AffordMatcher: Affordance Learning in 3D Scenes from Visual Signifiers

Supplementary Material

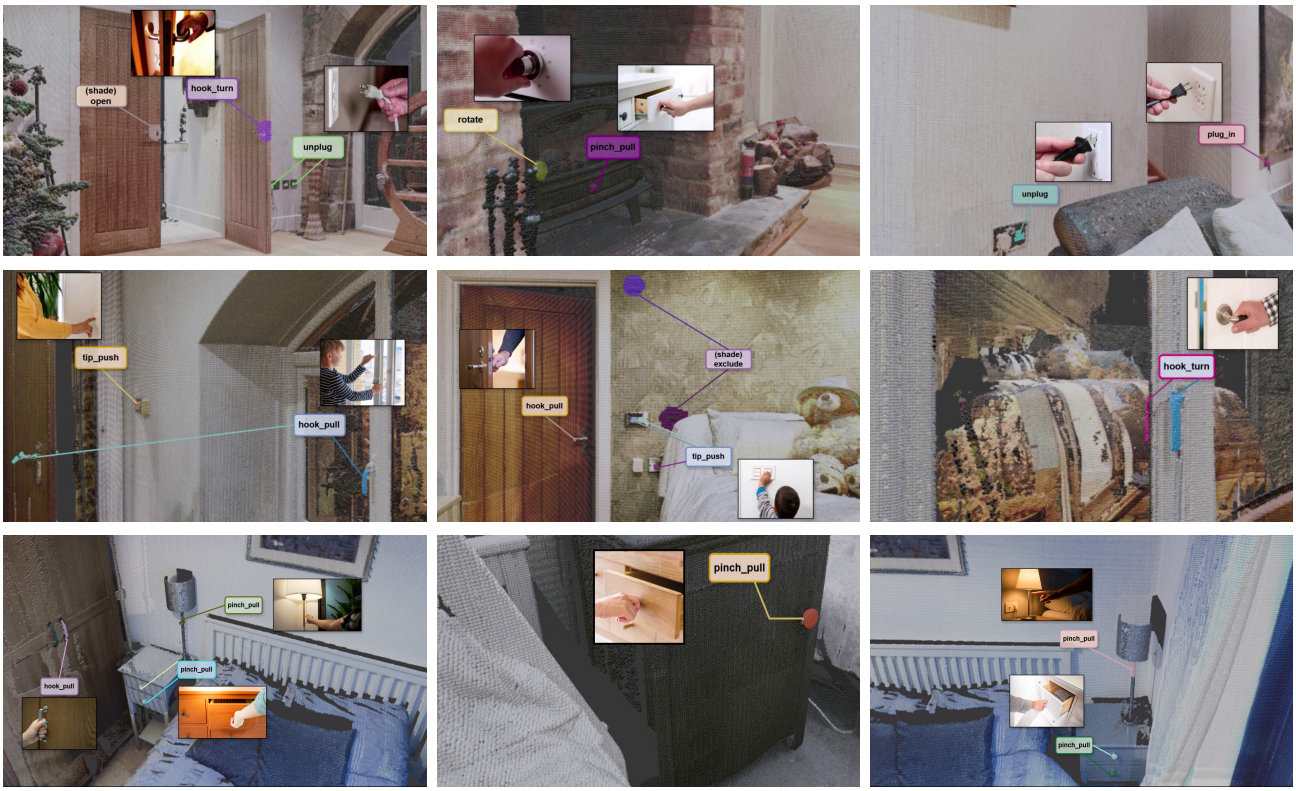


Figure 1. Sample visualization of *AffordBridge* dataset.

Abstract

This supplementary material provides additional dataset analysis, empirical insights, and qualitative evaluations to complement our paper. We begin by providing more details about our AffordBridge dataset, which features visualizations that illustrate affordance areas matched across multimodal inputs, including visual signifiers, textual descriptions, and action labels. We then discuss potential dataset usages, highlighting research directions in 3D scene understanding and human-scene interaction enabled by our annotations. Next, we present results from our user study evaluating the perceptual quality of our proposed AffordMatcher against baselines, demonstrating superior correctness and interpretability. Finally, we visualize representative failure cases of our method to underscore known limitations, such as handling sparse point clouds, complex actions, and ambiguous visuals, thereby reinforcing key observations and identifying opportunities for future work. Please see our video demonstration for a more interactive experience.

1. Dataset Visualization

Fig. 1 provides a sample visualization of the *AffordBridge* dataset, highlighting affordance areas and their corresponding action phrases across different modalities. This figure presents an input scene alongside affordance areas matched to a visual signifier, demonstrating how environmental cues are linked to potential interactions. For more details, please visit our demonstration video.

2. Potential Usages of Our Dataset

Our introduced *AffordBridge* dataset includes annotations for both 3D scenes and RGB images to support affordance reasoning. Below, we outline several exciting research directions that can benefit from leveraging our dataset:

- **3D Scene Understanding.** Traditional approaches to 3D scene analysis often focus on the instance level [2, 7]. By providing annotations for interactive elements on objects, our dataset opens opportunities for addressing various tasks in 3D scene understanding, such as 3D object detection and segmentation.

- **Robotic Manipulation.** Robots with affordance-aware systems can perform tasks more naturally, such as grasping [8], opening [9], or assembling objects in complex, unstructured environments [10]. The release of our dataset could help robotic systems to understand the purpose and function of objects in a 3D context.
- **Human-Scene Interaction.** With affordance masks for 3D indoor scenes and bounding boxes for RGB images, researchers can gain deeper insights into the functional regions of objects and their interactions with humans. This can contribute to the development of more robust human-object interaction models that integrate 2D and 3D data [3, 4], facilitating unified interaction reasoning [5].

3. *AffordMatcher* Analysis

User Study. We conducted a user study to evaluate the perceptual quality of semantic affordance masks produced by our proposed *AffordMatcher* versus three baselines, including Mask3D-F [1], PIAD [11], and Ego-SAG [6]. Twenty experts in 3D vision each reviewed 40 scenes (10 per method), rating the correctness of each affordance mask on a 5-point Likert scale (1 = completely incorrect, 5 = perfect) and selecting the single best segmentation per scene.

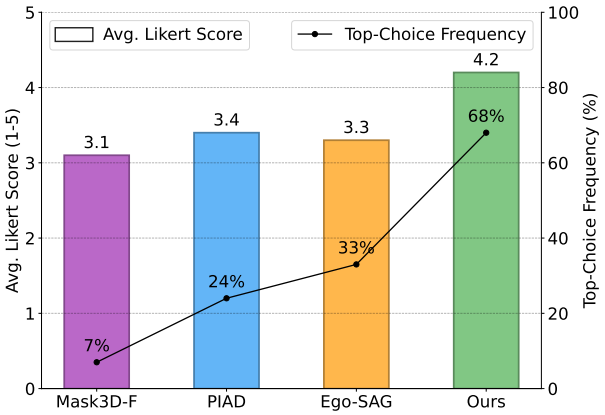


Figure 2. User study results in interaction matching criteria.

Fig. 2 presents the aggregated results: average Likert scores (bars) and the frequency each method was chosen as the top segmentation (line). Our approach achieved an average rating of 4.2, substantially higher than Mask3D-F (3.1), Ego-SAG (3.3), and PIAD (3.4), and was selected as best in 68% of trials, significantly outperforming all baselines ($p < 0.01$, paired t-test).

Participants noted that our masks more accurately captured fine-grained affordance regions, such as chair seats or door handles, and avoided spurious activations common in baseline outputs. This confirms that our *AffordMatcher* not only improves metric performance but also delivers actionable affordance in 3D point clouds.

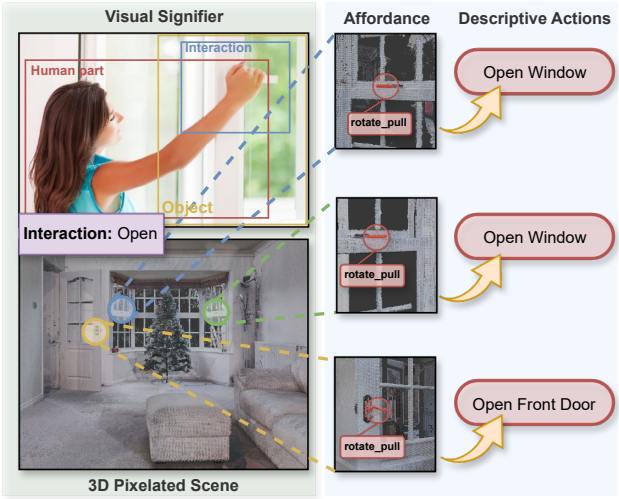
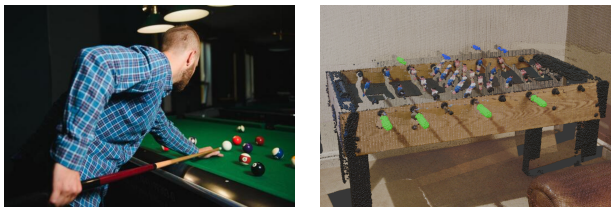


Figure 3. Affordance prediction for a specific interaction, but results in different objects.



(a)

(b)

Figure 4. Fail cases of our method: the visual signifier provides a complex action, which causes failure in reasoning. **Green** areas denote correct predictions; **red** and **blue** areas are false positives and false negatives, respectively.

One-to-many Analysis. In Fig. 3, we show the model's ability to localize the support regions required for a single interaction ("Open") across multiple object instances within the same scene. From the top to bottom rows, we feed the network the same 2D visual cue-an outstretched hand poised to open-with the corresponding 3D voxelized indoor environment. The resulting affordance predictions correctly highlight the window latch, the second window's handle, and finally the front door's knob, each delineated by high-response voxels in the 3D scene. These results show that our *AffordMatcher* flexibly generalizes the "pitch_pull" action to match with the interaction, successfully in semantically analogous parts on different objects, even when their appearance, scale, and orientation vary significantly.

Fail Cases. As outlined in the main paper, our method exhibits known limitations related to the challenge of semantic grounding, which are further exemplified through representative failure cases in Fig. 4. Specifically, such failure cases involve nuanced contextual cues or ambiguous object interactions that current models struggle to resolve without task-specific guidance. For example, in the Fig. 4a, the model failed to analyze the “push" action when the man is playing billiards is from which side, or in the Fig. 4b, the model failed to distinguish whether the action in the visual signifier is the “rotate" or the “push". Nonetheless, these examples serve to reinforce the overall generality of our approach under standard conditions, while motivating future work focused on enhancing semantic grounding and model adaptability in more challenging scenarios.

References

[1] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, et al. Scenefun3d: fine-grained functionality and affordance understanding in 3d scenes. In *CVPR*, 2024. 2

[2] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, et al. Oneformer3d: One transformer for unified point cloud segmentation. In *CVPR*, 2024. 1

[3] Minhao Li, Zheng Qin, Zhirui Gao, et al. 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In *ICCV*, 2023. 2

[4] Yong-Lu Li, Xinpeng Liu, Han Lu, et al. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 2

[5] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, et al. Multi-modal situated reasoning in 3d scenes. *NIPS*, 2024. 2

[6] Cuiyu Liu, Wei Zhai, Yuhang Yang, et al. Grounding 3d scene affordance from egocentric interactions. *arXiv*, 2024. 2

[7] Jiahao Sun, Chunmei Qing, Junpeng Tan, et al. Superpoint transformer for 3d scene instance segmentation. In *AAAI*, 2023. 1

[8] An Dinh Vuong, Minh Nhat Vu, Baoru Huang, et al. Language-driven grasp detection. In *CVPR*, 2024. 2

[9] Ruihai Wu, Kai Cheng, Yan Zhao, et al. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. *NIPS*, 2023. 2

[10] Kashu Yamazaki, Taisei Hanyu, Khoa Vo, et al. Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. In *ICRA*, 2024. 2

[11] Yuhang Yang, Wei Zhai, Hongchen Luo, et al. Grounding 3d object affordance from 2d interactions in images. In *ICCV*, 2023. 2