

GraspFoM: Towards Reconstruction-Driven Robotic Grasping with 3D Foundation Priors

Dongli Wu¹, Xiaobao Wei², Hao Wang², Qiaochu Dong², Ying Li^{2,3}, Qingpo Wu², Ming Lu², Wufan Zhao¹

¹The Hong Kong University of Science and Technology (Guangzhou) ²Peking University ³The Hong Kong University of Science and Technology

Robotic grasping is a fundamental capability in robotic manipulation. Yet grasping remains challenging under partial observations. Reliable grasping depends on both local contact cues and object-level 3D structure. Existing geometry-aware grasping methods recognize the value of reconstruction, but they typically treat geometry as an intermediate prediction rather than a reusable object prior for grasping. In this paper, we present GraspFoM, a unified framework that leverages 3D foundation priors (SAM3D) to build a shared 3D object latent for both reconstruction and grasp pose prediction. Built on this shared object latent, we introduce an anchor-initialized truncated pose-reasoning diffuser that predicts continuous and multimodal grasp poses without directly relying on discrete grasp candidates. We further investigate the interaction between reconstruction and grasping through a reconstruction-aware scorer and a residual latent updater. Reconstruction provides grounded geometric cues, while grasp supervision refines the shared object latent toward grasp-relevant affordances. GraspFoM jointly predicts grasp poses and reconstructs high-fidelity 3D assets in mesh and 3DGS forms. Comprehensive experiments demonstrate that GraspFoM achieves state-of-the-art results on both reconstruction and grasping. Notably, these improvements require only a small number of additional trainable parameters. Component-wise ablation studies also demonstrate the contribution of each component. Codes will be released.

Date: June 5, 2026

Code: Codes will be released.

Keywords: Reconstruction, Grasping Pose Prediction, Embodied Intelligence

1 Introduction

Robotic grasping is a fundamental capability in robotic manipulation, yet it remains challenging in real-world scenes (Du et al., 2021; Dong and Zhang, 2023; Newbury et al., 2023; Wang et al., 2025). A robust grasping system (Xu et al., 2023; Deng et al., 2025; Li et al., 2024, 2026) must jointly reason about object geometry and multimodal action policies under clutter, occlusion, and partial observations. This makes accurate and generalizable grasping difficult in practice.

Existing grasping methods have evolved along several major directions as shown in Fig. 1. Early learning-based approaches (Fang et al., 2020; Mousavian et al., 2019; Wang et al., 2021) mainly formulate grasping as direct pose prediction from RGB-D observations or point clouds, establishing the basic paradigm of data-driven grasp detection. Building on this line, subsequent works (Sundermeyer et al., 2021; Zhou et al., 2024) increasingly incorporate richer geometric cues, such as contact reasoning, graspness estimation, and object-centric 3D features, which improve grasp generation in cluttered scenes and for novel objects. More recently, several works (Chisari et al., 2024; Fan et al., 2025; Iwase et al., 2025; Wang et al., 2026) have moved toward unified frameworks that couple object reconstruction with grasp pose prediction, bringing explicit 3D supervision into grasp learning. In parallel, Vision-language-model (VLM) based policies (Shao et al., 2025; Deng et al., 2025; Cao et al., 2026a,b) further push grasping toward broader task-agnostic generalization by leveraging the strong semantic reasoning, prior knowledge, and instruction-following capabilities inherited from large language models. These methods highlight the growing role of large-scale foundational priors in improving the generalization ability of robotic grasping.

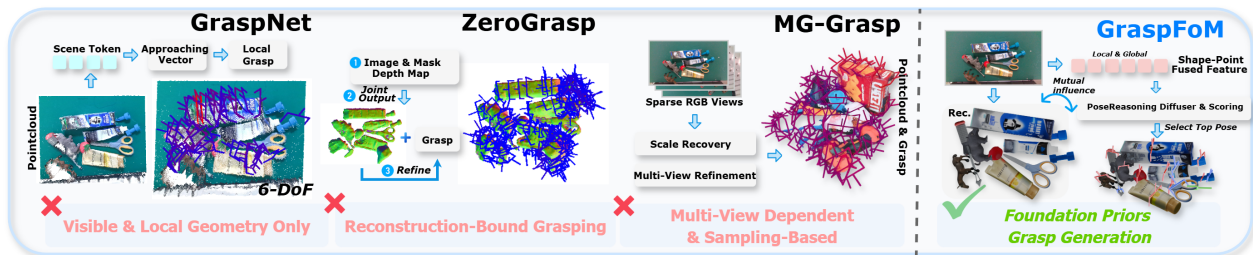


Figure 1 Representative paradigms for robotic grasping. GraspNet (Fang et al., 2020) relies on visible local geometry, ZeroGrasp (Iwase et al., 2025) couples grasping with task-specific reconstruction, and MG-Grasp (Wang et al., 2026) depends on sparse multi-view sampling and refinement. In contrast, GraspFoM builds grasping on transferable 3D foundation priors, using a shared object latent to unify reconstruction, continuous grasp generation, and reconstruction-aware scoring.

Despite these advances, robotic grasping under partial observations remains challenging because the underlying object structure is often incomplete or ambiguous. Reconstruction-aided grasping improves geometric grounding, but existing methods (Chisari et al., 2024; Fan et al., 2025; Iwase et al., 2025) still rely on task-specific representations learned from scratch. By contrast, VLM-based policies (Shao et al., 2025; Deng et al., 2025; Noh et al., 2025) improve task-level generalization, yet they inherit 2D semantic knowledge and rely on time-consuming LLM decoding, leading to poor and slow 3D grasping pose prediction.

Meanwhile, general 3D object reconstruction has progressed rapidly from task-specific geometric regression (Wang et al., 2021; Yariv et al., 2021; Cen et al., 2023; Wei et al., 2024; Huang et al., 2024) and optimization pipelines to large-scale reconstruction (Tochilkin et al., 2024; Yang et al., 2024; Xu et al., 2024; Wei et al., 2025; Wu et al., 2025) and generation (Qian et al., 2023; Shi et al., 2023; Deitke et al., 2023; Zeng et al., 2025; Chen et al., 2025) models that learn reusable shape priors from massive data. Recent methods (Hong et al., 2023; Zhang et al., 2024; Chen et al., 2025) such as LRM, GS-LRM and SAM3D show that high-quality 3D geometry can now be recovered in a feed-forward manner from sparse or incomplete observations, producing faithful object assets in forms such as meshes, radiance fields (Mildenhall et al., 2021), and 3D Gaussians (Kerbl et al., 2023). These advances suggest that strong object-level 3D priors can be learned and transferred across novel instances and observation conditions. Yet existing methods mainly optimize the fidelity of digital 3D assets, rather than the object representations needed for robotic manipulation. Consequently, their priors cannot be directly applied to grasping, which requires not only accurate geometry but also manipulation-relevant structure such as graspable regions and affordance cues (Pohl et al., 2020; Zeng et al., 2022; Zhong et al., 2025; Wei et al., 2025).

Motivated by this gap, we propose GraspFoM, a unified framework for reconstruction-driven robotic grasping built on top of SAM3D (Yang et al., 2023) priors. Given RGB images, object masks, and point maps, GraspFoM first uses the sparse-structure stage of SAM3D to initialize a shared 3D object latent. Specifically, we query surface points from the object and fuse point-wise local cues with grid-sampled global shape context to construct shared 3D object latents. Based on these latents, we introduce a pose-reasoning diffuser initialized with an anchored Gaussian distribution, which models continuous and multimodal grasp pose distributions through a reasoning trunk and a truncated denoising process. Unlike existing methods that rely on discrete grasp candidates, our method predicts grasp poses directly from the shared object latents.

Beyond joint prediction, GraspFoM enforces bidirectional interaction between reconstruction and grasp learning through a reconstruction-aware scorer and a residual latent updater. The scorer predicts graspness, quality, and affordance directly on reconstructed geometry, grounding grasp evaluation in object structure rather than pose-only cues. The updater takes grasp supervision back into the shared latent through surface-centered offsets, injecting manipulation-relevant signals into the reconstruction process. This establishes a bidirectional link between the two tasks: reconstruction improves grasp prediction, while grasp learning refines the underlying object representation. As a result, GraspFoM jointly outputs accurate grasp poses along with high-fidelity 3D assets, achieving state-of-the-art performance on the widely used GraspNet-1B (Fang et al., 2020) benchmark.

Our contributions are summarized as follows:

- We propose GraspFoM, a unified framework that leverages 3D foundation priors to build a shared 3D object latent for both object reconstruction and grasp pose prediction.
- We introduce an anchor-initialized pose-reasoning diffuser as the grasping head to model continuous 3D grasp pose distributions, thereby avoiding discrete candidates.
- We design a reconstruction-aware scorer and a residual latent updater to couple reconstruction and grasp learning, enabling geometric grounding and grasp-relevant affordance enhancement within the shared latent.
- GraspFoM achieves state-of-the-art performance on GraspNet-1B dataset. Comprehensive experiments demonstrate the effectiveness of proposed techniques.

2 Related Work

2.1 3D Object Reconstruction

3D object reconstruction aims to recover faithful geometry and appearance from sparse observations. Early works (Wang et al., 2021; Yariv et al., 2021; Cen et al., 2023; Wei et al., 2024; Huang et al., 2024; Wang et al., 2025) mainly rely on implicit neural representations, signed distance fields, or optimization-based formulations to model object surfaces with high fidelity. Recent progress (Tochilkin et al., 2024; Yang et al., 2024; Xu et al., 2024; Wei et al., 2025; Wu et al., 2025; Qian et al., 2023; Shi et al., 2023; Deitke et al., 2023; Zeng et al., 2025; Chen et al., 2025) shifts toward scalable reconstruction and generation models trained on large 3D corpora, enabling feed-forward recovery of object geometry from sparse or incomplete observations. In particular, large reconstruction models such as LRM (Hong et al., 2023), GS-LRM (Zhang et al., 2024), and SAM3D (Chen et al., 2025; Yang et al., 2023) show that strong object-level priors can be learned from massive data and transferred across novel instances and observation conditions. These methods can produce high-quality 3D assets in forms such as meshes, radiance fields (Mildenhall et al., 2021), and 3D Gaussians (Kerbl et al., 2023). However, existing reconstruction methods mainly optimize the fidelity of digital assets rather than the manipulation-oriented object structure required for robotic interaction. This leaves a gap between recent advances in general 3D reconstruction and the needs of robotic manipulation.

2.2 Robotic Reconstruction for Grasping

Accurate 3D geometric representation is foundational to robust robotic grasping (Williams et al., 2024). Traditional approaches rely on discrete formats (e.g., voxel grids and point clouds) to model object shapes. While widely adopted in early grasp pipelines, these methods often trade resolution for efficiency or lack generalizability across diverse geometries (Wang et al., 2023; Williams et al., 2024; Eppner et al., 2021; Park et al., 2019). For cluttered real-world environments, recent methods emphasize open-world adaptability. For instance, SceneComplete (Agarwal et al., 2024) fills missing geometries via 3D scene completion, and SAM-based approaches (Qu et al., 2024) leverage powerful segmentation models to reconstruct unseen objects. Beyond traditional discrete formats, recent advances in continuous 3D representations, such as neural occupancy fields and 3D Gaussians (Wei et al., 2024; Li et al., 2025, 2026; Wang et al., 2025; Wei et al., 2025), have significantly enhanced embodied spatial perception, scene understanding, and trajectory planning. Although these works highlight the critical role of complete 3D modeling, they focus primarily on scene understanding or spatial navigation without explicit integration into end-to-end grasp reasoning. Similarly, while implicit encoders enhance shape details (Huang et al., 2023; Heppert et al., 2023; Chan et al., 2023), and fast reconstruction frameworks prioritize speed for immediate grasp planning (Avigal et al., 2020), their task-specific designs or limited precision restrict their adaptability to dynamic, open-world scenarios. Furthermore, recent advancements like Sharp-It (Edelstein et al., 2025) refine low-fidelity 3D geometry into high-precision multi-view representations via diffusion models, but they still lack synergy with downstream grasp optimization.

2.3 Grasping Pose Prediction

Early grasp pose prediction methods typically output 6D poses directly from RGB-D inputs, often bypassing explicit geometric modeling, which leads to unstable contacts or collisions in cluttered scenes (Fang et al., 2020; Mousavian et al., 2019; Wang et al., 2021). To address these flaws, subsequent works have advanced grasping

in targeted scenarios. For instance, Contact-GraspNet (Sundermeyer et al., 2021) enables efficient 6-DoF grasp generation in clutter, while recent dynamic scene reconstruction pipelines (Zhou et al., 2024) facilitate the grasping of novel objects. However, these methods often rely on task-specific designs and lack generalizable 3D perception support. To tightly couple perception and planning, unified reconstruction-grasp frameworks attempt to co-optimize geometric modeling and pose prediction. For example, CenterGrasp (Chisari et al., 2024) learns object-aware implicit representations for simultaneous shape reconstruction and grasp estimation. Nevertheless, such approaches face inherent trade-offs among reconstruction detail, inference speed, and cross-scene adaptability (Ma et al., 2024; Liu et al., 2023; Chan et al., 2023; Chisari et al., 2024). Recent large-scale efforts, such as ZeroGrasp (Iwase et al., 2025), leverage massive datasets (e.g., 11.3B annotations) to achieve zero-shot grasping via unified octree-based reconstruction, but they remain susceptible to resolution dependency under occlusions. Concurrently, Vision-Language-Action (VLA) models (Shao et al., 2025) and their variants like GraspVLA (Deng et al., 2025) have shown remarkable generalization in dynamic scenarios. Yet, these VLM-based approaches typically prioritize semantic generalization over the high-precision geometric alignment required for stable grasping.

In contrast, we build GraspFoM on top of SAM3D priors and use them to form a shared 3D object latent for both reconstruction and grasp pose prediction. This design differs fundamentally from recent unified methods such as ZeroGrasp (Iwase et al., 2025), which learns from scratch and improves grasping through reconstruction. By leveraging transferable 3D foundation priors, our method focuses on reusable object-level representation rather than task-specific geometry learning.

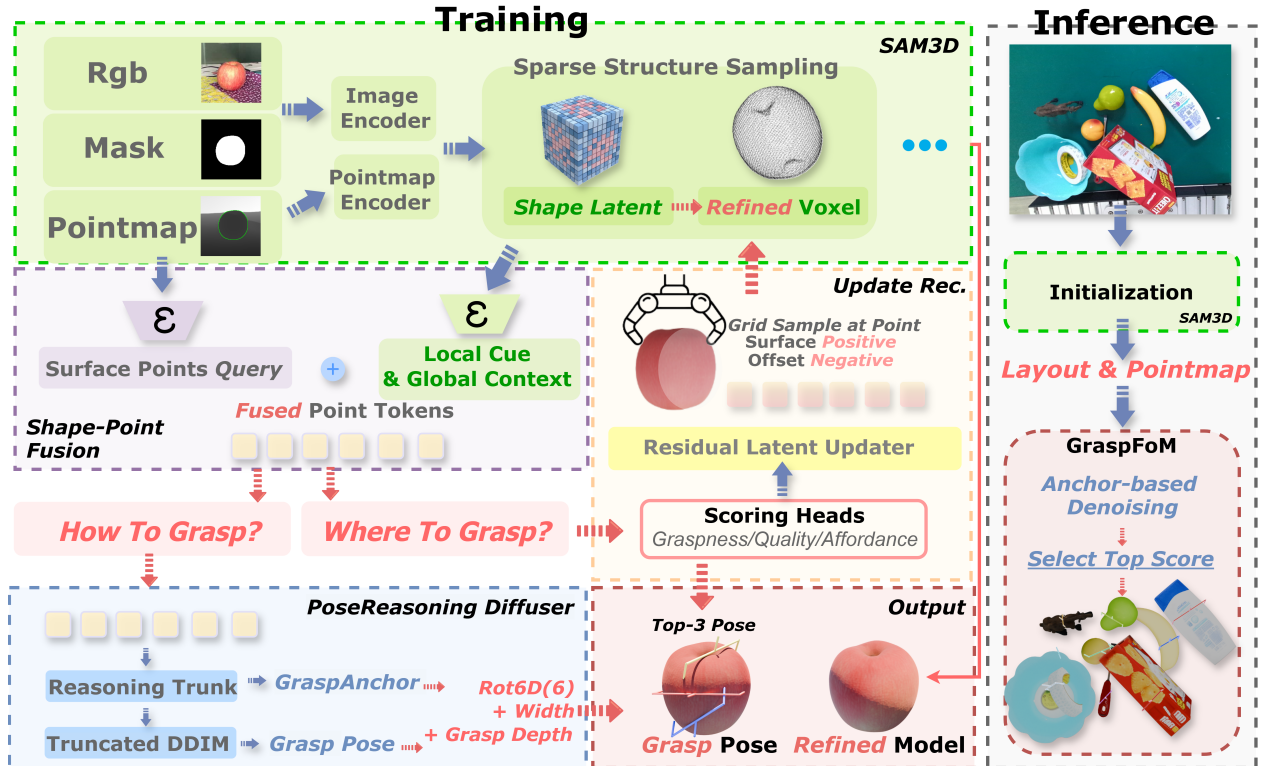


Figure 2 Overview of GraspFoM. (a) During training, we first fuse a shared shape latent. Based on these latents, the pose-reasoning diffuser predicts grasp poses and further refines the latents. (b) During inference, GraspFoM performs anchor-based denoising, score-based selection, and returns the top-ranked grasp poses together with high-fidelity 3D object assets.

3 Methodology

The overall pipeline of GraspFoM is illustrated in Fig. 2. We first introduce the problem setup in Sec. 3.1. We then describe how 3D foundational priors are transformed into a shared object representation in Sec. 3.2, followed by the pose generation process based on anchor-conditioned denoising in Sec. 3.3. The reconstruction-

aware scoring and latent refinement are presented in Sec. 3.4.

3.1 Preliminaries

Let $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ denote an RGB image, $\mathbf{M} \in \{0, 1\}^{H \times W}$ the object mask, and $\mathbf{P} \in \mathbb{R}^{H \times W \times 3}$ the object-centric point map in the camera frame. Our goal is to jointly reconstruct an object-centric 3D representation and predict a set of feasible grasp poses from the same observation. We denote the queried visible surface points by

$$\mathcal{Q} = \{(\mathbf{p}_i, \mathbf{n}_i)\}_{i=1}^N, \quad \mathbf{p}_i, \mathbf{n}_i \in \mathbb{R}^3, \quad (1)$$

where \mathbf{p}_i is a visible 3D point and \mathbf{n}_i is its corresponding surface normal.

We adopt the standard two-finger parallel gripper parameterization. For each queried point, a grasp pose is represented as

$$\mathbf{g}_i = [\mathbf{r}_i, w_i, d_i] \in \mathbb{R}^8, \quad (2)$$

where $\mathbf{r}_i \in \mathbb{R}^6$ is the continuous 6D rotation representation, w_i is the finger opening width, and d_i is the grasp depth. Following the implementation, the diffusion branch models only rotation and depth, while the width is recovered from local geometry during inference. We therefore define the diffusion target as

$$\mathbf{x}_i^0 = [\bar{\mathbf{r}}_i, \bar{d}_i] \in [-1, 1]^7, \quad (3)$$

where $\bar{\mathbf{r}}_i$ and \bar{d}_i are normalized versions of the 6D rotation and depth, respectively.

In addition to pose prediction, our model predicts three point-wise scores,

$$s_i \in \mathbb{R}, \quad q_i \in \mathbb{R}, \quad a_i \in \mathbb{R}, \quad (4)$$

which correspond to graspness, grasp quality, and affordance. These scores are used both for training and for reconstruction-aware ranking at inference time.

3.2 Shared Object Latent Preparation

A key challenge is that reconstruction and grasping require both fine-grained local geometry and coherent global object structure, yet existing pipelines usually encode them in separate task-specific representations. We therefore seek a shared object latent that inherits strong 3D priors from SAM3D while remaining directly usable for downstream grasp reasoning.

We build GraspFoM on top of the object-centric stage-1 output of SAM3D (Yang et al., 2023). Given $(\mathbf{I}, \mathbf{M}, \mathbf{P})$, the SAM3D sparse-structure pipeline produces a compact object latent together with the coordinate alignment parameters used by the point-map preprocessing:

$$(\mathbf{z}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \Phi_{\text{SAM3D}}(\mathbf{I}, \mathbf{M}, \mathbf{P}), \quad (5)$$

where $\mathbf{z} \in \mathbb{R}^{G^3 \times C_z}$ denotes the shared 3D object latent, and $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ denote the scale-shift parameters for aligning camera-space points with the latent coordinate system.

To make this compact latent directly usable for downstream grasp reasoning, we first convert it into a structured 3D feature volume and then lift it into a shape-aware grid representation:

$$\tilde{\mathbf{Z}} = E_{\text{shape}}(\text{Reshape}(\mathbf{z})) \in \mathbb{R}^{C \times G^3}, \quad (6)$$

where E_{shape} denotes a lightweight 3D encoder shared by local feature sampling and global token extraction. This shape grid serves as the common geometric representation for both reconstruction and grasp prediction.

Given the shape-aware grid $\tilde{\mathbf{Z}}$, we next build a latent-conditioned representation for each queried surface point $\mathbf{p}_i \in \mathcal{Q}$, where \mathcal{Q} is obtained by the surface-point query module from the visible object surface. Specifically, each queried point is first transformed from the camera coordinate system to the latent-aligned coordinate system and then normalized to the sampling range of the latent grid:

$$\bar{\mathbf{p}}_i = \mathcal{N}(\mathcal{T}_{\text{ssi}}(\mathbf{p}_i; \boldsymbol{\gamma}, \boldsymbol{\beta})), \quad (7)$$

where $\mathcal{N}(\cdot)$ denotes normalization to $[-1, 1]^3$.

Based on $\bar{\mathbf{p}}_i$, we extract local geometric cues and latent-conditioned shape features to form the point-wise local representation:

$$\mathbf{f}_i^{\text{pt}} = E_{\text{pt}}([\mathbf{p}_i, \mathbf{n}_i]), \quad \mathbf{f}_i^{\text{loc}} = E_{\text{loc}}(\mathcal{S}(\tilde{\mathbf{Z}}, \bar{\mathbf{p}}_i)), \quad (8)$$

where \mathbf{n}_i is the normal of \mathbf{p}_i , and $\mathcal{S}(\cdot, \cdot)$ denotes trilinear sampling on the shared 3D latent grid.

To inject global object context, we further convert the encoded latent grid into shape tokens $\mathbf{T}_{\text{shape}} = E_{\text{tok}}(\tilde{\mathbf{Z}})$ and update each point feature through cross-attention. The resulting feature is fused with the local point and latent features to form the final fused point token:

$$\mathbf{f}_i = E_{\text{fuse}}[\mathbf{f}_i^{\text{pt}}; \mathbf{f}_i^{\text{loc}}; \text{MHA}(\mathbf{f}_i^{\text{pt}} + \mathbf{f}_i^{\text{loc}}, \mathbf{T}_{\text{shape}}, \mathbf{T}_{\text{shape}})]. \quad (9)$$

Those fused point tokens jointly encode local surface cues and global shape context, and serves as the shared representation for the subsequent pose-reasoning, scoring, and reconstruction-update branches.

3.3 Pose-reasoning Diffuser

The grasp space is diverse and inherently multimodal. Direct regression tends to collapse to a single mode. We therefore formulate grasp prediction as latent-conditioned denoising in a normalized pose space.

We first build an anchor set in diffusion space by clustering valid training poses:

$$\mathcal{A} = \{\mathbf{a}_k\}_{k=1}^K, \quad \mathbf{a}_k \in [-1, 1]^7. \quad (10)$$

Each anchor is obtained by applying K-means to the normalized ground-truth pose targets over the training set. Given a training target \mathbf{x}_i^0 , we assign it to the nearest anchor,

$$k_i^* = \arg \min_k \|\mathbf{x}_i^0 - \mathbf{a}_k\|_2^2. \quad (11)$$

Instead of diffusing from pure Gaussian noise, we adopt a truncated anchored diffusion process. For a randomly sampled timestep $t \in \{0, \dots, T_{\text{tr}} - 1\}$ and Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we construct the noisy state by

$$\mathbf{x}_i^t = \sqrt{\bar{\alpha}_t} \mathbf{a}_{k_i^*} + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad (12)$$

where $\bar{\alpha}_t$ is the cumulative noise coefficient. This design follows the intuition of truncated diffusion: a more informative initialization improves denoising efficiency, while the anchor-conditioned formulation still supports multimodal grasp generation.

Conditioned on the fused latent \mathbf{f}_i , the pose branch first computes a pose reasoning feature

$$\mathbf{h}_i = E_{\text{pose}}(\mathbf{f}_i), \quad (13)$$

and predicts both the anchor logits and the denoised pose target:

$$\boldsymbol{\pi}_i = E_{\text{mode}}(\mathbf{h}_i) \in \mathbb{R}^K, \quad \hat{\mathbf{x}}_i^0 = D_\theta(\mathbf{h}_i, \mathbf{x}_i^t, t), \quad (14)$$

where $\boldsymbol{\pi}_i$ denotes the unnormalized scores over K grasp anchors, and D_θ is the pose-reasoning diffuser with sinusoidal timestep embedding. Unlike candidate-based grasp generation, the diffuser predicts grasp poses directly from the shared object latent.

3.4 Scorer and Reconstruction

The pose-reasoning diffuser generates continuous grasp hypotheses from the shared point-wise object latent. However, generative pose prediction alone does not determine which hypotheses are geometrically valid, locally stable, or semantically compatible with the target object. We further introduce a reconstruction-aware scorer to assess each queried surface point from three complementary aspects: graspability, grasp quality, and affordance.

Concretely, given the fused point-wise latent \mathbf{f}_i , we predict three point-wise logits:

$$\hat{s}_i = H_{\text{grasp}}(\mathbf{f}_i), \quad \hat{q}_i = H_{\text{qual}}(\mathbf{f}_i), \quad \hat{a}_i = H_{\text{aff}}(\mathbf{f}_i). \quad (15)$$

Here, \hat{s}_i estimates whether the queried surface point \mathbf{p}_i is graspable, \hat{q}_i measures the local grasp quality around that point, and \hat{a}_i captures affordance-oriented cues for manipulation. Unlike pose-only scoring, these predictions are grounded in the reconstructed object structure encoded in \mathbf{f}_i , which makes them suitable for both grasp ranking and reconstruction-aware latent refinement.

Beyond grasp scoring, we further use the predicted grasp signals to refine the shared reconstruction latent. Specifically, we first compute a confidence weight for each queried point,

$$\omega_i = \sigma(\hat{s}_i) \sigma(\hat{q}_i) \sigma(\hat{a}_i), \quad (16)$$

and aggregate the point-wise features into a manipulation-aware summary,

$$\bar{\mathbf{f}} = \frac{\sum_i \omega_i \mathbf{f}_i}{\sum_i \omega_i + \epsilon}. \quad (17)$$

This summary feature is then fed to a lightweight latent updater $U_{\text{lat}}(\cdot)$ to predict a residual feature volume,

$$\Delta \mathbf{Z} = U_{\text{lat}}(\bar{\mathbf{f}}, \mathbf{Z}), \quad (18)$$

which is added back to the original latent grid:

$$\mathbf{Z}' = \mathbf{Z} + \Delta \mathbf{Z}. \quad (19)$$

where $\Delta \mathbf{Z}$ is produced by a lightweight latent updater conditioned on $\bar{\mathbf{f}}$. In this way, grasp supervision is written back into the reconstruction branch, so that the latent gradually captures manipulation-relevant structure in addition to geometric fidelity.

To preserve geometric consistency after latent updating, we impose a surface-aware regularization on the refined latent. For each queried point, we evaluate a surface prediction head at the surface location and at a small offset along the normal direction:

$$\ell_i^+ = H_{\text{surf}}(\mathcal{S}(\mathbf{Z}', \bar{\mathbf{p}}_i)), \quad \ell_i^- = H_{\text{surf}}(\mathcal{S}(\mathbf{Z}', \bar{\mathbf{p}}_i + \delta \mathbf{n}_i)), \quad (20)$$

where δ is a small scalar offset. We then optimize

$$\mathcal{L}_{\text{rec-upd}} = \frac{\sum_i \omega_i \mathcal{L}_{\text{bce}}(\ell_i^+, 1)}{\sum_i \omega_i + \epsilon} + \frac{\sum_i \omega_i \mathcal{L}_{\text{bce}}(\ell_i^-, 0)}{\sum_i \omega_i + \epsilon}, \quad (21)$$

where \mathcal{L}_{bce} denotes the binary cross-entropy loss. This design makes the refined latent more favorable for reconstruction while aligning it with grasp-relevant affordance cues.

3.5 Optimization

Training. We train GraspFoM with point-wise supervision on the queried visible surface points, using the SAM3D stage-1 latent extracted online or loaded from cache. Let Ω denote the set of valid queried points, and let $\Omega^+ \subseteq \Omega$ denote the subset with valid positive grasp supervision. For each point \mathbf{p}_i , the model predicts graspness \hat{s}_i , quality \hat{q}_i , affordance \hat{a}_i , anchor logits $\boldsymbol{\pi}_i$, and the denoised diffusion target $\hat{\mathbf{x}}_i^0$.

For compactness, we denote binary cross-entropy and cross-entropy losses by $\ell_{\text{bce}}(\cdot, \cdot)$ and $\ell_{\text{ce}}(\cdot, \cdot)$, respectively. The point-wise scoring losses are:

$$\mathcal{L}_{\text{grasp}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \ell_{\text{bce}}(\hat{s}_i, s_i), \quad (22)$$

$$\mathcal{L}_{\text{qual}} = \frac{1}{|\Omega^+|} \sum_{i \in \Omega^+} \|\hat{q}_i - q_i\|_2^2, \quad (23)$$

$$\mathcal{L}_{\text{aff}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \ell_{\text{bce}}(\hat{a}_i, a_i). \quad (24)$$

For the diffusion branch, we supervise both the denoised pose target and the anchor assignment:

$$\mathcal{L}_{\text{diff}} = \frac{1}{|\Omega^+|} \sum_{i \in \Omega^+} \|\hat{\mathbf{x}}_i^0 - \mathbf{x}_i^0\|_2^2, \quad (25)$$

$$\mathcal{L}_{\text{mode}} = \frac{1}{|\Omega^+|} \sum_{i \in \Omega^+} \ell_{\text{ce}}(\boldsymbol{\pi}_i, k_i^*), \quad (26)$$

where k_i^* is the index of the assigned anchor for point i .

When the residual latent updater is enabled, we additionally optimize the reconstruction-update loss $\mathcal{L}_{\text{rec-upd}}$ defined in Sec . 3.4. The overall training objective is

$$\begin{aligned} \mathcal{L} = & \lambda_g \mathcal{L}_{\text{grasp}} + \lambda_q \mathcal{L}_{\text{qual}} + \lambda_a \mathcal{L}_{\text{aff}} \\ & + \lambda_d \mathcal{L}_{\text{diff}} + \lambda_m \mathcal{L}_{\text{mode}} + \lambda_r \mathcal{L}_{\text{rec-upd}}. \end{aligned} \quad (27)$$

In practice, the grasp-anchor bank is built once by K-means clustering over normalized valid training poses and then cached for subsequent training and inference.

Inference. At inference time, we first run the SAM3D sparse-structure stage to obtain $(\mathbf{z}, \gamma, \beta)$, and then query visible surface points and normals from the point map. Given the fused point-wise latent, the pose-reasoning diffuser performs truncated DDIM sampling from each grasp anchor. Specifically, let $\mathbf{x}_{i,k}^{t_0}$ denote the initial noisy sample around anchor \mathbf{a}_k at a truncated starting timestep t_0 . Starting from $\mathbf{x}_{i,k}^{t_0}$, the sampler iteratively updates

$$\mathbf{x}_{i,k}^{t-1} = \text{DDIMStep}(\hat{\mathbf{x}}_{i,k}^0, \mathbf{x}_{i,k}^t, t), \quad (28)$$

for a small number of denoising steps, and the final denoised state is converted back to metric pose parameters by

$$[\hat{\mathbf{r}}_{i,k}, \hat{d}_{i,k}] = \mathcal{T}_{\text{denorm}}(\hat{\mathbf{x}}_{i,k}^0). \quad (29)$$

The predicted pose parameters $\hat{\mathbf{r}}_{i,k}$ and $\hat{d}_{i,k}$ are further converted into full grasp poses by estimating the finger opening width from local object geometry, rather than diffusing it directly. Specifically, the width $\hat{w}_{i,k}$ is computed from the local extent of the neighborhood $\mathcal{N}(i)$ around \mathbf{p}_i along the width axis induced by $\hat{\mathbf{r}}_{i,k}$, followed by clipping to the valid gripper range.

The scorer ranks sampled grasp hypotheses in two stages. We first compute a point-level score that reflects whether the queried surface point is graspable, reliable, and manipulation-relevant:

$$\phi_i = \sigma(\hat{s}_i) \hat{q}_i (1 + \lambda_{\text{aff}} \sigma(\hat{a}_i)). \quad (30)$$

Here, ϕ_i combines graspness, local grasp quality, and affordance cues into a reconstruction-aware point score.

We then lift this point score to each grasp hypothesis by incorporating the view-alignment term and the anchor probability. For the grasp pose $\hat{\mathbf{g}}_{i,k}$ associated with point i and anchor k , we define

$$\text{Score}_{i,k} = \phi_i (1 + \lambda_{\text{view}} \rho_{i,k}) \text{softmax}(\boldsymbol{\pi}_i)_k, \quad (31)$$

where $\rho_{i,k}$ denotes the alignment between the predicted approach direction and the camera view ray, and $\text{softmax}(\boldsymbol{\pi}_i)_k$ is the confidence of the k -th grasp mode. In this way, the final ranking jointly accounts for point-level graspability, geometric compatibility with the viewing condition, and the mode confidence predicted by the diffuser.

The top-ranked hypotheses are optionally filtered by collision checking. Finally, the same latent can be decoded into mesh or 3D Gaussian outputs, so the pipeline returns both grasp poses and high-fidelity 3D assets.

Table 1 Quantitative evaluation of grasp pose prediction on the GraspNet-1B. G and R in the output columns indicate whether the method predicts grasp poses and reconstructions, respectively. Best results are shown in **bold**.

Method	Output		Seen			Similar			Novel		
	G	R	AP	AP _{0.8}	AP _{0.4}	AP	AP _{0.8}	AP _{0.4}	AP	AP _{0.8}	AP _{0.4}
GG-CNN (Morrison et al., 2018)	✓		15.48	21.84	10.25	13.26	18.37	4.62	5.52	5.93	1.86
Chuet al. (Chu et al., 2018)	✓		15.97	23.66	10.80	15.41	20.21	7.06	7.64	8.69	2.52
CenterGrasp (Chisari et al., 2024)	✓	✓	16.46	20.24	11.74	9.52	11.92	5.71	1.60	1.89	1.12
GPD (Ten Pas et al., 2017)	✓		22.87	28.53	12.84	21.33	27.83	9.64	8.24	8.89	2.67
Lianet al. (Liang et al., 2019)	✓		25.96	33.01	15.37	22.68	29.15	10.76	9.23	9.89	2.74
GraspNet (Fang et al., 2020)	✓		27.56	33.43	16.59	26.11	34.18	14.23	10.55	11.25	3.98
GSNet (Wang et al., 2021)	✓		67.12	78.46	60.90	54.81	66.72	46.17	24.31	30.52	14.23
Maet al. (Ma and Huang, 2023)	✓		63.83	74.25	58.66	58.46	70.05	51.32	24.63	31.05	12.85
HGGD (Chen et al., 2024)	✓		64.45	72.81	61.16	53.59	64.12	45.91	24.59	30.46	15.58
EconomicGrasp (Wu et al., 2024)	✓		68.21	79.60	63.54	61.19	73.60	53.77	25.48	31.46	13.85
MG-Grasp (Wang et al., 2026)	✓	✓	66.80	-	-	57.35	-	-	23.22	-	-
ZeroGrasp (Iwase et al., 2025)	✓	✓	72.43	83.12	65.57	65.45	78.32	55.48	28.49	34.21	15.80
GraspFoM (Ours)	✓	✓	78.87	89.60	71.13	73.34	87.84	64.12	54.32	64.01	29.36

4 Experiments

4.1 Setup

Settings and metrics. Following prior works, we evaluate GraspFoM on two complementary tasks: grasp pose prediction and 3D object reconstruction. For grasp pose prediction, we follow the official GraspNet-1Billion benchmark and report AP, AP_{0.8}, and AP_{0.4} on the Seen, Similar, and Novel splits. For reconstruction, we evaluate the quality of the predicted 3D object geometry on GraspNet-1B using Chamfer Distance (CD), F1-Score@10mm (F1), and Normal Consistency (NC).

Datasets. We conduct all experiments on the GraspNet-1Billion benchmark (Fang et al., 2020). Unlike recent methods (Iwase et al., 2025) that rely on additional synthetic datasets, GraspFoM is trained using only the official GraspNet-1B training split. For grasp pose prediction, we evaluate on the standard Seen, Similar, and Novel benchmark splits. For reconstruction, we use the object-level 3D supervision available in GraspNet-1B to compare predicted geometry against ground-truth shapes. The ReOcS dataset collected in ZeroGrasp (Iwase et al., 2025) has not been released, which prevents us from further comparison.

Baselines. For grasp pose prediction, we compare with representative scene-level grasping methods, including GG-CNN (Morrison et al., 2018), Chuet al. (Chu et al., 2018), GPD (Ten Pas et al., 2017), PointNet-GPD (Liang et al., 2019), GraspNet (Fang et al., 2020), GSNet (Wang et al., 2021), HGGD (Chen et al., 2024), EconomicGrasp (Wu et al., 2024), and Maet al. (Ma and Huang, 2023). We also compare with reconstruction-aware or unified pipelines, including CenterGrasp (Chisari et al., 2024), ZeroGrasp (Iwase et al., 2025), and MG-Grasp (Not Open Sourced) (Wang et al., 2026). For 3D reconstruction, we compare with sparse-structure baselines including MinkowskiNet (Choy et al., 2019), OCNN (Wang et al., 2020), OctMAE (Iwase et al., 2024), and ZeroGrasp (Iwase et al., 2025). These baselines cover direct grasp detection, geometry-aware grasping, and unified reconstruction-grasping approaches.

Due to space limitations, we defer model details and training details to the supplementary material. **Please refer to the appendix for additional experiments and visualization results.**

4.2 Main Results

Grasp Pose Prediction. Tab. 1 reports quantitative results on the GraspNet-1B benchmark. GraspFoM achieves the strongest overall performance across all benchmark splits and all reported friction settings. More importantly, the gain is especially pronounced on the Novel split, which is the most challenging setting because

Table 2 3D reconstruction on GraspNet-1B. Seg. indicates segmented reconstruction. Best results are in **bold**.

Method	Segmented Reconstruction	GraspNet-1B		
		CD↓	F1↑	NC↑
Minkowski (Choy et al., 2019)	✓	6.84	81.45	77.89
OCNN (Wang et al., 2020)	✓	7.23	82.22	78.44
OctMAE (Iwase et al., 2024)		7.57	78.38	75.19
ZeroGrasp (Iwase et al., 2025)	✓	6.05	84.08	78.46
GraspFoM (Ours)	✓	2.74	96.08	87.18

Table 3 Ablation study. We evaluate anchor number, diffusion horizon, refinement steps, and reconstruction update. Light green rows denote default settings.

Grasp Parameter Variants	Seen AP↑	Similar AP↑	Novel AP↑
Default setting ($K = 8, T_{\max} = 50, N_{\text{ref}} = 4$)	78.87	73.34	54.32
<i>Anchor Number (K)</i>			
$K = 4$	75.12	68.45	48.60
$K = 12$	77.05	71.12	51.95
<i>Diffusion Horizon (T_{\max})</i>			
$T_{\max} = 40$	74.58	67.90	47.85
$T_{\max} = 60$	77.40	71.85	52.10
<i>Refinement Steps (N_{ref})</i>			
$N_{\text{ref}} = 2$	73.20	65.50	45.10
$N_{\text{ref}} = 6$	77.65	72.05	52.80
Reconstruction Ablation	CD↓	F1↑	NC↑
<i>Reconstruction Update</i>			
w/o Reconstruction Update	3.58	91.25	80.45
Full GraspFoM	2.74	96.08	87.18

the object categories are not observed during training. This result suggests that the learned representation is not merely fitting training-set grasp patterns, but instead transfers to previously unseen objects through reusable object-level 3D priors.

Existing reconstruction-aware grasping methods typically learn task-specific latent spaces from scratch, whereas GraspFoM starts from transferable SAM3D priors and further builds shape-point fused features that combine local surface evidence with global shape context. This design is particularly beneficial under occlusion and partial observations, where purely local point neighborhoods are often insufficient to determine stable grasps. The strong improvement on novel objects supports this view.

Fig. 3 leads to the same conclusion. GraspFoM produces more reliable grasp poses with better contact regions and more stable approach directions. We attribute this behavior to the shared object latent, which jointly captures point-level local geometry and object-level global structure. As a result, the model can reason about graspable regions beyond visible local fragments and better preserve the overall object layout when predicting grasps.

Another notable aspect of this comparison is the training protocol. ZeroGrasp reports a stronger setting by pretraining on its additional ZeroGrasp-11B dataset and then fine-tuning on GraspNet-1B, while MG-Grasp uses sparse multi-view RGB observations with known camera poses and a different input setup. By contrast, GraspFoM is trained only on the official GraspNet-1B training split. The strong performance indicates that foundation priors can provide a more data-efficient route to robust grasp generalization.

3D Reconstruction. Tab. 2 reports reconstruction performance on GraspNet-1B. GraspFoM also achieves the best overall geometry quality, demonstrating that the proposed framework improves not only grasp prediction but also object reconstruction itself. This result is important because our method does not treat reconstruction merely as an auxiliary branch. Instead, reconstruction and grasping are optimized jointly through the shared latent, the reconstruction-aware scorer, and the residual latent updater.

Compared with sparse-structure baselines such as MinkowskiNet, OCNN, and OctMAE, GraspFoM benefits from stronger object priors and a reconstruction process that is explicitly shaped by grasp supervision. Compared with ZeroGrasp, which also performs joint reconstruction and grasp prediction, our method is built on top of transferable 3D foundation priors rather than a task-specific latent space learned from scratch. As a result, the predicted geometry is more favorable for both accurate surface recovery and downstream grasp reasoning.

Another notable observation is that the reconstruction improvement is consistent with the gain in grasp prediction. This suggests that the learned latent is not simply better for one task at the expense of the other. Instead, the two tasks reinforce each other: better reconstruction provides more grounded geometric

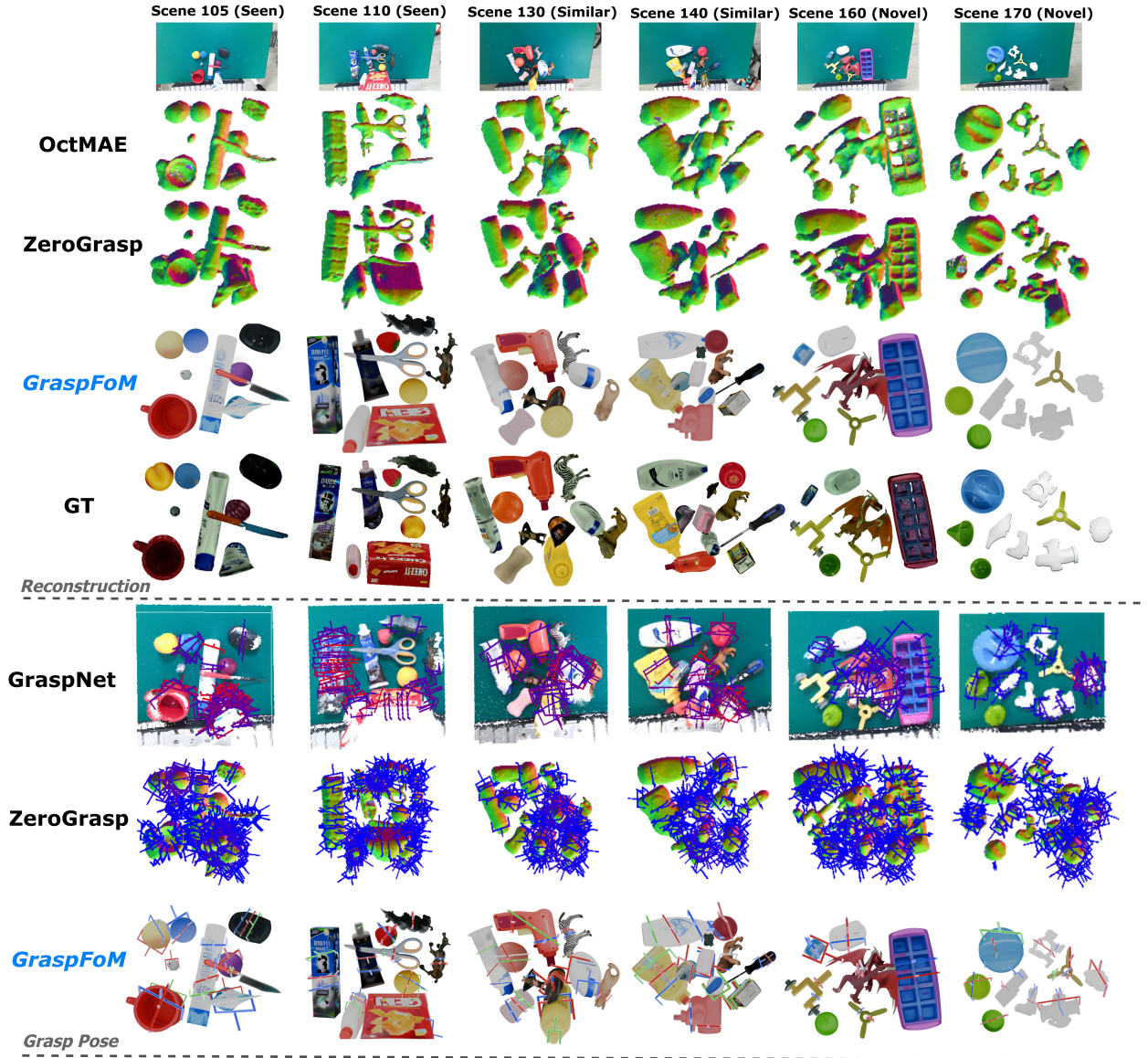


Figure 3 Reconstruction and grasping pose visualization comparison.

support for grasp generation, while grasp supervision encourages the latent to capture manipulation-relevant structure, such as graspable regions and affordance-aware cues. This mutual reinforcement is a key reason why GraspFoM is able to outperform prior unified frameworks on both tasks simultaneously.

4.3 Ablation Study

We conduct ablation studies on GraspNet-1B to examine the contributions of several key design choices in GraspFoM, including the anchor number K , the diffusion horizon T_{\max} , the number of refinement steps N_{ref} , and the reconstruction update. For the grasp generation branch, we report AP on the Seen, Similar, and Novel splits. For the reconstruction branch, we report CD, F1, and NC. Unless otherwise specified, the default configuration uses $K = 8$, $T_{\max} = 50$, and $N_{\text{ref}} = 4$, which is highlighted in light green in Tab. 3.

The results show that all three hyperparameters have a clear impact on grasp prediction performance. For anchor number, using $K = 8$ gives the best AP across all three splits, outperforming both a smaller setting ($K = 4$) and a larger one ($K = 12$). This suggests that too few anchors limit the diversity of candidate grasp hypotheses, while too many anchors introduce redundancy and make optimization less effective. A



Figure 4 In-the-wild reconstruction and grasp pose visualization comparison.

similar trend is observed for the diffusion horizon. Setting $T_{\max} = 50$ achieves the best results, whereas a shorter horizon ($T_{\max} = 40$) leads to clear degradation, indicating insufficient denoising capacity, and a longer horizon ($T_{\max} = 60$) also slightly hurts performance, suggesting diminishing returns from overly long diffusion trajectories. For iterative refinement, $N_{\text{ref}} = 4$ performs best. Reducing the number of refinement steps to 2 significantly weakens performance, especially on the Novel split, while increasing it to 6 does not bring further gains. This indicates that a moderate refinement budget is sufficient to capture most of the benefit, while excessive refinement yields limited improvement.

We further ablate the reconstruction update to verify the role of reconstruction-aware learning in GraspFoM. Removing this module degrades reconstruction quality from 2.74 to 3.58 in CD, from 96.08 to 91.25 in F1, and from 87.18 to 80.45 in NC. These results show that the reconstruction update is important for learning more accurate and geometrically consistent object representations. In turn, this stronger 3D representation also benefits grasp generation, supporting our core design that grasp reasoning should be built on top of high-quality object-centric 3D priors rather than only partial observations.

4.4 In-the-wild Evaluation

We further conduct qualitative evaluation on several object scenes selected from the WildRGB-D (Xia et al., 2024) dataset, using only the RGB images as input, as shown in Fig. 4. Compared with ZeroGrasp, GraspFoM produces more complete reconstructions and stable grasp poses. This advantage comes from the shared object latent, which combines local point cues with global shape context from 3D foundation priors for more robust reconstruction and grasp reasoning in open-world scenes.

5 Conclusion

In this paper, we present GraspFoM, a novel unified framework for robotic grasping with 3D foundation priors. We form a shared 3D object latent that jointly supports object reconstruction and grasp pose prediction. GraspFoM achieves SOTA performance on GraspNet-1B and simultaneously produces high-fidelity 3D assets.

Limitations and Future Work. The current framework mainly focuses on rigid objects. Incorporating structured kinematic priors, such as URDF (Li et al., 2025; Le et al., 2024), is a possible direction for articulated-object manipulation.

References

- G. Du, K. Wang, S. Lian, and K. Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, 54(3):1677–1734, 2021.
- M. Dong and J. Zhang. A review of robotic grasp detection technology. *Robotica*, 41(12):3846–3885, 2023.
- R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, et al. Deep learning approaches to grasp synthesis: A review. *IEEE Transactions on Robotics*, 39(5):3994–4015, 2023.
- H. Wang, X. Wei, Y. Li, Q. Wuwu, D. Wu, J. Cao, M. Lu, W. Zheng, and S. Zhang. Roboarms: High-quality robotic arm splatting via b-spline curve refinement. *arXiv preprint arXiv:2511.17961*, 2025.
- K. Xu, S. Zhao, Z. Zhou, Z. Li, H. Pi, Y. Zhu, Y. Wang, and R. Xiong. A joint modeling of vision-language-action for target-oriented grasping in clutter. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11597–11604. IEEE, 2023.
- S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, W. Zhang, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025.
- P. Li, Z. Wang, M. Liu, H. Liu, and C. Chen. Clickdiff: click to induce semantic contact map for controllable grasp generation with diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 273–281, 2024.
- Y. Li, X. Wei, X. Chi, Y. Li, Z. Zhao, H. Wang, N. Ma, M. Lu, and S. Han. Manipdreamer3d: Synthesizing plausible robotic manipulation video with occupancy-aware 3d trajectory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6644–6652, 2026.
- H.-S. Fang, C. Wang, M. Gou, and C. Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.
- A. Mousavian, C. Eppner, and D. Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2901–2910, 2019.
- C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15964–15973, 2021.
- M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13438–13444. IEEE, 2021.
- L. Zhou, H. Wang, Z. Zhang, Z. Liu, F. E. Tay, and M. H. Ang. You only scan once: A dynamic scene reconstruction pipeline for 6-dof robotic grasping of novel objects. In *2024 IEEE international conference on robotics and automation (ICRA)*, pages 13891–13897. IEEE, 2024.
- E. Chisari, N. Heppert, T. Welschhold, W. Burgard, and A. Valada. Centergrasp: Object-aware implicit representation learning for simultaneous shape reconstruction and 6-dof grasp estimation. *IEEE Robotics and Automation Letters*, 9(6):5094–5101, 2024.
- Q. Fan, Y. Cai, C. Li, W. He, X. Zheng, T. Lu, B. Liang, and S. Wang. Neugrasp: Generalizable neural surface reconstruction with background priors for material-agnostic object grasp detection. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3197–3203. IEEE, 2025.
- S. Iwase, M. Z. Irshad, K. Liu, V. Guizilini, R. Lee, T. Ikeda, A. Amma, K. Nishiwaki, K. Kitani, R. Ambrus, et al. Zerograsp: Zero-shot shape reconstruction enabled robotic grasping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17405–17415, 2025.
- K. Wang, S. Chen, C. Jiang, S. Shen, Y. Dai, and G. Wang. Mg-grasp: Metric-scale geometric 6-dof grasping framework with sparse rgb observations. *arXiv preprint arXiv:2603.16270*, 2026.
- R. Shao, W. Li, L. Zhang, R. Zhang, Z. Liu, R. Chen, and L. Nie. Large vlm-based vision-language-action models for robotic manipulation: A survey. *arXiv preprint arXiv:2508.13073*, 2025.

- J. Cao, Q. Zhang, P. Jia, X. Zhao, B. Lan, X. Zhang, X. Wei, S. Chen, L. Li, X. Liu, et al. Fastdrivevla: Efficient end-to-end driving via plug-and-play reconstruction-based token pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 2571–2579, 2026a.
- J. Cao, X. Zhang, X. Wei, L. Huang, W. Zijian, H. Zhang, Z. Jia, W. Mao, H. Wang, X. Liu, et al. Evodrivevla: Evolving autonomous driving vision-language-action model via collaborative perception-planning distillation. *arXiv preprint arXiv:2603.09465*, 2026b.
- S. Noh, J. Kim, D. Nam, S. Back, R. Kang, and K. Lee. GraspSAM: When segment anything model meets grasp detection. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14023–14029. IEEE, 2025.
- P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.
- L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. Volume rendering of neural implicit surfaces. *Advances in neural information processing systems*, 34:4805–4815, 2021.
- J. Cen, Z. Zhou, J. Fang, W. Shen, L. Xie, D. Jiang, X. Zhang, Q. Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36:25971–25990, 2023.
- X. Wei, R. Zhang, J. Wu, J. Liu, M. Lu, Y. Guo, and S. Zhang. Nto3d: Neural target object 3d reconstruction with segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20352–20362, 2024.
- N. Huang, X. Wei, W. Zheng, P. An, M. Lu, W. Zhan, M. Tomizuka, K. Keutzer, and S. Zhang. S3gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024.
- D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao. TripoSR: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.
- C. Yang, S. Li, J. Fang, R. Liang, L. Xie, X. Zhang, W. Shen, and Q. Tian. Gaussianobject: High-quality 3d object reconstruction from four views with gaussian splatting. *arXiv preprint arXiv:2402.10259*, 2024.
- J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.
- X. Wei, P. Chen, G. Li, M. Lu, H. Chen, and F. Tian. Gazegaussian: High-fidelity gaze redirection with 3d gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13293–13303, 2025.
- D. Wu, H. Li, and X. Wei. Dnrselect: Active best view selection for deferred neural rendering. *arXiv preprint arXiv:2501.12150*, 2025.
- G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.
- R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.
- M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.
- K. Zeng, Z. Wu, K. Xiong, X. Wei, X. Guo, Z. Zhu, K. Ho, L. Zhou, B. Zeng, M. Lu, et al. Rethinking driving world model as synthetic data generator for perception tasks. *arXiv preprint arXiv:2510.19195*, 2025.
- T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Z. Li, Q. Liang, X. Lin, Y. Ge, Z. Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.
- Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.
- K. Zhang, S. Bi, H. Tan, Y. Xiangli, N. Zhao, K. Sunkavalli, and Z. Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024.
- X. Chen, F.-J. Chu, P. Gleize, K. J. Liang, A. Sax, H. Tang, W. Wang, M. Guo, T. Hardin, X. Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

- B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.
- C. Pohl, K. Hitzler, R. Grimm, A. Zea, U. D. Hanebeck, and T. Asfour. Affordance-based grasping and manipulation in real world applications. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9569–9576. IEEE, 2020.
- A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal of Robotics Research*, 41(7):690–705, 2022.
- Y. Zhong, Q. Jiang, J. Yu, and Y. Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22584–22594, 2025.
- Y.-L. Wei, M. Lin, Y. Lin, J.-J. Jiang, X.-M. Wu, L.-A. Zeng, and W.-S. Zheng. Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11818–11828, 2025.
- Y. Yang, X. Wu, T. He, H. Zhao, and X. Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023.
- Y. Wang, X. Wei, M. Lu, and G. Kang. Plgs: Robust panoptic lifting with 3d gaussian splatting. *IEEE Transactions on Image Processing*, 2025.
- F. Williams, J. Huang, J. Swartz, G. Klar, V. Thakkar, M. Cong, X. Ren, R. Li, C. Fuji-Tsang, S. Fidler, et al. fvdb: A deep-learning framework for sparse, large scale, and high performance spatial intelligence. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024.
- P. Wang, Y. Liu, and X. Tong. Octree-based convolutional neural network, July 18 2023. US Patent 11,704,537.
- C. Eppner, A. Mousavian, and D. Fox. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE, 2021.
- J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.
- A. Agarwal, G. Singh, B. Sen, T. Lozano-Pérez, and L. Pack Kaelbling. Scenecomplete: Open-world 3d scene completion in complex real world environments for robot manipulation. *arXiv e-prints*, pages arXiv–2410, 2024.
- H. Qu, Y. Zhang, C.-H. Zhang, M. Huang, and C. Hua. Sam-assisted 3d reconstruction: A novel framework for unseen object modeling in unknown environments. In *2024 China Automation Congress (CAC)*, pages 6732–6737. IEEE, 2024.
- Y. Li, X. Wei, X. Chi, Y. Li, Z. Zhao, H. Wang, N. Ma, M. Lu, and S. Zhang. Manipdreamer: Boosting robotic manipulation world model with action tree and visual guidance. *arXiv preprint arXiv:2504.16464*, 2025.
- H. Wang, X. Wei, X. Zhang, J. Li, C. Bai, Y. Li, M. Lu, W. Zheng, and S. Zhang. Embodiedocc++: Boosting embodied 3d occupancy prediction with plane regularization and uncertainty sampler. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 925–934, 2025.
- X. Wei, X. Zhang, H. Wang, Q. Wuwu, M. Lu, W. Zheng, and S. Zhang. Omniindoor3d: Comprehensive indoor 3d reconstruction. *arXiv preprint arXiv:2505.20610*, 2025.
- J. Huang, Z. Gojcic, M. Atzmon, O. Litany, S. Fidler, and F. Williams. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4369–4379, 2023.
- N. Heppert, M. Z. Irshad, S. Zakharov, K. Liu, R. A. Ambrus, J. Bohg, A. Valada, and T. Kollar. Carto: Category and joint agnostic reconstruction of articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21201–21210, 2023.
- E. R. Chan, K. Nagano, M. A. Chan, A. W. Bergman, J. J. Park, A. Levy, M. Aittala, S. De Mello, T. Karras, and G. Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023.

- Y. Avigal, S. Paradis, and H. Zhang. 6-dof grasp planning using fast 3d reconstruction and grasp quality cnn. *arXiv preprint arXiv:2009.08618*, 2020.
- Y. Edelstein, O. Patashnik, D. Cohen-Bar, and L. Zelnik-Manor. Sharp-it: A multi-view to multi-view diffusion model for 3d synthesis and manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21458–21468, 2025.
- H. Ma, M. Shi, B. Gao, and D. Huang. Generalizing 6-dof grasp detection via domain prior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18102–18111, 2024.
- Z. Liu, Y. Feng, M. J. Black, D. Nowrouzezahrai, L. Paull, and W. Liu. Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133*, 2023.
- D. Morrison, P. Corke, and J. Leitner. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. *arXiv preprint arXiv:1804.05172*, 2018.
- F.-J. Chu, R. Xu, and P. A. Vela. Real-world multiobject, multigrasp detection. *IEEE Robotics and Automation Letters*, 3(4):3355–3362, 2018.
- A. Ten Pas, M. Gualtieri, K. Saenko, and R. Platt. Grasp pose detection in point clouds. *The International Journal of Robotics Research*, 36(13-14):1455–1473, 2017.
- H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang. Pointnetgpd: Detecting grasp configurations from point sets. In *2019 international conference on robotics and automation (ICRA)*, pages 3629–3635. IEEE, 2019.
- S. Chen, W. Tang, P. Xie, W. Yang, and G. Wang. Efficient heatmap-guided 6-dof grasp detection in cluttered scenes. *arXiv preprint arXiv:2403.18546*, 2024.
- X.-M. Wu, J.-F. Cai, J.-J. Jiang, D. Zheng, Y.-L. Wei, and W.-S. Zheng. An economic framework for 6-dof grasp detection. In *European Conference on Computer Vision*, pages 357–375. Springer, 2024.
- H. Ma and D. Huang. Towards scale balanced 6-dof grasp detection in cluttered scenes. In *Conference on robot learning*, pages 2004–2013. PMLR, 2023.
- C. Choy, J. Gwak, and S. Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019.
- P.-S. Wang, Y. Liu, and X. Tong. Deep octree-based cnns with output-guided skip connections for 3d shape and scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 266–267, 2020.
- S. Iwase, K. Liu, V. Guizilini, A. Gaidon, K. Kitani, R. Ambrus, and S. Zakharov. Zero-shot multi-object scene completion. In *European Conference on Computer Vision*, pages 96–113. Springer, 2024.
- H. Xia, Y. Fu, S. Liu, and X. Wang. Rgb-d objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22389, 2024.
- Z. Li, X. Bai, J. Zhang, Z. Wu, C. Xu, Y. Li, C. Hou, and S. Zhang. Urdf-anything: Constructing articulated objects with 3d multimodal language model. *arXiv preprint arXiv:2511.00940*, 2025.
- L. Le, J. Xie, W. Liang, H.-J. Wang, Y. Yang, Y. J. Ma, K. Vedder, A. Krishna, D. Jayaraman, and E. Eaton. Articulate-anything: Automatic modeling of articulated objects via a vision-language foundation model. *arXiv preprint arXiv:2410.13882*, 2024.

A Supplementary Material

A.1 Overview

The supplementary material includes the following components:

- Data Preprocessing
- Training Details
- Implementation Details
- Additional Visualization Results

A.2 Data Preprocessing

Our pipeline is built on top of object-centric outputs from the SAM3D reconstruction. For each training step, we use the RGB image together with the object mask to obtain an object-aligned 3D representation, including a pointmap and a compact volumetric shape latent. To reduce training overhead, these intermediate outputs are cached offline and reused during grasp training. As a result, the grasp model is trained on shared 3D object latents rather than repeatedly invoking the reconstruction backbone online.

For grasp supervision, we use GraspNet-1B annotations and associate them with object-level query points. In the current implementation, the query points are expressed in the camera coordinate frame. To control memory and improve training stability, the raw object point cloud is further reduced by voxel-based downsampling, with a default voxel size of 5 mm.

The grasp pose labels are converted into a diffusion-friendly representation before training. Specifically, each grasp pose is parameterized by a 6D rotation representation and a grasp-depth term, which are then normalized into a bounded target space. The finger opening is not directly included in the diffusion target; instead, it is recovered later from local object geometry during inference. To model multimodal pose structure, we additionally construct a compact anchor set by collecting normalized ground-truth pose targets from the training set and applying offline k -means clustering. These anchors serve as coarse pose modes for subsequent classification and initialization.

A.3 Implementation Details

Architecture. Our grasp branch is trained on top of cached SAM3D pointmaps and shape latents, so that all predictions are built on a shared 3D object representation. The shared latent is represented as an 8-channel volumetric grid with resolution $16 \times 16 \times 16$. For each query point, we concatenate its 3D coordinates and surface normal as input to a lightweight point encoder. Local latent features are obtained by trilinear sampling from the latent grid, while global shape context is introduced through cross-attention with latent tokens. The resulting fused point token is used for graspness, quality, affordance, and pose prediction.

Training. For pose learning, we adopt anchor-initialized truncated diffusion. Each positive query point is assigned to its nearest anchor in the normalized diffusion space, which provides supervision for anchor classification. Given a clean pose target x_0 , we sample a truncated timestep and generate a noisy state x_t , and train the pose reasoner to recover x_0 from (x_t, t) . The overall objective is a weighted combination of graspness, grasp quality, affordance, anchor classification, and diffusion denoising losses. The implementation additionally supports an optional reconstruction-aware latent refinement branch, denoted as *update_Reconstruction*, with an auxiliary surface consistency loss.

Inference. Optimization is performed with AdamW for 20 epochs using a learning rate of 10^{-4} and a batch size of 32. Training is conducted with distributed data parallelism on 8 GPUs. At inference time, grasp generation starts from the selected anchor with Gaussian perturbation and is refined by truncated DDIM. Unless otherwise stated, the default setup uses a diffusion horizon of 50, initializes reverse denoising from $t_{\text{init}} = 16$, and performs 5 refinement steps. The finger opening is estimated separately from local object geometry along the predicted gripper width axis, rather than being directly generated by the diffusion branch.



Figure 5 Additional qualitative results on occluded and poor-view objects.

A.4 Additional Visualization Results

We further provide qualitative results on several challenging object scenes with severe occlusion or unfavorable viewpoints, as shown in Fig. 5.

These examples further verify the robustness of GraspFoM under partial observations. Even when only limited visible evidence is available, our method still reconstructs geometrically coherent object shapes and predicts stable grasp poses with reasonable contact regions and approach directions. This advantage comes from the shared object latent, which combines local point cues with global shape context from transferable 3D foundation priors, making grasp reasoning less dependent on directly visible local fragments. As a result, GraspFoM not only produces more complete reconstructions but also yields grasps that are better aligned with the recovered object structure.

A.5 Additional Material Description

In the supplementary folder, we additionally provide representative case materials, including the input RGB images, corresponding output results, and mesh files for qualitative visualization. We also include a compact release of the core method code for presentation and discussion, covering the main modules of point-shape fusion, pose reasoning, and diffusion-based grasp refinement, so that the overall input-output pipeline and the key implementation of our method can be more clearly understood.