TOUCH-HOI: ACTIVE TACTILE 3D RECONSTRUCTION FROM HAND–OBJECT INTERACTIONS

Anonymous authors
Paper under double-blind review

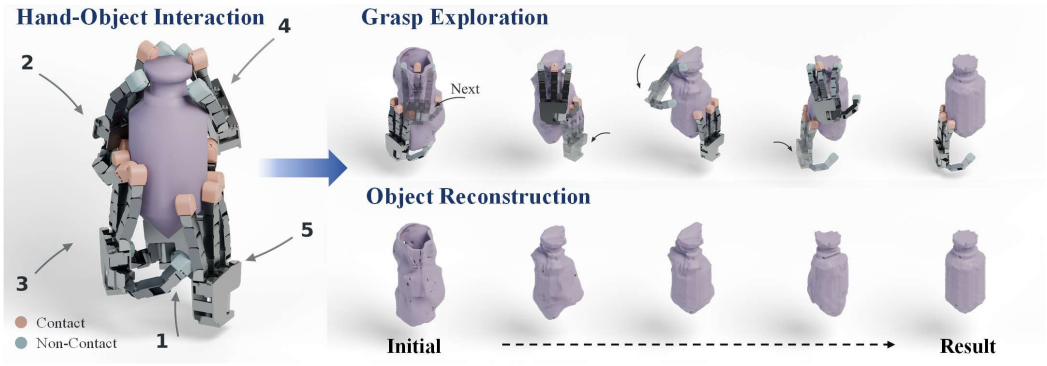

Figure 1: **Active tactile 3D reconstruction with Touch-HOI.** Sparse tactile observations admit multiple plausible object shapes. Touch-HOI integrates tactile sensing and proprioceptive cues to infer complete geometry and select subsequent grasps that resolve the remaining shape ambiguity.

ABSTRACT

Humans can infer an object's shape through touch, even when hidden inside an opaque container. For robots, tactile 3D reconstruction without visual input remains challenging because sparse contacts reveal only local surface patches. Existing methods underexploit structured hand–object interaction cues and lack explicit posterior representations to reason about shape uncertainty. In this paper, we present Touch-HOI, a unified generative framework for tactile 3D reconstruction with active exploration. Specifically, we formally unify tactile sensing and proprioception as **Hand–Object Interaction Evidence**, and encode their spatial distribution and statistics as **Hand–Object Interaction Encoding**. Building upon this, we construct a flow matching model with a pretrained 3D prior, establishing an explicit generative belief state over several shape hypotheses. Therefore, belief-guided grasp selection evaluates prospective tactile feedback via Bayesian reweighting to maximize expected geometric risk reduction without training an extra exploration policy. Extensive experiments on the ABC dataset demonstrate that Touch-HOI achieves mean Chamfer Distances of 2.399 with five random grasps and 2.093 with active selection, which verifies the effectiveness of the proposed Touch-HOI. The code will be made publicly available upon acceptance.

1 INTRODUCTION

Humans use eyes to perceive the world every day, and robot vision aiming to achieve this perception capability has been studied for decades. However, we also find that humans can infer the shape of an object through a series of touch attempts without sight, e.g., when it is hidden inside an opaque pocket or container. Can robots perceive the 3D world through touch as well?

Achieving the goal is challenging as sparse tactile contacts constrain only local surface patches, which casts touch-based 3D object shape perception as an inherently ill-posed problem. Therefore, existing methods (Dragiev et al., 2011; Comi et al., 2024; Suresh et al., 2024; Swann et al., 2024;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Gu et al., 2026) mainly use touch as a complementary cue to vision, where per-object optimization remains heavily constrained by sparse tactile feedback, thereby yielding suboptimal or inaccurate 3D geometries. Alternatively, learning-based methods address the ambiguity of sparse touch by learning geometric priors across objects. Specifically, deformation-based methods (Smith et al., 2020; 2021) use local tactile signals to guide mesh deformation for 3D shape reconstruction. Recently, generative approaches extend shape completion beyond a single prediction: Touch2Shape (Wang et al., 2025) combines latent diffusion with a learned exploration policy, and TouchDiffusion (Zhang et al., 2026) generates complete point clouds from tactile observations.

While the aforementioned methodologies have demonstrated efficacy, they primarily formulate reconstruction as a mapping from sparse tactile observations to 3D object shapes, leaving the underlying geometric evidence provided by hand-object interactions underexploited. As a result, these methods fail to fully exploit the structured geometric information inherent in each grasp, which includes both surface tactile cues and proprioceptive feedback from non-contacting fingers and the palm. Beyond individual grasps, the spatial distribution and contact frequency of successive interactions offer complementary evidence for assessing geometric uncertainty. Consequently, shape exploration naturally becomes a process of accumulating 3D spatial information from both free and occupied space.

Furthermore, active exploration selection in existing methods often lacks explicit reasoning over the probabilistic belief of plausible 3D shapes. Current active strategies predominantly fall into two paradigms, both failing to handle the inherent ambiguity of sparse touch. On one hand, deterministic coverage-based approaches (Kim et al., 2026) select actions to maximize surface coverage based on a single point-estimate prediction, which overlooks competing structural hypotheses compatible with the sparse observations. On the other hand, reinforcement learning frameworks (Smith et al., 2021; Wang et al., 2025) optimize interaction policies via implicit reconstruction or denoising rewards, lacking explicit posterior representations to directly measure how prospective interactions reduce shape uncertainty. Fundamentally, effective active touch requires evaluating how prospective interactions reduce shape ambiguity across candidate predictions. However, existing paradigms cannot directly measure posterior uncertainty, which limits their ability to target interactions that efficiently resolve geometric ambiguities.

To address these limitations, we present Touch-HOI, a unified generative framework for active 3D shape reconstruction driven by hand-object interactions. Rather than treating tactile cues in isolation, we observe that hand-object interactions naturally impose holistic information: tactile sensing delineates contact surface, and proprioception (non-contacting digits and the palm) constrains free space. We formally unify these cues into **Hand-Object Interaction Evidence**. Furthermore, to implicitly capture the varying uncertainty of tactile and proprioceptive cues across successive interactions, we encode the spatial distribution of tactile observations and contact statistics as **Hand-Object Interaction Encoding**. Building upon the encoding of hand-object interaction evidence, we construct a flow matching model (Lipman et al., 2023) with a pretrained 3D shape prior (Xiang et al., 2024), which effectively represents shape ambiguity in under-observed regions, thereby establishing an explicit generative belief state over candidate 3D shapes.

Building upon the generative belief states, we propose **belief-guided grasp selection** to actively resolve remaining geometric ambiguities, as shown in Figure 1. Specifically, for candidate actions in a predefined grasp pool, we predict tactile feedback across shape hypotheses to directly evaluate their expected impact on posterior uncertainty via Bayesian reweighting. We then select the interaction that guides grasp selection using expected geometric risk reduction under the remaining grasp budget, systematically prioritizing actions that best disambiguate competing shape hypotheses. Crucially, we evaluate candidate grasps strictly over internal belief states, requiring neither physical rollouts before execution nor separately trained reinforcement learning policies. Ultimately, the proposed belief-guided strategy enables sequential active grasp selection, thereby achieving efficient tactile-based 3D shape reconstruction without training an extra active selection policy.

Extensive experiments on the challenging ABC and ShapeNet datasets demonstrate the superiority of the proposed method under sparse tactile observations. On the ABC dataset, the proposed method achieves a Mean CD of 2.399 with five random grasp selections and 2.093 with five active grasp selections. Moreover, comprehensive ablations further show that preserving hand–object interaction cues is particularly valuable when direct finger-contact evidence is scarce.

In summary, the principal contributions of this work can be summarized as follows:

- We introduce Touch-HOI, a unified framework that couples tactile 3D reconstruction and active exploration through a shared conditional generative model.
- We establish a unified formulation of hand–object interactions for tactile 3D shape reconstruction, introducing Hand–Object Interaction Evidence and Encoding to capture geometric relationships within grasps and evidence accumulation across interactions.
- We propose a belief-guided grasp selection strategy to systematically disambiguate multiple shape hypotheses, enabling efficient tactile 3D reconstruction within a limited grasp budget.

2 Related Work

Optimization-based tactile 3D reconstruction. Optimization-based methods fit object-specific representations directly to observations at inference time. For instance, Gaussian-process implicit surfaces (Dragiev et al., 2011) model probabilistic geometries from sparse contact measurements, whereas TouchSDF (Comi et al., 2024) optimizes a DeepSDF latent vector against local tactile geometry. Similarly, NeuralFeels (Suresh et al., 2024) learns a neural implicit field online while jointly optimizing object pose from visuo-tactile and proprioceptive cues, and Touch-GS (Swann et al., 2024) fits 3D Gaussian splats under visual and tactile supervision. TouchAnything (Gu et al., 2026) reconstructs test instances through test-time optimization, enforcing sparse contact constraints alongside pretrained 2D visual diffusion priors and semantic guidance. Consequently, despite incorporating learned priors, these paradigms still depend on instance-level test-time optimization. Although these methods effectively exploit measured local geometry, they leave under-observed regions ambiguous when observations are sparse.

Feed-forward tactile 3D reconstruction. Feed-forward models learn a direct mapping from tactile inputs to 3D geometry, enabling efficient reconstruction without per-object optimization at test time. Early deterministic frameworks employ chart-based networks (Smith et al., 2020; 2021) to deform surface meshes or learn implicit neural fields like VTacO (Xu et al., 2023) from visuotactile observations. To characterize geometric uncertainty under sparse observations, Touch2Shape (Wang et al., 2025) employs a multi-stage latent diffusion model, whereas TouchDiffusion (Zhang et al., 2026) generates 3D point clouds conditioned on registered tactile geometry. However, existing methods overlook the crucial geometric constraints embedded in tactile-generating hand–object interactions and lack a systematic way to incorporate them into tactile reconstruction. To address this, we fully exploit these interaction-level geometric relationships within tactile 3D reconstruction in this paper.

Active exploration for tactile reconstruction. Active sensing selects informative observations to maximize reconstruction accuracy under limited grasps. Early methods utilize Gaussian process implicit surfaces (Dragiev et al., 2011) to guide tactile exploration. Subsequent haptic frameworks target unobserved or uncertain regions, including Act-VH (Rustler et al., 2022) for visuo-haptic completion, VISHAC (Rustler et al., 2023) with freespace guidance, coverage-based planning (Kim et al., 2026) across multi-modal interactions, and Next Best Sense (Strong et al., 2025) for information-driven selection. Another research line learns exploration policies via reinforcement learning: ActiveVT (Smith et al., 2021) maximizes reconstruction gains, whereas Touch2Shape (Wang et al., 2025) employs latent diffusion denoising rewards. However, coverage-based strategies focus only on unexplored areas of a single shape estimate, whereas existing RL-based methods rely on reconstruction or denoising rewards without explicitly modeling how upcoming tactile observations will eliminate ambiguity across plausible shape hypotheses.

3 Methodology

3.1 Problem Formulation and Overview

We consider reconstructing an unknown object shape $\mathcal{S}$ from tactile feedback and hand proprioception, without visual input. After K grasps, the accumulated evidence $\mathcal{O}_K$ is registered in a common, normalized object frame. Given a budget $K_{\max}$, the robot reconstructs the object and,

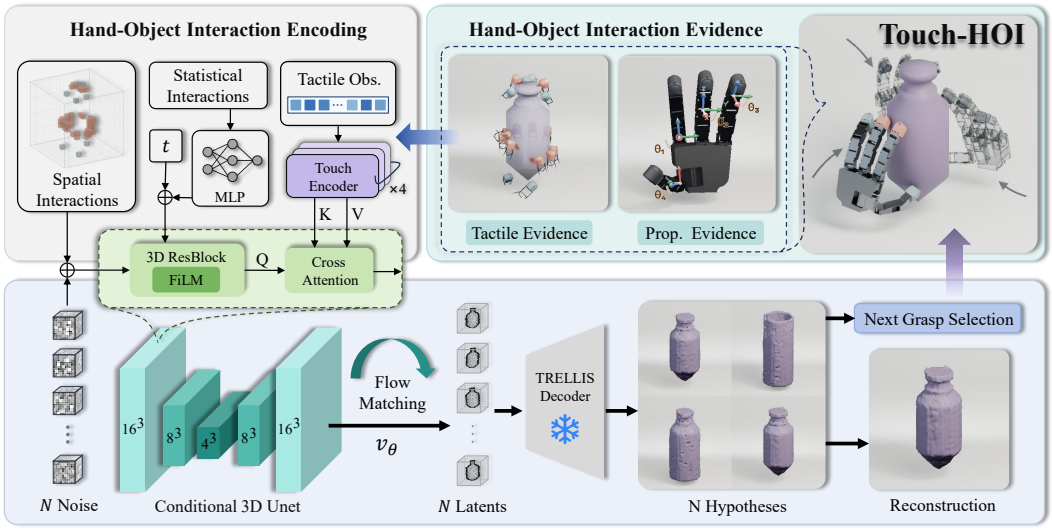

Figure 2: **Overview of Touch-HOI.** Interaction evidence conditions a latent flow model through tactile sensing and proprioceptive cues. The frozen TRELLIS Xiang et al. (2024) sparse-structure decoder and geometry extraction convert generated latents into shape hypotheses. Separate samples support reconstruction and belief-guided grasp selection. The selected grasp supplies new evidence for the next iteration.

when $K < K_{\max}$, selects the next grasp from the available candidate grasps $\mathcal{A}_K$. To accommodate ambiguity under sparse observations, we learn a conditional shape distribution $p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathcal{S} \mid \mathcal{O}_K)$, where $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ parameterize the generator and trainable interaction encoders.

As shown in Figure 2, we formulate **Hand–Object Interaction Evidence** from hand-object interaction and proprioceptive cues. Our **Hand–Object Interaction Encoding** integrates tactile sensing, statistical interactions, and spatial interactions to condition a latent flow-matching model. The generated shape hypotheses support reconstruction and belief-guided grasp selection, which evaluates prospective tactile outcomes to reduce geometric ambiguity.

3.2 Hand–Object Interaction Evidence

A grasp reveals geometry through the relationship between the object and the entire articulated hand. Contacting fingers observe local surfaces, non-contacting fingers provide negative observations at tested configurations, and proprioception situates both within the posed-hand geometry. We retain these complementary cues as **Hand–Object Interaction Evidence**:

$$\mathcal{O}_K = \left\{ \left(\{\mathcal{C}_{k,f}\}_{f=1}^{F}, \boldsymbol{h}_k \right) \right\}_{k=1}^{K}, \qquad \mathcal{C}_{k,f} = \{(\boldsymbol{p}_{k,f,j}, s_{k,f,j})\}_{j=1}^{P}, \tag{1}$$

where F represents the number of fingers, and each finger observation $\mathcal{C}_{k,f}$ contains P position–status pairs, following the tactile-chart representation (Smith et al., 2021; Wang et al., 2025). Positions $\boldsymbol{p}_{k,f,j} \in \mathbb{R}^3$ use the shared object frame; $s_{k,f,j}$ distinguishes contact, valid non-contact, and invalid samples. The configuration $\boldsymbol{h}_k$ records the hand base pose and joint state.

Tactile Evidence. We retain both contact and non-contact observations, as the absence of contact at a tested fingertip configuration also provides localized geometric evidence. A contacting finger provides a local surface chart, whereas a valid non-contact observation records the fingertip position with a distinct status, repeated to match the chart size P. Their shared-frame positions preserve spatial relationships across fingers and grasps. Invalid or missing observations are masked separately and are not treated as non-contact evidence.

Proprioceptive Evidence. The hand configuration $\boldsymbol{h}_k$ determines the poses of the palm and finger links, providing geometric context beyond fingertip observations. We approximate finger links by capsules and the palm by an oriented box, and represent the posed hand using signed-distance and soft occupancy fields. Two signed-distance channels retain truncated distances to the fingers and

palm, respectively, taking the minimum across grasps. Two occupancy channels aggregate whole-hand occupancy by the maximum and mean, capturing its spatial extent and repeated occupancy. To accommodate approximation errors, we attenuate maximum-occupancy values near observed contacts. These fields encode geometric exclusion information under the non-penetration assumption and provide spatial context for shape generation.

3.3 Hand–Object Interaction Encoding

The evidence in Eq. 1 varies in local geometry, contact frequency, and spatial distribution. Our **Hand–Object Interaction Encoding** represents these properties through finger-observation features, accumulated interaction statistics, and spatial fields.

Tactile Encoding. A shared PointNet-style encoder (Qi et al., 2017) maps each $\mathcal{C}_{k,f}$ to a feature describing its sampled positions and statuses. We augment this feature with a positional embedding of its mean valid position. A Transformer models relationships among observations to produce local tokens $\boldsymbol{L}$, masking invalid inputs. A frozen tactile encoder pretrained through tactile–shape alignment (Wang et al., 2025) additionally supplies a global feature $\boldsymbol{q}_{\mathrm{G}}$.

Statistical Interaction Encoding. Equal grasp counts can yield different amounts of surface evidence: some grasps establish several contacts, whereas others produce mostly non-contact observations. We summarize the accumulated grasp count, number of valid finger observations, and contact incidence at both grasp and finger levels in a vector $\boldsymbol{s}_K$. These statistics distinguish how many interactions were performed from how much contact they produced. An MLP maps $\boldsymbol{s}_K$ to $\boldsymbol{c}_{\mathrm{stat}}$, providing an explicit summary of evidence accumulation alongside the observation features.

Spatial Interaction Encoding. Contact statistics alone cannot distinguish repeated observations of the same region from observations distributed across the object. We therefore rasterize contact and valid non-contact observations into separate channels on a shared 3D grid, retaining sub-voxel contact offsets to preserve their spatial distribution. We align these grid-based observation channels with the proprioception-derived signed-distance and occupancy fields defined above. Separate 3D encoders extract features from the observation channels and hand geometry fields, which are then fused at multiple spatial resolutions. The fused features capture the spatial relationship between tactile observations and accumulated hand geometry. Combined with the local tokens, global tactile feature, and interaction statistics, they constitute the Hand–Object Interaction Encoding.

3.4 Interaction-Conditioned Shape Generation

Sparse interaction evidence may admit multiple different shapes. We formulate these alternatives via conditional flow matching in the latent space of the frozen TRELLIS sparse-structure autoencoder (Xiang et al., 2024).

Conditional Latent Flow Matching. As shown in Figure 2, a 3D U-Net parameterizes the velocity field $v_{\boldsymbol{\theta}}$ in the pretrained TRELLIS (Xiang et al., 2024) latent space, conditioned on $\mathcal{E}_{\boldsymbol{\psi}}(\mathcal{O}_K)$. Global tactile features and interaction statistics, together with the time embedding t, modulate residual blocks through FiLM. Global and local features provide cross-attention context, while spatial features $\boldsymbol{G}_r$ are added at matching encoder resolutions and the bottleneck.

Training and Inference. During training, we encode each ground-truth 3D shape $\mathcal{S}$ using the frozen TRELLIS (Xiang et al., 2024) encoder and standardize the resulting latent to obtain $\widetilde{\boldsymbol{z}}$, the target endpoint for flow matching. We independently sample $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $t \sim \mathcal{U}[0,1]$, and construct $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\widetilde{\boldsymbol{z}}$. The velocity network and trainable interaction encoders are jointly optimized to predict the target velocity $\widetilde{\boldsymbol{z}} - \boldsymbol{x}_0$ through flow matching (Lipman et al., 2023):

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{(\mathcal{S},\mathcal{O}_K),\boldsymbol{x}_0,t}\left[\left\|v_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \mathcal{E}_{\boldsymbol{\psi}}(\mathcal{O}_K)) - (\widetilde{\boldsymbol{z}} - \boldsymbol{x}_0)\right\|_2^2\right]. \tag{2}$$

Training uses varying grasp budgets, including empty observations, and joint condition dropout for classifier-free guidance (Ho & Salimans, 2022). The pretrained TRELLIS autoencoder (Xiang et al., 2024) and global tactile encoder remain frozen.

At inference, we sample independent Gaussian noise vectors and integrate the learned velocity field over multiple steps across $t \in [0,1]$, optionally applying classifier-free guidance. The resulting latents are inverse-standardized and decoded by the frozen TRELLIS decoder (Xiang et al., 2024) to yield a set of 3D shape hypotheses.

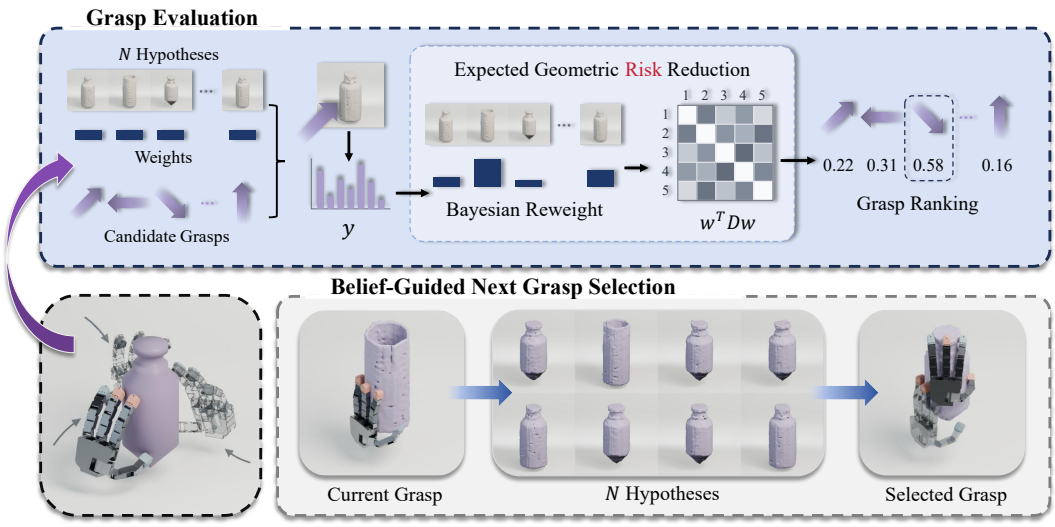

Figure 3: **Generative active exploration. Upper:** Predicted joint finger outcomes induce Bayesian reweighting of shape hypotheses, allowing candidate grasps to be ranked by expected geometric risk reduction. **Lower:** Shape hypotheses from the current interaction evidence guide the next grasp.

3.5 Belief-Guided Grasp Selection

Given a set of tactile evidence, our model can generate a set of shape hypotheses representing competing 3D geometries of the target object. Intuitively, a good candidate grasp is the one whose anticipated tactile feedback helps distinguish among competing shape hypotheses. We therefore propose a belief-guided grasp selection strategy that predicts tactile outcome hypotheses and ranks candidate grasps by their expected reduction in geometric ambiguity.

Geometric Risk-based Grasp Evaluation. Given observations $\mathcal{O}_K$ from K grasps, we generate N shape hypotheses $\{\widehat{\mathcal{S}}_i\}_{i=1}^N$ with uniform initial weights $w_i = 1/N$. For each candidate grasp $a \in \mathcal{A}_K$, a geometric contact model predicts the distribution $L_{ia}(\boldsymbol{y})$ over tactile outcome hypotheses $\boldsymbol{y}$ under shape hypothesis $\widehat{\mathcal{S}}_i$. Here, $\mathcal{A}_K$ denotes the available candidate grasps, and $\boldsymbol{y}$ specifies each finger's contact stage or absence of contact along a nominal closing trajectory. As illustrated in Figure 3, we compute the predictive tactile feedback y and apply Bayesian reweighting to the shape hypotheses:

$$P_a(\boldsymbol{y} \mid \boldsymbol{w}) = \sum_{i=1}^{N} w_i L_{ia}(\boldsymbol{y}), \qquad w_i^{a,\boldsymbol{y}} = \frac{w_i L_{ia}(\boldsymbol{y})}{P_a(\boldsymbol{y} \mid \boldsymbol{w})}, \tag{3}$$

where $\boldsymbol{w}$ collects the current shape-hypothesis weights, $P_a(\boldsymbol{y} \mid \boldsymbol{w})$ marginalizes the outcome likelihood over shape hypotheses, and $w_i^{a,\boldsymbol{y}}$ is the hypothetical posterior weight of hypothesis i if $\boldsymbol{y}$ were observed. Bayesian reweighting favors shapes consistent with the hypothesized tactile feedback, estimating how each possible outcome would distinguish existing shape hypotheses without regenerating them. We then quantify the remaining geometric disagreement under each hypothetical posterior and average over possible tactile outcomes:

$$\mathcal{U}(\boldsymbol{w}) = \sum_{i,j=1}^{N} w_i w_j D_{ij}, \qquad \overline{\mathcal{U}}_1(a, \boldsymbol{w}) = \sum_{\boldsymbol{y} \in \mathcal{Y}_a} P_a(\boldsymbol{y} \mid \boldsymbol{w}) \mathcal{U}(\boldsymbol{w}^{a,\boldsymbol{y}}), \tag{4}$$

where D_{ij} measures occupancy and truncated-distance disagreement near the predicted surfaces of shape hypotheses i and j, $\mathcal{Y}_a$ contains the tactile outcome hypotheses with positive predictive probability for grasp a, and $\boldsymbol{w}^{a,\boldsymbol{y}}$ collects the corresponding posterior shape weights. Here, $\mathcal{U}$ serves as a geometric risk proxy, while $\overline{\mathcal{U}}_1$ estimates the risk expected to remain after grasp a. We score each candidate grasp by its normalized risk reduction, $\rho(a, \boldsymbol{w}) = \max\{\mathcal{U}(\boldsymbol{w}) - \overline{\mathcal{U}}_1(a, \boldsymbol{w}), 0\} / \max\{\mathcal{U}(\boldsymbol{w}), \epsilon\}$, where $\epsilon > 0$ ensures numerical stability. This criterion prioritizes candidate grasps whose anticipated tactile feedback is expected to resolve greater geometric ambiguity among shape hypotheses.

Table 1: Tactile-only reconstruction on the ABC dataset after five grasps under random and active grasp selection.

Method	Mean CD ↓	Median CD ↓	EMD ↓	F@2mm ↑	F@5mm ↑
Random grasp selection					
ActiveVT (Smith et al., 2021)	6.422	1.862	78.17	15.43	46.00
Touch2Shape (Wang et al., 2025)	3.134	–	–	–	–
TouchHOI(Ours)	**2.399**	**0.677**	**17.73**	**25.10**	**73.31**
Active grasp selection					
ActiveVT (Smith et al., 2021)	5.928	1.631	78.40	15.88	47.31
Touch2Shape (Wang et al., 2025)	2.671	–	–	–	–
TouchHOI(Ours)	**2.093**	**0.517**	**16.56**	**25.54**	**74.32**

Grasp Selection and Update. Based on normalized scores $\rho(a, \boldsymbol{w})$, we select candidate grasps maximizing expected risk reduction. Candidate selection integrates multi-step evaluation and directional novelty to balance action quality with spatial exploration (details in Appendix B.5). Executing the chosen grasp yields new observations to update the grasp sequence. After executing the selected grasp, we generate new shape hypotheses conditioned on the updated evidence. Iterating this cycle couples active perception with shape reconstruction, eliminating the need for a separately trained exploration policy.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate tactile-only reconstruction on ABC (Koch et al., 2019) and ShapeNet (Chang et al., 2015). The ABC dataset contains diverse mechanical geometries, and the ShapeNet dataset covers six categories, including bottle, mug, bowl, camera, guitar, and jar. Tactile and proprioceptive data are collected following the Allegro-hand protocol of ActiveVT (Smith et al., 2021), with four finger charts per grasp and up to five grasps per object.

Implementation Details. We implement our method in PyTorch and jointly optimize the interaction encoders and velocity network with AdamW, while keeping the TRELLIS autoencoder (Xiang et al., 2024) frozen. Following Touch2Shape (Wang et al., 2025), we adopt classifier-free guidance and randomly sample zero to five grasps per object during training. We train the models for 500 epochs in all experiments. At inference, we generate eight hypotheses using an Euler solver over 50 steps with a guidance scale of 5 and select their geometric medoid based on pairwise Chamfer distance. For ABC evaluation, we sample 30,000 surface points per hypothesis for comparison. For active exploration, we generate 16 separate planning hypotheses with a guidance scale of 1, retaining the same reconstruction model.

Evaluation Metrics. Following previous works (Smith et al., 2021; Wang et al., 2025), we report symmetric squared Chamfer distance (CD) on ABC, scaled by 9,000 and computed from 30,000 surface points per shape. We additionally evaluate Earth Mover's Distance (EMD) using 4,096 sampled points and report it in millimeters. In addition, we report F-score@2mm and F-score@5mm, which measure the harmonic mean of precision and recall at the corresponding distance thresholds and are expressed as percentages. On ShapeNet, we report EMD following the reporting convention of the compared methods (Comi et al., 2024). For active exploration, we report the mean CD over $K = 1, \ldots, 5$, with the area under the curve (AUC) of mean CD as an additional metric, which further illustrates the variation across different grasp budgets.

4.2 Evaluation on Tactile Reconstruction Performance

Results on the ABC dataset. As shown in Table 1, the proposed method achieves the lowest mean CD (2.399) after five randomly selected grasps and optimal performance on both EMD and F-scores. Table 1 further reports that the proposed active exploration strategy reduces mean CD from 2.399 to 2.093, which verifies the effectiveness of the proposed active exploration strategy. Figure 4 presents a qualitative comparison between the proposed method and ActiveVT (Smith et al., 2021).

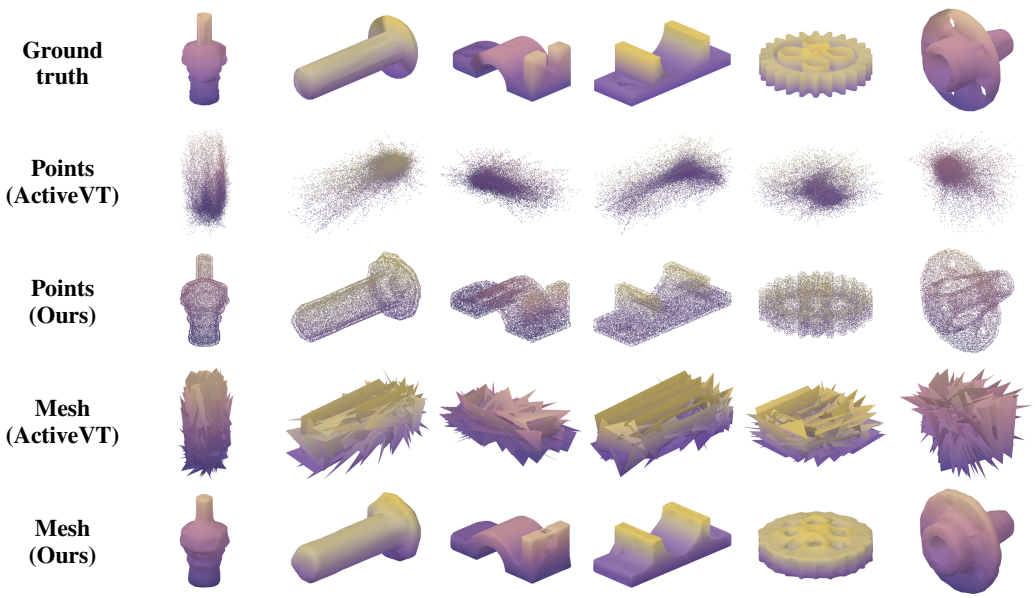

Figure 4: Qualitative tactile reconstruction comparison on ABC dataset. Rows show the ground-truth mesh, original predicted point clouds from ActiveVT and Touch-HOI, the original ActiveVT mesh, and the Touch-HOI mesh, respectively.

Table 2: Tactile-only reconstruction on ShapeNet measured by EMD. TouchSDF (Comi et al., 2024) and Touch2Shape (Wang et al., 2025) use fixed 10 or 20 point touches, whereas our method uses five four-finger grasps and its results are grouped by the actual number of finger contacts for fair comparison.

Method	Sensing budget	Bottle	Mug	Bowl	Camera	Guitar	Jar	Ave.
TouchSDF (Comi et al., 2024)	10 touches	0.082	0.072	0.073	0.101	0.177	0.136	0.112
Touch2Shape (Wang et al., 2025)	10 touches	0.042	0.050	0.051	**0.058**	0.067	0.059	0.056
TouchHOI (Ours)	1–10 contacts	**0.037**	**0.038**	**0.032**	0.061	**0.043**	**0.044**	**0.042**
TouchSDF (Comi et al., 2024)	20 touches	0.047	0.066	0.048	0.092	0.155	0.071	0.081
Touch2Shape (Wang et al., 2025)	20 touches	0.041	0.049	0.049	0.056	0.064	0.055	0.053
TouchHOI (Ours)	11–20 contacts	**0.027**	**0.036**	**0.027**	**0.040**	**0.031**	**0.033**	**0.030**

Results on the ShapeNet dataset. The compared methods use a single-contact robotic arm to acquire a fixed budget of 10 or 20 touches. To facilitate a fair comparison, we limit Touch-HOI to at most five grasps with a four-finger hand and group the results by the number of effective finger contacts: 1–10 and 11–20. These groups contain 7.5 and 13.1 contacts on average, respectively, both below the corresponding baseline budgets. Despite fewer effective contacts, Touch-HOI achieves the lowest average EMD in both groups, leading in five of six categories in the 1–10-contact group and all six categories in the 11–20-contact group (Table 2). These results support the effectiveness of explicitly modeling hand–object interactions, which incorporates non-contact observations and hand geometry alongside surface measurements to reconstruct shapes from limited tactile evidence.

4.3 Evaluation of Active Exploration Strategies

Table 3 presents a comparison of Random, Even, an RL-based policy (Wang et al., 2025), and our belief-guided active exploration strategy with the same initial grasp. As shown in Table 3, the proposed strategy achieves the lowest Mean CD at every subsequent grasp budget. These consistent gains demonstrate the effectiveness of our active exploration strategy. Importantly, different with existing methods (Smith et al., 2021; Wang et al., 2025), our active strategy requires no additional

Table 3: Comparison of exploration strategies across grasp budgets on ABC dataset.

Exploration Policy	Mean CD ↓					AUC ↓
	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	
Random	8.660	5.400	3.976	3.056	2.399	4.490
Even	8.660	5.480	3.993	2.922	2.520	4.496
RL (Wang et al., 2025)	8.660	4.671	3.227	2.803	2.605	4.083
Ours	8.660	**4.012**	**2.945**	**2.323**	**2.093**	**3.664**

Table 4: Ablation study of hand–object interaction evidence and encoding on ABC dataset.

Ablation Setting	Mean CD ↓	Mean CD by Finger-Contact Count ↓		
		1–4 ($n=81$)	5–8 ($n=230$)	9–20 ($n=628$)
Hand–Object Interaction Evidence:				
w/o Tactile observation	4.898 (+2.499)	9.508 (+6.258)	4.457 (+1.913)	4.329 (+2.149)
w/o non-contact finger	3.202 (+0.803)	6.246 (+2.996)	3.262 (+0.718)	2.767 (+0.586)
w/o Proprioception	2.619 (+0.219)	3.697 (+0.447)	3.138 (+0.593)	2.252 (+0.071)
Hand–Object Interaction Encoding:				
w/o Spatial Interaction Encoding	2.579 (+0.179)	3.870 (+0.620)	3.096 (+0.552)	2.193 (+0.012)
w/o Statistical Interaction Encoding	2.567 (+0.168)	3.648 (+0.398)	3.028 (+0.483)	2.202 (+0.021)
Touch-HOI (Full)	**2.399**	**3.250**	**2.545**	**2.181**

model training and uses the shape uncertainty among samples from the trained generative model to assess shape uncertainty and guide exploration.

4.4 Ablation Study

Table 4 ablates the contributions of **Hand–Object Interaction Evidence** and **Hand–Object Interaction Encoding**. As shown in Table 4, excluding tactile observations increases overall Mean CD by 2.499. Omitting non-contact evidence raises Mean CD by 2.996 in the 1–4 contact group, whereas the absence of proprioception incurs increases of 0.447 and 0.593 in the 1–4 and 5–8 contact groups, respectively. Moreover, without spatial interaction encoding, Mean CD increases by 0.620 in the 1–4 contact group but only 0.012 in the 9–20 contact group, and disabling statistical interaction encoding yields increases of 0.483 and 0.021 in the 5–8 and 9–20 contact groups, respectively. The pronounced degradation under sparse contact underscores the importance of supplementing information with non-contact evidence and hand proprioception, as well as preserving the spatial distribution and statistics of hand–object interactions.

5 Conclusion

In this paper, we present Touch-HOI for active tactile 3D shape reconstruction through explicit modeling of hand–object interactions. The key contributions of our work are threefold: (1) a unified generative framework for reconstruction and active exploration; (2) structured hand–object interaction modeling with spatial and statistical encoding; (3) belief-guided grasp selection based on expected geometric risk reduction. Comprehensive experiments on ABC and ShapeNet datasets demonstrate the effectiveness of Touch-HOI in tactile 3D reconstruction.

Limitations and Future Work. Although Touch-HOI achieves promising tactile 3D reconstruction performance, our current experiments are limited to simulated data, and exploration selects from a finite set of discrete grasps rather than planning continuous motion and touch. Our future work will focus on deploying the framework on dexterous robotic hands equipped with tactile sensors and enabling continuous tactile exploration.

AI USE STATEMENT

We used GPT to assist with drafting, translation, and language polishing of the manuscript, as well as brainstorming ideas for the proposed approach and interpreting selected experimental results. The authors critically evaluated the AI-assisted ideas and interpretations against relevant literature and experimental evidence, and reviewed and revised all AI-assisted text for accuracy and consistency with the work presented. All final methodological decisions and conclusions were made by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The potential ethical implications of this work depend on its application context. When applying the proposed method, practitioners should assess relevant risks, including safety, privacy, and fairness, and establish appropriate safeguards before deployment.

REPRODUCIBILITY STATEMENT

The proposed method is described in Section 3. The experimental settings and evaluation protocols are provided in Section 4, with additional implementation details in Appendix. These descriptions are intended to facilitate reproduction of the reported results. The code will be made publicly available upon acceptance.

REFERENCES

Mårten Björkman, Yasemin Bekiroglu, Virgile Högman, and Danica Kragic. Enhancing visual perception of shape through tactile glances. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3180–3186, 2013. doi: 10.1109/IROS.2013.6696808. URL https://research.chalmers.se/publication/518857.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996. URL https://arxiv.org/abs/cs/9603104.

Mauro Comi, Yijiong Lin, Alex Church, Alessio Tonioni, Laurence Aitchison, and Nathan F. Lepora. TouchSDF: A DeepSDF approach for 3d shape reconstruction using vision-based tactile sensing. *IEEE Robotics and Automation Letters*, 9(6):5719–5726, 2024. doi: 10.1109/LRA.2024.3396054.

Stanimir Dragiev, Marc Toussaint, and Michael Gienger. Gaussian process implicit surfaces for shape estimation and grasping. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2845–2850, 2011. doi: 10.1109/ICRA.2011.5980395.

Danny Driess, Peter Englert, and Marc Toussaint. Active learning with query paths for tactile object shape exploration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017. doi: 10.1109/IROS.2017.8202139. URL https://dannydriess.github.io/papers/17-driess-IROS.pdf.

Danny Driess, Daniel Hennes, and Marc Toussaint. Active multi-contact continuous tactile exploration with Gaussian process differential entropy. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019. URL https://dannydriess.github.io/papers/19-driess-ICRA.pdf.

Langzhe Gu, Hung-Jui Huang, Mohamad Qadri, Michael Kaess, and Wenzhen Yuan. TouchAnything: Diffusion-guided 3D reconstruction from sparse robot touches. arXiv preprint arXiv:2604.08945, 2026. URL https://arxiv.org/abs/2604.08945.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

Chung Hee Kim, Shivani Kamtikar, Tye Brady, Taskin Padir, and Joshua Migdal. Grasp, slide, roll: Comparative analysis of contact modes for tactile-based shape reconstruction. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. ABC: A big CAD model dataset for geometric deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9601–9611, 2019.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21(4):163–169, 1987. doi: 10.1145/37401.37422. URL https://www.cs.toronto.edu/~jacobson/seminar/lorenson-and-cline-1987.pdf.

Jens Lundell, Francesco Verdoja, and Ville Kyrki. Robust grasp planning over uncertain shape completions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1526–1532, 2019. doi: 10.1109/IROS40897.2019.8967816. URL https://arxiv.org/abs/1903.00645.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4460–4470, 2019.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 165–174, 2019.

Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision – ECCV 2020*, pp. 523–540, 2020. doi: 10.1007/978-3-030-58580-8_31. URL https://is.mpg.de/avg/publications/peng2020eccv.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 652–660, 2017. doi: 10.1109/CVPR.2017.16.

Lukas Rustler, Jens Lundell, Jan Kristof Behrens, Ville Kyrki, and Matej Hoffmann. Active visuo-haptic object shape completion. *IEEE Robotics and Automation Letters*, 7(2):5254–5261, 2022. doi: 10.1109/LRA.2022.3152975.

Lukas Rustler, Jiri Matas, and Matej Hoffmann. Efficient visuo-haptic object shape completion for robot manipulation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3121–3128, 2023. doi: 10.1109/IROS55552.2023.10342200.

Brad Saund and Dmitry Berenson. CLASP: Constrained latent shape projection for refining object shape from robot contact. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1391–1400. PMLR, 2022. URL https://proceedings.mlr.press/v164/saund22a.html.

Edward J. Smith, Roberto Calandra, Adriana Romero, Georgia Gkioxari, David Meger, Jitendra Malik, and Michal Drozdzal. 3D shape reconstruction from vision and touch. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14193–14206, 2020.

Edward J. Smith, David Meger, Luis Pineda, Roberto Calandra, Jitendra Malik, Adriana Romero-Soriano, and Michal Drozdzal. Active 3d shape reconstruction from vision and touch. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Matthew Strong, Boshu Lei, Aiden Swann, Wen Jiang, Kostas Daniilidis, and Monroe Kennedy III. Next best sense: Guiding vision and touch with FisherRF for 3d gaussian splatting. In *IEEE International Conference on Robotics and Automation*, 2025.

Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, Joseph Ortiz, and Mustafa Mukadam. Neural feels with neural fields: Visuo-tactile perception for in-hand manipulation. *Science Robotics*, 9(96), 2024.

Aiden Swann, Matthew Strong, Won Kyung Do, Gadiel Sznaier Camps, Mac Schwager, and Monroe Kennedy III. Touch-GS: Visual-tactile supervised 3D gaussian splatting. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10511–10518, 2024. doi: 10.1109/IROS58592.2024.10802412.

Yuanbo Wang, Zhaoxuan Zhang, Jiajin Qiu, Dilong Sun, Zhengyu Meng, Xiaopeng Wei, and Xin Yang. Touch2Shape: Touch-conditioned 3d diffusion for shape exploration and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5656–5665, 2025.

David Watkins-Valls, Jacob Varley, and Peter Allen. Multi-modal geometric learning for grasping and manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7339–7345, 2019. doi: 10.1109/ICRA.2019.8794233. URL https://www.cs.columbia.edu/~allen/PAPERS/multimodal_icra_2019.pdf.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480, 2025. doi: 10.1109/CVPR52734.2025.02000.

Wenqiang Xu, Zhenjun Yu, Han Xue, Ruolin Ye, Siqiong Yao, and Cewu Lu. Visual-tactile sensing for in-hand object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8803–8812, 2023.

Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3DShape2VecSet: A 3D shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics*, 42(4):1–16, 2023. doi: 10.1145/3592442. URL https://1zb.github.io/3DShape2VecSet/.

Han Zhang, Xiaohui Zhang, Jun Huang, Zhao Feng, and Xiaohui Xiao. End-to-end diffusion-based 3d object reconstruction from robotic tactile sensing. *IEEE Robotics and Automation Letters*, 11 (2):1434–1441, 2026. doi: 10.1109/LRA.2025.3641115.

Sheng Zhong, Nima Fazeli, and Dmitry Berenson. RUMI: Rummaging using mutual information. *IEEE Transactions on Robotics*, 41:5431–5450, 2025. URL https://arxiv.org/abs/2408.10450.

A Additional Results

A.1 Proprioception Representation and Spatial Injection

Table 5 evaluates the spatial representation described in Appendix B.3. Combining signed distance field and occupancy channels yields the lowest overall Mean CD, although occupancy-only conditioning performs slightly better in the 5–8-contact group. Under sparse contact, occupancy-only conditioning increases Mean CD from 3.250 to 3.995. Restricting spatial injection to the input scale (16^3) raises overall Mean CD from 2.399 to 2.620, with the largest degradation in the 1–4-contact group (3.250 to 4.211). These results support combining complementary proprioceptive cues with multiscale spatial injection, particularly when tactile information is limited.

Table 5: Ablation study of proprioception representation and multiscale interaction injection on ABC dataset.

Ablation Setting	Mean CD ↓	Mean CD by Finger-Contact Count ↓		
		1–4 ($n = 81$)	5–8 ($n = 230$)	9–20 ($n = 628$)
Proprioception Representation:				
SDF-only (finger + palm)	2.530 (+0.131)	3.448 (+0.198)	2.753 (+0.209)	2.283 (+0.102)
Occupancy-only (maximum + mean)	2.472 (+0.073)	3.995 (+0.745)	**2.518** (−0.026)	2.238 (+0.058)
Spatial Interaction Injection:				
Input-scale addition (16^3 only)	2.620 (+0.221)	4.211 (+0.961)	2.961 (+0.416)	2.223 (+0.042)
Touch-HOI (Full)	**2.399**	**3.250**	2.545	**2.181**

A.2 Effect of the 3D Latent Representation

3D shape Autoencoding. Before evaluating reconstruction from sparse tactile observations, we examine how accurately each latent representation preserves complete object geometry. Table 6 compares the encode–decode reconstruction of ground-truth ABC shapes using our ABC-trained SS-VAE and the pretrained TRELLIS autoencoder (Xiang et al., 2025). Providing complete ground-truth shapes to the encoders measures autoencoding fidelity without requiring shape inference from sparse tactile observations. Both representations are evaluated on the same 947 objects using identical normalization, isosurface extraction, and surface-point sampling protocols. No tactile conditioning is involved in this comparison.

TRELLIS reduces Mean CD by 0.014 and yields lower reconstruction error on 858 of 947 objects, although our task-specific SS-VAE also preserves complete geometry with low error. The higher effective rank of TRELLIS indicates more distributed latent-channel variation on this evaluation set; it does not by itself establish greater capacity or generalization. Both autoencoding errors are more than an order of magnitude below their corresponding tactile-conditioned errors, suggesting that inferring geometry from sparse observations remains a substantially greater challenge than encoding and decoding complete shapes.

Tactile 3D Reconstruction. We also train a conditional flow model using our task-specific SS-VAE. Table 7 compares this system with Touch-HOI under the same evaluation protocol. The systems differ in their spatial conditioning architectures, so this comparison assesses overall reconstruction performance and does not isolate the effect of the autoencoder.

Our task-specific system achieves approximately 4.0% lower Mean CD (2.304 versus 2.399) and lower Median CD (0.626 versus 0.686). Touch-HOI achieves a 2.13% higher F@2mm and a slightly lower P90 CD (6.106 versus 6.128). These mixed results demonstrate that our independently trained SS-VAE supports competitive tactile-conditioned reconstruction: strong performance is attainable without the pretrained TRELLIS representation. TRELLIS nevertheless offers comparable overall accuracy without ABC-specific autoencoder training. Thus, the results do not support attributing reconstruction quality solely to TRELLIS pretraining, although architectural differences preclude isolating its contribution.

Table 6: Comparison of latent representations for 3D shape autoencoding. Latent rank is the effective rank of the latent-channel covariance.

3D latent representation	Latent rank	Mean CD ↓	Median CD ↓	Decode wins ↑
Task-specific SS-VAE	5.650	0.174	0.174	89
Pretrained TRELLIS (Xiang et al., 2024)	**7.778**	**0.160**	**0.159**	**858**

Table 7: Tactile 3D Reconstruction using task-specific SS-VAE and pretrained TRELLIS representations on ABC dataset.

3D latent representation	Mean CD ↓	Median CD ↓	F@2mm (%) ↑
Task-specific SS-VAE	**2.304**	**0.626**	22.98
Pretrained TRELLIS (Xiang et al., 2024)	2.399	0.686	**25.11**

A.3 More Analysis on Active Reconstruction

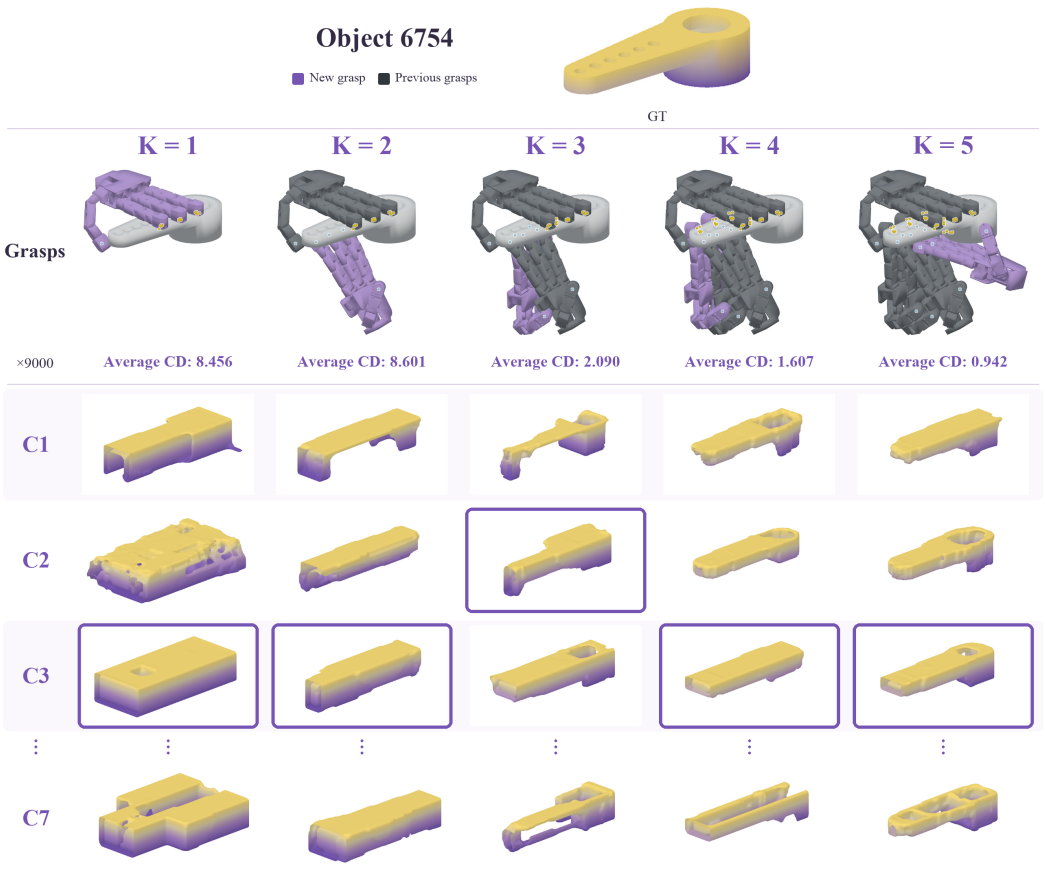

Figure 5: Active tactile 3D reconstruction of ABC dataset on object 6754 over five actively selected grasps.

Table 8 reports each method's reconstruction trajectory relative to its own zero-grasp error. For Mean CD e_K after K grasps, the percentage is defined as:

$$r_K = 100\, e_K / e_0. \tag{5}$$

Table 8: More analysis on active reconstruction.

Method	$K = 0$	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
Touch2Shape (Wang et al., 2025)	100	16.90	9.62	7.62	6.88	6.63
Touch-HOI (Ours)	100	**14.84**	**6.88**	**5.05**	**3.98**	**3.59**

The zero-grasp Mean CD values are 40.283 for Touch2Shape (Wang et al., 2025) and 58.336 for Touch-HOI. Touch2Shape follows its reported sensing protocol (Wang et al., 2025).

As shown in Table 8, the proposed Touch-HOI consistently achieves lower relative CD than Touch2Shape at every subsequent step ($K = 1$–5). After five grasps, Touch-HOI reduces CD to 3.59% of its initial value, compared with 6.63% for Touch2Shape. These results demonstrate more effective relative error reduction across the evaluated interaction budgets.

Active exploration progressively constrains candidate geometries as interaction evidence accumulates. Figures 5 and 6 illustrate this process. For object 6754, subsequent grasps recover the elongated body and circular end feature, reducing the average candidate CD from 8.456 to 0.942. For object 19731, initially ambiguous box-like candidates resolve into the notched geometry after the fourth and fifth grasps, reducing the average candidate CD from 17.039 to 0.525.

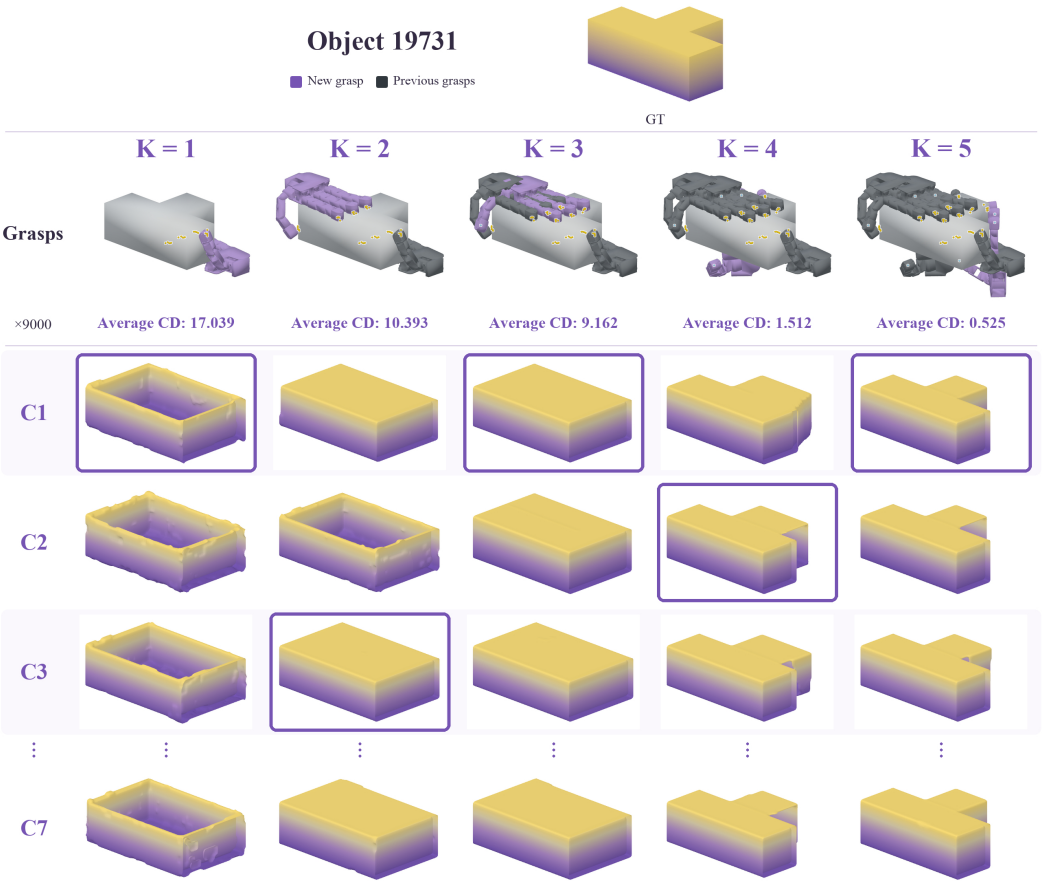

Figure 6: Active tactile 3D reconstruction of ABC dataset on object 19731 over five actively selected grasps.



Figure 7: Selected failure cases from the five-grasp ABC evaluation. Touch-HOI reconstructions are shown in blue, ground-truth shapes in gray, and observed contacts in orange. Each example reports CD scaled by 9,000 and the cumulative number of contacting fingers. These cases were selected post hoc from the range $5 < \text{CD} \leq 10$ to illustrate characteristic errors and do not represent average test-set performance.

A.4 Failure Cases

Figure 7 presents ABC failure cases under the same five-grasp reconstruction protocol as the main evaluation. The displayed cases have CD values between 5.463 and 9.914 and include examples with zero to 15 contacting fingers, showing that failure is not determined by contact count alone.

With no contact or only a few contacting fingers, the observations constrain little of the object surface, and the returned shape can follow a plausible but incorrect global topology. This behavior is visible when a thin ring or a branched part is reconstructed as a much thicker, simpler shape. Failures also occur with more contacts when those contacts cover only a limited region. These cases expose two remaining limitations. Sparse interaction evidence may not distinguish globally different shapes, and geometric medoid selection can favor a shared but incorrect completion when it dominates the finite hypothesis set. Broader spatial coverage and more diverse hypotheses are therefore likely to be most useful for objects with thin parts or complex topology.

A.5 Qualitative Results on ShapeNet

The ShapeNet dataset uses ShapeNetCoreV2 (Chang et al., 2015) models from six categories: bottle, bowl, camera, guitar, jar, and mug. After excluding objects for which preprocessing failed or no simulated finger contact was available, we randomly select 1,650 objects in proportion to the category frequencies and divide them into 1,100 training, 200 validation, and 350 test objects with seed 131. ShapeNet and ABC use the same simulated Allegro-hand interaction protocol and coordinate conventions. Each object has up to 50 recorded grasp candidates, and every grasp provides four 25-sample tactile observations and the corresponding hand configuration. At most five grasps are used for reconstruction.

As shown in Figure 8, Touch-HOI recovers the overall 3D geometry and recognizable category-specific structures of the objects. These examples illustrate that the same Touch-HOI framework

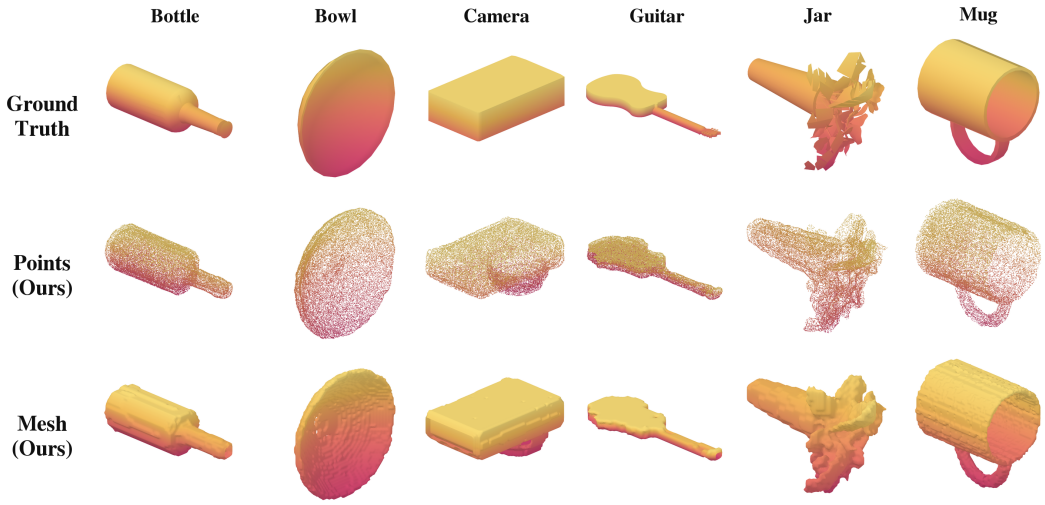

Figure 8: Qualitative ShapeNet results. Columns correspond to the six evaluated categories, and rows show the ground-truth shape, reconstructed surface points, and the corresponding mesh visualization.

can be trained and applied to the category-level objects in ShapeNet as well as the mechanical geometries in ABC, and show excellent reconstruction performance at the same time.

B Implementation Details

This section expands the tactile 3D reconstruction and active exploration pathways in Figures 2 and 3. We specify the evidence representation, interaction encoders, conditional generator, and budget-aware action selection used by Touch-HOI.

B.1 Observation Construction and Coordinate Conventions

Tactile measurements have previously been used to refine probabilistic shape estimates initialized from vision (Björkman et al., 2013) and to condition learned depth–tactile shape completion (Watkins-Valls et al., 2019). Our observation construction instead follows the tactile-only benchmark protocol described below.

Our ABC interaction data follow the simulated Allegro-hand protocol of ActiveVT (Smith et al., 2021). Each object has up to 50 recorded grasp candidates. For every grasp, the dataset stores four tactile observations and the corresponding hand configuration; we represent these records with $\mathcal{O}_K$ and $\mathcal{C}_{k,f}$ as defined in Eq. 1. Each tactile chart contains $P = 25$ point–status pairs. Status values 0, 1, and 2 denote invalid, valid non-contact, and contact samples, respectively. Contact samples describe the sensed local surface. For a valid non-contact observation, the fingertip position is repeated across the chart to maintain a fixed input size; these repeated entries are not independent measurements. Invalid and padded samples are masked during chart encoding and attention. With at most $K_{\max} = 5$ executed grasps, the encoder receives at most $M = 4K_{\max} = 20$ tactile observations.

The positions recorded in tactile observations, posed-hand geometry, and candidate grasp trajectories share the same normalized object frame. The TRELLIS latent is defined on the corresponding normalized lattice. For a metric point $\boldsymbol{p}$, preprocessing maps it to normalized coordinates $s_o(\boldsymbol{p} - \boldsymbol{c}_o)$ using the object center $\boldsymbol{c}_o$ and scale s_o. These normalized coordinates are used for spatial encoding and grasp planning. The corresponding TRELLIS latent lattice spans $[-0.5, 0.5]^3$, and the inverse transformation is applied before metric evaluation. The object frame and scale are supplied rather than estimated from touch; consequently, the experiments evaluate reconstruction under known spatial alignment and do not measure pose or scale recovery. This setting differs from contact-based active pose estimation, such as RUMI, which plans informative interactions to localize a known movable object (Zhong et al., 2025).

B.2 Statistical Interaction Encoding

The statistical interaction encoder summarizes how much interaction evidence has accumulated. It counts executed grasps, valid tactile observations, grasps containing at least one contact, and valid tactile observations containing contact. Each count is provided in both a linearly normalized form and a normalized $\log(1+\text{count})$ form; the count ranges are five grasps and twenty chart slots. Three additional ratios describe the fraction of valid chart slots among the four slots per executed grasp, the fraction of executed grasps that contain contact, and the fraction of valid charts that contain contact. A final binary indicator records whether no valid tactile chart contains contact. These twelve statistics are mapped to the statistical embedding $\boldsymbol{c}_{\mathrm{stat}}$ by an MLP with two GELU-activated hidden layers and a 768-dimensional output, followed by layer normalization.

B.3 Spatial Interaction Encoding

Tactile spatial representation. We rasterize contact and valid non-contact samples on the 16^3 latent lattice. If a sample lies outside the normalized cube, we clamp its voxel index to 0 or 15 along each out-of-range axis, so it is stored in the corresponding edge voxel rather than discarded. For voxel $\boldsymbol{v}$, $o_{\mathrm{c}}(\boldsymbol{v})$ indicates whether it contains a contact sample, $n_{\mathrm{nc}}(\boldsymbol{v})$ counts its valid non-contact entries, and $\boldsymbol{\delta}_{\mathrm{c}}(\boldsymbol{v})$ is the mean contact displacement from the voxel center normalized by the voxel width. We set the offset to zero in voxels without contact. With normalized voxel center $\boldsymbol{r}_{\boldsymbol{v}}$, the eight-channel tactile representation can be described as:

$$\boldsymbol{G}_{\mathrm{T}}(\boldsymbol{v}) = \left[o_{\mathrm{c}}(\boldsymbol{v}),\ \boldsymbol{\delta}_{\mathrm{c}}(\boldsymbol{v})^{\mathsf{T}},\ \eta_{\mathrm{nc}} \log(1+n_{\mathrm{nc}}(\boldsymbol{v})),\ \boldsymbol{r}_{\boldsymbol{v}}^{\mathsf{T}}\right]^{\mathsf{T}}, \tag{6}$$

where $\eta_{\mathrm{nc}} = 0.1$ scales the non-contact channel. This channel distinguishes observed free space from voxels with no observation, while the offset retains contact locations below the voxel resolution. Because a valid non-contact chart repeats one fingertip position P times, n_{nc} counts stored chart entries rather than independent measurements.

Posed-hand representation. Let $\mathcal{K}_{\mathrm{H}}$ contain the executed grasps with valid hand configurations. For each $k \in \mathcal{K}_{\mathrm{H}}$, we approximate the posed finger links by capsules and the palm by an oriented box. Their signed-distance fields are $d_{\mathrm{f}}^{(k)}$ and $d_{\mathrm{p}}^{(k)}$, respectively, with negative values inside the represented geometry. Their minimum, $d_{\mathrm{h}}^{(k)} = \min(d_{\mathrm{f}}^{(k)}, d_{\mathrm{p}}^{(k)})$, represents the union of the fingers and palm. To account for voxel size, we dilate the primitives by $\Delta_{\mathrm{H}} = 0.35\ell_{\mathrm{vox}}/2$, where ℓ_{vox} is the metric voxel width. We then clip the distances and convert the whole-hand distance into a continuous occupancy score:

$$\begin{aligned} \widetilde{d}_q^{(k)}(\boldsymbol{v}) &= \mathrm{clip}\Big(d_q^{(k)}(\boldsymbol{v}) - \Delta_{\mathrm{H}}, -\rho, \rho\Big), \\ o_{\mathrm{h}}^{(k)}(\boldsymbol{v}) &= \sigma\Big(-\widetilde{d}_{\mathrm{h}}^{(k)}(\boldsymbol{v})/\tau_{\mathrm{H}}\Big), \end{aligned} \tag{7}$$

where $q \in \{\mathrm{f}, \mathrm{p}, \mathrm{h}\}$ indexes the finger, palm, and whole-hand fields, $\rho = 0.08\,\mathrm{m}$ is the distance truncation threshold, $\tau_{\mathrm{H}} = 0.008\,\mathrm{m}$ controls the occupancy transition, and σ is the logistic function.

Across executed grasps, we retain the minimum normalized distance to the fingers and palm, together with the maximum and mean whole-hand occupancy at each voxel:

$$\begin{aligned} D_q(\boldsymbol{v}) &= \min_{k \in \mathcal{K}_{\mathrm{H}}} \widetilde{d}_q^{(k)}(\boldsymbol{v})/\rho, \quad q \in \{\mathrm{f}, \mathrm{p}\}, \\ O_{\max}(\boldsymbol{v}) &= \max_{k \in \mathcal{K}_{\mathrm{H}}} o_{\mathrm{h}}^{(k)}(\boldsymbol{v}), \qquad O_{\mathrm{mean}}(\boldsymbol{v}) = \frac{1}{|\mathcal{K}_{\mathrm{H}}|} \sum_{k \in \mathcal{K}_{\mathrm{H}}} o_{\mathrm{h}}^{(k)}(\boldsymbol{v}). \end{aligned} \tag{8}$$

Approximate hand primitives can overlap a measured contact surface. We therefore set $O_{\max}$ to zero at contact voxels and within their one-voxel neighborhood, obtaining $\overline{O}_{\max}$. This contact relief modifies only the maximum-occupancy channel; the distance and mean-occupancy channels are retained. The resulting four-channel posed-hand representation is defined as:

$$\boldsymbol{G}_{\mathrm{H}}(\boldsymbol{v}) = \left[D_{\mathrm{f}}(\boldsymbol{v}), D_{\mathrm{p}}(\boldsymbol{v}), \overline{O}_{\max}(\boldsymbol{v}), O_{\mathrm{mean}}(\boldsymbol{v})\right]^{\mathsf{T}}. \tag{9}$$

The distance channels record proximity to any executed hand pose, whereas the occupancy channels describe the strongest and repeated hand presence across those poses. If no executed grasp has a

valid hand configuration, $\boldsymbol{G}_{\mathrm{H}}$ is set to zero. These channels provide spatial context to the learned generator rather than imposing a hard exclusion mask. Contact and free-space evidence also constrain sampled shape beliefs in CLASP (Saund & Berenson, 2022). Here, posed-hand geometry supplies learned spatial conditioning rather than a latent projection objective.

Multiscale additive injection. Separate 3D encoders transform $\boldsymbol{G}_{\mathrm{T}}$ and $\boldsymbol{G}_{\mathrm{H}}$ into features at resolutions $r \in \{16, 8, 4\}$. The channel dimensions match the corresponding velocity-network features $\boldsymbol{F}_r$. Their sum gives the multiscale spatial condition:

$$\boldsymbol{G}_r = \phi_r^{\mathrm{T}}(\boldsymbol{G}_{\mathrm{T}}) + \phi_r^{\mathrm{H}}(\boldsymbol{G}_{\mathrm{H}}), \qquad \boldsymbol{F}_r \leftarrow \boldsymbol{F}_r + \boldsymbol{G}_r, \tag{10}$$

where ϕ_r^{T} and ϕ_r^{H} are the tactile and posed-hand 3D encoders at resolution r. We add $\boldsymbol{G}_r$ at the 16^3 and 8^3 encoder stages and at the 4^3 bottleneck. Spatially structured local features also underpin Convolutional Occupancy Networks (Peng et al., 2020); in our model, they condition the latent velocity network at multiple resolutions.

B.4 Feature Interfaces, Training, and Shape Sampling

Chart features and global context. A shared PointNet-style encoder (Qi et al., 2017) applies pointwise transformations and masked pooling to the position–status samples in each tactile chart. We add an embedding of the mean valid chart position, then use a Transformer to model relationships among charts. Invalid and padded charts are masked. The resulting local tokens are $\boldsymbol{L} \in \mathbb{R}^{M \times d}$, where $M = 20$ is the maximum number of tactile observations and $d = 768$ is the conditioning width. Following Touch2Shape (Wang et al., 2025), a separately pretrained tactile encoder and projection provide the normalized global feature $\boldsymbol{q}_{\mathrm{G}}$ and remain frozen during conditional flow training.

Two trainable adapters A_{FiLM} and A_{attn} map $\boldsymbol{q}_{\mathrm{G}}$ to the modulation and attention interfaces. With time embedding $E_{\mathrm{time}}(t)$, the conditions are as follows:

$$\begin{aligned} \boldsymbol{c}_t &= E_{\mathrm{time}}(t) + A_{\mathrm{FiLM}}(\boldsymbol{q}_{\mathrm{G}}) + \boldsymbol{c}_{\mathrm{stat}}, \\ \boldsymbol{C} &= \begin{bmatrix} A_{\mathrm{attn}}(\boldsymbol{q}_{\mathrm{G}})^{\mathsf{T}} \\ \boldsymbol{L} \end{bmatrix}, \end{aligned} \tag{11}$$

where $\boldsymbol{c}_t$ controls feature-wise affine modulation following FiLM (Perez et al., 2018) in the 3D U-Net, while $\boldsymbol{C}$ supplies one global token and the local chart tokens for cross-attention. Spatial features enter separately through Eq. 10.

Latent representation and optimization. Learned shape representations include continuous signed-distance functions in DeepSDF (Park et al., 2019) and occupancy decision boundaries in Occupancy Networks (Mescheder et al., 2019). Models such as 3DShape2VecSet further combine shape autoencoding with latent generative modeling (Zhang et al., 2023). We use a pretrained sparse-structure representation as specified below.

The frozen TRELLIS sparse-structure encoder (Xiang et al., 2025) maps a 64^3 surface-support occupancy volume to $\boldsymbol{z} \in \mathbb{R}^{8\times16\times16\times16}$. The occupied voxels are obtained from surface points rather than by filling the object's interior. Training-set statistics $\boldsymbol{\mu}_{\mathrm{z}}$ and $\boldsymbol{\sigma}_{\mathrm{z}}$ standardize the target:

$$\widetilde{\boldsymbol{z}} = (\boldsymbol{z} - \boldsymbol{\mu}_{\mathrm{z}})/\boldsymbol{\sigma}_{\mathrm{z}}, \qquad \boldsymbol{z} = \widetilde{\boldsymbol{z}} \odot \boldsymbol{\sigma}_{\mathrm{z}} + \boldsymbol{\mu}_{\mathrm{z}}, \tag{12}$$

where $\widetilde{\boldsymbol{z}}$ is the standardized latent and all arithmetic is elementwise, including multiplication denoted by $\odot$. The statistics $\boldsymbol{\mu}_{\mathrm{z}}$ and $\boldsymbol{\sigma}_{\mathrm{z}}$ are computed from the training set and then fixed. Eq. 2 jointly trains the tactile encoder, the statistical interaction encoder, the spatial interaction encoders, the proprioception encoders, and the velocity U-Net $v_{\boldsymbol{\theta}}$. The pretrained global tactile encoder and projection, together with the TRELLIS sparse-structure encoder and decoder, remain frozen. The straight interpolation and endpoint-velocity regression in Eq. 2 are also related to rectified flow (Liu et al., 2023); our training does not use its iterative reflow procedure.

During training, K is sampled uniformly from $\{0, \ldots, 5\}$ and K distinct grasps are drawn from those containing at least one valid tactile chart; successful contact is not required. Joint condition dropout removes the tactile features, interaction statistics, and spatial representations together, yielding the empty-observation condition $\mathcal{O}_0$ used for classifier-free guidance.

Numerical integration and guidance. For guidance scale ω, classifier-free guidance (Ho & Salimans, 2022) is as follows:

$$\boldsymbol{v}_{\mathrm{cfg}} = \boldsymbol{v}_{\mathrm{uncond}} + \omega(\boldsymbol{v}_{\mathrm{cond}} - \boldsymbol{v}_{\mathrm{uncond}}), \tag{13}$$

where $\boldsymbol{v}_{\mathrm{uncond}}$ and $\boldsymbol{v}_{\mathrm{cond}}$ evaluate the same latent state and time under the empty and observed conditions, respectively. Both reconstruction and planning use Euler integration on 50 equally spaced time points from zero to one. With $\boldsymbol{x}_m$ denoting the latent state at time t_m, the 49 updates are $\boldsymbol{x}_{m+1} = \boldsymbol{x}_m + (t_{m+1} - t_m)\boldsymbol{v}_{\mathrm{cfg}}(\boldsymbol{x}_m, t_m)$ for $m = 0, \ldots, 48$. Generated latents are destandardized and decoded to occupancy logits.

Planning uses $N_{\mathrm{A}} = 16$ samples with $\omega = 1$, whereas final reconstruction uses $N_{\mathrm{F}} = 8$ with $\omega = 5$ at every evaluated grasp budget. The two sets use separate noise streams with base seed 131. Planning operates on decoded fields, while final reconstruction extracts surfaces by marching cubes (Lorensen & Cline, 1987) at zero logit.

Geometric Medoid. For extracted candidate surfaces $\{\widehat{\mathcal{S}}_n\}_{n=1}^{N_{\mathrm{F}}}$, let d_{CD} denote symmetric squared Chamfer distance between sampled surface point clouds. We select the following hypothesis as the reconstruction result:

$$n^\star = \arg\min_{n \in \{1,\ldots,N_{\mathrm{F}}\}} \frac{1}{N_{\mathrm{F}}} \sum_{m=1}^{N_{\mathrm{F}}} d_{\mathrm{CD}}(\widehat{\mathcal{S}}_n, \widehat{\mathcal{S}}_m). \tag{14}$$

Pairwise medoid distances are computed in the normalized object frame, whereas reported CD is computed after transforming the points back to metric coordinates and applying the benchmark scaling factor of 9,000. Ground-truth geometry is used only for evaluation and does not affect medoid selection.

B.5 Generative Active Exploration

This section details the two stages summarized in Section 3.5: geometric risk-based grasp evaluation, followed by grasp selection and update. We first define the available candidate grasps, then describe the geometric contact model, two-step evaluation, directional novelty, and the update after the selected grasp is executed. Earlier tactile exploration work selects continuous query paths on an unknown surface (Driess et al., 2017); our action space consists of the discrete grasp candidates specified below.

B.5.1 Candidate Grasp Set and Shared Initialization

For object o, preprocessing provides a fixed set $\mathcal{A}_o \subseteq \{0, \ldots, 49\}$ of candidate grasps with valid tactile records and base poses. Grasps that produce valid non-contact observations are retained; contact is not required for inclusion. Let $\mathcal{E}_K$ contain the indices of the K executed grasps. The available candidate grasps are as follows:

$$\mathcal{A}_K = \mathcal{A}_o \setminus \mathcal{E}_K. \tag{15}$$

The active comparison uses 947 eligible objects and a shared random initial grasp with seed 131. All methods therefore start from the same observations and reconstruction at $K = 1$, and active selection determines grasps two through five.

Candidate poses and feasibility are supplied by the offline benchmark. The planner evaluates each available grasp without accessing its unrevealed tactile charts, realized contact states, or ground-truth reconstruction error. The recorded tactile observations and hand configuration are revealed only after a grasp is selected. The experiment therefore evaluates selection from a fixed candidate set rather than unconstrained grasp synthesis.

B.5.2 Geometric Risk of Shape Hypotheses

Evaluating grasps across multiple plausible shapes has a precedent in uncertainty-aware grasp planning (Lundell et al., 2019). Here, hypotheses support the selection of informative observations rather than the maximization of grasp robustness.

At each decision round, the generator produces $N = 16$ shape hypotheses with uniform initial weights. Thresholding each decoded field at zero gives a binary shape volume. For hypothesis i, we compute the following quantities:

$$\begin{aligned} d_i(\boldsymbol{x}) &= \mathrm{EDT}(\mathrm{inside}_i)(\boldsymbol{x}) - \mathrm{EDT}(\mathrm{outside}_i)(\boldsymbol{x}), \\ \bar{d}_i(\boldsymbol{x}) &= \mathrm{clip}(d_i(\boldsymbol{x})/\kappa, -1, 1), \end{aligned} \tag{16}$$

where $\boldsymbol{x}$ indexes a voxel in the decoded volume, EDT is the Euclidean distance transform, $\kappa = 3$ voxels, inside_i denotes occupied voxels, and $\mathrm{outside}_i$ denotes their complement. Positive distances lie inside the thresholded shape volume. Because the decoded field represents sparse surface support rather than a filled solid, d_i is used as a geometric contact surrogate rather than an exact physical object SDF. Its sign convention is opposite to the negative-inside hand-primitive convention in Eq. 7.

To measure the geometric ambiguity represented by the hypotheses, we first define a pairwise disagreement D_{ij} for each pair of decoded shapes. The geometric risk $\mathcal{U}(\boldsymbol{w})$ then averages these values according to the current hypothesis weights. Since most of the decoded volume is empty, a comparison over the entire volume would be dominated by empty space shared by both hypotheses. We therefore restrict each comparison to the union of the b-voxel neighborhoods around their predicted surfaces:

$$\mathcal{V}_{ij} = \{\boldsymbol{x} : |d_i(\boldsymbol{x})| \le b \ \vee \ |d_j(\boldsymbol{x})| \le b\}, \qquad b = 2 \text{ voxels}. \tag{17}$$

Within this surface band, the pairwise geometric disagreement is defined as:

$$\begin{aligned} D_{ij} = \frac{\lambda}{|\mathcal{V}_{ij}|} \sum_{\boldsymbol{x} \in \mathcal{V}_{ij}} \mathbf{1}[(d_i(\boldsymbol{x}) > 0) \neq (d_j(\boldsymbol{x}) > 0)] \\ + \frac{1-\lambda}{2|\mathcal{V}_{ij}|} \sum_{\boldsymbol{x} \in \mathcal{V}_{ij}} |\bar{d}_i(\boldsymbol{x}) - \bar{d}_j(\boldsymbol{x})|, \qquad \lambda = 0.65, \end{aligned} \tag{18}$$

where $\mathbf{1}[\cdot]$ is the indicator function and $|\mathcal{V}_{ij}|$ is the number of voxels in the surface band. The first term measures inside–outside disagreement, while the second compares normalized, truncated distances to the two predicted surfaces. We set $D_{ii} = 0$ and define the disagreement as zero if $\mathcal{V}_{ij}$ is empty. The resulting symmetric matrix is computed once per decision round and reused to evaluate all candidate grasps. This planning-specific measure is distinct from the surface-point Chamfer distance used for medoid selection and reconstruction evaluation. Selecting observations through their expected effect on predictive uncertainty is a classical principle in statistical active learning (Cohn et al., 1996). Our criterion instantiates this principle with a task-specific geometric disagreement measure; it is not Shannon entropy or mutual information.

B.5.3 Tactile Outcome Prediction

Nominal Closing Trajectory. Each candidate base pose determines a nominal Allegro closing trajectory. We sample the four distal finger segments at five closing stages, using nine points per segment, and denote normalized sample ℓ on finger f at stage s for candidate grasp a by $\boldsymbol{p}_{afs\ell}$. For shape hypothesis i, trilinear interpolation of the untruncated distance field gives the contact evidence e_{iafs}, which is converted into the conditional probability h_{iafs} that the finger makes contact at stage s given that it has not contacted at an earlier stage:

$$e_{iafs} = \max_{\ell} d_i(\boldsymbol{p}_{afs\ell}), \qquad h_{iafs} = \sigma((e_{iafs} + r_{\mathrm{vox}})/\tau), \tag{19}$$

where σ is the logistic function and $\tau = 0.75$ voxels controls the transition from separation to contact. At decoded resolution $R = 64$, the physical finger radius $r_{\mathrm{m}} = 0.012\,\mathrm{m}$ is converted to $r_{\mathrm{vox}} = r_{\mathrm{m}} s_o R$ so that contact accounts for the finger's finite thickness. Trajectory samples outside the decoded volume are assigned distance $-R$, corresponding to negligible contact probability, and h_{iafs} is clipped to $[10^{-6}, 1 - 10^{-6}]$ for numerical stability.

Each finger has six possible outcomes along the trajectory: first contact at one of the five stages, or no contact at any stage. Their probabilities are defined as follows:

$$\pi_{iaf}(s) = h_{iafs} \prod_{u<s} (1 - h_{iafu}), \qquad \pi_{iaf}(\text{no contact}) = \prod_{u=1}^{5} (1 - h_{iafu}), \tag{20}$$

where u indexes a closing stage. For contact at stage s, the first product requires no contact at any earlier stage, while the no-contact probability requires the finger to remain out of contact throughout all five stages. This geometric model therefore predicts whether and when contact could first occur along the nominal trajectory; it does not synthesize tactile images or simulate contact dynamics.

Tactile Outcome Hypotheses and Likelihood. For each finger, the five stages are grouped into early contact (stages 1–2), middle contact (stages 3–4), late contact (stage 5), and no contact. Let $q_{iaf}(y_f)$ denote this four-category distribution. Assuming that finger outcomes are conditionally independent given the shape hypothesis and candidate grasp, the tactile outcome hypothesis $\boldsymbol{y} = (y_1, \ldots, y_4)$ has a probability as follows:

$$L_{ia}(\boldsymbol{y}) = \frac{\left[\prod_{f=1}^{4} q_{iaf}(y_f)\right]^{\alpha}}{\sum_{\boldsymbol{y}'} \left[\prod_{f=1}^{4} q_{iaf}(y'_f)\right]^{\alpha}}, \qquad \alpha = 0.75, \tag{21}$$

where $\boldsymbol{y}'$ ranges over all $4^4 = 256$ joint outcomes, which form $\mathcal{Y}_a$, and α tempers the distribution to reduce overconfidence in the approximate geometric contact model. All outcomes are included in the one-step expectation. Eq. 3 uses $L_{ia}(\boldsymbol{y})$ to compute each outcome probability and the corresponding posterior shape-hypothesis weights; Eq. 4 then computes the expected remaining geometric risk. Numerical denominators use $\epsilon = 10^{-12}$.

B.5.4 Grasp Selection and Update

One-step Risk Constraint. The normalized score $\rho(a, \boldsymbol{w})$ defined in Section 3.5 measures the expected one-step geometric risk reduction of candidate grasp a. Before applying two-step evaluation or directional novelty, we retain:

$$\mathcal{C}_K = \left\{ a \in \mathcal{A}_K : \rho(a, \boldsymbol{w}) \geq (1 - \delta_K) \max_{a' \in \mathcal{A}_K} \rho(a', \boldsymbol{w}) \right\}, \tag{22}$$

where δ_K is the round-dependent tolerance and $(\delta_1, \delta_2, \delta_3, \delta_4) = (0.02, 0.05, 0.10, 0.15)$. Every grasp in $\mathcal{C}_K$ therefore remains within the prescribed fraction of the largest one-step risk reduction. Ground-truth CD is not used in this selection.

Two-step Evaluation. Nonmyopic tactile exploration has previously been formulated using Gaussian-process differential entropy for simultaneous multiple contacts (Driess et al., 2019). Our two-step evaluation instead uses hypothetical reweighting of generated shape hypotheses under discrete grasp outcomes.

For a first grasp $a \in \mathcal{C}_K$, let $\mathcal{Y}_a^+$ contain the most probable tactile outcome hypotheses until their cumulative probability reaches 0.995 or 32 outcomes have been retained. Outcomes with probability at most 10^{-12} are omitted. For each retained outcome, we select the second grasp that minimizes the expected geometric risk remaining after its execution under the corresponding posterior weights:

$$\begin{aligned} \overline{\mathcal{U}}_2(a, \boldsymbol{w}) = &\sum_{\boldsymbol{y} \in \mathcal{Y}_a^+} P_a(\boldsymbol{y} \mid \boldsymbol{w}) \min_{a' \in \mathcal{A}_K \setminus \{a\}} \overline{\mathcal{U}}_1(a', \boldsymbol{w}^{a,\boldsymbol{y}}) \\ &+ \sum_{\boldsymbol{y} \notin \mathcal{Y}_a^+} P_a(\boldsymbol{y} \mid \boldsymbol{w}) \mathcal{U}(\boldsymbol{w}^{a,\boldsymbol{y}}), \\ \rho_2(a, \boldsymbol{w}) = &\frac{\max\{\mathcal{U}(\boldsymbol{w}) - \overline{\mathcal{U}}_2(a, \boldsymbol{w}), 0\}}{\max\{\mathcal{U}(\boldsymbol{w}), \epsilon\}}, \end{aligned} \tag{23}$$

where $\overline{\mathcal{U}}_1(a', \boldsymbol{w}^{a,\boldsymbol{y}})$ is the one-step expected risk after the second grasp a' under the posterior induced by first-grasp outcome $\boldsymbol{y}$. Outcomes not expanded in $\mathcal{Y}_a^+$ retain their posterior risk after the first grasp in the second sum. Thus, $\overline{\mathcal{U}}_2(a, \boldsymbol{w})$ evaluates a together with the best second grasp for each retained outcome, and $\rho_2(a, \boldsymbol{w})$ is the corresponding normalized risk reduction used to select the first grasp. The second-grasp minimization considers every remaining grasp in $\mathcal{A}_K \setminus \{a\}$, whereas $\mathcal{C}_K$ restricts only the first-grasp candidates.

At $K = 1, 2$, we rank grasps in $\mathcal{C}_K$ by decreasing ρ_2, then decreasing ρ, decreasing directional novelty as defined below, and increasing grasp index. Only the first grasp of the selected two-step plan is executed. The likelihoods and pairwise geometric disagreements are reused during

hypothetical reweighting, so evaluating the second step does not invoke additional conditional shape generation.

Directional Novelty. In later rounds, directional novelty ranks the grasps that satisfy the one-step risk constraint. Let $\boldsymbol{p}_a$ be the recorded hand base position in the object frame and $\widehat{\boldsymbol{p}}_a = \boldsymbol{p}_a / \|\boldsymbol{p}_a\|_2$. We define the novelty score as the smallest normalized angular separation from any executed grasp:

$$\nu_K(a) = \min_{e \in \mathcal{E}_K} \frac{\arccos\big(\mathrm{clip}(\widehat{\boldsymbol{p}}_a^{\mathsf{T}} \widehat{\boldsymbol{p}}_e, -1, 1)\big)}{\pi}. \tag{24}$$

We choose the grasp in $\mathcal{C}_K$ with the largest ν_K, breaking ties by decreasing ρ and increasing grasp index. The later-round rule therefore combines near-optimal one-step geometric risk reduction with a new approach direction; it does not minimize expected posterior risk over the full candidate set.

Post-execution Update. After choosing $a_K^\star$, we execute the grasp, reveal its recorded tactile charts and hand configuration, and update the tactile observation and executed grasp set as follows:

$$\mathcal{O}_{K+1} = \mathcal{O}_K \cup \mathcal{O}(a_K^\star), \qquad \mathcal{E}_{K+1} = \mathcal{E}_K \cup \{a_K^\star\}, \tag{25}$$

where $\mathcal{O}(a_K^\star)$ denotes the tactile observations and hand configuration revealed by the executed grasp. Prospective grasp evaluation changes only the weights of the current shape hypotheses. After execution, the conditional generator uses $\mathcal{O}_{K+1}$ to produce 16 new shape hypotheses with uniform initial weights for the next decision; the hypothetical posterior weights are discarded, and the realized tactile observations replace the tactile outcome hypotheses used during evaluation.