Ges3ViG: Incorporating Pointing Gestures into Language-Based 3D Visual Grounding for Embodied Reference Understanding

Supplementary Material

In this supplementary material, we provide additional details about several aspects discussed in the paper. To begin, we would like to highlight that our dataset and code is made publicly available. Readers can access the code for both Ges3ViG and the Imputer framework, as well as information on downloading the complete ImputeRefer dataset, via https://github.com/AtharvMane/Ges3ViG.

The following sections elaborate on various components of the ImputeRefer dataset.

A. Language Description Generation

In our Imputer framework, we introduced a Language Description Generation step designed to augment the original language descriptions in ScanRefer, detailed in Section 3.2. This step augments the verbal descriptions by integrating information about the existence of additional pointing gestures. Specifically, we used Gemini [15] to augment the existing verbal descriptions by incorporating this context of the presence of additional pointing gestures. To achieve this, we utilized a carefully designed textual prompt, which we describe in detail below:

1. **Context Paragraph:** Provide a detailed explanation of the task that serves as the context for the LLM.
2. **Output Specifications:** Specify clear instructions for the LLM on how to structure its response, including the required format.
3. **Query:** The original query sourced from ScanRefer.

A.1. Context Paragraph

In the context paragraph, we provided a detailed explanation of the task at hand to the LLM. Specifically, we provide context for Gemini to augment the existing ScanRefer description by considering the presence of additional pointing gestures. The following phrase was used as the context paragraph.

There is a scene where a human is pointing at a target object. There is an external description of the target object which was given to point at the target object in case the human was not present. Assume that you are that human. What would you say to point at the object while doing the pointing gesture?

Some examples of queries provided to Gemini are presented in Section A.5.

A.2. Output Specifications

We then provide the LLM with specific instructions on how to structure its response using the following prompt.

Give special attention to the following while answering:

1. *There may be multiple objects of the same class as the target object.*
2. *The human is pointing at the target object.*
3. *Do not add information that can not be directly inferred from the query.*
4. *Give any 3 possible distinct expressions and no other text.*
5. *Do not use any special characters or punctuation marks other than period and comma.*
6. *The output format is as follows:*
   a *Output 1*
   b *Output 2*
   c *Output 3*

A.3. Query

To guide Gemini in generating a language description that incorporates the presence of pointing gestures, we referenced the existing queries in the original ScanRefer dataset [5]. The approach involves taking these original queries and augmenting them to include considerations for an additional pointing gesture in the generated language descriptions.

Here are some examples of queries from the existing ScanRefer dataset:

- A black TV, in the direction from the entrance and from the outside, will be on the right side of the blue curtain . on the left of the tv is a small bike.
- There is a beige wooden working table. placed on the side of the room

A.4. Complete Textual Prompt

The following shows the complete textual prompt format that was used as an input to Gemini.

There is a scene where a human is pointing at a target object. There is an external description of the target object which was given to point at the target object in case the human was not present. Assume that you are that human. What would you say to point at the object while doing the pointing gesture?

Give special attention to the following while answering:

1. *There may be multiple objects of the same class as the target object.*
2. *The human is pointing at the target object.*

3. *Do not add information that can not be directly inferred from the query.*
4. *Give any 3 possible distinct expressions and no other text.*
5. *Do not use any special characters or punctuation marks other than period and comma.*
6. *The output format is as follows:*
   a *Output 1*
   b *Output 2*
   c *Output 3*

   CURRENT QUERY:<ScanRefer Query>

A.5. Some Example Prompts

A.5.1. Example 1

Prompt to the Gemini:

There is a scene where a human is pointing at a target object. There is an external description of the target object which was given to point at the target object in case the human was not present. Assume that you are that human. What would you say to point at the object while doing the pointing gesture?

Give special attention to the following while answering:

1. There may be multiple objects of the same class as the target object.
2. The human is pointing at the target object.
3. Do not add information that can not be directly inferred from the query.
4. Give any 3 possible distinct expressions and no other text.
5. Do not use any special characters or punctuation marks other than period and comma.
6. The output format is as follows:
   a Output 1
   b Output 2
   c Output 3

   CURRENT QUERY: A black TV, in the direction from the entrance and from the outside, will be on the right side of the blue curtain . on the left of the tv is a small bike.

Response from Gemini:

1. That black TV there
2. The television set to your right
3. The black TV over there

A.5.2. Example 2

Prompt to the Gemini:

There is a scene where a human is pointing at a target object. There is an external description of the target object which was given to point at the target object in case the human was not present. Assume that you are that human. What would you say to point at the object while doing the pointing gesture?

Give special attention to the following while answering:

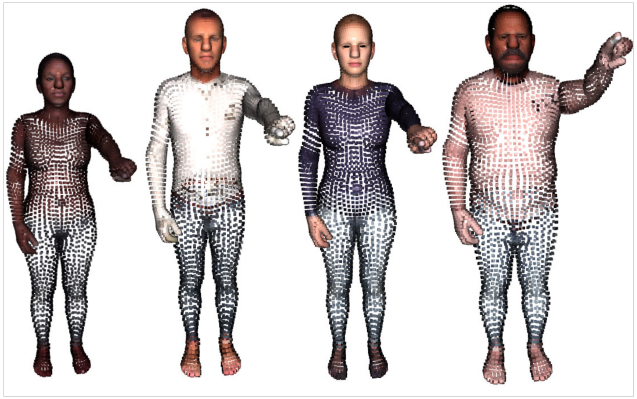


Figure 1. Human avatars used in ImputeRefer dataset.

1. There may be multiple objects of the same class as the target object.
2. The human is pointing at the target object.
3. Do not add information that can not be directly inferred from the query.
4. Give any 3 possible distinct expressions and no other text.
5. Do not use any special characters or punctuation marks other than period and comma.
6. The output format is as follows:
   a Output 1
   b Output 2
   c Output 3

   CURRENT QUERY: There is a beige wooden working table. placed on the side of the room.

Response from Gemini:

1. The wooden table over there.
2. The beige table in the corner.
3. That wooden table by the wall.

A.6. Observations from the (Re)Generated Referring Expressions:

We observed that the original descriptions in ScanRefer, which were generally longer in word count, were condensed in the ImputeRefer dataset. Additionally, we noticed an increased use of prepositions such as “here,” “that,” and “this,” indicating that, alongside the verbal descriptions, there is also a reliance on pointing gestures to identify the target object.

B. Example Generations from Imputer

As described in Section 3.1, and presented in Figure 2, to generate the ImputeRefer dataset, we used multiple human avatars. We present examples of the human avatars used in Figure 1. In the complete ImputeRefer dataset, we ensure

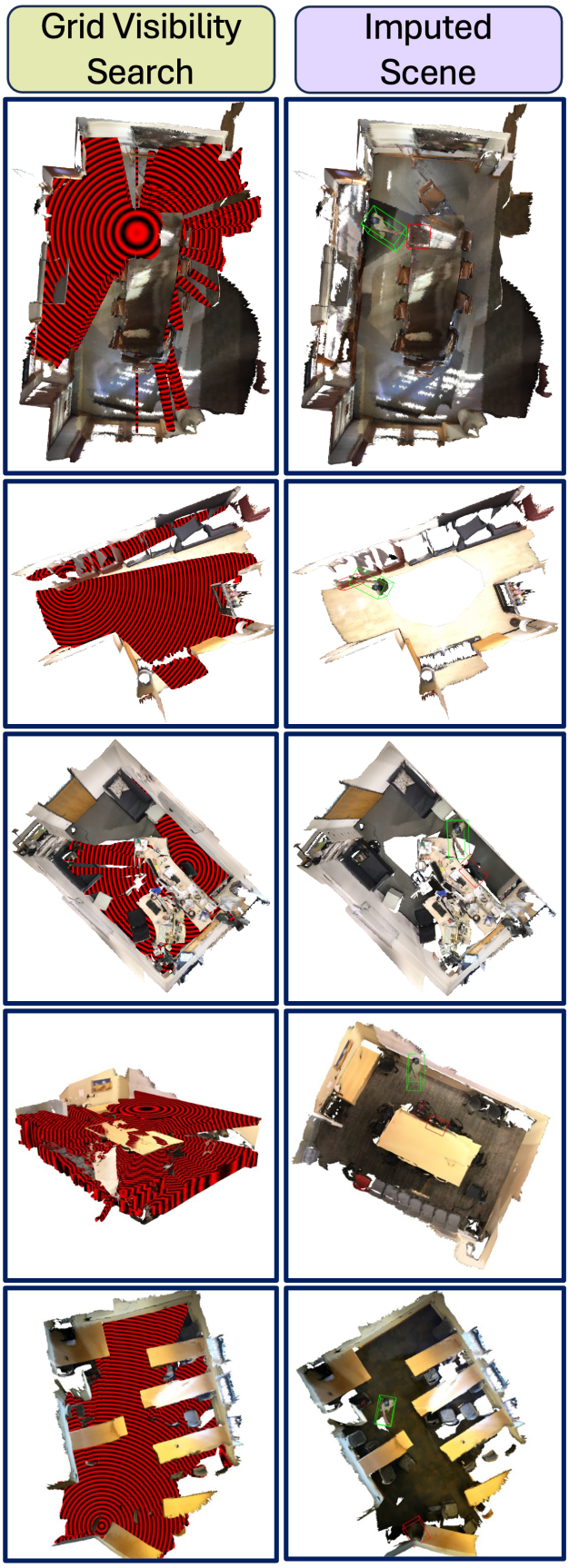


Figure 2. Examples of Imputer Framework

Table 1. Ablation studies for Ges3ViG in ImputeRefer dataset

Model	unique		multiple		overall	
	IoU @0.25	IoU @0.5	IoU @0.25	IoU @0.5	IoU @0.25	IoU @0.5
Ges3ViG $_{\text{w/o Gestures}}$	69.26	48.84	52.58	37.32	55.79	39.54
Ges3ViG $_{\text{noHumanLoss}}$	68.76	48.57	58.75	42.31	60.68	43.51
Ges3ViG $_{\text{noEF\_onlyLF}}$	69.43	49.28	54.81	39.02	57.62	41.00
Ges3ViG $_{\text{onlyEF\_noLF}}$	83.71	70.09	66.47	54.92	69.93	58.05
Ges3ViG $_{\text{random\_LF}}$	84.0	70.6	66.1	54.6	69.6	57.7
Ges3ViG $_{\text{onlyGest}}$	15.29	11.81	12.46	9.80	13.0	10.18
Ges3ViG $_{\text{ConstantLang}}$	51.05	43.87	44.89	36.78	46.08	38.15
Ges3ViG $_{\text{GT Poses}}$	84.73	71.75	67.99	55.89	70.97	58.99
Ges3ViG	**84.60**	**71.03**	**67.57**	**55.77**	**70.85**	**58.71**

that it is balanced in terms of gender, height, and body mass.

B.1. Deployment steps for humans in the scene

Grid visibility search is one of the key intermediate steps of our Imputer framework. We showcase additional examples demonstrating the identification of visible regions and the imputation of a human avatar into the existing point-cloud scene in Figure 2. In this figure, the images on the left depict the grid-based visibility search used to identify regions with a clear line of sight. The images on the right show the imputed scene after adding the human avatar. The green bounding box highlights the human avatar, while the red bounding box indicates the target object. Visibility values were calculated using the grid-path counting method proposed by Goldstein et al. [14]. Regions with visibility values greater than 0.33 and located within the scene were identified as suitable for human imputation, as detailed in the paper.

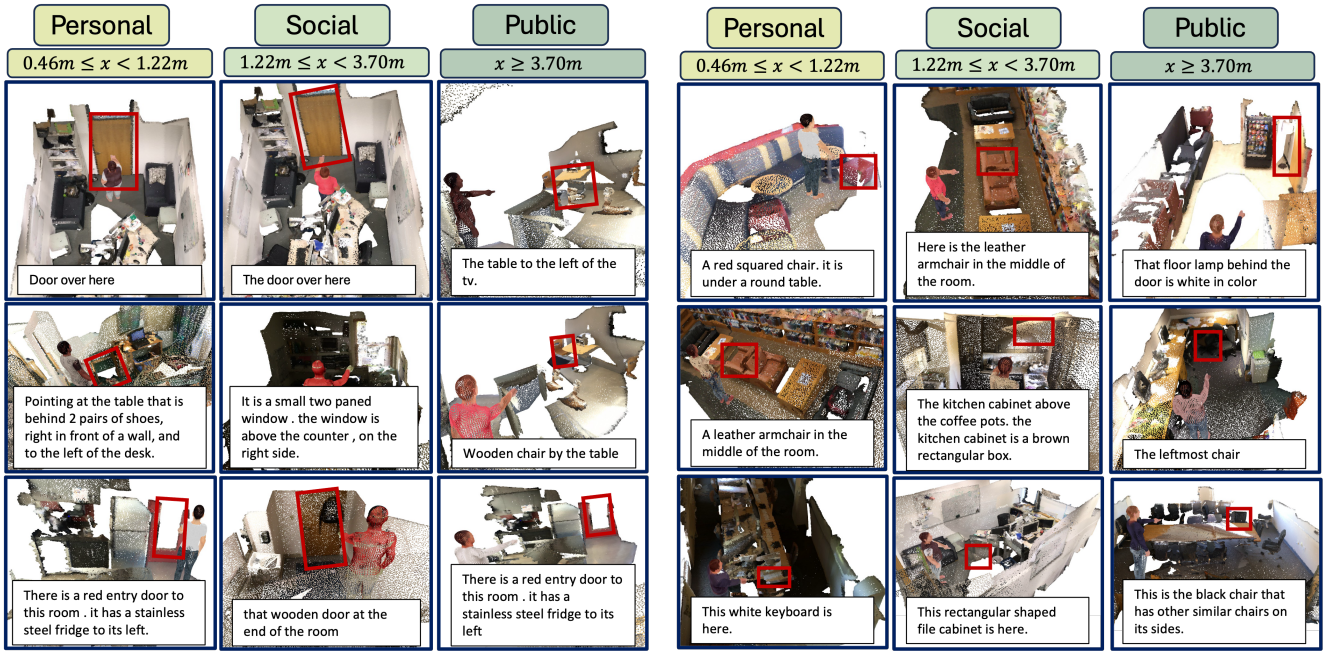
B.2. Examples of deployment in Personal, Social, and Public settings for Unique and Multiple objects

In the paper, we presented one example from the ImputeRefer for each combination. Here, we provide additional examples of scenes along with the corresponding queries associated with them. Figure 3a illustrates scenes with 'unique' objects, where only one object from the same object class as the target object is present. We include three additional samples that vary based on the distances between the human and the target object: 'Personal' (0.46m to 1.22m), 'Social' (1.22m to 3.70m), and 'Public' (greater than 3.70m).

Similarly, Figure 3b illustrates scenes with 'multiple' objects, where there may be more than one object (distractors) belonging to the same class as the target object. As with the 'unique' samples, we provide three additional examples for each distance range.

C. Additional Ablation Studies

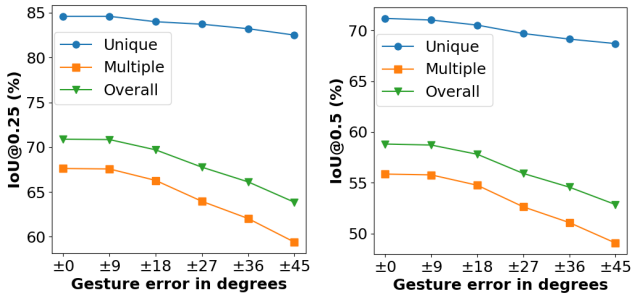
We show additional ablation studies in Table 1. In particular, Ges3ViG $_{\text{GT Poses}}$ evaluates the performance of Ges3ViG when ground truth poses were used for the late fusion instead of the predicted poses from the proposed model.



(a) Samples from ImputeRefer with 'unique' object scenes

(b) Samples from ImputeRefer with 'multiple' object scenes

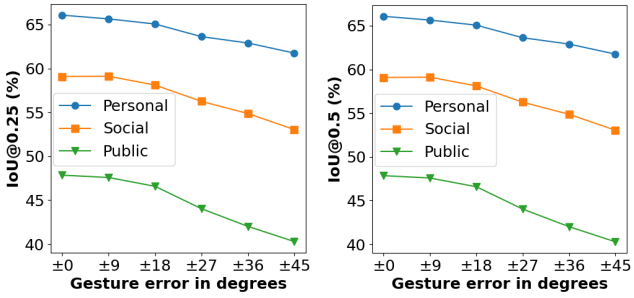
Figure 3. Samples from ImputeRefer dataset. The imputed human is placed at different distances from the target



(a) IoU@0.25 (%)

(b) IoU@0.5 (%)

Figure 4. Accuracy for multiple vs unique samples for 3D-ERU with imprecise pointing of the human avatar



(a) IoU@0.25 (%)

(b) IoU@0.5 (%)

Figure 5. Accuracy at different distance ranges for 3D-ERU with imprecise pointing of the human avatar

Even when the ground truth pose data is used, Ges3ViG $_{\text{GT Poses}}$ saw only a minor boost in accuracy, implying that Ges3ViG's approach of jointly learning human pose is effective.

As mentioned in Section 3, we introduced a uniformly sampled rotational jitter $(0 - 9°)$ in the pointing direction to account for natural human error and variations. In Figure 4 and 5, we extend this analysis by varying the rotational noise and plotting the resulting accuracy. In Figure 4, we find that there is only a minor drop in accuracy for the unique object scenarios, even under very high rotational noise. However, for multiple object scenarios, we observe a significant drop in accuracy when rotational noise exceeds $18°$. Similarly, from Figure 5, we see that the drop in accuracy due to imprecise gestures is significant when the human-to-object distance is higher (public distance). Thus, we find that the effect of imprecise gestures is more pronounced at larger distances and in the presence of multiple similar objects. In general, these figures show that Ges3ViG is able to tolerate rotational noise of up to $18°$, reasonably well. A rotational error beyond this range can be deemed infrequent in practical situations.