AffordGrasp: Cross-Modal Diffusion for Affordance-Aware Grasp Synthesis

Xiaofei Wu[1], Yi Zhang[1], Yumeng Liu[3], Yuexin Ma[1], Yujiao Shi[1*], Xuming He[1,2*]
[1]ShanghaiTech University, Shanghai, China
[2]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[3]University of Science and Technology of China

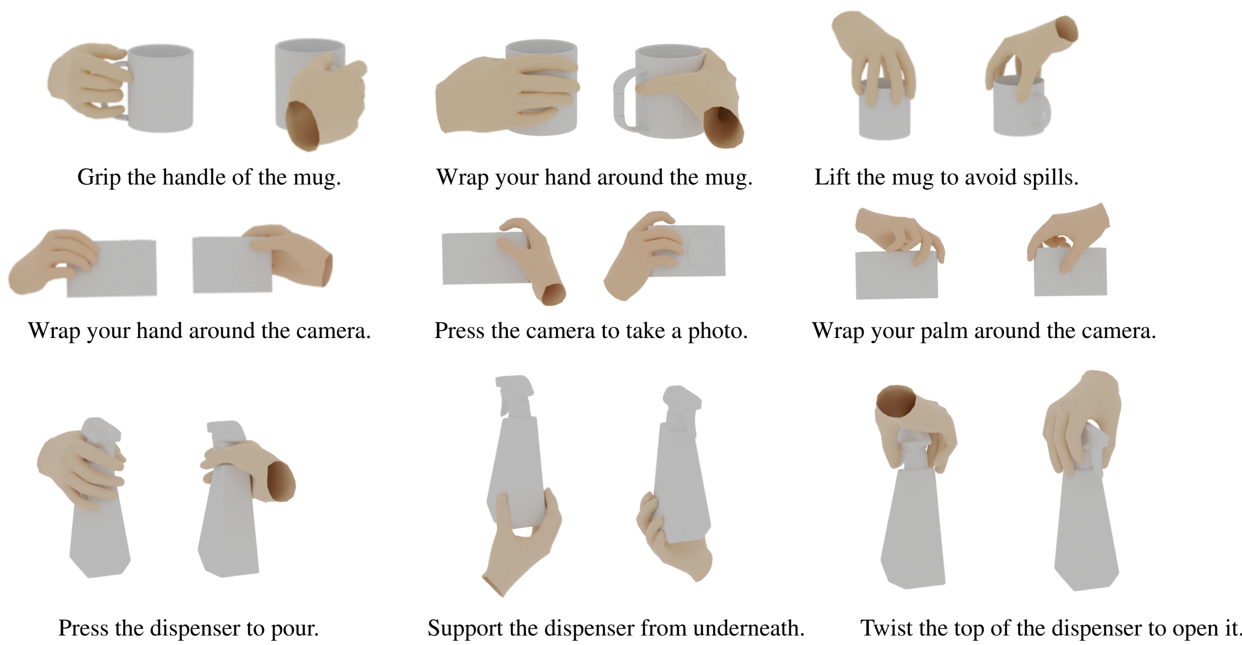

Figure 1. **AffordGrasp** produces realistic and instruction-aligned hand grasps directly from textual descriptions. For each object, we demonstrate the model's capacity to generate semantically diverse and physically plausible grasps across three textual instructions, visualized from two viewpoints to reveal the precision and robustness of the synthesized hand poses.

Abstract

Generating human grasping poses that accurately reflect both object geometry and user-specified interaction semantics is essential for natural hand–object interactions in AR/VR and embodied AI. However, existing semantic grasping approaches struggle with the large modality gap between 3D object representations and textual instructions, and often lack explicit spatial or semantic constraints, leading to physically invalid or semantically inconsistent grasps. In this work, we present AffordGrasp, a diffusion-based framework that produces physically stable and semantically faithful human grasps with high precision. We first introduce a scalable annotation pipeline that automatically enriches hand–object interaction datasets with fine-grained structured language labels capturing interaction intent. Building upon these annotations, AffordGrasp integrates an affordance-aware latent representation of hand poses with a dual-conditioning diffusion process, enabling the model to jointly reason over object geometry, spatial affordances, and instruction semantics. A distribution adjustment module further enforces physical contact consistency and semantic alignment. We evaluate AffordGrasp across four instruction-augmented benchmarks derived from HO-3D, OakInk, GRAB, and AffordPose, and observe substantial improvements over state-of-the-art methods in grasp quality, semantic accuracy, and diversity. Project Website: AffordGrasp

*Corresponding authors.

1. Introduction

Semantic-based grasp generation aims to synthesize human hand poses that interact with objects according to user instructions, enabling natural and intuitive interactions for AR/VR and robotic systems. However, traditional grasp generation approaches [9, 10, 16, 25, 28] rely solely on 3D object geometry and thus fail to reflect the user's intended interaction. For instance, grasping a teacup by its rim versus holding its handle requires distinct semantic intent despite identical geometry. This underscores the need to jointly model object shape, linguistic intent, and interaction context to generate meaningful and physically valid grasps.

Recent semantic grasping frameworks combine point-cloud embeddings with textual instructions to condition diffusion models [3, 12]. While effective to a degree, these models still struggle to produce high-precision grasps due to two key challenges: (1) the substantial modality gap between raw 3D geometry and natural language makes direct fusion insufficient for fine-grained geometric–semantic alignment (e.g, distinguishing "grasp the handle" from "hold the rim"); and (2) current diffusion pipelines lack explicit spatial and instruction-driven constraints, often yielding semantically incompatible contacts or physically unrealistic poses. Although VLM-based annotation pipelines [12, 19] seek to enhance semantic grounding through multi-turn question answering, these procedures remain susceptible to inconsistency and reduced controllability arising from error propagation across multiple reasoning steps, divergent reasoning paths, and context dependencies.

To address these challenges, we propose a method that directly generates fine-grained, structured language annotations for scalable and consistent dataset enrichment. Specifically, we automatically augment existing public hand–object interaction datasets with language labels that explicitly capture interaction intent. Building on this enriched data, we introduce *AffordGrasp*, an efficient cross-modal generative framework designed to synthesize diverse grasp poses that satisfy both physical constraints and textual semantic instructions.

Our approach leverages a latent diffusion model [26] augmented with an affordance-aware representation of hand poses within a compact latent space. The diffusion process employs a dual-conditioning mechanism that systematically integrates physical plausibility and semantic guidance, effectively modeling the conditional distribution of hand poses given object properties and instructional prompts. Concretely, *AffordGrasp* first generates local spatial cues aligned with instructions through an Affordance Generator and learns a low-dimensional latent representation of hand posture parameters using a Variational AutoEncoder. To ensure consistency with physical contact and semantic intent, we introduce a Distribution Adjustment Module that refines the latent representation during sampling based on object contact constraints and instruction semantics.

To validate our method, we construct four benchmarks based on HO-3D [6], OakInk [31], GRAB [25], and AffordPose [8], each enhanced with fine-grained textual instructions describing specific grasp poses and object interactions. Extensive experiments demonstrate that *AffordGrasp* significantly outperforms state-of-the-art baselines across all evaluation metrics, establishing a robust framework for advancing human grasp synthesis and embodied intelligence research.

In summary, our contributions are as follows:

- We introduce *AffordGrasp*, a diffusion-based framework that generates physically stable and semantically meaningful grasps with high precision, without requiring test-time adaptation.
- We propose the use of object affordance as complementary guidance for cross-modal fusion, bridging linguistic semantics and geometric representations to improve grasp intention understanding.
- We develop a distribution adjustment module that maintains diffusion sampling stability while enforcing strict physical and semantic constraints on grasp poses.
- Our method establishes a new state-of-the-art performance across multiple benchmarks through comprehensive quantitative and qualitative evaluations.

2. Related Work

Grasp Synthesis. Grasp synthesis is critical for robot manipulation, animation, and human motion analysis [19, 35]. We address the challenge of synthesizing realistic human grasps under two key constraints: semantic alignment with object functionality and physical plausibility. Existing methods typically predict MANO parameters [25, 28, 31] or joint positions [10] using generative models, while Liu *et al.*[3, 12] suffer from a modality gap between 3D geometry and language, weakening the alignment with semantic intent and reducing attention to critical affordances.To overcome this, we introduce object affordances as cross-modal cues that bridge geometry and language. By leveraging affordance-aware features, our architecture improves semantic grounding while preserving fine-grained hand-object geometric fidelity.

Affordance in Hand-Object Interaction. Understanding hand-object interaction is critical for applications in AR, VR, and embodied AI. While existing methods emphasize contact cues to ensure physical plausibility and detailed modeling [2, 5, 9, 18**?**], they often overlook object functionality and affordances. Recent efforts provide object affordance annotations [8, 31], but focus mainly on geometric properties, lacking interaction-centric semantic reasoning essential for task-oriented grasping. SemGrasp [12]

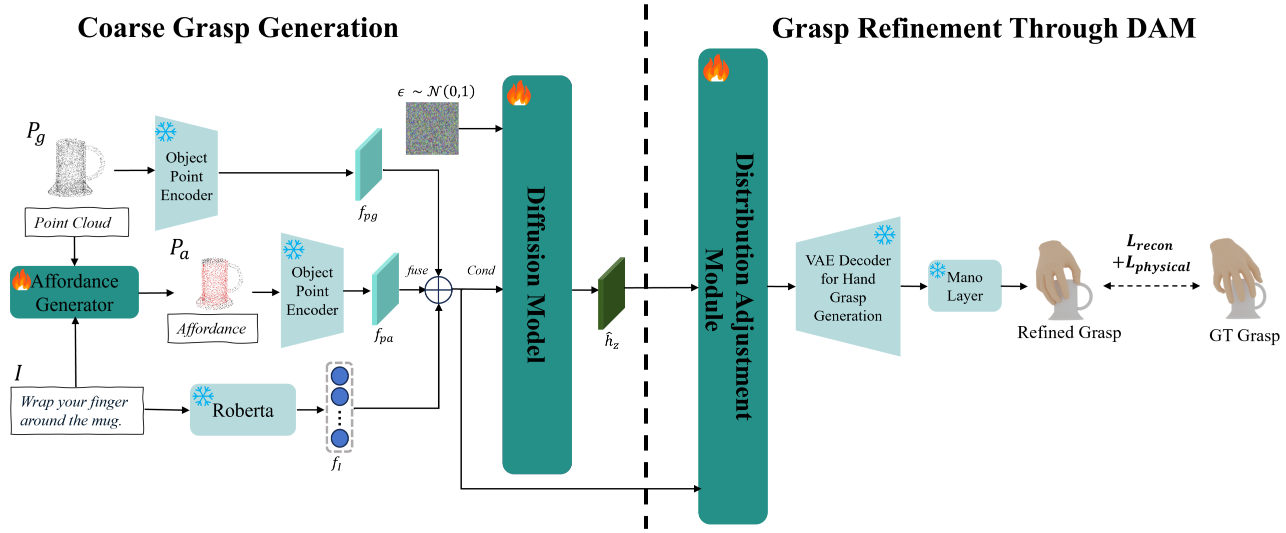

Figure 2. **Overview of AffordGrasp.** We integrate language instructions with object point clouds and employ the Affordance Generator to predict point-wise confidence features, which are aggregated into the final affordance map to enhance spatial detail and align linguistic semantics with 3D structures.The right part employs a DAM module, which ensures that the synthesized grasping poses generated by the LDM model align with physical constraints and language semantics.

leverages multi-modal language models to unify object, grasp, and language representations, but relies on 2D projections of 3D data, which introduces occlusions and limits instruction fidelity. In this work, we propose a 3D-native, interaction-aware framework that directly utilizes 3D spatial features to predict affordance categories. These features are integrated with a language model to generate precise, context-aware task instructions. Our method bridges spatial geometry and semantic understanding, enabling more accurate and robust task-driven grasp synthesis.

Denoising Diffusion Probabilistic Models. Denoising diffusion models [11, 15, 21, 23, 29] generate data by reversing a structured noise corruption process through a learned forward-backward stochastic framework. Recent methods incorporate gradient-based objectives during sampling to steer generation [27, 32], but often struggle under high-noise conditions. In particular, aggressive gradient updates in early diffusion steps can push samples off the learned data manifold, resulting in irreversible distribution shifts and degraded output quality.

3. Approach - AffordGrasp

We present AffordGrasp, a diffusion-based framework for generating human grasp poses that are both physically plausible and semantically aligned with user instructions. Given an object point cloud P_g and a textual instruction I, the goal is to generate a functional grasp pose h_p, represented by MANO parameters [22].

AffordGrasp explicitly models the interplay between instruction semantics, object geometry, and grasp intent through three integrated components: (1) an **Affordance Generator** that predicts interaction-relevant object regions from language–geometry inputs, (2) a **Cross-Modal Diffusion Model** that synthesizes grasp poses conditioned on multi-modal cues, and (3) a **Distribution Adjustment Module** (DAM) that refines the denoised latent representation to enhance contact realism and semantic precision.

3.1. Affordance Generator

Semantic grasp generation requires aligning the user instruction with geometric cues on the object. We therefore train an Affordance Generator that estimates point-wise affordance probabilities, indicating the relevance of each object point to the instruction. Given (P_g, I), the generator predicts an affordance map P_a, highlighting a semantically grounded local region in the object point cloud. This region serves as an intermediate representation that explicitly links language semantics to a geometric structure, reducing the cross-modal difficulty faced by prior methods.

Training the affordance generator is hindered by two main challenges: the lack of large-scale and diverse affordance datasets, and the severe imbalance between affording and non-affording object points. Among existing datasets, only AffordPose [8] provides affordance annotations, while datasets such as GRAB [25] and OakInk [31] lack such labels. This limitation reduces the geometric diversity of objects within the training data.

To address the first challenge, we generate pseudo-labels using a self-looping annotation engine. Specifically, we adopt the network architecture proposed by LASO [13] as our affordance generator and initially train it on the AffordPose [8] dataset. The trained model is then used in a self-training loop to annotate OakInk and GRAB, thereby expanding the dataset and enriching the geometric diversity of the training objects.

To address the class imbalance issue in affordance prediction, we employ a combined objective of Focal Loss [14] and Dice Loss [24] to optimize the model:

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_1 \mathcal{L}_{\text{dice}}, \tag{1}$$

where λ_1 is a balancing hyperparameter. Details of $\mathcal{L}_{\text{focal}}$ and $\mathcal{L}_{\text{dice}}$ are presented in the supplementary material Sec. 9.2.

3.2. Text and Affordance Guided Grasp Generation

Generating functional hand–object interactions requires jointly modeling semantic intent derived from language and geometric feasibility informed by object shape and affordance cues. To this end, we propose a cross-modal latent diffusion model conditioned on the triplet $\mathcal{C} = \{I, P_g, P_a\}$, which encodes the textual instruction, object geometry, and predicted affordance, respectively.

We first apply modality-specific encoders to extract feature representations from each input. Following LASO [13], the language instruction I is encoded using RoBERTa [17]. The object point cloud P_g and the affordance point cloud P_a are processed by two independent PointNet encoders(E_g, E_a) [20]. The resulting features are then fused to construct a unified conditioning vector $f = \{f_I, f_{pg}, f_{pa}\}$. The formal definition is:

$$f_I = RoBERTa(I), f_{pg} = E_g(P_g), f_{pa} = E_a(p_a) \tag{2}$$

To better preserve spatial structure in grasp poses, we encode the ground-truth hand mesh vertices $h_v^{gt} \in \mathbb{R}^{778\times 3}$ into a compact latent representation $h_z = \mathcal{E}(h_v^{gt})$ using a pre-trained autoencoder [28]. A conditional diffusion model [21] is then trained to learn the distribution of latent hand embeddings conditioned on f. The forward diffusion process is defined as:

$$z^t = \sqrt{\alpha_t} z^0 + \sqrt{1-\alpha_t}\epsilon, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and α_t is a predefined noise schedule. At each timestep t, a noise prediction network ϵ_θ estimates the additive Gaussian noise ϵ. The decoder maps the denoised latent representation to MANO parameters $h_p \in \mathbb{R}^{61}$, which are further passed through a differentiable MANO layer to reconstruct the hand mesh h_m. The overall training objective is defined as:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(h_v), \epsilon\sim\mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta(z^t, f, t)\|_2^2\right], \tag{4}$$

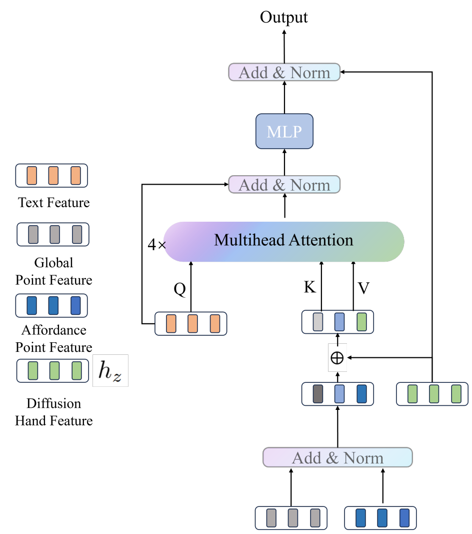

Figure 3. **Distribution Adjustment Module (DAM) Architecture.** Hand and object features are fused and aligned with language instructions to produce stable, instruction-consistent grasps.

where $\epsilon_\theta(z^t, f, t)$ denotes the conditional denoising U-Net, and z^t is the noisy latent obtained by perturbing the encoded hand representation h_z with Gaussian noise at timestep t. Through iterative denoising, the model learns to reconstruct realistic and semantically consistent hand meshes h_m from noise, guided by both linguistic and geometric cues.

3.3. Distribution Adjustment Module

To generate more accurate and physically plausible grasping poses, we introduce the Distribution Adjustment Module (DAM). As illustrated in Fig. 3, DAM is a lightweight fusion framework that refines the latent grasp representation $\hat{h}_z$ predicted by the diffusion module. It integrates the conditioning feature f with $\hat{h}_z$ to produce a refined latent $\tilde{h}_z$, ensuring the final pose aligns closely with both contextual and physical constraints.

According to Eq. 4, the diffusion model ϵ_θ predicts the noise term ϵ during training, which prevents its output from being directly used as input to the Distribution Adjustment Module (DAM). To address this issue, we propose a simple yet efficient approximation strategy that converts the noise prediction of the diffusion model into the corresponding latent hand pose representation. This process can be expressed as:

$$\hat{h}_z = \frac{1}{\sqrt{\alpha_t}}(z^t - \sqrt{1-\alpha_t}\epsilon_\theta(z^t, f, t)). \tag{5}$$

As shown in Fig. 3, we first construct the spatial representation f_{spatial} by fusing the global point feature (f_{pg}) and the affordance point feature (f_{pa}) with the diffusion hand feature $\hat{h}_z$ (from Eq. 5). To effectively balance geometric detail and task intent, this representation interacts with the instruction embedding f_I through a multi-head attention (MHA) module. This process employs a dual residual mechanism: the first residual connection preserves the instruction semantics (f_I) after the MHA, and the second preserves the original hand representation ($\hat{h}_z$) after the MLP module. This design contributes to improved performance while enhancing the network's expressive capacity. Unlike training-free methods [27, 32] that increase inference time, our DAM is a lightweight, single-pass refinement module applied post-sampling, resulting in minimal inference overhead. The process is formulated as:

$$\begin{aligned} f_{\text{spatial}} &= \text{Norm}(f_{pg} + f_{pa}) + \hat{h}_z, \\ f_{\text{align}} &= \text{Attention}(f_I, f_{\text{spatial}}, f_{\text{spatial}}) + f_I, \\ \tilde{h}_z &= \text{Norm}(\text{MLP}(f_{\text{align}}) + \hat{h}_z), \end{aligned} \tag{6}$$

where f_{pg}, f_{pa}, and f_I denote the embeddings extracted from P_g, P_a, and I using pretrained models. According to Eq. 6, the proposed DAM module integrates the latent hand feature with both the instruction and the object geometry, thereby improving physical feasibility and instruction adherence.

During DAM training, the diffusion model ϵ_θ remains frozen. The predicted latent representation $\hat{h}_z$ from Eq. 5 is used as the input to the DAM module. Since $\hat{h}_z$ may contain inaccuracies, the DAM further refines it to produce a more reliable grasp latent.

$$h_m = \text{MANO}(\text{Decoder}(\tilde{h}_z)), \quad \tilde{h}_z = \text{DAM}(\hat{h}_z, f). \tag{7}$$

The overall training objective, detailed in Eq. 8, combines reconstruction and physical constraint losses following prior works [9, 28].

$$\mathcal{L} = \lambda_2 \mathcal{L}_{\text{recon}}(h_v, h_p, h_v^{\text{gt}}, h_p^{\text{gt}}) + \mathcal{L}_{\text{physical}}(h_m, h_m^{\text{gt}}, P_g), \tag{8}$$

Here, λ_2 is a balancing coefficient. The Decoder reconstructs the refined latent code $\tilde{h}_z$ into hand parameters h_p, which are then passed through the MANO layer to produce the hand mesh h_m. The loss $\mathcal{L}_{\text{recon}}$ supervises the reconstruction of the hand pose by encouraging the predicted MANO parameters, vertices to align with their corresponding ground truth. In contrast, $\mathcal{L}$physical enforces physically plausible hand–object interactions by penalizing object–object interpenetration and promoting consistent contact patterns.

By jointly optimizing reconstruction accuracy and physical constraints, the DAM learns to refine the diffusion model's latent predictions, yielding final grasp poses h_m that are both task-compliant and physically plausible. Additional implementation details are provided in Appendix Sec. 11.

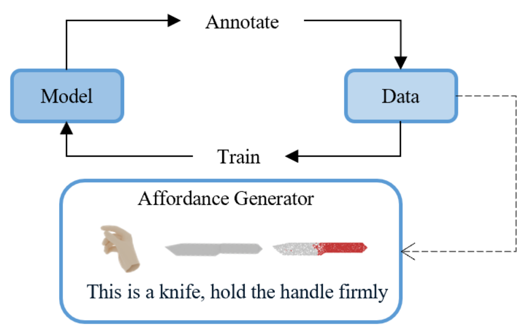

Figure 4. **Affordance Annotation.** Implement an automated self-training pipeline that first assigns pseudo-labels to unlabeled data, then iteratively optimizes the model using these refined annotations.

3.4. Inference

Our inference process commences by sampling noise $\epsilon \sim \mathcal{N}(0, 1)$ and condition it on the f, which guides the diffusion model to synthesize coarse grasp poses in the latent space. To improve sampling efficiency, we adopt the DDIM framework [23], governed by the following update rule:

$$\begin{aligned} \boldsymbol{z}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\boldsymbol{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{z}_t, f)}{\sqrt{\alpha_t}} \right) \\ + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(\boldsymbol{z}_t, f) + \sigma_t \epsilon, \end{aligned} \tag{9}$$

where α_t represents the DDIM scheduling parameters, and ϵ_θ corresponds to the conditional denoising U-Net [26].

To further strengthen the model's conditional dependence on both physical constraints and linguistic semantics, we refine the generated sample using DAM. The refinement is defined as:

$$\tilde{z}_1 = \text{DAM}(z_1, f), \tag{10}$$

where $\tilde{z}_1$ represents the optimized hand representation in the latent space, and z_1 denotes the DDIM output. The refined latent vector is then decoded into the hand mesh h_m through a decoder and MANO layer. This process can be formally expressed as follows:

$$h_m = \text{MANO}(\text{Decoder}(\tilde{z}_1)). \tag{11}$$

Our framework generates hand mesh h_m satisfies both linguistic plausibility and physical feasibility. In contrast to prior approaches that fuse the instruction and object point cloud to generate the final pose, our method mitigates cross-modal challenges. By narrowing the gap between language and geometric modalities, it produces hand poses that are more physically plausible and semantically aligned with the instructions.

Dataset	Method	Penetration Volume ↓	Simulation Displacement ↓	Contact Ratio ↑	Entropy ↑	Cluster Size ↑	ACC ↑
OakInk [31]	TTA [9]	8.21	2.33	97	2.82	2.81	60.83%
	FastGrasp [28]	7.88	2.27	88	2.88	3.42	78.05%
	D-VQVAE [37]	7.33	2.44	91	2.88	3.57	76.4%
	Ours(ControlNet)	8.08	3.11	81	2.87	3.44	76.94%
	Ours	**7.31**	**1.43**	**98**	**2.94**	**3.74**	**80.08%**
GRAB [25]	TTA [9]	6.51	1.22	91	2.73	1.41	55.00%
	FastGrasp [28]	4.61	**1.20**	94	2.76	1.96	61.50%
	D-VQVAE [37]	8.04	1.82	89	2.88	3.41	57.50%
	Ours(ControlNet)	6.78	1.57	93	2.76	3.47	60.50%
	Ours	**3.06**	**1.66**	**94**	**2.91**	**3.53**	**62.50%**

Table 1. Quantitative comparison on the **OakInk** and **GRAB** dataset. We compare our results with baselines as well as with a framework where the DAM module is replaced by the ControNet [36] structure. Our method achieves the best performance on all evaluation metrics.

Dataset	Method	Penetration Volume ↓	Simulation Displacement ↓	Contact Ratio ↑	Entropy ↑	Cluster Size ↑	ACC ↑
HO-3D [6]	TTA [9]	12.55	3.22	95	2.55	2.99	66%
	FastGrasp [28]	14.45	2.73	96	2.81	2.23	52.00%
	D-VQVAE [37]	13.12	**2.33**	95	2.78	3.63	64.00%
	Ours(ControlNet)	16.06	2.38	97	2.84	3.52	51.00%
	Ours	**7.38**	**2.33**	**97**	**2.85**	**3.70**	**72.00%**
AffordPose [8]	TTA [9]	19.41	4.32	91	2.89	2.97	42.56%
	FastGrasp [28]	22.75	3.77	88	2.83	3.77	54.08%
	D-VQVAE [37]	21.43	4.18	91	2.91	3.64	63.99%
	Ours(ControlNet)	24.77	4.78	88	2.88	3.55	52.45%
	Ours	**10.36**	**3.59**	**91**	**2.92**	**3.93**	**69.71%**

Table 2. Comparison with previous methods on the **HO-3D** and **AffordPose** dataset, where our model is trained on the GRAB [25] dataset. Our model achieves state-of-the-art performance on two out-of-domain dataset, setting new benchmarks.

4. Experiment

4.1. Automated Labeling for Dataset Enrichment

To enrich the OakInk [31] and GRAB [25] datasets—both centered on hand-object interactions—we introduce an automated pipeline for generating instruction annotations. Starting with the AffordPose [8] dataset, we perform cold-start training of a classifier [33] to produce initial language labels, which are iteratively refined via error analysis to improve consistency. These refined annotations, combined with the initial dataset, support multiple rounds of training and labeling for full annotation coverage. Finally, we incorporate large language models [30] to generate task-oriented, step-by-step instructional text, further enhancing semantic richness. Implementation details are in the Appendix.

We evaluate our method on four benchmarks: OakInk [31], GRAB [25], HO-3D [6], and AffordPose [8], following standard experimental protocols. For in-domain evaluation, we train and test on OakInk and GRAB; the latter includes 51 objects grasped by 10 subjects, while OakInk offers a larger-scale dataset with 1,700 objects manipulated by 12 subjects. For cross-dataset generalization, we evaluate on HO-3D and the out-of-domain object split of AffordPose under zero-shot settings, consistent with prior work [9, 25, 28, 36]. We exclude AffordPose from training due to its missing MANO parameters and partial noise, which are incompatible with our MANO-based differentiable model. Quantitative analysis justifying this exclusion is provided in the Appendix.

4.2. Evaluation Metrics

Following established evaluation protocols [10, 25, 28, 36], we assess generated grasping poses using four criteria: (1) physical plausibility, (2) stability, (3) pose diversity, and (4) semantic alignment with language specifications.

Physical Plausibility Assessment. We evaluate physical plausibility through two metrics: (1) hand-object mutual penetration volume calculated by voxelizing both meshes at $1mm^3$ resolution and measuring intersection regions, and (2) contact ratio, which measures the percentage of grasp poses maintaining persistent surface contact.

Grasp Stability Assessment. Stability evaluation follows prior physics-based approaches [9, 25, 28] through simulated grasp executions. We quantify stability by measuring the gravitational displacement of the object's center of mass, with lower displacement indicating greater robustness.

Diversity Assessment. Following diversity metrics from

Dataset	Method	Penetration Volume ↓	Simulation Displacement ↓	Contact Ratio ↑	Entropy ↑	Cluster Size ↑	ACC ↑
OakInk [31]	w/o object affordance	8.27	**1.22**	97	2.88	3.81	76.56%
	w/o DAM	8.12	1.77	97	2.89	**4.07**	79.11%
	Whole pipeline	**7.31**	1.43	**98**	**2.94**	3.74	**80.08%**
GRAB [25]	w/o object affordance	4.32	**1.55**	92	2.79	3.61	55.00%
	w/o DAM	4.91	1.71	90	2.84	**3.77**	**63.00%**
	Whole pipeline	**3.06**	1.66	**94**	**2.91**	3.53	62.50%
HO-3D [6]	w/o object affordance	8.88	**2.29**	95	**2.85**	**3.71**	71.00%
	w/o DAM	11.21	2.44	91	2.79	3.64	69.00%
	Whole pipeline	**7.38**	2.33	**97**	**2.85**	3.70	**72.00%**
AffordPose [8]	w/o object affordance	11.29	3.78	88	2.84	3.87	67.81%
	w/o DAM	12.32	3.64	**91**	2.88	**3.94**	70.21%
	Whole pipeline	**10.36**	**3.59**	**91**	**2.92**	3.83	**72.43%**

Table 3. Ablation study results on the **GRAB, OakInk, HO-3D, AffordPose** datasets [6, 8, 25, 31]. The evaluation of the HO-3D and AffordPose is an out-of-domain generalization test, where the model is trained using the GRAB dataset.

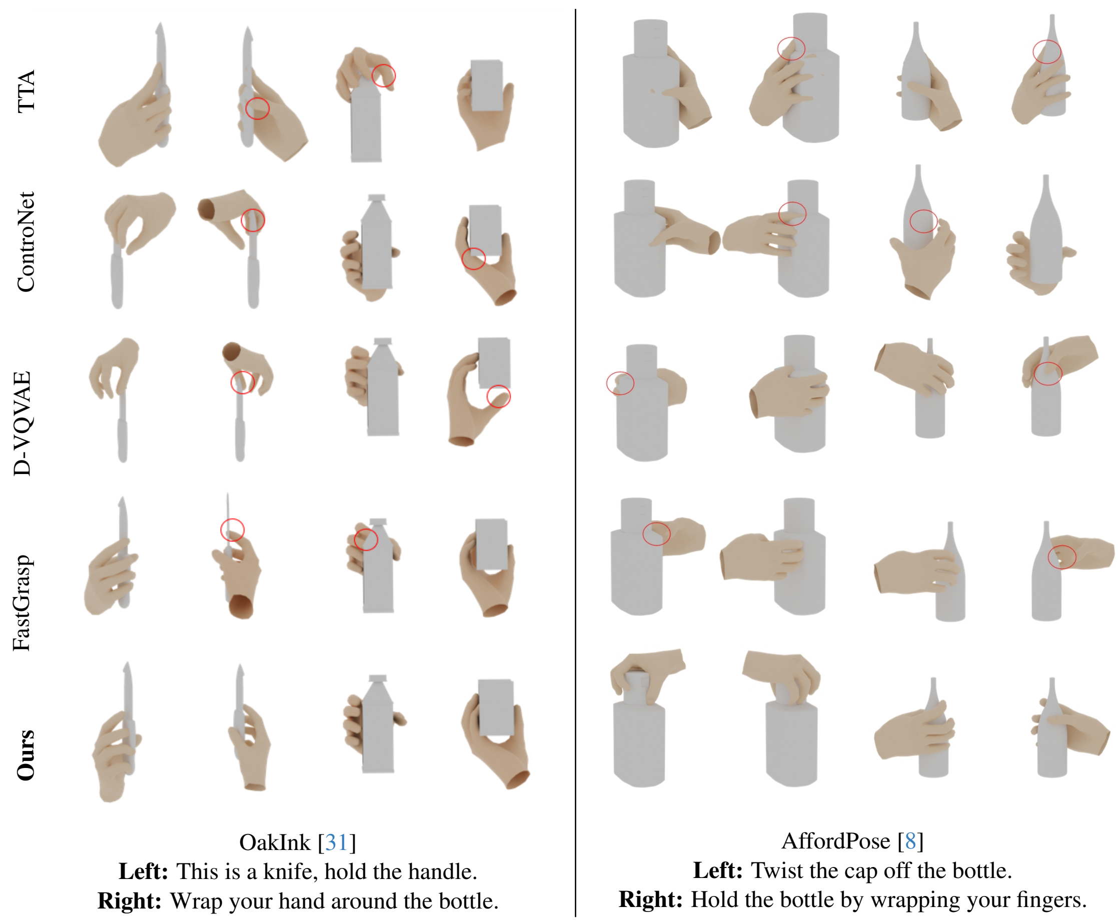

Figure 5. Qualitative comparisons with state-of-the-art methods on GRAB, OakInk datasets. Each pair (two columns) visualizes the generated grasps from two different views. Our method demonstrates a significant reduction in object penetration compared to other methods.

grasp generation literature [10, 16, 28], we apply K-means clustering (k=20) to 3D hand keypoints across all methods. Diversity is assessed via two measures: (1) entropy of cluster assignments, where higher values indicate more diverse distributions across clusters, and (2) average cluster size, reflecting grasp space coverage. While larger average cluster sizes reflect better coverage of the grasp space.

Semantic Accuracy Assessment(ACC). To evaluate semantic consistency, we categorize hand-object interactions into ten affordance classes like AffordPose [8]. A classi-

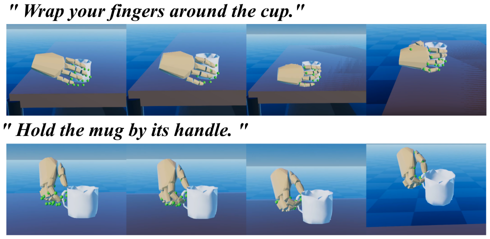

Figure 6. **Simulation environment: grasping a single object under different instructions.**

fier trained during the data annotation phase is used to assess whether the generated hand poses align with the corresponding semantic affordance instructions. To validate the effectiveness of our classifier, we report its accuracy on unseen objects in the appendix.

4.3. Grasp Generation Performance

In-Domain Evaluation. Our approach, which combines DAM and hierarchical spatial fusion, outperforms all competitors in the in-domain evaluation (see Tab. 1 and Fig. 5). It shows superior performance on the OakInk [31] and GRAB [25] datasets across all metrics: intrusion volume, simulation distance, grasp pose diversity, and semantic accuracy. These results highlight the effectiveness of our method in generating complex grasp poses, surpassing leading methods such as FastGrasp [28], D-VQVAE [37], and ControlNet [36].
Out-of-Domain Evaluation. We further evaluated our model's universal applicability on the HO-3D [6] and AffordPose [8] datasets. As depicted in Tab. 2 and Fig. 5, out-of-domain evaluation validates our method's outstandingly accurate semantic results, in addition to physical generalization and generation diversity. Our method hence emerges as an effective solution that can transcend domain boundaries.

4.4. Ablation Study

We conduct a comprehensive ablation study to evaluate the impact of individual components on our framework's performance. This analysis provides empirical evidence of each component's contribution, offering crucial context for subsequent experiments.

Tab. 3 summarizes the key findings. Excluding object affordances results in a slight improvement in displacement distance but increases volume intrusion. This suggests that the model depends on object affordances to better capture spatial relationships and minimize hand-object collisions. Improved displacement occurs when object invasion causes immobilization, reducing movement. Thus, incorporating object affordances as cross-modal cues enhances the model's ability to capture spatial details of objects.

Removing the DAM module increases cluster sizes, as the diffusion model better captures global distributions, weakening the conditioning. In contrast, incorporating the DAM module results in a more concentrated output distribution, improving thematic coherence, adherence to physical constraints, and object contact rates. This novel use of the adaptive module transformation in Out-of-Distribution scenarios highlights robust generalization capabilities.

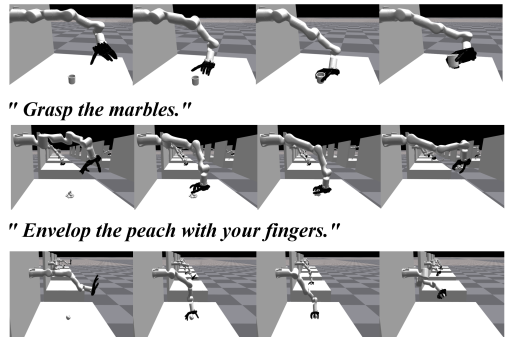

Figure 7. **Simulation environment: grasping across multiple objects.**

4.5. Downstream Applications

To evaluate the downstream effectiveness of our generated grasps, we conducted experiments in the RaiSim [7] physics simulator. We employed both the D-Grasp [4] framework and the CrossDex [34] platform to execute dynamic grasp trajectories.

In our framework, AffordGrasp predicts a reference grasp pose $\bar{G}$ conditioned on the target object O and the language instruction L. In D-Grasp, this reference is used to generate physically plausible and temporally consistent grasp motions. In CrossDex, the same reference grasp serves as a target goal configuration, guiding the policy to achieve it in simulation. During evaluation, each grasped object is lifted against gravity to assess stability. Fig. 6 and Fig. 7 illustrate grasps of the same object under different instructions, as well as grasps of multiple objects.

5. Conclusion and Discussion

This paper presents an automated annotation engine that enriches hand-object interaction datasets with linguistic instructions. By leveraging object affordance for cross-modal alignment and introducing a Distribution Adjustment Module, our method captures both spatial geometry and instruction semantics. It enhances physical plausibility through geometric grounding and improves alignment between language and 3D representation.

Limitations. Our current framework primarily focuses on data-driven learning and does not explicitly incorporate physical priors such as gravity or friction. As illustrated in appendix Fig. 16, certain real-world effects may not be fully reflected in the generated results. Future work could benefit from integrating physics-based reasoning or simulation to further enhance grasp stability and realism.

References

[1] Daich Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19107–19117, 2021. 3

[2] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019. 2

[3] Xiaoyun Chang and Yi Sun. Text2grasp: Grasp synthesis by text prompts of object grasping parts. *ArXiv*, abs/2404.15189, 2024. 2

[4] Sammy Christen et al. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *CVPR*, 2022. 8

[5] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5041, 2020. 2

[6] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6, 7, 8, 4

[7] Jemin Hwangbo, Joonho Lee, and Marco Hutter. Per-contact iteration method for solving contact dynamics. *IEEE Robotics and Automation Letters*, 3(2):895–902, 2018. 8

[8] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14713–14724, 2023. 2, 3, 4, 6, 7, 8, 1

[9] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the International Conference on Computer Vision*, 2021. 2, 5, 6, 1, 4

[10] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *2021 International Conference on 3D Vision (3DV)*, pages 11–21. IEEE, 2021. 2, 6, 7

[11] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *ArXiv*, abs/2210.10960, 2022. 3

[12] Kailin Li, Jingbo Wang, Lixin Yang, Cewu Lu, and Bo Dai. Semgrasp: Semantic grasp generation via language aligned discretization. *ArXiv*, abs/2404.03590, 2024. 2

[13] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14251–14260, 2024. 4, 3

[14] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2017. 4

[15] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. *ArXiv*, abs/2206.01714, 2022. 3

[16] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 7, 1

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. 4

[18] Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. *ArXiv*, abs/2411.14280, 2024. 2

[19] Yumeng Liu, Yaxun Yang, Youzhuo Wang, Xiaofei Wu, Jiamin Wang, Yichen Yao, Sören Schwertfeger, Sibei Yang, Wenping Wang, Jingyi Yu, Xuming He, and Yuexin Ma. Realdex: Towards human-like grasping for robotic dexterous hand. *ArXiv*, abs/2402.13853, 2024. 2

[20] C. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2016. 4

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3, 4

[22] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 3

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. 3, 5

[24] Carole Helene Sudre, Wenqi Li, Tom Kamiel Magda Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep learning in medical image analysis and multimodal learning for clinical decision support : Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, held in conjunction with MICCAI 2017 Quebec City, QC,...*, 2017:240–248, 2017. 4

[25] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 6, 7, 8, 1, 4

[26] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 2, 5

[27] Shijie Wu, Yihang Zhu, Yunao Huang, Kaizhen Zhu, Jiayuan Gu, Jingyi Yu, Ye Shi, and Jingya Wang. Afforddp: Generalizable diffusion policy with transferable affordance. *ArXiv*, abs/2412.03142, 2024. 3, 5, 1

[28] Xiaofei Wu, Tao Liu, Caoji Li, Yuexin Ma, Yujiao Shi, and Xuming He. Fastgrasp: Efficient grasp synthesis with diffusion. *ArXiv*, abs/2411.14786, 2024. 2, 4, 5, 6, 7, 8, 1

[29] Xiaofei Wu, Guozhen Zhang, Zhiyong Xu, Yuan Zhou, Qinglin Lu, and Xuming He. Pack and force your memory: Long-form and consistent video generation. *arXiv preprint arXiv:2510.01784*, 2025. 3

[30] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 6

[31] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 6, 7, 8, 1, 4

[32] Lingxiao Yang, Shutong Ding, Yifan Cai, Jingyi Yu, Jingya Wang, and Ye Shi. Guidance with spherical gaussian constraint for conditional diffusion. *ArXiv*, abs/2402.03201, 2024. 3, 5, 1

[33] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 1

[34] Haoqi Yuan, Bohan Zhou, Yuhui Fu, and Zongqing Lu. Cross-embodiment dexterous grasping with reinforcement learning. *arXiv preprint arXiv:2410.02479*, 2024. 8, 7

[35] Hui Zhang, Sammy Joe Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. *ArXiv*, abs/2403.19649, 2024. 2

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. 6, 8

[37] Zhe Zhao, Mengshi Qi, and Huadong Ma. Decomposed vector-quantized variational autoencoder for human grasp generation. In *European Conference on Computer Vision*, 2024. 6, 8, 1