Fixed Reality, Diffused Possibility: Disentangling Stochastic and Deterministic Latent for Cluttered Grasping

Hritam Basak and Zhaozheng Yin

Stony Brook University, NY, USA
hritambasak48@gmail.com

Abstract. Recent diffusion-based generative models have shown strong capabilities in 3D scene understanding and affordance reasoning. However, existing methods predominantly rely on deterministic global encoders or fully stochastic diffusion backbones, which fail to disentangle structural certainty from contextual uncertainty, a key requirement for reliable affordance-grasp prediction in cluttered, partially observed environments. We propose a split-latent hierarchical diffusion framework that explicitly decomposes the global scene representation into two complementary subspaces: a *fixed latent* encoding deterministic geometric and semantic priors, and a *diffused latent* capturing stochastic scene-level variations through a Global Conditional Diffusion guided by RGB context. To propagate global uncertainty to fine-grained reasoning, we introduce a Local Conditional Diffusion that predicts dense, point-wise affordance and grasp fields conditioned jointly on both latents, producing geometrically precise yet contextually diverse predictions. Furthermore, we incorporate Harmonic Grasp Field, a differential geometric regularization that enforces smooth, obstacle-aware grasp manifolds via harmonic potential constraints. This unified formulation bridges deterministic geometry with probabilistic reasoning, yielding physically consistent and uncertainty-aware grasp generation. Extensive experiments on cluttered 3D affordance and grasping benchmarks demonstrate superior stability, generalization, and diversity compared to deterministic and single-latent diffusion approaches. Project Website.

Keywords: Diffusion · Cluttered Grasping · Robotics

1 Introduction

Dexterous manipulation in unstructured 3D environments, particularly in densely cluttered scenes, remains a fundamental challenge for embodied intelligence. Unlike conventional pick-and-place tasks, dexterous grasping demands a precise understanding of intricate object interactions, high-dimensional control, and strategic long-horizon reasoning. These challenges are exacerbated by severe occlusions, physical interlocking, and ambiguous spatial configurations, as documented in recent benchmarks [1]. In such cases, a successful grasp often requires

a sequence of coordinated actions that combine *singulation* (non-prehensile rearrangement) with final grasp execution [44].

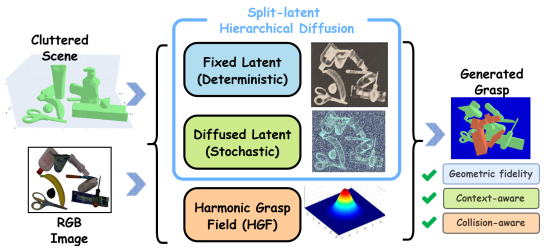

Fig. 1: Our proposed split-latent cluttered grasping framework.

This complex interplay introduces two problems: (i) **scene-level reasoning**: deciding *which* object to act upon, and (ii) **grasp generation**: determining *how* to grasp it robustly in cluttered, partially observed environments. While prior works have explored high-level decision-making through reinforcement learning or vision-language models (VLMs) [26, 44], we aim to solve these problems by developing a geometry- and uncertainty-aware grasp generation framework that can infer stable, collision-free grasps directly from noisy 3D observations without explicit planning. Specifically, we aim to model both global structural priors and local contextual variability to achieve physically consistent and generalizable grasp predictions in cluttered, real-world settings. On the other end, grasp generation methods attempt to directly model the distribution of valid grasps. Regression-based approaches collapse under multimodality, converging to invalid average poses [44], while recent diffusion-based models [46] improve diversity but remain *myopic* (i.e., they condition primarily on local geometry without reasoning about global scene structure or task-level context). As a result, they often produce locally feasible yet globally incoherent grasps.

The core limitation of existing frameworks originates from their inability to **disentangle structural certainty from contextual uncertainty**. Within a scene, certain facts are geometrically deterministic (e.g., "this is a handle"), whereas others are inherently stochastic and strategy-dependent (e.g., "should the mug or the block be moved first?"). Current encoders either conflate these aspects within a single latent space, losing interpretability and robustness to context variations, or model them uniformly as noise, destroying geometric consistency. This entanglement limits both the controllability of grasp generation under uncertainty and the geometric stability.

We introduce a hierarchical diffusion framework that explicitly decomposes the scene representation into two complementary latent subspaces: **Fixed latent** (h_{fixed}): a deterministic embedding capturing invariant geometric and semantic priors from the observed point cloud; **Diffused latent** (h_{diffused}): a stochastic embedding generated via a **Global Conditional Diffusion** from the observed RGB image, This disentangled latent architecture (as shown in Figure 1) enables the model preserve geometric stability through h_{fixed} while sampling over plausible strategic contexts through h_{diffused}, bridging deterministic understanding with probabilistic reasoning. To propagate global uncertainty into fine-grained

prediction, we propose a **Local Conditional Diffusion** that predicts dense point-wise affordance and grasp fields conditioned jointly on $[h_{\text{fixed}}, h_{\text{diffused}}]$. Unlike previous diffusion models, our local conditional diffusion couples global stochastic reasoning with local geometric consistency, yielding grasp predictions that are both contextually coherent and geometrically precise. Furthermore, to impose physical smoothness and geometric stability, we integrate **Harmonic Grasp Fields (HGF)**, a differential geometric formulation where harmonic potentials define continuous, obstacle-aware grasp manifolds on object surfaces. Our key contributions are:

- **Split-latent Hierarchical Diffusion:** a hierarchical architecture that disentangles deterministic geometric priors from stochastic strategic uncertainty for robust reasoning in cluttered 3D scenes.
- **Global and Local Conditional Diffusion:** a novel conditioning mechanism that fuses global stochastic context with local point-wise geometric features to produce uncertainty-aware grasp fields.
- **Integration with Harmonic Grasp Fields (HGF):** a principled coupling of probabilistic diffusion and differential geometry that yields smooth, continuous, and obstacle-aware dexterous grasp manifolds.
- Comprehensive evaluations across multiple cluttered scene benchmarks validate the effectiveness, robustness, and generalizability of our framework.

2 Related Works

2.1 Grasp Detection and 6-DoF Grasp Estimation

Grasp detection has progressed from geometric heuristics to deep learning-based formulations. Early methods like GQ-CNN [22] established end-to-end learning of grasp quality from RGB-D data, while [5] achieved real-time performance with lightweight architectures. The shift to 6-DoF grasping began with GPD [35], which scored sampled grasp candidates using deep features, and PointGPD [19], which directly processed point clouds via PointNet-style encoders. The large-scale GraspNet-1B dataset [6] enabled data-driven 6-DoF grasp learning, while GSNet [37] and follow-ups like [27] leveraged graph and transformer backbones for richer spatial reasoning. Recent approaches such as [24] and ContactGraspNet [32] explored self-supervised and contact-aware modeling. Yet, most methods remain deterministic, failing to capture multimodal (RGB and point-cloud) uncertainty in cluttered, partially observable scenes.

2.2 Diffusion Models

Diffusion models have become a cornerstone of modern generative modeling. Foundational works: DDPM [9], DDIM [31], and LDM [30] established efficient formulations, while classifier-free guidance [10] improved controllability. Beyond 2D images, diffusion has advanced 3D generation across shapes [36], point clouds [21], and 6D pose estimation [43]. In robotics, diffusion has been

applied to modeling end-effector affordances [39], motion planning [2], and conditional spatial prediction [33]. Our method extends these efforts through a split-latent hierarchical diffusion design that disentangles deterministic geometry from stochastic context, enabling stable yet uncertainty-aware affordance generation.

2.3 Cluttered Scene Grasping

Grasping in dense, occluded scenes remains a core challenge in embodied AI due to complex inter-object interactions and uncertain visibility. Early heuristic and sampling-based methods struggle with robustness and scalability. Deep learning frameworks like GPD [17] and PointGPD [19] have improved grasp scoring from 3D data, while GraspNet-1B [6] has enabled large-scale learning of diverse grasps. Recent architectures, including GSNet [37] and S4G [27], combine global context and spatial reasoning to better handle clutter. Masked pretraining [28] and contact-based reasoning [32] further improve grasp stability. Other notable efforts include solving the problem using VLM [26], diffusion models [38, 46], policy learning [48], etc.

However, existing approaches typically predict a single deterministic grasp and overlook scene-level uncertainty. Emerging generative frameworks, particularly diffusion-based models, begin to capture multimodal grasp distributions but often entangle geometric priors with contextual variability. Our split-latent diffusion formulation explicitly separates these components, yielding coherent, uncertainty-aware grasp synthesis in cluttered environments.

3 Proposed Method

To achieve robust grasp poses in cluttered 3D environments, we introduce a *hierarchical* (global+local) diffusion framework that disentangles stable geometric priors from variable contextual cues, enabling both global coherence and local precision in grasp reasoning. As illustrated in Figure 2, our approach begins with a multi-view pre-trained global encoder (subsection 3.1), extracting fixed global feature. This, along with stochastic global diffused feature from Global Conditional Diffusion (GCD) form the split-latent global scene representation (subsection 3.2). The global diffused latent conditions the Local Conditional Diffusion (LCD) (subsection 3.3) to model local latent, which is further smoothened using Harmonic Grasp Field (HGF) (subsection 3.4). Together, these modules enable coherent, uncertainty-aware grasp prediction (subsection 3.5) that generalizes across cluttered and unseen environments.

3.1 Multi-view Encoder Pre-training

To endow the global encoder with robust, view-consistent geometric priors, we adopt a **multi-view encoder pretraining** strategy inspired by masked autoencoding, directly in 3D space unlike using projected images [25]. Given a N-point point cloud or partial scan $\mathbf{X} \in \mathbb{R}^{N\times 3}$, we sample K depth views $\{\mathbf{X}_k\}_{k=1}^{K}$ from

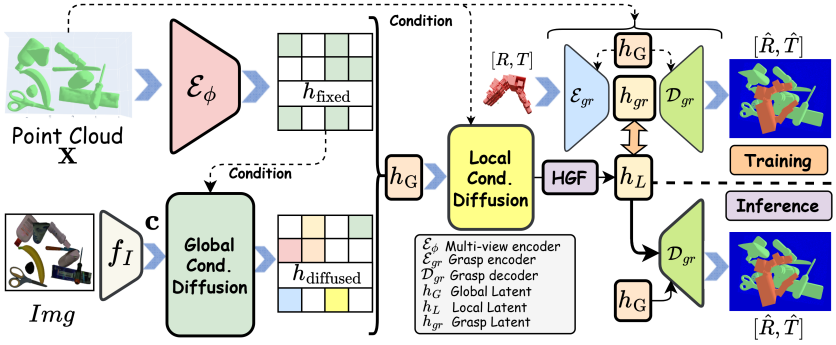

Fig. 2: Overview of the proposed method. A multi-view pre-trained encoder $\mathcal{E}_\phi$ extracts a fixed latent $\mathbf{h}_{\text{fixed}}$ capturing stable geometric and semantic priors, and a diffused latent $\mathbf{h}_{\text{diffused}}$ modeling contextual uncertainty is extracted via a Global Conditional Diffusion process guided by RGB input and h_{fixed}. The combined global latent $\mathbf{h}_G = [\mathbf{h}_{\text{fixed}}; \mathbf{h}_{\text{diffused}}]$ conditions the Local Codnitional Diffusion to generate local latent h_L through Harmonic Grasp Field (HGF) that ensures spatial smoothing. The global and local diffusions together are coined as *hierarchical diffusion*. During training, a grasp encoder $\mathcal{E}_{gr}$ encodes the grasp pose $[R, T]$ into latent h_{gr}, aligned with h_L, which are then decoded using grasp decoder $\mathcal{D}_{gr}$ to generate the grasp output $[\hat{R}, \hat{T}]$. During inference, $\mathcal{D}_{gr}$ decodes $[h_L, h_G]$ into grasp prediction.

uniformly distributed virtual camera poses $\{p_k\}_{k=1}^K$. Each view corresponds to point cloud capturing visible geometry under self-occlusion.

Each partial view $\mathbf{X}_k$ is divided into M local patches $\{\mathbf{x}_{k,j}\}_{j=1}^M$ and encoded using a point transformer encoder $\mathcal{E}_\phi$, yielding per-view latent embeddings $\mathbf{z}_k = \mathcal{E}_\phi(\mathbf{X}_k) \in \mathbb{R}^{M \times d}$. A random subset (50–70%) of patches is masked, and the model reconstructs the missing regions through a lightweight point decoder $\mathcal{D}_\phi$:

$$\hat{\mathbf{X}}_k = \mathcal{D}_\phi(\mathcal{E}_\phi(\mathbf{X}_k^{\text{masked}})). \tag{1}$$

The reconstruction is supervised using the *Chamfer distance (CD)* with respect to the original point sets:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{K} \sum_{k=1}^{K} \text{CD}(\hat{\mathbf{X}}_k, \mathbf{X}_k), \tag{2}$$

where the Chamfer distance is defined as

$$\text{CD}(\hat{\mathbf{X}}_k, \mathbf{X}_k) = \frac{1}{|\hat{\mathbf{X}}_k|} \sum_{\hat{\mathbf{x}}_i \in \hat{\mathbf{X}}_k} \min_{\mathbf{x}_j \in \mathbf{X}_k} \|\hat{\mathbf{x}}_i - \mathbf{x}_j\|_2^2 + \frac{1}{|\mathbf{X}_k|} \sum_{\mathbf{x}_j \in \mathbf{X}_k} \min_{\hat{\mathbf{x}}_i \in \hat{\mathbf{X}}_k} \|\mathbf{x}_j - \hat{\mathbf{x}}_i\|_2^2. \tag{3}$$

This encourages the network to reconstruct spatially complete geometry from partial observations while maintaining geometric smoothness and fine-grained

details. To enhance viewpoint invariance, we also apply a *latent contrastive alignment* loss across views of the same object:

$$\mathcal{L}_{\text{align}} = -\frac{1}{K}\sum_{i\neq j}\log\frac{\exp(\text{sim}(\mathbf{z}_i,\mathbf{z}_j)/\tau)}{\sum_l \exp(\text{sim}(\mathbf{z}_i,\mathbf{z}_l)/\tau)}, \tag{4}$$

where $\text{sim}(\cdot)$ denotes cosine similarity and τ is the temperature coefficient. This enforces that latent embeddings z from different viewpoints $\{i, j\}$ of the same object remain consistent, facilitating cross-view feature consistency.

The pretrained encoder $\mathcal{E}_\phi$ thus learns occlusion-tolerant, geometry-consistent 3D information, enabling accurate reasoning under partial visibility. After pretraining, $\mathcal{E}_\phi$ is used as *global encoder* used in our split-latent framework (subsection 3.2), by extracting deterministic global feature ($\mathbf{h}_{\text{fixed}}$) that conditions the stochastic latent ($\mathbf{h}_{\text{diffused}}$) generation through global conditional diffusion.

3.2 Split-Latent Hierarchical Diffusion

Conventional diffusion models often conflate stable geometric priors with stochastic contextual variations, resulting in limited controllability and poor generalization in cluttered 3D scenes. In contrast, we introduce a **Split-Latent Hierarchical Diffusion** framework that explicitly disentangles deterministic geometric structure from stochastic scene context. Specifically, the global scene representation is decomposed into two complementary subspaces: a *fixed latent* that encodes invariant geometric and semantic priors, and a *diffused latent* that models stochastic global variations through a conditional diffusion process guided by RGB observations. Given an input point cloud $\mathbf{X} \in \mathbb{R}^{N\times 3}$ and corresponding image features $\mathbf{c} = f_I(\text{Img})$, the global encoder $\mathcal{E}_\phi$ first produces the fixed latent embedding $\mathbf{h}_{\text{fixed}} = \mathcal{E}_\phi(\mathbf{X})$, which captures the deterministic, geometry-aware structure of the scene. Then, a stochastic latent $\mathbf{h}_{\text{diffused}}$, representing the uncertain and context-dependent aspects of the scene, is generated through a Global Conditional Diffusion (GCD) process conditioned on both $\mathbf{h}_{\text{fixed}}$ and $\mathbf{c}$. The two components are concatenated to form the composite global latent $\mathbf{h}_G = [\mathbf{h}_{\text{fixed}}; \mathbf{h}_{\text{diffused}}]$, which jointly encodes stable structure and probabilistic variability. This composite latent then conditions the Local Conditional Diffusion (LCD), which performs fine-grained, point-level reasoning to model dense affordance and grasp fields under uncertainty. Together, GCD and LCD constitute a hierarchical diffusion architecture that seamlessly bridges global stochastic reasoning with local geometric precision, enabling robust and uncertainty-aware grasp synthesis in complex, cluttered environments.

3.3 Global and Local Conditional Diffusion

As explained in subsection 3.2, the **Global Conditional Diffusion (GCD)** encodes the stochastic global context, conditioned on $\mathbf{h}_{\text{fixed}}$ and c. During training, an initial $\mathbf{h}^0_{\text{diffused}}$ is sampled from a Gaussian distribution whose parameters

are derived from existing scene feature:

$$\mathbf{h}^0_{\text{diffused}} \sim q(\mathbf{h}^0_{\text{diffused}}|\mathbf{h}_{\text{fixed}}, \mathbf{c}) = \mathcal{N}(\mu(\mathbf{h}_{\text{fixed}}, \mathbf{c}), \sigma^2(\mathbf{h}_{\text{fixed}}, \mathbf{c})\mathbf{I}), \tag{5}$$

where $\mu(\mathbf{h}_{\text{fixed}}, \mathbf{c})$ and $\sigma^2(\mathbf{h}_{\text{fixed}}, \mathbf{c})$ are learned scalars derived from $(\mathbf{h}_{\text{fixed}}, \mathbf{c})$. They determine the mean and variance of the stochastic latent, centering it around scene-specific global context while allowing controlled variability. This ensures that $\mathbf{h}^0_{\text{diffused}}$ represents meaningful, scene-dependent stochastic information. The forward process gradually perturbs the stochastic latent for T steps[1]:

$$\mathbf{h}^t = \sqrt{\bar{\alpha}_t}\,\mathbf{h}^0_{\text{diffused}} + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{6}$$

where $\alpha_t = 1 - \beta_t$, β_t is the variance schedule, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. In the reverse process, denoising (U-Net like) network ϵ_θ predicts the noise at each step conditioned on $\mathbf{h}_{\text{fixed}}$ and RGB feature c:

$$p_\theta(\mathbf{h}^{t-1}|\mathbf{h}^t, \mathbf{h}_{\text{fixed}}, \mathbf{c}) = \mathcal{N}\Big(\mu_\theta(\mathbf{h}^t, \mathbf{h}_{\text{fixed}}, \mathbf{c}, t), \sigma_t^2\mathbf{I}\Big), \tag{7}$$

with mean parameterized as

$$\mu_\theta(\mathbf{h}^t, \mathbf{h}_{\text{fixed}}, \mathbf{c}, t) = \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{h}^t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{h}^t, \mathbf{h}_{\text{fixed}}, \mathbf{c}, t)\Big). \tag{8}$$

The diffusion objective is

$$\mathcal{L}_{\text{GCD}} = \mathbb{E}_{t, \mathbf{h}^0_{\text{diffused}}, \boldsymbol{\epsilon}^t}\Big[\big\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{h}^t, \mathbf{h}_{\text{fixed}}, \mathbf{c}, t)\big\|_2^2\Big]. \tag{9}$$

During inference, $\mathbf{h}^T_{\text{diffused}} \sim \mathcal{N}(0, \mathbf{I})$ is iteratively denoised conditioned on $\mathbf{h}_{\text{fixed}}$ and $\mathbf{c}$ to produce $\mathbf{h}^0_{\text{diffused}}$. The final global latent $\mathbf{h}_G = [\mathbf{h}_{\text{fixed}}; \mathbf{h}^0_{\text{diffused}}]$ serves as a scene-informed prior for LCD.

We introduce the **Local Conditional Diffusion (LCD)**, a point-level diffusion process generating the local latent manifold $\mathbf{h}_L$ conditioned on the global latent $\mathbf{h}_G$. Given $\mathbf{h}_G$ and the input point cloud $\mathbf{X}$, it models a time-indexed local latent field $\mathbf{h}^t$:

$$q(\mathbf{h}^t|\mathbf{h}^{t-1}, \mathbf{h}_G) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{h}^{t-1}, \beta_t\mathbf{I}), \mathbf{h}^t = \sqrt{\bar{\alpha}_t}\mathbf{h}_L + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}^t, \quad \boldsymbol{\epsilon}^t \sim \mathcal{N}(0, \mathbf{I}), \tag{10}$$

The reverse process denoises $\mathbf{h}^t$ conditioned on $\mathbf{h}_G$ and timestep t:

$$\begin{aligned} p_\psi(\mathbf{h}^{t-1}|\mathbf{h}^t, \mathbf{h}_G) &= \mathcal{N}\big(\mu_\psi(\mathbf{h}^t, \mathbf{h}_G, t), \sigma_t^2\mathbf{I}\big), \\ \mu_\psi(\mathbf{h}^t, \mathbf{h}_G, t) &= \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{h}^t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\psi(\mathbf{h}^t, \mathbf{h}_G, t)\Big), \end{aligned} \tag{11}$$

where ϵ_ψ is a U-Net–like network with cross-attention injecting global latent features into local updates. Each local latent is updated via spatial conditioning f_ψ on point coordinates and global context:

$$\epsilon_\psi(\mathbf{h}^t, \mathbf{h}_G, t) = f_\psi\Big([\mathbf{h}^t, \text{PE}(t)], \text{CrossAttn}(\mathbf{h}^t, \mathbf{h}_G)\Big), \tag{12}$$

[1] $\mathbf{h}^t_{\text{diffused}}$ is represented as $\mathbf{h}^t$ for simplicity

where PE(t) encodes the diffusion timestep and CrossAttn($\cdot$) aligns per-point features with the global latent. The conditional diffusion loss is:

$$\mathcal{L}_{\text{LCD}} = \mathbb{E}_{t,\mathbf{h}_L,\boldsymbol{\epsilon}^t}\left[\left\|\boldsymbol{\epsilon}^t - \epsilon_\psi(\mathbf{h}^t, \mathbf{h}_G, t)\right\|_2^2\right]. \tag{13}$$

After denoising, $\mathbf{h}_L$ is decoded to generate grasps (see subsection 3.5).

3.4 Harmonic Grasp Field

While LCD produces spatially coherent grasp hypotheses, these predictions may still exhibit discontinuities or local instability under clutter and contact noise. To ensure smooth, geometry-consistent grasp manifolds, we introduce the **Harmonic Grasp Field (HGF)**, a continuous potential field over the object surface that regularizes grasp poses via the intrinsic properties of harmonic functions. Let the object surface be represented by a set of points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$. We define a grasp potential function Φ whose extrema correspond to graspable configurations. The potential field $\Phi(\mathbf{x})$ is constrained to be *harmonic* within the object region:

$$\nabla^2\Phi(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega, \tag{14}$$

subject to Dirichlet boundary conditions:

$$\Phi(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in \partial\Omega_o \text{ (obstacle or background)}, \\ 1, & \mathbf{x} \in \partial\Omega_g \text{ (grasp contact surface)}, \end{cases} \tag{15}$$

where Ω denotes the object volume, $\partial\Omega_o$ the obstacle boundary, and $\partial\Omega_g$ the contact boundary estimated from the local diffusion prediction. The resulting field defines smooth potential gradients that naturally guide end-effector trajectories toward stable, collision-free grasp regions. To couple the harmonic potential with the latent diffusion process, each predicted local latent vector $\mathbf{h}_{L,i}$ for point x_i is projected into harmonic-consistent space:

$$\mathbf{h}'_{L,i} = \mathbf{h}_{L,i} + \lambda_h \nabla\Phi(\mathbf{x}_i), \tag{16}$$

where $\nabla\Phi(\mathbf{x}_i)$ denotes the local harmonic gradient encoding grasp flow direction, and λ_h controls the strength of field alignment. This augmentation ensures that the local diffusion features follow smooth geometric trajectories defined by the harmonic manifold rather than arbitrary latent variations. The harmonic regularization loss is:

$$\mathcal{L}_{\text{harm}} = \mathbb{E}_{\mathbf{X}}\left[\,|\nabla^2\Phi(\mathbf{x})|_2^2\,\right] + \gamma\left[\,|\mathbf{h}'_L - \mathbf{h}_L|_2^2\,\right], \tag{17}$$

which penalizes non-harmonic curvature in the potential and enforces alignment between diffused local latents and the harmonic field gradients.

3.5 Grasp Generation

A pretrained variational encoder $\mathcal{E}_{gr}$ maps the grasp pose $\mathbf{y} = [\mathbf{R}, \mathbf{T}]$ into a latent distribution:

$$q(\mathbf{h}_{gr}|\mathbf{X}, \mathbf{y}) = \mathcal{N}\big(\boldsymbol{\mu}_{gr}(\mathbf{X}, \mathbf{y}), \boldsymbol{\sigma}^2_{gr}(\mathbf{X}, \mathbf{y})\mathbf{I}\big), \tag{18}$$

where $\mathbf{h}_{gr}$ represents the fine-grained local latent space aligned to grasp-relevant regions. A decoder $\mathcal{D}_{gr}$ reconstructs the grasp pose $\hat{\mathbf{y}}$ from the sampled latent $\mathbf{h}_{gr}$ as $\hat{\mathbf{y}} = \mathcal{D}_{gr}(\mathbf{h}_{gr})$. To ensure consistency between the local diffusion latent $\mathbf{h}_L$ and the grasp latent $\mathbf{h}_{gr}$, we adopt an evidence lower bound objective (ELBO):

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{\mathbf{h}_{gr}\sim q(\mathbf{h}_{gr}|\mathbf{X},\mathbf{y})}\big[\log p(\mathbf{y}|\mathbf{h}_{gr}, \mathbf{h}_L)\big] - D_{\text{KL}}\Big(q(\mathbf{h}_{gr}|\mathbf{X}, \mathbf{y}) \,\Big\|\, p(\mathbf{h}_L|\mathbf{h}_G)\Big), \tag{19}$$

where the first term enforces accurate reconstruction of the grasp pose, and the second term regularizes the posterior to align with the diffusion-conditioned prior. Assuming Gaussian distributions, this simplifies to:

$$\mathcal{L}_{\text{ELBO}} \approx \mathbb{E}_q\big[\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2\big] D_{\text{KL}}\Big(\mathcal{N}(\boldsymbol{\mu}_{gr}, \boldsymbol{\sigma}^2_{gr}\mathbf{I}) \,\Big\|\, \mathcal{N}(\boldsymbol{\mu}_L, \boldsymbol{\sigma}^2_L\mathbf{I})\Big), \tag{20}$$

where $(\boldsymbol{\mu}_L, \boldsymbol{\sigma}_L)$ are the predicted mean and variance from the local diffusion latent $\mathbf{h}_L$. Finally, the decoded grasp field is obtained as $\hat{\mathbf{y}}_i = \mathcal{D}_{gr}(\mathbf{h}_{gr,i})$, where $\hat{\mathbf{y}}_i = [\hat{\mathbf{R}}_i, \hat{\mathbf{T}}_i]$ denotes the predicted 6-DoF grasp pose.

3.6 Overall Learning Objective

The entire framework is optimized through a staged training procedure that progressively aligns global geometric priors, local diffusion dynamics, and grasp generation objectives. First, the multi-view point encoder $\mathcal{E}_\phi$ is pretrained via masked point reconstruction and contrastive alignment losses, $\mathcal{L}_{\text{MAE}} + \mathcal{L}_{\text{align}}$, enabling it to capture view-consistent and occlusion-tolerant 3D geometry. The pre-trained global encoder $\mathcal{E}_\phi$ is frozen to serve as conditioning priors for the hierarchical diffusion training. In the second training stage, **(i)** the global diffusion module is trained to disentangle deterministic structural priors ($\mathbf{h}_{\text{fixed}}$) from stochastic contextual features ($\mathbf{h}_{\text{diffused}}$), providing uncertainty-aware global representations. **(ii)** LCD is trained to propagate the global uncertainty into fine-grained point-level reasoning to generate coherent local latents $\mathbf{h}_L$. **(iii)** HGF regularization is applied to LCD to enforce geometric smoothness and manifold continuity. **(iv)** A grasp variational autoencoder $(\mathcal{E}_{gr}, \mathcal{D}_{gr})$ is trained, ensuring grasp-consistent manifold formation. The overall learning objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GCD}} + \lambda_1\mathcal{L}_{\text{LCD}} + \lambda_2\mathcal{L}_{\text{harm}} + \lambda_3\mathcal{L}_{\text{ELBO}}, \tag{21}$$

where $\{\lambda_i\}$ are scalar weights controlling the balance between diffusion, and regularization objectives. This two-stage optimization ensures that the model first learns stable 3D priors, then progressively integrates stochastic reasoning and geometric regularization to achieve robust, uncertainty-aware grasp generation in cluttered scenes.

4 Experiments and Results

4.1 Dataset Description

We evaluate our model on four complementary 3D grasping and manipulation datasets across three tasks (dexterous grasp, gripper grasp and policy-rollout). (1) DexGraspNet-2.0 [46] is a large-scale synthetic benchmark for dexterous hand grasping in cluttered scenes, with two subsets: GraspNet-1B and ShapeNet. Together they comprise $1,319$ objects, $8,270$ cluttered scenes, and $\sim 427M$ annotated grasps for a multi-finger hand model. (2) GraspClutter6D [1] is a large-scale real-world dataset for parallel-jaw (gripper) grasping in highly cluttered scenes. It contains $52K$ samples, scanned from $1,000$ real scenes (14.1 objects per scene, $\sim 62.6\%$ occlusion), annotated with $736K$ 6D object poses and $\sim 9.3B$ feasible 6-DoF grasps across 200 objects in 75 environment configurations. (3) GraspNet-1Billion [6] is a large real-world benchmark for general object grasping (grippers) in cluttered tabletop scenes. It includes $97,280$ samples spanning 190 scenes, and over $1.1B$ annotated grasp poses for 88 objects. (4) $\mathbf{A^2}$ Simulation Dataset [41] is a simulation-based task-rollout dataset developed for evaluating pick, place, and pick-n-place policies.

4.2 Implementation Details

Our multi-view encoder $\mathcal{E}_\phi$ is built upon a four-layer Point Transformer [47] backbone, each followed by set abstraction and feature propagation blocks for geometric aggregation. GCD uses a 12-layer residual MLP with sinusoidal timestep embeddings of 128 dimensions, and adaptive layer normalization. LCD employs a 4-scale U-Net with point-voxel convolutions (PVConv) at each stage, propagating a local latent field $\mathbf{h}_L \in \mathbb{R}^{N\times 32}$. Positional encodings for each time step are injected into all diffusion modules. $\mathcal{E}_{gr}, \mathcal{D}_{gr}$ consist of three FC layers, followed by layer normalization and ReLU activation. The two-stage training protocol follows a multi-view encoder pretraining (8k epochs), then final training (24k epochs) using fixed global encoder. Adam optimizer is used throughout, with learning rate 1×10^{-4} for all modules and cosine annealing schedule. $\lambda_h = 0.5$ (Equation 16), $\gamma = 0.05$ (Equation 17), $\{\lambda_1, \lambda_2, \lambda_3\} = \{0.8, 1.0, 0.5\}$ Equation 21 are set following validation. All experiments are performed on 8 NVIDIA V100 GPUs, with VRAM 32GB. We closely follow [46], [1], and [41] for training and evaluation settings. Please refer to **supplementary file** for more description.

4.3 Evaluation of Dexterous Grasping

Table 1 reports quantitative results on DexGraspNet-2.0 under varying clutter densities. Our method consistently surpasses all SoTA approaches across all settings, achieving grasp success rates of 93.1/85.9/76.1 on three GraspNet-1Billion splits and 83.2/87.8/74.9 on ShapeNet, corresponding to significant improvements over the strongest baselines (DexGraspVLA [49] and ClutterDexGrasp [3]). The gains are most pronounced in Dense scenes, where collision-free

Table 1: Quantitative grasp success comparison for cluttered dexterous grasping on the DexGraspNet-2.0 dataset with two subsets: GraspNet-1Billion and ShapeNet. Dense scenes contain 8–11 objects; Random scenes contain 1–10; Loose scenes include 1–2 objects, following [46]. * indicates decomposed pose modeling version. The best and second-best results are highlighted in red and blue.

Method	GraspNet-1Billion			ShapeNet		
	Dense	Random	Loose	Dense	Random	Loose
HGC-Net [18]	46.0	37.8	26.7	46.4	44.8	30.4
DDGC [20]	47.9	38.2	30.4	50.9	48.7	38.5
GraspTTA [13]	62.5	54.1	42.8	56.6	57.8	46.4
ISAGrasp [4]	63.4	60.7	51.4	64.0	56.3	52.7
DexGraspNet-2* [46]	84.2	80.7	71.3	74.9	72.5	66.4
DexGraspNet-2 [46]	90.6	83.7	73.2	81.0	85.4	74.2
DexMGNet [40]	89.5	84.5	74.9	81.4	86.7	73.9
ClutterDexGrasp [3]	91.6	85.1	75.2	80.7	86.9	73.5
DexGraspVLA [49]	91.7	85.0	75.5	81.8	85.4	72.8
Ours	**93.1**	**85.9**	**76.1**	**83.2**	**87.8**	**74.9**

Table 2: Quantitative grasp success comparison of our proposed method for cluttered gripper grasping on the GraspClutter6D benchmark dataset. We report grasp success rate (GSR) and declutter rate (DR) in different cluttered scenes.

Method	Packed		Pile					
	5 Objects		5 Objects		10 Objects		15 Objects	
	GSR	DR	GSR	DR	GSR	DR	GSR	DR
Contact-GraspNet [32]	84.9	86.1	77.6	75.4	77.2	64.7	77.3	54.0
MVGrasp [14]	83.7	84.4	75.8	76.1	76.2	63.9	77.5	54.2
BiGraspFormer [16]	82.9	84.4	77.2	75.4	76.9	64.2	77.1	54.7
Q-Grasp [11]	80.9	84.1	77.5	76.2	76.8	64.5	78.2	54.9
ThinkGrasp [26]	84.9	86.8	78.0	76.2	76.8	65.1	77.9	56.2
AffordGrasp [34]	85.1	86.7	77.6	77.1	76.4	65.2	77.3	55.9
Ours	**86.2**	**87.3**	**78.6**	**78.6**	**77.6**	**66.1**	**79.0**	**56.8**

planning is the hardest, indicating that our harmonic field formulation effectively mitigates grasp-interference in clutter. Unlike prior models that depend on discrete proposal sampling or deterministic regression, our approach maintains performance across all clutter regimes.

Competing methods underperform for complementary reasons: geometry-only pipelines (e.g., HGC-Net [18], DDGC [20]) collapse under occlusion due to limited contextual reasoning; diffusion-free regressors (e.g., DexGraspNet-2 [46]) lack multimodal pose representation; and recent diffusion-based models (DexMGNet [40], DexGraspVLA [49]) still treat collision-feasibility implicitly. In contrast, our Harmonic Grasp Field constructs a continuous, obstacle-aware potential guiding feasible approach trajectories, while a hierarchical latent diffusion backbone captures both deterministic global structure and residual local uncertainty. This coupling of geometric regularity and probabilistic diversity reduces collision-induced failures and enhances stability.

4.4 Evaluation of Gripper Grasping

As shown in Table 2, our method achieves the highest grasp success rate (GSR) and declutter rate (DR) across all clutter configurations of the GraspClutter6D [1] benchmark, with consistent improvements of $0.9 \sim 1.3$ in GSR and $0.5 \sim 1.6$ in DR over the best competing approaches (AffordGrasp [34] and ThinkGrasp [26]), with the largest margins observed under Pile-15 clutter where inter-object occlusion and contact constraints dominate. Existing methods either rely on local geometry priors (e.g., Contact-GraspNet [32], MVGrasp [14]) or discrete affordance sampling (e.g., BiGraspFormer [16], AffordGrasp [34]), leading to unstable performance as clutter density increases.

Table 3: Quantitative comparison for gripper grasping on the GraspNet-1Billion benchmark dataset using RealSense setting. * denotes using grasp scores ≥ 0.9. † means rendered depth images used.

Method	AP Seen	AP Similar	AP Novel
GG-CNN [23]	15.5	13.3	5.5
Chu et al. [5]	16.1	15.4	7.6
GPD [35]	22.9	21.3	8.2
PointGPD [19]	26.0	22.7	9.2
GraspNet1B [6]	27.6	26.1	10.6
GSNet [37]	67.1	54.8	24.3
DexGraspNet-2 [46]	56.0	53.2	23.2
Ours	**68.4**	**55.7**	**25.1**
GSNet* [37]	68.7	59.8	24.6
DexGraspNet-2* [46]	66.6	61.7	27.4
Ours*	**70.8**	**62.5**	**28.9**
DexGraspNet-2† [46]	77.4	72.5	39.1
GSNet† [37]	78.4	69.7	36.7
Ours†	**79.8**	**74.1**	**40.9**

Table 4: Quantitative evaluation of our method on the A^2 simulation dataset. Metrics of pick and pick-n-place are presented as Success Rate ↑ / Planning Steps ↓. Improvements are in red.

Category	Method	Seen	Unseen
Pick	LERF-TOGO [29]	83.3/3/37	76.0/2.01
	GraspSplats [12]	58.0/2.05	37.3/1.67
	VLG [42]	74.3/4.11	78.7/3.98
	ThinkGrasp [20]	84.7/2.55	57.3/4.11
	A^2-G-Pick [41]	83.3/3.78	84.7/3.85
	A^2 [41]	95.3/2.55	97.3/2.57
	Ours+A^2-G-Pick [41]	**86.9/2.97** (4.3%↑/16.4%↓)	**86.6/3.11** (2.4%↑/19.2%↓)
	Ours+A^2	**96.3/2.14** (1.1%↑/16.7%↓)	**98.2/2.08** (1.1%↑/19.1%↓)
Place	VLP [45]	40.0	20.0
	A^2-G-Place [41]	32.3	29.3
	A^2 [41]	89.3	74.0
	A^2-PA [41]	89.0	76.0
	Ours+A^2-G-Place	**35.7** (10.3%↑)	**33.6** (14.7%↑)
	Ours+A^2	**91.3** (2.2%↑)	**76.5** (3.4%↑)
	Ours+A^2-PA	**89.9** (1.1%↑)	**77.2** (1.6%↑)
Pick-n-Place	Act3D [7]	0.0/-	0.0/-
	RVT-2 [8]	0.83/4.00	0.0/-
	3D-Diff-Act [15]	1.67/6.13	0.0/-
	A^2-G [41]	30.7/2.42	28.3/2.00
	A^2 [41]	87.5/2.45	71.7/3.02
	A^2-PA [41]	91.3/2.06	76.7/3.22
	Ours+A^2-G	**32.8/2.23** (6.9%↑/7.8%↓)	**30.9/1.84** (9.2%↑/8%↓)
	Ours+A^2	**88.7/2.23** (1.6%↑/8.9%↓)	**72.8/2.94** (2.5%↑/2.1%↓)
	Ours+A^2-PA	**92.1/2.04** (1.1%↑/1.0%↓)	**79.8/2.99** (4.1%↑/7.1%↓)

We further evaluate the performance of our proposed method in non-clutter grasping benchmark on the GraspNet-1Billion [6] under the RealSense setting in Table 3. As seen in the table, our model consistently achieves the highest average precision (AP) across Seen, Similar, and Novel object splits, marking absolute gains of 1.4 $\sim$ 2.4 over the strongest baseline GSNet [37] and 1.6 $\sim$ 1.7 over DexGraspNet-2 [46]. Earlier analytic or regression-based models (e.g., GPD [35], PointGPD [19]) struggle to generalize across novel geometries due to their reliance on local point-cloud geometry and discrete candidate scoring, while recent diffusion or sampling-based methods still lack consistent global shape priors. In contrast, our HGF encodes grasp feasibility as a smooth potential function that unifies surface normal, curvature, and reachability cues, allowing the network to generalize to unseen object categories.

Overall, our unified framework demonstrates strong generalization across both cluttered and non-cluttered settings, as seen in Table 1, 2, and 3. In non-clutter scenes, the harmonic potential enables precise, geometry-consistent grasp localization, while in cluttered environments, its continuous obstacle-aware formulation, coupled with hierarchical diffusion sampling, preserves physical feasibility under occlusion and contact constraints. Together, they highlight the scalability of our method. Qualitative visualizations are shown in Figure 3.

4.5 Evaluation on Policy Rollout

To further assess the downstream utility of our grasp generation framework, we extend our evaluation beyond grasp prediction to full policy-level manipulation in Table 4 on the A^2 simulation benchmark [41], encompassing pick, place and pick-and-place tasks under both seen and unseen object settings. We integrate

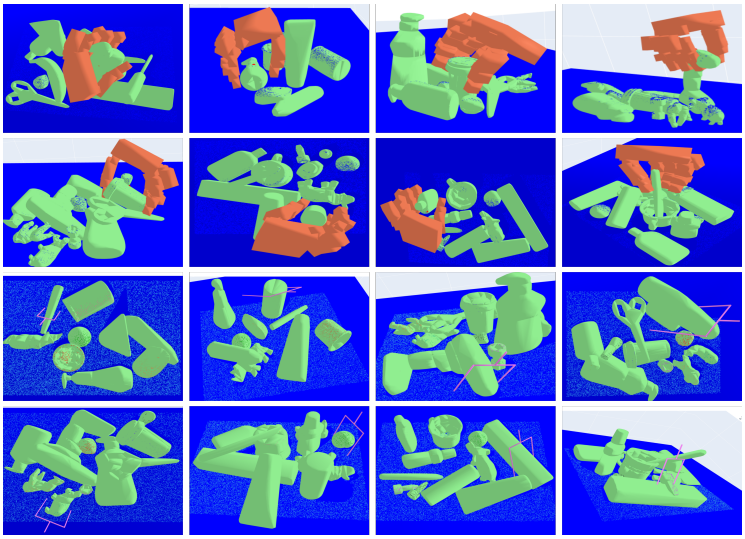

Fig. 3: Qualitative visualization of dexterous grasp (top two rows) and gripper grasp (bottom two rows) in cluttered scenes.

our proposed grasping model with several existing policy architectures, without modifying their control or planning modules. This setup isolates the contribution of grasp prediction quality to overall policy performance.

Across all task categories in Table 4, integrating our grasp generation significantly enhances both success rate and planning efficiency. For Pick tasks, our model boosts success by +4.3% (Seen) and +2.4% (Unseen) for A^2-G-Pick[2], while reducing planning steps by 16 $\sim$ 19%, and further improves A^2 to 96.3/2.14 (Seen) and 98.2/2.08 (Unseen). For Place tasks, performance gains reach $+10.3\%/+14.7\%$ for A^2-G-Place and remain consistent across A^2-PA and A^2 variants, demonstrating better spatial alignment between grasp and target pose. Similarly, for Pick-n-Place policies, our approach yields stable improvements of $1 \sim 9\%$ in success and up to 9% fewer planning steps, with the highest efficiency observed for A^2-G and A^2-PA. Overall, as seen in Table 1, 2, 3, 4, our method not only enhances grasp-level success but also translates to policy-level improvements. Please see **supplementary files** for additional results.

4.6 Ablation Experiments

Table 5 presents a comprehensive ablation analysis of our framework across dexterous and gripper grasping tasks, illustrating the cumulative effect of each component. **(1)** Instead of splitting the latent representation, we utilize a single global feature vector, generating a baseline performance (e.g., 87.0/79.6/69.1

[2] A^2-G-Pick, A^2, A^2-G, A^2-PA are different model variants in [41]

Table 5: Ablation experiment to analyze the contribution of individual components for different grasping conditions. Note, we only report GSR for cluttered gripper grasping on the GraspClutter6D dataset.

Latent Decomposition					DexGraspNet2 [GraspNet1B]			DexGraspNet2 [ShapeNet]			GraspClutter6D				GraspNet-1Billion		
No Split	Fixed	Diffused	LCD	HGF	Dense	Random	Loose	Dense	Random	Loose	Packed (5)	Piled (5)	Piled (10)	Piled (15)	AP Seen	AP Similar	AP Novel
✓	✗	✗	✗	✗	87.0	79.6	69.1	79.5	83.0	68.5	84.5	75.5	74.5	75.0	75.4	69.7	34.3
✗	✓	✗	✗	✗	86.5	78.4	68.2	79.1	83.1	68.2	84.0	76.1	75.0	76.4	74.2	68.5	33.5
✗	✓	✓	✗	✗	88.5	80.0	69.8	80.1	84.0	69.2	85.0	77.8	76.2	77.1	76.5	70.2	35.8
✗	✓	✓	✓	✗	89.3	82.0	73.6	81.2	85.4	71.8	85.8	78.1	76.8	78.2	77.9	72.9	37.4
✗	✓	✓	✓	✓	**93.1**	**85.9**	**76.1**	**83.2**	**87.8**	**74.9**	**86.2**	**78.6**	**77.6**	**79.0**	**79.8**	**74.1**	**40.9**

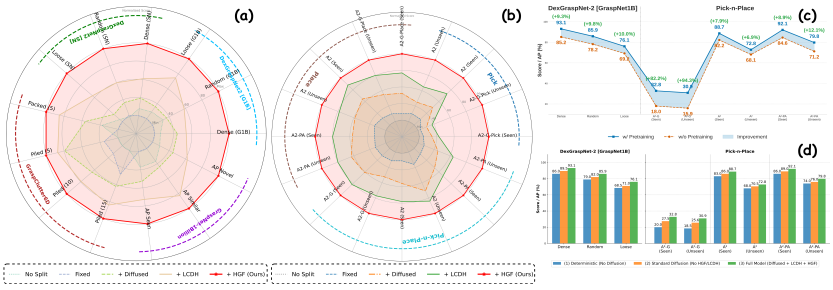

Fig. 4: Ablation Visualization: **(a)** Grasping and **(b)** Pick-n-Place results illustrating the incremental contribution of each component in our framework. **(c)** Impact of encoder pre-training. **(d)** Comparison of our split-latent diffusion process against alternative diffusion variants, demonstrating improved grasp feasibility and policy success.

on GraspNet1B) in row 1. **(2)** Fixed Latent slightly stabilizes training but restricts adaptability, showing minimal gains or slight drops (Dense DexGraspNet2 [GraspNet1B split] success rate 86.5, AP_{Seen} of GraspNet1B gripper grasping 74.2), since deterministic encoding cannot capture multiple valid grasp hypotheses. **(3)** Diffused Latent by GCD introduces stochastic sampling, leading to consistent $1.5 \sim 2$ improvement across datasets (e.g., Dense 88.5, AP_{Novel} 35.8) by better exploring multimodal grasp distributions and reducing overfitting to dominant modes. **(4)** Adding LCD enhances hierarchical global-local consistency, further improving performance by $1.5 \sim 2.5$ (e.g., Loose 69.8 $\rightarrow$ 73.6), as it aligns local contact reasoning with global object geometry, yielding structurally coherent grasps. **(5)** Finally, incorporating the Harmonic Grasp Field (HGF) yields the largest boost (+3.8 pp Dense, +3.5 pp AP_{Novel}) by enforcing a continuous, obstacle-aware potential that naturally guides grasp generation along collision-free manifolds. Together, these components collectively transform grasp synthesis from discrete candidate sampling into a geometrically grounded, probabilistic field representation, resulting in consistent performance gains across cluttered, dexterous, and parallel-jaw grasping conditions. We visualize ablation results for grasping and policy-rollout in Figure 4(a) and Figure 4(b), effectiveness of encoder pre-training in Figure 4(c) and a quantitative comparison of diffusion variants in Figure 4(d), justifying the effectiveness of split-latent diffusion model.

5 Conclusion

In this work, we introduce a unified differential geometry and diffusion-based framework for dexterous and gripper grasp generation in both cluttered and non-cluttered environments. By integrating harmonic manifold potentials with a split-latent diffusion hierarchy, our method achieves continuous, obstacle-aware, and multimodal grasp synthesis that generalizes across diverse object categories and manipulation settings.

Acknowledgments

The authors were supported by NSF grants IIS-2331769, CMMI-2246673, and ECCS-2025929.

References

1. Back, S., Lee, J., Kim, K., Rho, H., Lee, G., Kang, R., Lee, S., Noh, S., Lee, Y., Lee, T., et al.: Graspclutter6d: A large-scale real-world dataset for robust perception and grasping in cluttered scenes. IEEE Robotics and Automation Letters (2025)
2. Carvalho, J., Le, A.T., Kicki, P., Koert, D., Peters, J.: Motion planning diffusion: Learning and adapting robot motion planning with diffusion models. IEEE Transactions on Robotics (2025)
3. Chen, Z., Yan, Q., Chen, Y., Wu, T., Zhang, J., Ding, Z., Li, J., Yang, Y., Dong, H.: Clutterdexgrasp: A sim-to-real system for general dexterous grasping in cluttered scenes. arXiv preprint arXiv:2506.14317 (2025)
4. Chen, Z.Q., Van Wyk, K., Chao, Y.W., Yang, W., Mousavian, A., Gupta, A., Fox, D.: Learning robust real-world dexterous grasping policies via implicit shape augmentation. arXiv preprint arXiv:2210.13638 (2022)
5. Chu, F.J., Xu, R., Vela, P.A.: Real-world multiobject, multigrasp detection. IEEE Robotics and Automation Letters **3**(4), 3355–3362 (2018)
6. Fang, H.S., Wang, C., Gou, M., Lu, C.: Graspnet-1billion: A large-scale benchmark for general object grasping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11444–11453 (2020)
7. Gervet, T., Xian, Z., Gkanatsios, N., Fragkiadaki, K.: Act3d: 3d feature field transformers for multi-task robotic manipulation. arXiv preprint arXiv:2306.17817 (2023)
8. Goyal, A., Blukis, V., Xu, J., Guo, Y., Chao, Y.W., Fox, D.: Rvt-2: Learning precise manipulation from few demonstrations. arXiv preprint arXiv:2406.08545 (2024)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
11. Huang, C.Y., Shao, Y.H.: Integration of deep q-learning with a grasp quality network for robot grasping in cluttered environments. Journal of Intelligent & Robotic Systems **110**(3), 97 (2024)
12. Ji, M., Qiu, R.Z., Zou, X., Wang, X.: Graspsplats: Efficient manipulation with 3d feature splatting. arXiv preprint arXiv:2409.02084 (2024)

13. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11107–11116 (2021)
14. Kasaei, H., Kasaei, M.: Mvgrasp: Real-time multi-view 3d object grasping in highly cluttered environments. Robotics and Autonomous Systems **160**, 104313 (2023)
15. Ke, T.W., Gkanatsios, N., Fragkiadaki, K.: 3d diffuser actor: Policy diffusion with 3d scene representations. arXiv preprint arXiv:2402.10885 (2024)
16. Kim, K., Back, S., Lee, G., Lee, S., Noh, S., Lee, K.: Bigraspformer: End-to-end bimanual grasp transformer. arXiv preprint arXiv:2509.19142 (2025)
17. Lenz, I., Lee, H., Saxena, A.: Deep learning for detecting robotic grasps. The International Journal of Robotics Research **34**(4-5), 705–724 (2015)
18. Li, Y., Wei, W., Li, D., Wang, P., Li, W., Zhong, J.: Hgc-net: Deep anthropomorphic hand grasping in clutter. In: 2022 International conference on robotics and automation (ICRA). pp. 714–720. IEEE (2022)
19. Liang, H., Ma, X., Li, S., Görner, M., Tang, S., Fang, B., Sun, F., Zhang, J.: Pointnetgpd: Detecting grasp configurations from point sets. In: 2019 international conference on robotics and automation (ICRA). pp. 3629–3635. IEEE (2019)
20. Lundell, J., Verdoja, F., Kyrki, V.: Ddgc: Generative deep dexterous grasping in clutter. IEEE Robotics and Automation Letters **6**(4), 6899–6906 (2021)
21. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2837–2845 (2021)
22. Mahler, J., Liang, J., Niyaz, S., Laskey, M., Doan, R., Liu, X., Ojea, J.A., Goldberg, K.: Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. arXiv preprint arXiv:1703.09312 (2017)
23. Morrison, D., Corke, P., Leitner, J.: Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. arXiv preprint arXiv:1804.05172 (2018)
24. Peng, G., Ren, Z., Wang, H., Li, X., Khyam, M.O.: A self-supervised learning-based 6-dof grasp planning method for manipulator. IEEE Transactions on Automation Science and Engineering **19**(4), 3639–3648 (2021)
25. Qian, S., Mo, K., Blukis, V., Fouhey, D.F., Fox, D., Goyal, A.: 3d-mvp: 3d multi-view pretraining for manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22530–22539 (2025)
26. Qian, Y., Zhu, X., Biza, O., Jiang, S., Zhao, L., Huang, H., Qi, Y., Platt, R.: Thinkgrasp: A vision-language system for strategic part grasping in clutter. arXiv preprint arXiv:2407.11298 (2024)
27. Qin, Y., Chen, R., Zhu, H., Song, M., Xu, J., Su, H.: S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes. In: Conference on robot learning. pp. 53–65. PMLR (2020)
28. Radosavovic, I., Xiao, T., James, S., Abbeel, P., Malik, J., Darrell, T.: Real-world robot learning with masked visual pre-training. In: Conference on Robot Learning. pp. 416–426. PMLR (2023)
29. Rashid, A., Sharma, S., Kim, C.M., Kerr, J., Chen, L.Y., Kanazawa, A., Goldberg, K.: Language embedded radiance fields for zero-shot task-oriented grasping. In: 7th Annual Conference on Robot Learning (2023)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
31. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

32. Sundermeyer, M., Mousavian, A., Triebel, R., Fox, D.: Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In: 2021 IEEE international conference on robotics and automation (ICRA). pp. 13438–13444. IEEE (2021)
33. Tan, A.H., Narasimhan, S., Nejat, G.: 4cnet: A diffusion approach to map prediction for decentralized multi-robot exploration. IEEE Transactions on Robotics (2026)
34. Tang, Y., Zhang, S., Hao, X., Wang, P., Wu, J., Wang, Z., Zhang, S.: Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9433–9439. IEEE (2025)
35. Ten Pas, A., Gualtieri, M., Saenko, K., Platt, R.: Grasp pose detection in point clouds. The International Journal of Robotics Research **36**(13-14), 1455–1473 (2017)
36. Wang, C., Peng, H.Y., Liu, Y.T., Gu, J., Hu, S.M.: Diffusion models for 3d generation: A survey. Computational Visual Media **11**(1), 1–28 (2025)
37. Wang, C., Fang, H.S., Gou, M., Fang, H., Gao, J., Lu, C.: Graspness discovery in clutters for fast and accurate grasp detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15964–15973 (2021)
38. Weng, Z., Lu, H., Kragic, D., Lundell, J.: Dexdiffuser: Generating dexterous grasps with diffusion models. IEEE Robotics and Automation Letters **9**(12), 11834–11840 (2024)
39. Wu, S., Zhu, Y., Huang, Y., Zhu, K., Gu, J., Yu, J., Shi, Y., Wang, J.: Afforddp: Generalizable diffusion policy with transferable affordance. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6971–6980 (2025)
40. Xie, Z., Xie, G., Liu, Y., Zhang, Y., Cao, B., Ji, Y., Wang, Z., Liu, H.: Dexmgnet: Multi-mode dexterous grasping in cluttered scenes with generative models. IEEE Robotics and Automation Letters (2025)
41. Xu, K., Xia, X., Wang, K., Yang, Y., Mao, Y., Deng, B., Ye, J., Xiong, R., Wang, Y.: Efficient alignment of unconditioned action prior for language-conditioned pick and place in clutter. IEEE Transactions on Automation Science and Engineering (2025)
42. Xu, K., Zhao, S., Zhou, Z., Li, Z., Pi, H., Zhu, Y., Wang, Y., Xiong, R.: A joint modeling of vision-language-action for target-oriented grasping in clutter. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 11597–11604. IEEE (2023)
43. Xu, L., Qu, H., Cai, Y., Liu, J.: 6d-diff: A keypoint diffusion framework for 6d object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9676–9686 (2024)
44. Xu, L., Liu, Z., ZHEWEI, G., Guo, J., Jiang, Z., Xu, Z., Gao, C., Shao, L.: Dexsingrasp: Learning a unified policy for dexterous object singulation and grasping in cluttered environments. In: ICRA 2025 Workshop"Handy Moves: Dexterity in Multi-Fingered Hands"Paper Submission (2025)
45. Xu, Z., Xu, K., Xiong, R., Wang, Y.: Object-centric inference for language conditioned placement: A foundation model based approach. In: 2023 International Conference on Advanced Robotics and Mechatronics (ICARM). pp. 203–208. IEEE (2023)
46. Zhang, J., Liu, H., Li, D., Yu, X., Geng, H., Ding, Y., Chen, J., Wang, H.: Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In: 8th Annual Conference on Robot Learning (2024)

47. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021)
48. Zheng, X., Yuan, S., Chen, P.: Robotic autonomous grasping strategy and system for cluttered multi-class objects. International Journal of Control, Automation and Systems **22**(8), 2602–2612 (2024)
49. Zhong, Y., Huang, X., Li, R., Zhang, C., Chen, Z., Guan, T., Zeng, F., Lui, K.N., Ye, Y., Liang, Y., et al.: Dexgraspvla: A vision-language-action framework towards general dexterous grasping. arXiv preprint arXiv:2502.20900 (2025)