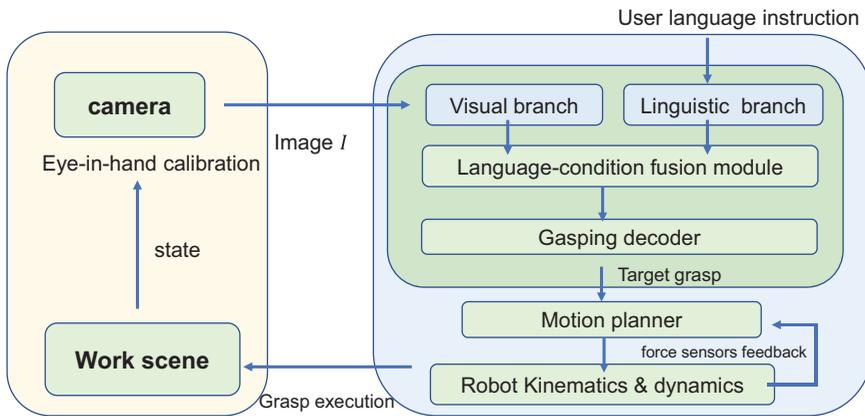


Figure 4. System diagram of our architecture.

The terms $L_{\text{grasp-score}}$, L_{angle} , and L_{width} represent individual loss components related to specific aspects of the robotic manipulation task. $L_{\text{grasp-score}}$ is a loss term related to the grasp quality. It measures how well the predicted grasp aligns with the actual grasp location on the object. L_{angle} is the loss related to the orientation or angle of the object, indicating how accurately the model predicts the object’s orientation. L_{width} is the loss associated with the width of the gripper, indicating how well the model estimates the size of the object. For each component of the loss function, \mathcal{L}_i is the mean square error between the corresponding value of the model and the ground truth. For instance, the first term of $\mathcal{L}_{\text{grasp}}$ is defined as $\mathcal{L}_{\text{grasp-score}} = \sum_{i=1}^N \|\tilde{G}_i - G_i^*\|^2$, where \tilde{G}_i is the output of the grasp quality head and G_i^* is the corresponding ground truth.

4. System design

4.1. Robotic grasping setting

This section provides an overview of the implementation details of our robotic system. The configuration of the entire system is depicted in Fig. 4. We utilize a Franka Emika Panda Gen 1 robot equipped with a two-finger parallel gripper for performing grasping tasks. The robot boasts 7 degrees of freedom and can handle a maximum payload of 3 kg. To enable real-time grasp detection, an Intel RealSense D435 camera is mounted on the robot’s wrist. The Panda robot has an operational range of 85.5 cm. Additionally, a desktop computer with a GPU is connected to the robot to execute our model and send commands for robot manipulation. The training process is conducted on the Ubuntu 18.04 desktop with Intel Core i9 CPU and 4 NVIDIA 3090 GPU. Training the vision-language transformer-based robotic manipulation system took approximately 15 h.

4.2. Grasp execution pipeline

The grasping flow is visualized in Fig. 4. Upon camera calibration, the captured scene image, along with the user’s language instruction, is inputted into the model. Subsequently, the model generates grasping configuration parameters, which are then forwarded to the controller for grasping trajectory planning. In this planning phase, the position information of the grasping center is transformed from image space to the world coordinate system. During the grasping process, the robot follows a series of steps. Initially, it moves to the designated initial position while keeping the gripper aligned with the specified grasping orientation. Then, the robot commences the grasping operation based on the trajectory calculated through inverse kinematics. This process continues until successful grasping is achieved or a collision is detected, which depends on the feedback from contact force sensors.

A successful grasp is achieved as the robot successfully gripping the object specified by the user's language instruction and subsequently placing it in the desired location.

4.3. Grasping scenarios

In order to thoroughly evaluate the effectiveness of our model, we conduct real robot experiments using a diverse range of everyday objects. These objects encompass both familiar items present in the datasets and entirely unseen objects. The object categories include fruits, tools, and more. To assess the robustness of our grasping system, the grasped objects exhibit various shapes and sizes. During each trial, the objects are randomly positioned within the robot's working space. This random placement ensures that the grasping system encounters different spatial configurations, contributing to a comprehensive assessment of its performance.

5. Experiment

In this section, we evaluate our model in both datasets and real-world applications. Experiments are carried out to investigate (1) the performance of integrating visual-linguistic transformer models in building effective and general multi-modal human-robot interfaces, and (2) the applicability of the transformer-based approach to physical robotic systems.

5.1. Dataset and evaluation metric

To facilitate comparison with other methods, we train and test our model on two main grasping datasets: the *Cornell* dataset [16] and the *Jacquard* dataset [35]. The *Cornell* dataset [16] consists of more than 800 RGB-D images, containing 240 graspable objects. The *Jacquard* dataset, on the other hand, is a large synthetic grasp dataset comprising over 50,000 images and featuring 11,000 objects. Its synthetic nature provides a diverse and extensive collection of data for evaluation purposes. In addition to grasping evaluation, we also assess the visual-linguistic understanding ability of our model using the *RefCOCO/RefCOCO+/RefCOCOg* benchmarks. The *RefCOCO* dataset [28], *RefCOCO+* dataset [28], and *RefCOCOg* dataset [27] are widely recognized visual grounding benchmarks. These datasets utilize images from the MS COCO dataset and provide annotations for referring expressions. The dataset is divided into training, validation, and test sets, with the test set further split into testA and testB subsets. Furthermore, we utilize the *Flickr30K Entities* dataset [36], which extends the original Flickr30K dataset [37] by incorporating phrase annotations.

5.2. Evaluation metrics

To make a fair comparison with previous grasp detection works [38–39], this work also adopts the rectangle criterion to evaluate the grasping quality. A predicted grasp is regarded as a success when the following conditions are satisfied. (1) The orientation angle difference between the predicted grasp and the ground truth is under 30° . (2) The Jacquard index of the predicted grasping rectangle and the ground truth should be greater than 25%.

5.3. Implementation details

The entire architecture is optimized using the AdamW optimizer and the batch size is set to 64. The parameters of the language branch in our model are initialized using pre-trained BERT [10]. The visual branch is initialized using the weights from pre-trained DETR [34]. The sentence length processed by the linguistic branch is fixed at 40. For shorter expressions, empty tokens are padded to match the sentence length. Longer sentences, exceeding a length of 40, are split up into smaller segments. The

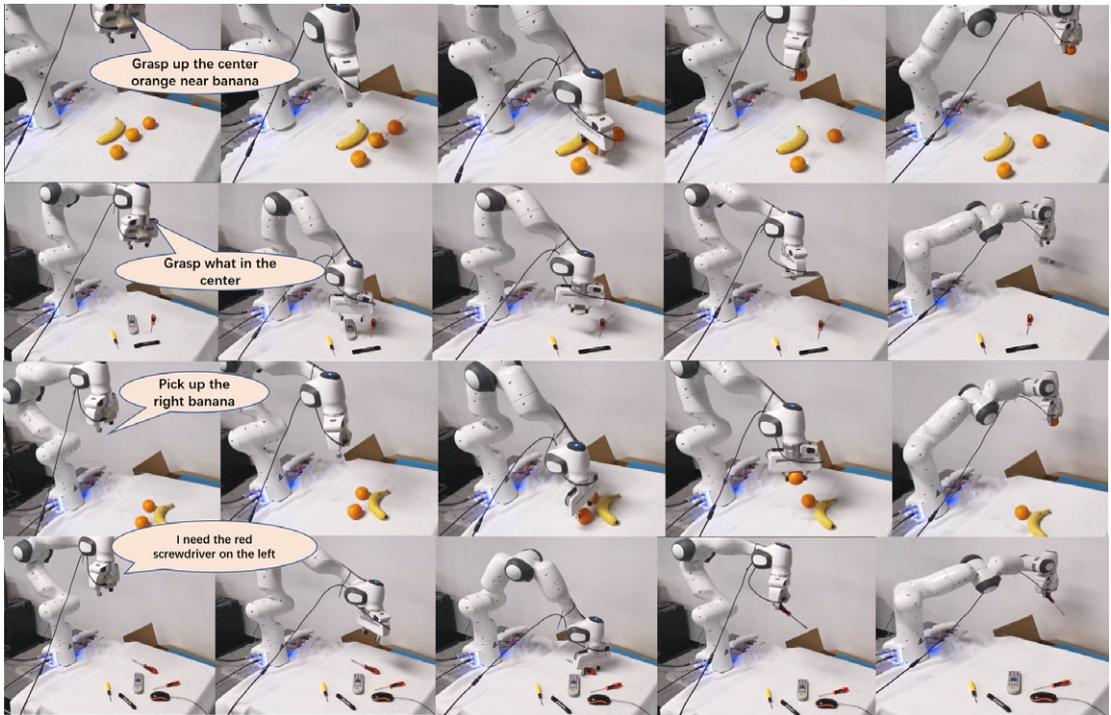


Figure 5. Our multi-modal grasping system demonstrates the understanding of novel sentences and enables the end-effector to generate stable grasps.

visual encoder incorporates attention mechanisms to effectively retain the geometric details of objects. The visual-linguistic fusion module consists of six stacked cross-attention transformer blocks. Each transformer block comprises two fully connected layers with 512 and 2048 neurons, respectively.

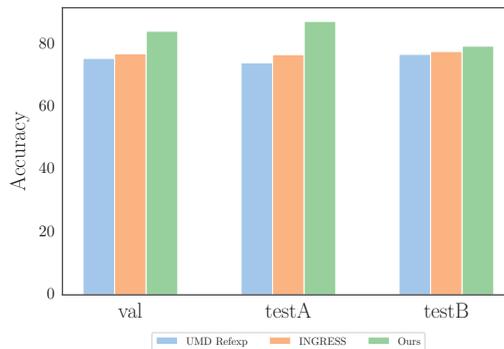
5.4. Results

To investigate the generalization of our model to unseen language commands, we conducted tests using non-templated phrases and objects that were not present in the training set. A screenshot of physics experiments is shown in Fig. 5. The results, presented in Table I, indicate that our method achieved successful completion in 16 out of 20 trials when using novel language instructions to grasp familiar objects. In Fig. 7, we provide visual examples of robot manipulation using diverse linguistic commands. The first row shows images capturing the grasping scenes, while the second row visualizes the attention heatmaps generated by the vision-language transformer encoder. These qualitative examples demonstrate that our method effectively models queries with complex relationships. Specifically, our model learns spatial relationships between objects for instructions such as “grasp what in the center” or “pick the banana on the left of the orange.” This indicates that our multi-modal robotic system can comprehend language in a human-like manner, rather than relying solely on native methods like keyword retrieval or template-based approaches.

We also supplement quantitative results to showcase the visual accuracy in recognizing objects. The accuracy of object recognition refers to how effectively the model can identify and position objects within the scene by language instructions. We compare the accuracy of INGRESS [40] and UMD Refexp [41] with our vision-language network as illustrated in Fig. 6 on RefCOCO benchmark. It is evident that our language grounding model performs better than other two baseline methods.

Table I. Experimental results for evaluating the generalization of the model.

Types of generalization	Success rate
New language commands for seen objects	16/20
New language commands for unseen objects	13/20
Grasp unseen objects in the clutter	18/20
Grasp known objects with unseen poses	15/20

**Figure 6.** Results of recognition and positioning accuracy.

We further test our method with new language commands applied to unseen objects, achieving a success rate of 13 out of 20 trials. Additionally, our method performs well even when objects are cluttered in the environment or in poses that were not seen during training. It is worth noting that understanding these instructions requires not only the comprehension of the phrase-referred landmarks but also the implicit disambiguation.

The results presented in Table II demonstrate the strong performance of our visual understanding model compared to the baselines, showcasing clear superiority over previous CNN-based approaches. The experimental findings highlight the importance of training the visual model for language understanding on a dataset consisting of text-image pairs. Furthermore, the diversity and size of the dataset are found to have a significant impact on the overall performance of the model. We also observed that initializing the visual and linguistic encoders with pre-training weights played a crucial role in the training process. This initialization greatly facilitated the model in establishing correlations between images and their corresponding captions. In Table I, we present experimental results that showcase the generalization capability of our model to different natural language instructions and various objects.

5.5. Ablation studies

To assess the effectiveness of each component in our model, we conduct independent empirical analyses of the encoder and decoder on the relevant datasets to evaluate the performance of vision-language understanding and grasping. Specifically, we perform the following ablation studies:

Encoder analysis: We solely utilize the encoder of our model and use the language input to predict the bounding box of the corresponding object. This analysis allows us to evaluate the performance of the vision and text encoder.

Decoder analysis: In this analysis, we disable the language component and focus on evaluating the grasping performance of the decoder. By excluding the language input, we could assess the capabilities and effectiveness of the decoder component.

Table II. The accuracy on mainstream visual grounding dataset.

Models	Venue	Backbone	RefCOCO			RefCOCO+			RefCOCOg		
			Val	TestA	TestB	Val	TestA	TestB	Val-g	Val-u	Test-u
CMN [42]	CVPR'17	VGG16	–	71.03	65.77	–	54.32	47.76	57.47	–	–
VC [25]	CVPR'18	VGG16	–	73.33	67.44	–	58.40	5.318	62.30	–	–
ParalAttn [26]	CVPR'18	VGG16	–	75.31	65.52	–	61.34	50.86	58.03	–	–
MAttNet [29]	CVPR'18	ResNet-101	76.65	81.14	69.99	65.33	71.62	56.02	–	66.58	67.27
LGRANs [23]	CVPR'19	VGG16	–	76.60	66.40	–	64.00	53.40	61.78	–	–
DGA [43]	ICCV'19	VGG16	–	78.42	65.53	–	69.07	51.99	–	–	63.28
RvG-Tree [44]	TPAMI'19	ResNet-101	75.06	78.61	69.85	63.51	67.45	56.66	–	66.95	66.51
NMTree [45]	ICCV'19	ResNet-101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
Ref-NMS [46]	AAAI'21	ResNet-101	80.70	84.00	76.04	68.25	73.68	59.42	–	70.55	70.62
CMRE [47]	TPAMI'21	ResNet-101	–	82.53	68.58	–	75.76	57.27	–	–	67.38
CM-A-E [48]	CVPR'19	ResNet-101	78.35	83.14	71.32	68.09	73.65	58.03	–	67.99	68.67
SSG [30]	arXiv'18	DarkNet-53	–	76.51	67.50	–	62.14	49.27	47.47	58.80	–
FAOA [32]	ICCV'19	DarkNet-53	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36
RCCF [49]	CVPR'20	DLA-34	–	81.06	71.85	–	70.35	56.32	–	–	65.73
ReSC-Large [50]	ECCV'20	DarkNet-53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
LBYL-Net [31]	CVPR'21	DarkNet-53	79.67	82.91	74.15	68.64	73.38	59.49	62.70	–	–
Transformer-based:											
VGTR [51]	ICME'22	ResNet-101	79.30	82.16	74.38	64.40	70.85	55.84	64.05	66.83	67.28
Refformer [52]	NeurIPS'21	ResNet-101	82.23	85.59	76.57	71.58	75.96	62.16	–	69.41	69.40
TransVG [53]	ICCV'21	ResNet-101	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73
<i>Ours</i>											
Language-Grasp		ResNet-50	84.19	87.31	79.42	74.20	79.50	64.86	71.88	73.53	73.90
Language-Grasp		ResNet-101	84.55	88.34	80.36	75.04	79.54	65.72	72.72	74.26	74.09

Table III. The accuracy on Flickr30K Entities dataset.

Models	Backbone	Flickr30K accuracy (%)
Similarity Net [54]	ResNet-101	60.89
CITE [24]	ResNet-101	61.33
PIRC [55]	ResNet-101	72.83
DDPN [56]	ResNet-101	73.30
LCMCG [57]	ResNet-101	76.74
ZSGNet [58]	ResNet-50	63.39
FAOA [32]	DarkNet-53	68.71
ReSC-Large [50]	DarkNet-53	69.28
TransVG [53]	ResNet-50	78.47
TransVG [53]	ResNet-101	79.10
Ours	ResNet-50	79.41
Ours	ResNet-101	80.12

5.5.1. Visual understanding

To evaluate the effectiveness of our visual understanding module, we conduct validation on several widely recognized visual grounding benchmarks and compare our results with state-of-the-art methods. We follow standard evaluation criteria to report the performance, considering a prediction as positive when the Jaccard index between the predicted region and ground truth exceeds 0.5. The comparison results with recent methods on the RefCOCO, RefCOCO+, and RefCOCOg datasets are presented in Table II. We categorized the methods into two groups: two-stage and one-stage. Our approach demonstrates competitive results across all data splits when compared to recent methods. Notably, we achieved an accuracy of 84.55% on the validation set of RefCOCO, surpassing the prior best method by an improvement of 2.32%. Moreover, by utilizing the more powerful ResNet-101 backbone instead of ResNet-50, the performance of our module is further enhanced. Even when dealing with long expressions, our visual-linguistic module continues to perform well. Table III presents the performance of our model on the Flickr30K Entities test set. With the ResNet-50 and ResNet-101 backbones, our model achieves accuracies of 79.41% and 80.12%, respectively.

The ablation studies reveal that the integration of visual and textual information in the transformer layers is the key component of our approach. This integration enables the extraction of visual features specifically tailored to the textual information. Our visual-linguistic module facilitates homogeneous inter-modal contextual reasoning by embedding both visual and language information into a shared semantic space. The superior results presented in Tables II and III provide evidence of the effectiveness of our visual and linguistic encoder and fusion modules. In Fig. 7, it is evident that our model is capable of capturing the intricate relationships between objects as described in language expressions, such as “the orange on the right of the banana.” The attention context awareness allows our model to take a holistic perspective of the scene, enabling it to focus more accurately on the region where the referred object is located.

5.5.2. Grasping performance

In this section, we analyze the performance of our grasping module. To determine the number of attention layers in each branch, we conducted ablation experiments. The ablation studies show the impact of varying the number of attention layers in the encoder and decoder stages of the model. The two metrics reported are “GFLOPS” (floating-point operations per second) and “Accuracy” in percentage. From the results, it’s evident that increasing the number of encoder layers improves accuracy up to a certain point. The accuracy steadily increases as more layers are employed until it reaches the saturation point. In Tables IV and V, the results highlight that a configuration with four attention layers in encoder

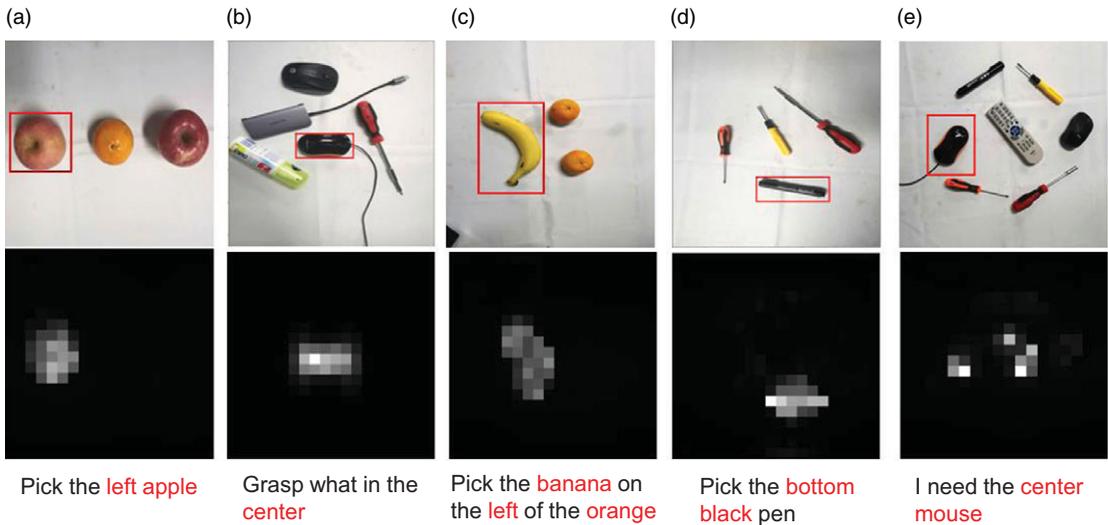


Figure 7. Visualization of the attention heatmaps of the model guided by the language. The first row shows the grasping scene captured from the camera and the red rectangle box indicates the roughly generated grasping candidate area. The second row demonstrates the visualization of discriminative features learned by attention.

and six in decoder yields the highest accuracy while maintaining a reasonable level of computational efficiency. Table VI presents a comparison of our model with recent state-of-the-art approaches on the Cornell grasping dataset. Notably, our model achieves significant improvements, such as increasing the best accuracy on the Cornell dataset from 97.7% to 98.74%. Our approach directly predicts the grasping center, grasping angle, and opening width of the gripper, eliminating the need for designing anchors to match different targets. The highest-scoring grasp candidate is selected as the final grasp prediction. To further assess the performance of our model compared to recent methods, we conduct evaluation of the Jacquard dataset, the results of which are summarized in Table VII.

Comparison with other methods: We surpass the ROI-GD method by 4.1% in terms of grasping accuracy rate due to several key enhancements in our approach. Our approach demonstrates enhanced performance due to the inherent benefits of the transformer-based architecture, enabling the robot to better understand contextual cues from both visual and textual inputs. This facilitates more accurate perception of object attributes, scene elements, and contextual instructions, resulting in improved grasp planning and execution. Furthermore, in comparison to the TF-Grasp method, our method leverages a cross-attention mechanism within a vision-language transformer architecture. The cross-modal interactions between vision and language data enable more effective fusion of information and context, resulting in a higher level of understanding and manipulation accuracy.

Costs in terms of training and execution: Regarding the potential costs associated with training, execution time, and computational resources, we acknowledge that the utilization of transformer-based architectures can indeed introduce higher computational requirements compared to traditional methods. This is primarily due to the increased model complexity and the demand for larger training datasets to capture the diverse range of interactions and instructions. However, it's important to note that the improvements in grasping accuracy and overall task performance, as demonstrated in our experiments, justify the investment in computational resources. The enhanced accuracy directly translates to reduced instances of failed grasps, ultimately leading to a more efficient and reliable robotic manipulation system. Moreover, the scalability of transformer-based models allows for parallel processing and optimization techniques that can mitigate some of the resource constraints. In summary, while our proposed method may require comparatively more computational resources during training and execution, the achieved

Table IV. The ablation studies of the attention layer numbers in encoder.

The encoder stages (N)	GFLOPS	Accuracy (%)
Encoder layers (1)	1.13	75.64
Encoder layers (2)	2.26	78.05
Encoder layers (3)	3.39	82.70
Encoder layers (4)	4.52	84.19
Encoder layers (6)	6.78	84.18

Table V. The ablation studies of the attention layer numbers in decoder.

The decoder stages (N)	GFLOPS	Accuracy (%)
Decoder layers (1)	2.34	63.64
Decoder layers (2)	4.68	72.05
Decoder layers (4)	9.36	80.70
Decoder layers (6)	14.04	84.19
Decoder layers (8)	18.72	84.22

Table VI. The grasping performance on Cornell grasping dataset.

Authors	Algorithm	Accuracy (%)
Jiang [59]	Fast Search	60.5
Lenz [16]	SAE, struct. reg.	73.9
Redmon [60]	AlexNet, MultiGrasp	88.0
Wang [14]	Two-stage closed-loop	85.3
Asif [61]	STEM-CaRFs	88.2
Kumra [39]	ResNet-50x2	89.2
Morrison [62]	GG-CNN	73.0
Guo [17]	ZF-net	93.2
Zhou [63]	FCGN, ResNet-101	97.7
Karaoguz [64]	GRPN	88.7
Asif [38]	GraspNet	90.2
Our	GraspDecoder	98.74

advancements in grasping accuracy and human-robot interaction effectiveness substantiate the benefits and justify the associated costs.

5.6. Discussion and limitations

Our proposed method has been evaluated on various objects with different language instructions, demonstrating its effectiveness on previously unseen objects such as remote controls, oranges, and screwdrivers. It exhibits good generalization capabilities by maintaining a high grasping success rate even when objects are placed in orientations that were not seen during training. The results are summarized in Table VI. Through extensive empirical validation, our model proves to be a powerful solution for language-driven robotic understanding tasks, offering a seamless and effective human-robot interface.

While our approach has demonstrated promise in language-driven grasping, it is constrained to a two-dimensional plane. Real-world applications encompass scenarios well beyond this scope. Grasping tasks

Table VII. The accuracy on Jacquard grasping dataset.

Authors	Method	Input	Accuracy (%)
Depierre [35]	Jacquard	RGB-D	74.2
Morrison [62]	GG-CNN2	D	84
Zhou [63]	FCGN, ResNet-101	RGB	91.8
Alexandre [65]	GQ-STN	D	70.8
Zhang [66]	ROI-GD	RGB	90.4
Stefan [15]	Det Seg	RGB	92.59
Stefan [15]	Det Seg Refine	RGB	92.95
Kumra [67]	GR-ConvNet	D	93.7
Kumra [67]	GR-ConvNet	RGB	91.8
Wang [18]	TF-Grasp	D	93.1
Ours	GraspDecoder	D	94.5

in vertical spaces, such as hierarchical shelving, introduce additional layers of complexity, including height considerations, spatial positioning, and object orientation. Notably, our framework's performance in these more intricate settings remains unverified, representing a prominent limitation that warrants dedicated attention in future work. In summary, our forthcoming research will focus on surmounting the challenges associated with vertical space grasping. Research avenues will encompass enhancing 3D perception, advancing 6DOF pose estimation, addressing object occlusion, and refining motion planning precision. By addressing these facets, we aim to extend our framework's applicability to real-world scenarios where vertical space grasping is an essential requirement.

6. Conclusions

This paper introduces a novel robotic system designed to comprehend human intentions and execute user-specified object manipulation commands through natural language. By leveraging natural language as a means of expressing human intentions, our system enables robots to understand object relationships and generate desired actions. A key feature of our model is its capability to comprehend flexible language instructions using a multi-modal transformer. Through extensive evaluations of various datasets and real-world robotic systems, we demonstrate that our proposed method is user-friendly and exhibits promising performance. The flexibility of our framework allows for effective multi-modal understanding.

Financial support. This work was supported in part by National Key R&D Program of China under Grant 2022YFB4701400/4701403, National Natural Science Foundation of China under Grant U2013601, U19B2044, U22A2060, and National Key R&D Program Program of China under Grant 2021YFF0501600.

Author contributions. Author 1: Designed methodology, conducted experiments, and managed data collection. Author 2: Conducted extensive literature review and implementation of specific algorithms. Author 3: Co-wrote the sections related to experimental setup and results. Author 4: Refining the discussion and conclusion sections. Author 5: Provided expertise in robotics and automation and contributed to the overall research in the paper.

Competing interests. The authors declare none.

References

- [1] W. Yan and A. Mehta, "Towards one-dollar robots: An integrated design and fabrication strategy for electromechanical systems," *Robotica* **41**(1), 31–47 (2023).
- [2] A. Spielberg, T. Du, Y. Hu, D. Rus and W. Matusik, "Advanced soft robot modeling in chainqueen," *Robotica* **41**(1), 74–104 (2023).

- [3] X. Zhou, J. Ye, C. Wang, J. Zhong and X. Wu, "Time-frequency feature transform suite for deep learning-based gesture recognition using senn signals," *Robotica* **41**(2), 775–788 (2023).
- [4] H. Su, Y. Schmirander, S. E. Valderrama-Hincapié, W. Qi, S. E. Ovrur and J. Sandoval, "Neural-learning-enhanced cartesian admittance control of robot with moving RCM constraints," *Robotica* **41**(4), 1231–1243 (2023).
- [5] Z. Li, G. Li, X. Wu, Z. Kan, H. Su and Y. Liu, "Asymmetric cooperation control of dual-arm exoskeletons using human collaborative manipulation models," *IEEE Trans. Cybern.* **52**(11), 12126–12139 (2022).
- [6] L. D. Rosenblum, *See What I'm Saying: The Extraordinary Powers of Our Five Senses* (WW Norton & Company, New York, 2011).
- [7] J. Krishnamurthy and T. Kollar, "Jointly learning to parse and perceive: Connecting natural language to the physical world," *Trans. Assoc. Comput. Linguist.* **1**, 193–206 (2013).
- [8] Y. Chen, R. Xu, Y. Lin and P. A. Vela, "A Joint Network for Grasp Detection Conditioned on Natural Language Commands," **In: Proc. IEEE Int. Conf. Robot. Automat.** (2021) pp. 4576–4582.
- [9] L. Shao, T. Migimatsu, Q. Zhang, K. Yang and J. Bohg, "Concept2robot: Learning Manipulation Concepts From Instructions and Human Demonstrations," **In: Robotics: Science and Systems** (2020).
- [10] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," **In: Proc. Assoc. Comput. Linguistics** (2019) pp. 4171–4186.
- [11] J. Arkin, D. Park, S. Roy, M. R. Walter, N. Roy, T. M. Howard and R. Paul, "Multimodal estimation and communication of latent semantic knowledge for robust execution of robot instructions," *Int. J. Robot. Res.* **39**(10-11), 1279–1304 (2020).
- [12] Z. Li, Q. Li, P. Huang, H. Xia and G. Li, "Human-in-the-loop adaptive control of a soft exo-suit with actuator dynamics and ankle impedance adaptation," *IEEE Trans. Cybern.* 1–13 (2023).
- [13] Z. Li, K. Zhao, L. Zhang, X. Wu, T. Zhang, Q. Li, X. Li and C.-Y. Su, "Human-in-the-loop control of a wearable lower limb exoskeleton for stable dynamic walking," *IEEE ASME Trans. Mechatron.* **26**(5), 2700–2711 (2021).
- [14] Z. Wang, Z. Li, B. Wang and H. Liu, "Robot grasp detection using multimodal deep convolutional neural networks," *Adv. Mech. Eng.* **8**(9), 1687814016668077 (2016).
- [15] S. Ainetter and F. Fraundorfer, "End-to-End Trainable Deep Neural Network for Robotic Grasp Detection and Semantic Segmentation from RGB," **In: Proc. IEEE Int. Conf. Robot. Automat.** (IEEE, 2021) pp. 13452–13458.
- [16] I. Lenz, H. Lee and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.* **34**(4-5), 705–724 (2015).
- [17] G. Di, F. Sun, H. Liu, T. Kong, B. Fang and N. Xi, "A Hybrid Deep Architecture for Robotic Grasp Detection," **In: Proc. IEEE Int. Conf. Robot. Automat.** (2017) pp. 1609–1614.
- [18] S. Wang, Z. Zhou and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE Robot. Autom. Lett.* **7**(3), 8170–8177 (2022).
- [19] Z. Zhou, S. Wang, Z. Chen, M. Cai, H. Wang, Z. Li and Z. Kan, "Local observation based reactive temporal logic planning of human-robot systems," *IEEE Trans. Autom. Sci. Eng.* 1–13 (2023).
- [20] S. Wang, Z. Zhou, H. Wang, Z. Li and Z. Kan, "Unsupervised Representation Learning for Visual Robotics Grasping," **In: International Conference on Advanced Robotics and Mechatronics (ICARM)** (2022) pp. 57–62.
- [21] K. Chen, S. Wang, B. Xia, D. Li, Z. Kan and B. Li, "Tode-Trans: Transparent Object Depth Estimation with Transformer," **In: IEEE Int. Conf. Robot. Autom.** (2023) pp. 4880–4886.
- [22] S. Wang, W. Zhang, Z. Zhou, J. Cao, Z. Chen, K. Chen, B. Li and Z. Kan, "What you see is what you grasp: User-friendly grasping guided by near-eye-tracking," *arXiv preprint arXiv:2209.06122* (2022).
- [23] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao and A. van den Hengel, "Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks," **In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.** (2019) pp. 1960–1968.
- [24] B. A. Plummer, P. Kordas, M. H. Kiapour, S. Zheng, R. Piramuthu and S. Lazebnik, "Conditional Image-Text Embedding Networks," **In: Proc. Eur. Conf. Comput. Vis.**, vol. **11216** (2018) pp. 258–274.
- [25] H. Zhang, Y. Niu and S.-F. Chang, "Grounding Referring Expressions in Images by Variational Context," **In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.** (2018) pp. 4158–4166.
- [26] B. Zhuang, Q. Wu, C. Shen, I. D. Reid and A. van den Hengel, "Parallel Attention: A Unified Framework for Visual Object Discovery Through Dialogs and Queries," **In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.** (2018) pp. 4252–4261.
- [27] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille and K. Murphy, "Generation and Comprehension of Unambiguous Object Descriptions," **In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.** (2016) pp. 11–20.
- [28] L. Yu, Poirson P., Yang S., Berg A. C. and Berg T. L., "Modeling Context in Referring Expressions," **In: Proc. Eur. Conf. Comput. Vis.**, vol. **9906** (2016) pp. 69–85.
- [29] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal and T. L. Berg, "Modular Attention Network for Referring Expression Comprehension," **In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.** (2018) pp. 1307–1315.
- [30] X. Chen, L. Ma, J. Chen, Z. Jie, W. Liu and J. Luo, "Real-time referring expression comprehension by single-stage grounding network," *CoRR*, abs/1812.03426 (2018).
- [31] B. Huang, D. Lian, W. Luo and S. Gao, "Look Before You Leap: Learning Landmark Features for One-Stage Visual Grounding," **In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.** (2021) pp. 16888–16897.
- [32] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu and J. Luo, "A Fast and Accurate One-stage Approach to Visual Grounding," **In: Proc. IEEE Int. Conf. Comput. Vis.** (2019) pp. 4682–4692.
- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," **In: Proc. Adv. Neural Inf. Process. Syst.**, vol. **30** (2017).
- [34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-End Object Detection with Transformers," **In: Proc. Eur. Conf. Comput. Vis.** (Springer, Cham, 2020) pp. 213–229.

- [35] A. Depierre, E. Dellandréa and L. Chen, “Jacquard: A Large Scale Dataset for Robotic Grasp Detection,” *In: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* (2018) pp. 3511–3516.
- [36] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier and S. Lazebnik, “Flicker30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” *Int. J. Comput. Vis.* **123**(1), 74–93 (2017).
- [37] P. Young, A. Lai, M. Hodosh and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Trans. Assoc. Comput. Linguist.* **2**, 67–78 (2014).
- [38] U. Asif, J. Tang and S. Harrer, “Graspnet: An Efficient Convolutional Neural Network for Real-Time Grasp Detection for Low-Powered Devices,” *In: IJCAI*, vol. 7 (2018) pp. 4875–4882.
- [39] S. Kumra and C. Kanan, “Robotic Grasp Detection Using Deep Convolutional Neural Networks,” *In: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* (2017) pp. 769–776.
- [40] M. Shridhar, D. Mittal and D. Hsu, “INGRESS: Interactive visual grounding of referring expressions,” *Int. J. Robot. Res.* **39**(2-3), 217–232 (2020).
- [41] V. K. Nagaraja, V. I. Morariu and L. S. Davis, “Modeling Context Between Objects for Referring Expression Understanding,” *In: Proc. Eur. Conf. Comput. Vis.*, vol. **9908** (2016) pp. 792–807.
- [42] R. Hu, M. Rohrbach, J. Andreas, T. Darrell and K. Saenko, “Modeling Relationships in Referential Expressions with Compositional Modular Networks,” *In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (2017) pp. 4418–4427.
- [43] S. Yang, G. Li and Y. Yu, “Dynamic Graph Attention for Referring Expression Comprehension,” *In: Proc. IEEE Int. Conf. Comput. Vis.* (2019) pp. 4643–4652.
- [44] R. Hong, D. Liu, X. Mo, X. He and H. Zhang, “Learning to compose and reason with language tree structures for visual grounding,” *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(2), 684–696 (2022).
- [45] D. Liu, H. Zhang, Z.-J. Zha and F. Wu, “Learning to Assemble Neural Module Tree Networks for Visual Grounding,” *In: Proc. IEEE Int. Conf. Comput. Vis.* (2019) pp. 4672–4681.
- [46] L. Chen, W. Ma, J. Xiao, H. Zhang and S.-F. Chang, “Ref-NMS: Breaking Proposal Bottlenecks in Two-Stage Referring Expression Grounding,” *In: AAAI Conf. Artif. Intell.* (2021) pp. 1036–1044.
- [47] S. Yang, G. Li and Y. Yu, “Relationship-embedded representation learning for grounding referring expressions,” *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(8), 2765–2779 (2021).
- [48] X. Liu, Z. Wang, J. Shao, X. Wang and H. Li, “Improving Referring Expression Grounding with Cross-Modal Attention-Guided Erasing,” *In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (2019) pp. 1950–1959.
- [49] Y. Liao, S. Liu, G. Li, F. Wang, Y. Chen, C. Qian and B. Li, “A Real-Time Cross-Modality Correlation Filtering Method for Referring Expression Comprehension,” *In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (2020) pp. 10877–10886.
- [50] Z. Yang, T. Chen, L. Wang and J. Luo, “Improving One-Stage Visual Grounding by Recursive Sub-Query Construction,” *In: Proc. Eur. Conf. Comput. Vis.*, vol. **12359** (2020) pp. 387–404.
- [51] Y. Du, Z. Fu, Q. Liu and Y. Wang, “Visual Grounding with Transformers,” *In: Proc. IEEE Conf. Multi. Exp.* (2022) pp. 1–6.
- [52] M. Li and L. Sigal, “Referring Transformer: A One-Step Approach to Multi-Task Visual Grounding,” *In: Proc. Int. Conf. Neural Inf. Process. Syst.* (2021) pp. 19652–19664.
- [53] J. Deng, Z. Yang, T. Chen, W. Zhou and H. Li, “TransVG: End-to-End Visual Grounding with Transformers,” *In: Proc. IEEE Int. Conf. Comput. Vis.* (2021) pp. 1749–1759.
- [54] L. Wang, Y. Li, J. Huang and S. Lazebnik, “Learning two-branch neural networks for image-text matching tasks,” *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(2), 394–407 (2019).
- [55] R. Kovvuri and R. Nevatia, “PIRC Net: Using Proposal Indexing, Relationships and Context for Phrase Grounding,” *In: Asian Conf. Comput. Vis.*, vol. **11364** (2018) pp. 451–467.
- [56] Z. Yu, C. Xiang, Z. Zhao, Q. Tian and D. Tao, “Rethinking Diversified and Discriminative Proposal Generation for Visual Grounding,” *In: Int. Joint Conf. Artif. Intell.* (2018) pp. 1114–1120.
- [57] Y. Liu, B. Wan, X. Zhu and X. He, “Learning Cross-Modal Context Graph for Visual Grounding,” *In: AAAI Conf. Artif. Intell.* (2020) pp. 11645–11652.
- [58] A. Sadhu, K. Chen and R. Nevatia, “Zero-Shot Grounding of Objects from Natural Language Queries,” *In: Proc. IEEE Int. Conf. Comput. Vis.* (2019) pp. 4693–4702.
- [59] Y. Jiang, S. Moseson and A. Saxena, “Efficient Grasping from RGBD Images: Learning Using a New Rectangle Representation,” *In: Proc. IEEE Int. Conf. Robot. Automat.* (2011) pp. 3304–3311.
- [60] J. Redmon and A. Angelova, “Real-Time Grasp Detection Using Convolutional Neural Networks,” *In: Proc. IEEE Int. Conf. Robot. Autom.* (2015) pp. 1316–1322.
- [61] U. Asif, M. Bennamoun and F. A. Sohel, “RGB-D object recognition and grasp detection using hierarchical cascaded forests,” *IEEE Trans. Robot.* **33**(3), 547–564 (2017).
- [62] D. Morrison, P. Corke and J. Leitner, “Learning robust, real-time, reactive robotic grasping,” *Int. J. Robot. Res.* **39**(2-3), 183–201 (2020).
- [63] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang and N. Zheng, “Fully Convolutional Grasp Detection Network with Oriented Anchor Box,” *In: Proc. IEEE Int. Conf. Intell. Robots Syst.* (2018) pp. 7223–7230.
- [64] H. Karaoguz and P. Jensfelt, “Object Detection Approach for Robot Grasp Detection,” *In: Proc. IEEE Int. Conf. Robot. Automat.* (2019) pp. 4953–4959.
- [65] A. Gariépy, J.-C. Ruel, B. Chaib-Draa and P. Giguere, “GQ-STN: Optimizing One-Shot Grasp Detection Based on Robustness Classifier,” *In: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* (2019) pp. 3996–4003.

- [66] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian and N. Zheng, “ROI-Based Robotic Grasp Detection for Object Overlapping Scenes,” **In: Proc. IEEE Int. Conf. Intell. Robots Syst.** (2019) pp. 4768–4775.
- [67] S. Kumra, S. Joshi and F. Sahin, “Antipodal Robotic Grasping Using Generative Residual Convolutional Neural Network,” **In: Proc. IEEE Int. Conf. Intell. Robots Syst.** (IEEE, 2020) pp. 9626–9633.