AdaDexGrasp: Adaptive Dexterous Grasping via 3D Visuo-Tactile Representation Fusion

Xirui Liang[1], Jiaqi Liang[1*], Jingkai Xu[1*], Yuran Wang[1*], Ruochong Li[2], Yuanpei Chen[1], Masayoshi Tomizuka[3], Wei Zhan[3], and Ruihai Wu[3†]

[1] Peking University, Beijing, China
[2] The Hong Kong University of Science and Technology, Hong Kong, China
[3] University of California, Berkeley, CA, USA

Abstract. Humans achieve stable and adaptive grasps by seamlessly integrating visual perception and tactile feedback, a capability that remains challenging to replicate in robotic systems. Existing robotic grasping approaches predominantly rely on visual inputs and lack mechanisms for tactile-guided adaptation after contact, limiting robustness and generalization. To address this challenge, we propose a unified visuo-tactile-fusion grasping framework that integrates grasp generation, feasibility prediction, and adaptive refinement. At its core, our method introduces an efficient visuo-tactile representation that tightly fuses object geometry with tactile feedback by associating tactile signals with finger identities. This unified representation supports contact-aware grasp pose generation during planning and tactile-guided refinement after contact, enabling the system to reason about fine-grained finger-object interactions and adjust grasps dynamically. Comprehensive experiments in both simulation and real-world environments demonstrate that our approach significantly enhances grasp success rates and generalization across diverse objects.

Keywords: Dexterous Grasping · Adaptation · Visuo-Tactile Fusion

1 Introduction

In daily life, humans skillfully use dexterous hands to grasp a wide range of objects. This capability arises not only from the ability to plan an appropriate grasp pose based on visual perception, but also from the competence in adjusting actions in real time through tactile feedback, ensuring stable and reliable grasping. Transferring such visuo-tactile-fusion grasping capabilities to robotic systems with dexterous hands is highly important yet technically challenging.

Most existing robotic grasping methods [2, 8, 33, 38, 39, 45, 50, 53] rely purely on visual information (e.g. RGB, point cloud) for planning. Such approaches determine the hand pose based on an initial observation, without the ability to refine or adapt the grasp according to feedback after contact. Some works [10, 13] explore tactile-only strategies for adaptive grasping under occlusions, but these typically require multiple trial attempts and cannot utilize available visual information, resulting in reduced efficiency.

*Equal contribution. †Corresponding author.

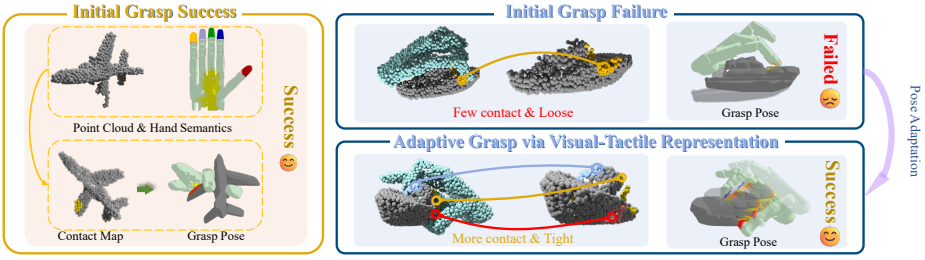

Fig. 1: Overview. (Left) An initial grasping module predicts a contact map from the raw point cloud with finger/palm identities, aiming to generate physically successful hand poses for grasp planning informed by hand semantic information. **(Right-Top)** The initially generated hand pose does not guarantee a successful grasp and may fail due to sparse or unstable contact formations. **(Right-Bottom)** Our method further leverages a visuo-tactile-fusion representation to adapt the initial pose, resulting in more stable contact patterns that enable robust grasp execution.

To enhance visuo-tactile integration for improved performance, recent studies [11, 15, 40, 44, 46] have explored a variety of multimodal fusion strategies. However, most of these efforts concentrate on object manipulation rather than achieving generalizable grasping across a wide range of categories and shapes. Several works [34, 48, 49] have taken initial steps toward visuo-tactile generalization in dexterous grasping. Nevertheless, their approaches mainly emphasize feature-level fusion, aiming to encourage the network to exploit both complementary and shared information from visual and tactile modalities for a richer state representation, while overlooking the spatial alignment and interaction between the two inputs, which are crucial for accurate and adaptive grasp control.

To more effectively bridge the gap between visual and tactile modalities, we propose a novel framework that integrates tactile and object geometric information to enable grasp pose generation, feasibility prediction, and adaptive refinement, which are essential for achieving efficient and robust dexterous grasping but have been largely overlooked in previous research.

Inspired by the human ability to plan which parts of the hand should interact with specific regions of an object to achieve a stable grasp, our approach introduces a more expressive contact representation during initial grasp pose generation. Unlike traditional methods that only indicate contact map without semantic information [21,26,29], we bind contact map with finger/palm identities through tactile signals (Figure 1, Top), allowing it to represent not only contact occurrence but also the expected hand–object associations. This semantic contact map is incorporated into the object point cloud features for grasp pose decoding, enabling the model to anticipate successful tactile responses during the initial planning stage and thereby improving grasp success rates.

However, for complex or unseen objects, the initially generated grasp pose may still fail to ensure a stable grasp [17, 52]. To address this, we leverage tactile feedback obtained during execution to assess and refine the grasp (Figure 1,

Middle and Bottom). Our core idea is to generate tactile contact markers (which represent different contact states such as no contact or specific finger/palm contact) based on tactile signals, and integrate them into the current hand–object point cloud to create a tactile-mapped geometry representation. These point clouds are then used by a Classifier to judge the final grasp success. When a grasp is deemed suboptimal, an Adapt Model is invoked to adjust hand pose based on the current observation, aiming to achieve a successful grasp.

The absence of suitable simulation environments partly obstructs the research on visuo-tactile dexterous grasping. Consequently, we developed a tactile-equipped dexterous hand manipulation environment in IsaacGym [23], enabling efficient simulation-based training and evaluation. Furthermore, we also deployed our visuo-tactile-fusion method to real-world robotic systems. Comprehensive experiments were conducted to assess the generalization and robustness of our method in both simulation and real world. Both quantitative and qualitative results consistently demonstrate the effectiveness and reliability of our approach.

Our main contributions can be summarized as follows:

- We introduce an effective visuo-tactile representation that tightly fuses geometric and tactile information through a contact-aware encoding, enabling fine-grained reasoning over finger–object interactions.
- We develop a unified visuo-tactile grasping framework that integrates grasp generation, feasibility prediction, and adaptive refinement based on tactile-guided feedback, enabling more robust and efficient dexterous grasping.
- Extensive experiments in both simulation and real world demonstrate the effectiveness of our framework.

2 Related Work

2.1 Dexterous Grasping

Dexterous hand has received extensive attention for its potential for human-like manipulation in robotics [4,6,7,16,27,28]. Current dexterous grasping approaches can be broadly categorized into three major paradigms: reinforcement learning (RL), imitation learning (IL), and object-centric policies. RL methods [30,32,42] acquire dexterous manipulation skills through trial-and-error exploration, while IL frameworks [1, 35–37, 41] transfer human demonstrations to robotic systems for efficient policy learning. Object-centric approaches [24, 25, 33], on the other hand, leverage robot proprioception and RGB-D observations to learn visually guided grasping policies. Despite these advancements, most existing methods rely on visual information and lack tactile perception to identify contact instabilities such as slipping or rolling on smooth surfaces [22]. This limitation often leads to grasp failures even when the visual geometry indicates a feasible configuration. Recently, several studies [34, 48, 49] have begun to explore visuo-tactile integration for dexterous grasping. However, these approaches primarily focus on encoding and directly fusing the two modalities, without investigating deeper spatial and semantic relationships between vision and touch. To address this gap,

we propose a visuo-tactile grasping framework that integrates tactile feedback with object geometry, enabling more robust and efficient dexterous grasping.

2.2 Visuo-Tactile Fusion for Manipulation

Fusing tactile and visual information is essential for robust dexterous manipulation [14, 20], as visual perception provides priors about object properties and affordances, while tactile feedback captures contact quality to guide action generation. Existing visuo-tactile fusion approaches fall into two categories: learning-based and raw-data-based methods. Learning-based approaches typically encode each modality into latent representations and combine them through concatenation [40] or pretraining strategies [3, 5, 9, 12, 18, 19]. However, they mainly emphasize global feature alignment while overlooking spatial correspondence and interaction between visual and tactile inputs. Raw-data-based methods [46, 47] attempt direct signal-level fusion, but they likewise fail to capture the geometric and contact-level relationships crucial for coordinated perception. In contrast, our approach explicitly aligns modalities by integrating tactile signals into the object point cloud for grasp generation and constructing tactile-mapped geometric representations for adaptive refinement, effectively bridging visual and tactile spaces to achieve more stable, efficient, and robust dexterous grasping.

3 Method

3.1 Problem Formulation

We consider the problem of learning stable and adaptive grasping with a tactile-equipped dexterous hand. $\mathcal{O}$ denotes the target object, represented by a point cloud $\mathcal{P}_{obj} = \{(x_i, y_i, z_i, r_i, g_i, b_i)\}_{i=1}^{N_0}$ captured by an RGB-D camera sensor.

The robot hand state is defined as $s = [p, q, \mathbf{j}]$, where $p \in \mathbb{R}^3$ and $q \in \mathbb{R}^4$ denote the palm position and orientation, and $\mathbf{j} \in \mathbb{R}^{22}$ represents the joint angles of the 22 actuated joints. During interaction, tactile sensors on the fingertips and palm provide readings $\tau = \{\tau_i\}_{i=1}^{N_t}$ corresponding to local contact pressures, where N_t denotes the number of tactile sensors (six in our setup: five fingertips and one palm).

Our objective is to find a grasp configuration s^* that maximizes the probability of grasp success conditioned on visual and tactile observations:

$$s^* = \arg\max_s \; p(y = 1 \mid \mathcal{P}_{obj}, \tau, s), \tag{1}$$

where $y \in \{0, 1\}$ indicates grasp success. The key challenge is modeling the relationship between visuo-tactile observations and grasp stability.

3.2 Overview

To address this challenge, we propose a unified framework that integrates object geometry and tactile feedback to support grasp pose generation, feasibility

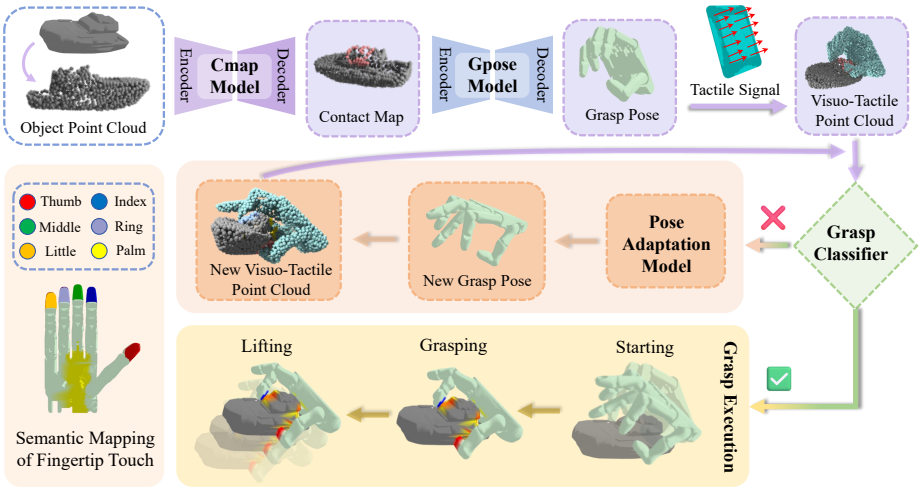

Fig. 2: Pipeline of the proposed visuo-tactile grasping framework. Given an object point cloud and semantic fingertip labels, the system first predicts a contact map using the Cmap model and generates an initial grasp pose with the Gpose model. The predicted hand configuration and tactile feedback are fused into a visuo–tactile point cloud that captures both geometric and contact cues. A grasp classifier evaluates grasp stability, and if failure is predicted, a pose adaptation module iteratively refines the grasp by generating updated poses and visuo–tactile observations until success is predicted or a maximum number of iterations is reached. The final grasp is executed through a staged process of reaching, grasping, and lifting, enabling stable interaction with diverse objects.

prediction, and adaptive refinement. The overall pipeline is illustrated in Figure 2. The architecture of the proposed method (Figure 3) consists of three components:

- **Contact-Driven Grasp Pose Generator** (Section 3.3), which produces the initial grasp pose.
- **Grasp Pose Classifier Model** (Section 3.4), which predicts grasp feasibility and stability.
- **Grasp Pose Adaptation Model** (Section 3.5), which refines suboptimal grasp poses.

Closed-Loop Grasp Optimization (Section 3.6) integrates these components into an iterative refinement process. Section 3.7 further describes the training data collection procedure.

3.3 Contact-Driven Grasp Pose Generator

Given the point cloud of an object, we generate an initial grasp hypothesis by predicting a contact distribution over the object surface. Model h_ψ predicts a probabilistic contact map $\mathcal{M}$:

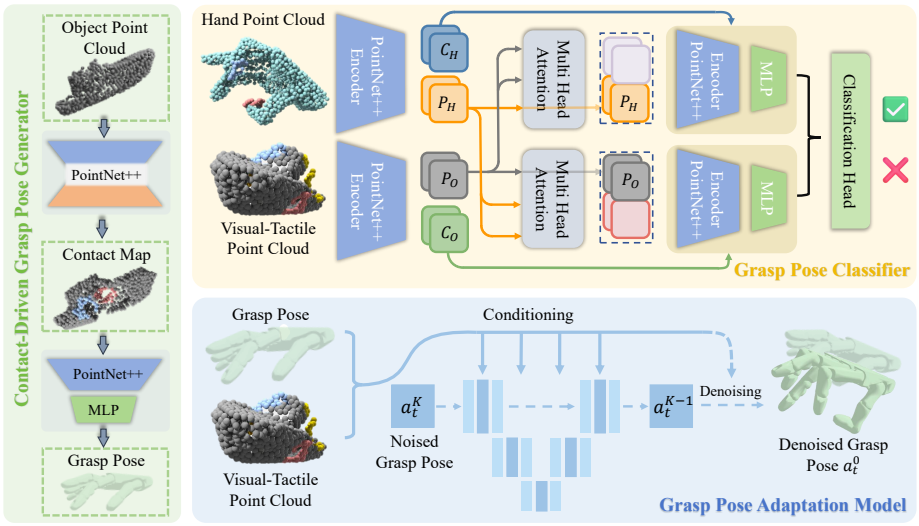

Fig. 3: Architecture of the grasping framework. The system consists of a contact-driven grasp generator, a grasp classifier, and a grasp adaptation module for stable dexterous grasping. **Green:** the object point cloud is processed by PointNet++ to predict a visuo–tactile contact map and generate an initial grasp pose. **Yellow:** a dual-encoder PointNet++ architecture encodes the hand and visuo–tactile point clouds, followed by multi-head attention for grasp success prediction. **Blue:** if failure is predicted, a diffusion-based adaptation module iteratively refines the grasp pose conditioned on visuo–tactile features.

$$\mathcal{M} = \{\mathcal{M}_i \mid \mathcal{M}_i \in \{1,2,3,4,5,6\}, \\ i = 1,2,\ldots,N_0\} = h_{\psi_1}(\mathcal{P}_{obj}) \tag{2}$$

where each label $\mathcal{M}_i$ corresponds to a specific hand part (finger or palm). Thus, the predicted contact map encodes potential tactile interactions between the hand and the object.

The grasp generator h'_ψ then conditions on $(\mathcal{P}_{obj}, \mathcal{M})$ to predict an initial grasp state:

$$s^{(0)} = h'_{\psi_2}(\mathcal{P}_{obj}, \mathcal{M}). \tag{3}$$

This estimate s_0 serves as an initial grasp pose that often succeeds directly or provides a good starting point for subsequent refinement.

During initial grasp generation, tactile inputs are unavailable because no contact has occurred. Each training sample includes hand joint states, grasp success labels, and visuo–tactile point clouds. The contact map prediction is supervised using cross-entropy loss:

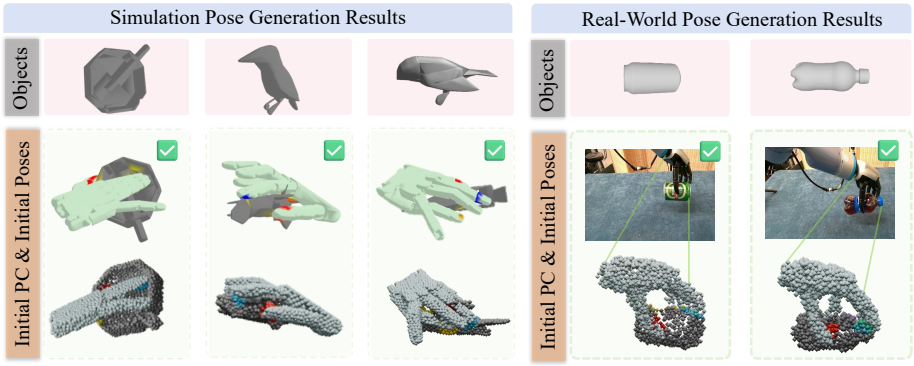

Fig. 4: Visualization of Grasp Pose Generation. Row 1 presents object samples in both simulation and the real world. As shown in Row 2 and Row 3, the predicted contact map with semantic information (finger/palm identities) enables grasp pose generator to generate successful pose for target object grasping.

$$\mathcal{L}_{CE}(\mathcal{M}, \mathcal{M}_s) = -\sum_i M_{s,i} \log M_i \tag{4}$$

while the grasp pose is supervised using mean-squared error:

$$\mathcal{L}_{MSE} = \|s^{(0)} - s^s\|_2^2 \tag{5}$$

Training data are collected through RL rollouts and can be adapted to different dexterous hands by training the policy on the corresponding hand models. Figure 4 visualizes the initial grasp generation process.

3.4 Grasp Pose Classifier Model

Due to the diversity of hand poses and object-dependent grasp requirements, an initially generated grasp may not always succeed. We therefore predict grasp feasibility based on the current hand–object interaction state.

After executing the initial grasp pose s_0, we annotate the object point cloud with tactile signals derived from contact forces at the five fingertips and the palm. Each contact region is assigned a unique ID corresponding to the contacting hand part, producing a tactile-mapped point cloud $\mathcal{P}_{vt}$.

It is worth clarifying the distinction between the contact map $\mathcal{M}$ and the tactile-mapped point cloud $\mathcal{P}_{vt}$, as both encode contact-related information but play different roles. The contact map $\mathcal{M}$ is a *predicted* per-point categorical prior inferred from the object point cloud and used *before* execution to guide initial grasp generation. In contrast, $\mathcal{P}_{vt}$ is the *post-contact* object point cloud annotated with *measured* tactile signals, consumed by the classifier and the adaptation model.

Given $\mathcal{P}_{vt}$, the classifier estimates the probability that the current configuration yields a stable grasp. Formally, the model learns a discriminative mapping:

$$\hat{y} = f_\theta(\mathcal{P}_{vt}, s), \tag{6}$$

where $\hat{y} \in [0, 1]$ denotes the predicted grasp success probability and f_θ is a PointNet-based network. The model is trained using binary cross-entropy:

$$\mathcal{L}_{cls} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]. \tag{7}$$

By leveraging both geometric structure and tactile contact patterns, f_θ provides a differentiable estimate of grasp stability that can be evaluated in real time.

3.5 Grasp Pose Adaptation Model

To recover from failed grasps, we introduce a grasp pose adaptation model g_ϕ that transforms unstable grasps into stable ones.

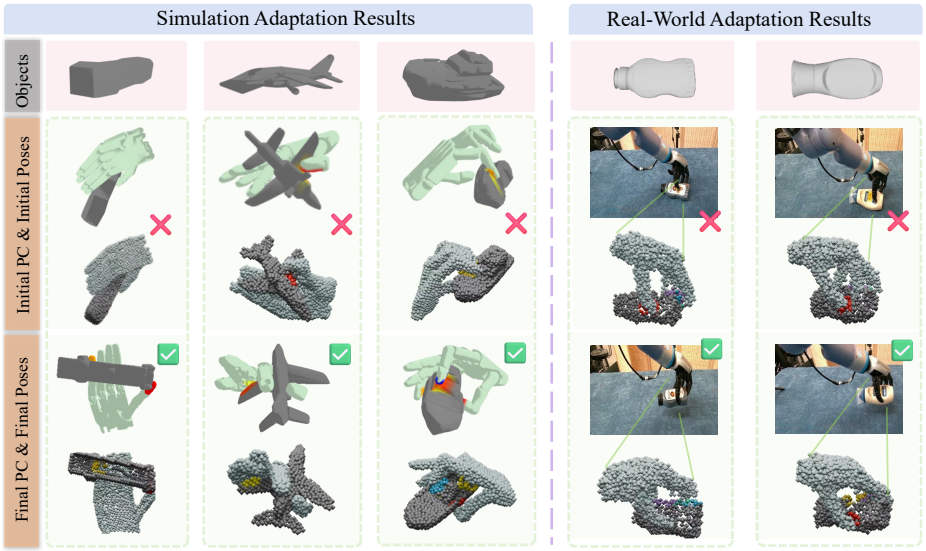

Fig. 5: Visualization of Dexterous Grasp Pose Adaptation. Row 1 presents object samples in both simulation and the real world. As shown in Row 2 (hand–object interaction mesh) and Row 3 (hand–object interaction point cloud), the initially generated grasp poses may fail to securely grasp the object, as evidenced by sparse or weak contact regions that lead to loose or unstable grasps. Row 4 and Row 5 illustrate the refined grasp poses and updated contact states after adaptation, where hand–object interaction becomes significantly more stable, with denser contact regions and a larger support area in the point cloud. These observations further validate the effectiveness of the proposed adaptation module in correcting suboptimal initial grasps.

For training, we construct paired tactile-enhanced states:

$$(\mathcal{P}_{vt}^{f}, s^{f}, y=0), \quad (\mathcal{P}_{vt}^{s}, s^{s}, y=1),$$

where the successful grasp s^s is geometrically and temporally close to the failed grasp s^f. Concretely, for each object we pool successful and failed states from the PPO rollouts immediately before lifting, so that all candidates share the same interaction stage, and transform them into an object-centric frame to remove any dependence on absolute object placement. Each failed state s^f is then paired with its nearest successful counterpart s^s under the distance

$$d(s, s') = 10\sqrt{\|p - p'\|_2^2 + \theta_{q,q'}^2} + \|\mathbf{j} - \mathbf{j}'\|_2, \tag{8}$$

where $(p, q, \mathbf{j})$ denote the palm translation, palm orientation, and the 22 joint angles, and $\theta_{q,q'}$ is the geodesic distance between the two orientations.

The model learns a corrective mapping:

$$\Delta s = g_\phi(\mathcal{P}_{vt}^{f}, s^{f}), \tag{9}$$

where Δs represents adjustments to the hand pose and joint angles. The refined grasp is:

$$s^{f,\text{new}} = s^{f} + \Delta s. \tag{10}$$

The training objective encourages the corrected pose to approach the successful grasp while maximizing predicted grasp stability:

$$\mathcal{L}_{gen} = \|\Delta s - (s^{s} - s^{f})\|_2^2 + \lambda(1 - f_\theta(\mathcal{P}_{vt}^{f}, s^{f} + \Delta s)). \tag{11}$$

This objective enables the model to learn corrective actions guided by both tactile perception and grasp success prediction. Figure 5 visualizes the adaptation process.

3.6 Closed-Loop Grasp Optimization

The contact-driven grasp generator, grasp classifier, and grasp adaptation model are integrated into a closed-loop refinement process. Starting from the initial pose s_0, the system iteratively evaluates and updates the grasp:

$$\hat{y}_k = f_\theta(\mathcal{P}_{vt}^{k}, s^{(k)}), \tag{12}$$

$$s_{k+1} = \begin{cases} s_k, & \text{if } \hat{y}_k > \tau_{succ}, \\ s_k + g_\phi(\mathcal{P}_k, s_k), & \text{otherwise}, \end{cases} \tag{13}$$

$$\text{stop if } \ \hat{y}_k > \tau_{succ} \text{ or } k = K. \tag{14}$$

Here τ_{succ} is the success threshold. The process terminates when the predicted success probability exceeds τ_{succ} or the maximum iteration number K is reached. The resulting grasp configuration s^* satisfies both geometric feasibility and tactile stability, enabling robust grasping under diverse contact conditions.

3.7 Tactile-Enhanced Data Collection via Reinforcement Learning

We collect training data using reinforcement learning following the first-stage PPO setup in UniDexGrasp [43]. This strategy generates diverse grasp poses on the same object while maintaining stable grasps with minimal refinement.

A dexterous hand is trained to grasp across various objects using Proximal Policy Optimization (PPO) [31]. Each interaction produces a trajectory

$$\mathcal{T} = \{(s_t, a_t, r_t, \tau_t)\}_{t=1}^{T},$$

where a_t and r_t denote the control command and grasp reward respectively. Both successful and failed attempts are collected to ensure diverse tactile data.

During data collection, tactile information is annotated for each interaction by monitoring contact forces on the fingertips and palm. If the force on a finger exceeds a predefined threshold (0.01N in our experiments), that finger is marked as active. For each detected contact, we select nearby object points and assign them corresponding contact labels.

Based on these annotations, we construct a tactile-mapped point cloud $\mathcal{P}$ with contact identifiers for each hand part. The representation is defined as

$$\mathcal{P}_{vt} = \{(x_i, y_i, z_i, r_i, g_i, b_i, k_i, \mathbf{c}_i) \mid i = 1, \ldots, N\}, \tag{15}$$

where (x_i, y_i, z_i) denote 3D coordinates and (r_i, g_i, b_i) are color channels. The scalar $k_i \in \mathbb{R}$ indicates the contact state between the hand and the object, while $\mathbf{c}_i \in [0, 1]^6$ represents a six-dimensional tactile intensity vector corresponding to the five fingertips and the palm.

This fused representation captures both geometric contact distribution and tactile intensity, and serves as the unified input for subsequent models. Each dataset sample is represented as $(\mathcal{P}_{vt}, s, y)$, where the binary label y indicates grasp success or failure.

4 Experiments

We evaluate the proposed visuo-tactile fusion grasping framework in both simulation and real-world settings using a tactile-enabled dexterous hand. The simulation setup is described in Section 4.1, followed by baseline comparisons (Section 4.2) and ablation study (Section 4.3). Real-world experiments (Section 4.4) further validate the feasibility and practicality of the proposed approach.

4.1 Simulation Experimental Setup

Our simulation environment is built upon IsaacGym [23], enabling large-scale parallel collection of grasp trajectories and tactile feedback under realistic physics. The grasping task consists of three subtasks: reaching, grasping, and lifting.

Assets. The dexterous hand is based on ShadowHand, featuring five fingers with 22 degrees of freedom (DoF). Including the 6-DoF root pose controlling spatial motion, the system has 28 DoF in total. In simulation, the hand is moved

to target positions by directly modifying the root pose. Object assets are taken from UniDexGrasp++ [33], including 50 objects from 6 categories. Among them, 20 objects are used for training and 30 unseen-category objects for testing.

Metrics. A grasp is considered successful if the object remains stable without slippage for at least 1 second after lift-off. We report the **grasp success rate** and evaluate performance on three object sets:

- Seen Objects. Objects used in training.
- Unseen Objects. Objects from the same categories but unseen in training.
- Unseen Categories. Objects from categories absent in training.

4.2 Baselines

We compare our method against ten state-of-the-art dexterous grasping frameworks, all baselines are evaluated under identical simulation settings using the same grasp success metric.

- **DexGraspAnything** [54]: a large-scale grasp generator trained on massive grasp–object pairs, primarily exploiting visual geometry.
- **UniDexGrasp++** [33]: a diffusion-based framework that predicts grasp poses for multiple robotic hands without tactile feedback.
- **DexDiffuser** [39]: a model that samples diverse grasp configurations from object point clouds.
- **UniDexGrasp(PPO)** [43]: a PPO policy based on UniDexGrasp using our visuo-tactile inputs.
- **Robot Synesthesia**: a tactile distillation approach using tactile point clouds.
- **Intuitive Closed-Loop**: a heuristic closed-loop policy following the rule "move closer when no contact is detected".
- **ContactDexNet** [51]: a method that proposes grasp poses from predicted contact maps.
- **DexGraspVLA** [53]: a VLA-based grasping model using RGB input.
- **UniDexGrasp-CL** [43]: UniDexGrasp augmented with a failure classifier that resamples a new grasp whenever failure is predicted, giving it closed-loop refinement capability.
- **DexDiffuser-VT** [39]: DexDiffuser extended with tactile conditioning, where tactile features are concatenated with its visual input during evaluation and refinement.

From Table 1, we observe that our method consistently outperforms all baseline approaches across the three object sets, demonstrating superior robustness and generalization capability. Through analyzing the grasp poses generated by the baselines during experiments, we identify four major issues.

1) During the initial grasp pose generation, generative methods (Row 1, Row 2, Row 3) rely solely on visual geometry; without intermediate semantic correspondence, they often produce infeasible grasp configurations for complex objects, resulting in grasp failures. 2) While end-to-end models (Row 3) directly

Table 1: Comparison Results between Baselines and Our Method.

Method	Seen Objects	Unseen Objects	Unseen Categories
UniDexGrasp++	55%	49%	42%
DexGraspAnything	77%	72%	67%
DexDiffuser	71%	69%	55%
UniDexGrasp (PPO)	52%	46%	40%
Robot Synesthesia	33%	29%	22%
Intuitive Closed-Loop	76%	71%	65%
ContactDexNet	72%	68%	63%
DexGraspVLA	69%	60%	58%
UniDexGrasp-CL	59%	52%	46%
DexDiffuser-VT	75%	71%	57%
Ours	**91%**	**82%**	**83%**

map RGB inputs to continuous actions, they lack precise understanding of geometric structures, making them fail easily when handling hard-to-grasp objects. 3) Lacking real-time tactile feedback, even offline contact modeling (Row 7) yields only static predictions. Consequently, it cannot perceive actual contact states to perform secondary closed-loop adjustments against initial spatial misalignments. 4) In the refinement stage, existing visuo-tactile methods struggle with stability. Pure RL policy (Row 4) suffers from high-dimensional exploration noise; distilled model(Row 5) remains brittle to local geometry shifts; and the Intuitive Closed-Loop policy blindly tightens its fingers toward the object without geometric awareness, inevitably leading to failure. Resampling-based refinement (Row 9, UniDexGrasp-CL) stays confined to the same generative distribution and, lacking contact information from the failed attempt, cannot escape the original failure modes; and DexDiffuser-VT (Row 10) fuses modalities weakly, as its Basis Point Set encoding captures only global object features while destroying the local point-cloud structure that tactile signals depend on.

In contrast, our approach incorporates a contact-aware representation that encodes finger and palm identities into grasp generation, along with tactile-guided iterative refinement. This enables feasible contact prediction and real-time pose correction. The high success rates on both unseen objects and unseen categories highlight the strong generalization capability of our framework, suggesting that combining visual geometry with semantic and tactile cues is critical for reliable dexterous grasping.

Furthermore, the performance gap between our method and the baselines increases for unseen categories, indicating that baselines struggle to generalize to novel object shapes and configurations. This emphasizes the importance of incorporating hand–object semantic awareness and closed-loop tactile feedback in developing robust grasping policies. The gains also persist at scale: on the DexGrasp Anything dataset [54] (15k+ objects), our method still reaches 87%.

4.3 Ablations

To investigate the contribution of each component, we conduct ablation experiments focusing on the tactile encoding and adaptation. Our ablations include:

- **w/o contact ID in generation.** All fingertips share a unified contact ID during grasp pose generation, meaning that the contact map feature of input point cloud does not contain any semantic information.
- **w/o contact ID in adaptation.** All fingertips share a unified contact ID during both the classifier and adaptation models, meaning that the contact map feature of input point cloud does not contain any semantic information.
- **w/o adaptation.** The objects are directly lifted after the initial pose generation without adaptation.
- **Object-only.** Only object point cloud input with tactile mapping.
- **Full Point Cloud w/o tactile.** The combined hand-object point cloud is used, excluding all tactile-related features.
- **w/o PC.** Use images labeled with tactile signals instead of point clouds labeled with same tactile information.

Table 2: Ablation Results about Our Method.

Method	Seen Objects	Unseen Objects	Unseen Categories
w/o contact id in adaptation	84%	72%	73%
w/o contact id in generation	79%	75%	69%
w/o adaptation	81%	78%	59%
Object-only	72%	67%	61%
Full Point Cloud w/o tactile	78%	74%	67%
w/o PC	74%	71%	64%
Ours	**91%**	**82%**	**83%**

Table 2 shows the ablation results about our method. As shown in Table 2 (Row 1, Row 2, and Row 7), the semantic information provided by contact ID proves essential for both the initial hand pose generation and subsequent refinement. It enables the model to be aware of the current prediction or contact states, thereby producing more stable hand poses. Moreover, the comparison between Row 3 and Row 7 highlights the critical role of the adaptation module, which effectively refines suboptimal initial poses. This improvement is especially notable on unseen objects (from 78% to 82%) and unseen categories (from 59% to 83%), demonstrating the strong robustness and generalization capability of our approach. Furthermore, the comparison between Row 4, Row 5 and Row 7 shows that without hand PC, we cannot have hand-object relation, and without tactile signals, the adaptation will turn harder. Additionally, the comparison between Row 6 and Row 7 demonstrates that without PC, the network cannot fully understand the geometric structure of objects through images. We further show hand pose generation performance empowered by contact id and adaptation performance in Figure 4 and Figure 5. Figure 4 shows that under the guidance of the contact ID, the contact-driven grasp pose generator successfully produced initial poses that are geometrically consistent and meet the task requirements,

and Figure 5 demonstrates the dynamic refinement process, where the adaptation module iteratively transforms initial unstable grasps into secure ones.

4.4 Real-World Experiment

Table 3: Real World Evaluation Results.

Method	Seen Objects	Unseen Objects	Unseen Categories
DexGraspAnything	81%	71%	73%
DexDiffuser	77%	65%	52%
DexGraspVLA	64%	57%	55%
Ours	**90%**	**87%**	**81%**

To evaluate our approach in real-world settings, we conduct physical experiments using a Psibot robotic hand equipped with high-resolution tactile sensors. The real-world environment setting is shown in Figure 6.

We evaluate our method on more than 20 objects with distinct geometric and material properties. The evaluation metrics are consistent with those in simulation. A grasp is considered successful if the object remains stably held for at least three seconds after lifting. Table 3 reports the final evaluation results. Figure 4 and Figure 5 also illustrate the hand pose generation and adaptation performance in real world. More visualization results can be found in supplementary material.

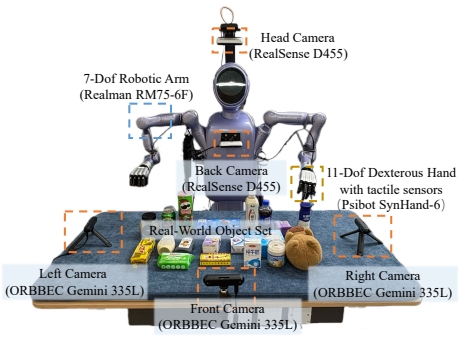

Fig. 6: Real-World Setup.

4.5 Representation Choice

Selecting an effective state representation is a critical trade-off between informational density and the feasibility of sim-to-real transfer. Our tactile-mapped point cloud, denoted as $\mathcal{P}_{vt} = \{(x_i, y_i, z_i, r_i, g_i, b_i, k_i, \mathbf{c}_i)\}$, simultaneously provides RGB, 3D geometry, and local contact semantics derived from tactile feedback.

While we explored richer tactile modalities such as force directions and shear forces, we found that the denser representation resulted in much greater data requirements, while our current tactile representation can already tackle most scenarios for dexterous grasping with less data. In small data tests, our accuracy and the accuracy when using dense modality were 88% and 81%, respectively. Moreover, choosing rich modalities places significant demands on both the real-world hardware and the sim-to-real gap of tactile representations, making it difficult to align the configurations in simulation and the real world. We also tested

accuracy in real world, and the accuracy of our method and the method using the dense modality were 85% and 72%, respectively. Therefore, we eventually selected the relatively simple representations due to the effectiveness and efficiency, real-world and simulation alignment, achieving good experimental performance.

4.6 Deep Analysis in Tactile Signals

Beyond quantitative success rates, we analyze the physical interpretability of the learned contact semantics to verify their correlation with grasp stability. Specifically, since our tactile signals are normalized to a range of $[0, 1]$, we conducted detailed experiments to explore the relationship between tactile signals and stable grasping. Figure 7 shows the detailed analysis.

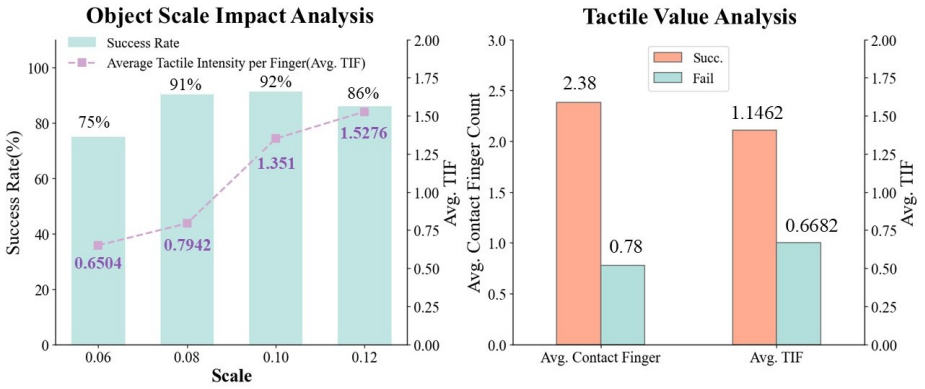

Fig. 7: (Left) As the size of an object increases, Avg. TIF also increases, while grasping success rate first rises and then slightly decreases. This is because small objects are prone to slipping, whereas too large objects are difficult to grasp. **(Right)** The Avg. TIF and the number of contact fingers in successful grasping postures are higher than those in failed grasping postures.

5 Conclusion

In this work, we presented a unified visuo-tactile-fusion grasping framework that enables robust and generalizable dexterous grasping by tightly coupling tactile feedback with visual geometric perception. Through a contact-aware representation that binds tactile signals with finger identities and a tactile-guided adaptation mechanism for real-time refinement, our method enables fine-grained reasoning over finger–object interactions, enhancing both initial grasp feasibility and adaptive stability in simulation and real-world settings, further providing the research community with a feasible and scalable technical route for integrating visual and tactile modalities in dexterous manipulation.

References

1. Alessi, C., Vasile, F., Ceola, F., Pasquale, G., Boccardo, N., Natale, L.: Hannesimitation: Grasping with the hannes prosthetic hand via imitation learning (2025), https://arxiv.org/abs/2508.00491
2. Chen, S., Bohg, J., Liu, C.K.: Springgrasp: Synthesizing compliant, dexterous grasps under shape uncertainty (2024), https://arxiv.org/abs/2404.13532
3. Chen, Y., Sipos, A., der Merwe, M.V., Fazeli, N.: Visuo-tactile transformers for manipulation (2022), https://arxiv.org/abs/2210.00121
4. Dafle, N.C., Rodriguez, A., Paolini, R., Tang, B., Srinivasa, S.S., Erdmann, M., Mason, M.T., Lundberg, I., Staab, H., Fuhlbrigge, T.: Extrinsic dexterity: In-hand manipulation with external forces. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). pp. 1578–1585. IEEE (2014)
5. Dave, V., Lygerakis, F., Rueckert, E.: Multimodal visual-tactile representation learning through self-supervised contrastive pre-training (2024), https://arxiv.org/abs/2401.12024
6. Dogar, M.R., Srinivasa, S.S.: Push-grasping with dexterous hands: Mechanics and a method. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2123–2130. IEEE (2010)
7. Duan, H., Wang, P., Huang, Y., Xu, G., Wei, W., Shen, X.: Robotics dexterous grasping: The methods based on point cloud and deep learning. Frontiers in Neurorobotics **15**, 658280 (2021)
8. Fang, H.S., Yan, H., Tang, Z., Fang, H., Wang, C., Lu, C.: Anydexgrasp: General dexterous grasping for different hands with human-level learning efficiency (2025), https://arxiv.org/abs/2502.16420
9. George, A., Gano, S., Katragadda, P., Farimani, A.B.: Vital pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies (2024), https://arxiv.org/abs/2403.11898
10. Guzey, I., Evans, B., Chintala, S., Pinto, L.: Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play (2023), https://arxiv.org/abs/2303.12076
11. Huang, B., Wang, Y., Yang, X., Luo, Y., Li, Y.: 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing (2025), https://arxiv.org/abs/2410.24091
12. Kerr, J., Huang, H., Wilcox, A., Hoque, R., Ichnowski, J., Calandra, R., Goldberg, K.: Self-supervised visuo-tactile pretraining to locate and follow garment features (2023), https://arxiv.org/abs/2209.13042
13. Lee, K.W., Qin, Y., Wang, X., Lim, S.C.: Dextouch: Learning to seek and manipulate objects with tactile dexterity. IEEE Robotics and Automation Letters **9**(12), 10772–10779 (Dec 2024). https://doi.org/10.1109/lra.2024.3478571, http://dx.doi.org/10.1109/LRA.2024.3478571
14. Lee, M.A., Zhu, Y., Srinivasan, K., Shah, P., Savarese, S., Fei-Fei, L., Garg, A., Bohg, J.: Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks (2019), https://arxiv.org/abs/1810.10191
15. Li, J., Wu, T., Zhang, J., Chen, Z., Jin, H., Wu, M., Shen, Y., Yang, Y., Dong, H.: Adaptive visuo-tactile fusion with predictive force attention for dexterous manipulation (2025), https://arxiv.org/abs/2505.13982
16. Li, M., Hang, K., Kragic, D., Billard, A.: Dexterous grasping under shape uncertainty. Robotics and Autonomous Systems **75**, 352–364 (2016)

17. Li, P., Wang, Z., Liu, M., Liu, H., Chen, C.: Clickdiff: Click to induce semantic contact map for controllable grasp generation with diffusion models. Proceedings of the 32nd ACM International Conference on Multimedia (2024), https://api.semanticscholar.org/CorpusID:271533543
18. Li, Y., Zhu, J.Y., Tedrake, R., Torralba, A.: Connecting touch and vision via cross-modal prediction (2019), https://arxiv.org/abs/1906.06322
19. Liu, Q., Ye, Q., Sun, Z., Cui, Y., Li, G., Chen, J.: Masked visual-tactile pre-training for robot manipulation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 13859–13875 (2024). https://doi.org/10.1109/ICRA57147.2024.10610933
20. Liu, Z., Chi, C., Cousineau, E., Kuppuswamy, N., Burchfiel, B., Song, S.: Maniwav: Learning robot manipulation from in-the-wild audio-visual data (2024), https://arxiv.org/abs/2406.19464
21. Lu, J., Kang, H., Li, H., Liu, B., Yang, Y., Huang, Q., Hua, G.: Ugg: Unified generative grasping (2023)
22. Luo, S., Bimbo, J., Dahiya, R., Liu, H.: Robotic tactile perception of object properties: A review. Mechatronics **48**, 54–67 (2017)
23. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., State, G.: Isaac gym: High performance gpu-based physics simulation for robot learning (2021), https://arxiv.org/abs/2108.10470
24. Mandikal, P., Grauman, K.: Learning dexterous grasping with object-centric visual affordances. In: 2021 IEEE international conference on robotics and automation (ICRA). pp. 6169–6176. IEEE (2021)
25. Mandikal, P., Grauman, K.: Dexvip: Learning dexterous grasping with human hand pose priors from video. In: Conference on Robot Learning. pp. 651–661. PMLR (2022)
26. Martinez-Gonzalez, P., Mulero-Perez, D., Oprea, S., Benavent-Lledo, M., Orts-Escolano, S., Garcia-Rodriguez, J.: Synthetic contact maps to predict grasp regions on objects. In: 2022 International Joint Conference on Neural Networks (IJCNN). pp. 1–6 (2022). https://doi.org/10.1109/IJCNN55064.2022.9892548
27. Nagabandi, A., Konolige, K., Levine, S., Kumar, V.: Deep dynamics models for learning dexterous manipulation. In: Conference on robot learning. pp. 1101–1112. PMLR (2020)
28. Okamura, A.M., Smaby, N., Cutkosky, M.R.: An overview of dexterous manipulation. In: Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065). vol. 1, pp. 255–262. IEEE (2000)
29. Pokhariya, C., Shah, I.N., Xing, A., Li, Z., Chen, K., Sharma, A., Sridhar, S.: Manus: Markerless grasp capture using articulated 3d gaussians (2024), https://arxiv.org/abs/2312.02137
30. Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., Levine, S.: Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. arXiv preprint arXiv:1709.10087 (2017)
31. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms (2017), https://arxiv.org/abs/1707.06347
32. Singh, R., Wyk, K.V., Abbeel, P., Malik, J., Ratliff, N., Handa, A.: End-to-end rl improves dexterous grasping policies (2025), https://arxiv.org/abs/2509.16434
33. Wan, W., Geng, H., Liu, Y., Shan, Z., Yang, Y., Yi, L., Wang, H.: Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware cur-

riculum and iterative generalist-specialist learning (2023), https://arxiv.org/abs/2304.00464
34. Wang, K., Lu, B., Cheng, Z., Zhang, H., Song, L.: D3grasp: Diverse and deformable dexterous grasping for general objects (2025), https://arxiv.org/abs/2509.19892
35. Wang, Y., Wu, R., Chen, Y., Wang, J., Liang, J., Zhu, Z., Geng, H., Malik, J., Abbeel, P., Dong, H.: Dexgarmentlab: Dexterous garment manipulation environment with generalizable policy (2025), https://arxiv.org/abs/2505.11032
36. Wang, Z., Chen, J., Chen, Z., Xie, P., Chen, R., Yi, L.: Genh2r: Learning generalizable human-to-robot handover via scalable simulation, demonstration, and imitation (2024), https://arxiv.org/abs/2401.00929
37. Wei, D., Xu, H.: A wearable robotic hand for hand-over-hand imitation learning (2023), https://arxiv.org/abs/2309.14860
38. Wei, Z., Xu, Z., Guo, J., Hou, Y., Gao, C., Cai, Z., Luo, J., Shao, L.: D (r, o) grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. arXiv preprint arXiv:2410.01702 (2024)
39. Weng, Z., Lu, H., Kragic, D., Lundell, J.: Dexdiffuser: Generating dexterous grasps with diffusion models (2024), https://arxiv.org/abs/2402.02989
40. Wu, T., Li, J., Zhang, J., Wu, M., Dong, H.: Canonical representation and force-based pretraining of 3d tactile for dexterous visuo-tactile policy learning (2025), https://arxiv.org/abs/2409.17549
41. Wu, Y.H., Wang, J., Wang, X.: Learning generalizable dexterous manipulation from human grasp affordance. In: Conference on Robot Learning. pp. 618–629. PMLR (2023)
42. Xu, W., Zhao, Y., Guo, W., Sheng, X.: Hierarchical reinforcement learning for articulated tool manipulation with multifingered hand (2025), https://arxiv.org/abs/2507.06822
43. Xu, Y., Wan, W., Zhang, J., Liu, H., Shan, Z., Shen, H., Wang, R., Geng, H., Weng, Y., Chen, J., Liu, T., Yi, L., Wang, H.: Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy (2023), https://arxiv.org/abs/2303.00938
44. Xue, H., Ren, J., Chen, W., Zhang, G., Fang, Y., Gu, G., Xu, H., Lu, C.: Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation (2025), https://arxiv.org/abs/2503.02881
45. Ye, J., Wang, K., Yuan, C., Yang, R., Li, Y., Zhu, J., Qin, Y., Zou, X., Wang, X.: Dex1b: Learning with 1b demonstrations for dexterous manipulation. In: Robotics: Science and Systems (RSS) (2025)
46. Yin, J., Qi, H., Malik, J., Pikul, J., Yim, M., Hellebrekers, T.: Learning in-hand translation using tactile skin with shear and normal force sensing (2025), https://arxiv.org/abs/2407.07885
47. Yin, Z.H., Huang, B., Qin, Y., Chen, Q., Wang, X.: Rotating without seeing: Towards in-hand dexterity through touch (2023), https://arxiv.org/abs/2303.10880
48. Zhang, H., Lyu, J., Zhou, C., Liang, H., Tu, Y., Sun, F., Zhang, J.: Adg-net: A sim2real multimodal learning framework for adaptive dexterous grasping. IEEE Transactions on Cybernetics **55**(2), 840–853 (2025). https://doi.org/10.1109/TCYB.2024.3518975
49. Zhang, H., Wu, Z., Huang, L., Christen, S., Song, J.: Robustdexgrasp: Robust dexterous grasping of general objects (2025), https://arxiv.org/abs/2504.05287

50. Zhang, J., Liu, H., Li, D., Yu, X., Geng, H., Ding, Y., Chen, J., Wang, H.: Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes (2024), https://arxiv.org/abs/2410.23004
51. Zhang, L., Bai, K., Huang, G., Bing, Z., Chen, Z., Knoll, A., Zhang, J.: Contactdexnet: Multi-fingered robotic hand grasping in cluttered environments through hand-object contact semantic mapping (2025), https://arxiv.org/abs/2404.08844
52. Zhao, F., Tsetserukou, D., Liu, Q.: Graingrasp: Dexterous grasp generation with fine-grained contact guidance (2024), https://arxiv.org/abs/2405.09310
53. Zhong, Y., Huang, X., Li, R., Zhang, C., Chen, Z., Guan, T., Zeng, F., Lui, K.N., Ye, Y., Liang, Y., Yang, Y., Chen, Y.: Dexgraspvla: A vision-language-action framework towards general dexterous grasping (2025), https://arxiv.org/abs/2502.20900
54. Zhong, Y., Jiang, Q., Yu, J., Ma, Y.: Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness (2025), https://arxiv.org/abs/2503.08257

Supplementary Material

To further demonstrate the robustness and generalization capabilities of our method, this supplementary material provides extensive qualitative results, visualizing both grasp pose generation (Sec. A) and the adaptive adjustment process (Sec. F) following initial failures across a wider range of objects in the simulation and real-world settings. Note that the real-world results presented here utilize the robot's right arm, in contrast to the left arm used in the main text, which validates that our method is agnostic to the manipulator's specific mounting position and kinematic configuration (left vs. right), further confirming its capability to achieve stable grasping independent of the hardware setup.

A Visualizations of Grasp Pose Generation

We present more visualizations of the generated grasping poses by our method on various challenging objects in Figure 8 and Figure 9. The Contact-Driven Grasp Pose Generator is able to produce reasonable and stable grasping poses for complex and irregular objects such as the toy figure (1st col in Figure 8), the tape measure (3rd col in Figure 8), and the brush (2nd row, 3rd col in Figure 9).

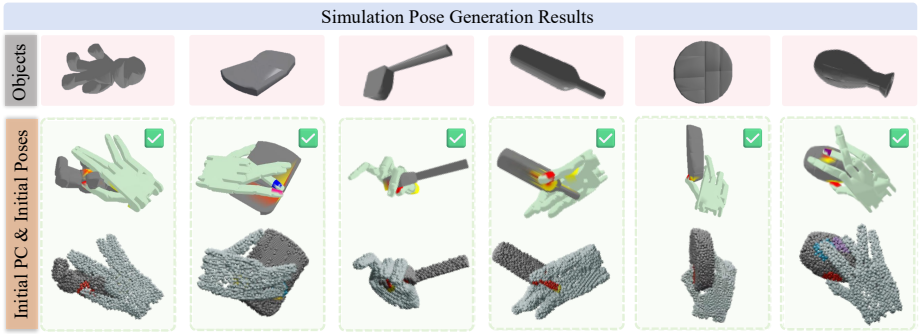

Fig. 8: Visualization of successful grasp pose generation in simulation experiments.

B Training and Inference Details

We train four models in our pipeline—the CMap generator, pose generator, classifier, and adapter—all on a single NVIDIA RTX 4090 GPU (24 GB), each fully utilizing the available memory. The CMap and pose generators both use a batch size of 32 and a learning rate of 1×10^{-3}, taking about 3 hours each; the classifier uses a batch size of 128 and a learning rate of 3×10^{-4} (around 10 hours); and the adapter uses a batch size of 256 and a learning rate of 1×10^{-4}

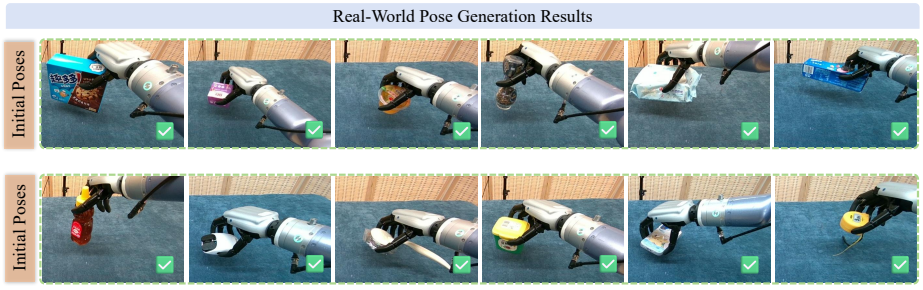

Fig. 9: Visualization of successful grasp pose generation in real-world experiments.

(around three days). These settings collectively ensure stable convergence across all components.

As for inference, the entire pipeline—including all four components—occupies approximately 16 GB of GPU memory when running inside IsaacGym, and only about 2 GB when executed on the real robot without the simulation environment.

In addition, regarding the hyper-parameter settings, during closed-loop refinement, we use a success threshold of $\tau_{succ} = 0.5$ and a maximum iteration count of $K = 5$. In practical evaluations, we observe that grasp refinement is triggered in roughly 30% of the trials. Among those cases, over 90% converge to a successful grasp within fewer than 3 refinement iterations, demonstrating both the efficiency and effectiveness of the adaptive optimization process.

C Real-Time Feasibility

Due to the speed optimization of the Diffusion Policy backbone, our system frequency is 20–25 Hz, feasible for real-time inference. Thus, if the object moves, our model can timely refine the grasp pose.

D Additional Quantitative Analysis

Effect of the intermediate contact map. Our generator first predicts a contact map $\mathcal{M}$ and then conditions the grasp pose on it. To assess the value of this intermediate representation, we train a variant that predicts the grasp pose directly from the object point cloud, without the contact map. This lowers the success rate to 82%/78%/77% on the seen, unseen-object, and unseen-category splits, compared with 91%/82%/83% for our full model. The consistent drop confirms that $\mathcal{M}$ provides a useful semantic prior that guides the generator toward feasible hand–object contacts.

Grasp-proximity metric for adaptation pairs. When constructing the failed–successful state pairs used to train the adaptation model, we match each failed state to its nearest successful counterpart under a combined $SE(3)$-and-joint distance. We ablate this design by matching with only the $SE(3)$ term or

only the joint-angle term. Both variants degrade performance, to 82%/75%/61% and 84%/72%/57% respectively, compared with 91%/82%/83% for the combined metric. Using either term alone yields degenerate pairs—similar root poses with entirely different finger configurations, or similar finger postures under inconsistent global placement—whereas the combined metric captures both where the hand is and how the fingers are configured, producing physically meaningful pairs and stable corrective targets.

Tactile representation choice. Here, the *dense modality* replaces our six-dimensional tactile-intensity vector with a per-contact dense force-field vector, while the *small-data* setting trains on 50% of the full collected dataset under an identical framework. Under the small-data setting, our representation reaches 88% accuracy versus 81% for the dense modality, and in the real world it reaches 85% versus 72%. Richer modalities demand substantially more data and widen the sim-to-real gap of tactile signals, making it harder to align simulation and reality; we therefore adopt the simpler representation for its efficiency and tighter simulation–real alignment.

E Real-World Evaluation Protocol

We evaluate 26 real-world objects, split into 10/9/7 seen, unseen-object, and unseen-category groups, with 10 randomized trials per object, yielding 260 trials per method. The real-world trends mirror the simulation results: because the baselines are either open-loop or lack tactile feedback, they cannot detect or correct unstable contact once the initial pose is executed, whereas our tactile-driven closed-loop adaptation continuously monitors contact and recovers from such errors. Residual failures stem mainly from slippage during lifting, insufficient enclosure for objects near the hand's size limit, and unstable contact on small support areas.

F Visualizations of Grasp Pose Adaptation

We provide additional visualizations for Grasp Pose Adaptation in Figure 10 and Figure 11. As observed, the initial hand poses may fail to yield a stable grasp. Common failure reasons include insufficient contact points (e.g., 3rd col in Figure 10; 3rd row, 6th col in Figure 11), limited contact surface area (e.g., 5th col in Figure 10; 3rd row, 3rd col in Figure 11), and geometric mismatch between the contact points and the object geometries (e.g., 6th col in Figure 10; 1st row, 6th col, and 3rd row, 4th col in Figure 11). However, our classifier can accurately identify these potential failure cases, while the adaptation module can correct them to a reasonable one for a successful grasp.

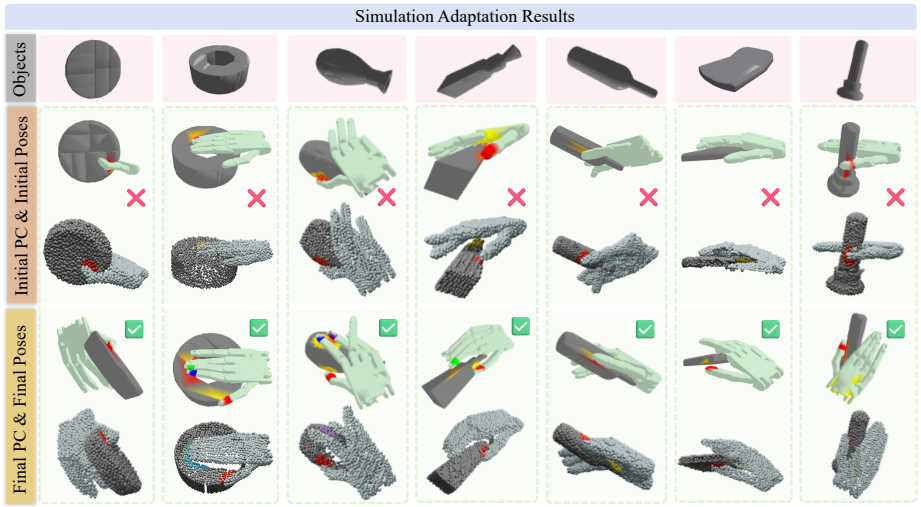

Fig. 10: Visualization of grasp pose adaptation in simulation experiments.

Fig. 11: Visualization of grasp pose adaptation in real-world experiments.