MaskDexGrasp: Generative Masked Modeling for Part-Aware Dexterous Grasp Synthesis

Supplementary Material

A. Overview

This supplementary material is organized in the following sections:

- Section B: Details of the network and implementation.
- Section C: Dataset construction and energy functions.
- Section D: Definitions of metrics.
- Section E: Details of real-world experiments.
- Section F: User study of generated grasps.
- Section G: Additional qualitative results.
- Section H: Failure cases and limitations.

B. More Details

Network Architecture. Our model employs a part-aware VQ-VAE that decomposes the Shadow Hand into six anatomical components, including one palm and five fingers. For each part i, we sample a vertex set $\mathbf{H}_i$ and feed it into a PointNet-based encoder to extract a 1024-dimensional point feature $\mathbf{f}_i = \textit{PointNetEncoder}(\mathbf{H}_i)$, which is mapped into a 512-dimensional latent embedding through linear layers $\mathbf{z}_i = \text{MLP}(\mathbf{f}_i)$. Each $\mathbf{z}_i$ is subsequently quantized via a part-specific quantizer. For the decoder, it consists of a fully-connected architecture with hidden layers $[1024, 256, 128, 64]$, followed by a linear alignment layer. It reconstructs the complete Shadow Hand parameters from the concatenated latent embeddings via $\hat{\mathbf{g}} = \text{Decoder}(\hat{\mathbf{z}})$.

Implementation Detail. For the grasp tokenizer, the codebook dimension is set to 256×512, and the exponential moving average coefficient used for updating the codebook is set to 0.99. For the masked grasp transformer, we adopt a learnable task embedding module to extract the task embedding of dimension 512, adopt a CLIP encoder ("ViT-B/32") to extract the text embedding of dimension 512, and adopt a PointNet-based encoder to extract the object embedding of dimension 1024. During inference, we use the guidance scale $s = 4$ and adopt $T = 4$ iterations for both subsets. Meanwhile, we uniformly sample 2000 points around the surface of the object.

C. TDG Dataset

C.1. Dataset Structure

In this paper, we introduce a large-scale, **T**ext-rich **D**exterous **G**rasp dataset, named *TDG*, which is constructed from the AffordPose [24] and the OakShape [72] datasets. 1) For the AffordPose, it comprises 26,709 hand-object interaction instances across 641 object models and 8 task categories. The released version provides both hand meshes and 16-DoF hand parameters to represent a human hand. All grasping poses, along with corresponding objects, are used to form our *Subset 1*. 2) For the OakShape, it is a subset of OakInk and contains 62,046 grasping poses with 1,801 objects. Four tasks including "use", "hold", "lift-up", and "handover" are included in this dataset. Since our method is designed for single-hand grasp generation, "handover" that involves collaborations between two hands is excluded, leaving 38,182 grasping poses to form our *Subset 2*. MANO parameters [48] are available to represent hand models in the OakShape. In Tab. 4, we summarize the scale and attribute of our dataset. And in Fig. 8, we illustrate the distribution characteristics of both two subsets.

Table 4. Scale and attribute of the TDG dataset.

Name	grasps	objects	tasks	texts
Subset 1	26,709	641	8	106,836
Subset 2	38,182	1,655	3	152,728
Total	64,891	2,296	11	259,564

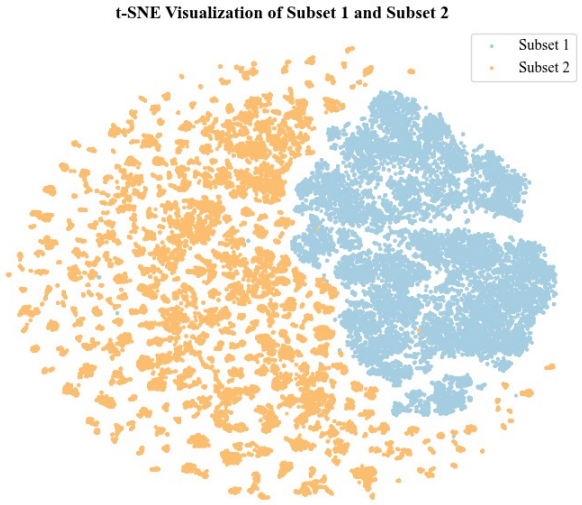


Figure 8. t-SNE visualization of the two subsets in our dataset.

C.2. Grasp Initialization

As we mentioned previously, the grasp representation is different in the AffordPose and OakShape datasets. Therefore,

we adopt hand joint positions $\mathbf{J}_k$ to establish a unified representation. Based on it, a joint-level retargeting solution [44] is used to map human hand joints onto the corresponding poses of the Shadow Hand. This procedure is formulated as:

$$\arg\min_{\mathbf{g}_0} \sum_{k=0}^{K} \|\mathbf{J}_k - f_k(\mathbf{g}_0)\|_2 \tag{9}$$

where $\mathbf{g}_0$ denotes the initial grasp, f_k is a forward kinematics function that maps the grasp $\mathbf{g}_0$ into the i-th keypoint of the Shadow Hand.

C.3. Optimization Functions

Directly combining the initial grasp $\mathbf{g}_0$ with the object often results in obstacles such as penetration or floating. Therefore, to optimize initial grasps, we utilize a composite energy function to encourage physically plausible interactions. The overall optimization objective is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_{\text{tip}}\mathcal{L}_{\text{tip.}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg.}} \\ + \lambda_{\text{spen}}\mathcal{L}_{\text{self-pen.}} + \lambda_{\text{pen}}\mathcal{L}_{\text{pen.}},\end{aligned} \tag{10}$$

where $\mathcal{L}_{\text{tip.}}$ supervises the fingertips with a weight of $\lambda_{\text{tip}} = 10.0$, $\mathcal{L}_{\text{reg.}}$ penalizes extreme grasps with a weight of $\lambda_{\text{reg}} = 0.01$, $\mathcal{L}_{\text{self-pen.}}$ and $\mathcal{L}_{\text{pen.}}$ penalize penetrations with weights $\lambda_{\text{spen}} = 10.0$ and $\lambda_{\text{pen}} = 10.0$, respectively. The detailed description of each term is as follows.

Fingertip Alignment Loss $\mathcal{L}_{\text{tip.}}$. To enforce consistency between the fingertips of the Shadow Hand and the target MANO hand, we minimize the weighted ℓ_2 distance between their respective fingertip coordinates:

$$\mathcal{L}_{\text{tip.}} = \frac{1}{5}\sum_{i=1}^{5} w_i \left\|\mathbf{J}_i^{\text{dex}} - \mathbf{J}_i^{\text{mano}}\right\|_2, \tag{11}$$

where $\mathbf{J}_i^{\text{dex}}$ and $\mathbf{J}_i^{\text{mano}}$ represent the fingertip positions of the i-th finger, and w_i is a manually assigned finger weight, reflecting the observation that the thumb and index fingers contribute more to achieving a plausible grasp:

$$[w_1, w_2, w_3, w_4, w_5] = [30.0, 10.0, 1.2, 1.0, 0.2].$$

Pose Regularization Loss $\mathcal{L}_{\text{reg.}}$. To prevent unnatural hand poses, we apply a loss term to restrict out-of-range joint angles, and apply an ℓ_2 regularization to discourage extreme poses:

$$\begin{aligned}\mathcal{L}_{\text{reg.}} = \sum_j \left(\max(0, \theta_j - \theta_j^{\text{up}}) + \max(0, \theta_j^{\text{low}} - \theta_j)\right) \\ + \|\theta\|_2,\end{aligned} \tag{12}$$

where θ denotes the joint angles, and $\theta^{\text{low}}, \theta^{\text{up}}$ represent the valid angle range defined by the Shadow Hand model.

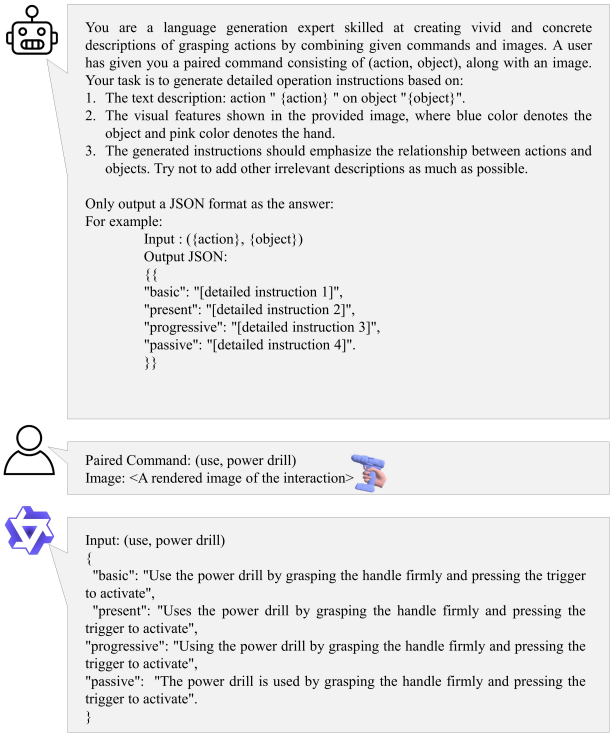


Figure 9. VLM prompt used for generating textual descriptions.

Self-Penetration Loss $\mathcal{L}_{\text{self-pen.}}$. To preserve physical plausibility within the hand structure, we compute the intersection volumes among different hand segments and penalize any overlapping regions:

$$\mathcal{L}_{\text{self-pen.}} = \sum_{(i,j)} \max(0, \tau - \|\mathbf{p}_i - \mathbf{p}_j\|_2), \tag{13}$$

where $\{\mathbf{p}_i, \mathbf{p}_j | i \neq j\}$ refers to pre-defined penetration points, τ is set to 0.02 in practice.

Object Penetration Loss $\mathcal{L}_{\text{pen.}}$. To prevent interpenetration between the hand mesh and the object surface, we compute signed distances to penalize implausible penetrations:

$$\mathcal{L}_{\text{pen.}} = \sum_{v \in \mathcal{V}_{\text{hand}}} \max(0, -d(v, \mathcal{S}_{\text{obj}})), \tag{14}$$

where $d(v, \mathcal{S}_{\text{obj}})$ denotes the signed distance from the hand vertex v to the object surface $\mathcal{S}_{\text{obj}}$.

C.4. VLM Prompt

As discussed in Sec. 4, to enrich our dataset with detailed textual descriptions, we develop a language grounding system based on the Qwen VLM [1]. By carefully designing the prompt, this system integrates the given image and concise user command to produce vivid grasp descriptions. More details regarding the prompt are provided in Fig. 9.

D. Metrics

We first supplement detailed definitions for certain metrics in the main paper, and then introduce additional perceptual metrics to assess the authenticity of the generated grasps.

Detailed Definitions. 1) *Q1* is defined as the radius of the inscribed sphere within the convex hull, representing the norm of the smallest wrench that can destabilize the grasp. A larger Q1 value indicates a stronger resistance to perturbations. Following the standard practice, we apply a contact threshold of 1cm and limit each link to at most one contact point to ensure computational efficiency. 2) H_{mean} and H_{std} quantify the diversity through joint angle entropy. For each angle, we discretize its range into 100 bins and use all generated samples to estimate a probability distribution. The diversity score is then computed as the entropy of this distribution. We use H_{mean} to reflect the overall diversity of generated grasps, and H_{std} to measure how uniformly this diversity is distributed across different angles.

Perceptual Scores. To report the perceptual quality of generated grasps, we require participants to score them based on the following criteria. 1) *Plausibility* reflects how physically plausible the generated grasps appear to human observers. It quantifies the plausibility through two aspects, including the correctness of interaction physics and the conformity to typical human grasping behavior. 2) *Semantics* assesses the consistency between the generated grasp and the provided semantic description. It specifically evaluates how well the grasp matches the grasping style and position selection requested in the text.

E. Real-World Experiments

E.1. Experimental Settings

Our real-world platform, as illustrated in Fig. 10, is composed of an XArm7 robotic arm, a RealSense D435 camera, a Freedom five-fingered dexterous hand, and manipulated objects. In addition to discussions in the main paper, we supplement more details on how to transform the generated Shadow Hand poses into the Freedom five-fingered dexterous hand. Additional experiments conducted on this real-world platform are also described.

E.2. Retargeting Details

Our generative framework is based on the Shadow Hand model, which offers rich articulation for precise grasp synthesis. In contrast, our real-world platform employs a cost-effective five-fingered robotic hand equipped with only six actuated DoFs. Although this discrepancy reduces the expressiveness of the real-world platform, the overall hand structure remains sufficiently similar. Consequently, we perform a simplified retargeting procedure that maps the root-joint DoFs of each Shadow Hand finger to the corresponding controllable joints of the physical hand. At the same time, the remaining articulation is implicitly achieved through mechanical coupling. Although the reduced dexterity limits fine-grained manipulation, we find that such coarse finger-posture alignment is sufficient for the evaluated tasks. From a task-completion perspective, the performance decline is negligible, demonstrating that our framework produces robust grasps even on hardware with substantially constrained articulation.

Figure 10. Real-world robotic platform. (Left) A real robotic platform comprises an XArm7 robot and a Freedom five-fingered dexterous hand. (Right) 8 real objects are used for evaluation.

E.3. Experimental Results

The summary of real-world experiments regarding the number of success is presented in Tab. 5, where each object is randomly placed on the table and subjected to 10 independent grasping attempts. We observe that several objects present challenges when performing our method. For instance, when picking up a screwdriver, due to the thin handle, it always slides out of the hand.

Table 5. Results of the real-world experiments, where each object is subjected to 10 grasping attempts.

Name								
Total	7	6	10	8	8	7	5	4

F. User Study

Definitions. Given the critical role of human perception in assessing a generative model, we perform a dedicated user study to investigate the plausibility and semantic alignment of the generated grasps. This study involves 12 distinct objects, with six selected from *Subset 1* and six from *Subset 2*. For each object, we consider 2 randomly sampled grasps from our method and partial baselines [62, 83], as they also obtain results guided by additional semantic conditions. To ensure comprehensive evaluation, we organize 15 volunteers to participate in our questionnaire and ask them to

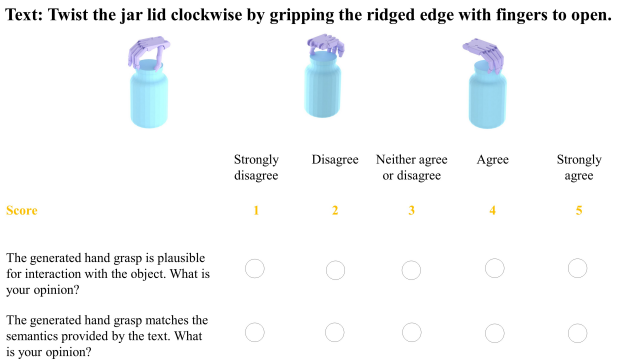


Figure 11. Questionnaire layout. A higher score reflects increased plausibility and stronger semantic consistency.

Table 6. User study results. Higher scores reflect greater plausibility and more accurate semantic alignment of the generated grasps.

Method	plausibility	semantics
DexGYS	3.60 ± 1.752	3.60 ± 1.657
DexGraspAnything	3.76 ± 1.583	3.82 ± 1.476
Ours	3.90 ± 1.216	3.90 ± 1.164

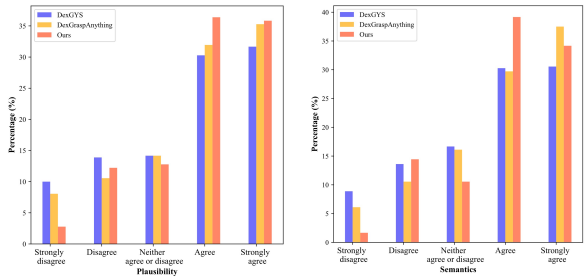


Figure 12. Distribution of perceptual scores. A higher distribution in the "Strongly agree" indicates that our method achieves satisfactory plausibility and semantic alignment compared to baselines. The left side corresponds to performance in terms of plausibility, while the right corresponds to semantic consistency.

answer two questions about the plausibility and semantics of each grasp. These questions include: 1) *"The generated hand grasp is plausible for interaction with the object. What is your opinion?"*; 2) *"The generated hand grasp matches the semantics provided by the text. What is your opinion?"* Participants are required to score both questions on a scale from 1 (Strongly disagree) to 5 (Strongly agree), where a higher score indicates greater plausibility and more accurate alignment with the specified semantic requirement.

Results. The layout of our questionnaire is illustrated in Fig. 11, which contains 3 rendered views for each interaction. We report the mean and variance of perceptual scores

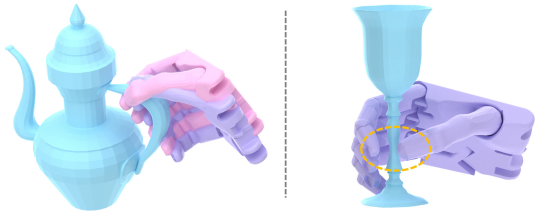


Figure 13. Failure cases of our method.

in Tab. 6. From these results, we conclude that our method receives higher recognition from the participants, indicating that our framework effectively generates grasps with both high plausibility and semantic alignment. Fig. 12 depicts the distribution of human perceptual scores for plausibility (left) and semantics (right). For items corresponding to "xxx agree", a higher percentage demonstrates superior performance, whereas for items corresponding to "xxx disagree", a higher percentage reflects poorer generation.

G. Additional Results

In addition to visualizations presented in the main paper, we exhibit more qualitative results and comparisons in this appendix, including additional generated grasps, diversity, and comparisons.

Fig. 14 presents the generated grasps from our method on *Subset 1* (left) and *Subset 2* (right). Three views are provided for each instance.

Fig. 15 illustrates the diversity of generated grasps, which are conditioned on consistent object point clouds and textual descriptions. Three views and two grasps are provided for each instance.

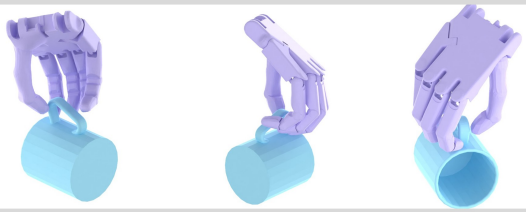
Fig. 16 provides additional comparisons between our method and other baselines.

H. Failure Cases and Limitations

Although our method demonstrates satisfactory performance, several limitations remain. First, as illustrated in Fig. 13 (left), our method occasionally generates similar grasps under identical guidance. As we mentioned in Sec. 5, this deficiency arises partly from the inherent limitations of the VQ-VAE in capturing diversity, and partly from the limited dimension of our tokenizer. Due to the fact that the designed tokenizer only operates 6 tokens in total, the obtained token indices sequence and even editable performance are constrained. Second, the current framework lacks physical supervision, making certain implausible interactions difficult to avoid. The right side of Fig. 13 shows that our method produces floating grasps on objects with extremely thin shapes. To address this limitation, incorporating affordance cues and additional physics-based constraints would be a promising direction for future works.

" Wrap-grasp the mug by encircling its body, positioning fingers along the side "

" Use the mug by grasping the handle with your fingers "



" Twist the jar lid clockwise by gripping the ridged edge with fingers to open "

" Use the knife by grasping the handle firmly and positioning the blade towards the cutting surface "

" Press the dispenser by applying downward force on the top nozzle with the thumb finger "

" Use the headphones by grasping the headband and placing it over your head "

" Wrap-grasp the body of the bottle by curling the fingers around it securely "

" Use the cylinder bottle by grasping it around the middle section "

" Support the pot by placing the hand underneath its base to stabilize it "

" Lift up the eyeglasses by grasping the temples with your fingers "

Figure 14. Additional grasps generated by our framework.

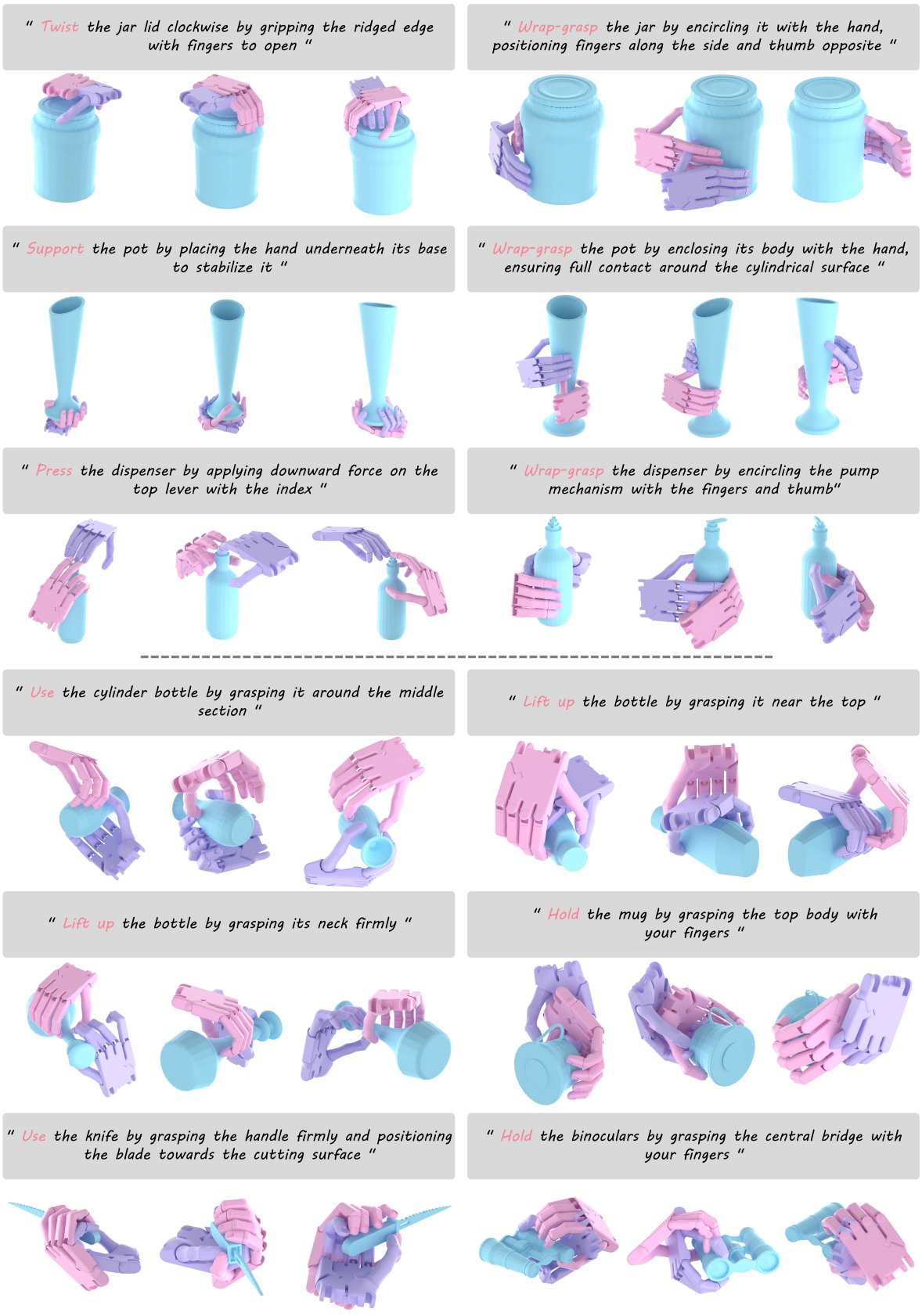


Figure 15. Additional diverse grasps generated by our framework.

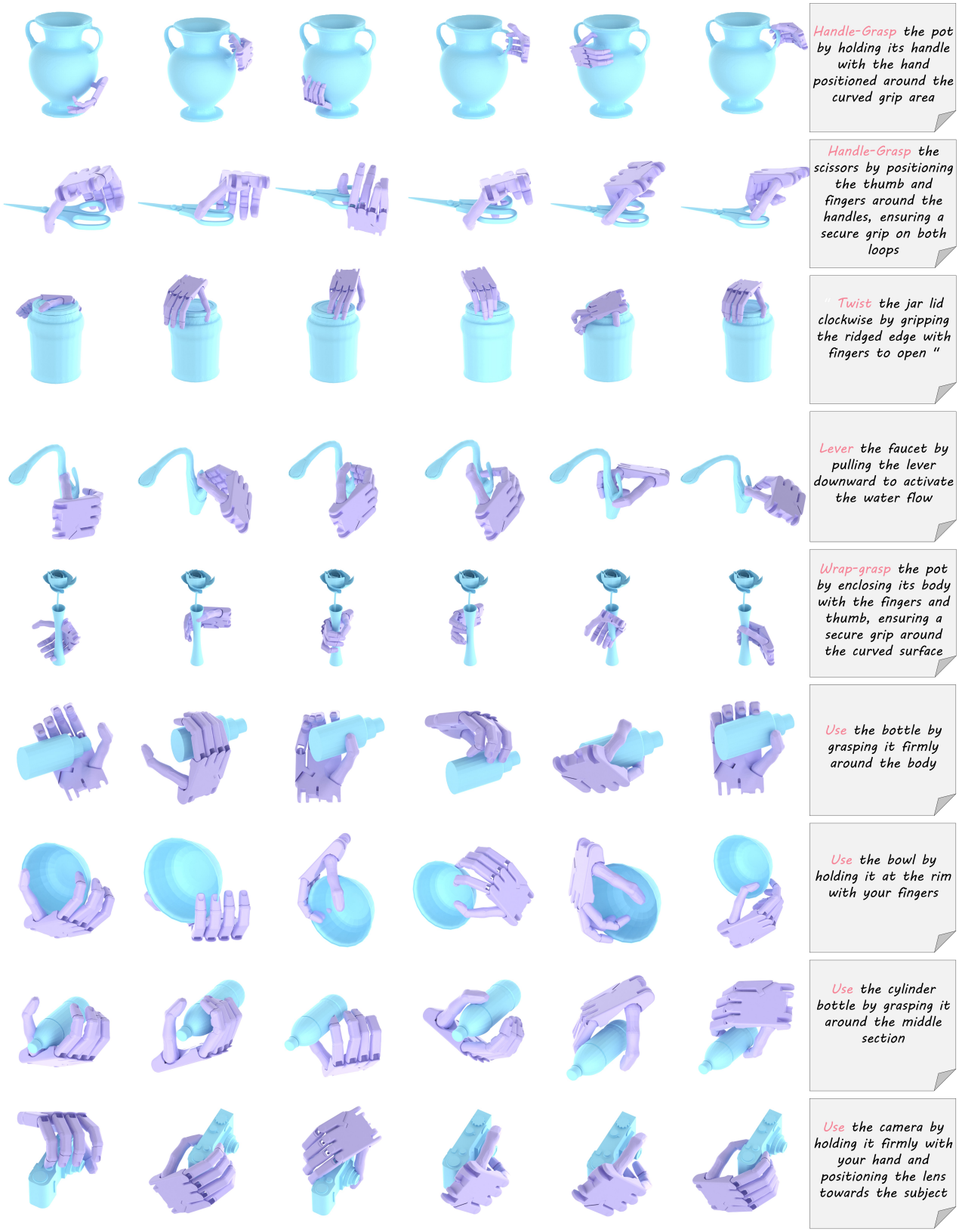


Figure 16. Additional comparisons between baselines and our framework.