2022 IEEE International Conference on Robotics and Automation (ICRA)
May 23-27, 2022. Philadelphia, PA, USA

Audio-Visual Grounding Referring Expression for Robotic Manipulation

Yefei Wang†, Kaili Wang†, Yi Wang†, Di Guo, Huaping Liu*, Fuchun Sun

Abstract— **Referring expressions are commonly used when referring to a specific target in people's daily dialogue. In this paper, we develop a novel task of audio-visual grounding referring expression for robotic manipulation. The robot leverages both the audio and visual information to understand the referring expression in the given manipulation instruction and the corresponding manipulations are implemented. To solve the proposed task, an audio-visual framework is proposed for visual localization and sound recognition. We have also established a dataset which contains visual data, auditory data and manipulation instructions for evaluation. Finally, extensive experiments are conducted both offline and online to verify the effectiveness of the proposed audio-visual framework. And it is demonstrated that the robot performs better with the audio-visual data than with only the visual data.**

I. INTRODUCTION

Referring expressions are commonly used when people talking with each other specifying some particular objects in the scene. For example, "the cup next to the computer", "the brown bag on the chair", etc. By understanding the referring expression, the target object can be localized in the scene given natural language description. Different from traditional visual perception tasks which have predefined object labels, the referring expression task is faced with more complex language and visual semantics making it a more challenging task. Currently, the referring expression task has aroused the attention from both the computer vision and natural language communities. And various methods and datasets for referring expression tasks are proposed [1][2][3][4].

However, existing referring expression tasks are mostly subjected to static scenarios and the referred objects have to be within the image. Considering the navigation ability of the robot, it will be of more practical usage if the referring expression task can be implemented while the robot exploring the environment. Recently, Ref. [5] introduces a new task named *Remote Embodied Visual referring Expression in Real Indoor Environments (REVERIE)*, in which the target object might not in sight at first and the agent needs to explore the environment to identify the target object given natural language instruction. It is a more challenging task as it additionally introduces the action into referring expression task and requires the joint learning of vision, language and action together. Thereafter, to achieve the REVERIE task, Ref. [6] exploits the linguistic and visual clues together with commonsense knowledge to generate the exploration action for the agent.

Although the navigation ability of the agent is considered in REVERIE, the manipulation ability, one of the most important charateristics of robot, is ignored. Ref.[7] makes some early attempts on this topic, and Ref.[8] develops temporal grounding graphs for language understanding with manipulators. To solve the referring expression problem, a robotic system INGRESS [9] is proposed, in which the robot is able to pick and place objects following natural language instructions. And a POMDP model is used to eliminate ambiguity and facilitate to ground natural language referring expressions in the robotic manipulation scenario. Furthermore, Ref. [10] introduces a robot system INVIGORATE which enables the robot to grasp a target object from the clutter where occlusions might exist given human instructions. Both the model-based POMDP planning and data-driven deep learning methods are leveraged for the robot to continuously interact with the environment and human. Ref. [11] additionally takes the manipulation history into consideration when visually grounding a series of text instructions for robotic manipulations.

All the above mentioned tasks are only based on visual information. However, in the real-world environment, other sensory modalities can also provide complementary perception information [12][13][14][15], among which sound is widely used in many tasks. In [16][17], the sound of the object is used to distinguish identical bottles which have different contents in them. Besides sound information, Ref. [18] also introduces the tactile information to improve the auditory classification performance. In [19], the sound information can work together with visual information to detect failure in robotic manipulation. Additionally, a type of audio-visual embodied navigation task is proposed recently, in which the agent navigates to a sounding object by leveraging both visual and auditory data [20][21].

In this paper, we propose a new robotic manipulation task for audio-visual embodied referring expression (Fig. 1). Both the audio and visual information are leveraged to interpret the referring expression in robotic manipulation. For example, "find the bottle with the capsule and put it in the left bowl", which requires the robot to firstly localize all the bottles on the table and then identify whether there is capsule in the bottle by listening to the sound when manipulating the bottle. After finding the target bottle, the bottle is put in a target

†denotes the equal contributions. *denotes the corresponding author (hpliu@tsinghua.edu.cn). The authors are with the Beijing National Research Center for Information Science and Technology, Institute for Artificial Intelligence, and the Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China. Yefei Wang is also with School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Kaili Wang is also with School of Physics and Electronic Information, Yantai University. This work was supported in part by the National Key Research and Development Program under Grant 2018YFB1305102 and in part by the Seed Fund of Tsinghua University (Department of Computer Science and Technology)-Siemens Ltd., China Joint Research Center for Industrial Intelligence and Internet of Things.

2022 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9681-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRA46639.2022.9811895

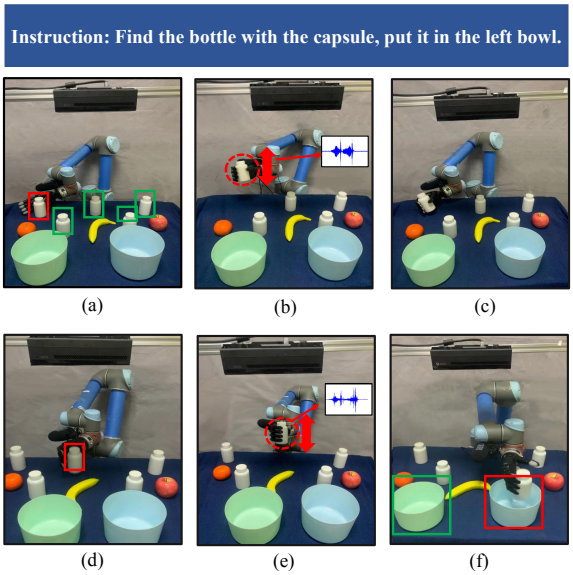

Fig. 1: An illustration of the proposed Audio-Visual Grounding Referring Expression for Robotic Manipulation. In this example, the robot receives a natural language instruction and manipulate with the objects accordingly. (a) Start to pick up the bottle. (b) Perform actions to collect audio information for judgment. (c) If it is not the target object, put the bottle back. (d) Pick up the second bottle in the middle. (e) Perform actions to collect audio information for judgment. (f)If it is the target object, put it in the target bowl.

place. The main contributions of the paper are summarized as the following:

- We develop a novel task of audio-visual grounding referring expression for robotic manipulation, where both audio and visual information are used for interpret the referring expressions.
- We propose an audio-visual framework which can be used for both visual localization and audio recognition for the robot to implement manipulation instructions.
- We collect a multi-modal dataset and conduct extensive experiments to verify the effectiveness of the proposed framework both offline and online.

The remainder of the paper is organized as follows. In Section II, the problem formulation is described. And in Section III, we present an overview of the task architecture. Section IV describes the implementation details of the proposed method. Section V presents the establishment of the dataset. And then, the experimental results are analyzed in Section VI. Finally, we conclude the paper in Section VII.

II. PROBLEM FORMULATION

The goal of the proposed task is to enable the robot to accomplish manipulation tasks following complex natural language instructions, and both the audio and visual information are leveraged to interpret the referring expression.

Concretely speaking, we denote the given natural language instruction as $\mathscr{I} = \{w_1, w_2, ..., w_T\}$, where w_i denotes the i-th word in the instruction. The agent is expected to follow the instruction to manipulate properly. Specifically, both the visual and audio information are required to fully interpret

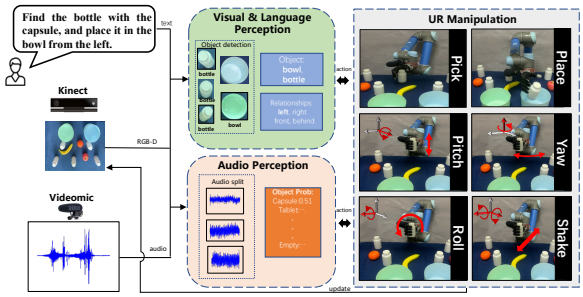

Fig. 2: System Block Diagram. The entire system relies on language, vision, and audio as input, combined with visual-language and audio recognition modules for processing, and uses different manipulation actions to interact with the environment.

the referring expression in the instruction. For example, in the referring expression "a bottle with pills in it", the robot firstly localizes the bottles in the scene by understanding the captured visual information, while it is impossible to know whether there is pills in the bottle. And then the robot could shake the bottle and identify the target bottle by analyzing the sound of shaking. With correct comprehension of the referring expression in the given instruction, the robot is able to implement the manipulations accordingly.

III. ARCHITECTURE

The architecture of the proposed system is demonstrated in Fig. 2, which is composed of a visual-language perception module, audio perception module and manipulation module. At first, the visual information together with the textual instruction is fed into the visual-language module to localize possible targets of the referred objects. As it is difficult to identify the target object with only visual information, the manipulation module will then be activated to implement the different actions to generate sound information. Meanwhile, the sound can be recorded by the equipped auditory sensor. With the collected audio information, the audio recognition module can analyze the audio to recognize the target object that is referred in the instruction. Finally, the robot is able to correctly ground the textual instruction into the manipulation scenario using both the audio and visual information, and the manipulation module will execute proper actions accordingly.

IV. METHOD

As shown in Fig. 3, the proposed framework is composed of a visual-language module, an auditory module and a manipulation module. To understand the referring expression in the given instruction, the visual-language module is firstly implemented to localize possible target objects, and the audio module is used to further recognize the target object. The manipulation module is used to throughout the process to achieve the manipulation instruction.

A. Language model

Inspired by [22], we use a Bi-GRU to encode each word in the natural language instruction $\mathscr{I} = \{w_1, w_2, ..., w_T\}$.

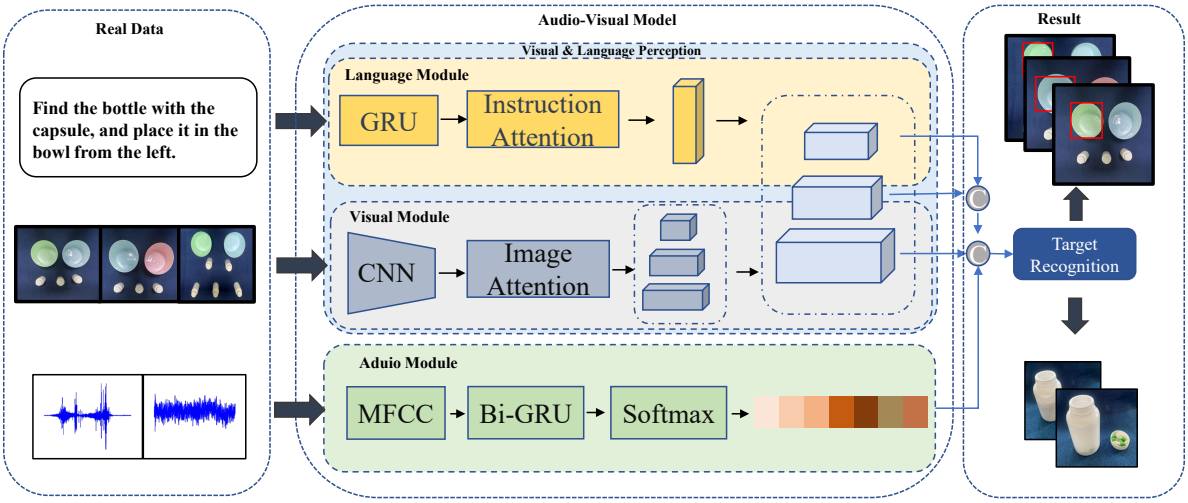

Fig. 3: The structure of the proposed model for audio-visual model. This model consists of three main modules: the language module, the visual module and the auditory module.

Specifically, the one-hot embedding is firstly used to embed each word, and then a Bi-GRU is adopted to further encode the text, where the concatenation of the hidden vectors from both directions represents each word.

$$e_t = \text{One-hot}\ (w_t) \tag{1}$$

$$\overrightarrow{h}_t = \text{GRU}\left(e_t, \overrightarrow{h}_{t-1}\right) \tag{2}$$

$$\overleftarrow{h}_t = \text{GRU}\left(e_t, \overleftarrow{h}_{t+1}\right) \tag{3}$$

$$h_t = \left[\overrightarrow{h}_t, \overleftarrow{h}_t\right] \tag{4}$$

Instead of regarding the instruction as a whole, we try to decompose the textual instruction into subject appearance, location and relationship information for comprehension. Therefore, we define the trainable vectors s_m, where $m \in \{subj, loc, rel\}$, to represent the attention on each word from each perspective and the feature vector for the entire instruction can be obtained as a weighted sum of each word's embedding.

$$a_{m,t} = \frac{\exp\left(s_m h_t\right)}{\sum_{k=1}^{T} \exp\left(s_m h_k\right)} \tag{5}$$

$$q_m = \sum_{t=1}^{T} a_{m,t} e_t \tag{6}$$

And the concatenation of the instruction features from each perspective of $m \in \{subj, loc, rel\}$ forms the textual instruction feature $f_t = [q_{subj}, q_{loc}, q_{rel}]$.

B. Visual Model

In terms of visual features, we use Darknet-53 and feature pyramid network to extract the hierarchical visual features $f_{v_i}, i = 1,2,3$ of the input image. f_{v_i}. The size of the input image is 256×256, and the resolution of the three spatial feature vectors is $8 \times 8 \times 1024, 16 \times 16 \times 512$, and $32 \times 32 \times 256$ respectively.

For each level, the visual-language feature can be fused as:

$$f_{m_i} = \sigma\left(f_t \mathbf{W}_t\right) \odot \sigma\left(f_{v_i} \mathbf{W}_{v_i}\right) \tag{7}$$

where $\mathbf{W}_t$ and $\mathbf{W}_{v_i}$ are projection weight matrices, σ is the Leaky RelU function and $\odot$ denotes the dot-multiplication. For features of different level, we use a 2×2 upsampling operation to map them to the same dimension [23].

With the visual-language fused feature, we would like to localize the target object in the scene for the manipulation. Firstly, we use the Yolo-V3 network to detect possible objects in the frame and generate the visual feature r_{loc}. And then the visual-language feature from different levels are concatenated to the visual-language feature f_m and visual feature r_{loc} are used to calculate the attention weight for each area. The area with the largest score is the most suitable location for the manipulation.

$$t = \varphi\left(\left(W_v f_m + b_v\right) \otimes \left(W_t r_{loc} + b_t\right)\right) \tag{8}$$

$$\beta = \text{softmax}(t) \tag{9}$$

$$\mathrm{u}_{loc} = \beta \otimes f_m \tag{10}$$

$$s_{loc} = \mathscr{D}\left(\mathrm{u}_{loc}, \mathrm{r}_{loc}\right) \tag{11}$$

We use the center point as predicted location for the target object and the image coordinates are converted into the robotic coordinates for the manipulation.

C. Auditory Model

We recognize the type of the target object by analyzing the sound generated when the robot is manipulating the object. The MFCC features [24] are selected to represent the collected sound for it shows robustness to suppress the noise generated during the manipulation. To obtain the MFCC feature, we use the Hamming window with a window size of 30ms and a step size of 15ms. Eventually, 21 Mel coefficients corresponding to the spectrum are obtained. And then the

extracted MFCC features are fed into a Bi-GRU network following a Softmax classifier to recognize the material.

In real experimental environment, there is a lot of noise generated when the robot is running, so a noise suppression algorithm is designed to filter the collected audio data. We set a threshold that is dertermined by some common signal envelope method to remove the signal that is under the threshold.

D. Manipulate Model

With the manipulation model, the robot is able to manipulate with the target object in the scene. The action space is defined as $\mathrm{A} = \{a_{pick}, a_{\mathrm{yaw}}, a_{\mathrm{roll}}, a_{\mathrm{pitch}}, a_{\mathrm{shake}}, a_{\mathrm{place}}\}$, among which $a_{\mathrm{yaw}}, a_{\mathrm{roll}}, a_{\mathrm{pitch}}, a_{\mathrm{shake}}$ are four actions that the robot can use to collect the sound of the object.

When the system detects a target object referred in the instruction, the a_{pick} action is selected. And then the five-finger robotic hand picks up the target object, and implement different actions to generate the sound of the object. After the object is recognized, the a_{place} action is executed to place the object to the expected location.

To collect the sound of the object, the robot performs the four actions $a_{\mathrm{yaw}}, a_{\mathrm{roll}}, a_{\mathrm{pitch}}, a_{\mathrm{shake}}$ in sequence. And the sound clips corresponding to each action are fed into the sound model for recognition. The predicted class with the largest number of occurrence is taken as the class for the object.

V. DATASET

A. Hardware System

The hardware system to collect the dataset is composed of a UR5 robotic arm, sound sensor, Kinect camera, and a five-finger robotic hand, running under the ROS environment with the NVIDIA 2070 GPU. An overview of the system is shown in Fig. 4.

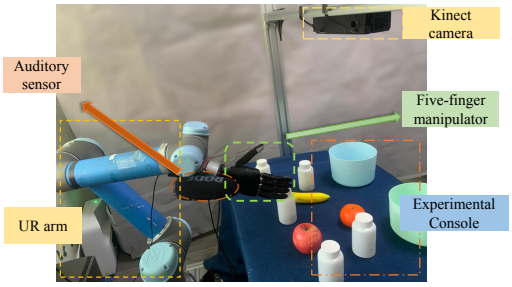

Fig. 4: Hardware architecture. The hardware devices mainly include UR arms, five-finger manipulator, auditory sensor, Kinect camera, and related containers and target objects.

The UR5 robotic arm is a six-degree-of-freedom tandem manipulator which is used to generate different actions to interact with the bottles. A self-developed five-finger robotic hand is attached to the end of the robotic arm which is flexible to execute required manipulation actions. In this work, the robotic hand is used to grasp the bottle in the scenario. The Kinect camera is placed on the top of the table and is responsible for capturing RGB-D images for the scene. And a auditory sensor is equipped near the robotic hand. It is used to receive and record the sound of the objects when they are manipulated by the robot.

B. Auditory Dataset

As is shown in Fig. 5, we consider 12 types of common objects that are usually stored in the bottle. As the bottles are identical in appearance, it is difficult to distinguish different objects in bottles according to the visual information alone. Therefore, we design various actions for the robot to interact with the bottles to generate different sounds for object recognition.

Four actions, namely yaw, roll, pitch, and shake, are designed for the robot to interact with the bottle. Firstly, the bottle is manually handed over to the robotic hand and the bottle is grasped with the thumb and four fingers opposed to each other. And then the bottle is manipulated by the robotic hand. At the same time, the sound of the object will be recorded. It is noted that when implementing the yaw, roll, pitch and shake actions, the angular velocity of the hand is set to be 3.14 rad/s. For 12 different objects, they have been manipulated by the four actions respectively, and each with 20 times. In Fig. 6, we have visualized the sound waves of some typical objects under different actions. It can be seen that different objects have different sound characteristics and the same object could generate different sounds under different actions.

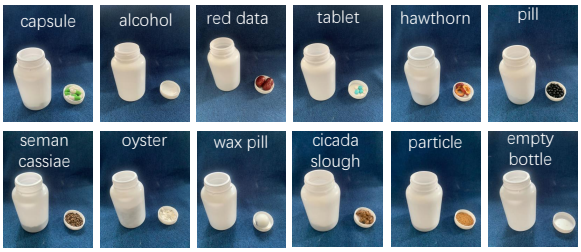

Fig. 5: Object dataset. There are multiple types of object including both liquid and solid to ensure the diversity of sound data.

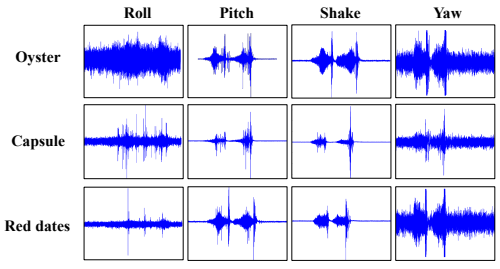

Fig. 6: Sound waves analysis.

C. Manipulation Instruction Design

We consider the scenario where some identical bottles containing different objects are on the table. Three types of manipulation instructions are generated as shown in TABLE I, and are named as existence instruction, classification

TABLE I: Instruction type setting

Instruction type	Instruction template	Example sentence	Dataset size
Existence Instruction	Find the bottle with the <obj>, put it in the <relation>/<color> bowl.	Find the bottle with the hawthorn, put it in the left/green bowl.	288
Classification Instruction	Find all the <obj>, and put it in the <relation>/<color> bowl .	Find all the hawthorns and put them in the left/green bowl.	288
Exploratory Instruction	Check the bottle <relation> the <obj1> for <obj2>.	Check the bottle on the banana for hawthorn.	36

instruction and exploratory instruction respectively. The instructions are corresponded to different scenes. In our scene setting, one scene is associated with at least one type of instruction depending on the complexity of the scene.

A varying degree of interaction between the robot and the environment is required for different manipulation instructions. For the existence instruction, the robot only needs to recognize one target object. And for the classification instruction, the robot needs to explore all bottles until all the target objects are placed in the referred location, which is likely to involve more actions than the existence instruction. For the exploratory instruction, the robot is required to recognize the spatial relationship and manipulate with the referred object.

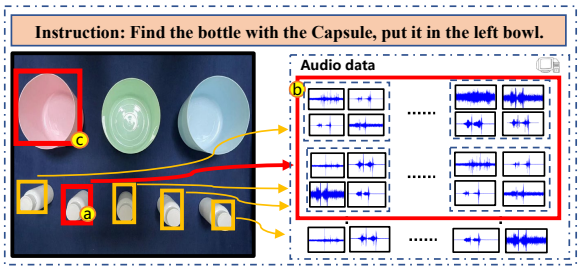

Fig. 7: Audio-visual dataset. For each bottle **a** in the scene, we can pick its corresponding audio data **b** from a pool of its sound data. And the target location **c** is able to be obtained by the given instruction and the visual data of the scene.

VI. EXPERIMENTAL RESULTS

We conduct both the offline and online experiments to verify the performance of the proposed task.

A. Offline experiment

To conduct the offline experiment, we utilize the collected dataset for verification. Fig. 7 gives an overview of the collected dataset, which includes visual information of the scene, collected sound data of each bottle, and given instructions. For each bottle **a** in the scene, we can pick its corresponding audio data **b** from a pool of its sound data. And the target location **c** is able to be obtained by the given instruction and the visual data of the scene. Therefore, we design the following offline experiments with the collected datset.

1) Auditory model recognition experiment: To evaluate the recognition performance of the proposed auditory model, a sound recognition experiment is conducted. In the test phase, the bottles are filled with 12 types of objects with a random capacity between 20% to 80%. And then the robot manipulates with each bottle with the four predefined actions, namely yaw, roll, pitch, and shake. And the sound of the object during the manipulation is recorded. For each object, we collect the sound data for 20 times. The collected data is then fed into the auditory model for recognition.

TABLE II: Accuracy corresponding to different actions

	all actions	Pitch	Yaw	Roll	Shake
Capsule	**71.50%**	63.60%	66.70%	60.90%	59.10%
Alcohol	**62.20%**	56.30%	57.90%	52.10%	49.90%
Red Dates	**75.50%**	62.10%	62.70%	59.30%	58.20%
Tablet	**78.50%**	71.50%	76.90%	72.10%	73.30%
Hawthorn	**82.30%**	73.10%	68.40%	70.30%	64.20%
Pill	**85.40%**	76.10%	81.20%	73.60%	68.50%
Seman Cassiae	**65.70%**	59.30%	61.50%	63.60%	56.10%
Oyster	**72.40%**	61.80%	64.80%	59.70%	56.30%
Wax Pill	**76.30%**	65.80%	63.60%	65.50%	66.10%
Cicada Slough	**67.20%**	65.10%	62.10%	58.70%	63.30%
Particle	**76.40%**	72.10%	74.50%	67.20%	68.40%
Empty	**61.30%**	53.20%	59.70%	56.70%	61.80%
Average	**72.90%**	65.00%	66.70%	63.30%	62.10%

After evaluation, an average accuracy of 72.9% (TABEL II) is obtained for the sound recognition task. It can be seen that the recognition accuracy of different objects varies. The cicada slough and empty bottle have a relatively low accuracy. It is because that the sound generated by them is weak and might be covered by the noise generated when the robot is running. We believe that a stronger denoising method could improve the accuracy for these two situations.

2) Auditory action comparison experiment: In the proposed framework, we use the sound generated by all the four actions to recognize the object. In this experiment, we compare the recognition results that generated by separate actions and the proposed method which utilizes a combination of four actions. For separate actions, only the sound generated by one action is used to recognize the object. The results are demonstrated in TABEL II.

It can be seen that the results obtained from all actions is superior to that from single action. It is because that with the execution of all actions, more audio information can be collect which helps the robot identify the target object more accurately. It is noted that although the accuracy of a single action is not as good as all actions, it can still recognize some visually indistinguishable object illustrating the effectiveness of auditory information.

3) Audio-visual system verification experiment: To evaluate the performance of the entire audio-visual system, we have designed some test scenes as shown in Fig. 9. Also, corresponding manipulation instructions are generated. Given the manipulation instruction, we record the scene information and the sound data generated when the robot is manipulating with the object. And three metrics are designed for a quantitative evaluation. The detailed definition is as follows.

- Target recognition accuracy (TRA): According to the

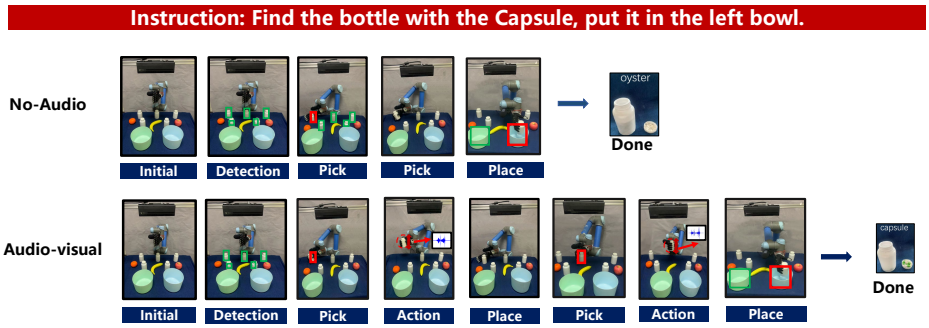

Fig. 8: Online experiment comparison. The no-Audio module is a process of randomly selecting a bottle. Our model is combined with the audio-visual model, and it will not stop picking until the target object is found.

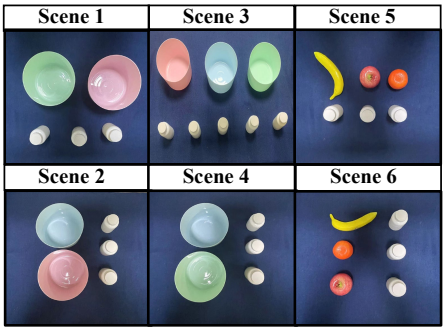

Fig. 9: Example test scene. According to the position and attribute relationship between objects, it is divided into 6 types of typical scenes

visual pictures and natural language instructions in all scenes, the corresponding specific target position is generated through our instruction expression model, and the target recognition accuracy rate is calculated.

- Audio recognition accuracy (ARA): According to the multiple bottles that appear in the corresponding scene, we select the audio data of the bottle at the corresponding position, and calculate the audio recognition accuracy rate under different command tasks through the audio model.
- Overall task success rate (OTSR): This indicator is expressed as the success rate of the entire task when the first two tasks are successful.

Fig. 9 demonstrates six typical test scenes. For each scene and each instruction type, 144 manipulation instructions are given. The results of the three metrics are demonstrated in TABLE III.

TABLE III: Audio-visual system comparison of overall performance

Scene type	manipulation type	TRA	ARA	OTSA
scene1	Existence	87.5%	71.4%	68.3%
scene2	Existence	62.5%	68.7%	54.2%
scene3	Classification	51.7%	42.4%	39.6%
scene4	Classification	70.8%	56.4%	48.5%
scene5	Exploratory	45.8%	64.2%	40.1%
scene6	Exploratory	41.6%	66.5%	35.4%

For existence instruction, it only requires to recognize one target object, and thus its audio recognition accuracy is higher than that of classification instruction which requires to recognize all the target objects. For exploration instruction, it is faced with a more complex scene, and the target recognition accuracy is not as good as that of existence or classification instruction. As only one target object is required, the audio recognition accuracy is still better than the classification instruction.

B. Online experiment

We apply the proposed audio-visual framework on the real world robotic platform. To further illustrate the performance of the proposed framework, we also design a non-auditory baseline. In the non-auditory baseline, a uniform sampling method is used to select the item and complete the task.

Fig. 8 demonstrates the manipulation process when the instruction is given with both the no-audio and the proposed audio-visual approach. We also calculate the OTSR score for each method in Scene 1 and Scene 2. The audio-visual method has the OTSR of 45.4% and 41.2% respectively. And the no-audio method has the OTSR of 24.7% and 22.3% respectively. It can be seen that without the auditory model, the robot randomly selects objects, which significantly reduces the success rate of the entire task. The experimental system with auditory module can guarantee a relatively stable task success rate even in the actual environment.

VII. CONCLUSION

In this paper, we develop a novel task of audio-visual grounding referring expression for robotic manipulation. Both the audio and visual information is used for understanding the referring expression in the manipulation instruction. And an audio-visual framework is proposed. We have also establish a dataset which contains auditory data, visual datat and manipulation instructions. Extensive experiments are conducted both offline and online demonstrating that with the multi-modal data, the robot performs better than with only visual data. For the future work, we would like to extend the proposed task to a more complex scenario, and propose an end-to-end framework for the robot to follow manipulation instructions.

REFERENCES

[1] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 2020.

[2] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[3] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[4] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1124, 2017.

[5] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020.

[6] Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3064–3073, 2021.

[7] Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300, 2016.

[8] Rohan Paul, Andrei Barbu, Sue Felshin, Boris Katz, and Nicholas Roy. Temporal grounding graphs for language understanding with accrued visual-linguistic context. *arXiv preprint arXiv:1811.06966*, 2018.

[9] Mohit Shridhar, Dixant Mittal, and David Hsu. Ingress: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research*, 39(2-3):217–232, 2020.

[10] Hanbo Zhang, Yunfan Lu, Cunjun Yu, David Hsu, Xuguang La, and Nanning Zheng. Invigorate: Interactive visual grounding and grasping in clutter. *arXiv preprint arXiv:2108.11092*, 2021.

[11] Hyemin Ahn, Obin Kwon, Kyoungdo Kim, Dongheui Lee, and Songhwai Oh. Visually grounding instruction for history-dependent manipulation. *arXiv preprint arXiv:2012.08977*, 2020.

[12] Vivian Chu, Reymundo A. Gutierrez, Sonia Chernova, and Andrea L. Thomaz. Real-time multisensory affordance-based control for adaptive object manipulation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7783–7790, 2019.

[13] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021.

[14] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. A multimodal classifier generative adversarial network for carry and place tasks from ambiguous language instructions. *IEEE Robotics and Automation Letters*, 3(4):3113–3120, 2018.

[15] Gyan Tatiya and Jivko Sinapov. Deep multi-sensory object category recognition using interactive behavioral exploration. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7872–7878, 2019.

[16] Erik Strahl, Matthias Kerzel, Manfred Eppe, Sascha Griffiths, and Stefan Wermter. Hear the egg-demonstrating robotic interactive auditory perception. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5041–5041. IEEE, 2018.

[17] Shaowei Jin, Huaping Liu, Bowen Wang, and Fuchun Sun. Open-environment robotic acoustic perception for object recognition. *Frontiers in neurorobotics*, 13:96, 2019.

[18] Yannick Jonetzko, Niklas Fiedler, Manfred Eppe, and Jianwei Zhang. Multimodal object analysis with auditory and tactile sensing using recurrent neural networks. In *International Conference on Cognitive Systems and Signal Processing*, pages 253–265. Springer, 2020.

[19] Arda Inceoglu, Gökhan Ince, Yusuf Yaslan, and Sanem Sariel. Failure detection using proprioceptive, auditory and visual modalities. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2491–2496. IEEE, 2018.

[20] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE, 2020.

[21] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020.

[22] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[23] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10034–10043, 2020.

[24] Feng Wang, Di Guo, Huaping Liu, Junfeng Zhou, and Fuchun Sun. Sound-indicated visual object detection for robotic exploration. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8070–8076, 2019.