Physically Grounded 3D Generative Reconstruction under Hand Occlusion using Proprioception and Multi-Contact Touch

Gabriele M. Caddeo[1], Pasquale Marra[1], and Lorenzo Natale[1]

Istituto Italiano di Tecnologia
{gabriele.caddeo,pasquale.marra,lorenzo.natale}@iit.it
https://hsp.iit.it/

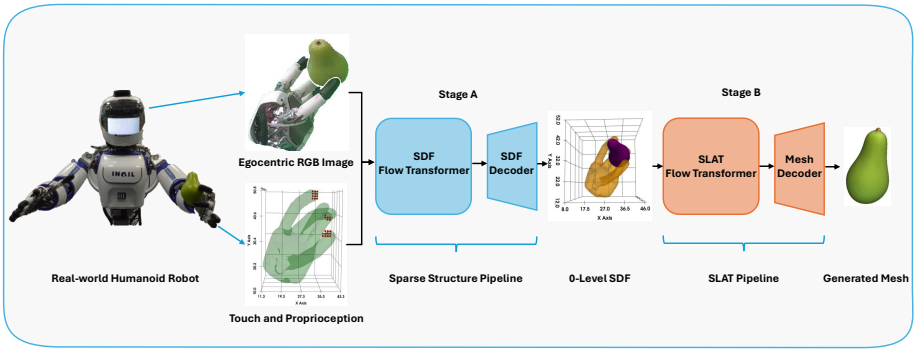

Fig. 1: Inference pipeline. We present a method for physically plausible 3D shape reconstruction by fusing vision and contact. Specifically, we use active contact information from tactile sensors distributed on the hand, and negative information, i.e., non-interpenetration, from the hand geometry. In addition, the method requires an egocentric RGB image of the object, together with the hand and object masks. A Flow Transformer conditioned on multiple senses then generates a physically consistent SDF on a 3D grid in the same geometric domain as the hand. Finally, a Structured Latents (**SLat**) pipeline can generate a refined textured mesh, directly at metric scale.

Abstract. We propose a multimodal, physically grounded approach for metric-scale amodal object reconstruction and pose estimation under severe hand occlusion. Unlike prior occlusion-aware 3D generation methods that rely only on vision, we leverage physical interaction signals: proprioception provides the posed hand geometry, and multi-contact touch constrains where the object surface must lie, reducing ambiguity in occluded regions. We represent object structure as a pose-aware, camera-aligned signed distance field (SDF) and learn a compact latent space with a Structure-VAE. In this latent space, we train a conditional flow-matching diffusion model, pretraining on vision-only images and finetuning on occluded manipulation scenes while conditioning on visible RGB evidence, occluder/visibility masks, the hand latent representation, and tactile information. Crucially, we incorporate physics-based objectives and differentiable decoder-guidance during finetuning and inference to reduce hand–object interpenetration and to align the

reconstructed surface with contact observations. Because our method produces a metric, physically consistent structure estimate, it integrates naturally into existing two-stage reconstruction pipelines, where a downstream module refines geometry and predicts appearance. Simulation experiments show that adding proprioception and touch substantially improves completion under occlusion and yields physically plausible reconstructions at correct real-world scale compared to vision-only baselines; we further validate transfer by deploying the model on a real humanoid robot with an end-effector different from those used during training. See `https://github.com/hsp-iit/physical-generative-reconstruction`

Keywords: Physical AI · 3D reconstruction · Robotics

1 Introduction

Accurate object geometry is important to guide robotic manipulation. This includes knowledge of the object surface to compute stable contacts, free space for collision avoidance, and object shape to plan subsequent actions. Yet the moments when geometry matters most are also when vision is least reliable. During grasping and in-hand manipulation, the hand severely occludes the object, and single-view RGB reconstruction becomes fundamentally underconstrained. Vision-only reconstructions may look plausible but remain unusable for control, e.g., intersecting the hand, missing true contact regions, or drifting in metric scale. Manipulation naturally produces *physical evidence*. Contacts and known hand kinematics provide useful cues to resolve the ambiguities caused by occlusion. Tactile sensing has long been used to estimate object properties [3, 41], pose [4], and to improve robustness of grasping [31, 53] and in-hand manipulation [47, 58], and recent advances in tactile hardware and learning-based models have made touch increasingly practical as a perception signal. Yet these cues are rarely integrated as *geometric constraints* for dense 3D reconstruction under occlusion. Critically, contact provides direct constraints on geometry exactly where vision is missing: observed touches indicate where the surface must be, while physical feasibility (*e.g.*, non-interpenetration with the end-effector) constrains where the object cannot be.

Despite their relevance, such interaction cues are still not commonly integrated as explicit geometric constraints in dense 3D reconstruction under severe occlusion. In parallel, recent diffusion and flow-based 3D generative models have improved single-view 3D inference by learning strong shape priors, but they are typically optimized for visual plausibility and are rarely grounded in physical constraints [35]. Under heavy occlusion, this can produce samples that violate contact feasibility or drift in scale, limiting their impact for downstream robotic tasks (*e.g.*, insertion [64] or pose estimation [72]).

We address this gap by treating in-hand 3D reconstruction as a **physically grounded** generative inference problem. Building upon [75], we propose a multimodal approach for metric-scale amodal object reconstruction under severe hand occlusion from a single egocentric RGB observation, proprioceptive hand pose,

and multi-contact tactile feedback. Our key contribution operates at the structure generation stage: we represent object geometry as a pose-aware, camera-aligned signed distance field (SDF), providing a continuous and differentiable representation that enables gradient-based physical constraints. We learn a compact latent space for SDF grids with a Structure-VAE and train a conditional flow-matching model to infer this latent. The model is pretrained on vision-only object images to learn a strong shape prior, then finetuned on occluded manipulation scenes while conditioning on visible RGB evidence and masks, a latent encoding of the posed hand, and a learned volumetric touch representation. To enforce physical validity, we incorporate two interaction-derived objectives: a contact-consistency loss that pulls the reconstructed surface toward observed contacts, and a non-interpenetration loss that pushes the object out of the hand interior. We further use these same terms as decoder-based physics guidance during sampling, steering generation toward shapes that satisfy contact and feasibility constraints. Since we reconstruct the object in the same 3D domain as the end-effector, the pose is implicitly estimated. Moreover, our method integrates naturally into state-of-the-art shape–appearance pipelines: the physically consistent structure can be passed to a downstream refinement/texturing module (made pose- and occlusion-aware) to obtain textured reconstructions consistent with the camera view and the occluding hand. Alternatively, our structure estimate can serve as input to other recent reconstruction frameworks (*e.g.*, [65]). Fig. 1 shows the inference pipeline of our approach.

Across simulation experiments, we show that integrating proprioception and touch substantially improves amodal completion under severe occlusion, producing reconstructions that are both visually plausible and physically consistent in real-world scale compared to vision-only baselines. Finally, we validate real-world transfer by deploying the model on a humanoid robot with an end-effector different from those used during training, demonstrating real-world applicability and generalization across embodiments. In summary, our contributions are:

- A method to inject physical constraints into a generative reconstruction pipeline.
- A pose-aware, camera-aligned SDF representation and multimodal conditioning that align 2D evidence, hand pose, and tactile contacts in a shared 3D grid.
- Simulation and real-robot results demonstrating improved amodal completion and cross-embodiment transfer relative to vision-only baselines.

2 Related Works

2.1 3D Generative Models

Early 3D generative approaches used GANs [26] to synthesize various 3D representations such as point clouds [29,36], NeRF-style radiance fields [6,7,46,52], or SDF [24] from visual inputs, but often struggled to generalize beyond the training distribution. With the emergence of Diffusion Models [28,59], subsequent works

explored representations such as point clouds [42, 43, 74], voxel grids [34, 44], Triplanes [9, 55, 80], or Gaussian mixtures [81], improving fidelity but frequently at high computational cost. To address efficiency, several recent methods generate shapes in a compact latent space [13, 50, 51, 66, 82], and rectified-flow formulations [1, 37, 38] have further accelerated sampling [75]. However, all these methods assume the objects are fully visible, without considering occlusions, limiting their use in real-world scenarios. In parallel, a growing body of work studies partial observability, reconstructing or generating 3D shape from incomplete inputs [16, 20]. Recent approaches [15, 65, 73] explicitly model occlusions by incorporating foreground masks into the generation pipeline. However, these methods typically rely on visual cues alone and do not leverage physical interaction signals to resolve ambiguities induced by occlusion. Physical constraints have also been incorporated into human–object and hand–object interaction, using human topology, hand pose, interaction priors, pose-aligned implicit representations, or optimization-based geometric constraints for monocular reconstruction [2, 12, 14, 39, 49, 79]. Related physics-based optimization has also been introduced in category-level 3D reconstruction [45,78] and object-agnostic reconstruction [11,27,35]. Separately, robotic tactile sensing has been used to improve 3D generation by fusing vision with contact observations [25]. In contrast, our approach integrates proprioception and multi-contact touch as 3D-space conditioning signals to guide occlusion-aware generation, enabling physically consistent completion and pose estimation.

2.2 3D Reconstruction in Robotics

In robotics, touch has long been used to estimate object properties [40]. Some works [30,67,77,83,84] performed extensive tactile exploration for object recognition and pose estimation. Later methods [18,54] inferred 3D shape from multiple contacts. While effective, purely tactile approaches are inherently local and lack the global context provided by vision. Accordingly, subsequent work fused vision and touch, initially on simple shapes [10, 61, 63, 70] and later in more complex scenes using modern neural 3D representations [18, 19, 62]. Robotic in-hand reconstruction has also been studied with RGB-D tracking and modeling of a grasped object [33], multi-finger contacts [56], active strategies that select informative interactions [57], and humanoid hands for SLAM-like reconstruction of novel objects during manipulation [60]. Conversely, we use touch and proprioception in the context of 3D generation under occlusion from a single RGB image, while remaining agnostic to the specific robotic hand and additionally estimating object scale and pose.

3 Method

Given a single egocentric observation of a robot manipulating an object, our goal is to reconstruct the object's 3D geometry at metric scale and recover its appearance despite severe occlusions induced by the hand, as shown in Fig. 1. For

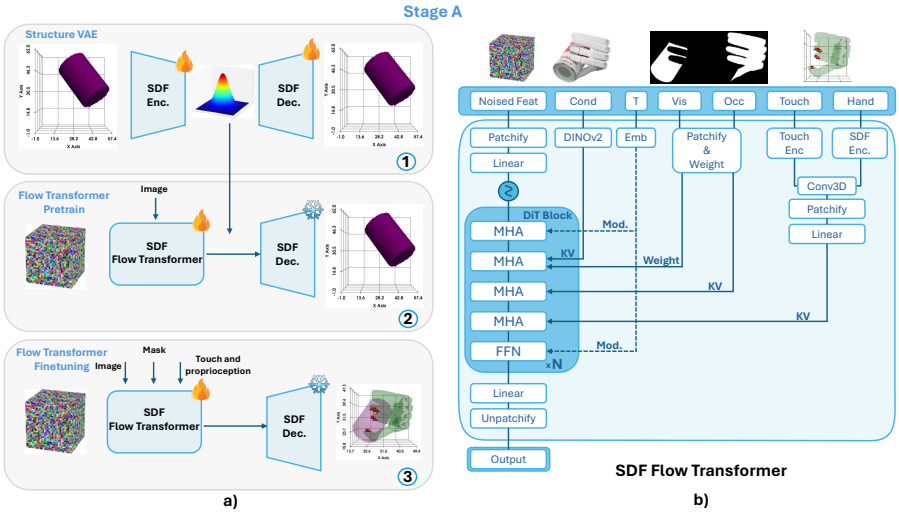

Fig. 2: Training procedure and Architecture. **a1)**: We train a Structure-VAE autoencoder that reconstructs pose-aware object SDFs. **a2)**: Using the frozen VAE encoder, we build latent datasets and train a conditional flow transformer from scratch using pose-consistent, unoccluded object images. **a3)**: We finetune the Structure-flow on occluded manipulation scenes, conditioning on visible RGB evidence, occluder/visibility masks, hand latent, and tactile features. **b)** The architecture of the Flow Transformer. Figure design inspired by [73].

each observation we are given: (i) an RGB image $I \in \mathbb{R}^{H\times W\times 3}$, (ii) joint angles $q \in \mathbb{R}^N$ defining the robot kinematic chain (and thus the camera–hand relative pose), (iii) a hand mesh $\mathcal{M}_h$ obtained via forward kinematics, and (iv) fingertip tactile readings u. Our approach focuses on **physically grounded** structure inference in a generative model. **Stage A** predicts a coarse object shape as a continuous signed distance field (SDF), conditioning on visible RGB evidence and, during finetuning and inference, leveraging proprioception and touch to impose contact and non-interpenetration constraints. To obtain a final textured reconstruction, we optionally pass the Stage-A output to **Stage B**, a SLat-style refinement module (Structured LAtens introduced in [75]) adapted to preserve pose and occlusion consistency with Stage A, or we pass the sampled voxelization to Sam3D [65] refinement stage (further discussion in Supp. Mat.). Fig. 2 shows the training procedure and the architecture of the Flow Transformer.
We next describe our dataset generation procedure (Sec. 3.1), then detail Stage A (Sec. 3.2), briefly describe Stage B (Sec. 3.3), and finally outline the end-to-end inference pipeline (Sec. 3.4). Additional details can be found in the Supp. Mat.

3.1 Data Generation

We train the two stages using synthetic object renders and simulated grasp scenes. For Stage A and Stage B pretraining we use isolated objects with clean

renders, while the finetuning stages use grasp scenes. All data are generated to enforce a deterministic correspondence between a rendered view and a canonical $\mathbb{R}^{R,R,R}$ 3D grid.

Rendering Clean and Occluded Conditioning Images For each object, we render RGBA images from a predefined set of camera viewpoints sampled on a sphere around the object. We normalize meshes to a canonical bounding box before rendering, and each rendered sample provides (i) an RGB(A) image, (ii) the camera pose, and (iii) the grid-alignment metadata used to construct the camera-aligned SDF grid. For finetuning, we generate simulated grasp scenes by placing the object in a posed hand configuration using the network introduced by [71]. We render an RGB image I together with instance segmentation masks for the visible object and the hand/occluder, denoted (M_o, M_h).

Pose-Aware Metric SDF Grids To ground the 3D generation in physics, we need a continuous representation of the scene. We represent object geometry with a signed distance field (SDF) $S_o : \Omega \to \mathbb{R}$ over a cubic domain $\Omega \subset \mathbb{R}^3$ discretized on an R^3 grid (we use $R = 64$), where the zero level set $\{\mathbf{x} : S_o(\mathbf{x}) = 0\}$ defines the surface. To deal with non-watertight meshes, we follow the approach of [69]. To maintain consistency between 2D evidence and the 3D representation, we define the grid frame to be *camera-aligned*: the grid $+z$ axis matches the viewing direction of the camera used to render the conditioning image, and the grid rotation is stored as metadata for each sample. The grid orientation is defined based on the views used for rendering the images, so that projecting the SDF along the grid z axis is consistent with the conditioning image.

During dataset generation we further apply random in-grid similarity transforms while ensuring the surface remains inside Ω. Concretely, we sample a downscaling factor $s_{aug} \sim \mathcal{U}(0.5, 1.0)$ (clipped to the maximum feasible value given the rotation and margin) and an in-plane translation (t_x, t_y) while centering along z, obtaining the augmented SDF. The augmented SDF is obtained by evaluating the canonical field at the inverse-warped coordinates and scaling distances accordingly:

$$S_o^{\text{aug}}(\mathbf{x}) = s_{aug} S_o\left(\frac{\mathbf{R}_g^\top(\mathbf{x} - \mathbf{t})}{s_{aug}}\right), \tag{1}$$

where $\mathbf{R}_g$ is the grid rotation and $\mathbf{t}$ is the translation in grid coordinates. These poses variations are necessary to match finetuning conditions: in grasp scenes, the object and hand share the same grid, so the object's size and location are determined by real-world scale and the hand–object relative pose, and are therefore not necessarily centered or tightly fit to the grid.

3.2 Stage A

Structure-VAE for SDF Reconstruction We learn a compact latent space for SDF grids using a variational autoencoder. Let $S \in \mathbb{R}^{1\times R\times R\times R}$ be an SDF

grid over $\Omega = [-1, 1]^3$. A 3D convolutional encoder E_ϕ parameterizes a diagonal Gaussian posterior $q_\phi(\mathbf{z} \mid S) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$ and produces a latent grid $\mathbf{z} \in \mathbb{R}^{C \times R' \times R' \times R'}$:

$$(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2) = E_\phi(S), \qquad \mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

A 3D convolutional decoder D_θ reconstructs the SDF grid:

$$\widehat{S} = D_\theta(\mathbf{z}). \tag{3}$$

Structure-VAE Objective We train the Structure-VAE to reconstruct SDF grids while encouraging valid signed-distance behavior. The reconstruction term is an ℓ_1 loss over the grid,

$$\mathcal{L}_{\mathrm{L1}} = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \left| \widehat{S}(\mathbf{x}) - S(\mathbf{x}) \right|. \tag{4}$$

We additionally impose two geometric regularizers computed from finite-difference SDF gradients (implemented with a fixed 3×3×3 convolution stencil). To avoid boundary artifacts from padding, we evaluate gradient-based terms only on an interior mask (dropping a one-voxel rim).

Eikonal Regularization. A signed distance field satisfies $\|\nabla S(\mathbf{x})\|_2 \approx 1$ almost everywhere, thus we minimize

$$\mathcal{L}_{\mathrm{eik}} = \mathbb{E}_{\mathbf{x}} \left(\left\| \nabla \widehat{S}(\mathbf{x}) \right\|_2 - 1 \right)^2, \tag{5}$$

where the expectation is taken over interior voxels.

Normal Consistency (Near-Surface). We encourage the predicted and ground-truth normal directions to agree in a narrow band $2 \times h$ around the surface:

$$\mathcal{L}_{\mathrm{n}} = \mathbb{E}_{\mathbf{x} \,:\, |S(\mathbf{x})| < 2h} \left(1 - \frac{\nabla \widehat{S}(\mathbf{x})}{\|\nabla \widehat{S}(\mathbf{x})\|_2} \cdot \frac{\nabla S(\mathbf{x})}{\|\nabla S(\mathbf{x})\|_2} \right), \tag{6}$$

again evaluated on interior voxels.

KL Divergence. We regularize the posterior $q_\phi(\mathbf{z} \mid S)$ toward $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with the standard KL term $\mathcal{L}_{\mathrm{KL}}$.
The overall training objective is

$$\mathcal{L}_{\mathrm{VAE}} = \lambda_{\mathrm{L1}} \mathcal{L}_{\mathrm{L1}} + \lambda_{\mathrm{eik}} \mathcal{L}_{\mathrm{eik}} + \lambda_{\mathrm{n}} \mathcal{L}_{\mathrm{n}} + \lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}}. \tag{7}$$

Flow Model Pre-train We train a conditional flow-matching model in the Structure-VAE latent space to predict an object latent grid $x_0 \in \mathbb{R}^{C\times R'\times R'\times R'}$ from observations. Following standard flow matching, we sample a timestep $t \in [0,1]$, draw noise $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, form a noisy latent x_t by linear interpolation between x_0 and ϵ (with a small minimum noise $\sigma_{\min}$), and train a denoiser f_θ to predict the corresponding velocity field. We rescale the continuous time $t \in [0,1]$ to $\tau = \alpha t$ before the timestep embedding. During pretraining, f_θ is conditioned on a clean rendered object image I that is cropped, placed to match the in-grid pose used to generate x_0, and then encoded using [48]. We optimize the flow-matching objective

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{x_0,t,\epsilon}\left[\left\| f_\theta(x_t, \alpha t;\, I) \;-\; \big((1-\sigma_{\min})\epsilon - x_0\big)\right\|_2^2\right], \tag{8}$$

Finetuning with Occlusion, Proprioception, and Touch We finetune the flow model on simulated grasp scenes, where the object is partially occluded by the hand. Each sample provides an RGB image I, a visible-object mask M_o, and a hand/occluder mask M_h. We condition the flow model on the visible evidence $I_o = I \odot M_o$ together with the masks (M_o, M_h), following [73] approach.

Proprioception determines the posed hand geometry in the same canonical grid as the object. We rasterize the posed hand as an SDF grid S_h and encode it with the same Structure-VAE encoder used for objects, obtaining a hand latent grid $x_{0,h} = E_\phi(S_h)$. Touch is represented as a two-channel 3D tensor $T = [C, D]$, where C is a binary contact-occupancy grid and D stores, for each voxel, its distance to the nearest contact voxel. While C provides sparse surface constraints, D supplies a dense, smooth field that helps learning by propagating contact information beyond the contacted voxels. To integrate touch, we encode T with a lightweight 3D CNN g_ψ and fuse the resulting feature volume with $x_{0,h}$ via concatenation followed by a $1\times1\times1$ projection:

$$\widetilde{x}_{0,h} = \mathrm{Conv}_{1\times1\times1}\big(\big[x_{0,h},\; g_\psi(T)\big]\big), \tag{9}$$

so that $\widetilde{x}_{0,h}$ reduces to $x_{0,h}$ when touch is disabled. The fused volume is patchified into tokens and provided to the denoiser via an additional cross-attention stream, alongside the image-based conditioning.

Physics-aware finetuning losses. In addition to the flow-matching regression loss $\mathcal{L}_{\mathrm{FM}}$, we impose auxiliary physical losses computed on decoded SDFs. Given the denoiser prediction, we form a clean latent estimate and decode it with the *frozen* Structure-VAE decoder to obtain the predicted object SDF $\widehat{S}_o$; we decode the hand latent $x_{0,h}$ once to obtain the hand SDF S_h. Because predictions at large diffusion timesteps are highly noisy, we downweight physics terms as a function of time using

$$w(t) = (1-t)^2, \tag{10}$$

and compute a weighted batch mean.

We penalize object mass inside the hand volume using a smooth interior saturation to avoid unstable gradients from deep penetration. Let τ be a saturation

threshold (we use $\tau = 0.1$) and define

$$\psi_\tau(s) = \tau \tanh\Big(\mathrm{ReLU}(s)/\tau\Big). \tag{11}$$

We compute the saturated interior field for the object and a binary hand-volume mask: $A_o(\mathbf{x}) = \psi_\tau(-\widehat{S}_o(\mathbf{x}))$ and $M_h(\mathbf{x}) = \mathbf{1}[\psi_\tau(-S_h(\mathbf{x})) > 0]$. The non-interpenetration loss is the mean object interior mass inside the hand volume,

$$\mathcal{L}_{\mathrm{NI}} = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{\mathbf{x}} A_{o,b}(\mathbf{x})\, M_{h,b}(\mathbf{x})}{\max(1, \sum_{\mathbf{x}} M_{h,b}(\mathbf{x}))}, \tag{12}$$

and we apply the time-weighted reduction described above.
Let $C(\mathbf{x}) \in \{0, 1\}$ denote the binary contact grid (first channel of T). We encourage the reconstructed surface to pass through contact locations by penalizing the SDF magnitude at contact voxels:

$$\mathcal{L}_{\mathrm{C}} = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{\mathbf{x}} C_b(\mathbf{x})\, |\widehat{S}_{o,b}(\mathbf{x})|}{\max(1, \sum_{\mathbf{x}} C_b(\mathbf{x}))}, \tag{13}$$

again using the time-weighted batch reduction.
The overall finetuning objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{NI}}\, \mathcal{L}_{\mathrm{NI}} + \lambda_{\mathrm{C}}\, \mathcal{L}_{\mathrm{C}}, \tag{14}$$

where λ_{NI} is warmed up during training and λ_{C} is fixed.

3.3 Stage B

Stage B is included primarily as an integration step to demonstrate that the metric-scale, physically consistent structure predicted by Stage A can be plugged into a standard refinement/texturing pipeline. We adopt a SLat-style representation following [75], with minimal adaptations for manipulation. We derive posed sparse voxel features from the Stage A prediction in the same camera-aligned canonical grid. SLAT decoding is performed in this pose-aligned frame and mapped to the world/camera frame using the known similarity transform from the pose/normalization metadata before rendering supervision. To predict the SLAT latent from an observation, we condition on the visible object evidence $I_o = I \odot M_o$ together with the hand/occluder mask M_h (as in [73]), and train with the same flow-matching objective as Stage A (Eq. (8)).

3.4 Physical guidance at inference

At inference time, we sample the Stage A latent with an explicit Euler solver driven by the learned velocity field $v_\vartheta(\cdot)$ under conditioning *cond* (occlusion-aware visual cues and, when available, hand/touch signals). To enforce physical

plausibility at inference time, we introduce an additive control term θ_k in the denoising process:

$$x_{k+1} = x_k - \Delta t_k \Big(v_\vartheta(x_k, t_k; \text{cond}) + \theta_k \Big). \tag{15}$$

When guidance is enabled, we decode the current latent estimate with the frozen Structure-VAE decoder to obtain $\widehat{S}_o$ (and decode the fixed hand latent once to obtain S_h), and define the physics energy

$$E(x_k) = \lambda_{\text{NI}} \, \mathcal{L}_{\text{NI}}\big(\widehat{S}_o, S_h\big) + \lambda_{\text{C}} \, \mathcal{L}_{\text{C}}\big(\widehat{S}_o, T\big). \tag{16}$$

We compute $g_k = \nabla_{x_k} E(x_k)$ and update the control term with an exponential moving average,

$$\theta_{k+1} \;=\; \beta \, \theta_k \;+\; \eta \, g_k, \tag{17}$$

where $\beta \in (0, 1)$ is a decay factor and $\eta > 0$ controls guidance strength. We apply standard stabilization heuristics (gradient normalization, a trust-region on $\|\theta_k\|$ relative to $\|v_\vartheta\|$, and optional projection to prevent increasing the contact energy) inspired by [68].

4 Experiments

4.1 Dataset and Metrics.

For VAE training and Flow Transformer pretraining, we use 3D-FUTURE [23], HSSD [32], ABO [17], and a subset of ObjaverseXL [21](100k objects). For the Flow Transformer finetuning, we use Google Scanned Objects [22]. We test on a subset of the YCB dataset [5] consisting of 36 objects. Grasps are generated as in Sec. 3.1, yielding 22,680 test images. We stratify results by occlusion level using K=5 equal-width bins over $x \in [0, 1]$ (for simplicity, we refer to them as B_k) with a non-even distribution of samples. We evaluate reconstruction using metrics commonly used in the literature: Chamfer Distance (**CD**), Normal consistency (**NC**), F-score with a 0.02 threshold (**F@0.02**), Voxel Intersection over Union (**Voxel-IoU**), and Earth Mover's Distance (**EMD**). For pose estimation, we use 3D Intersection over Union (**3D IoU**), **ICP-Rot** as defined in [65], Average Distance with Symmetry (**ADD-S**) from [76], and at a threshold of 0.1 **ADD-S@0.1**. Only valid output meshes are considered. More information and comparison with [2, 39] can be found in the Supp. Mat.

4.2 Simulation Results

3D reconstruction We compare our method against two representative approaches for 3D reconstruction under occlusion: Amodal3R [73] and SAM3D [65]. Amodal3R is a natural baseline for our setting, as it builds upon the same TRELLIS-style architecture, but relies on *vision-only* conditioning. SAM3D serves as a strong contemporary baseline for single-image 3D reconstruction. Both baselines are trained with larger datasets than ours, and SAM3D uses a larger model

Table 1: 3D reconstruction results in simulation.

		CD ↓	NC ↑	F@0.02 ↑	Voxel IoU ↑	EMD ↓
Amodal3R [73]	B_1	0.126	0.704	0.178	0.386	0.287
	B_2	0.150	0.674	0.188	0.331	0.291
	B_3	0.212	0.644	0.156	0.302	0.333
	B_4	0.263	0.624	0.120	0.295	0.374
	B_5	0.439	0.602	0.078	0.262	0.406
	All	0.188	0.669	0.162	0.339	0.314
Sam3D [65]	B_1	**0.020**	0.837	**0.265**	0.558	0.172
	B_2	0.025	0.811	**0.248**	0.514	0.184
	B_3	0.034	0.081	**0.198**	0.490	0.201
	B_4	0.057	0.775	**0.137**	0.433	0.239
	B_5	0.153	0.732	0.085	0.337	0.342
	All	0.039	0.803	**0.228**	0.504	0.202
Ours	B_1	0.021	**0.868**	0.215	**0.616**	**0.158**
	B_2	**0.022**	**0.844**	0.208	**0.585**	**0.172**
	B_3	**0.027**	**0.834**	0.180	**0.576**	**0.188**
	B_4	**0.055**	**0.805**	**0.137**	**0.526**	**0.229**
	B_5	**0.109**	**0.803**	**0.101**	**0.532**	**0.276**
	All	**0.033**	**0.844**	0.189	**0.586**	**0.184**

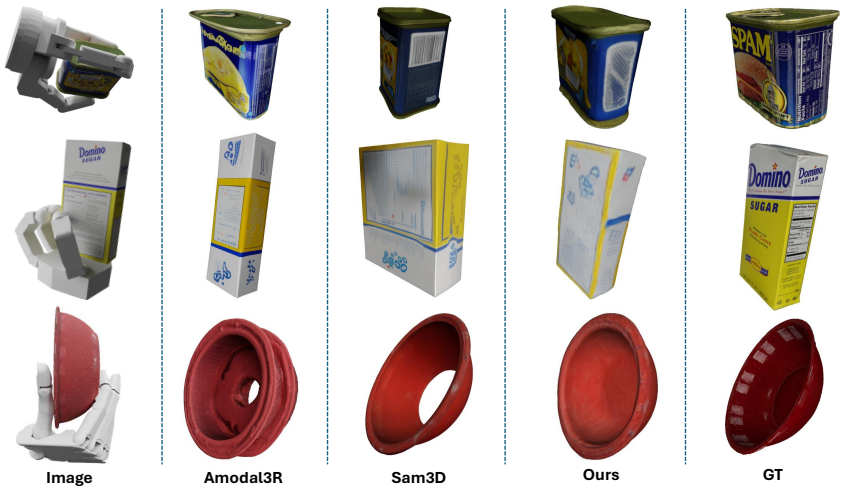

Fig. 3: Qualitative results in simulation. Vision-only baselines often exhibit artifacts (*e.g.*, holes or inconsistent relative dimensions) under occlusion. By leveraging contact cues, our method produces more physically plausible reconstructions.

capacity. Tab. 1 reports 3D reconstruction performance across occlusion bins and on the full evaluation set. Overall, our method and SAM3D outperform Amodal3R across metrics, highlighting the limitations of vision-only conditioning in heavily occluded in-hand scenarios. Compared to SAM3D, incorporating proprioception, touch, and physics-based constraints improves performance on all metrics, except in the least-occluded regime for **F@0.02**. This suggests that

when occlusion is mild, a strong vision-only model can match or slightly exceed our reconstruction quality, whereas its performance degrades more substantially as occlusion increases. Qualitative comparisons are shown in Fig. 3.

Table 2: Pose estimation results in simulation.

		3D IoU ↑	ICP-Rot ↓	ADD-S ↓	ADD-S@0.1 ↑
Sam3D [65]	B_1	0.453	15.848	0.080	0.776
	B_2	0.460	18.593	0.078	0.778
	B_3	0.413	19.610	0.091	0.656
	B_4	0.325	22.672	0.11	0.370
	B_5	0.150	35.013	0.19	0.099
	All	0.406	18.875	0.095	0.658
Ours	B_1	**0.595**	**7.067**	**0.062**	**0.864**
	B_2	**0.551**	**9.922**	**0.071**	**0.811**
	B_3	**0.512**	**12.434**	**0.082**	**0.721**
	B_4	**0.456**	**16.939**	**0.102**	**0.553**
	B_5	**0.356**	**18.199**	**0.143**	**0.258**
	All	**0.53**	**10**	**0.07**	**0.73**

Pose estimation For pose estimation, we compare against SAM3D [65]. Because SAM3D requires 3D information, we compare against the best case in which point maps are computed from ground-truth depth; this is because in noisy conditions, pose estimates of SAM3D are significantly less accurate. In contrast, our method does not require depth, as it leverages robot encoders to recover the camera–hand configuration, and uses touch as an additional constraint. Tab. 2 summarizes the results. Across all reported metrics, incorporating physical constraints during generation improves pose estimation accuracy. This is expected: proprioception constrains the hand pose directly, and multi-contact touch further reduces the set of object poses consistent with the observation by imposing contact and non-interpenetration feasibility.

4.3 Real-world Results

We evaluate real-world transfer by collecting data on a humanoid robot (Fig. 1) equipped with an end-effector different from the one used in simulation and instrumented with fingertip tactile sensors. We perform experiments on real counterparts of a subset of the objects used in simulation. For each trial, we acquire an egocentric RGB image and use calibrated camera–hand extrinsics (together with forward kinematics) to place the posed end-effector into the shared sparse-structure grid, alongside the tactile measurements. Tab. 3 reports quantitative results on reconstruction. Overall, SAM3D exhibits comparable performance under noisy real-world imagery w.r.t. the simulated scenario, while Amodal3R significantly suffers in the noisy scenario. Our method attains slightly lower scores in this setting, which we attribute primarily to encoder and calibration noise

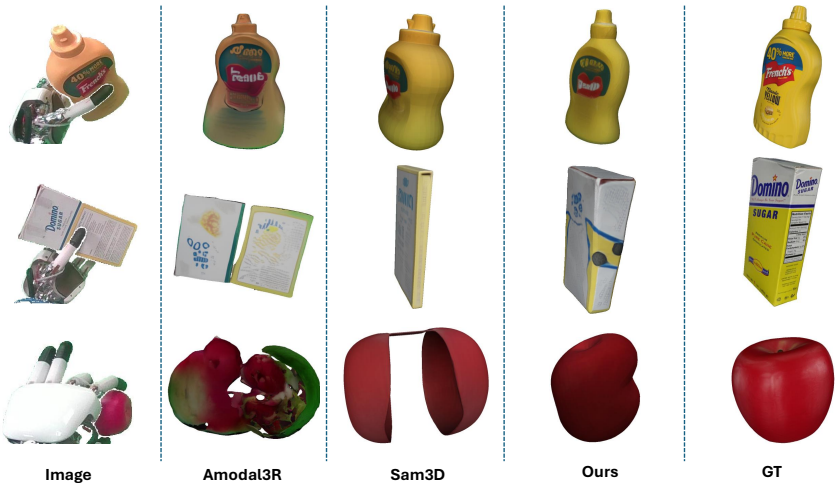

Fig. 4: Qualitative results on real-world data. Contact cues improve physical plausibility under occlusion. Artifacts can arise when camera–hand calibration/forward kinematics are inaccurate, leading to hand–object misalignment (third row).

affecting the hand pose and touch alignment. Qualitative examples are shown in Fig. 4.

Table 3: 3D reconstruction results in the real-world scenario.

	CD ↓	NC ↑	F@0.02 ↑	Voxel IoU ↑	EMD ↓
Amodal3R [73]	0.252	0.501	0.083	0.352	0.397
Sam3D [65]	0.037	0.888	**0.225**	0.510	**0.185**
Ours	**0.035**	**0.903**	0.186	**0.648**	0.190

4.4 Ablation Study

We perform ablations to quantify the contribution of sensing modalities, robustness to noisy inputs, and the impact of the main training and inference components. Results are summarized in Table 4, with aggregate metrics over all object categories. Additional qualitative examples and experiments on an unseen end-effector [8] are provided in the Supp. Mat.

Sensing ablation. We isolate each sensing modality by training reduced-input variants of Stage A. The *Vision-Only* variant uses only visual observations, making it comparable in information content to Amodal3R [73], up to differences in pose-consistent conditioning and mask distribution. It performs substantially worse than the full model, confirming that vision alone is insufficient under severe

Table 4: Ablation study for 3D reconstruction in simulation. We report aggregate results over all object categories.

Setting	CD ↓	NC ↑	F@0.02 ↑	Voxel IoU ↑	EMD ↓
Sensing modalities					
Vision-Only	0.142	0.696	0.116	0.357	0.296
No-Touch	0.085	0.760	0.126	0.430	0.264
Input noise					
Mask noise, level 1	0.033	0.835	0.180	0.572	0.187
Mask noise, level 2	0.035	0.806	0.178	0.537	0.199
Kinematic noise, level 1	0.037	0.809	0.177	0.541	0.190
Kinematic noise, level 2	0.042	0.798	0.171	0.501	0.212
Tactile noise, 3 mm	0.033	0.845	0.188	0.575	0.187
Tactile noise, 5 mm	0.036	0.854	0.186	0.563	0.183
Model components					
Inference without guidance	0.036	0.811	0.183	0.530	0.194
Training without physics losses	0.052	0.774	0.149	0.487	0.210
Binary contact alone	0.046	0.798	0.153	0.505	0.198
Distance contact alone	0.037	0.816	0.181	0.532	0.191
Ours	**0.033**	**0.844**	**0.189**	**0.586**	**0.184**

hand-object occlusion. Adding proprioception, i.e., vision plus posed end-effector geometry but no touch, clearly improves over Vision-Only by providing global geometric constraints. However, the *No-Touch* variant remains below the full model, showing that tactile cues provide complementary local surface evidence at contacts.

Robustness to noisy observations. We evaluate robustness by perturbing the inputs at test time. For tactile sensing, we add random spatial noise of up to 3 mm and 5 mm to the contact locations. A 3 mm perturbation has negligible impact, while 5 mm noise causes a small degradation in CD and voxel IoU. This behavior is consistent with the discretization of our representation: the hand–object scene spans tens of centimeters and is represented on a fixed R^3 grid, so small perturbations are partially absorbed by voxel quantization.

We further perturb the visual masks and kinematics to identify the main sources of error in the real-world setting. For masks, levels 1 and 2 correspond to erosion/dilation and jitter with maximum perturbations of 2 and 5 pixels, respectively. For kinematics, levels 1 and 2 correspond to maximum hand-eye calibration errors of 5 mm and 1 cm. The results show that the model is more sensitive to kinematic noise than to mask noise, which explains part of the performance drop observed in the real-world experiments.

Training and inference components. We also ablate inference guidance and the physics losses used during training. Removing inference guidance produces only a limited degradation in 3D reconstruction, suggesting that the learned model already captures most of the reconstruction signal. In contrast, removing the physics losses causes a substantial drop across all metrics, showing

that they are a crucial component of the training recipe. We then isolate the contribution of the tactile representation by using either binary contact or distance-based contact. Distance-based contact explains most of the tactile improvement, while combining it with binary contact yields the best overall performance.

5 Limitations

One limitation of our method is that it relies on accurate pose and calibration: camera–hand extrinsics, forward kinematics, and reliable occluder/visibility masks are required to align RGB images with the 3D grid, and errors in these signals can bias the reconstruction. In addition, operating on a fixed-resolution grid (64^3) limits geometric fidelity, particularly for thin structures and concavities, which may be lost in Stage A and only partially recovered downstream. This effect is amplified in manipulation scenes because the object is reconstructed in a shared grid with the end-effector, reducing the effective resolution available for the object. Finally, we notice that an object-centric frame could simplify fusing multiple views in dynamic scenes compared with the proposed camera-aligned frame. Further studies could investigate this potential issue.

6 Conclusion

We presented a multimodal method for metric-scale amodal object reconstruction under severe hand occlusion, and a practical way to inject physics into generative sparse-structure inference. Our approach fuses egocentric RGB evidence with proprioceptive hand geometry and multi-contact touch, representing shape as a pose-aware, camera-aligned SDF learned with a Structure-VAE. A conditional flow-matching model is pretrained on vision-only data and finetuned on occluded manipulation scenes, while contact-consistency and non-interpenetration objectives are used both as training losses and as decoder-guided sampling terms at inference time, steering generation toward physically feasible shapes. Experiments in simulation show that incorporating proprioception and touch improves occlusion-robust geometry reconstruction over vision-only baselines and yields physically plausible shapes at the correct real-world scale. We further demonstrate transfer to a real humanoid robot with an unseen end-effector, highlighting the benefit of expressing interaction cues in a shared 3D frame.

Future work includes scaling to larger, more diverse object–interaction datasets, developing more efficient conditioning than multi-stream cross-attention, and integrating contact cues more tightly in a single-stage pipeline. Beyond geometry, touch could also be used to infer physical properties such as friction or compliance, enabling reconstructions that better predict interaction dynamics.

Acknowledgements. This work was supported by the National Institute for Insurance against Accidents at Work (INAIL) project ergoCub-core. We thank Lorenzo Paiola for the inspiring and valuable discussions.

References

1. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: The Eleventh International Conference on Learning Representations (2023), https://arxiv.org/abs/2209.15571
2. Aytekin, A.I., Rhodin, H., Dabral, R., Theobalt, C.: Follow my hold: Hand-object interaction reconstruction through geometric guidance. International Conference on 3D Vision (2026)
3. Caddeo, G.M., Maracani, A., Alfano, P.D., Piga, N.A., Rosasco, L., Natale, L.: Sim2surf: A sim2real surface classifier for vision-based tactile sensors with a bilevel adaptation pipeline. IEEE Sensors Journal **25**(5), 8697–8709 (2025). https://doi.org/10.1109/JSEN.2025.3530712
4. Caddeo, G.M., Piga, N.A., Bottarel, F., Natale, L.: Collision-aware in-hand 6d object pose estimation using multiple vision-based tactile sensors. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 719–725 (2023). https://doi.org/10.1109/ICRA48891.2023.10160359
5. Calli, B., Walsman, A., Singh, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. IEEE Robotics Automation Magazine **22**(3), 36–52 (2015). https://doi.org/10.1109/MRA.2015.2448951
6. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16123–16133 (2022)
7. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5799–5809 (2021)
8. Chao, Y.-W., e.a.: Dexycb: A benchmark for capturing hand grasping of objects. In: CVPR 2021. pp. 9040–9049 (2021)
9. Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L., Su, H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2416–2425 (2023)
10. Chen, Y., Tekden, A.E., Deisenroth, M.P., Bekiroglu, Y.: Sliding touch-based exploration for modeling unknown object shape with multi-fingered hands. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8943–8950. IEEE (2023)
11. Chen, Y., Xie, T., Zong, Z., Li, X., Gao, F., Yang, Y., Wu, Y.N., Jiang, C.: Atlas3d: Physically constrained self-supporting text-to-3d for simulation and fabrication. Advances in Neural Information Processing Systems **37**, 102501–102530 (2024)
12. Chen, Z., Hasson, Y., Schmid, C., Laptev, I.: AlignSDF: Pose-Aligned signed distance fields for hand-object reconstruction. In: ECCV (2022)
13. Chen, Z., Tang, J., Dong, Y., Cao, Z., Hong, F., Lan, Y., Wang, T., Xie, H., Wu, T., Saito, S., et al.: 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26576–26586 (2025)
14. Chi, S., Sachdeva, E., Huang, P.H., Lee, K.: Contact-aware amodal completion for human-object interaction via multi-regional inpainting (2025), https://arxiv.org/abs/2508.00427

15. Cho, J., Youwang, K., Yang, H., Oh, T.H.: Robust 3d shape reconstruction in zero-shot from a single image in the wild. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22786–22798 (2025)
16. Chu, R., Xie, E., Mo, S., Li, Z., Nießner, M., Fu, C.W., Jia, J.: Diffcomplete: Diffusion-based generative 3d shape completion. Advances in Neural Information Processing Systems (2023)
17. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., et al.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21126–21136 (2022)
18. Comi, M., Lin, Y., Church, A., Tonioni, A., Aitchison, L., Lepora, N.F.: Touchsdf: A deepsdf approach for 3d shape reconstruction using vision-based tactile sensing. IEEE Robotics and Automation Letters **9**(6), 5719–5726 (2024)
19. Comi, M., Tonioni, A., Tremblay, J., Yang, M., Blukis, V., Lin, Y., Lepora, N.F., Aitchison, L.: Snap-it, tap-it, splat-it: Tactile-informed 3d gaussian splatting for reconstructing challenging surfaces. In: 2025 International Conference on 3D Vision (3DV). pp. 1134–1143. IEEE (2025)
20. Cui, R., Liu, W., Sun, W., Wang, S., Shang, T., Li, Y., Song, X., Yan, H., Wu, Z., Chen, S., et al.: Neusdfusion: A spatial-aware generative model for 3d shape completion, reconstruction, and generation. In: European Conference on Computer Vision. pp. 1–18. Springer (2024)
21. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems **36**, 35799–35813 (2023)
22. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. Ieee (2022)
23. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. International Journal of Computer Vision **129**(12), 3313–3337 (2021)
24. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. Advances in neural information processing systems **35**, 31841–31854 (2022)
25. Gao, R., Deng, K., Yang, G., Yuan, W., Zhu, J.Y.: Tactile dreamfusion: Exploiting tactile sensing for 3d generation. Advances in Neural Information Processing Systems **37**, 29839–29863 (2024)
26. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
27. Guo, M., Wang, B., Ma, P., Zhang, T., Owens, C., Gan, C., Tenenbaum, J., He, K., Matusik, W.: Physically compatible 3d object modeling from a single image. Advances in Neural Information Processing Systems **37**, 119260–119282 (2024)
28. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
29. Huang, Z., Yu, Y., Xu, J., Ni, F., Le, X.: Pf-net: Point fractal network for 3d point cloud completion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7662–7670 (2020)

30. Jamali, N., Ciliberto, C., Rosasco, L., Natale, L.: Active perception: Building objects' models using tactile exploration. pp. 179–185 (2016). https://doi.org/10.1109/HUMANOIDS.2016.7803275
31. Jiang, J., Cao, G., Butterworth, A., Do, T.T., Luo, S.: Where shall i touch? vision-guided tactile poking for transparent object grasping. IEEE/ASME Transactions on Mechatronics **28**, 233–244 (2022), https://api.semanticscholar.org/CorpusID:251718948
32. Khanna, M., Mao, Y., Jiang, H., Haresh, S., Shacklett, B., Batra, D., Clegg, A., Undersander, E., Chang, A.X., Savva, M.: Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16384–16393 (2024)
33. Krainin, M., Henry, P., Ren, X., Fox, D.: Manipulator and object tracking for in-hand 3d object modeling. I. J. Robotic Res. **30**, 1311–1327 (10 2011). https://doi.org/10.1177/0278364911403178
34. Li, M., Duan, Y., Zhou, J., Lu, J.: Diffusion-sdf: Text-to-shape via voxelized diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12642–12651 (2023)
35. Li, R., Zheng, C., Rupprecht, C., Vedaldi, A.: Dso: Aligning 3d generators with simulation feedback for physical soundness. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6772–6783 (2025)
36. Lin, C.H., Kong, C., Lucey, S.: Learning efficient point cloud generation for dense 3d object reconstruction. In: proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
37. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t
38. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
39. Liu, Y., Long, X., Yang, Z., Liu, Y., Habermann, M., Theobalt, C., Ma, Y., Wang, W.: EasyHOI: Unleashing the Power of Large Models for Reconstructing Hand-Object Interactions in the Wild . In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7037–7047. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2025). https://doi.org/10.1109/CVPR52734.2025.00660, https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00660
40. Luo, S., Lepora, N.F., Yuan, W., Althoefer, K., Cheng, G., Dahiya, R.: Tactile robotics: An outlook. IEEE Transactions on Robotics (2025)
41. Luo, S., Yuan, W., Adelson, E., Cohn, A.G., Fuentes, R.: Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 2722–2727 (2018). https://doi.org/10.1109/ICRA.2018.8460494
42. 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)
43. Melas-Kyriazi, L., Rupprecht, C., Vedaldi, A.: Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12923–12932 (2023)

44. Müller, N., Siddiqui, Y., Porzi, L., Bulo, S.R., Kontschieder, P., Nießner, M.: Diffrf: Rendering-guided 3d radiance field diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4328–4338 (2023)
45. Ni, J., Chen, Y., Jing, B., Jiang, N., Wang, B., Dai, B., Li, P., Zhu, Y., Zhu, S.C., Huang, S.: Phyrecon: Physically plausible neural scene reconstruction. Advances in Neural Information Processing Systems **37**, 25747–25780 (2024)
46. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11453–11464 (2021)
47. Oller, M., i Lisbona, M.P., Berenson, D., Fazeli, N.: Manipulation via membranes: High-resolution and highly deformable tactile sensing and control. In: 6th Annual Conference on Robot Learning (2022), https://openreview.net/forum?id=ft8IeFe4-8e
48. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193
49. Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
50. Ren, X., Huang, J., Zeng, X., Museth, K., Fidler, S., Williams, F.: Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4209–4219 (2024)
51. 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)
52. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. Advances in neural information processing systems **33**, 20154–20166 (2020)
53. Shahidzadeh, A.H., Caddeo, G.M., Alapati, K., Natale, L., Fermüler, C., Aloimonos, Y.: Feelanyforce: Estimating contact force feedback from tactile sensation for vision-based tactile sensors. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 251–257 (2025). https://doi.org/10.1109/ICRA55743.2025.11127723
54. Shahidzadeh, A.H., Yoo, S.J., Mantripragada, P., Singh, C.D., Fermüller, C., Aloimonos, Y.: Actexplore: Active tactile exploration on unknown objects. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 3411–3418. IEEE (2024)
55. Shue, J.R., Chan, E.R., Po, R., Ankner, Z., Wu, J., Wetzstein, G.: 3d neural field generation using triplane diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20875–20886 (2023)
56. Smith, E., Calandra, R., Romero, A., Gkioxari, G., Meger, D., Malik, J., Drozdzal, M.: 3d shape reconstruction from vision and touch. Advances in Neural Information Processing Systems **33**, 14193–14206 (2020)
57. Smith, E., Meger, D., Pineda, L., Calandra, R., Malik, J., Romero Soriano, A., Drozdzal, M.: Active 3d shape reconstruction from vision and touch. Advances in Neural Information Processing Systems **34**, 16064–16078 (2021)

58. Sodhi, P., Kaess, M., Mukadanr, M., Anderson, S.: Patchgraph: In-hand tactile tracking with learned surface normals. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2164–2170 (2022). https://doi.org/10.1109/ICRA46639.2022.9811953
59. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
60. Suresh, S., Qi, H., Wu, T., Fan, T., Pineda, L., Lambeta, M., Malik, J., Kalakrishnan, M., Calandra, R., Kaess, M., et al.: Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation. Science Robotics **9**(96), eadl0628 (2024)
61. Suresh, S., Si, Z., Mangelson, J.G., Yuan, W., Kaess, M.: Shapemap 3-d: Efficient shape mapping through dense touch and vision. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 7073–7080. IEEE (2022)
62. Swann, A., Strong, M., Do, W.K., Camps, G.S., Schwager, M., Kennedy, M.: Touchgs: Visual-tactile supervised 3d gaussian splatting. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10511–10518. IEEE (2024)
63. Tahoun, M., Tahri, O., Ramón, J.A.C., Mezouar, Y.: Visual-tactile fusion for 3d objects reconstruction from a single depth view and a single gripper touch for robotics tasks. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6786–6793. IEEE (2021)
64. Tang, B., Akinola, I., Xu, J., Wen, B., Handa, A., Wyk, K.V., Fox, D., Sukhatme, G.S., Ramos, F., Narang, Y.: Automate: Specialist and generalist assembly policies over diverse geometries (2024), https://arxiv.org/abs/2407.08028
65. Team, S.D., Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: Sam 3d: 3dfy anything in images (2025), https://arxiv.org/abs/2511.16624
66. Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K., et al.: Lion: Latent point diffusion models for 3d shape generation. Advances in Neural Information Processing Systems **35**, 10021–10039 (2022)
67. Vezzani, G., Pattacini, U., Battistelli, G., Chisci, L., Natale, L.: Memory unscented particle filter for 6-dof tactile localization. IEEE Transactions on Robotics **33**(5), 1139–1155 (2017). https://doi.org/10.1109/TRO.2017.2707092
68. Wang, L., Cheng, C., Liao, Y., Qu, Y., Liu, G.: Training free guided flow matching with optimal control. arXiv preprint arXiv:2410.18070 (2024)
69. Wang, P.S., Liu, Y., Tong, X.: Dual octree graph networks for learning adaptive volumetric shape representations. ACM Transactions on Graphics (SIGGRAPH) **41**(4) (2022)
70. Wang, S., Wu, J., Sun, X., Yuan, W., Freeman, W.T., Tenenbaum, J.B., Adelson, E.H.: 3d shape perception from monocular vision, touch, and shape priors. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1606–1613. IEEE (2018)
71. Wei, Z., Xu, Z., Guo, J., Hou, Y., Gao, C., Cai, Z., Luo, J., Shao, L.: $\mathcal{D}(\mathcal{R}, \mathcal{O})$ grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 4982–4988 (2025). https://doi.org/10.1109/ICRA55743.2025.11127754
72. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects (2024), https://arxiv.org/abs/2312.08344

73. Wu, T., Zheng, C., Guan, F., Vedaldi, A., Cham, T.J.: Amodal3r: Amodal 3d reconstruction from occluded 2d images (2025), https://arxiv.org/abs/2503.13439
74. Wu, Z., Wang, Y., Feng, M., Xie, H., Mian, A.: Sketch and text guided diffusion model for colored point cloud generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8929–8939 (2023)
75. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. arXiv preprint arXiv:2412.01506 (2024)
76. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes (2018), https://arxiv.org/abs/1711.00199
77. Xu, J., Lin, H., Song, S., Ciocarlie, M.T.: Tandem3d: Active tactile exploration for 3d object recognition. 2023 IEEE International Conference on Robotics and Automation (ICRA) pp. 10401–10407 (2022), https://api.semanticscholar.org/CorpusID:252368205
78. Yang, Y., Jia, B., Zhi, P., Huang, S.: Physcene: Physically interactable 3d scene synthesis for embodied ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16262–16272 (2024)
79. Ye, Y., Gupta, A., Tulsiani, S.: What's in your hands? 3d reconstruction of generic objects in hands. In: CVPR (2022)
80. Zhang, B., Cheng, Y., Wang, C., Zhang, T., Yang, J., Tang, Y., Zhao, F., Chen, D., Guo, B.: Rodinhd: High-fidelity 3d avatar generation with diffusion models. In: European Conference on Computer Vision. pp. 465–483. Springer (2024)
81. Zhang, B., Cheng, Y., Yang, J., Wang, C., Zhao, F., Tang, Y., Chen, D., Guo, B.: Gaussiancube: A structured and explicit radiance representation for 3d generative modeling. arXiv preprint arXiv:2403.19655 (2024)
82. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. ACM Transactions on Graphics (TOG) **43**(4), 1–20 (2024)
83. Zheng, H., Caddeo, G.M., Jalba, A.C., IJsselsteijn, W.A., Natale, L., Cuijpers, R.H.: Bayesian active object recognition and 6d pose estimation from multimodal contact sensing (2026), https://arxiv.org/abs/2603.21410
84. Zheng, H., Jalba, A., Cuijpers, R.H., IJsselsteijn, W., Schoenmakers, S.: A bayesian framework for active tactile object recognition, pose estimation and shape transfer learning (2024), https://arxiv.org/abs/2409.06912