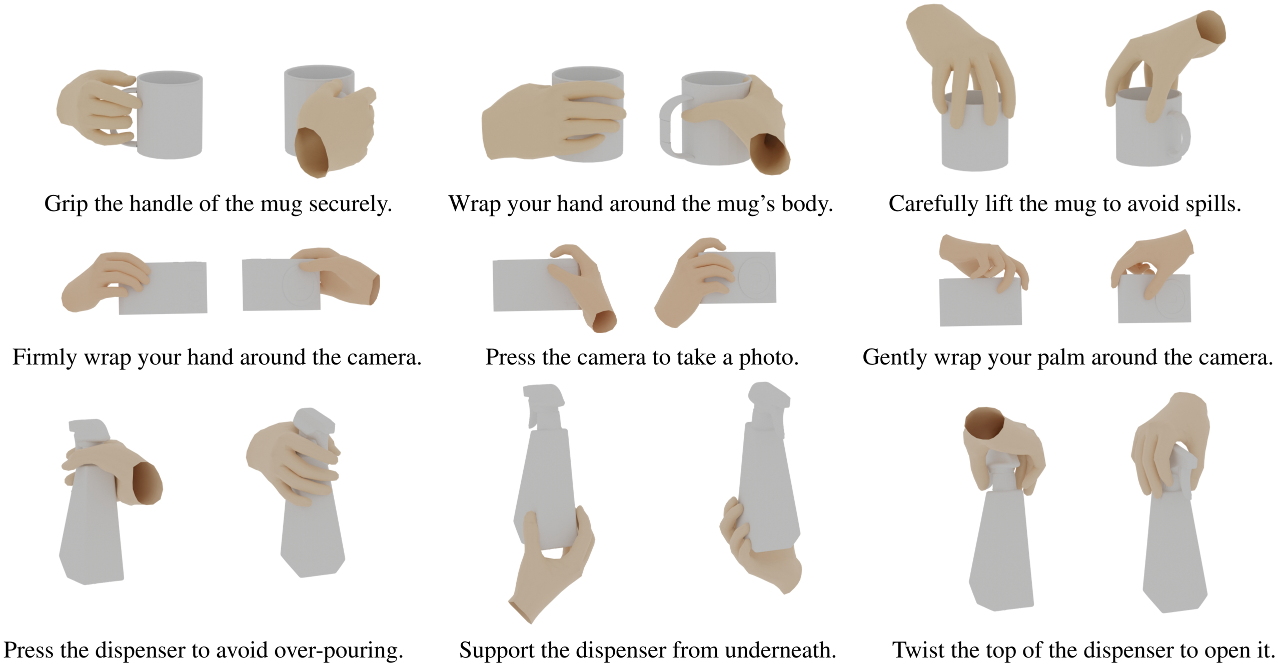
AffordGrasp: Cross-Modal Diffusion for Affordance-Aware Grasp Synthesis

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



Grip the handle of the mug securely.

Wrap your hand around the mug's body.

Carefully lift the mug to avoid spills.

Firmly wrap your hand around the camera.

Press the camera to take a photo.

Gently wrap your palm around the camera.

Press the dispenser to avoid over-pouring.

Support the dispenser from underneath.

Twist the top of the dispenser to open it.

Figure 1. **AffordGrasp** enables realistic dexterous hand grasping synchronized with textual instructions, with each grasp pair visualized from two distinct viewpoints to emphasize spatial details. For each object, three selected instructions guide the generation of diverse grasps.

Abstract

Accurately modeling human hand-object interactions remains challenging due to the complexity of aligning instructions with geometric features in applications. Previous approaches often leverage 3D geometric features and textual instructions to directly capture hand-object relationships. However, inherent modality gaps and the difficulty of learning fine-grained geometric constraints and instruction-semantic relationships, such strategies often result in suboptimal hand pose. To address this limitation, this work introduces a novel diffusion-based approach for generating grasping poses that adhere to both physical constraints and instruction semantics. In particular, we develop a latent diffusion model augmented with an object affordance module and a distribution adjustment module, conditioned on both objects and instructions. This enhancement improves the model's ability to capture fine-grained semantic instructions and geometric features effectively. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches in instruction semantic consistency, diversity, and pose quality.

1. Introduction

Semantic-based grasping generation, which aims to produce hand poses that interact with an object according to user instructions, plays an essential role for achieving natural and intuitive interactions in Augmented or Virtual Reality (AR/VR) and robotic systems. However, traditional grasp generation strategies [11, 12, 17, 27, 30] primarily rely on 3D geometric features extracted from objects and, as a result, are often unable to satisfy the semantic require-

[*] Corresponding authors.

ments specified in user instructions. Consider a teacup as an example: the grasping strategy for gently lifting the cup by its rim differs significantly from firmly gripping its handle, even though both interactions involve the same geometric structure. This highlights the necessity of jointly modeling object geometry and instruction semantics to infer the user intent and interaction context, enabling grasp generation tailored to real-world interaction needs.

Previous semantic-based grasping generation methods typically integrate object geometry with textual instructions as a condition to a diffusion model of hand poses [3, 14]. Despite their promising results, these approaches still struggle to achieve high precision in generating hand poses due to two main challenges: 1) The large modality gap between 3D object representations and textual instructions makes direct fusion uninformative for grasp synthesis, particularly in interactions requiring fine-grained geometric-semantic alignment (e.g., gripping a cup handle versus lifting its rim); 2) The lack of effective spatial and instruction-driven constraints in diffusion models often results in physically unrealistic or semantically incompatible grasping poses. While recent training-free techniques [8, 29, 34] partially mitigate this limitation through gradient-guided iterative optimization, they typically incur significant computational overhead, drastically reducing real-time performance.

To address these limitations, we propose an efficient cross-modal generation method, named *AffordGrasp*, to synthesize diverse grasping poses while satisfying physical constraints and semantic instructions. To accomplish this, we employ a latent diffusion model framework [28] to develop an affordance-aware representation of hand poses within a latent space and a diffusion-based generation process systematically integrates physical constraints with semantic guidance through a dual-conditioning mechanism, effectively modeling the probability distribution of hand poses conditioned on both object properties and instructional prompts.

Specifically, *AffordGrasp* first generates local spatial information aligned with instructions via an Affordance Generator and learns a low-dimensional representation of hand posture parameters based on an AutoEncoder (AE) network. It then encodes the object and affordance information with PointNet [23] and the instructions with an LLM, constructing a diffusion model in the latent space conditioned on object, affordance and instruction representations. To integrate physical constraints and semantic guidance for hand-object interaction, we introduce a distribution adjustment module, which refines the latent representation based on object contact information and instruction semantics. Finally, the generated hand-pose representation is decoded by the AE decoder into MANO [25] parameters.

To validate our method, we build four benchmark datasets based on HO-3D [9], OakInk [33], GRAB [27], and AffordPose [10], where we generate fine-grained textual instructions tailored to specific grasp poses and objects. Our experiments demonstrate that AffordGrasp significantly outperforms state-of-the-art methods across all evaluation metrics, establishing a robust framework for advancing human grasp generation and embodied intelligence research. In summary, our contributions are as follows:

- We introduce AffordGrasp, a diffusion-based framework that simultaneously generates physically stable and semantically meaningful grasps with high precision, eliminating test-time adaptation requirements.
- We propose object affordance as complementary guidance for cross-modal fusion, bridging language semantics and geometric representations to enhance grasp intention understanding.
- We develop a distribution adjustment module that maintains diffusion sampling stability while producing grasp poses satisfying strict physical constraints and semantic alignment.
- Our method establishes new state-of-the-art performance across all benchmarks through comprehensive quantitative and qualitative evaluations.

2. Related Work

2.1. Grasp Synthesis

Grasp synthesis is critical for robot manipulation, animation, and human motion analysis [20, 37]. We address authentic human grasp synthesis under dual constraints: semantic alignment with object function and physical plausibility. Current approaches predominantly predict MANO parameters [19, 27, 30, 33] or joint positions [12] via generative models, while Liu *et al*. [17] proposes a two-stage pipeline merging learned representations with iterative optimization. A key limitation persists: methods relying on object point clouds [3, 14] create a modality gap that impedes alignment with language semantics, hindering attention to crucial affordances. We bridge this gap by introducing object affordance as cross-modal compensation, enabling synergistic fusion of geometric and linguistic features. Our architecture enhances semantic precision in grasping gestures while modeling fine-grained geometric relationships with objects.

2.2. Affordance Learning

With advances in embodied AI, affordance [4, 21, 22] learning is expanding into 3D domains [1, 2, 5, 6]. 3D AffordanceNet pioneers the first benchmark dataset for learning affordances from point cloud geometry [7]. Yang *et al*. [20, 35] introduce image-guided learning of 3D affordance parts, these approaches rely exclusively on visual cues - tightly coupling geometric features with affordance labels while overlooking semantic dimensions. This visual-

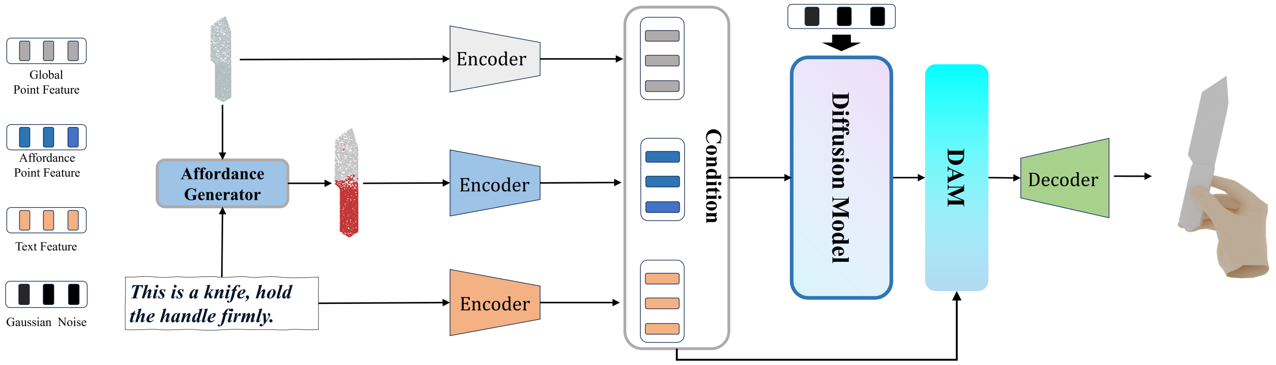


Figure 2. **Overview of AffordGrasp.** We use language and object point clouds to generate object affordance, enhancing the model's ability to capture local spatial details and learn the relationship between language semantics and space. The right part employs an LDM model for hand representation generation, while the DAM module ensures that the synthesized grasping poses align with physical constraints and language semantics.

only paradigm impedes integration with Large Language Models (LLMs) [31, 32] that could otherwise ground affordance reasoning in real-world deployment contexts. Furthermore, the common paradigm of treating affordance prediction as an auxiliary task [15] diverts focus from core affordance reasoning, potentially hindering structured affordance representation learning.

2.3. Denoising Diffusion Probabilistic Models

Denoising diffusion models [13, 16, 24, 26] learn data synthesis through a forward-backward stochastic process, where generation emerges from the iterative reversal of a structured noise corruption process. Recent advances integrate gradient-based objectives during sampling to steer denoising trajectories as exemplified by Wu *et al.* [29, 34]. These methods face inherent limitations when operating under high noise conditions. Specifically, aggressive gradient optimization at early diffusion steps risks displacing samples from the model's learned data manifold, leading to irreversible distribution shifts that degrade output quality.

Our work resolves this tension through a principled balance of stability and controllability. Rather than modifying core sampling dynamics. A strategy prone to error propagation in gradient-perturbed approaches. We preserve the intrinsic equilibrium of the diffusion process. Control is instead achieved through targeted enhancements to the conditional generation mechanism, ensuring guidance remains geometrically aligned with the underlying data manifold. This paradigm not only mitigates fidelity loss but demonstrates that sampling stability and precise controllability need not be competing objectives in diffusion-based synthesis.

3. Affordance-Guided Grasp Generation

Generating a natural and functionally efficient grasp pose from an object's shape and human-provided instructions is challenging, as it must balance ergonomic principles with precise linguistic constraints. To address this, we propose a framework that synthesizes physically plausible human hand poses while ensuring alignment with the given instructions. Our framework consists of two core components: (1) **Cross-Modal Diffusion Module**, a latent diffusion model that utilizes Cross-modal conditions to generate diverse grasp poses, and (2) **Distribution Adjustment Module (DAM)**, which refines the generated grasps by optimizing their alignment with both linguistic instructions and physical feasibility.

3.1. Model Architecture

Cross-Modal Diffusion Module. We develop a latent diffusion model based on the following conditioning set:

$$\mathcal{C} = \{I, P_g, P_a\}, \tag{1}$$

where I represents the instruction in natural language, P_g denotes the object's full point cloud, and P_a corresponds to the partial point cloud of the object's affordance region.

To process these inputs, we employ distinct encoders: RoBERTa [18] for text, following the LASO [15] framework, and separate PointNet [23] encoders without shared weights for P_g and P_a. These encoders extract meaningful cross-modal features that serve as conditioning inputs for the diffusion model. Additionally, for the input hand pose, we utilize FastGrasp's [30] autoencoder to extract its feature representation.

The extracted features f from $\mathcal{C}$ are then fed into the latent diffusion model, which progressively refines a latent grasp representation through iterative Gaussian denoising. Using DDIM sampling [26], the model approximates the conditional distribution $p(h_z|I, P_g, P_a)$, ensuring that the generated grasp poses align with the provided instructions and object affordances. The generated latent code h_z of

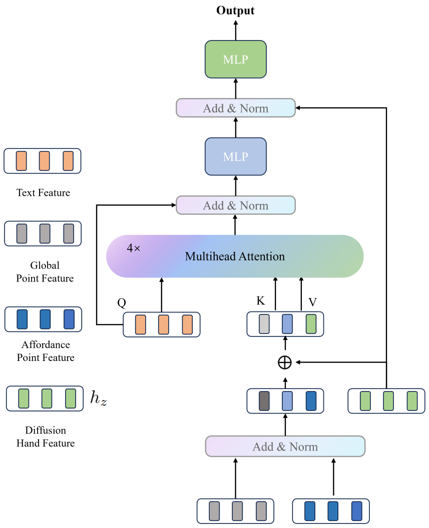


Figure 3. **Distribution Adjustment Module (DAM) Architecture.** We fuse hand pose embeddings with multi-scale object features to capture both local details and global context. The combined representation is aligned with language instructions through cross-attention, ensuring grasping actions correspond to given commands. A dual residual learning framework maintains semantic consistency with linguistic inputs while preserving the physical stability essential for grasp mechanics.

grasp poses then serves as inputs for downstream grasp execution and evaluation modules.

Distribution Adjustment Module. DAM is a fusion framework in Fig. 3 that integrates the conditioning set $\mathcal{C}$ and the latent code of the hand pose h_z as inputs to generate a refined latent representation $\hat{h}_z$. Its primary function is to adjust the distribution of the encoded hand pose based on the conditioning set, ensuring better alignment with the given context and physical constraints.

In DAM, we systematically align hand representations in latent space with both the full object and affordance point clouds to construct spatial features. By fusing these features with instruction semantics, our model effectively balances semantic and geometric information, while a dual residual mechanism maximizes the retention of both semantic context and hand representation, leading to superior performance. In contrast to training-free methods [29, 34], which increase inference time, our approach preserves diffusion sampling stability by maintaining the framework's structural integrity, ensures inference efficiency through a non-invasive adaptation strategy, and enhances grasp generation quality by aligning features in a distribution-aware manner to improve both physical plausibility and semantic consistency.

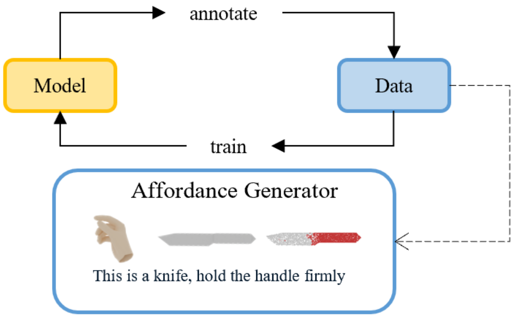


Figure 4. **Affordance Annotation.** Implement an automated self-training pipeline that first assigns pseudo-labels to unlabeled data, then iteratively optimizes the model using these refined annotations.

3.2. Model Training

Affordance Generator Training. To better align instruction semantics with geometric representations, we first train the Affordance Generator. This model takes both objects and instructions as inputs and estimates the probability that each point on the object's surface corresponds to a specific affordance, reflecting the relevance of that point to the given instruction. We then use the affordance map to extract a semantically aligned local geometric representation P_a from the object's global geometry. To address class imbalance in affordance prediction during training, we optimize the model using a combined objective of focal loss and Dice loss as follow:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{focal} + \mathcal{L}_{dice} \tag{2}$$

where λ_1 is balancing coefficients. Due to the absence of affordance labels, we implemented a self-looping annotation process, automatically generating pseudo-labels to obtain object affordance, as shown in Fig. 4. Implementation details are provided in the Appendix.

Latent Condition Diffusion Training. Our autoencoder takes a hand mesh vertex $h_v^{gt} \in \mathbb{R}^{778\times 3}$ as input, instead of directly using MANO [25] parameters, to better preserve the spatial shape characteristics of grasping poses, thereby enhancing both generalization and spatial detail retention. The decoder then maps the latent representation to MANO parameters $h_p \in \mathbb{R}^{61}$, which are passed through a differentiable MANO layer [25] to reconstruct the hand mesh h_m. The model is optimized with the composite loss function defined in Eq. 6.

We employ the diffusion model [24] for learning the distribution of latent hand representation generated by the autoencoder. Following standard diffusion methodology, we train a noise prediction network to anticipate the noise ϵ added in the t-th iteration. The training objective is formulated as:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(h_v),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z^t, f, t)\|_2^2\right] \tag{3}$$

where $\epsilon_\theta(z^t, f, t)$ denotes the conditional denoising U-Net used for training, where t ranges from 1 to T, the input z^t is the h_z mixed with ϵ_t, the f denotes the condition feature, the h_z is encoder output. Through training, the diffusion model learns to reconstruct the hand mesh h_m from Gaussian noise by denoising and decoding.

DAM Training. Based on the equation 3, we derive the original output distribution of the diffusion model through:

$$\hat{h_z} = h_z + \epsilon - \epsilon_\theta(z^t, f, t) \tag{4}$$

Here $\hat{h_z}$ signifies the diffusion model's output for the coarse-grained hand-grasping pose. The training objective encompasses both the reconstruction loss and physical constraints as follows [11, 30]:

$$h_m = Decoder(\mathbf{DAM}(\hat{h_z}, f)) \tag{5}$$

$$\mathcal{L} = \lambda_2\mathcal{L}_{recon} + \mathcal{L}_{physical} \tag{6}$$

where λ_2 is the balancing coefficient, the Decoder represents the network architecture that ultimately reconstructs $\hat{h_z}$ into the hand mesh h_m. We convert the noise prediction from the training diffusion model into both a reconstruction and physical constraints based on the DAM, enabling more efficient joint optimization. This approach enables the model to effectively learn both the physical interactions and the semantic context between hand meshes and objects, generating grasping poses that align with natural physical principles while meeting task-specific semantic requirements. Implementation details are provided in the Appendix.

3.3. Inference

During the inference process Fig. 2, we first sample noise $\epsilon \sim \mathcal{N}(0, 1)$ and condition it on C. This noise-conditioned input is transformed via Eq. 1 to produce a conditional feature representation f, which guides the diffusion model to synthesize coarse grasp poses. To enhance sampling efficiency, we employ the DDIM framework [26], governed by the rule:

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{x}_t, f)}{\sqrt{\alpha_t}}\right)}_{\text{predicted } \boldsymbol{x}_0} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta\boldsymbol{x}_t, f)}_{\text{direction pointing to } \boldsymbol{x}_t} + \underbrace{\sigma_t\epsilon}_{\text{random noise}} \tag{7}$$

where α_t represents the DDIM scheduling parameters, and ϵ_θ corresponds to the conditional denoising U-Net [28]. To enhance conditional dependence on physical constraints and linguistic semantics, we further refine this distribution through our DAM. The adjusted formulation is implemented as:

$$\hat{x}_0 = DAM(x_0, f) \tag{8}$$

where $\hat{x}_0$ denotes the optimized hand representation in latent space. We subsequently decode this refined latent vector into hand mesh h_m using a decoder. This process is formally expressed as:

$$h_m = Decoder(\hat{x}_0) \tag{9}$$

Our framework ensures both linguistic plausibility and physical feasibility of the generated hand mesh h_m. In contrast to previous approaches, such as [11, 17], which rely on test-time gradient optimization and face computational bottlenecks, our method enables robust one-stage generation without the need for Test-Time Adaptation.

4. Experiment

4.1. Automated Labeling for Dataset Enrichment

To enrich the OakInk [33] and GRAB [27] datasets, which focus on hand-object interactions, we propose an automated pipeline for instruction annotations. We begin by leveraging the AffordPose [10] dataset for cold-start training of a classifier [36]. This model generates initial language annotations, which are iteratively refined through error analysis to improve label consistency. The labeled data, along with the initial dataset, are then used for multiple rounds of training and labeling to ensure full dataset annotation. Finally, we integrate large language models [31] to generate task-oriented instructional text, enriching the annotations with step-by-step interaction guidance. Implementation details are provided in the Appendix.

We evaluate our method on four benchmarks: OakInk [33], GRAB [27], HO-3D [9], and AffordPose [10], following the experimental protocol in Sec. 4.2. For in-domain evaluation(Sec. 4.3), we train and test on OakInk and GRAB. GRAB contains 51 objects grasped by 10 subjects, while OakInk provides a larger-scale dataset with 1,700 objects manipulated by 12 subjects. For cross-dataset generalization (Sec. 4.3), we evaluate on HO-3D and AffordPose's out-of-domain object split under zero-shot settings, consistent with prior work [11, 27, 30, 38]. The AffordPose dataset was excluded from our training data due to its lack of MANO parameters and the presence of partial noise, both of which are incompatible with our MANO-based differentiable model. Quantitative validation supporting this decision is provided in the appendix.

4.2. Evaluation Metrics

Following established evaluation protocols [12, 27, 30, 38], we assess generated grasping poses using four criteria: (1)

Dataset	Method	Penetration Volume ↓	Simulation Displacement ↓	Contact Ratio ↑	Entropy ↑	Cluster Size ↑	ACC ↑
OakInk [33]	TTA [11]	8.54	2.17	95	2.88	2.79	73.17%
	FastGrasp [30]	6.08	2.47	77	2.83	3.62	66.77%
	D-VQVAE [39]	7.54	2.24	90	**2.89**	4.07	69.59%
	Ours(ControlNet)	5.28	3.31	80	2.88	4.07	75.71%
	Ours	**4.99**	**1.54**	**97**	**2.89**	**4.18**	**83.96%**
GRAB [27]	TTA [11]	6.84	1.44	**100**	2.86	1.61	63.00%
	FastGrasp [30]	4.61	**1.21**	96	2.76	1.96	55.00%
	D-VQVAE [39]	8.1	1.72	90	2.87	3.31	65.00%
	Ours(ControlNet)	6.58	1.56	93	2.81	3.44	67.50%
	Ours	**2.91**	**1.21**	**100**	**2.88**	**3.48**	**67.50%**

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 [38] 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 [9]	TTA [11]	12.34	3.12	95	2.75	2.98	58.33%
	FastGrasp [30]	15.45	2.73	98	**2.82**	2.25	51.00%
	D-VQVAE [39]	11.32	2.19	93	2.76	3.83	61.00%
	Ours(ControlNet)	15.06	2.41	97	2.81	3.58	51.00%
	Ours	**8.04**	**2.12**	**99**	**2.82**	**3.94**	**66.00%**
AffordPose [10]	TTA [11]	19.53	3.48	91	2.88	2.94	64.20%
	FastGrasp [30]	20.75	3.46	88	2.80	3.71	52.07%
	D-VQVAE [39]	21.72	4.34	93	2.91	3.64	63.57%
	Ours(ControlNet)	24.81	4.73	93	2.87	3.59	54.77%
	Ours	**11.12**	**3.39**	**95**	**2.92**	**3.83**	**72.43%**

Table 2. Comparison with previous methods on the **HO-3D** and **AffordPose** dataset, where our model is trained on the GRAB [27] dataset. Our model achieves state-of-the-art performance on two out-of-domain dataset, setting new benchmarks.

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 [11, 27, 30] 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 grasp generation literature [12, 17, 30], 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. We evaluate language alignment using the classifier from our data engine (Sec. 4.1), measuring how well-generated grasps correspond to their textual descriptions.

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 [33] and GRAB [27] 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 [30], D-VQVAE [39], and ControlNet [38].

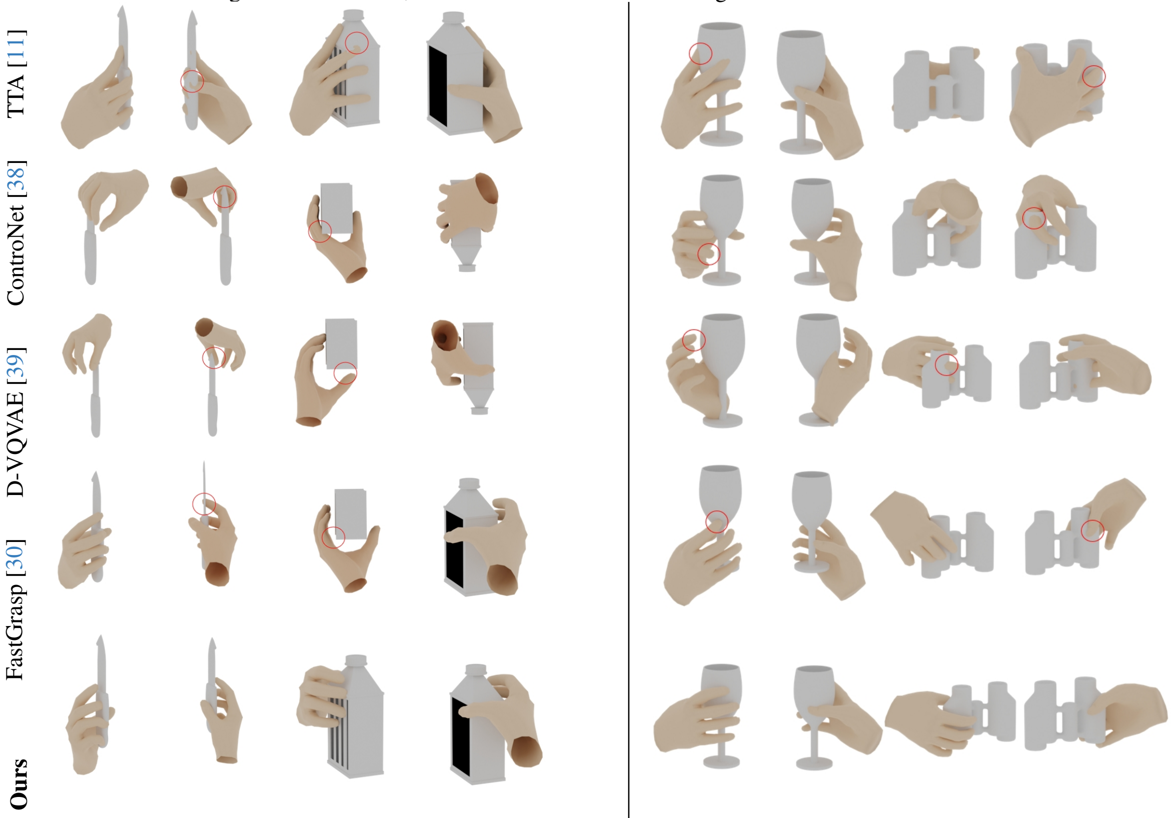
Out-of-Domain Evaluation. We further evaluated our model's universal applicability on the HO-3D [9] and AffordPose [10] datasets. As depicted in Tab. 2 and Fig . 6, out-of-domain evaluation validates our method's outstandingly accurate semantic results, in addition to physical generalization and generation 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.

Dataset	Method	Penetration Volume ↓	Simulation Displacement ↓	Contact Ratio ↑	Entropy ↑	Cluster Size ↑	ACC ↑
OakInk [33]	w/o object affordance	5.57	**1.26**	97	2.88	3.81	77.18%
	w/o DAM	5.86	1.69	97	**2.89**	**4.22**	82.01%
	Whole pipeline	**4.99**	1.54	**97**	**2.89**	4.18	**83.96%**
GRAB [27]	w/o object affordance	6.78	**0.61**	97	2.76	3.36	55.00%
	w/o DAM	4.94	1.54	96	2.83	**4.43**	63.33%
	Whole pipeline	**2.91**	1.21	**100**	**2.88**	3.48	**67.50%**
HO-3D [9]	w/o object affordance	12.52	**1.12**	99	2.76	3.66	**66.00%**
	w/o DAM	8.14	2.25	99	2.74	**4.28**	65.00%
	Whole pipeline	**8.04**	2.12	**99**	**2.82**	3.94	**66.00%**
AffordPose [10]	w/o object affordance	31.78	**1.56**	94	2.91	3.99	64.35%
	w/o DAM	16.22	3.07	89	2.90	**4.47**	66.74%
	Whole pipeline	**11.12**	3.39	**95**	**2.92**	3.83	**72.43%**

Table 3. Ablation study results on the **GRAB, OakInk, HO-3D, AffordPose** datasets [9, 10, 27, 33]. The evaluation of the HO-3D and AffordPose is an out-of-domain generalization test, where the model is trained using the GRAB dataset.



OakInk [33]
Left. This is a knife, hold the handle firmly.
Right. Wrap your hand around the bottle.

GRAB [27]
Left. Ensure a tight wrap around the wine-glass.
Right. Grip the handle of the telescope securely.

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.

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 geometric 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 distri-

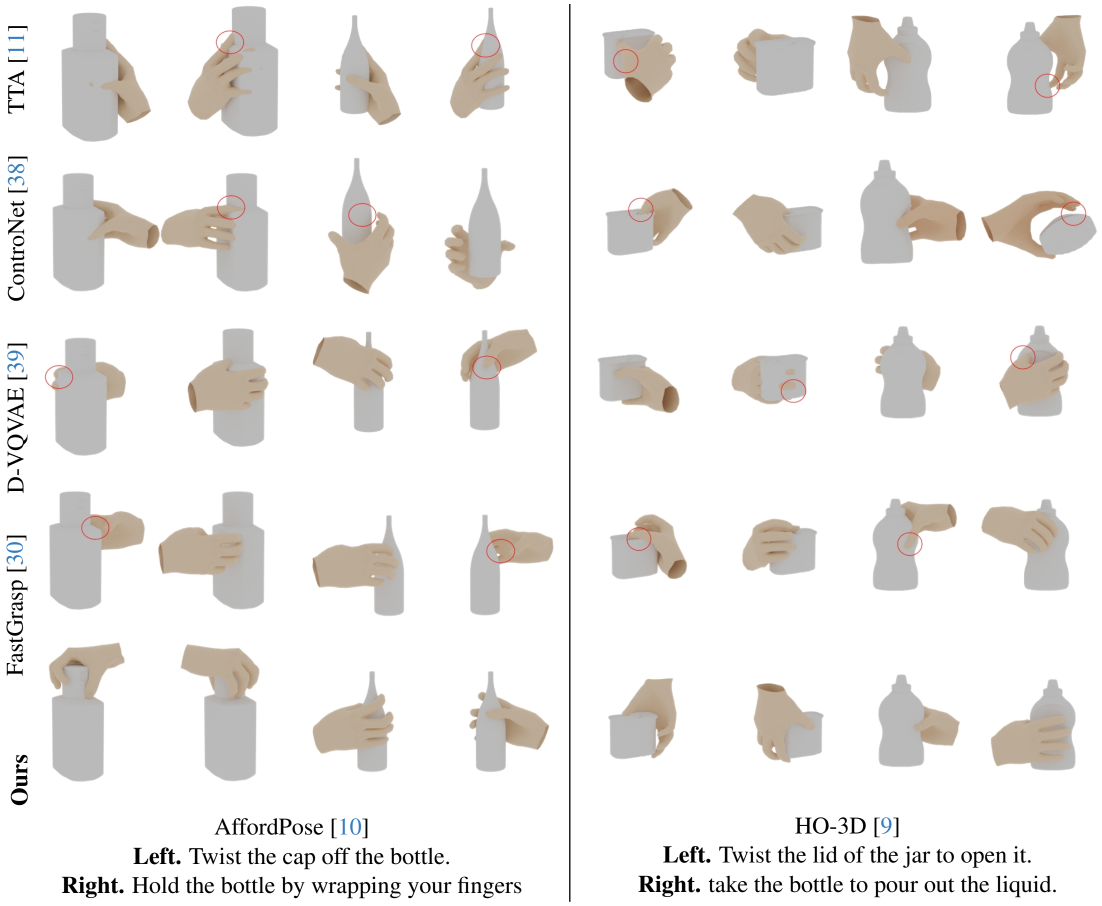

TTA [11]
ControNet [38]
D-VQVAE [39]
FastGrasp [30]
Ours
AffordPose [10]
Left. Twist the cap off the bottle.
Right. Hold the bottle by wrapping your fingers
HO-3D [9]
Left. Twist the lid of the jar to open it.
Right. take the bottle to pour out the liquid.

Figure 6. Qualitative comparisons with state-of-the-art methods on AffordPose [10] and HO-3d [9] datasets. Each pair (two columns) visualizes the generated grasps from two different views. Our method has stronger semantic accuracy compared to other methods.

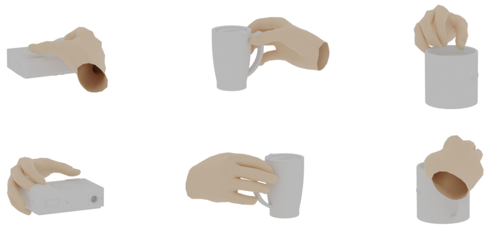


Figure 7. **Failure cases of our method.** Each pair displays a sample from two views.

bution, 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.

5. Conclusion and Discussion

This paper has introduced an automated annotation engine that improves hand-object interaction datasets by integrating linguistic instructions. Leveraging object affordance as cross-modal information compensation and a Distribution Adjustment Module (DAM), our approach enables the model to capture spatial details and instructional semantic information synergistically. The method strengthens adherence to physical constraints through geometric grounding and achieves superior linguistic-3D alignment compared to conventional approaches.

Limitations. Our data-driven pipeline does not yet account for environmental interactions, such as gravity and friction, limiting its realism and creating a gap between simulation and reality. As shown in Fig. 7, while the generated grasp posture ensures adequate object contact, it neglects gravity and friction, leading to instability. We believe that integrating a physics engine with the generative model's training pipeline could help bridge this gap.

6. Acknowledgement

This work was supported by NSFC 62350610269, Shanghai Frontiers Science Center of Human- centered Artificial Intelligence, and MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University).

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. *16th European Conference on Computer Vision (ECCV)*, 2020. 2

[2] 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. 2, 3

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

[4] Changmao Chen, Yuren Cong, and Zhen Kan. Worldafford:affordance grounding based on natural language instructions. *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 822–828, 2024. 2

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 202–221. Springer, 2020. 2

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 2

[7] Sheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1787, 2021. 2

[8] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. 2

[9] 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, 5, 6, 7, 8, 4

[10] 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, 5, 6, 7, 8, 1, 4

[11] 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. 1, 5, 6, 7, 8, 4

[12] 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. 1, 2, 5, 6

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

[14] 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

[15] 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. 3

[16] 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

[17] 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. 1, 2, 5, 6

[18] 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. 3

[19] 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

[20] 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

[21] Liangsheng Lu, Wei Zhai, Hongcheng Luo, Yu Kang, and Yang Cao. Phrase-based affordance detection via cyclic bilateral interaction. *IEEE Transactions on Artificial Intelligence*, 4:1186–1198, 2022. 2

[22] Hongcheng Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2242–2251, 2022. 2

[23] 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. 2, 3

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

[25] 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. 2, 4

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

[27] 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. 1, 2, 5, 6, 7, 4

[28] 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

[29] 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. 2, 3, 4, 1

[30] Xiaofei Wu, Tao Liu, Caoji Li, Yuexin Ma, Yujiao Shi, and Xuming He. Fastgrasp: Efficient grasp synthesis with diffusion. *ArXiv*, abs/2411.14786, 2024. 1, 2, 3, 5, 6, 7, 8, 4

[31] 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. 3, 5

[32] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 3

[33] 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, 5, 6, 7, 1, 4

[34] 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. 2, 3, 4, 1

[35] Yuhang Yang, Wei Zhai, Hongcheng Luo, Yang Cao, Jiebo Luo, and Zhengjun Zha. Grounding 3d object affordance from 2d interactions in images. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10871–10881, 2023. 2

[36] 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. 5, 1

[37] 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

[38] 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. 5, 6, 7, 8

[39] Zhe Zhao, Mengshi Qi, and Huadong Ma. Decomposed vector-quantized variational autoencoder for human grasp generation. In *European Conference on Computer Vision*, 2024. 6, 7, 8, 1