InstructPart: Affordance-based Part Segmentation from Language Instruction

Supplementary Material

Dataset Distribution

We follow ADE20K (Zhou et al. 2019) to provide the distribution of objects and parts within our InstructPart dataset. As shown in Fig. 1, the dataset comprises 700 data items, encompassing 54 object classes and 46 part classes, which together form 110 distinct object-part pair classes. Besides, we also provide a word cloud to visualize the object and part classes of our dataset, as depicted in Fig. 2. This diversity in class types offers robust criteria for analyzing the proficiency of current models in understanding instructions and segmenting parts.

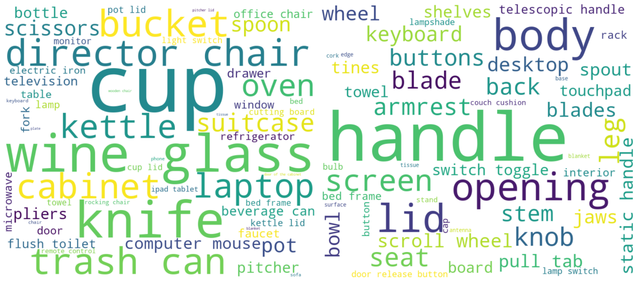


Figure 2: InstructPart dataset object and part classes. The left part shows the object class names and the right part shows the part class names.

Annotation Example

Fig. 3 presents two examples of annotations from our InstructPart dataset, focusing on the handle of a cup and the lid of a pod, respectively. In each JSON dictionary, the names of the object and its specific part are noted, aligned with an instruction that pertains to a particular part shown in the image. Additionally, both a low-level affordance name and a high-level action name are provided in relation to the instruction.

Results on different classes

In Fig. 4, we present the results for each part category of LISA (Lai et al. 2023) on the instruction-referring segmentation task. The average metric is the average of our four evaluation metrics, gIoU, cIoU, P@50:95, and P@50. The results reveal that parts with larger surfaces, such as boards, bed frames, and screens, tend to achieve higher average performance. Conversely, smaller parts like cords, switches, and buttons often yield poorer outcomes. This suggests that while the model is adept at handling simpler cases where parts are easily distinguishable, it still faces challenges with complex samples and smaller targets.

Training Details

In the discussion section of the main paper, we mention that we train LISA (Lai et al. 2023) with our dataset. Here are more details about the experimental setup.

The LISA (Lai et al. 2023) network is initialized with the LISA-7B-v1[1] model, previously trained using LISA's dataset's training and validation data. For this experiment, we employ 8 NVIDIA 3090 GPUs and conduct the training of 600 iterations using the Deepspeed (Rasley et al. 2020) engine. We randomly sample natural language data across four categories in both instruction reasoning and oracle referring tasks: human-annotated instruction, GPT-4 rewritten instruction, object-part name, and object-part-affordance. The latter two data types are integrated into templates to formulate sentences for querying LISA.

For the experimental dataset, we select 226 samples from the InstructPart dataset, ensuring representation according to the original category distribution for training purposes. The remaining 474 samples are reserved for testing. Originally, the LISA-7B-v1 model scored 26.18% in gIoU and 28.10% in cIoU on the test set. After fine-tuning, these metrics improved significantly to 42.01% and 48.75%, respectively. This improvement translates to substantial increases of 15.83% in gIoU and 20.65% in cIoU, demonstrating the remarkable impact of the fine-tuning process. This significant improvement in performance metrics serves as a strong indicator of the high quality of our dataset and its potential for enhancing further training processes.

References

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Rasley, J.; Rajbhandari, S.; Ruwase, O.; and He, Y. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3505–3506.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.

[1] https://huggingface.co/xinlai/LISA-7B-v1

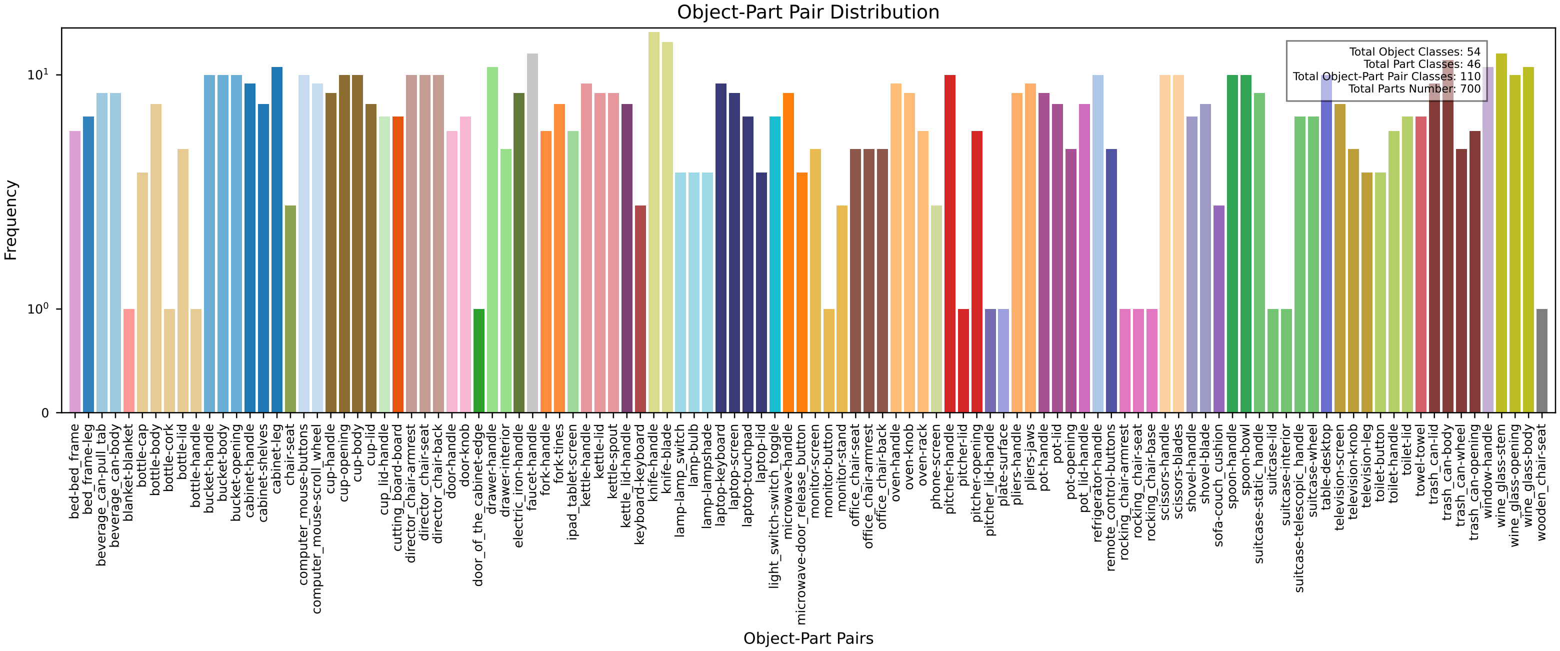
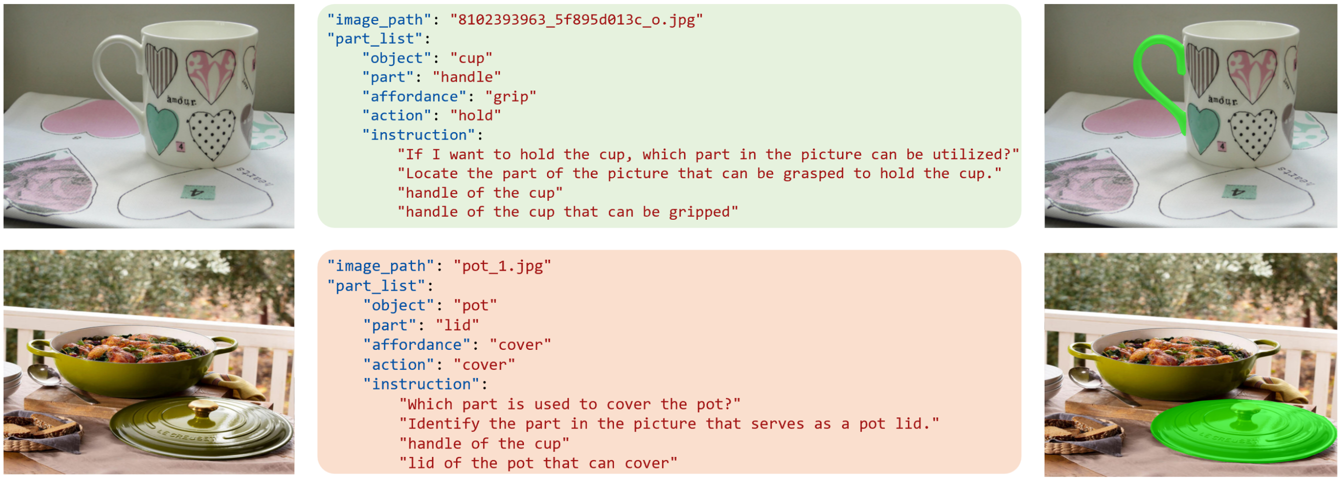


Figure 1: Object-part pair distribution. We collect 700 data pieces in total, containing 54 object classes and 46 part classes, constituting 110 different object-part pair classes. The x-axis shows the name of the object-part pairs, and the y-axis shows the frequency of each item. The parts belonging to the same object classes are highlighted with the same color in the bar chart.



```
"image_path": "8102393963_5f895d013c_o.jpg"
"part_list":
    "object": "cup"
    "part": "handle"
    "affordance": "grip"
    "action": "hold"
    "instruction":
        "If I want to hold the cup, which part in the picture can be utilized?"
        "Locate the part of the picture that can be grasped to hold the cup."
        "handle of the cup"
        "handle of the cup that can be gripped"
```

```
"image_path": "pot_1.jpg"
"part_list":
    "object": "pot"
    "part": "lid"
    "affordance": "cover"
    "action": "cover"
    "instruction":
        "Which part is used to cover the pot?"
        "Identify the part in the picture that serves as a pot lid."
        "handle of the cup"
        "lid of the pot that can cover"
```

Figure 3: Annotation Example: Each data item is represented by a JSON dictionary, which details the components involved. This includes the object to which these parts belong, the name of each part, a specific instruction related to these parts, a low-level affordance associated with the instruction, and a high-level action performed on the parts. Corresponding parts are highlighted in green in the images on the right.

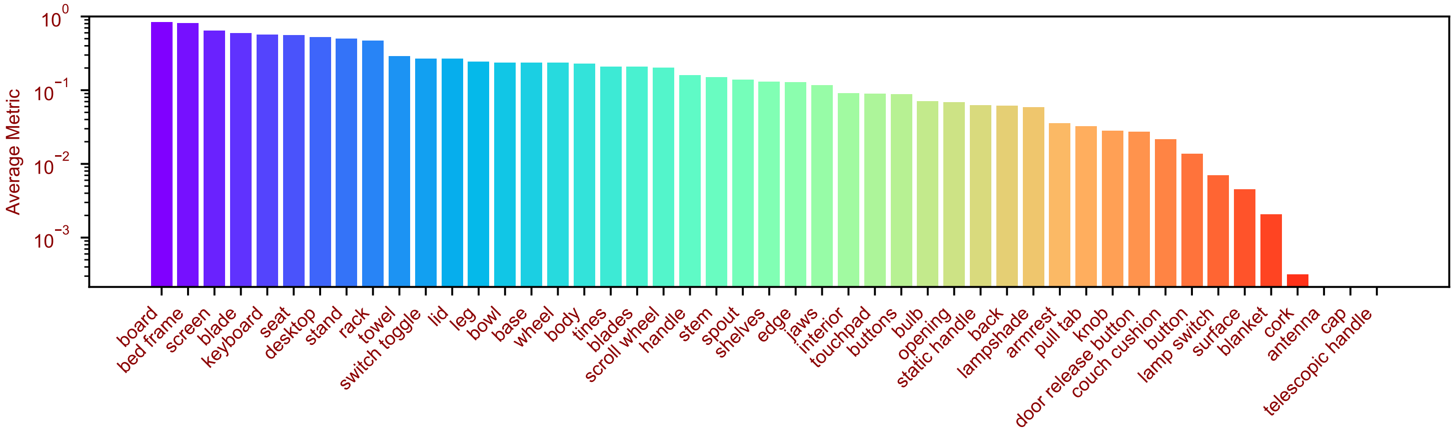


Figure 4: Performance of LISA (Lai et al. 2023) on differnent part classes. The average metric is the average of the four evaluation metrics, namely gIoU, cIoU, P@50:95, and P@50.