SNS-Grasp: Semantic-guided Noise Scaling for Grasp Generation

Zhenhua Tang[1], Yudian Zheng[1], Yuzhang Zhong[1], Haolun Li[2], Yanbin Hao[3], Chi-Man Pun[1]*

[1]University of Macau, Macau SAR, China
[2]Nanjing University of Posts and Telecommunications, Jiangsu, China
[3]Hefei University of Technology, Anhui, China
zhenhuat@foxmail.com, cmpun@um.edu.mo

Abstract

While diffusion models show promise for intent-based grasp generation, their isotropic noise schedules struggle with joint-specific sensitivity and task-aware variability. This limitation leads to grasps with suboptimal semantic alignment or physical feasibility. To address this challenge, we propose Semantic-guided Noise Scaling for grasp generation (SNS-Grasp), a novel framework that integrates two key innovations. First, the Semantic-guided Noise Scaling Diffusion (SNS-Diff) module generates intent-aware grasps by replacing isotropic noise with anisotropic modulation, dynamically adapting to task semantics and joint-specific sensitivity. Specifically, SNS-Diff leverages a pretrained Intent Recognizer to extract task-aware confidence scores and joint-specific gradient sensitivities from the interaction context. These signals adjust the noise scaling during denoising, downweighting perturbations for semantically critical joints to ensure semantic alignment. Second, the Fine-grained Grasp Refinement (FGR) module establishes dynamic joint-vertex coupling through fine-grained hand-object spatial relationships, enabling iterative optimization of physically executable grasps. Extensive experiments on OakInk and GRAB demonstrate SNS-Grasp's superior performance in semantic accuracy and physical feasibility, with robust generalization to unseen objects.

Introduction

Intent-based grasp generation aims to synthesize human-like hand poses that satisfy specific manipulation intents (e.g., holding a mug), with applications in human-computer interaction (Pollard and Zordan 2005), virtual reality (Höll et al. 2018; Wu et al. 2020), and robotics (Liu et al. 2024). Recent advances leverage generative frameworks (Taheri et al. 2020; Yang et al. 2022; Jian et al. 2023; Li et al. 2024; Tang et al. 2025) to generate intent-aware hand poses, particularly diffusion-based models (Wei et al. 2024; Wu et al. 2025; Tang et al. 2025), which exhibit superior performance due to their stable training dynamics and high sample diversity. Despite these advantages, their isotropic noise schedules fail to address two critical challenges in affordance-aware grasp generation. First, the biomechanical hierarchy of the MANO (Romero, Tzionas, and Black 2022) hand model requires differentiated control across joints, where distal joints (like fingertips) demand finer control than proximal joints (like the wrist) to maintain precise manipulation semantics. Second, the divergence in interaction purposes between exploratory actions (e.g., holding a camera body) and constrained manipulations (e.g., pressing camera button) necessitates task-adaptive noise profiles, as uniform noise application may either hinder precision or limit exploration.

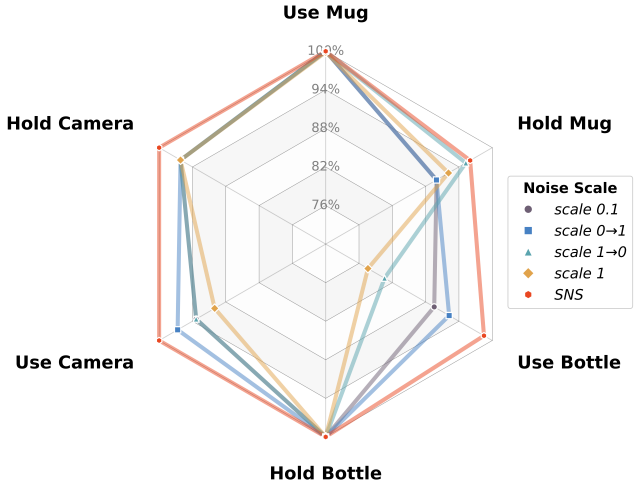


Figure 1: Performance comparison of grasp semantics under different noise scaling modes in the diffusion model: (1) *Scale 0.1* applies isotropic noise scaling (10% intensity) to all MANO joints; (2) *Scale* $0 \rightarrow 1$ progressively increases scaling from joint 0 (wrist) to joint 15 (pinky tip); (3) *Scale* $1 \rightarrow 0$ decreases scaling from joint 0 to joint 15; (4) *Scale 1* retains the original noise values. Our *SNS* dynamically scales noise based on manipulation intent and joint-specific criticality, demonstrating superior performance.

To explore biomechanical and task-semantic requirements in grasp generation, we experimentally investigate the diffusion model (Tang et al. 2025) with four noise configurations: isotropic noise (*scale 1*), weak isotropic noise (*scale 0.1*), increasing noise (*scale* $0 \rightarrow 1$), and decreasing noise (*scale* $1 \rightarrow 0$). We evaluate grasps generated under these different noise scaling modes across six affordance-aware

*Corresponding author.

interaction categories (Use/Hold Mug/Bottle/Camera). The evaluation employs an intent recognizer (see Intent Recognizer section) pre-trained on approximately 17,000 demonstration grasps, which achieves 98.7% test accuracy. As shown in Figure 1, the result comparison reveals:

- **Joint-specific sensitivity**: Joint-specific noise *scale* $0 \rightarrow 1$ outperforms isotropic noise *scale 1* by 14.63% in *Use Bottle* tasks;
- **Task-aware variability**: *scale* $0 \rightarrow 1$ yields best results for *Use Camera* and *Hold Bottle*, while *scale* $1 \rightarrow 0$ is optimal for *Hold Mug*.

These patterns suggest that semantically precise generation in diffusion model requires both: (1) *joint-specific noise adaptation* for critical joints, and (2) *task-aware noise modulation* to preserve manipulation constraints, revealing limitations in isotropic noise approaches (Samuel et al. 2023; Guo et al. 2024; Qi et al. 2024).

In this paper, we propose the Semantic-guided Noise Scaling for Grasp Generation (SNS-Grasp) framework, which jointly addresses semantic alignment and physical feasibility through two key innovations: a Semantic-guided Noise Scaling Diffusion (SNS-Diff) module and a Fine-grained Grasp Refinement (FGR) module. Since semantically precise generation necessitates both joint-specific adaptation and task-aware modulation (Figure 1), SNS-Diff introduces a novel semantic-guided noise scaling (SNS) mechanism. The key insight is that semantically critical joints receive reduced noise, limiting their output variance to ensure manipulation accuracy. Meanwhile, non-critical joints maintain higher noise levels, enabling anatomically valid exploration. Specifically, SNS-Diff implements three operations: First, a lightweight intent recognizer analyzes the interaction context to generate task-aware confidence scores and joint-specific gradient sensitivities. Second, these signals dynamically transform initial isotropic noise into anisotropic joint-specific profiles. This SNS process assigns reduced noise scales to semantically critical joints while preserving original noise magnitudes for non-critical joints. Finally, a Semantic Calibrator Transformer (SCT) (Tang et al. 2025) conditions the noise profiles on task instructions, ensuring semantic alignment throughout denoising.

To enhance physical feasibility, prior works (Taheri et al. 2020; Yang et al. 2022; Jian et al. 2023; Tang et al. 2025) optimize MANO parameters using vertex-object signed distance field (SDF) features. However, MANO's fixed skinning weights intrinsically limit deformation adaptability, potentially causing semantic drift and physical implausibility. Thus, our FGR module introduces a geometry-aware optimization pipeline with dynamic joint-vertex coupling. Concretely, FGR first explicitly models contact constraints through SDF evaluation. It then employs cross-attention to dynamically couple each joint with all hand mesh vertices by computing global attention weights from fine-grained SDF features, effectively mitigating MANO's rigid skinning weights limitations. Through an iterative process, physical penetrations are progressively minimized while maintaining manipulation semantics.

In summary, the key contributions of SNS-Grasp are threefold. First, the proposed SNS-Diff model pioneers joint-specific and task-aware anisotropic noise scheduling for grasp generation, where the noise profiles are adaptively adjusted to preserve manipulation semantics. Second, the FGR establishes dynamic joint-vertex coupling through SDF-guided cross-attention, overcoming MANO's rigid skinning constraints. Third, extensive experimental validation on two public datasets demonstrates superior performance in semantic controllability, physical feasibility, and generalization capability compared to state-of-the-art approaches.

Related Work

In this section, we first review 3D hand-object interaction synthesis, categorizing methods as general and intent-based grasp generation according to manipulation semantics. We then introduce diffusion noise scaling mechanism as a promising approach for semantic-controlled generation.

General Grasp generation. The general grasp generation task focuses on producing physically feasible hand poses for given objects without specific manipulation intent (Miller and Allen 2004; Hasson et al. 2019; Liu et al. 2019, 2020; Wang et al. 2023). Early regression-based approaches (Liu et al. 2019, 2020) directly predicted grasp parameters from object geometry, but often suffered from limited diversity and accuracy. To address these limitations, several studies have adopted various generative models (Corona et al. 2020; Taheri et al. 2020, 2022; Weng et al. 2024; Liu and Yi 2024), among which diffusion models have recently emerged as a promising paradigm. DexDiffuser (Weng et al. 2024) combines a conditional diffusion sampler with an evaluator to generate high-quality grasps from point clouds, while G-HOP (Ye et al. 2024) aligns hand skeletal distance fields with object geometry for coherent 3D representation. GeneOH Diffusion (Liu and Yi 2024) handles noisy interaction trajectories through a diffusion-based denoising process. Furthermore, contact map prior has proven to be an effective approach for enhancing physical feasibility in grasp generation (Brahmbhatt et al. 2019, 2020; Jiang et al. 2021; Li et al. 2022; Liu et al. 2023b). For instance, GraspTTA (Jiang et al. 2021) introduces a GraspCVAE for grasp generation coupled with a ContactNet for contact prediction. Contact2Grasp (Li et al. 2022) learns contact map distributions before mapping to grasps, and UGG (Lu et al. 2023) proposes a unified framework for jointly generating hands, objects, and their contact information. While these methods achieve stable and realistic hand poses, they fail to capture object affordances and manipulation semantics.

Intent-based grasp generation. To study how humans manipulate objects with specific intent, recent datasets such as OakInk (Yang et al. 2022) and AffordPose (Jian et al. 2023) have collected multimodal hand-object interaction data. OakInk (Yang et al. 2022) captures grasps based on diverse object meshes and action-related textual descriptions, including use, hold, lift-up, hand-out, and receive, thereby formalizing the task of intent-based grasp generation. Affordpose (Jian et al. 2023) further annotates specific part-level affordances on the objects, such as twist, pull, handle-grasp, and the corresponding parts. As a baseline, OakInk

(Yang et al. 2022) and AffordPose (Jian et al. 2023) inject the intent representation into GrabNet (Taheri et al. 2020) to guide generation.

Recent studies in intent-based grasp generation have proposed diverse methodologies to resolve manipulation ambiguity, where a single intent may correspond to multiple interactions. Language-guided approaches (Christen et al. 2024; Chang and Sun 2024; Li et al. 2024; Jian et al. 2025; Wu et al. 2025) such as SemGrasp (Li et al. 2024) leverage LLaVA (Liu et al. 2023a) for automatic grasp annotation and text-pose alignment, while PartDexTOG (Wu et al. 2025) employs GPT-4o (Achiam et al. 2023) to generate part-level manipulation descriptions. Retrieval-augmented methods like RAGG (Tang et al. 2025) demonstrate improved generalization through nearest-neighbor grasp retrieval. While effective for semantic conditioning, these methods maintain standard noise schedules, limiting joint-specific adaptation. This motivates our investigation of dynamic noise modulation, building on adaptive noise scaling successes in related domains (Samuel et al. 2023; Zhang et al. 2025).

Noise scaling Mechanism. Recent advances in diffusion models demonstrate their exceptional generative capability. While conventional approaches apply isotropic Gaussian noise, prior studies prove that semantic-aware generation requires adaptive noise mechanisms (Guo et al. 2024; Samuel et al. 2023; Qi et al. 2024). For instance, Samuel *et al.* (Samuel et al. 2023) proposed seed interpolation for rare concept synthesis, while Qi *et al.* (Qi et al. 2024) introduced noise selection via inversion stability. Zhang *et al.* (Zhang et al. 2025) further developed uncertainty-guided regional noise adjustment. For grasp generation, we extend these principles through SNS, specifically addressing the varying semantic criticality across hand joints during manipulation.

Proposed Method

Figure 2 illustrates our SNS-Grasp framework, which consists of two core components: (a) SNS-Diff, and (b) FGR. SNS-Diff first employs a pretrained intent recognizer to extract joint-specific gradients and classification confidence; it then applies SNS to convert isotropic noise to anisotropic noise, ultimately generating semantically valid grasps conditioned on the noise profiles. Subsequently, FGR refines the generated coarse grasps to ensure physical feasibility.

Intent Recognizer

In intent-based grasp generation, the semantic importance of hand joints varies significantly across different manipulation tasks (e.g., use camera vs. hold mug). However, prior works (Taheri et al. 2020; Yang et al. 2022) lack a principled framework to quantify joint-specific semantic importance. Inspired by the hand-object affordance understanding paradigm (Jian et al. 2023), we propose to learn these importance metrics through an intent-aware interaction classifier, where the gradient magnitude naturally indicates each joint's contribution to semantic intent fulfillment - larger gradients correspond to higher semantic-critical joints. Thus, we implement a lightweight intent recognizer that predicts the manipulation category from a hand mesh and an object point cloud.

The intent recognizer processes a composite input $\mathbf{S} = \mathrm{Concat}(\mathcal{M}, \mathcal{O}) \in \mathbb{R}^{4874\times 3}$, where $\mathcal{M} \in \mathbb{R}^{778\times 3}$ represents the hand mesh vertices generated from MANO parameters $\{\theta \in \mathbb{R}^{16\times 6}, P \in \mathbb{R}^3\}$ (encoding joint rotations and global translation respectively), and $\mathcal{O} \in \mathbb{R}^{4096\times 3}$ denotes the object point cloud. This composite input is then transformed through multiple channel mixers and spatial mixers parameterized, which are detailed in the appendix.

Finally the learned representation $\mathbf{F}$ is squeezed along the spatial dimension and processed by a classification head:

$$z = \mathrm{Softmax}(\mathbf{W}_2\mathrm{ReLU}(\mathbf{W}_1\mathbf{F} + \mathbf{b}_1) + \mathbf{b}_2), \quad (1)$$

where $\mathbf{W}_1 \in \mathbb{R}^{32\times 64}$, $\mathbf{b}_1 \in \mathbb{R}^{32}$, $\mathbf{W}_2 \in \mathbb{R}^{2\times 32}$, and $\mathbf{b}_2 \in \mathbb{R}^2$ are learnable parameters. With merely 0.414M parameters, this lightweight design achieves accurate intent discrimination, attaining 98.7% test classification accuracy.

Semantic-guided Noise Scaling

After pre-training the intent recognizer, we reformulate the denoising process using the SNS mechanism. As shown in Figure 2, we begin with an initial perturbed grasp $x_T \sim \mathcal{N}(0, \mathbf{I})$, where isotropic noise is applied uniformly to all joints. This initial denoising phase first generates an exploratory grasp proposal y_0 through standard denoising sampling. Semantically critical joints (e.g., fingertips when manipulating a camera) are identified through the intent recognizer's gradients and confidence scores. Since these joints require higher stability, we restart denoising from x_T with SNS. This mechanism adaptively rescales noise based on each joint's semantic criticality, ensuring stable generation of semantically critical joints.

Specifically, the SNS adjusts the noise magnitude for each joint based on its semantic-criticality. For joint j, it computes a joint-specific and task-aware coefficient:

$$\gamma_j = \underbrace{\left\| \frac{\partial \mathcal{L}_{int}}{\partial \theta_j} \right\|_2}_{\text{gradient magnitude}} \cdot \underbrace{\sigma(z)}_{\text{confidence}} \in [0, +\infty), \quad (2)$$

where $\mathcal{L}_{int}$ is the cross-entropy loss of the intent recognizer, θ_j denotes the rotation representation of joint j and $\sigma(z)$ is the logit value corresponding to the ground truth intent class. Higher γ_j indicates stronger semantic-criticality for intent preservation. Thus, the noise is scaled as:

$$x_{j,T}^{\text{SNS}} = x_{j,T} \cdot \left(\frac{1}{1+\gamma_j} \right) \in (0, 1]. \quad (3)$$

The SNS mechanism dynamically modulates noise based on joint semantic importance. For joints with high γ_j values, the noise scaling factor approaches zero ($\frac{1}{1+\gamma_j} \to 0$), preserving the intent-critical geometry. Conversely, for joints with low γ_j values, the scaling factor approaches 1 (when $\gamma_j = 0$), preserving most of the original noise magnitude to maintain natural kinematic diversity.

Theoretical Analysis. The SNS modulation induces anisotropic variance bounds in the generated distribution.

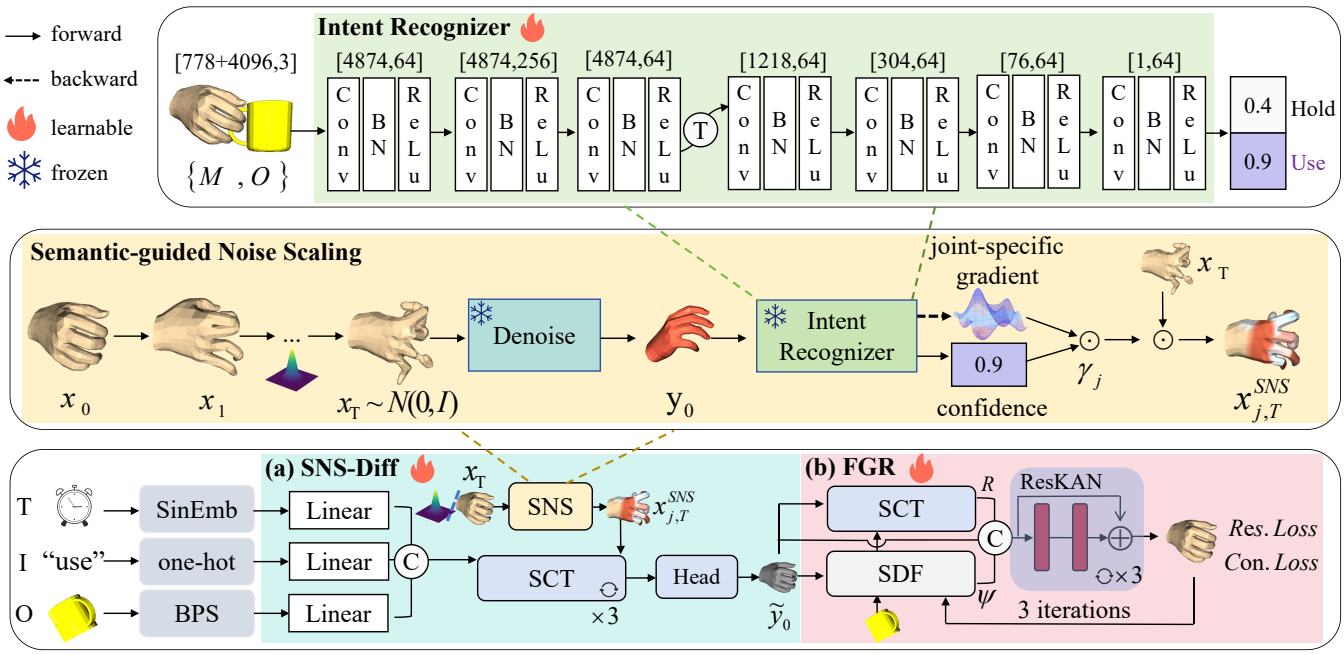


Figure 2: Overview of the SNS-Grasp framework: (a) SNS-Diff generates intent-aware grasps by converting isotropic to anisotropic noise via the SNS module, with scaling parameters guided by the pretrained Intent Recognizer. (b) FGR establishes dynamic joint-vertex coupling through SDF-based cross-attention and refines physical feasibility via ResKAN layers.

To quantify this effect, we derive the joint-specific variance upper bound under the DDPM framework (Ho, Jain, and Abbeel 2020), and the output variance of joint j satisfies:

$$\mathrm{Var}(p(x^{\mathrm{SNS}}_{j,0})) \leq \sum_{t=1}^{T} \frac{\beta_t(1-\bar{\alpha}_t)}{\bar{\alpha}_t(1+\gamma_j)^2} = \frac{C}{(1+\gamma_j)^2}, \quad (4)$$

where $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, $\alpha_t := 1 - \beta_t$, β_t is the cosine noise variance schedule and $C = \sum_{t=1}^{T} \beta_t(1-\bar{\alpha}_t)/\bar{\alpha}_t$. This bound implies that for high-γ_j joints, the variance reduction forces generated samples to concentrate tightly around task-optimal configurations, ensuring semantic alignment with manipulation intent; meanwhile, low-γ_j joints maintain higher variability to preserve natural kinematic diversity. Experimental results also demonstrate that SNS produces more compact distributions compared to isotropic noise baselines (see Experiment section and Appendix).

Semantic-guided Noise Scaling Diffusion

Building on the anisotropic noise distribution established by SNS, we formalize the conditioned denoising process where SNS-Diff reconstructs the clean pose x_0 from noise-modulated input x^{SNS}_T. The denoiser $\mathfrak{D}$ integrates three conditionally processed inputs: (1) diffusion timesteps T with sinusoidal encoding projected to $\mathbb{R}^{256}$;(2) manipulation intent I encoded as one-hot vectors through a trainable 16D lookup table and linearly projected to $\mathbb{R}^{256}$; and (3) object point clouds O represented as 4096D BPS features (Prokudin, Lassner, and Romero 2019) and compressed to $\mathbb{R}^{256}$ via linear transformation. This yields the generation:

$$\tilde{y}_0 = \mathfrak{D}\left(x^{\mathrm{SNS}}_T, \phi_t(T), \phi_i(I), \phi_o(O)\right), \quad (5)$$

where each $\phi_{(\cdot)}$ denotes the corresponding feature embedding layer.

The transformed features $\{\phi_t(T), \phi_i(I), \phi_o(O)\} \in \mathbb{R}^{256}$ are concatenated into a 3×256 conditioning matrix, which dynamically guides the denoising process through stacked Semantics Calibration Transformer (SCT) blocks (Tang et al. 2025). The final SCT output $h \in \mathbb{R}^{256}$ is decoded by a two-layer residual MLP to predict the clean MANO parameters $\tilde{y}_0 = \{\tilde{\theta}, \tilde{P}\}$. These parameters are then passed through the MANO layer (Romero, Tzionas, and Black 2022) to generate the semantic-aware hand mesh $\tilde{m}_0$. To train the model, we optimize a composite loss function that jointly enforces mesh reconstruction accuracy and contact feasibility:

$$\mathcal{L} = \underbrace{\lambda_1\|\tilde{m}_0 - m_0\|^2 + \lambda_2\|\tilde{e}_0 - e_0\|^2}_{\text{mesh reconstruction}} + \underbrace{\lambda_3\mathcal{L}_{h2o} + \lambda_4\mathcal{L}_{o2h}}_{\text{contact constraints}}, \quad (6)$$

where m_0 denotes the ground-truth hand mesh, $\tilde{e}_0$ and e_0 are the edge lengths computed on the predicted and ground-truth meshes over corresponding vertex pairs, and $\mathcal{L}_{h2o}$, $\mathcal{L}_{o2h}$ penalize hand-object interpenetration. The loss weights are set to $\lambda_1 = 34.825$, $\lambda_2 = 29.85$, $\lambda_3 = 34.825$, and $\lambda_4 = 29.85$.

Fine-grained Refinement

To enhance the physical feasibility of the coarse grasp $\tilde{y}_0 = \{\tilde{\theta}, \tilde{P}\}$ generated by SNS-Diff, our FGR module implements dynamic joint-vertex coupling via SDF-guided cross-attention. As illustrated in Figure 2(b), given the initial MANO parameters $\{\tilde{\theta}, \tilde{P}\}$ and corresponding mesh vertices

Algorithm 1: Fine-Grained Grasp Refinement (FGR)

Require: Initial hand pose $\{\tilde{\theta}^0, \tilde{P}^0\}$, object mesh $\mathcal{O}$, MANO parameters β
Ensure: Refined Grasp $\{\tilde{\theta}^*, \tilde{P}^*\}$
for $i \leftarrow 0$ **to** 2 **do**
  Hand Mesh Generation:
  $\mathbf{V}^i \leftarrow \mathcal{M}(\tilde{\theta}^i, \tilde{P}^i, \beta)$
  Signed Distance Field:
  $\Psi^i \leftarrow \{\text{sd}(\mathbf{V}_j^i, \mathcal{O})\}_{j=1}^{778}$
  Joint-vertex Coupling:
  $\mathbf{C}^i \leftarrow \text{SCT}\left(\phi_q(\{\tilde{\theta}^i, \tilde{P}^i\}), \phi_k(\Psi^i), \phi_v(\Psi^i)\right)$
  Grasp Refinement:
  $H_0^i \leftarrow \text{concat}(\{\tilde{\theta}^i, \tilde{P}^i\}, \Psi^i, \mathbf{C}^i)$
  $H_1^i \leftarrow \text{ResKAN}_1(H_0^i)$
  $H_2^i \leftarrow \text{ResKAN}_2(\text{concat}(H_1^i, H_0^i))$
  $H_3^i \leftarrow \text{ResKAN}_3(\text{concat}(H_2^i, H_0^i))$
  $\{\tilde{\theta}^{i+1}, \tilde{P}^{i+1}\} \leftarrow \text{Linear}(H_3^i)$ {Update grasp}
end for
return $\{\tilde{\theta}^*, \tilde{P}^*\} \leftarrow (\{\tilde{\theta}^3, \tilde{P}^3\}$

$\mathbf{V} \in \mathbb{R}^{778\times 3}$, we first compute per-vertex signed distance to the object surface $\mathcal{O}$:

$$\Psi = [\text{sd}(\mathbf{V}_1, \mathcal{O}), \ldots, \text{sd}(\mathbf{V}_{778}, \mathcal{O})] \in \mathbb{R}^{778\times 1} \quad (7)$$

where $\text{sd}(\mathbf{V}_j, \mathcal{O})$ computes the signed distance from hand vertex $\mathbf{V}_j$ to the object surface $\mathcal{O}$.

The key innovation over prior works (Taheri et al. 2020; Yang et al. 2022) is our SCT-based dynamic coupling:

$$\mathbf{C} = \text{SCT}\left(\phi_q(\{\tilde{\theta}, \tilde{P}\}), \phi_k(\Psi), \phi_v(\Psi)\right) \in \mathbb{R}^{17\times d}, \quad (8)$$

where $\phi_{q/k/v}$ are linear projections aligning dimensions for query (joint parameters), key (SDF features), and value (SDF features). Through fine-grained, geometry-aware attention, all 778 vertices adaptively respond to the 16 joints' movements, achieving more flexible deformation than MANO's linear blend skinning.

Subsequently, the hand representation $\{\tilde{\theta}, \tilde{P}\}$ is progressively optimized through successive ResKAN blocks (Tang et al. 2025), which jointly incorporate SDF constraints Ψ and dynamic joint-vertex couplings $\mathbf{C}$ to produce the refined grasp $\{\tilde{\theta}^*, \tilde{P}^*\}$. Algorithm 1 details our PyTorch implementation, with $\{\tilde{\theta}^{(0)}, \tilde{P}^{(0)}\}$ denoting the initial state. The SCT and ResKAN modules are implemented as described in (Tang et al. 2025). The network was optimized over three iterations by minimizing the contact and reconstruction losses.

Experiment

Datasets and Evaluation Metrics

OakInk (Yang et al. 2022) is a large-scale hand-object interaction dataset containing 1,800 object models across 32 categories. Following the experimental protocol of (Tang et al. 2025), we select 9 representative categories (bottles, cameras, cylinder bottles, eyeglasses, game controllers, lotion pumps, mugs, pens, and trigger sprayers), each annotated with two functional intents (*use* and *hold*). To evaluate generalization performance, we introduce 8 held-out categories (headphones, bowls, cups, hammers, frying pans, scissors, wine glasses, binoculars, and teapots) as unseen test cases.

GRAB (Taheri et al. 2020) provides real-world grasp demonstrations from 10 subjects interacting with 51 objects. For out-of-domain evaluation, we select objects manipulated by subject S1 in *pass* actions. This includes both: (1) categories overlapping with OakInk (camera and mug) for cross-dataset consistency validation, and (2) novel categories (wineglass and toothpaste) to test unseen object generalization.

Evaluation metrics. Following established protocols (Taheri et al. 2020; Yang et al. 2022; Tang et al. 2025), we assess grasp quality through five metrics evaluating both physical feasibility and semantic controllability. For physical feasibility, we measure: (1) *penetration depth*, (2) *solid intersection volume* (Yang et al. 2021), and (3) *simulation displacement* (Hasson et al. 2019). For semantic controllability, we report: (4) *intent recognition accuracy* using our pre-trained classifier, and (5) *user preference rate* assessed through a web-based platform, where 10 participants selected their preferred samples from 50 randomly presented trials based on evaluations of semantic intent accuracy and physical realism.

Performance Comparison

Table 1 presents a comprehensive evaluation across three experimental configurations: (1) seen object categories in OakInk, (2) unseen object categories in OakInk, and (3) object categories in GRAB. For OakInk's seen object categories, SNS-Grasp demonstrates significant improvements across all metrics. Our SNS-Grasp achieves 0.446 cm average penetration depth (4.5% lower than RAGG) and 1.411 cm average displacement (9.8% reduction), while maintaining the lowest interaction volume at 3.306 cm^3 (4.6% improvement over RAGG). The most notable advancement is SNS-Grasp's 97.39% controllability Auc. (5.1% higher than RAGG), which directly stems from our SNS strategy's ability to dynamically prioritize intent-critical joints while preserving natural variability. User evaluators confirmed these results with a 82.06% preference rate (vs. RAGG's 62.73%), validating our method's ability to generate physically stable and semantically precise grasps.

For unseen object categories in OakInk, SNS-Grasp exhibits strong cross-category generalization, achieving 0.459 cm average penetration depth (5.5% improvement over GrabNet) with superior stability (1.979 cm average displacement). While GrabNet achieves better interaction volumes (6.228 cm^3 vs. our 6.933 cm^3), this advantage stems from their tendency to generate simpler, semantically-unconstrained grasps that prioritize minimal intersection at the expense of functional precision. This trade-off explains their significantly lower controllability Auc. (GrabNet: 67.80%, AffordPose: 67.23%) compared to our 89.27%. The 67.50% user preference rate for functional grasps (vs. RAGG's 57.5%) demonstrates our SNS-Grasp's capability to: (1) transfer manipulation knowledge to novel objects

Data	Method	Physical Feasibility			Controllability	
		Pen. ↓	Ins. ↓	Dis.↓	Auc. ↑	User. ↑
OakInk seen	GrabNet (Yang et al. 2022)	0.527 ± 0.589	5.016 ±13.030	1.784 ± 2.184	89.50	23.96
	Affordpose (Jian et al. 2023)	0.499 ± 0.562	4.309 ± 9.898	1.945 ± 2.243	91.04	34.96
	RAGG (Tang et al. 2025)	0.467 ± 0.486	3.467 ± **7.156**	1.565 ± 1.736	92.24	62.73
	SNS-Grasp	**0.446** ± **0.467**	**3.306** ± 7.682	**1.411** ± **1.382**	**97.39**	**82.06**
OakInk unseen	GrabNet (Yang et al. 2022)	0.486 ± **0.469**	**6.228** ± **8.887**	2.036 ± 2.018	67.80	35.00
	Affordpose (Jian et al. 2023)	0.497 ± 0.571	7.007 ± 13.122	**1.945** ± 1.946	67.23	35.00
	RAGG (Tang et al. 2025)	0.532 ± 0.580	7.826 ± 11.099	2.071 ± 2.287	84.75	57.50
	SNS-Grasp	**0.459** ± 0.487	6.933 ± 11.842	1.979 ± **1.222**	**89.27**	**67.50**
GRAB	GrabNet (Yang et al. 2022)	0.401 ± 0.389	2.387 ± 2.705	1.734 ± 2.131	67.31	27.56
	Affordpose (Jian et al. 2023)	0.416 ± 0.402	2.590 ± 2.827	1.760 ± 2.095	68.96	19.03
	RAGG (Tang et al. 2025)	0.369 ± 0.408	2.291 ± 2.831	2.168 ± 2.534	62.11	51.93
	SNS-Grasp	**0.343** ± **0.328**	**1.734** ± **2.329**	**1.652** ± **1.760**	**73.44**	**84.46**

Table 1: Performance comparisons in terms of physical feasibility and controllability on OakInk dataset and GRAB dataset. The best result in each column is marked in **bold**. ↓ indicates lower is better; ↑ indicates higher is better.

while (2) maintain physical feasibility during generalization.

For objects in the GRAB dataset, SNS-Grasp demonstrates robust cross-dataset generalization capabilities, achieving state-of-the-art performance across all metrics. Our method establishes new benchmarks with 0.343 cm average penetration depth (7.0% improvement over RAGG) and 1.652 cm average displacement (23.8% lower than RAGG), while significantly reducing interaction volume to 1.734 cm^3 (24.3% improvement). In contrast, GrabNet and Affordpose suffer from excessively large interaction volumes (2.387 cm^3 and 2.590 cm^3, respectively), which are 37.7% and 49.4% higher than SNS-Grasp. Moreover, these methods exhibit significantly poorer controllability, with Auc. scores of only 67.31% and 68.96%. These results highlight SNS-Grasps ability to maintain both physical feasibility and semantic controllability when handling out-of-domain objects.

Ablation Study

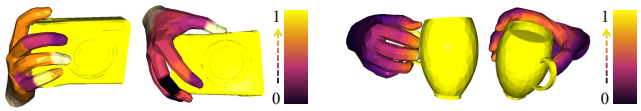
For a more in-depth analysis of our SNS-Grasp, we further conduct a series of ablation studies on the OakInk dataset. Table 2 provides a systematic comparison of three distinct noise scaling strategies in our SNS-Diff. The baseline diffusion generation (without SNS) achieves 0.608 cm penetration depth, 6.000 cm^3 interaction volume, and 1.141 cm displacement distance, with 93.38% controllability.

The three SNS variants demonstrate different behaviors under varying joint-specific noise. SNS^1 uses solely the classifier gradient norm $|\partial \mathcal{L}_{int}/\partial \theta_j|_2$, showing moderate improvements (Pen. 0.592 cm, -2.6%) but increased displacement (+9.2%), suggesting gradient information alone may disrupt physical stability. SNS^2 implements $x_{j,t}^{\text{SNS}} = x_{t,j} \cdot sigmoid(\gamma_j)$, yielding higher noise weights for semantically critical joints. Despite improving physical metrics (Pen. 0.589 cm) and accuracy (96.48%), the suboptimal performance confirms the need for noise reduction at critical joints. SNS^3 represents our full implementation, demonstrating superior performance with 6.5% lower penetration (0.568 cm), 19.7% reduced interaction volume (4.820 cm^3), and highest accuracy (97.74%), validating our approach of reducing noise for semantically critical joints.

Component	Physical Feasibility			Control.
	Pen.↓	Ins.↓	Dis.↓	Auc.↑
Generation	0.608	6.000	1.141	93.38%
SNS^1	0.592	5.260	1.246	94.70%
SNS^2	0.589	5.239	1.152	96.48%
SNS^3	0.568	4.820	1.154	97.74%
Grab_Refine	0.454	3.462	2.273	95.47%
ProKAN	**0.444**	3.308	1.699	95.60%
FGR	0.446	**3.306**	**1.411**	**97.39%**

Table 2: Performance contributions of each component in the proposed SNS-Grasp on the OakInk dataset. "Generation" refers to using only diffusion without SNS. $SNS^{1/2/3}$ indicates different semantic-guided noise scaling techniques. "Grab_Refine" denotes the refinement network used by GrabNet (Yang et al. 2022), and "ProKAN" represents the refinement used by RAGG (Tang et al. 2025).

The refinement stage comparisons in Table 2 show that existing methods like Grab_Refine (Yang et al. 2022) (Pen. 0.454 cm, Ins. 3.462 cm^3) and ProKAN (Tang et al. 2025) (Pen. 0.443 cm, Ins. 3.308 cm^3) are limited by their fixed skinning weights. While effective for reducing penetration depth, they fail to capture the global hand-object spatial context, leading to great displacement distance and semantic inconsistencies during refinement (Auc.: 95.47% and 95.60% respectively). Our FGR overcomes this limitation through dynamic joint-vertex coupling, iteratively optimizing coarse



(a) (b)

Figure 3: Visualizations of joint-specific gradients and attention patterns in SNS-Grasp. (a) Normalized gradient magnitudes across all hand joints for use/hold camera tasks. (b) Cross-attention maps between wrist joints and hand mesh vertices for use/hold mug tasks.

grasps to: (1) maintain competitive penetration metrics (Pen. 0.446 cm, Ins. 3.306 cm^3); (2) achieve 18.2% lower displacement (1.411 cm) than Grab_Refine; and (3) preserve hand-object semantic integrity (Auc. 97.39%). The significant improvements confirm FGR's capacity to balance physical constraints with semantic preservation, outperforming existing refinement approaches.

Qualitative Analysis

In this section, we validate SNS-Grasp through four visual analyses: (1) Noise scaling patterns in SNS modules, revealing joint-specific semantic sensitivity; (2) Cross-attention weights between wrist joints and vertex SDFs, demonstrating joint-vertex coupling dynamics; (3) t-SNE visualization of grasp distributions, showing more compact intra-class clustering and sharper inter-class separation; and (4) Rendered grasp comparisons, providing intuitive quality assessment.

Figure 3(a) visualizes SNS noise scaling patterns through intent recognizer joint gradients. For the use camera interaction, the recognizer shows higher gradient magnitudes at the index and ring fingertips, indicating their dominant role in manipulation recognition, whereas for hold camera, the thumb fingertip exhibits significantly larger gradients, establishing it as the primary determinant. Leveraging these semantic gradients, we construct the SNS module to generate joint-specific and task-aware anisotropic noise profiles.

Figure 3(b) visualizes the joint-to-vertex attention maps in FGR. For the use mug interaction, stronger attention weights concentrate on finger-contact vertices. For the hold mug interaction, attention distributes broadly over the palmar region. These distinct patterns demonstrate FGR's dynamic joint-vertex coupling.

Figure 4 illustrates the t-SNE visualization of grasp distributions generated by SNS-Grasp. The results exhibit two key characteristics: (1) compact intra-class clustering, indicating that grasps with the same manipulation semantic converge to highly consistent configurations, and (2) clear inter-class separation, confirming distinct embedding distributions for different manipulation contexts. This structured organization in the latent space reflects the semantic consistency enforced by SNS, which selectively reduces pose variability for semantically critical joints while retaining natural diversity in less constrained regions. Qualitative comparisons with baseline methods (Yang et al. 2022; Tang et al. 2025) are provided in the Appendix.

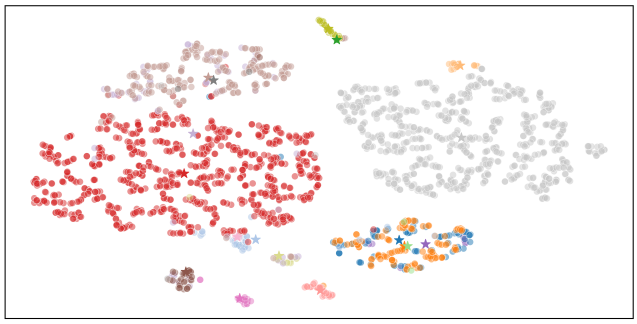


Figure 4: t-SNE visualization of grasp samples generated by SNS-Grasp. Different colors represent distinct interaction contexts, with star markers indicating class centroids.

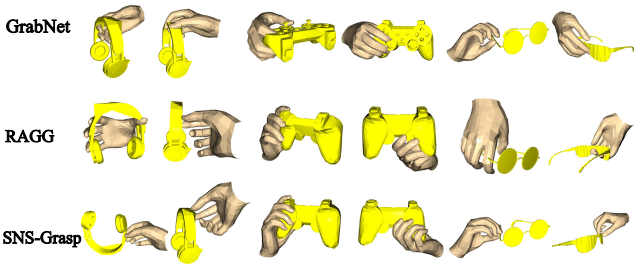


Figure 5: Examples of grasps generated by GrabNet, RAGG, and SNS-Grasp. Each object category is tested under both use and hold intents.

Figure 5 compares intent-based grasp generation by GrabNet (Yang et al. 2022), RAGG (Tang et al. 2025), and our SNS-Grasp on randomly selected objects (headphones, game controllers, eyeglasses) from OakInk. For seen categories (headphones, game controllers), GrabNet struggles to distinguish "hold" vs. "use" intents, yielding similar grasps. Both RAGG and SNS-Grasp achieve better semantic alignment, but RAGG exhibits minor penetration and non-contact failures. For unseen objects (e.g., eyeglasses), RAGG suffers from severe mesh penetration due to reference misalignment, while SNS-Grasp maintains superior performance in both semantic understanding and physical feasibility.

Conclusion

We present SNS-Grasp, a novel framework for intent-based grasp generation that integrates two key innovations: (1) SNS-Diff pioneers anisotropic noise scaling guided by joint-specific and task-aware criticality, enabling precise control over manipulation semantics while maintaining natural kinematic diversity; (2) FGR overcomes MANO's limitations via dynamic joint-vertex coupling, ensuring contact feasibility without compromising semantic alignment. Extensive experiments on OakInk and GRAB datasets demonstrate that SNS-Grasp outperforms state-of-the-art methods, achieving superior performance in semantic accuracy, physical feasibility, and generalization to unseen objects.

Acknowledgments

This work was supported in part by the Science and Technology Development Fund, Macau SAR, under Grant 0193/2023/RIA3 and 0079/2025/AFJ, and the University of Macau under Grant MYRG-GRG2024-00065-FST-UMDF.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Brahmbhatt, S.; Handa, A.; Hays, J.; and Fox, D. 2019. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *IROS*.

Brahmbhatt, S.; Tang, C.; Twigg, C. D.; Kemp, C. C.; and Hays, J. 2020. ContactPose: A dataset of grasps with object contact and hand pose. In *ECCV*.

Chang, X.; and Sun, Y. 2024. Text2Grasp: Grasp synthesis by text prompts of object grasping parts. *arXiv preprint arXiv:2404.15189*.

Christen, S.; Hampali, S.; Sener, F.; Remelli, E.; Hodan, T.; Sauser, E.; Ma, S.; and Tekin, B. 2024. DiffH2O: Diffusion-Based Synthesis of Hand-Object Interactions from Textual Descriptions. *arXiv preprint arXiv:2403.17827*.

Corona, E.; Pumarola, A.; Alenya, G.; Moreno-Noguer, F.; and Rogez, G. 2020. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*.

Guo, X.; Liu, J.; Cui, M.; Li, J.; Yang, H.; and Huang, D. 2024. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *CVPR*.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *CVPR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*.

Höll, M.; Oberweger, M.; Arth, C.; and Lepetit, V. 2018. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. In *VR*.

Jian, J.; Liu, X.; Chen, Z.; Li, M.; Liu, J.; and Hu, R. 2025. G-DexGrasp: Generalizable Dexterous Grasping Synthesis Via Part-Aware Prior Retrieval and Prior-Assisted Generation. *arXiv preprint arXiv:2503.19457*.

Jian, J.; Liu, X.; Li, M.; Hu, R.; and Liu, J. 2023. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *ICCV*.

Jiang, H.; Liu, S.; Wang, J.; and Wang, X. 2021. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*.

Li, H.; Lin, X.; Zhou, Y.; Li, X.; Huo, Y.; Chen, J.; and Ye, Q. 2022. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. *arXiv preprint arXiv:2210.09245*.

Li, K.; Wang, J.; Yang, L.; Lu, C.; and Dai, B. 2024. SemGrasp: Semantic Grasp Generation via Language Aligned Discretization. *arXiv preprint arXiv:2404.03590*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *NeuralIPS*.

Liu, M.; Pan, Z.; Xu, K.; Ganguly, K.; and Manocha, D. 2019. Generating grasp poses for a high-dof gripper using neural networks. In *IROS*.

Liu, M.; Pan, Z.; Xu, K.; Ganguly, K.; and Manocha, D. 2020. Deep differentiable grasp planner for high-dof grippers. *arXiv preprint arXiv:2002.01530*.

Liu, S.; Zhou, Y.; Yang, J.; Gupta, S.; and Wang, S. 2023b. ContactGen: Generative Contact Modeling for Grasp Generation. In *ICCV*.

Liu, X.; and Yi, L. 2024. GeneOH Diffusion: Towards Generalizable Hand-Object Interaction Denoising via Denoising Diffusion. *arXiv preprint arXiv:2402.14810*.

Liu, Y.; Chen, W.; Bai, Y.; Liang, X.; Li, G.; Gao, W.; and Lin, L. 2024. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.

Lu, J.; Kang, H.; Li, H.; Liu, B.; Yang, Y.; Huang, Q.; and Hua, G. 2023. UGG: Unified Generative Grasping. *arXiv preprint arXiv:2311.16917*.

Miller, A. T.; and Allen, P. K. 2004. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4).

Pollard, N. S.; and Zordan, V. B. 2005. Physically based grasping control from example. In *SIGGRAPH*.

Prokudin, S.; Lassner, C.; and Romero, J. 2019. Efficient learning on point clouds with basis point sets. In *ICCV*.

Qi, Z.; Bai, L.; Xiong, H.; and Xie, Z. 2024. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*.

Romero, J.; Tzionas, D.; and Black, M. J. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*.

Samuel, D.; Ben-Ari, R.; Darshan, N.; Maron, H.; and Chechik, G. 2023. Norm-guided latent space exploration for text-to-image generation. *NeuralIPS*, 36: 57863–57875.

Taheri, O.; Choutas, V.; Black, M. J.; and Tzionas, D. 2022. GOAL: Generating 4D whole-body motion for hand-object grasping. In *CVPR*.

Taheri, O.; Ghorbani, N.; Black, M. J.; and Tzionas, D. 2020. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*.

Tang, Z.; Zhu, B.; Hao, Y.; Ngo, C.-W.; and Hong, R. 2025. RAGG: Retrieval-Augmented Grasp Generation Model. In *AAAI*.

Wang, R.; Zhang, J.; Chen, J.; Xu, Y.; Li, P.; Liu, T.; and Wang, H. 2023. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *ICRA*.

Wei, Y.-L.; Jiang, J.-J.; Xing, C.; Tan, X.-T.; Wu, X.-M.; Li, H.; Cutkosky, M.; and Zheng, W.-S. 2024. Grasp as you say: Language-guided dexterous grasp generation. *arXiv preprint arXiv:2405.19291*.

Weng, Z.; Lu, H.; Kragic, D.; and Lundell, J. 2024. DexDiffuser: Generating Dexterous Grasps with Diffusion Models. *arXiv preprint arXiv:2402.02989*.

Wu, M.-Y.; Ting, P.-W.; Tang, Y.-H.; Chou, E.-T.; and Fu, L.-C. 2020. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *Journal of Visual Communication and Image Representation*, 70: 102802.

Wu, W.; Shi, Y.; Chen, Z.; and Cai, Z. 2025. PartDexTOG: Generating Dexterous Task-Oriented Grasping via Language-driven Part Analysis. *arXiv preprint arXiv:2505.12294*.

Yang, L.; Li, K.; Zhan, X.; Wu, F.; Xu, A.; Liu, L.; and Lu, C. 2022. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*.

Yang, L.; Zhan, X.; Li, K.; Xu, W.; Li, J.; and Lu, C. 2021. Cpf: Learning a contact potential field to model the hand-object interaction. In *ICCV*.

Ye, Y.; Gupta, A.; Kitani, K.; and Tulsiani, S. 2024. G-HOP: Generative Hand-Object Prior for Interaction Reconstruction and Grasp Synthesis. In *CVPR*.

Zhang, L.; You, W.; Shi, K.; and Gu, S. 2025. Uncertainty-guided Perturbation for Image Super-Resolution Diffusion Model. In *CVPR*.