G-DexGrasp: Generalizable Dexterous Grasping Synthesis Via Part-Aware Prior Retrieval and Prior-Assisted Generation

Supplementary Material

1. Prompt of MLLM

As mentioned in Section 3.2 of our main paper, we prompt the pre-trained MLLM to determine the fine-grained grasping arrangement, i.e. affordance type A and the name of contact part P. We employ GPT-4o from OpenAI API for the task instruction analysis. The prompt we use is shown in Figure 1. By carefully describing the analysis purpose and the definition of the pre-defined action library, we can leverage the pre-trained MLLM to analyze the user-specified input images and instructions in a zero-shot manner, and identify the interaction parts as well as the appropriate action type. This setup relies only on the powerful commonsense reasoning capabilities of MLLM, without the need for human-provided examples.

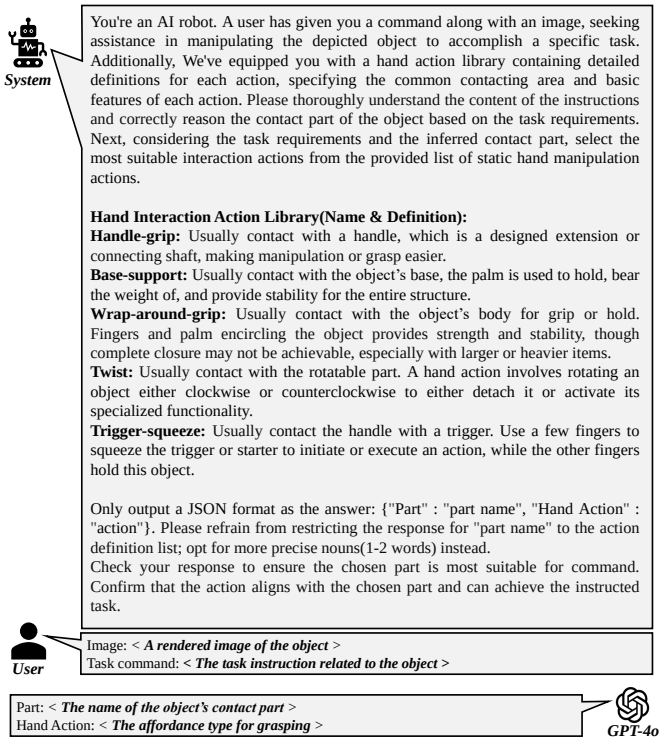

Figure 1. Prompt of MLLM for task instruction analysis.

2. Perceptual Evaluation

As mentioned in Section 4.1 of our main paper, we assess the perceptual score (Percep.Score %) as one evaluation metric for which we conduct a user study of 20 groups of results with 30 participants to judge the quality and task alignment of generated grasps. The participants invited to rate based on the key aspects of **task semantic alignment**, **naturalness**, **plausibility**, and **human-likeness** of the synthesized hand configurations. Specifically, we randomly selected 20 experimental results across 19 distinct object categories that are unseen in the training stage for visualization.

For each object, we provided the specified textual instruction (e.g. *"Please grasp the kettle to use it properly"*) and different experimental results: 5 from comparison experiments and 6 from ablation studies. Each result is rated independently. To facilitate observation, we rendered two images from different viewpoints for each experimental result. Volunteers were asked to rate each result on a scale of 0 to 5, where 5 indicates the highest score and 0 is the lowest. Fig. 2 is a partial screenshot of the user survey questionnaire, which shows two different results for the kettle.

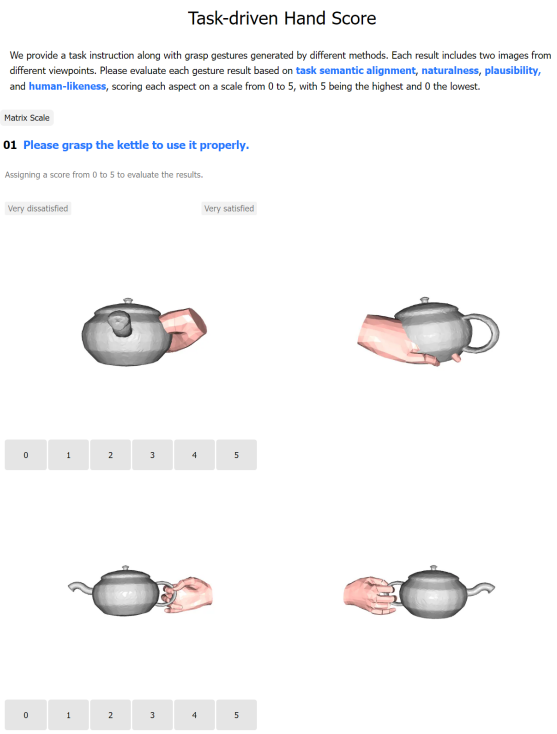

Figure 2. The partial screenshot of the user survey questionnaire.

Experiments	Quality and Stability			Semantic
	Solid.Intsec.Vol↓	Penet.Depth↓	Sim.Dis↓	Part Acc.
Ours-OBB	2.94	0.29	4.27	71.6
Ours-BPS	2.97	0.38	4.53	71.6

Table 1. Quantitative comparison. Ours uses pre-trained models to infer and localize the contact parts, while Our-OBB and Our-BPS refer to representing part with OBB size and encoded features using BPS, respectively.

3. More Results

3.1. Intermediate Results

To better show how G-DexGrasp worked, we take the results of a microwave as an example in Figure 3. First, we randomly color the parts of the object to improve visual perception for MLLM processing. Quantitative evaluations on unseen datasets using GPT-4o with our customized prompts achieve 96.99% accuracy in grasp-type parsing and 76.6% precision in combined part parsing and grounding. Subsequently, grasp priors associated with *Handle-grip* and *Handle* are accurately retrieved (OBB-Based). Notably, the grasp priors within a part of clusters have been showing, revealing significant variations strongly correlated with part semantics and tasks. Then, the Part-Aware Grasping Generation Network subsequently predicts intrinsic parameters, where initial poses achieve semantic-task alignment but exhibit penetration artifacts (e.g., pinky finger) and incomplete contacts (e.g., index finger). The first optimization stage refines extrinsic parameters with the grasp priors constraints, improving surface conformity while some penetrations and incomplete contacts remain. The second stage jointly adjusts intrinsic and extrinsic parameters, achieving a physically plausible, contact-compliant grasp.

Moreover, although the five affordance types already cover common grasp-related tasks and help avoid local optima during the optimization stage, we further introduce two additional hand-related affordance types, namely *point* and *press*. With these seven types, the affordance prediction accuracy reaches 96.66% and the part grounding accuracy 75.25%, both comparable to the original five-type setting.

3.2. BPS-based Results

In Section 3.2 of the main paper, we employ the part-level OBB size for dataset clustering and prior retrieval, which is simple yet effective as shown in the results. Notably, it is easily replaced with more advanced geometric descriptors. For example, we implemented the Basis Point Set (BPS) [4] to encode 3D part shapes, and measuring inter-part feature distances using cosine similarity. As shown in Table 1, the experimental results show that this BPS-based approach performs similarly to the OBB-based method while significantly outperforming other comparative approaches (shown in the main paper) across multiple evaluation metrics.

Diversity	T_{std} (↑)	R_{std} (↑)	J_{std} (↑)
AffordPose	0.49	40.91	12.94
GrabNet†	0.47	37.51	13.21
GraspTTA†	0.43	65.51	17.43
Ours	0.57	78.57	13.18

Table 2. Quantitative diversity metrics compared to the baseline based on the test-set of AffordPose dataset. T_{std}, R_{std} and J_{std} represent the standard deviations of translation, rotation, and joint angles, respectively.

Experiments	Quality and Stability			Semantic
	Solid.Intsec.Vol↓	Penet.Depth↓	Sim.Dis↓	Part Acc.
Objaverse-PartField	5.93	0.90	5.31	70.0
Few Intrinsics	2.35	0.27	5.91	70.23
Few Extrinsics	4.25	0.43	5.28	71.24
Fast Version	3.32	0.37	5.78	73.91
Ours	2.94	0.29	4.27	71.6

Table 3. Quantitative evaluation for generalization and scalability. Objaverse-PartField denotes the full model evaluated on the open-set Objaverse dataset and pre-segmented by PartField. "Few-Intrinsics" and "Few-Extrinsics" refer to reduced supervision settings with limited intrinsic and extrinsic parameters, respectively.

3.3. Diversity Results

Our work focuses on task-driven semantic grasping, aiming to synthesize anthropic and reasonable hand grasps that facilitate task completion. While we do not explicitly pursue grasp diversity - as excessive variation may result in unnatural hand configurations misaligned with human preferences or compromised task performance - we quantify the diversity of synthesized hands through the standard deviations analysis of translation, rotation, and intrinsic joint angle. As demonstrated in Table 2, our method exhibits competitive performance in translational and rotational diversity, while showing potential for improvement in joint angle variations. Notably, since baseline methods exhibit degraded performance on unseen test data, our diversity metrics are evaluated exclusively on the AffordPose test-set to ensure valid grasp references for comparative analysis.

3.4. Evaluation on Open-set Dataset

To better evaluate generalization in open-set scenarios, we select the first 500 objects from the Objaverse-tiny dataset [1] and filter out semantically ambiguous or non-graspable instances (e.g., boats). The remaining objects are resized to a canonical scale and pre-segmented using PartField [3] and subsequently processed through our proposed pipeline. The resulting performance on this setting, referred to as *Objaverse-PartField*, is comparable to that on our unseen test set (see Table 3), highlighting the robustness and generalizability of our approach.

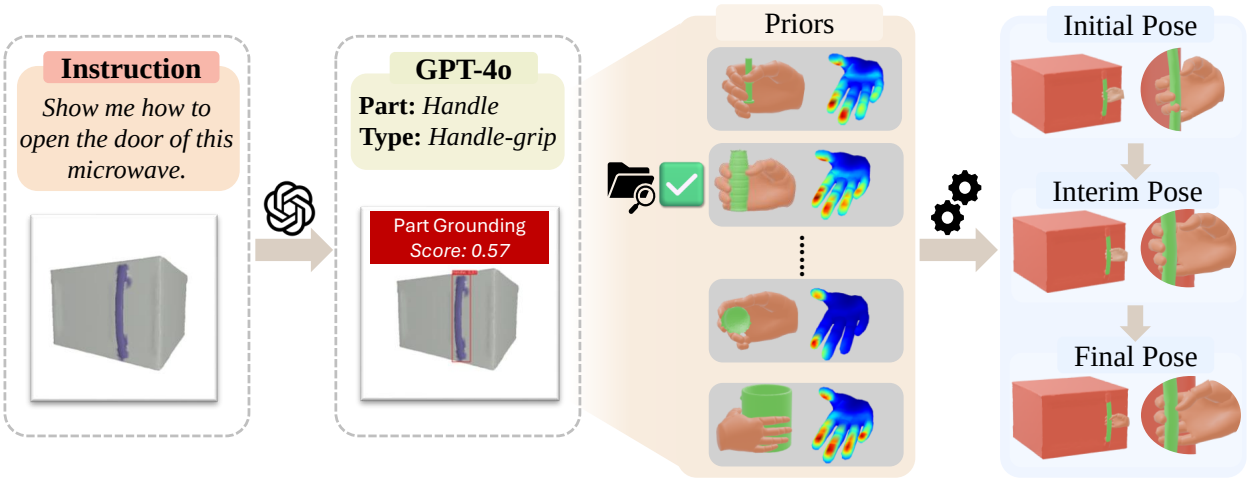

Figure 3. The intermediate results of our G-DexGrasp. (1) Model input (2) MLLM output and grounding results; (3) Grasp priors (including intrinsic params and hand contact maps) visualization (4) Predicted grasp poses in three stages

3.5. Evaluations for Scalability

We conduct the following experiments to validate the scalability of our method:

"Few-Intrinsics": We reserve only 30 intrinsic parameters per affordance type for the retrieval, which is approximately 1/100 of our original dataset.

"Few-Extrinsics": We use only 1/10 of the extrinsic parameters from the original dataset to train the network.

These experiments perform comparably with our full approach and much better than the baselines in our main paper, validating that our approach is scalable with a small set of fine-grained labels. However, once the retrieved prior is completely removed, the performance decreases dramatically as the experiment "w/o prior guide." in the main paper.

3.6. More Visual Results

Figure 4 presents comparison results on additional object categories. The first row shows grasps on a scissors from a seen category, where all methods generate reasonable poses and contact regions, except GraspTTA [2], which produces an unstable grasp. The remaining rows show unseen categories, where existing methods often yield unreasonable results with severe interpenetration and grasp failures when test data deviates from training data.

Figure 5 shows the ablation experiment results on the same set of objects from unseen categories. The first two experiments, i.e. Object-based Net. and Part Rand Init., produce obviously worse results. Once the dexterous hands are absurdly initialized around the object, it is difficult to optimize the extrinsic and intrinsic parameters to correct them, such as the umbrella case in the first row. By contrast, without a refinement optimization, i.e. W/O Optim., although the generated grasps look reasonable, it is noticeable that these grasp cannot stably hold the objects for the subsequent tasks, see the umbrella and kettle cases, while it remains severe penetration, see the bell pepper, coffee-machine cases. The last two experiments, W/O Prior Guid. and One-Stage Optim., produces nearly satisfactory results as ours. However, from the zoom-in visualizations, we can see that these generated results are actually un-optimal, also with remaining penetration (e.g., the W/O Prior Guid. results in the second and the fifth row), unstable hand intrinsics (e.g., the W/O Prior Guid. and the One-Stage Optim. result for the umbrella case and the screwdriver case in the fifth row), and discontaction between hand and object (e.g., the One-Stage Optim. results for bell pepper in the last row), etc.

References

[1] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 2

[2] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11087–11096, 2021. 3

[3] Minghua Liu, Mikaela Angelina Uy, Donglai Xiang, Hao Su, Sanja Fidler, Nicholas Sharp, and Jun Gao. Partfield: Learning 3d feature fields for part segmentation and beyond. *arXiv preprint arXiv:2504.11451*, 2025. 2

[4] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4331–4340, 2019. 2

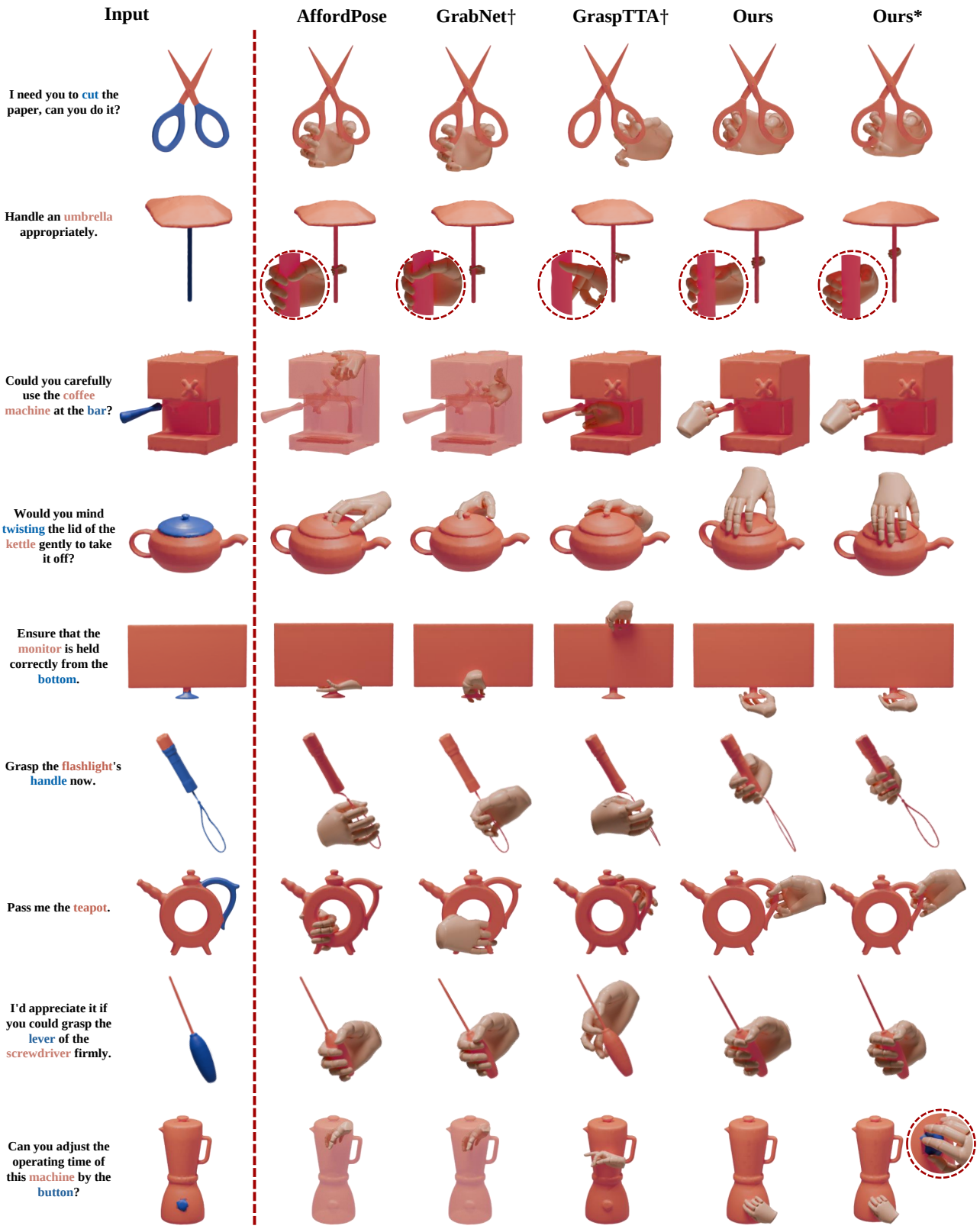

Figure 4. More visual results of comparison experiments. The parts that align with the task instructions are highlighted in blue.

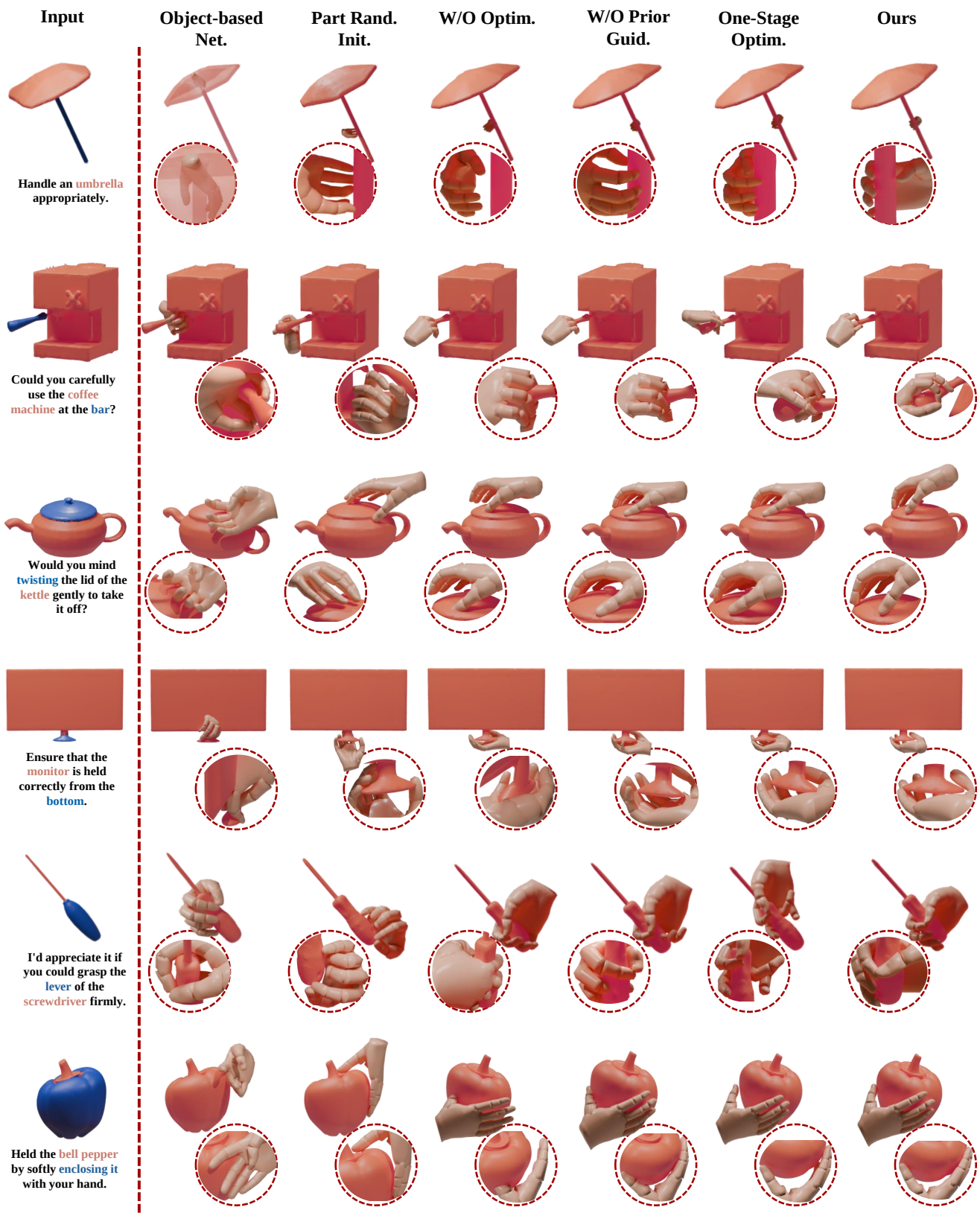

Figure 5. More visual results of ablation experiments. The parts that align with the task instructions are highlighted in blue. The dashed circles in the lower corner of each result show the corresponding partial enlarged views.