Language-driven Grasp Detection with Mask-guided Attention

Tuan Van Vo[1], Minh Nhat Vu[2,3,*], Baoru Huang[4], An Vuong[1], Ngan Le[5], Thieu Vo[6], Anh Nguyen[7]

***Abstract*— Grasp detection is an essential task in robotics with various industrial applications. However, traditional methods often struggle with occlusions and do not utilize language for grasping. Incorporating natural language into grasp detection remains a challenging task and largely unexplored. To address this gap, we propose a new method for language-driven grasp detection with mask-guided attention by utilizing the transformer attention mechanism with semantic segmentation features. Our approach integrates visual data, segmentation mask features, and natural language instructions, significantly improving grasp detection accuracy. Our work introduces a new framework for language-driven grasp detection, paving the way for language-driven robotic applications. Intensive experiments show that our method outperforms other recent baselines by a clear margin, with a 10.0% success score improvement. We further validate our method in real-world robotic experiments, confirming the effectiveness of our approach.**

I. INTRODUCTION

Grasp detection is the fundamental task in robotics, with widespread applications in manufacturing, logistics, and service robots [1]. Traditional grasping detection methods often struggle with object complexities and occlusions [2]. However, recent advances in computer vision, machine learning, and natural language processing have opened up new possibilities for addressing the challenge using deep networks [3]. However, most existing works focus on detecting grasp poses without language instruction [4]–[10]. In practice, language-driven grasping presents an intriguing and demanding task in robot manipulation [3], [11], where natural language can guide the robot to grasp on-demand objects. Developing language-driven grasping systems is not a rival task and requires the understanding of language instructions and visual information of scene [12], [13].

In recent years, there has been a surge in interest in language-driven robotic manipulation, enabling robots to comprehend natural language commands for executing manipulation tasks [3], [12], [14]–[16]. This paradigm shift brings numerous advantages, including enhanced human-robot interaction, adaptability to various environments, and improved task efficiency [17]. Language-driven robotic frameworks are gaining momentum, empowering robots to process natural language and bridging the gap between robotic manipulations and real-world human-robot interaction [17]. Embodied robots such as PaLM-E [16], Ego-COT [17], and ConceptFusion [18] have emerged with the capability to comprehend natural language by leveraging large foundation models like ChatGPT [19]. However, many existing works primarily focus on high-level robot actions, overlooking fundamental grasping actions, thus limiting generalization across robotic domains, tasks, and skills [20]. Despite the rising interest in language-driven grasp detection, current approaches struggle to handle object complexities effectively [20]. Challenges such as ambiguities in natural language instructions, limited vocabulary, and difficulties in contextual understanding impede the accurate interpretation of user commands [17]. Moreover, dependencies on precise language understanding and inefficiencies in noisy environments pose additional obstacles, potentially leading to difficulties in object comprehension [12].

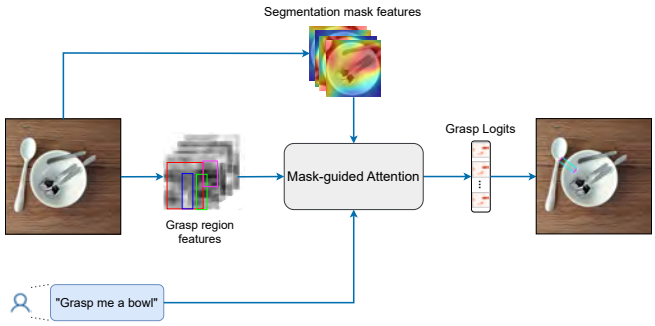

Fig. 1. We propose a mask-guided attention mechanism that learns the mask and language features to tackle the language-driven grasping task.

In this paper, we present a new approach to tackle the language-driven grasp detection task. Inspired by the Transformer network's powerful attention mechanism [21], our method capitalizes on recent advancements in multimodal learning to integrate visual information, segmentation mask features, and natural language instructions for robust grasp detection. Specifically, we propose a mask-guided attention mechanism for the language-driven grasp detection task to concurrently model grasp region features, segmentation mask features, and language embeddings. Our approach aims to enhance object understanding through attention to segmentation mask features, facilitating a better connection between attended language embeddings and the correct grasp region features, thereby improving the accuracy of the language-driven grasp detection task. Extensive experiments demonstrate that our method achieves an approximately 10% success score improvement over the baselines. Ablation studies and qualitative results on real-world robotic grasping

[1] FPT Software AI Center, Vietnam tuanvv7@fpt.com
[2] Automation & Control Institute, TU Wien, Austria
[3] Austrian Institute of Technology (AIT) GmbH, Austria
[4] Imperial College London, UK
[5] Department of Computer Science & Computer Engineering, University of Arkansas, USA
[6] National University of Singapore, Singapore
[7] Department of Computer Science, University of Liverpool, UK
[*] Corresponding author minh.vu@ait.ac.at

applications further validate the effectiveness of our approach and provide insights for future research directions.

Our main contributions are summarized as follows:

- We propose a mask-guided attention mechanism to enhance multimodal integration for the language-driven grasp detection task.
- We provide a comprehensive analysis of our proposed method, including experimental results on benchmark datasets and ablation studies to evaluate the effectiveness of different components.

II. Related Work

Grasp Detection. Traditional approaches to robotic grasp detection have included analytical methods [4], [5], [22], which focus on object geometry and contact forces, and convolutional neural networks (CNNs) [7], [23]–[27], trained on labeled datasets of grasping examples [7]–[9]. While attention mechanisms in Transformers have shown promise in sequence modeling for information fusion across global sequences [21], Wang *et al.* [10] introduced a Transformer-based visual grasp detection framework, leveraging attention's ability to aggregate information across input sequences for improved global representation. The design of this framework incorporates local window attention to capture local contextual information and detailed features of graspable objects. However, a significant drawback of both analytical, CNN-based, and Transformer-based methods is their limited scene understanding and inability to process language instructions, which hampers their effectiveness in dynamic, human-centric environments.

Language-driven Grasp Detection. Recent advances in large language models have facilitated the integration of language understanding to robotic tasks, enabling robots to execute manipulation tasks based on natural language instructions [28]–[35]. This transition towards language-guided grasp referral empowers robots to identify and manipulate objects according to user specifications, thereby enhancing their utility in complex scenarios [12], [17]. For instance, Tziafas *et al.* [36] concentrate on grasp synthesis based on linguistic references, predicting grasp poses for referenced objects using natural language in cluttered scenes, while Chen *et al.* [37] propose a method to jointly learn from visual and language features and predict 2D grasp boxes from RGB images.

Transformer Attention Mechanism. Initially for NLP tasks [38], the Transformer's multi-head attention mechanism excels in capturing long-term word correlations. While primarily used in NLP, attempts have been made to apply Transformers to vision tasks such as image super-resolution [39], object detection [40], and multimodal video understanding [41]–[43]. However, these methods still rely on CNN-extracted features. Recent advancements include convolution-free vision Transformers [44], which operate directly on raw images, achieving competitive performance. Further improvements in training data efficiency have been made by [45] through stronger data augmentations and knowledge distillation. The pure Transformer design has since been applied to various vision tasks, including semantic segmentation [46], point cloud classification [47], and action recognition [48]–[50]. In our work, we propose Mask-guided Attention as the Transformer attention model to learn visual inputs, object segmentation features, and text features for the language-driven grasp detection task.

Despite the burgeoning interest in language-driven grasp detection, extant methods grapple with the intricacies of object geometries, linguistic ambiguities, and contextual understanding challenges, impeding precise interpretation of user commands [16], [17]. This restricts robots' ability to understand nuanced, implicit instructions crucial for real-world interactions [12], [17], [20]. To this end, we introduce a new framework for language-driven grasp detection, harnessing recent strides in transformer multimodal learning and attention mechanisms [21]. Specifically, we advocate for a mask-guided attention mechanism, tailored to bolster grasp detection reliability through comprehensive multimodal integration. Our intuition is to utilize segmentation mask features to enhance the alignment between language embeddings and grasp region features for the language-driven grasping task.

III. Language-driven Grasping with Mask-guided Attention

A. Overview

Given an input RGB image I and a text prompt describing the object of interest, our goal is to detect the grasping pose on the image that best matches the text prompt input. We follow the popular *rectangle grasp* convention that is widely used in previous work to define the grasp pose [8]. In particular, each grasp pose is defined with five parameters: the (x, y) center coordinate, the width, height (w, h) of the rectangle, and the rotational angle identifies the orientation of the rectangle relative to the horizontal axis of the image. Fig. 2 shows an overview of our framework. We leverage a pretrained text encoder and segmentation mask features from the segmentation head, along with a grasp region proposal head for spatial feature extraction. Our method, mask-guided attention, improves grasp detection task by incorporating cross-attention from segmentation mask features and language embeddings, enhancing the connection between language embeddings and grasp region features through attention to segmentation mask features.

B. Visual and Language Feature Extraction

Grasp Region Feature Extraction. The first stage of our visual processing pipeline is to extract a set of grasp regions of interest (ROIs) and their feature representations as in [51], [52]. The visual grasp region features contain geometric information for determining grasp configuration and semantic information for reasoning with natural languages. At the end of the grasp region feature extraction pipeline, a fixed-size feature map is passed to a convolution neural network to produce a set of vectors, which are interpreted as the embedding for each candidate grasp region. Specially, the grasp region feature representation vector is defined as:

$$\mathcal{F}^{vis} = \{y_{\text{vis},i}\}_1^m = \{(\rho^i, r^i, y^i_{\text{img}})\}_1^m, \tag{1}$$

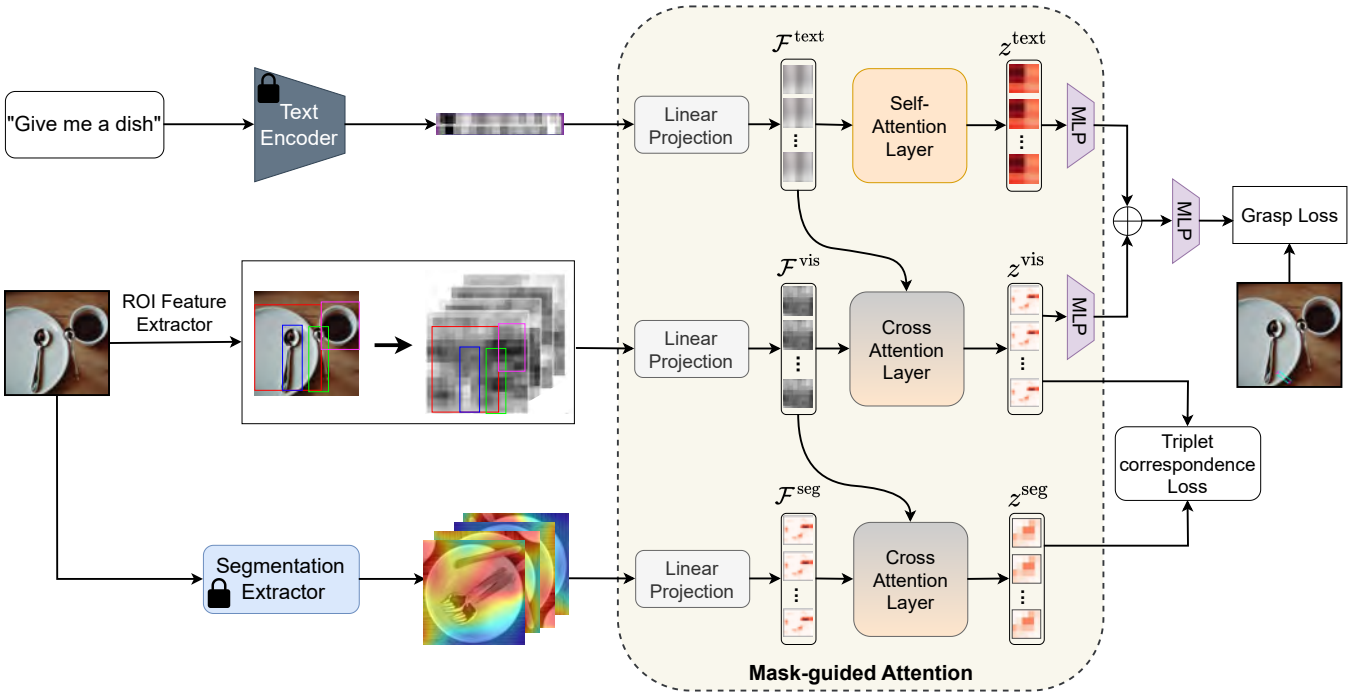

Fig. 2. The overview of our mask-guided attention framework for the language-driven grasp detection task.

where $\mathcal{F}^{vis} \in \mathbb{R}^{d\times h\times w}$, the coordinates of $(\rho, r, y_{\text{img}})$ consist of the proposal probability, the predicted position, and the image feature representation.

Segmentation Feature Extraction. We leverage an object instances segmentation network to acquire features that represent the "meaning" of objects within an image. By building upon the proven object instance segmentation architecture [53], [54], our network progressively analyzes the image, culminating in rich, high-dimensional segmentation mask features $\mathcal{F}^{\text{seg}} \in \mathbb{R}^{d\times h\times w}$. These features provide a granular understanding of the scene, from large shapes to fine details, leading to improved performance.

Text embedding. Given an input text query with K words (e.g., "grasp the blue bottle"), we embed the text input with a pre-trained BERT [21] or CLIP [55] into text embedding feature vectors $\mathcal{F}^{\text{text}} \in \mathbb{R}^{d}$. We note that the text encoder is frozen during the training.

C. Mask-guided Attention

Inspired by the Transformer network's powerful attention mechanism [21], we introduce a new architecture called mask-guided attention. Our approach merges information from various sources (grasp region features, text features, and segmentation mask features) to achieve a deeper understanding of grasping tasks. By employing cross-modal attention, our method focuses on critical features within each modality, ultimately fusing them into a unified representation that guides toward robust grasping. Our proposed cross attention mechanism jointly learns the $W_Q^{\text{text}}, W_K^{\text{text}}, W_V^{\text{text}}$ and $W_Q^{\text{vis}}, W_K^{\text{vis}}, W_V^{\text{vis}}$ and $W_Q^{\text{seg}}, W_K^{\text{seg}}, W_V^{\text{seg}}$, i.e., the query, key and value weight matrices for grasp region features, text features, and segmentation mask features, respectively. Here, all weight matrices have dimensions $d \times d$.

We first use a self-attention layer to compute the output S^{text} from the input features $\mathcal{F}^{\text{text}}$ by first transforming them into query, key, and value matrices using learned linear transformations. Specifically, we calculate $Q^{\text{text}} = \mathcal{F}^{\text{text}} \times W_Q^{\text{text}}$; $K^{\text{text}} = \mathcal{F}^{\text{text}} \times W_K^{\text{text}}$; $V^{\text{text}} = \mathcal{F}^{\text{text}} \times W_V^{\text{text}}$. Subsequently, we determine the value S^{text} as follows:

$$S^{\text{text}} = \text{softmax}(\frac{Q^{\text{text}} \cdot (K^{\text{text}})^\top}{\sqrt{d}}) \tag{2}$$

To understand the relationship between the text and grasp region features, we first calculate $Q^{\text{vis}} = \mathcal{F}^{\text{text}} \times W_Q^{\text{vis}}$; $K^{\text{vis}} = \mathcal{F}^{\text{vis}} \times W_K^{\text{vis}}$; $V^{\text{vis}} = \mathcal{F}^{\text{vis}} \times W_V^{\text{vis}}$, then calculate S^{vis}:

$$S^{\text{vis}} = \text{softmax}(\frac{Q^{\text{vis}} \cdot (K^{\text{vis}})^\top}{\sqrt{d}}) \tag{3}$$

Similarity, to understand the relationship between segmentation mask and grasp region features, we first obtain $Q^{\text{seg}} = \mathcal{F}^{\text{vis}} \times W_Q^{\text{seg}}$; $K^{\text{seg}} = \mathcal{F}^{\text{seg}} \times W_K^{\text{seg}}$; $V^{\text{seg}} = \mathcal{F}^{\text{seg}} \times W_V^{\text{seg}}$, then calculate S^{seg}:

$$S^{\text{seg}} = \text{softmax}(\frac{Q^{\text{seg}} \cdot (K^{\text{seg}})^\top}{\sqrt{d}}) \tag{4}$$

The overall attention outputs can then be computed as $S^{\text{text}} \times V^{\text{text}}$, $S^{\text{vis}} \times V^{\text{vis}}$ and $S^{\text{seg}} \times V^{\text{seg}}$, respectively, which can be applied to the original vectors $\mathcal{F}^{\text{text}}$, $\mathcal{F}^{\text{vis}}$ and $\mathcal{F}^{\text{seg}}$. The attention can be learned over multiple heads in parallel, as seen in the transformer architecture [38], for added context within the scaling. If the attention layer is splitted into H heads, the output of each head will have a dimension of $d_{\text{head}} = \frac{d}{H}$. Also, the weight matrices Q, K and V will now be of dimensions $d_{\text{head}} \times d$. The final output of the multi-head attention layer is then given by:

$$\text{MultiHeadAttn} = \text{concat}(head_1, ..., head_H)W_O \quad (5)$$

where $W_O \in \mathbb{R}^{d\times d}$ are weights to be learned. We use a layer normalization post-scaling to reduce the chances of overfitting. The final output of the scaling results in the following embeddings:

$$z^{\text{text}} = \text{LayerNorm}(S^{\text{text}} \times V^{\text{text}} + \mathcal{F}^{\text{text}}) \quad (6)$$

$$z^{\text{vis}} = \text{LayerNorm}(S^{\text{vis}} \times V^{\text{vis}} + \mathcal{F}^{\text{vis}}) \quad (7)$$

$$z^{\text{seg}} = \text{LayerNorm}(S^{\text{seg}} \times V^{\text{seg}} + \mathcal{F}^{\text{seg}}) \quad (8)$$

An output grasping module that consists of two Multi-Layer Perceptron (MLP) to fused the information of $\{z_1^{\text{text}}, \cdots, z_M^{\text{text}}\}$ and $\{z_1^{\text{vis}}, \cdots, z_M^{\text{vis}}\}$, then the final MLP layers projects fused features into a set of M grasp scores $\{S_1^{\text{vis}}, \cdots, S_M^{\text{vis}}\}$, respectively. The grasp object proposal $G^i, i \in M$ with the highest grasping score is selected as the final grasping prediction.

D. Training

Triplet Correspondence Loss. The introduced loss function for grasp region features correspondence aims to understand the relationship between grasp regions and those identified in the segmentation mask feature of objects. We formulate this correspondence loss as a triplet loss [56]–[60]:

$$\mathcal{L}_{\text{cor}} = \sum_{m=1}^{M} \Big\{ \big[\alpha - s(z_m^{\text{vis}}, z_m^{\text{seg}}) + s(z_m^{\text{vis}}, z_i^{\text{seg}})\big]_+ + \big[\alpha - s(z_m^{\text{vis}}, z_m^{\text{seg}}) + s(z_j^{\text{vis}}, z_m^{\text{seg}})\big]_+ \Big\},$$

where $\text{s}(\cdot)$ is the similarity function. We use the inner product over the L2 normalized feature z^{vis} and z^{seg} as $\text{s}(\cdot)$ in our experiments. α is the margin with a default value of 0.1. i, j are the index for the hard negatives where $i = \text{argmax}_{i\neq m} s(z_m^{\text{vis}}, z_i^{\text{seg}})$ and $j = \text{argmax}_{j\neq m} s(z_j^{\text{vis}}, z_m^{\text{seg}})$. We determine the grasp correspondence among the attention grasp region features proposals m and the attention segmentation mask features within each input image I.

Grasp Loss. The grasp loss function, denoted as $\mathcal{L}_{\text{grasp}}$, is informed by previous research on grasp detection [51], [52]. It combines grasp regression and classification losses to serve two main purposes. Firstly, grasp regression loss guarantees accurate grasp localization. Secondly, classification loss assists in accurately identifying successful grasps, crucial for distinguishing effective grasping strategies from ineffective ones.

$$\mathcal{L}_{\text{grasp}} = -\sum_{i\in\text{Positive}} \log(p_g^i) - \sum_{i\in\text{Negative}} \log(p_u^i) + \beta \sum_{i\in\text{Positive}} \text{smoothL1}(G^i, G_{gt}^i) \quad (9)$$

where p_g^i and p_u^i denote probabilities of grasp sample classification into "graspable" and "ungraspable", and G^i and G_{gt}^i denote predicted and ground truth grasps, respectively. In our experiment, β is set to 1.4 to balances the contributions of grasp regression and classification.

Finally, the overall training objective is the combination of both loss terms $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{grasp}} + \lambda_{\text{c}}\mathcal{L}_{\text{cor}} \quad (10)$$

In our experiment, λ_{c} is set to 0.8 to balance the loss.

IV. Experiments

We first conduct experiments to assess the effectiveness of our proposed method on a large-scale language-driven grasping dataset [11]. We further verify our method on real robot grasping experiments. Additionally, we showcase the ablation study of our approach in language-driven grasp detection tasks. Finally, we discuss the encountered challenges and outline open questions for future research.

A. Experimental Setup

Dataset. Our experimental setup utilizes the Grasp-Anything dataset [11], a large-scale compilation of grasp data synthesized from foundational models. This dataset boasts diversity and scale, comprising 1M images with textual descriptions and featuring over 3M objects. As in [11], [61], we categorize data into 'Seen' and 'Unseen' categories, allocating 70% of categories as 'Seen' and the remaining 30% as 'Unseen'. We also use the harmonic mean ('H') metric to measure overall success rates [61].

Evaluation Metrics. Our principal evaluation metric is the success rate, as defined similarly to [24]. This necessitates that the Intersection over Union (IoU) score of the predicted grasp exceeds 25% with the ground truth grasp, and the offset angle is less than 30°. During training, we keep the text encoder and segmentation extractor fixed and then train the rest of the network end-to-end.

Baselines. We compare our method (MaskGrasp) with GR-CNN [24], Det-Seg-Refine [62], GG-CNN [63], CLIPORT [3], and CLIP-Fusion [34], utilizing either a pretrained CLIP [55] model for text embedding.

TABLE I

Language-driven grasp detection results.

Baseline	Seen	UnSeen	H	#Params	Inference time
GR-ConvNet [24] + CLIP [55]	0.37	0.18	0.24	2.07M	0.022s
Det-Seg-Refine [62] + CLIP [55]	0.30	0.15	0.20	1.82M	0.200s
GG-CNN [63] + CLIP [55]	0.12	0.08	0.10	1.24M	0.040s
CLIPORT [3]	0.36	0.26	0.29	10.65M	0.131s
CLIP-Fusion [34]	0.40	0.29	0.33	13.51M	0.157s
MaskGrasp + BERT [21] (ours)	0.47	0.43	0.42	4.91M	0.127s
MaskGrasp + CLIP [55] (ours)	**0.50**	**0.46**	**0.45**	4.72M	0.116s

B. Language-driven Grasp Detection Results

Quantitative Results. Table I shows the comparison between our method and baselines on the Grasp-Anything dataset. Our approach consistently outperforms other baselines with a clear margin in both 'Seen' and 'Unseen' setups. Moreover, in the 'Unseen' setup, our method exhibits significant superiority, surpassing the runner-up, CLIP-Fusion [34] by 0.17 in success score.

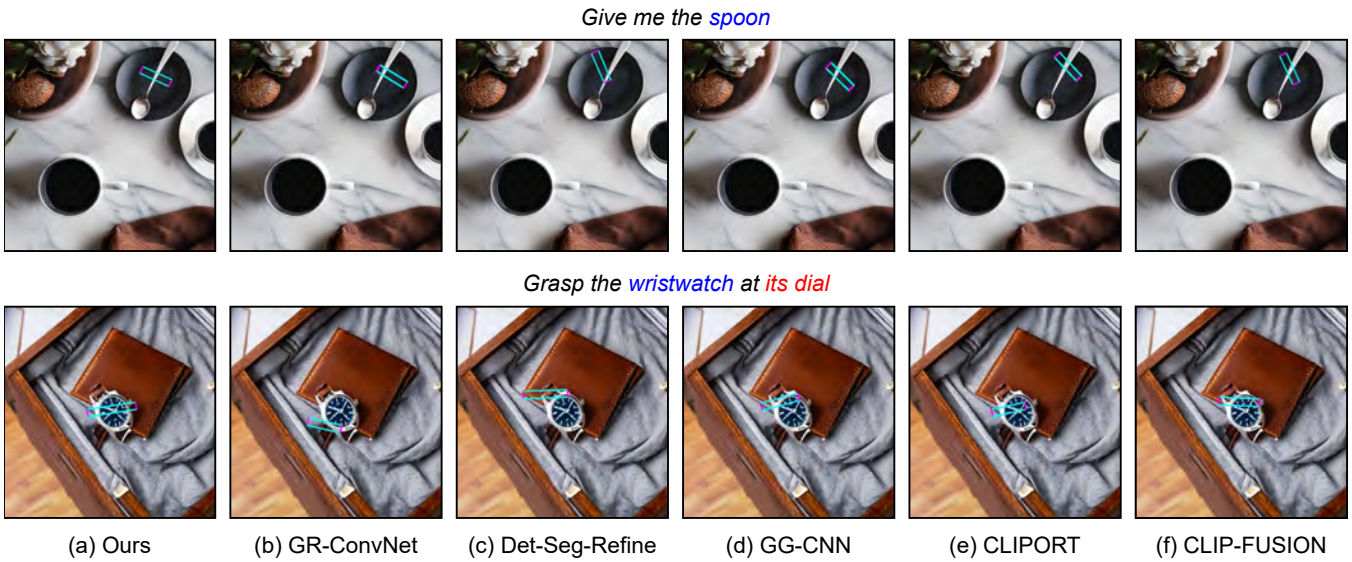

Fig. 3. Language-driven grasp detection results.

Qualitative Results. Fig. 3 shows the quantitative evaluation of our method and other baselines. This figure shows that our method produces semantically plausible results, particularly in cluttered scenes where we have occlusions.

C. Ablation Study

Mask-guided Attention Analysis. To understand how our mask-guided attention performs, we visualize the attention focus under different conditions and compare the model's prioritization of information when provided with both segmentation mask features and text instructions versus text alone. Our results, depicted in Fig. 4, indicate that our method concentrates attention on the target object more effectively when segmentation mask features are available, suggesting that these features guide the model's focus towards crucial regions and facilitate the extraction of richer contextual information, thereby improving grasp performance.

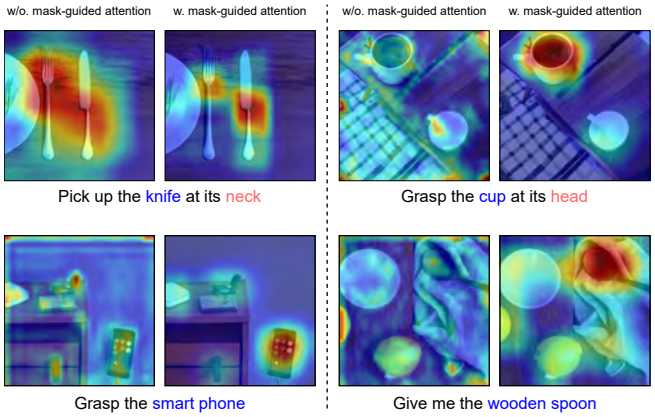

Fig. 4. The visualization comparison between using and not using our mask-guided attention.

Effectiveness of Segmentation Features. To assess the importance of segmentation mask features for grasping, we compared our model's performance with and without them. Using t-SNE [64], we cluster the grasp region feature representations z^{vis} (Equation 7) under both conditions. Fig. 5 illustrates that without correspondence loss with mask-guided features, the decision boundaries for most grasp region features are indistinct and challenging to discern during training. Conversely, applying correspondence loss with mask-guided features enhances both the accuracy and learned grasp region features of the network. These findings, supported by detailed results in Table II, underscore the positive impact of integrating mask-guided features and the corresponding loss function on grasping performance.

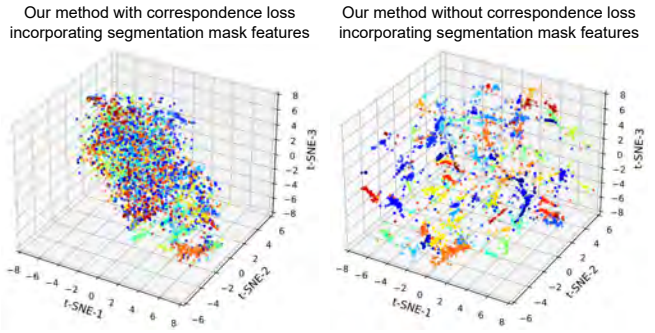

Fig. 5. **t-SNE visualization of the grasp object feature representations**. We apply t-SNE to cluster the grasp object feature representations z^{vis} of Equation 7 when using and not using the correspondence loss with mask feature objects in our method.

TABLE II
ABLATION STUDY.

Baseline	Seen	UnSeen	H
Ours w/o segmentation mask	0.432	0.314	0.349
Ours w/o correspondence loss	0.483	0.429	0.447
Ours	**0.500**	**0.460**	**0.451**

In the Wild Detection. Fig. 6 showcases visualizations produced by our method, trained solely on the Grasp-Anything dataset, when applied to random internet images and other dataset images. These results demonstrate our

TABLE III

ROBOTIC LANGUAGE-DRIVEN GRASP DETECTION RESULTS

Baseline	Single	Cluttered
GR-ConvNet [24] + CLIP [55]	0.33	0.30
Det-Seg-Refine [62] + CLIP [55]	0.30	0.23
GG-CNN [63] + CLIP [55]	0.10	0.07
CLIPORT [3]	0.27	0.30
CLIP-Fusion [34]	0.40	0.40
MaskGrasp (ours)	**0.43**	**0.42**

Fig. 6. In the wild detection results. Top row images are from GraspNet [65], YCB-Video [9] datasets; bottom row shows internet images.

model's robust generalization to real-world images, despite being trained solely on synthetic data from Grasp-Anything, without real image inputs.

D. Robotic Experiment

In Fig. 7, we showcase our robotic evaluation using a KUKA robot. Grasp detection, alongside other methods listed in Table III of the main paper, is evaluated using depth images from an Intel RealSense D435i depth camera following methodology in [24]. Our proposed method infers 4-DoF grasp poses, transformed into 6 DoF poses under the assumption of flat surface objects. Trajectory optimization detailed in [66], [67] guides the robot to target poses. The setup involves two computers: PC1 handles real-time control, camera, and gripper, while PC2 runs ROS on Ubuntu Noetic 20.04, communicating with the robot via EtherCAT protocol. PC2, equipped with an NVIDIA RTX 3080 Ti graphics card, manages the inference process. We assess performance across single-object and cluttered scenarios with a diverse set of real-world objects, repeating each experiment for all methods 30 times to ensure reliability.

Our proposed method, incorporating mask-guided attention guidance, outperforms other baselines, as shown in Table III. Remarkably, despite being trained solely on Grasp-Anything, a synthetic dataset generated by foundational models, it performs well on real-world objects.

V. DISCUSSION

Limitation. Despite significantly improving generalization capabilities, our method faces challenges when handling scenes with complex semantic object relationships and intricate geometries. Ambiguities arising from contextual cues

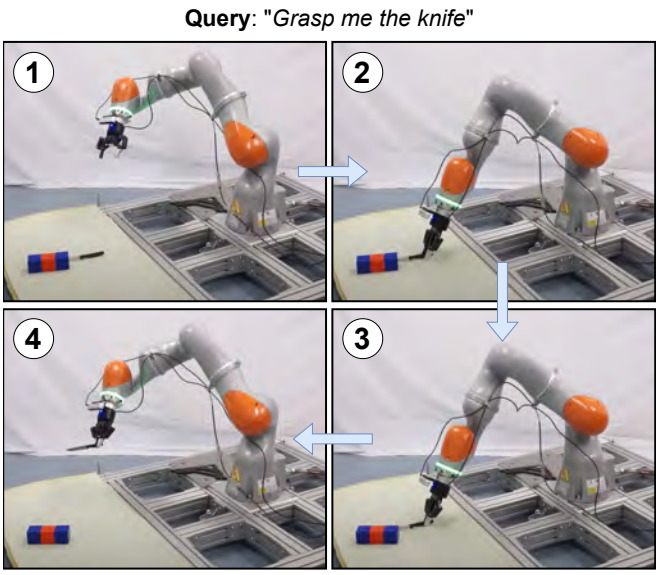

Fig. 7. The robotic experiment setup and sequence of grasping actions.

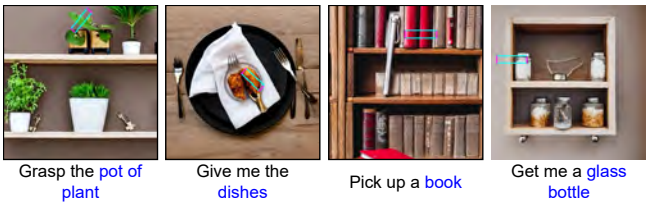

Fig. 8. Failure cases of our method.

can impede the accurate association of objects with language-driven instructions, as demonstrated by some failure cases in Fig. 8. Further refinement may be required to effectively address such scenarios.

Future work. While our study represents a new method in language-driven grasp detection with mask-guided attention, several promising avenues for future research warrant exploration. Further investigation into nuanced attention mechanisms and refinement of handling complex semantic relationships between objects and language instructions are essential [68]. Additionally, the integration of reinforcement learning techniques offers the potential for developing adaptive grasp strategies tailored to specific tasks and environments [69]. These directions hold promise for enhancing the capabilities and applicability of language-driven robotic grasping systems in real-world scenarios.

VI. CONCLUSION

We introduce a mask-guided attention mechanism to improve multimodal integration for the language-driven grasp detection task. By combining a transformer with segmentation mask-conditioned attention, our method effectively integrates visual and textual information, enhancing grasping accuracy and adaptability. This mechanism prioritizes crucial regions in both modalities, resulting in improved performance. The intensive experiments show that our method outperforms other baselines by a clear margin in vision-based benchmarks and real-world robotic grasping experiments. Our source code and trained model will be made publicly available to facilitate future studies.

REFERENCES

[1] Y. Sun, J. Falco, M. A. Roa, and B. Calli, "Research challenges and progress in robotic grasping and manipulation competitions," *IEEE Robotics and Automation Letters*, 2021.

[2] M. Gilles, Y. Chen, E. Z. Zeng, Y. Wu, K. Furmans, A. Wong, and R. Rayyes, "Metagraspnetv2: All-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping," *IEEE Transactions on Automation Science and Engineering*, 2023.

[3] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on Robot Learning*, 2022.

[4] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. Abbeel, "Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding," in *ICRA*, 2010.

[5] Y. Domae, H. Okuda, Y. Taguchi, K. Sumi, and T. Hirai, "Fast graspability evaluation on single depth maps for bin picking with general grippers," in *ICRA*, 2014.

[6] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Preparatory object reorientation for task-oriented grasping," in *IROS*, 2016.

[7] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *ICRA*, 2011.

[8] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *IROS*, 2018.

[9] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *CVPR*, 2020.

[10] S. Wang, Z. Zhou, and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE Robotics and Automation Letters*, 2022.

[11] A. D. Vuong, M. N. Vu, H. Le, B. Huang, B. Huynh, T. Vo, A. Kugi, and A. Nguyen, "Grasp-anything: Large-scale grasp dataset from foundation models," *ICRA*, 2024.

[12] R. Platt, "Grasp learning: Models, methods, and performance," *Annual Review of Control, Robotics, and Autonomous Systems*, 2023.

[13] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," in *ICCV*, 2023.

[14] T. Nguyen, M. N. Vu, A. Vuong, D. Nguyen, T. Vo, N. Le, and A. Nguyen, "Open-vocabulary affordance detection in 3d point clouds," in *IROS*, 2023.

[15] T. Van Vo, M. N. Vu, B. Huang, T. Nguyen, N. Le, T. Vo, and A. Nguyen, "Open-vocabulary affordance detection using knowledge distillation and text-point correlation," *ICRA*, 2024.

[16] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv*, 2023.

[17] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, and P. Luo, "Embodiedgpt: Vision-language pre-training via embodied chain of thought," in *NeurIPS*, 2024.

[18] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, A. Maalouf, S. Li, G. Iyer, S. Saryazdi, N. Keetha, *et al.*, "Conceptfusion: Open-set multimodal 3d mapping," *arXiv*, 2023.

[19] OpenAI, "Introducing ChatGPT," Software, accessed: February 6th 2023.

[20] Y. Mu, S. Yao, M. Ding, P. Luo, and C. Gan, "Ec2: Emergent communication for embodied control," in *CVPR*, 2023.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2018.

[22] M. A. Roa and R. Suárez, "Grasp quality measures: review and performance," *Autonomous Robots*, 2015.

[23] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International Journal of Robotics Research*, 2020.

[24] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *IROS*, 2020.

[25] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, 2015.

[26] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *ICRA*, 2016.

[27] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv*, 2017.

[28] 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 *IROS*, 2023.

[29] Q. Sun, H. Lin, Y. Fu, Y. Fu, and X. Xue, "Language guided robotic grasping with fine-grained instructions," in *IROS*, 2023.

[30] T. Nguyen, M. N. Vu, B. Huang, A. Vuong, Q. Vuong, N. Le, T. Vo, and A. Nguyen, "Language-driven 6-dof grasp detection using negative prompt guidance," in *ECCV*, 2024.

[31] C. Cheang, H. Lin, Y. Fu, and X. Xue, "Learning 6-dof object poses to grasp category-level objects by language instructions," in *ICRA*, 2022.

[32] A. D. Vuong, M. N. Vu, B. Huang, N. Nguyen, H. Le, T. Vo, and A. Nguyen, "Language-driven grasp detection," in *CVPR*, 2024.

[33] N. Nguyen, M. N. Vu, B. Huang, A. Vuong, N. Le, T. Vo, and A. Nguyen, "Lightweight language-driven grasp detection using conditional consistency model," in *IROS*, 2024.

[34] 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 *ICRA*, 2023.

[35] Y. Yang, X. Lou, and C. Choi, "Interactive robotic grasping with attribute-guided disambiguation," in *ICRA*, 2022.

[36] G. Tziafas, Y. Xu, A. Goel, M. Kasaei, Z. Li, and H. Kasaei, "Language-guided robot grasping: Clip-based referring grasp synthesis in clutter," *arXiv*, 2023.

[37] Y. Chen, R. Xu, Y. Lin, and P. A. Vela, "A joint network for grasp detection conditioned on natural language commands," in *ICRA*, 2021.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[39] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *CVPR*, 2020.

[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.

[41] C. Sun, F. Baradel, K. Murphy, and C. Schmid, "Learning video representations using contrastive bidirectional transformer," *arXiv*, 2019.

[42] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020.

[43] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *arXiv*, 2020.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2020.

[45] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICLR*, 2021.

[46] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *CVPR*, 2021.

[47] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *ICCV*, 2021.

[48] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, 2021.

[49] G. Sharir, A. Noy, and L. Zelnik-Manor, "An image is worth 16x16 words, what is a video worth?" *arXiv*, 2021.

[50] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *ICCV*, 2021.

[51] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robotics and Automation Letters*, 2018.

[52] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "Roi-based robotic grasp detection for object overlapping scenes," in *IROS*, 2019.

[53] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, "Learning rgb-d feature embeddings for unseen object instance segmentation," in *CoRL*, 2021.

[54] S. Back, J. Lee, T. Kim, S. Noh, R. Kang, S. Bak, and K. Lee, "Unseen object amodal instance segmentation via hierarchical occlusion modeling," in *ICRA*, 2022.

[55] 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 *ICML*, 2021.

[56] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[57] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv*, 2017.
[58] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
[59] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *ICCV*, 2019.
[60] Z. Yang, S. Zhang, L. Wang, and J. Luo, "Sat: 2d semantics assisted training for 3d visual grounding," in *ICCV*, 2021.
[61] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *CVPR*, 2022.
[62] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *ICRA*, 2021.
[63] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv*, 2018.
[64] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of Machine Learning Research*, 2008.
[65] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv*, 2017.
[66] F. Beck, M. N. Vu, C. Hartl-Nesic, and A. Kugi, "Singularity avoidance with application to online trajectory optimization for serial manipulators," *IFAC-PapersOnLine*, 2023.
[67] M. N. Vu, F. Beck, M. Schwegel, C. Hartl-Nesic, A. Nguyen, and A. Kugi, "Machine learning-based framework for optimally solving the analytical inverse kinematics for redundant manipulators," *Mechatronics*, 2023.
[68] M. Chen, I. Laina, and A. Vedaldi, "Training-free layout control with cross-attention guidance," in *WACV*, 2024.
[69] S. Nasiriany, H. Liu, and Y. Zhu, "Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks," in *ICRA*, 2022.