

In Situ 3D Scene Synthesis for Ubiquitous Embodied Interfaces

Appendix

ANONYMOUS AUTHOR(S)

1 METHOD

1.1 User intention inference

1.1.1 Text extractor $E_{user}(e_{user}, t_{user}, O_{user}|u)$. Language is the most commonly used medium of interaction. Users can express where they want to go through language. Generally speaking, user language is unstructured, that is, users are free to express their thoughts spontaneously, without adhering to a specific word order. LLMs have demonstrated remarkable capabilities in text understanding and generation, which allows them to be applied effectively in information extraction tasks [Kalyan 2023]. Therefore, a few-shot LLM is used to extract the place, season, and objects mentioned by the user. One example of the prompt is:

Prompt of text extractor

You are now an excellent model for sentence recognition.
The text I'm giving you is a sentence and you need to predict if place, season, and objects are mention in the sentence.
Text: I want stay in home with tables, sofas, chairs.

One example of the output is:

Output

Answer: "place": "home", "season": "null", "objects": ["tables", "sofas", "chairs"]

1.2 Physics-based inference

1.2.1 Affordance predictor $A(V_{\mathcal{A}}|\mathcal{A}, O)$. A few-shot LLM is used to infer the likelihood of each affordance that all objects have. **AFFORDANCES** is all affordances in the list. One example of the prompt is:

Prompt for affordance prediction

You are now an excellent model for object affordance recognition.
*Affordances list: **AFFORDANCES**.*
The text I'm giving you is the name and size of the object.
The size of object includes: length, width and height.
Please give the probabilities of each affordance in Affordances list for the object I give you.
Text: table, length: 1 meter, width: 0.6 meter, height: 0.8 meter.

One example of the output is:

Output

Answer: '{walkable}':0.00, '{supportable}':1.00, '{sitable}':0.60, '{drinkable}':0.00, '{eatable[4]}':0.00, '{graspable[5]}':0.10, '{breakable}':0.00, '{dangerous}':0.00, '{moveable}':0.20, '{obstructive}':1.00.

1.3 Virtual object-based inference

1.3.1 Place predictor $E(V_{\mathcal{E}}|O)$. A few-shot LLM is used to infer the likelihood of all objects appearing in each place. **PLACES** is all places in the list. One example of the prompt is:

Prompt for place prediction

You are now an excellent model for object place recognition.
*Places list: **PLACES**.*
The text I'm giving you is the name and the description of the object. Note that sometimes there is no description of the object.
Please give the probabilities of the object appear in each place in places list for the object I give you.
Text: table. Introduction: it is often seen at camping sites and other outdoor facilities;

One example of the output is:

Output

Answer: '{Library}':0.25, '{Conservatory}':0.25, '{Spa}':0.25, '{Lounge}':0.25, '{Observatory}':0.25, '{Suite}':0.25, '{Monastery}':0.25, '{Studio}':0.25, '{Bookstore}':0.25, '{Aquarium}':0.02, '{Beach}':1.00, '{Forest}':0.80, '{Garden}':1.00, '{Vineyard}':1.00, '{Yacht}':0.10, '{Rooftop}':0.25, '{Treehouse}':0.90, '{Reef}':0.02, '{Peak}':0.02, '{Rainforest}':0.02.

1.3.2 Season predictor $T(V_{\mathcal{T}}|O)$. A few-shot LLM is used to infer the likelihood of all objects appearing in each season. **SEASONS** is all seasons in the list. One example of the prompt is:

Prompt for season prediction

You are now an excellent model for season recognition.
*Season list: **SEASONS**.*
The text I'm giving you is the name and the description of the object. Note that sometimes there is no description of the object.
Please give the probabilities of the object can be appear in the season in season list for the object I give you.
Text: table. Introduction: the table with some snow.

One example of the output is:

Output

Answer: '{spring}':0.20, '{summer}':0.05, '{autumn}':0.20, '{winter}':1.00.

1.4 Whole scene synthesis

1.4.1 Size similarity module $S(o_i, o_j)$. When calculating size similarity, we default that two objects can rotate around the z-axis. For one virtual object o_m^{vir} and one physical object o_n^{phy} , their size similarity can be calculated using the object size bounding box size $s_i = (sx_i, sy_i, sz_i) \in \mathbb{R}^3$ as follows:

$$x_{diff_1} = \begin{cases} 0.01, & \text{if } \frac{sx_m^{vir}}{sx_n^{phy}} < 0.5 \\ \frac{sx_m^{vir}}{sx_n^{phy}}, & \text{if } 0.5 \leq \frac{sx_m^{vir}}{sx_n^{phy}} < 1 \\ -0.5 \times \frac{sx_m^{vir}}{sx_n^{phy}} + 1.5, & \text{if } 1 \leq \frac{sx_m^{vir}}{sx_n^{phy}} \leq 2 \\ 0.01, & \text{if } 2 < \frac{sx_m^{vir}}{sx_n^{phy}} \end{cases} \quad (1)$$

$$y_{diff_1} = \begin{cases} 0.01, & \text{if } \frac{sy_m^{vir}}{sy_n^{phy}} < 0.5 \\ \frac{sy_m^{vir}}{sy_n^{phy}}, & \text{if } 0.5 \leq \frac{sy_m^{vir}}{sy_n^{phy}} < 1 \\ -0.5 \times \frac{sy_m^{vir}}{sy_n^{phy}} + 1.5, & \text{if } 1 \leq \frac{sy_m^{vir}}{sy_n^{phy}} \leq 2 \\ 0.01, & \text{if } 2 < \frac{sy_m^{vir}}{sy_n^{phy}} \end{cases} \quad (2)$$

$$x_{diff_2} = \begin{cases} 0.01, & \text{if } \frac{sx_m^{vir}}{sx_n^{phy}} < 0.5 \\ \frac{sx_m^{vir}}{sx_n^{phy}}, & \text{if } 0.5 \leq \frac{sx_m^{vir}}{sx_n^{phy}} < 1 \\ -0.5 \times \frac{sx_m^{vir}}{sx_n^{phy}} + 1.5, & \text{if } 1 \leq \frac{sx_m^{vir}}{sx_n^{phy}} \leq 2 \\ 0.01, & \text{if } 2 < \frac{sx_m^{vir}}{sx_n^{phy}} \end{cases} \quad (3)$$

$$y_{diff_2} = \begin{cases} 0.01, & \text{if } \frac{sy_m^{vir}}{sy_n^{phy}} < 0.5 \\ \frac{sy_m^{vir}}{sy_n^{phy}}, & \text{if } 0.5 \leq \frac{sy_m^{vir}}{sy_n^{phy}} < 1 \\ -0.5 \times \frac{sy_m^{vir}}{sy_n^{phy}} + 1.5, & \text{if } 1 \leq \frac{sy_m^{vir}}{sy_n^{phy}} \leq 2 \\ 0.01, & \text{if } 2 < \frac{sy_m^{vir}}{sy_n^{phy}} \end{cases} \quad (4)$$

$$z_{diff} = \begin{cases} 0.01, & \text{if } \frac{sz_m^{vir}}{sz_n^{phy}} < 0.5 \\ \frac{sz_m^{vir}}{sz_n^{phy}}, & \text{if } 0.5 \leq \frac{sz_m^{vir}}{sz_n^{phy}} < 1 \\ -0.5 \times \frac{sz_m^{vir}}{sz_n^{phy}} + 1.5, & \text{if } 1 \leq \frac{sz_m^{vir}}{sz_n^{phy}} \leq 2 \\ 0.01, & \text{if } 2 < \frac{sz_m^{vir}}{sz_n^{phy}} \end{cases} \quad (5)$$

$$S(o_m^{vir}, o_n^{phy}) = \min(x_{diff_1} \times y_{diff_1} \times z_{diff}, x_{diff_2} \times y_{diff_2} \times z_{diff}) \quad (6)$$

2 BASELINES

2.1 LLM-based method

Similar to the work [Feng et al. 2023], the LLM-based method predicts the corresponding virtual object for each physical object by using its information as the prompt. Since the virtual objects predicted by LLM can be arbitrary, we use the language similarity module $L(\cdot, \cdot)$ to predict the similarity between the virtual objects obtained by LLM and all virtual objects we have. The object with the highest similarity will be used to synthesize the virtual scene. The virtual object prediction pipeline is shown in figure 1. After selecting the virtual objects corresponding to the physical objects, the following synthesis process is the same as our method.

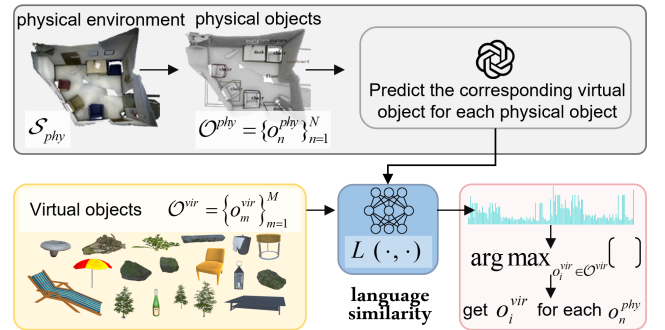


Fig. 1. The virtual object prediction pipeline of the LLM-based method without size consideration.

If only the category of virtual objects is considered, the size of the corresponding virtual objects and physical objects will be very different. Therefore, we considered the size similarity between the virtual object and the real object by using the size similarity module $S(o_i, o_j)$. the final probability is calculated by the size similarity results multiplied by the result from the language similarity module $L(\cdot, \cdot)$, as shown in the figure 2.

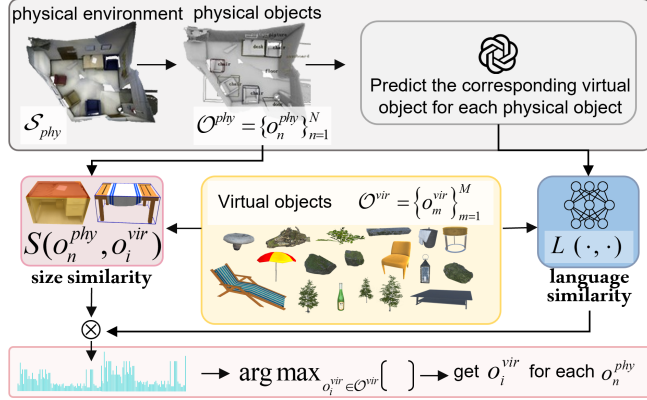


Fig. 2. The virtual object prediction pipeline of the LLM-based method with size consideration.

One example of the prompt for predicting the corresponding virtual object for each physical object is:

Prompt of LLM-based method

You are now an excellent model for object prediction.
The text I'm giving you is the name and size of the object, or the place, or the season.
The size of object includes: length, width and height.
Please tell me about an object that has a similar affordance to the given object in the given place and season (if any).
Text: table, length: 1 meter, width: 0.6 meter, height: 0.8 meter, forest, summer.

One example of the output is:

Output

Answer: big stone

2.2 Semantics-based method

Semantics-based method predicts the corresponding virtual object for each physical object based on the language similarity between the virtual objects and the physical object as shown in figure 3. In addition, similar to the LLM-based method, we also consider the effect of the size as shown in figure 4.

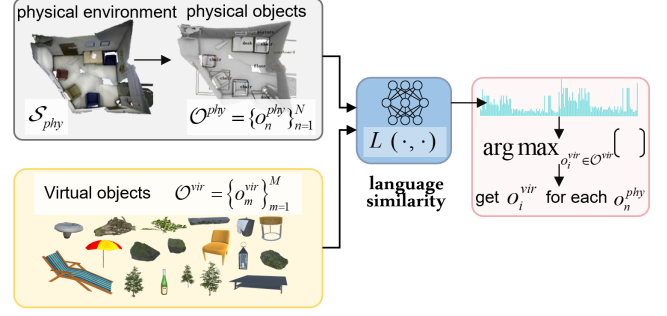


Fig. 3. The virtual object prediction pipeline of the semantics-based method with size consideration.

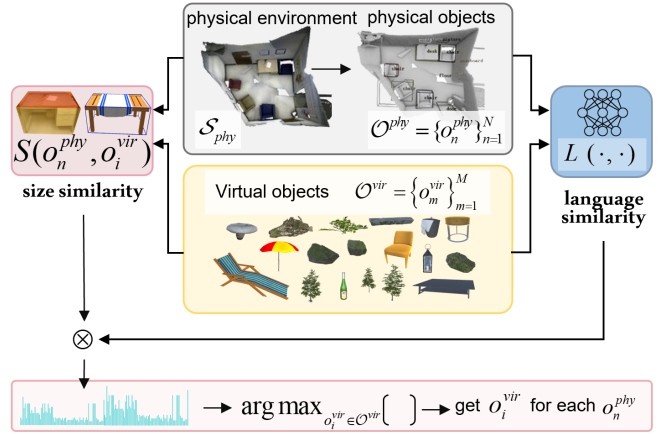


Fig. 4. The virtual object prediction pipeline of the semantics-based method with size consideration.

REFERENCES

- Weixi Feng, Wanrong Zhu, Tsu jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. arXiv:2305.15393 [cs.CV]
- Katikapalli Subramanyam Kalyan. 2023. A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4. arXiv preprint arXiv:2310.12321 (2023).