Visually Grounding Language Instruction for History-Dependent Manipulation

Hyemin Ahn[1*], Obin Kwon[2*], Kyungdo Kim[3], Jaeyeon Jeong[2], Howoong Jun[4], Hongjung Lee[4], Dongheui Lee[1,5], Songhwai Oh[2]

***Abstract*— This paper emphasizes the importance of a robot's ability to refer to its task history, especially when it executes a series of pick-and-place manipulations by following language instructions given one by one. The advantage of referring to the manipulation history can be categorized into two folds: (1) the language instructions omitting details or using expressions referring to the past can be interpreted, and (2) the visual information of objects occluded by previous manipulations can be inferred. For this, we introduce a history-dependent manipulation task whose objective is to visually ground a series of language instructions for proper pick-and-place manipulations by referring to the past. We also suggest a relevant dataset and model which can be a baseline, and show that our network trained with the proposed dataset can also be applied to the real world based on the CycleGAN. Our dataset and code are publicly available on the project website: https://sites.google.com/view/history-dependent-manipulation.**

I. Introduction

In this paper, our objective is to validate the benefits of the history-aware robots in manipulation, when the robot has to continuously receive language instructions one by one to perform a series of pick-and-place tasks. To solve this problem, we propose a task of *history-dependent manipulation*, whose example is shown in Figure 1. Given a shared workspace with rubber and metal blocks scattered around, the task of our robot is to manipulate the blocks one by one by following human language instructions to build a structure. While the human can observe the workspace from multiple perspectives, the robot can only observe the workspace with a single RGB camera at a fixed position.

In this setting, only relying on the current language and image inputs might be insufficient for robots to understand and execute the given task. First, the problem would occur if we verbally instruct the robot by omitting details or using expressions referring to the past. The language expression in blue fonts in Figure 1 shows this case. After the human instructs the robot to manipulate 'rear left red block', she orders to manipulate '*unmoved* red block'. To interpret this, the robot needs to recall that the previously manipulated object was the 'rear left red block', and understand what '*unmoved* red block' refers to.

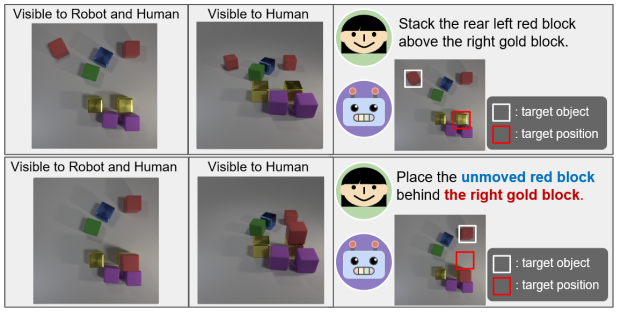

Fig. 1. An example scenario of history-dependent manipulation task, where the robot is required to visually ground the language instruction by referring to its task history. By referring to the past, the robot can understand the human language that assumes the robot's ability to recall the past (blue fonts), and can infer the visually occluded information (red fonts).

Second, the problem would occur if there is an occlusion from robot's viewpoint after stacking or closely arranging objects. Then, relying on the current visual observation would not be enough for robots to perform the task. The language expression with red fonts in Figure 1 shows an example of this case. After the human instructs the robot to stack the rear left red block above the right gold block, the gold block becomes invisible to the robot at the next stage. However, since the right gold block is still visible from human's perspective, she can instruct the robot to place the next target object behind the occluded gold block. Although the right gold block is invisible to the robot, this problem will be solved if the robot can recall the visual information obtained from the previous operation.

To solve these two problems, our robot needs to refer to its task history (1) for understanding the language expression assuming the robot's ability to recall the past, and (2) for overcoming the occlusion in its visual observation. For these challenges, we propose a dataset of various scenarios of the *history-dependent manipulation* task, and a deep neural network-based model to suggest a baseline for related future works. Our dataset provides various tasks of history-dependent manipulations, where each task is composed of a series of pick-and-place operations for building structures with given blocks. For each pick-and-place operation, a

*These authors are equally contributed to this work.

[1]Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Wessling, Germany (e-mail: hyemin.ahn@dlr.de). [2]Department of Electrical and Computer Enginnering and ASRI, Seoul National University, Seoul, Korea (e-mail: {obin.kwon, jaeyeon.jeong}@rllab.snu.ac.kr; songhwai@snu.ac.kr). [3]Transdisciplinary Institute of Medicine & Advanced Technology, Seoul National University Hopital, Seoul, Korea (e-mail: kimkyungdo08@gmail.com). [4]Graduate School of Artificial Intelligence (GSAI) and ASRI, Seoul National University, Seoul, Korea (e-mail: {howoong.jun, hongjung.lee}@rllab.snu.ac.kr).[5]Human-centered Assistive Robotics, Technical University of Munich, Munich, Germany (e-mail: dhlee@tum.de).

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01190, [SW Star Lab] Robot Learning: Efficient, Safe, and Socially-Acceptable Machine Learning), and partially supported by the Helmholtz Association.

set of relevant instructions and RGB images observing the workspace are provided. The proposed model takes an RGB image and language instruction as inputs, and generates two heatmaps that estimate target object positions before and after the manipulation. After each manipulation, a history vector that can abstract the obtained image, language, and time information is stored in the model's external memory. To visually ground the current language instruction, the stored history vectors are employed to perceive the current situation depending on the relevant manipulation history.

II. Related Work

a) History-Dependent Instruction Following: Instruction following for navigation [1]–[25] or manipulation [22]–[24], [26]–[33] requires robots to iteratively interacts with the environment. The robot needs to remember what it has done so far to understand the current state of the environment correctly. However, their instructions are history-independent (e.g., What color is the chair in the bedroom?), explicitly explaining the target without temporal deixis. They focused on mapping instruction to a series of robot actions, without having to recall the past to understand the current instruction.

Most of the history-dependent expressions in the existing datasets are simple sequencing expressions in the navigation task [10]–[21], [23], such as 'Go straight and turn right at the *second* intersection'. Dialogue-guided navigation dataset RobotSlang [25] contains few complex history-dependent expressions (below 0.1%, e.g., 'Go back to the brown wall *you were at before*'). Household tasks instructions dataset ALFRED [22] also contains some history-dependent expressions, referring to the objects previously seen but not present in the robot view. Compared to the previous tasks, our history-dependent manipulation task focuses more on the advanced understanding of history-dependent instructions. We explicitly define the difficulties when the language instructions can only be interpreted by referring to the task history, and suggest a method to resolve such difficulties.

Of course, there exist related works tackling the similar problem of understanding history-dependent instructions. [34] suggested a system to ground a given language instruction depending on the environment as well as the task context. One of the main challenges in [34] is to understand anaphoric references based on the history (i.e., Pick up the *snack*, and microwave *it*). In [35], authors proposed a probabilistic model, which can accumulate the knowledge based on the past visual and language information, and employ that knowledge to ground the current language instruction.

These works focus on understanding the context that happened one or two steps before the current instruction. In our history-dependent manipulation task, the instructions sometimes refer to the past far from the current step. For example, our language input can refer to "the block you moved at the *first time*", when the robot is conducting the *sixth* pick-and-place operation. This makes our task more challenging since robots need to recall task history through multiple interactions. Furthermore, our study suggests that understanding the task history can also help robots to infer the visual information about occluded objects, which is another crucial aspect not covered by previous relevant studies.

b) Visually Grounding Referring Expressions: Since our robot needs to ground the given language instructions in the RGB image, our history-dependent manipulation task has a similarity with several studies in the computer vision research field, which comprehend the referring expressions that describes a target object in the input image [36]–[42]. Related studies can be categorized by whether the proposed model finds the target based on a detection box [36], [37], segmentation map [38]–[40], or both [41], [42].

This task of visually grounding referring expressions also has been related to object manipulation [43], [44], for enabling robots to find the target object to manipulate. [43] proposed a dataset that can be used to train a robot or agent that can pick up the target object referred to by human language instructions. [44] developed a system that can distinguish the target object referred to by human language instruction when multiple objects belonging to the same class are given. These studies are similar to ours in that their models can retrieve the object referred to by a given language expression. However, since they do not assume that a human can instruct a robot over time, their objective does not include enabling robots to understand the *historical* information accrued through the past.

We found similar tasks as ours in natural language processing literature which studied a two-player collaborative scenario to collect the cards [24] or build an architecture [33] in a 3D environment. In [24], [33], the players use multiple history-dependent expressions to complete the given task. However, their focus is on language expression understanding, assuming the agent has full observability. In this paper, we study more practical issue arises from using a real robot with partial observability.

III. Dataset

As shown in Figure 2, our history-dependent manipulation task consists of several pick-and-place operations, which are instructed by a real human language. Each task starts with a shared workspace with several blocks of various colors and materials. Human gives a set of language instructions for the robot to move given blocks one by one for building a specific structure. The main objective of the robot is to estimate the location of the target block before and after executing the given instructions. During the task, it is assumed that the robot observes the given workspace vertically from above using an RGB camera. The human observes the workspace from multiple perspectives, but s/he provides the instruction from the front side to clarify directional words such as left and right. Since the robot's camera would be fixed at a predefined location, there can be blocks that are invisible or partially visible to the robot due to occlusion.

For each pick-and-place operation, our dataset provides synthetic RGB images of the workspace from two viewpoints, a set of language instructions from real humans, bounding boxes, and heatmaps showing the target object's

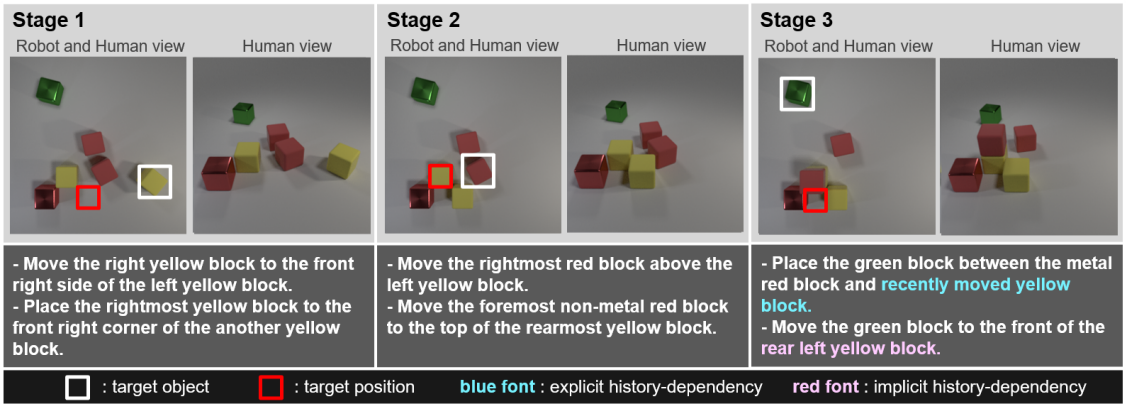

Fig. 2. The example of the proposed dataset for a history-dependent manipulation task consisting of three pick-and-place operations. The first row shows images from our robot and human perspectives, the second row shows the example language instructions. Ground truth positions of the target object before and after the manipulation are annotated with white and red boxes.

position before and after the manipulation. With our simulator based on the code from [45], we collected images reflecting this task scenario. After spawning blocks with different colors and materials, our simulator for image generation randomly chooses which block to move and where to put it. The target position of the block can be among the predefined positions (i.e., center, front left, rear right) inside the workspace, or the position nearby the other blocks.

After collecting synthetic images as shown in Figure 2, we recruited six human participants to annotate each pick-and-place operation with language instructions. The tasks are evenly distributed to the participants so that at least two people can annotate each task. The style of the collected language instructions is different from each human participant, which results in the challenging instruction dataset. In total, the dataset comprises 300 scenarios of history-dependent manipulation tasks (250 for training and 50 for the test), 1339×2 images from the robot and human viewpoints for 1339 pick-and-place operations, and 4642 language instructions.

In our dataset, the collected language instructions can be categorized as follows: (1) one that requires the robot to recall how blocks moved by using explicit expressions such as '*previous*', '*that you moved before*', (2) another that refers to blocks that are invisible or partially visible from the robot's viewpoint, and (3) others that do not belong to the previous two categories. We will call the first type of history-dependency in language instructions as *explicit history-dependency* (blue fonts in Figure 2), and the second type as *implicit history-dependency* (red fonts in Figure 2).

IV. Methodology

A. Overview

In our history-dependent manipulation task, let us assume that a human instructs our robot for n times. Then, let $\mathbf{I}_k \in \mathbb{R}^{w_I \times h_I \times 3}$ denote the k-th input RGB image, and $\mathbf{L}_k = \{w_1, \ldots, w_{n_L}\}$ denote the k-th input language sentence with length n_L, where w_i denotes an one-hot encoding vector representing the i-th word in the sentence. In addition, let $\mathbf{H}_{k-1} = \{h_1 \ldots h_{k-1}\}$ denote a set of history vectors accrued over $k-1$ times of pick-and-place operation. In our model, the visual, linguistic, and time information related to the k-th pick-and-place operations are abstracted into h_k, and stored in $\mathbf{H}_{k-1}$.

Let $\mathcal{F}$ denote our model. It takes the $\mathbf{I}_k$, $\mathbf{L}_k$, and $\mathbf{H}_{k-1}$ as inputs and results in two-dimensional positional heatmaps $\mathbf{M}_k^{pick}, \mathbf{M}_k^{place} \in \mathbb{R}^{w_M \times h_M}$, and one history vector h_k:

$$\mathbf{M}_k^{pick}, \mathbf{M}_k^{place}, h_k = \mathcal{F}(\mathbf{I}_k, \mathbf{L}_k, \mathbf{H}_{k-1})$$

Here, $\mathbf{M}_k^{pick}$ and $\mathbf{M}_k^{place}$ indicate the confidence in the position of the target object to *pick* and where to *place* the target object. Based on the obtained two heatmaps, the robot executes the pick-and-place operation based on the position where the value of each heatmap for pick/place is the highest. After the manipulation, the obtained history vector h_k is stored in $\mathbf{H}_{k-1}$ for the next operation.

B. Network Structure

Figure 3 shows the structure of the proposed model $\mathcal{F}$, and it is influenced by the network structure from [47]. The proposed model consists of (1) an hourglass network [46] to encode the image feature as well as to decode all information into the desired heatmaps (upper blue zone), (2) a bidirectional LSTM to encode the language feature (lower left yellow zone), and (3) a bidirectional LSTM to encode the history feature (lower right green zone).

a) Image Information: When $\mathbf{I}_k$ is given as an input, the *spatial coordinates* $\mathbf{S}$, whose size is same as $\mathbf{I}_k$ and including the xy-positional information of the each spatial location, is concatenated to the $\mathbf{I}_k$ first. The implementation of $\mathbf{S}$ is similar as the approach proposed in [48]. With concatenated $\mathbf{I}_k$ and $\mathbf{S}$ as an input, the hourglass network [46] encodes a set of spatial features from high to low resolution based on its bottom-up process with max-pooling layers and residual modules [49]. Let D denote the depth of the hourglass network, which implies that the bottom-up process is conducted for D times, and Figure 3 shows a model when $D = 3$. Let $\mathbf{F}_I = \{f_I^1 \ldots f_I^D\}$ denote our image information, which is a set of encoded spatial features from the all bottom-up processes. Here, $f_I^d \in \mathbb{R}^{w_d \times h_d \times l_d}$ denotes

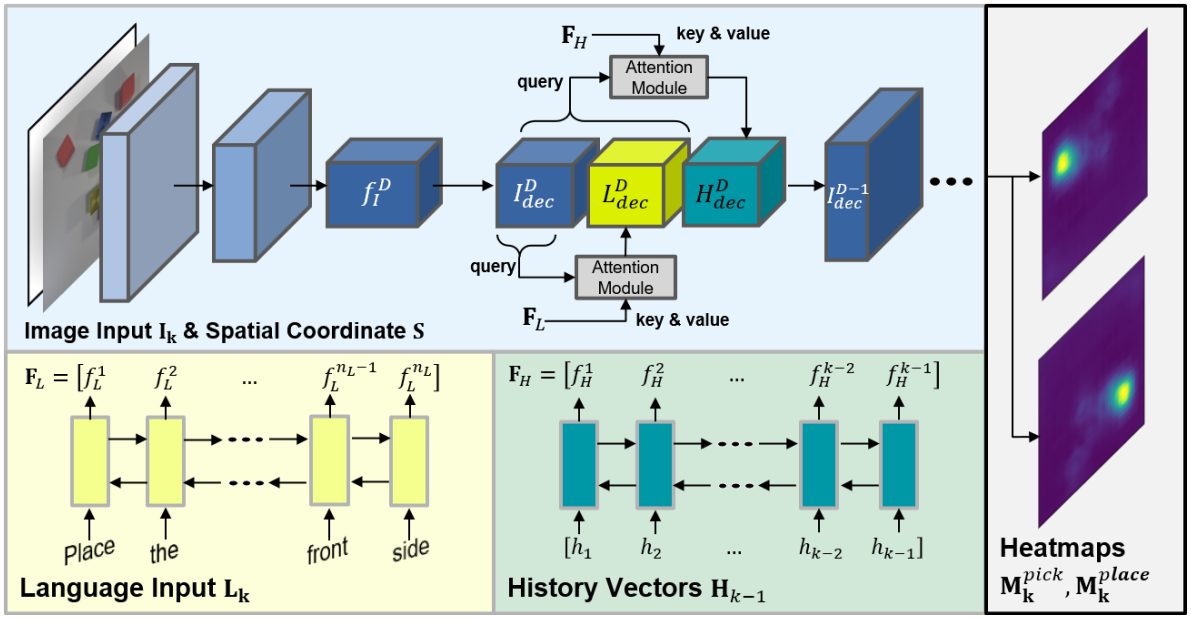

Fig. 3. Structure of the proposed model for the history-dependent manipulation task. The model consists of the hourglass network [46] (upper blue zone), bidirectional LSTM for language information (lower left yellow zone), and another bidirectional LSTM for history information (lower right green zone). In the bottom-up process, a set of spatial features is encoded from the input image and spatial coordinates. In the top-down process, the obtained language and history information are fused into the spatial features to obtain the desired heatmaps.

the spatial feature from the d-th bottom-up process. The top-down process of the hourglass network starts from f_I^D, and it is described in the last paragraph of this section.

b) Language Information: When the language sentence $\mathbf{L}_k = \{w_1 \dots w_{n_L}\}$ is given as an input, all one-hot encoding word vectors w_i are converted into the embedding representation e_i with the trainable embedding matrix E, such that $e_i = Ew_i$. When $\{e_1 \dots e_{n_L}\}$ is given to the bidirectional LSTM, a set of hidden state vectors of the LSTM is obtained and used as our language information $\mathbf{F}_L = \{f_L^1, \dots, f_L^{n_L}\}$. It is used in the top-down process of the hourglass network, which fuses all information to obtain two desired heatmaps.

c) History and Time Information: A set of history vectors $\mathbf{H}_{k-1} = \{h_1, \dots h_{k-1}\}$ is also given to the bidirectional LSTM. Then, the obtained set of hidden state vectors of the LSTM is used as our history information $\mathbf{F}_H = \{f_H^1, \dots, f_H^{k-1}\}$. This history information is fused with image and language information in the top-down process of the hourglass network.

In addition, we also encode the time information into a feature vector f_T, by projecting the one-hot encoded time index vector T into the higher dimensional space. For example, if the human instructs the robot for the second time, the time indexes would be $T = [0, 1, 0, 0, \dots]^T$, and T is projected into the higher dimensional time feature vector f_T by matrix multiplication.

d) Attention Module: Before describing the top-down process of the hourglass network, we would explain the attention module first, which is used in the top-down process to fuse image, language, and history information for obtaining desired heatmaps. Let $\mathbf{X} = \{x_1, \dots, x_N\}$ denote a set of feature vectors, where $x_i \in \mathbb{R}^{D_x}$. Based on the attention mechanism from [50], the attention module determines where to attend on the information in $\mathbf{X}$.

Let $\mathbf{q} \in \mathbb{R}^{D_{qkv}}$ denotes the given query vector, and denote $W_k, W_v \in \mathbb{R}^{D_{qkv} \times D_x}$ as matrices for projecting x_i into key and value vectors as $W_k x_i, W_v x_i \in \mathbb{R}^{D_{qkv}}$. Based on this, a feature vector $\mathbf{f}_X$ representing $\mathbf{X}$ is obtained by a weighted sum of values, where the weight for the i-th value $W_v x_i$ is a softmax score of the i-th key $W_k x_i$, computed based on the dot product between $\mathbf{q}$ as follows:

$$\mathbf{f}_X = \mathcal{A}(\mathbf{q}, \mathbf{X}; W_k, W_v) = \sum_{i=1}^{N} a_i (W_v x_i),$$

$$a_i = \frac{\exp(\text{score}(\mathbf{q}, W_k x_i))}{\sum_{j=1}^{N} \exp(\text{score}(\mathbf{q}, W_k x_j))}, \quad \text{score}(\mathbf{m}, \mathbf{n}) = \mathbf{m}^T \mathbf{n} / \tau$$

Here, τ represents a temperature parameter.

e) Top-down Process for Heatmap Generation: After obtaining the image, language, and history information, our model decodes all into the desired heatmaps based on the top-down process with upsampling and skip connection. The top-down process starts from f_I^D, the last spatial feature from the bottom-up process.

First, we obtain I_{dec}^D by passing f_I^D to a single residual module. Let $I_{dec}^D(i,j)$ denotes a feature vector from I_{dec}^D, at a spatial location (i,j). We fuse the language information $\mathbf{F}_L$ to each spatial location of I_{dec}^D based on the attention module. To do this, $I_{dec}^D(i,j)$ is used as a query and $\mathbf{F}_L$ is used as keys and values with matrices W_k^L and W_v^L, such that:

$$L_{dec}^D(i,j) = \mathcal{A}(I_{dec}^D(i,j), \mathbf{F}_L; W_k^L, W_v^L)$$

TABLE I

PICK-AND-PLACE ACCURACY (%) COMPARISON BETWEEN MODELS WITH AND WITHOUT HISTORY INFORMATION

	History Dependency		
	None	Explicit	Implicit
Proposed with H_{dec}	**73.0**	**62.2**	**68.3**
Proposed w/o H_{dec}	71.3	54.9	66.7

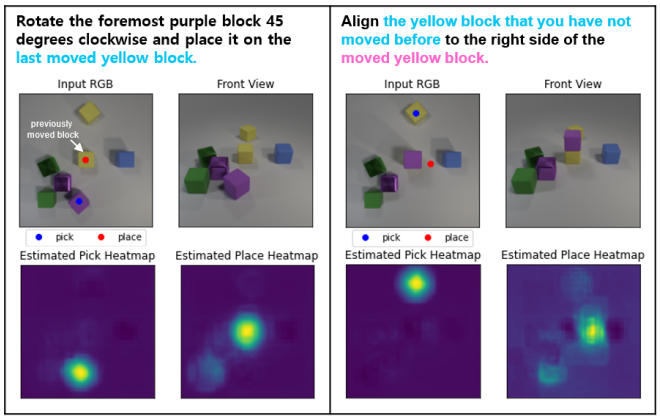

Fig. 4. Generated heatmap results from a history-dependent task in our test dataset. In input RGB images, the blue and red dots denote positions where the the values of the estimated heatmaps are the highest.

Then, every $L^D_{dec}(i,j)$ is concatenated with $I^D_{dec}(i,j)$, and used as a query to fuse the history information $\mathbf{F}_H$ for each spatial location (i,j):

$$H^d_{dec}(i,j) = \mathcal{A}([I^d_{dec}(i,j); L^d_{dec}(i,j)], \mathbf{F}_H; W^H_k, W^H_v)$$

After this process, I^D_{dec}, L^D_{dec}, and H^D_{dec} are concatenated again. The concatenated result passes the residual module as well as the skip connection with f^{D-1}_I, and I^{D-1}_{dec} is obtained. This process is repeated until the I_{dec} has a desired spatial dimension. The final I_{dec} from the top-down process is converted into the desired heatmaps $\mathbf{M}^{pick}_k$ and $\mathbf{M}^{place}_k$, after it passes one residual module and several convolutional layers. The history vector h_k summarizing the image, language, and time information of this pick-and-place operation is obtained as $h_k = [avgpool(I^D_{dec}); avgpool(L^D_{dec}); f_T]$, where $avgpool$ denotes the average pooling layer.

C. Network Training

Our dataset consists of 300 history-dependent manipulation tasks, where 250 tasks are for training, and 50 tasks are for test. Note that a task consists of a series of pick-and-place operations, and our dataset provides one to four language instructions describing each pick-and-place operation. Therefore, when training the model over one task, one sentence is randomly sampled as an input for each pick-and-place. For the data augmentation, we randomly flip images horizontally and change the sentence to correspond to the flipped image. After the model results in a series of heatmaps, a mean square error is calculated based on the ground truth heatmaps and it is minimized by Adam optimizer [51]. Note that the whole model is trained in an end-to-end way.

V. EXPERIMENTS

A. Qualitative Result

Figure 4 shows an example of the qualitative result from the proposed model and test dataset. Here, the blue and red dots in the input RGB images denote the positions where the values of the estimated heatmaps are maximum. In language instructions, the blue fonts show the *explicitly history-dependent* expression which includes language expressions explicitly referring to the past (i.e., '*last*', '*moved before*'), and the red fonts show the *implicitly history-dependent* expression which refers to the occluded object.

The first instruction orders the robot to stack the foremost purple block above the 'last moved yellow block'. Note that the center yellow block in the first input RGB image is the one that moved beforehand. The obtained heatmaps show that the model successfully estimated which yellow block is the previous one. In the second instruction, the human refers to the remained yellow block as "the yellow block that you have not moved before". Also, the center yellow block is referred to as "moved yellow block", but it is not visible to robot due to the occlusion. The generated heatmaps for this pick-and-place operation show that the model successfully estimates the position of the target object and where to place it, even if there are explicit/implicit history-dependent expressions in the same language sentence.

B. Quantitative Result

a) Ablation Study: To validate the effectiveness of exploiting the history information in our task, we compare the performance of the proposed model with ours which does not exploit the history information $\mathbf{F}_H$. When our model does not exploit the history feature $\mathbf{F}_H$, it neglects $\mathbf{F}_H$ in its top-down process, so that only the image and language information is employed to obtain the desired heatmap. To measure the model performance, we calculate the accuracy (%) of pick-and-place operations consisting of the history-dependent manipulation tasks in our test dataset. The accuracy for the single pick-and-place operation is measured by randomly sampling one sentence in test data as an input. The operation is defined as successful when the distance between the predicted pick/place position and ground truth position is less than 15 pixels, which is about the half size of the block when the image size is 256.

Table I shows the comparison results between proposed models with and without the history information. The results are classified based on the type of history-dependency in the input language instructions. When the instruction requires the robot to recall its task history, the performances are higher from the proposed model with history information. When there is an implicit history dependency, it is shown that the performance gap is lower. We expect this is because the model without the history feature could also work when the occluded object is 'partially' visible. But still, we claim that the result implies that the ability to refer to the task history can help the robot to solve the task better when a human explicitly/implicitly requires the robot to employ the task history information.

TABLE II

PICK-AND-PLACE ACCURACY (%) OF OUR MODEL AND MATTNET [41]

	History Dependency		
	None	Explicit	Implicit
Proposed	73.0	**62.2**	**68.3**
MAttNet [41]	**81.1**	61.6	56.5

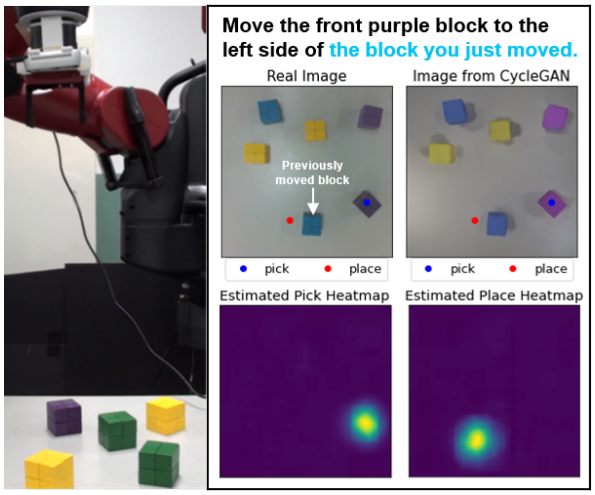

Fig. 5. An example of applying the trained model to the real world. Input language instruction is grounded to the image transferred by CycleGAN [52]. In images, blue and red dots denote the positions where the values of the estimated heatmaps are the highest. More real world demonstrations are included in our supplementary video.

b) Comparison with MAttNet [41]: In addition to the ablation study, we compared the performance of our proposed model with MAttNet [41], which is one of the representative studies of visually grounding referring expressions. We trained two separate MAttNets for pick/place operations. Note that the MAttNet does not have the ability to refer to its task history, but we add this to our comparison study to check how much our model is adequate to be a baseline.

Table II shows the pick-and-place accuracy (%) of the proposed model and MAttNet, where the accuracy is measured with the same method used in the ablation study. MAttNet outperforms when there is no history-dependent expression in the input language, since it is specialized in the task of visually grounding referring expression. It first detects the candidate region proposals from the image and classifies which region is referred to by the given sentence. This makes easier for MAttNet to solve the problem than heatmap-based approach like us, since the candidates are explicitly defined in advance. But note that our MAttNet implementation based on the official code has 52.7M network parameters, and takes 64.4 hours for training, with Nvidia Titan X GPU, 128GB RAM, Intel i7 CPU. Compared to this, ours has 39.6M network parameters, and takes 2 hours for training, with Nvidia RTX 2060 SUPER GPU, 32GB RAM, Intel i7 CPU.

Our model performs better when there is explicit/implicit history dependency. Especially, it is shown that the performance gap between the proposed model and MAttNet is the highest when there is an implicit history-dependency. This shows that exploiting the history information makes our model more effective in solving the history-dependent manipulation task. Based on this, we claim that the proposed model can be a comparable baseline for future studies relevant to the history-dependent manipulation task.

C. Real World Demonstration

We also show that our model can be also applied to the real world. As shown in Figure 5, we attach a RealSense D435 to the Baxter robot's right hand and place its arm at a fixed position to observe the workspace. To bridge the gap between RGB images from the real world and the dataset, we employ CycleGAN [52], which transfers images from different domains. To train CycleGAN, we collect 354 RGB images of real-world workspace and employ 660 images from our training dataset. As shown in Figure 5, the trained CycleGAN transforms the real-world RGB images to look like the ones from our dataset, so that our model can execute the task based on the transferred image.

With the transformed RGB image and the given language instruction, our model infers two heatmaps of pick-and-place. Then, we choose two pixel positions with the highest pick/place heatmap values. By using the depth image of the calibrated camera, a robot roughly estimates the real-world xyz-positions of the chosen pixel positions, and uses them for manipulation. We use the built-in inverse-kinematics solver of the Baxter robot to generate its proper arm trajectory. Here, we also leverage the inner-wrist RGB camera in the Baxter robot. This camera is used to adjust the gripper position, after reaching the roughly estimated xyz-position from the RealSense camera. After reaching the estimated xyz-position, we calculate the contours of the target block from the RGB image and re-estimate the precise pose of the block. Please refer to our supplementary video for more real world demonstration cases.

VI. CONCLUSION

In this paper, we define a task of *history-dependent manipulation*, which aims to enable a robot to refer to its task history when executing a series of pick-and-place operations instructed by language. In this task, two main challenges can arise: (1) the language instructions referring to the past can be given, and (2) the objects occluded after pick-and-place operations need to be inferred. To solve this, we propose a dataset and neural network-based model, and apply the model to the real world to suggest the scalability of our work. However, since we simplified several factors to focus on the scenario when understanding the task history becomes most crucial factor to perform a task, there exist limitations such as (1) images are synthetic, and (2) the robot behavior is limited to pick-and-place. But still, our results show that the proposed end-to-end-trainable model can successfully learn how to understand the history by abstracting the visual, language, and time information from the previous pick-and-place operations. Our work will provide a reasonable baseline for studies of history-aware manipulation robots, and improve the chance of realizing long-term human-robot collaboration in future.

REFERENCES

[1] H. Yu, X. Lian, H. Zhang, and W. Xu, "Guided Feature Transformation (GFT): A Neural Language Grounding Module for Embodied Agents," in *Proc. of International Conf. on Robotics and Automation (ICRA)*, 2019.

[2] H. Yu, H. Zhang, and W. Xu, "Interactive Grounded Language Acquisition and Generalization in a 2D World," in *Proc. of International Conf. on Learning Representations (ICLR)*, 2018.

[3] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov, "Gated-Attention Architectures for Task-Oriented Language Grounding," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[4] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied Question Answering," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. van den Hengel, "REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] V. Blukis, Y. Terme, E. Niklasson, R. A. Knepper, and Y. Artzi, "Learning to Map Natural Language Instructions to Physical Quadcopter Control using Simulated Flight," in *Proc. of the Conf. on Robot Learning (CoRL)*, 2020.

[8] V. Blukis, D. Misra, R. A. Knepper, and Y. Artzi, "Mapping Navigation Instructions to Continuous Control Actions with Position-Visitation Prediction," in *Proc. of the Conf. on Robot Learning (CoRL)*.

[9] V. Blukis, N. Brukhim, A. Bennett, R. A. Knepper, and Y. Artzi, "Following High-level Navigation Instructions on a Simulated Quadcopter with Imitation Learning," in *Robotics: Science and Systems (RSS)*, 2018.

[10] X. Zang, A. Pokle, M. Vázquez, K. Chen, J. C. Niebles, A. Soto, and S. Savarese, "Translating Navigation Instructions in Natural Language to a High-Level Plan for Behavioral Robot Navigation," in *Proc. of the Conf. on Empirical Methods for National Language Processing (EMNLP)*, 2018.

[11] V. Jain, G. Magalhaes, A. Ku, A. Vaswani, E. Ie, and J. Baldridge, "Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation," in *Proc. of the Association for Computational Linguistics (ACL)*, 2019.

[12] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge, "Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding," in *Proc. of the Conf. on Empirical Methods for National Language Processing (EMNLP)*, 2020.

[13] W. Zhu, H. Hu, J. Chen, Z. Deng, V. Jain, E. Ie, and F. Sha, "BabyWalk: Going Farther in Vision-and-Language Navigation by Taking Baby Steps," in *Proc. of the Association for Computational Linguistics (ACL)*, 2020.

[14] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi, "Touchdown: Natural language navigation and spatial reasoning in visual street environments," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] T. Paz-Argaman and R. Tsarfaty, "RUN through the Streets: A New Dataset and Baseline Models for Realistic Urban Navigation," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing and the International Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[16] K. M. Hermann, M. Malinowski, P. Mirowski, A. Banki-Horvath, K. Anderson, and R. Hadsell, "Learning to follow directions in street view," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

[17] H. Kim, A. Zala, G. Burri, H. Tan, and M. Bansal, "ArraMon: A Joint Navigation-Assembly Instruction Interpretation Task in Dynamic Environments," in *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

[18] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-Dialog Navigation," in *Proc. of the Conf. on Robot Learning (CoRL)*, 2019.

[19] Y. Zhu, F. Zhu, Z. Zhan, B. Lin, J. Jiao, X. Chang, and X. Liang, "Vision-Dialog Navigation by Exploring Cross-modal Memory," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] K. Nguyen and H. Daumé III, "Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning," in *Proc. of the Conf. on Empirical Methods for National Language Processing (EMNLP)*, 2019.

[21] T.-C. Chi, M. Shen, M. Eric, S. Kim, and D. Hakkani-tur, "Just Ask:An Interactive Learning Framework for Vision and Language Navigation," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

[22] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[23] D. K. Misra, A. Bennett, V. Blukis, E. Niklasson, M. Shatkhin, and Y. Artzi, "Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction," in *Proc. of the Conf. on Empirical Methods for National Language Processing (EMNLP)*, 2018.

[24] A. Suhr, C. Yan, J. Schluger, S. Yu, H. Khader, M. Mouallem, I. Zhang, and Y. Artzi, "Executing Instructions in Situated Collaborative Interactions," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing and the International Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[25] S. Banerjee, J. Thomason, and J. J. Corso, "The RobotSlang Benchmark: Dialog-guided Robot Localization and Navigation," in *Proc. of the Conf. on Robot Learning (CoRL)*, 2020.

[26] C. Paxton, Y. Bisk, J. Thomason, A. Byravan, and D. Foxl, "Prospection: Interpretable Plans from Language By Predicting the Future," in *Proc. of International Conf. on Robotics and Automation (ICRA)*, 2019.

[27] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. Ben Amor, "Language-Conditioned Imitation Learning for Robot Manipulation Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[28] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, "Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions," in *Proc. of International Conf. on Robotics and Automation (ICRA)*, 2018.

[29] Y. Bisk, K. J. Shih, Y. Choi, and D. Marcu, "Learning Interpretable Spatial Operations in a Rich 3D Blocks World." in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[30] M. Shridhar and D. Hsu, "Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction," in *Robotics: Science and Systems (RSS)*, 2018.

[31] C. Lynch and P. Sermanet, "Language Conditioned Imitation Learning over Unstructured Data," *Robotics: Science and Systems (RSS)*, 2021.

[32] Y. Bisk, D. Yuret, and D. Marcu, "Natural Language Communication with Robots," in *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016.

[33] P. Jayannavar, A. Narayan-Chen, and J. Hockenmaier, "Learning to execute instructions in a Minecraft dialogue," in *Proc. of the Association for Computational Linguistics (ACL)*, 2020.

[34] D. K. Misra, J. Sung, K. Lee, and A. Saxena, "Tell me dave: Context-sensitive grounding of natural language to manipulation instructions," *The International Journal of Robotics Research (IJRR)*, vol. 35, no. 1-3, pp. 281–300, 2016.

[35] R. Paul, A. Barbu, S. Felshin, B. Katz, and N. Roy, "Temporal grounding graphs for language understanding with accrued visual-linguistic context," in *Proc. of the International Joint Conf. on Artificial Intelligence*, 2017, pp. 4506–4514.

[36] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural language object retrieval," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4555–4564.

[37] L. Yu, H. Tan, M. Bansal, and T. L. Berg, "A joint speaker-listener-reinforcer model for referring expressions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[38] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *European Conf. on Computer Vision (ECCV)*. Springer, 2016, pp. 108–124.

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

[40] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 502–10 511.

[41] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1307–1315.

[42] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, and R. Ji, "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 034–10 043.

[43] R. Scalise, S. Li, H. Admoni, S. Rosenthal, and S. S. Srinivasa, "Natural language instructions for human–robot collaborative manipulation," *The International Journal of Robotics Research (IJRR)*, vol. 37, no. 6, pp. 558–565, 2018.

[44] V. Cohen, B. Burchfiel, T. Nguyen, N. Gopalan, S. Tellex, and G. Konidaris, "Grounding language attributes to objects using bayesian eigenobjects," in *Proc. of International Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1187–1194.

[45] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2901–2910.

[46] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conf. on Computer Vision (ECCV)*. Springer, 2016, pp. 483–499.

[47] H. Ahn, S. Choi, N. Kim, G. Cha, and S. Oh, "Interactive text2pickup networks for natural language-based human–robot collaboration," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3308–3315, 2018.

[48] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *European Conf. on Computer Vision (ECCV)*. Springer, 2016, pp. 108–124.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems (NeurIPS)*, 2017, pp. 5998–6008.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conf. on Computer Vision*, 2017.