WSGS: A Speech-Driven Zero-Shot System for 6D Robotic Arm Grasping

Yitong Ge, Lin Zhang*, Yang Chen, Ying Shen
School of Computer Science and Technology, Tongji University, Shanghai, China
{2411499, cslinzhang, 2011439, yingshen}@tongji.edu.cn

***Abstract*—In the robotic vision industry, recent years have witnessed a growing interest in object detection and pose estimation for precise robotic arm grasping. In fact, various unfavorable factors, such as the size limitations of robotic grippers, the diversity and potential complexity of object shapes and poses, and the cluttered environment, make robotic arm grasping based on 6D object poses much harder than it seems. In this paper, to solve these issues to some extent, we proposed a speech-driven zero-shot system for robotic arm grasping, called WSGS (Whisper-SAM6D Grasping System). It enables speech-driven human-interactive grasping with the Franka Emika robotic arm by performing instance segmentation and pose estimation based on speech instructions. Specifically, WSGS accurately recognizes the 6D pose of an unknown object and adapts to its size to find the most suitable position for grasping. Comprehensive experiments on our real-world scenarios demonstrate that WSGS can produce high-accuracy instance segmentation and pose estimation results, achieving adaptive robotic arm grasping of unknown objects based on speech commands.**

***Index Terms*—Robotic arm grasping, instance segmentation, pose estimation, semantic understanding**

I. Introduction

In recent years, the rapid advancement of robotics and autonomous systems has significantly expanded the scope of robotic applications in tasks such as assembly, advanced manufacturing, human-robot collaboration, and inter-robot cooperation. To effectively complete these tasks, robots with excellent grasping capabilities are necessary [1]. Current robotic arm grasping techniques are primarily classified into two major categories: analytical methods [2]–[4] and data-driven methods [5]. The former relies on the geometric information of objects to compute grasp poses. However, its performance significantly deteriorates when applied to unknown objects or unstructured scenarios. The latter, grounded in deep learning, learns grasping strategies from large-scale datasets. Nevertheless, it is highly dependent on extensive training data and computational resources.

To successfully grasp objects, accurate instance segmentation and pose estimation are indispensable challenges. Over the past few years, several innovative frameworks have been proposed to address these challenges, such as the Segment Anything Model (SAM) [6] and SAM-6D [7]. These frameworks have demonstrated significant potential in zero-shot instance segmentation and pose estimation tasks. Despite the excellent performance of SAM-6D, its application and integration in practical grasping tasks remain relatively limited. Specifically, existing research has yet to provide an effective solution for the seamless integration of pose estimation of unknown objects in the real world with robotic arm grasping tasks. Moreover, the current grasping systems necessitate the necessary flexibility, especially in extracting user intent from natural language commands and executing corresponding grasping tasks. Most existing methods are insufficient to simultaneously handle arbitrary, non-fixed template speech commands and grasp unknown objects. Therefore, a more advanced user interaction mechanism is urgently required, enabling users to directly control robotic arms via speech commands to accomplish zero-shot grasping tasks.

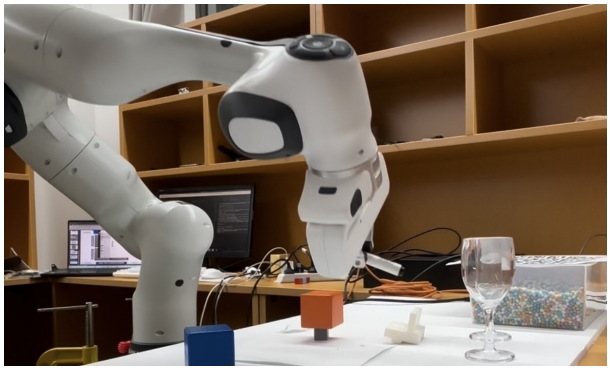

Fig. 1. The robotic arm successfully executed the grasping task in response to the speech command "Help me grasp the whiteboard marker on the desk".

To address the above limitations, this paper introduces a speech-driven zero-shot robotic arm grasping system, named Whisper-SAM6D Grasping System (**WSGS**). Designed specifically for human-interactive grasping tasks using the *Franka Emika* robotic arm, WSGS integrates speech command recognition, semantic understanding, instance segmentation, pose estimation, and robotic arm grasping into a unified framework. WSGS first employs the Whisper model to transcribe speech commands into text and then extracts key objects from the instructions based on a Transformer-based natural language processing model, BERT [8]. Secondly, in the segmentation module of WSGS, the Semantic-based Instance Segmentation Model assigns matching scores by combining semantic, appearance, geometric, and contextual information to candidate objects, enabling the identification of novel objects in scenarios where objects are occluded. After that, the Semantic-

* Corresponding author: cslinzhang@tongji.edu.cn.

enhanced Pose Estimation Model uses the point registration approach to estimate the 6D bounding box of the target object relative to the candidate instance. Notably, the semantic embeddings from Whisper and BERT are combined with geometric features to enhance the pose estimation process, using attention mechanisms to weigh both semantic and geometric features for improved accuracy. Finally, considering the 6D bounding box of the object and the gripper size limitation, a proper grasping angle for the robotic arm can be determined, guaranteeing that WSGS can achieve grasping at a high successful rate.

Our contributions are summarized as follows:

- A novel speech-driven zero-shot 6D grasping pipeline for the robotic arm, WSGS. WSGS performs zero-shot instance segmentation by incorporating semantic information prompts. The segmentation results are then semantic-enhanced and matched with objects to obtain precise 6D pose estimation, enabling 6D grasping of unknown objects in complex environments.
- The first robotic arm grasping strategy that considers object size and gripper size limitations. Traditional grasping strategies often neglect the size constraints of robotic arm grippers. WSGS estimates the object's pose using stereo bounding boxes by combining 6D pose information and CAD models. The narrowest edge of the bounding cube is then selected for grasping, effectively handling a wide range of complex object shapes and orientations.
- Extensive experimental results demonstrate the effectiveness of WSGS. Based on WSGS, the *Franka Emika* robotic arm successfully executed 6D grasping tasks for unknown objects in various situations, showcasing the robustness and practicality of the proposed method in real-world scenarios.

II. RELATED WORK

A. Object Detection and Segement

Object detection is a key task in computer vision, with significant progress driven by deep learning. Methods like YOLO [9], [10] and Faster R-CNN [11] are widely used for their accuracy and efficiency. Transformer-based models such as DETR [12] have further advanced the field by enabling end-to-end detection with attention mechanisms. Despite advancements in accuracy and efficiency, existing methods face challenges. YOLO and Faster R-CNN struggle with generalization to unknown objects or cross-domain scenarios. Traditional methods provide bounding box information but lack precise object shapes or semantic segmentation. While DETR performs well in complex scenes, its high computational cost limits real-time use. These issues have led research towards object segmentation, which improves scene understanding. The Segment Anything Model (SAM) [6] has made breakthroughs in zero-shot object segmentation, but it focuses on segmentation rather than directly addressing pose estimation or grasp planning for robotic manipulation.

B. Pose Estimation of Unknown Objects

Pose estimation is vital for robotic manipulation and augmented reality, especially with unknown objects. Traditional methods like Gen6D [13] and MegaPose [14] perform well for known objects but struggle with unseen ones. Approaches like OnePose [15] and ZeroPose [16] improve generalization using structure-from-motion and geometric matching, but face challenges with occlusion, segmentation errors, and efficiency. The SAM-6D framework [7] combines SAM's segmentation with pose estimation, enhancing accuracy through a two-stage process and a Sparse-to-Dense point transformer. Despite its success in benchmarks, its application in robotic arm tasks remains limited.

C. Adaptive Grasping

Adaptive grasping has been challenging due to complex object shapes. Early methods predicted grasp points based on object geometry [17], but were computationally expensive. Data-driven deep learning approaches have become dominant [18]. Bohg *et al*. [5] and Kroemer *et al*. [19] used 3D model databases and active learning for better accuracy, while supervised learning [20], [21] enabled grasp planning for unknown objects from RGB images. However, these methods require large computational resources and data. WSGS improves grasping by using predefined points and pose estimation to adapt to object position and orientation, offering robust, low-complexity grasping.

III. METHOD

In this section, the proposed WSGS framework will be described in detail, and its structural diagram is shown in Fig. 2. WSGS first utilizes the Whisper model for speech recognition, converting speech commands into text. The BERT model is then applied to extract key object features, which is subsequently used to guide the Semantic-based Instance Segmentation Model (SISM) in generating more accurate segmentation prompts. During the segmentation phase, the framework calculates matching scores by combining semantic, appearance, geometric, and contextual information. Following this, the Semantic-enhanced Pose Estimation Model (SPEM) is employed to estimate the 6D pose of the target object. Finally, WSGS computes a compact 6D bounding box for the object and selects the optimal grasping direction and position.

A. Speech Recognition and Semantic Understanding

In practice, complex robotic arm grasping tasks in real-world scenarios require not only static image information but also the ability to understand and respond to external commands. Traditional ISM methods [7] rely on prompt generation based on image content, which can lead to redundant or missed segmentations when confronted with cluttered environments. To achieve more precise recognition and segmentation of target objects, WSGS introduces Whisper and BERT as front-end processing modules, using speech commands and natural language descriptions to generate prompts that are better suited to grasping instructions. Due to its outstanding performance

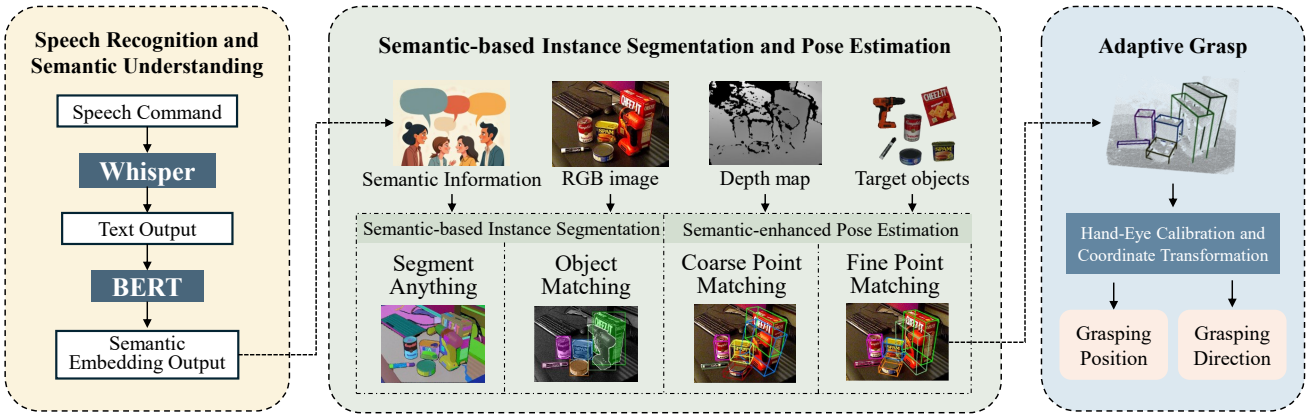

Fig. 2. The architecture sketch of the proposed WSGS system. In WSGS, semantic information is extracted from speech commands and used to guide instance segmentation and improve pose estimation. Finally, the results of pose estimation are used to guide the robotic arm grasping.

in multilingual speech recognition and strong adaptability to background noise and dialects, the Whisper model was utilized to generate real-time textual outputs from speech commands, providing a reliable data source for semantic understanding of the target object:

$$\mathcal{P}_{\text{speech}} = \text{Whisper}(\mathcal{S}), \tag{1}$$

where $\mathcal{S}$ represents the input speech signal, and $\mathcal{P}_{\text{speech}}$ is the textual instruction output by Whisper.

Subsequently, WSGS utilizes a pre-trained BERT model to perform Named Entity Recognition (NER). BERT is used to interpret and embed the text output from Whisper semantically:

$$\mathcal{P}_{\text{sem}} = \text{BERT}(\mathcal{P}_{\text{speech}}), \tag{2}$$

where $\mathcal{P}_{\text{sem}}$ represents the semantic embedding output by BERT, which encodes key features of the target object, such as color and location. This semantic embedding can be used as input for the improved prompt generation strategy, which, in combination with prior information from speech commands, guides the instance segmentation model to focus more accurately on the region containing the target object, thereby generating dynamic and dense prompts tailored to the target object.

B. Instance Segmentation and Pose Estimation

Semantic-based Instance Segmentation Model (SISM). In the instance segmentation stage, WSGS employs SISM, an improved Instance Segmentation Model (ISM) proposed by us, for semantic-guided segmentation of target object instances in cluttered scenes. Specifically, for an input RGB image $\mathcal{I}$, SISM performs segmentation with the assistance of the Segment Anything Model (SAM) [6] based on prompts $\mathcal{P}_r$, using three modules: an image encoder Φ_{Image}, a prompt encoder Φ_{Prompt}, and a mask decoder Ψ_{Mask}. The segmentation process is formalized as:

$$\mathcal{M}, \mathcal{C} = \Psi_{\text{Mask}}(\Phi_{\text{Image}}(\mathcal{I}), \Phi_{\text{Prompt}}(\mathcal{P}_r)), \tag{3}$$

where $\mathcal{M}$ and $\mathcal{C}$ denote the predicted proposals and their corresponding confidence scores, respectively. Traditional ISM methods usually employ zero-shot transfer to generate all possible segmentation prompts $\mathcal{P}_r$ with a uniform 2D grid, which may lead to redundancy or omissions. In contrast, our SISM generates more focused proposals by incorporating semantic-guided generation, leveraging target object information from speech commands. Specifically, by integrating semantic embeddings, (3) can be rewritten as:

$$\widetilde{\mathcal{M}}, \widetilde{\mathcal{C}} = \Psi_{\text{Mask}}(\Phi_{\text{Image}}(\mathcal{I}), \Phi_{\text{Prompt}}(\mathcal{P}_r, \mathcal{P}_{\text{sem}})). \tag{4}$$

From the candidate set $\mathcal{M}$, ISM assigns a matching score s_m to each candidate $m \in \mathcal{M}$, which is used to identify instances matching the target object $\mathcal{O}$. The matching score s_m is evaluated based on three aspects [7]: semantic similarity s_{sem}, appearance similarity s_{appe}, and geometric similarity s_{geo}. To account for occlusion, a visibility ratio r_{vis} is introduced to adjust the weight of the geometric score dynamically.

For the appearance matching score s_{appe}, a weighting strategy is introduced in SISM to enhance the precision of matching for regions surrounding the target object, as such information is often provided in speech commands. The updated weighted appearance matching score is formulated as follows:

$$\widetilde{s_{\text{appe}}} = \frac{1}{N_{\mathcal{I}_m}^{\text{patch}}} \sum_{j=1}^{N_{\mathcal{I}_m}^{\text{patch}}} \max_{i=1,\dots,N_{T_{\text{best}}}^{\text{patch}}} \frac{\boldsymbol{w}_j \cdot \langle \boldsymbol{f}_{\mathcal{I}_m,j}^{\text{patch}}, \boldsymbol{f}_{T_{\text{best}},i}^{\text{patch}} \rangle}{\| \boldsymbol{f}_{\mathcal{I}_m,j}^{\text{patch}} \| \cdot \| \boldsymbol{f}_{T_{\text{best}},i}^{\text{patch}} \|}, \tag{5}$$

where $\boldsymbol{w}_j$ is a weighting factor emphasizing regions around the target object, $N_{\mathcal{I}_m}^{\text{patch}}$ is the number of patches in the candidate region $\mathcal{I}_m$, and $\boldsymbol{f}_{\mathcal{I}_m,j}^{\text{patch}}$ is the feature vector of the j-th patch in $\mathcal{I}_m$. Similarly, $N_{T_{\text{best}}}^{\text{patch}}$ denotes the number of patches in the target template T_{best}, with $\boldsymbol{f}_{T_{\text{best}},i}^{\text{patch}}$ as the feature vector of the i-th patch in T_{best}.

In addition, SISM incorporates contextual information from speech commands, such as object positions, to optimize segmentation and matching, significantly improving success rates in complex environments. For commands with positional information, SISM focuses on the relevant region for precise segmentation and matching of the target object. The degree of

alignment between the target object and its relative position to other objects is represented by the contextual matching score:

$$s_{\text{context}} = \exp\left(-\frac{d(\boldsymbol{p}, \boldsymbol{p}_k)^2}{2\sigma^2}\right), \tag{6}$$

where $\boldsymbol{p}$ represents the target object's position, $\boldsymbol{p}_k$ the position of a known object, and σ the standard deviation of the gaussian function, controlling the spatial distance influence on the contextual matching score. A smaller σ increases the impact of closer objects. The improved matching score calculation is thus formulated as follows:

$$\widetilde{s_m} = \frac{s_{\text{sem}} + \widetilde{s_{\text{appe}}} + w_{\text{context}} \cdot s_{\text{context}} + r_{\text{vis}} \cdot s_{\text{geo}}}{1 + 1 + w_{\text{context}} + r_{\text{vis}}}, \tag{7}$$

where w_{context} is the weighting coefficient for contextual information, reflecting the importance of the contextual object in matching. Higher weights are assigned to objects that are spatially closer to the target.

Semantic-enhanced Pose Estimation Model (SPEM). After the instance segmentation, WSGS utilizes an improved Pose Estimation Model (PEM) which further incorporates semantic features from speech commands to predict the 6D pose of the target object $\mathcal{O}$ corresponding to a candidate region. For each candidate object m, PEM employs a point registration approach to estimate the 6D pose of m relative to $\mathcal{O}$. To incorporate the semantic information from speech commands, SPEM first concatenates the semantic embeddings $\mathcal{P}_{\text{sem}}$ with the geometric features $\boldsymbol{F}_m$ and $\boldsymbol{F}_o$ to form an enhanced feature representation. The attention matrix $\boldsymbol{A}$ is then designed to include the semantic features as follows:

$$\boldsymbol{A} = \left[\boldsymbol{f}_m^{\text{bg}}, \boldsymbol{F}_m, \mathcal{P}_{\text{sem}}^m\right] \times \left[\boldsymbol{f}_o^{\text{bg}}, \boldsymbol{F}_o, \mathcal{P}_{\text{sem}}^o\right]^T \in \mathbb{R}^{(N_m+1)\times(N_o+1)}, \tag{8}$$

where $\boldsymbol{f}_m^{\text{bg}} \in \mathbb{R}^C$ and $\boldsymbol{f}_o^{\text{bg}} \in \mathbb{R}^C$ are background feature vectors for candidate object m and target object $\mathcal{O}$, respectively. $\boldsymbol{F}_m \in \mathbb{R}^{N_m \times C}$ and $\boldsymbol{F}_o \in \mathbb{R}^{N_o \times C}$ are feature matrices of the point sets with N_m and N_o as the number of points, respectively. $\mathcal{P}_{\text{sem}}^m$ and $\mathcal{P}_{\text{sem}}^o$ are the semantic features of the candidate and target objects, respectively.

Next, the attention matrix $\boldsymbol{A}$ undergoes a normalization procedure to obtain the soft assignment matrix $\tilde{\boldsymbol{A}}$:

$$\tilde{\boldsymbol{A}} = \text{Softmax}_{\text{row}}\left(\frac{\boldsymbol{A}}{\tau}\right) \cdot \text{Softmax}_{\text{col}}\left(\frac{\boldsymbol{A}}{\tau}\right), \tag{9}$$

where $\boldsymbol{A}$ is the semantic-enhanced attention matrix, $\text{Softmax}_{\text{row}}()$ and $\text{Softmax}_{\text{col}}()$ apply Softmax along the rows and columns, respectively, with τ as a constant temperature. The values in each row of $\tilde{\boldsymbol{A}}$ (excluding the background row) represent matching probabilities between point $\boldsymbol{p}_m \in \mathcal{P}_m$ and points in $\mathcal{P}_o$, with the matching point $\boldsymbol{p}_o \in \mathcal{P}_o$ determined by the maximum value's index.

After obtaining $\tilde{\boldsymbol{A}}$, the matching pairs $\{(\boldsymbol{p}_m, \boldsymbol{p}_o)\}$ and their scores are collected, and Weighted Singular Value Decomposition (SVD) is applied to compute the final pose:

$$\boldsymbol{R}, t = \text{SVD}(\boldsymbol{W}_{ij}(\mathcal{P}_m^f, \mathcal{P}_o^f)), \tag{10}$$

where $\boldsymbol{W}_{ij}$ is the weight matrix representing the weighted relationships between the point pairs, which combines both geometric similarity and semantic similarity:

$$\boldsymbol{W}_{ij} = \alpha \cdot s_{\text{geo}}(\boldsymbol{p}_m^i, \boldsymbol{p}_o^j) + \beta \cdot s_{\text{sem}}(\boldsymbol{p}_m^i, \boldsymbol{p}_o^j), \tag{11}$$

where α and β are the weight coefficients that balance the contributions of each similarity measure.

C. Adaptive Grasp

With SPEM, the compact 6D bounding box of the detected target object can be obtained. This allows for the approximate description of the object's shape and pose using a simple cubic approximation, which can then be employed to guide the selection of grasping positions and angles. Unlike traditional grasping strategies that rely solely on detailed geometric features of the object, the adaptive grasping strategy of WSGS offers a solution to the grasping problem for objects with complex shapes. Through the compact bounding box, the standard grasping point is defined as the cube center, denoted by $\boldsymbol{c}$. Specifically, the initial direction vector $\boldsymbol{a}$ of the object is selected, where $\boldsymbol{a}$ represents the direction vector of the longest edge of the bounding box. Using the rotation matrix $\boldsymbol{R}$ output by SPEM, the object direction vector $\boldsymbol{a}$ is updated in real-time to the current direction $\boldsymbol{a'}$:

$$\boldsymbol{a'} = \boldsymbol{R} \cdot \boldsymbol{a}, \tag{12}$$

where $\boldsymbol{a'}$ represents the current direction vector. Considering the limited size of the gripper, the plane of the grasping direction is selected to be perpendicular to the current direction vector. This ensures that the gripper can grasp the object at its minimum cross-sectional area. To ensure grasping stability, the robotic arm is directed to choose a direction that is perpendicular to the current direction vector $\boldsymbol{a'}$ and close to the gravity direction as the grasping direction. This approach effectively maximizes the stability of the gripper-object contact, avoiding sliding or instability due to an improper grasping angle. Additionally, selecting a direction close to gravity helps reduce interference from external forces during the grasping process, ensuring that the object remains firmly in place after being grasped, thereby reducing mechanical strain and improving both the success rate and accuracy of the grasping. Next, we will introduce our grasping strategy mathematically in detail.

Assuming the gravity direction unit vector is $\boldsymbol{g} = [0, 0, -1]^T$, $\boldsymbol{g}$ should be projected onto $\boldsymbol{a'}$ to obtain the component of $\boldsymbol{g}$ in the direction of $\boldsymbol{a'}$, denoted by $\boldsymbol{proj}_{\boldsymbol{a'}}(\boldsymbol{g})$:

$$\boldsymbol{proj}_{\boldsymbol{a'}}(\boldsymbol{g}) = \frac{\boldsymbol{g} \cdot \boldsymbol{a'}}{\boldsymbol{a'} \cdot \boldsymbol{a'}} \boldsymbol{a'}, \tag{13}$$

where $\cdot$ represents the dot product operation. The part of $\boldsymbol{g}$ parallel to $\boldsymbol{a'}$ is removed, resulting in

$$\boldsymbol{g'} = \boldsymbol{g} - \boldsymbol{proj}_{\boldsymbol{a'}(\boldsymbol{g})}, \tag{14}$$

where $\boldsymbol{g'}$ is perpendicular to $\boldsymbol{a'}$. Finally, normalizing $\boldsymbol{g'}$ yields the grasping vector $\boldsymbol{v}$:

$$\boldsymbol{v} = \frac{\boldsymbol{g'}}{\|\boldsymbol{g'}\|}. \tag{15}$$

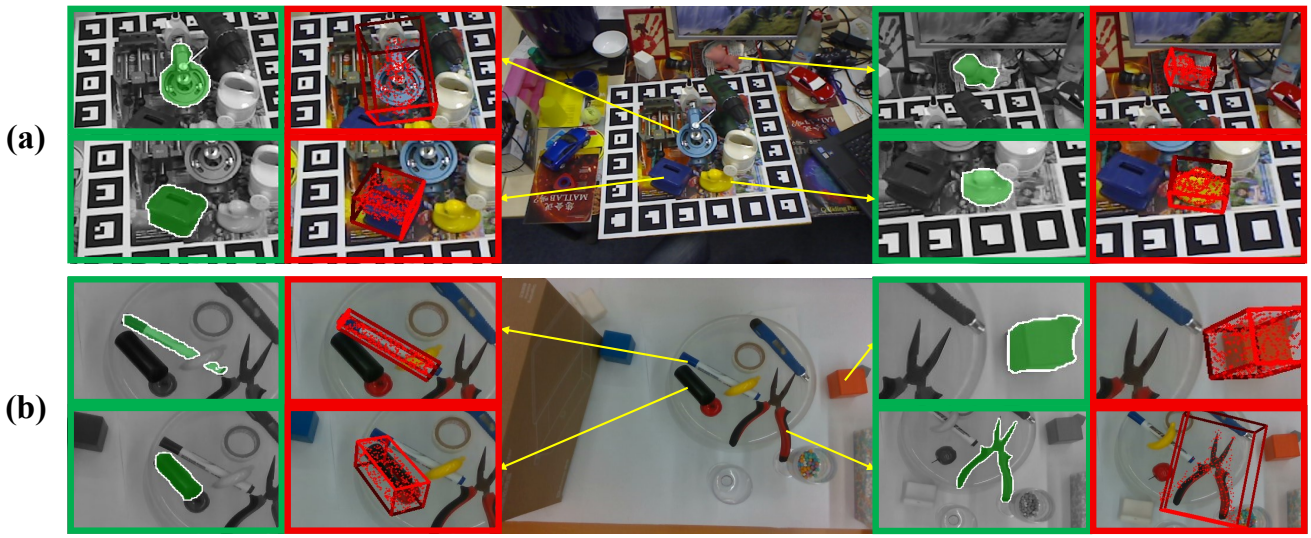

Fig. 3. The results of instance segmentation (within green bounding boxes) and pose estimation (within red bounding boxes). Specifically, (a) illustrates the performance of WSGS on the BOP dataset and (b) presents the results applied to unknown objects in a real laboratory environment.

TABLE I
THE PERFORMANCE OF COMPARED METHODS ON INSTANCE SEGMENTATION, POSE ESTIMATION, AND ROBOTIC ARM GRASPING IN REAL-WORLD SCENARIOS.

Methods	S-AOI (%)	E-CPE (cm)	RE (°)	U-GSR (%)	O-GSR (%)
CNOS	90.4	-	-	-	-
ZeroPose	89.5	0.9	6.1	-	-
MegaPose	91.2	0.7	5.8	-	-
O3DGSA	-	1.2	-	78.5	71.0
GraspNeRF	-	0.8	5.9	89.0	81.5
WSGS (Ours)	**91.7**	**0.6**	**5.6**	**92.5**	**90.0**

With the cube center $\boldsymbol{c}$ and the grasping vector $\boldsymbol{v}$, accurate grasping of the target object can then be conducted effectively.

IV. EXPERIMENT

A. Experimental Setup

The experiments were conducted in an environment configured with Ubuntu 20.04 as the operating system and Python 3.8.10 as the development language. The hardware included an NVIDIA Corporation Device 2204 (rev a1) GPU, running with CUDA 11.3 and PyTorch version 1.10.1+cu113. The *Franka Emika* robotic arm and the Realsense D435i depth camera were utilized. The control of the robotic platform was implemented using the Franka Control Interface (FCI). Experiments were performed on the core dataset LM-O of the BOP benchmark [22] and in our real laboratory environment equipped with a Franka robotic arm.

B. Quantitative Experiments

We compare our method with the following baselines:

- CNOS [23]: a novel method for segmenting unseen objects in RGB images using their CAD models.
- ZeroPose [16]: a zero-shot 6D pose estimation framework that enables fast, model-free pose estimation of novel objects using a Discovery-Orientation-Registration (DOR) pipeline.
- MegaPose [14]: a method for 6D pose estimation of novel objects, using a render&compare strategy and a coarse pose estimation network, without retraining.
- O3DGSA [17]: a computational algorithms designed to generate stable and effective 3D object grasps for autonomous multi-fingered robotic hands.
- GraspNeRF [24]: a multiview RGB-based 6-DoF grasp detection network.

We measure the performance of compared methods above by Segmentation Accuracy with Occluded Instances (**S-AOI**), Error in Centre Position Estimation (**E-CPE**), Rotation Error (**RE**), Unobstructed Grasp Success Rate (**U-GSR**), and Obstructed Grasp Success Rate (**O-GSR**).

The experiment was conducted on 40 instances at five test positions on three aspects, involving instance segmentation, pose estimation, and robotic arm grasping trials. Two of these test positions involved occlusion of the target objects, for which speech instructions with positional information were provided during the grasping process. The results are summarized in Table I. It is observed that WSGS outperforms other methods in segmentation accuracy when handling occluded instances and achieves higher precision in estimating the target object's center position. Additionally, WSGS shows the smallest 6D pose estimation error and significantly improves the grasping success rate compared to previous methods. Moreover, in the presence of occlusion, WSGS shows minimal performance degradation, maintaining a 90.0% success rate in object grasping. These findings demonstrate that WSGS provides high robustness and accuracy across multiple evaluation tasks, particularly excelling in handling grasping tasks with occlusion.

Fig. 4. The robotic arm successfully achieves 6D grasping of unknown objects with different poses in complex real-world environments.

C. Qualitative Experiments

Based on our SISM model, WSGS demonstrates outstanding object segmentation and matching capabilities in complex backgrounds and scenarios where objects are occluded. Besides, leveraging the SPEM model, the system achieves excellent point cloud matching and 6D pose estimation performance in challenging environments. The visualization results of both SISM and SPEM are shown in Fig. 3, where it can be observed that the estimated results of WSGS are highly consistent with the ground truth. Based on accurate estimation results, WSGS, using an adaptive grasping strategy, determines the optimal grasping angle and endpoint position. An example of grasping in a real laboratory environment is shown in Fig. 4. The results indicate that WSGS is capable of generating feasible and reasonable grasp positions and poses, enabling successful grasping of objects with a maximum width exceeding the gripper's size, even in cases where the gripper's size is limited. However, a failure case was observed when attempting to grasp soft and easily deformable objects. In the absence of a force feedback mechanism, the application of excessive grasping force frequently resulted in damage to the objects.

V. CONCLUSION

In this paper, the Whisper-SAM6D Grasping System (WSGS) was proposed as a robust framework for object manipulation based on speech commands in real-world environments. In WSGS, initially, speech recognition and semantic understanding are performed. Following that, semantic information guides the Semantic-based Instance Segmentation Model (SISM) to generate accurate segmentation prompts and optimize matching scores. Then, the Semantic-enhanced Pose Estimation Model (SPEM) uses point registration, combining semantic embeddings with geometric features of candidate and target objects to estimate the object's 6D pose. Ultimately, WSGS uses an adaptive grasping mechanism based on the object's compact bounding box for stable and efficient handling of objects with varying shapes. Real-world experiments validate the system's effectiveness, achieving high accuracy in diverse grasping tasks.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62272343 and in part by the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] H. Hodson, "A gripping problem: Designing machines that can grasp and manipulate objects with anything approaching human levels of dexterity is first on the to-do list for robotics," *Nature*, pp. 24–26, 2018.

[2] A. Sahbani, S. El-Khoury, and P. Bidaud, "An overview of 3D object grasp synthesis algorithms," *Rob. Auton. Syst.*, pp. 326–336, 2012.

[3] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *ICRA*, 2000, pp. 348–353.

[4] K. B. Shimoga, "Robot grasp synthesis algorithms: A survey," *Int. J. Rob. Res.*, pp. 230–266, 1996.

[5] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—A survey," *IEEE Trans. Robot.*, pp. 289–309, 2014.

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, et al., "Segment anything," in *ICCV*, 2023, pp. 3992–4003.

[7] J. Lin, L. Liu, D. Lu, and K. Jia, "SAM-6D: Segment anything model meets zero-shot 6D object pose estimation," in *CVPR*, 2024, pp. 27906–27916.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[10] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding Yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, pp. 91–99, 2015.

[12] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020, pp. 213–229.

[13] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T Komura, and W. Wang, "Gen6D: Generalizable model-free 6-DoF object pose estimation from RGB images," in *ECCV*, 2022, pp. 298–315.

[14] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, et al., "MegaPose: 6D pose estimation of novel objects via render & compare," in *PoRL*, 2023, pp. 715–725.

[15] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou, "OnePose: One-shot object pose estimation without CAD models," in *CVPR*, 2022, pp. 6825–6834.

[16] J. Chen, M. Sun, T. Bao, R. Zhao, L. Wu, and Z. He, "3D model-based zero-shot pose estimation pipeline," *arXiv preprint arXiv:2305.17934*, 2023.

[17] A. Sahbani, S. El-Khoury, and P. Bidaud, "An overview of 3D object grasp synthesis algorithms," *Rob. Auton. Syst.*, pp. 326–336, 2012.

[18] A. Ghodake, P. Uttam, and B. B. Ahuja, "Accurate 6-DOF grasp pose detection in cluttered environments using deep learning," in *I4Tech*, 2022, pp. 1–6.

[19] O. Kroemer, C. Daniel, G. Neumann, H. van Hoof, and J. Peters, "Active learning using mean shift optimization for robot grasping," in *IROS*, 2009.

[20] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Rob. Res.*, pp. 157–173, 2008.

[21] L. Montesano and M. Lopes, "Active learning of visual descriptors for grasping using non-parametric smoothed beta distributions," *Rob. Auton. Syst.*, pp. 452–462, 2012.

[22] M. Sundermeyer, T. Hodan, Y. Labbe, G. Wang, E. Brachmann, B. Drost, et al., "BOP challenge 2022 on detection, segmentation and pose estimation of specific rigid objects," in *CVPR*, 2023, pp. 2784–2793.

[23] V. N. Nguyen, T. Groueix, G. Ponimatkin, V. Lepetit, and T. Hodan, "CNOS: A strong baseline for CAD-based novel object segmentation," in *ICCVW*, 2023, pp. 2126–2132.

[24] Q. Dai, Y. Zhu, Y. Geng, C. Ruan, J. Zhang, and H. Wang, "GraspNeRF: Multiview-based 6-DoF grasp detection for transparent and specular objects using generalizable NeRF," in *ICRA*, 2023, pp. 1757–1763.