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], 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
[*]Equal contribution [†]Corresponding author

***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.**

I. Introduction

Humans grasp diverse objects by planning poses from vision and adjusting them through tactile feedback after contact. Transferring this visuo-tactile capability to dexterous robotic hands remains important yet challenging.

Most existing grasping methods [4, 34, 1, 24, 36, 20, 29, 23] rely solely on visual inputs such as RGB or point clouds, predicting hand poses from initial observations without post-contact refinement. Tactile-only strategies [7, 6] adapt under occlusion but require multiple trials and discard available visual cues. Recent visuo-tactile works [32, 21, 33] improve generalization, but mainly fuse modality features without modeling their spatial alignment and interaction, which are crucial for adaptive grasp control.

We propose a framework integrating tactile and geometric information for grasp generation, feasibility prediction, and adaptive refinement. Inspired by how humans assign hand parts to object regions, we introduce an expressive semantic contact representation. Unlike contact maps without semantics [15, 10, 16], ours associates each contact with a finger or palm identity (Fig. 1, top), encoding both contact occurrence and expected hand–object associations. Fusing this map into the object point cloud enables pose decoding with improved initial grasp success.

Initial poses may nevertheless fail on complex or unseen objects. During execution, we therefore use tactile feedback to assess and refine the grasp (Fig. 1, bottom). Tactile contact markers are integrated into the hand–object point cloud to construct a tactile-mapped representation. A classifier predicts grasp success, while an adaptation model adjusts suboptimal poses. We build a tactile-equipped dexterous-hand environment in IsaacGym [12] for training and evaluation, and deploy the method on a real robot. Our contributions are:

- A visuo-tactile representation that fuses geometric and tactile information through contact-aware encoding for fine-grained finger–object interaction reasoning.
- A unified framework for grasp generation, feasibility prediction, and tactile-guided adaptive refinement, enabling robust and efficient dexterous grasping.
- Extensive simulation and real-world experiments demonstrating the framework's effectiveness.

II. Related Work

Dexterous grasping. Dexterous hands have received extensive attention for human-like manipulation. Existing methods mainly follow three paradigms: reinforcement learning [19, 27, 17], imitation learning [22, 26], and object-centric policies [14, 13, 20] that map proprioception and RGB-D observations to grasps. Most rely on vision alone and cannot detect tactile instabilities such as slipping [11], causing failures even when the visual geometry appears feasible. Recent studies [32, 21, 33] explore visuo-tactile integration, but mainly encode and directly fuse the modalities without modeling their spatial and semantic relationships.

Visuo-tactile fusion for manipulation. Vision and touch are complementary for robust manipulation [8, 9]: vision provides geometric priors, while touch captures contact quality. Existing approaches fall into two categories. Learning-based methods [25, 3, 5, 2] encode each modality into latent features and combine them through concatenation or pretraining, emphasizing global alignment over spatial correspondence. Raw-data-based methods [31, 30] directly fuse signals but likewise overlook geometric and contact-level relationships. In contrast, we integrate tactile signals into the object point cloud to

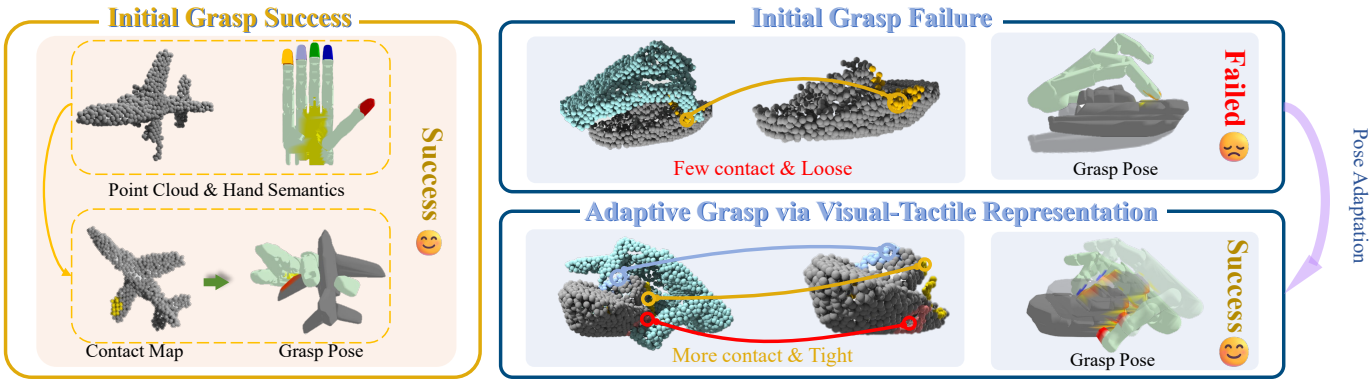

Fig. 1: **Overview. (Left)** An initial grasping module predicts a contact map from the raw point cloud with finger/palm identities, generating physically plausible hand poses informed by hand semantics. **(Right-Top)** The initial pose does not guarantee success and may fail due to sparse or unstable contact. **(Right-Bottom)** Our visuo-tactile-fusion representation adapts the initial pose into more stable contact patterns for robust execution.

construct tactile-mapped geometric representations for contact-aware generation and adaptation.

III. Method

A. Problem Formulation

The target object $\mathcal{O}$ is represented by a point cloud $\mathcal{P}_{obj} = \{(x_i, y_i, z_i, r_i, g_i, b_i)\}_{i=1}^{N_0}$ from an RGB-D camera. The hand state is $s = [p, q, \mathbf{j}]$, where $p \in \mathbb{R}^3$ and $q \in \mathbb{R}^4$ are palm position and orientation and $\mathbf{j} \in \mathbb{R}^{22}$ the joint angles. Tactile sensors on the fingertips and palm provide readings $\tau = \{\tau_i\}_{i=1}^{N_t}$ ($N_t = 6$: five fingertips and one palm). We seek a configuration s^* maximizing grasp success given 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 success.

B. Overview

Our framework (Fig. 2) integrates object geometry and tactile feedback through three components: a **Contact-Driven Grasp Pose Generator** (Sec. III-C) producing the initial pose; a **Grasp Pose Classifier** (Sec. III-D) predicting feasibility; and a **Grasp Pose Adaptation Model** (Sec. III-E) refining suboptimal poses. These form a closed loop (Sec. III-F); Sec. III-G describes data collection. Each module uses a PointNet++ backbone; the classifier adopts a dual-encoder design fused by multi-head attention, and the adaptation module is a diffusion model conditioned on the visuo–tactile features.

C. Contact-Driven Grasp Pose Generator

Given the object point cloud, a model h_{ψ_1} predicts a probabilistic contact map $\mathcal{M}$ in which each point is labeled with a hand part:

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

The generator h'_{ψ_2} then conditions on $(\mathcal{P}_{obj}, \mathcal{M})$ to predict an initial grasp state $s^{(0)} = h'_{\psi_2}(\mathcal{P}_{obj}, \mathcal{M})$. The contact map is supervised with cross-entropy $\mathcal{L}_{CE} = -\sum_i \mathcal{M}_{s,i} \log \mathcal{M}_i$ and the pose with $\mathcal{L}_{MSE} = \|s^{(0)} - s^s\|_2^2$.

D. Grasp Pose Classifier

After executing $s^{(0)}$, we annotate the object point cloud with tactile signals from the five fingertips and palm, assigning each contact region the ID of the contacting hand part to form a tactile-mapped point cloud $\mathcal{P}_{vt}$. A PointNet-based classifier f_θ estimates the success probability $\hat{y} = f_\theta(\mathcal{P}_{vt}, s) \in [0, 1]$, trained with binary cross-entropy

$$\mathcal{L}_{cls} = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})]. \tag{3}$$

By leveraging both geometric structure and tactile contact patterns, f_θ provides a differentiable, real-time estimate of grasp stability.

E. Grasp Pose Adaptation Model

To recover from failures, an adaptation model g_ϕ maps unstable grasps toward stable ones. Using paired states $(\mathcal{P}_{vt}^f, s^f, y{=}0)$ and $(\mathcal{P}_{vt}^s, s^s, y{=}1)$ where s^s is close to s^f, it predicts a correction $\Delta s = g_\phi(\mathcal{P}_{vt}^f, s^f)$, giving the refined grasp $s^{f,\text{new}} = s^f + \Delta s$. The objective drives the corrected pose toward the successful one while maximizing predicted stability:

$$\mathcal{L}_{gen} = \|\Delta s - (s^s - s^f)\|_2^2 + \lambda\,(1 - f_\theta(\mathcal{P}_{vt}^f, s^f + \Delta s)). \tag{4}$$

F. Closed-Loop Grasp Optimization

The three components form a closed loop. Starting from $s^{(0)}$, the system iteratively evaluates and updates the grasp:

$$\hat{y}_k = f_\theta(\mathcal{P}_{vt}^k, s^{(k)}), \tag{5}$$

$$s_{k+1} = \begin{cases} s_k, & \text{if } \hat{y}_k > \tau_{succ}, \\ s_k + g_\phi(\mathcal{P}_k, s_k), & \text{otherwise}, \end{cases} \tag{6}$$

stopping when $\hat{y}_k > \tau_{succ}$ or $k = K$. The result s^* satisfies both geometric feasibility and tactile stability.

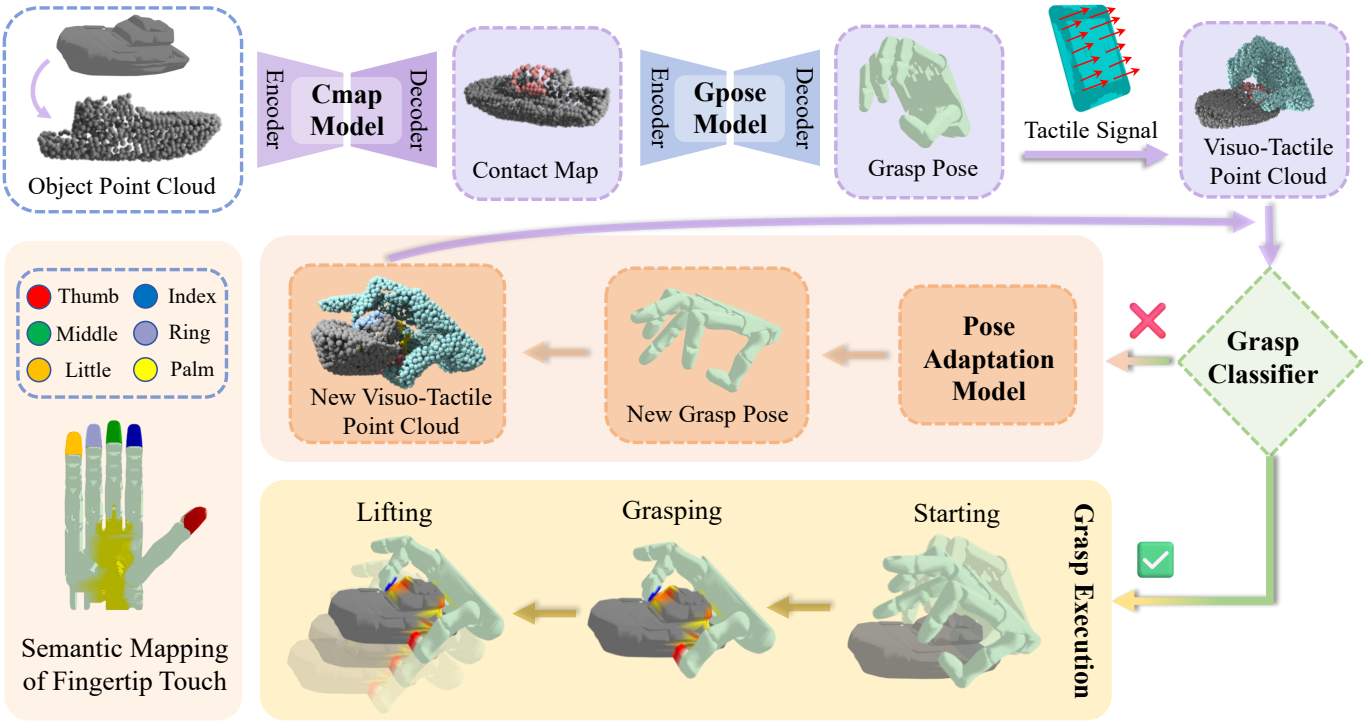

Fig. 2: **Pipeline of the framework.** From an object point cloud and semantic fingertip labels, the system predicts a contact map and an initial grasp pose, fuses hand configuration and tactile feedback into a visuo–tactile point cloud, classifies stability, and—if failure is predicted—iteratively refines the grasp before execution.

G. Tactile-Enhanced Data Collection

We collect data via PPO [18] following the first-stage setup of UniDexGrasp [28], recording both successful and failed attempts for diverse tactile data. A finger is marked active when its contact force exceeds 0.01N, and nearby object points receive the corresponding contact label, yielding a tactile-mapped point cloud

$$\mathcal{P}_{vt} = \{(x_i, y_i, z_i, r_i, g_i, b_i, k_i, \mathbf{c}_i) \mid i = 1, \ldots, N\}, \quad (7)$$

where k_i encodes the contact state and $\mathbf{c}_i \in [0,1]^6$ is the tactile-intensity vector over the five fingertips and palm. Each sample is $(\mathcal{P}_{vt}, s, y)$.

IV. Experiments

We evaluate the framework in both simulation (Sec. IV-A–IV-C) and the real world (Sec. IV-D) using a tactile-enabled dexterous hand.

A. Simulation Experimental Setup

Our simulation is built on IsaacGym [12] for large-scale parallel collection of grasp trajectories and tactile feedback; the task consists of reaching, grasping, and lifting. The hand is a ShadowHand with five fingers and 22 DoF (28 DoF including the 6-DoF root). Objects come from UniDexGrasp++ [20]: 50 objects over 6 categories, 20 for training and 30 unseen-category objects for testing. A grasp succeeds if the object stays stable for at least 1 s after lift-off. We report grasp success rate on three sets: *Seen Objects*, *Unseen Objects* (same categories), and *Unseen Categories*.

B. Baselines

We compare against eight state-of-the-art frameworks under identical settings: **DexGraspAnything** [37] and **DexDiffuser** [24] (visual grasp generators), **UniDexGrasp++** [20] (diffusion-based, no tactile), **UniDexGrasp(PPO)** [28] (PPO with our visuo-tactile inputs), **Robot Synesthesia** (tactile distillation), an **Intuitive Closed-Loop** heuristic, **ContactDexNet** [35] (contact-map-based), and **DexGraspVLA** [36] (RGB VLA).

TABLE I: 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%
Ours	**91%**	**82%**	**83%**

From Table I, our method consistently outperforms all baselines. We identify four failure modes: (1) generative methods (Rows 1–3) rely on visual geometry without semantic correspondence, giving infeasible configurations for complex

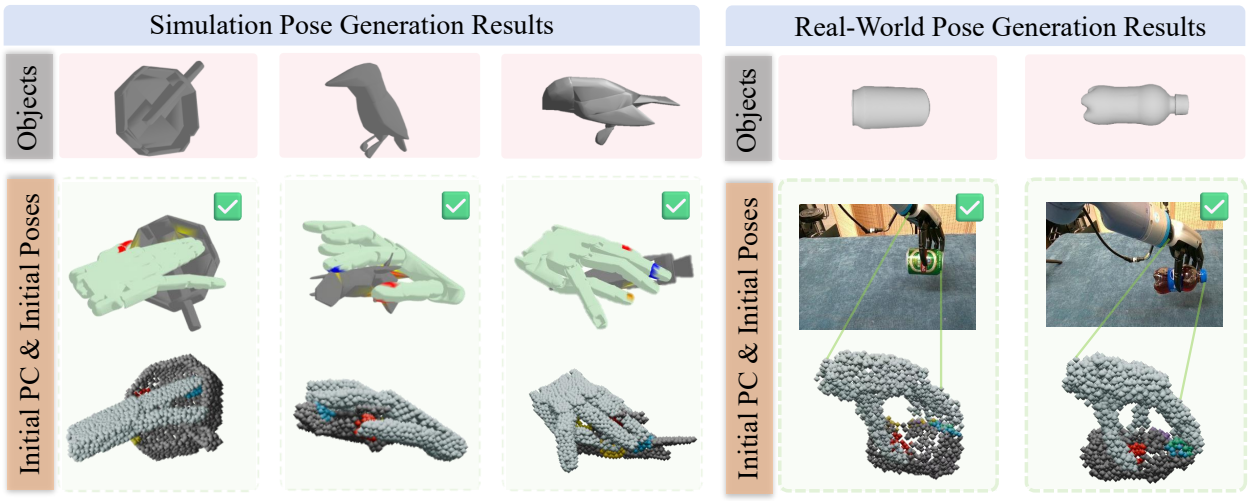

Fig. 3: **Qualitative results.** For objects in both simulation and the real world, the contact-driven generator and adaptation module produce stable grasps with geometrically consistent, contact-rich hand poses.

objects; (2) the end-to-end RGB-to-action model (Row 8) lacks geometric understanding; (3) offline contact modeling (Row 7) yields only static predictions and cannot correct in closed loop; and (4) prior visuo-tactile refinement is unstable, as RL (Row 4) suffers from exploration noise, distillation (Row 5) is brittle, and the heuristic tightens fingers without geometric awareness. By encoding finger/palm identities into generation and applying tactile-guided refinement, our method avoids these, with the gap widening on unseen categories and underscoring the value of hand–object semantics and closed-loop tactile feedback.

C. Ablations

We ablate the tactile encoding and adaptation: **w/o contact ID in generation** (all fingertips share a unified ID during generation, removing semantic information); **w/o contact ID in adaptation** (unified ID in the classifier and adaptation models); **w/o adaptation** (objects are lifted directly after the initial pose); **Object-only** (only the object point cloud with tactile mapping); **Full Point Cloud w/o tactile** (hand–object point cloud without tactile features); and **w/o PC** (images labeled with tactile signals instead of point clouds).

TABLE II: 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%**

As shown in Table II (Rows 1, 2, 7), the semantic contact ID is essential for both generation and refinement. Comparing Rows 3 and 7 highlights the adaptation module, with gains on unseen objects ($78\% \rightarrow 82\%$) and unseen categories ($59\% \rightarrow 83\%$). Rows 4–7 show that removing the hand point cloud, removing tactile signals, or replacing point clouds with images all hurt. Figure 3 visualizes these effects: guided by contact IDs, the generator produces geometrically consistent initial poses, and the adaptation module iteratively turns unstable grasps into secure ones.

D. Real-World Experiment

TABLE III: 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%**

We deploy on a **Psibot** robotic hand with high-resolution tactile sensors, transferring the simulation-trained model without per-object tuning by mapping live tactile readings into the same visuo-tactile representation. We evaluate on more than 20 objects of distinct geometry and material, counting a grasp successful if held stably for at least 3 s after lifting. As reported in Table III, our method transfers robustly and outperforms the baselines; more visualizations are in the supplementary material.

V. Conclusion

We presented a unified visuo-tactile-fusion grasping framework that couples tactile feedback with visual geometry through a contact-aware representation binding tactile signals to finger identities and a tactile-guided adaptation module for real-time refinement. It improves both initial feasibility and adaptive stability in simulation and the real world, offering a scalable route for integrating vision and touch in dexterous manipulation.

References

[1] Sirui Chen, Jeannette Bohg, and C. Karen Liu. Springgrasp: Synthesizing compliant, dexterous grasps under shape uncertainty, 2024. URL https://arxiv.org/abs/2404.13532.

[2] Yizhou Chen, Andrea Sipos, Mark Van der Merwe, and Nima Fazeli. Visuo-tactile transformers for manipulation, 2022. URL https://arxiv.org/abs/2210.00121.

[3] Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training, 2024. URL https://arxiv.org/abs/2401.12024.

[4] Hao-Shu Fang, Hengxu Yan, Zhenyu Tang, Hongjie Fang, Chenxi Wang, and Cewu Lu. Anydexgrasp: General dexterous grasping for different hands with human-level learning efficiency, 2025. URL https://arxiv.org/abs/2502.16420.

[5] Abraham George, Selam Gano, Pranav Katragadda, and Amir Barati Farimani. Vital pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies, 2024. URL https://arxiv.org/abs/2403.11898.

[6] Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play, 2023. URL https://arxiv.org/abs/2303.12076.

[7] Kang-Won Lee, Yuzhe Qin, Xiaolong Wang, and Soo-Chul Lim. Dextouch: Learning to seek and manipulate objects with tactile dexterity. *IEEE Robotics and Automation Letters*, 9(12):10772–10779, December 2024. ISSN 2377-3774. doi: 10.1109/lra.2024.3478571. URL http://dx.doi.org/10.1109/LRA.2024.3478571.

[8] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks, 2019. URL https://arxiv.org/abs/1810.10191.

[9] Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Benjamin Burchfiel, and Shuran Song. Maniwav: Learning robot manipulation from in-the-wild audio-visual data, 2024. URL https://arxiv.org/abs/2406.19464.

[10] Jiaxin Lu, Hao Kang, Haoxiang Li, Bo Liu, Yiding Yang, Qixing Huang, and Gang Hua. Ugg: Unified generative grasping, 2023.

[11] Shan Luo, Joao Bimbo, Ravinder Dahiya, and Hongbin Liu. Robotic tactile perception of object properties: A review. *Mechatronics*, 48:54–67, 2017.

[12] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021. URL https://arxiv.org/abs/2108.10470.

[13] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 6169–6176. IEEE, 2021.

[14] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022.

[15] Pablo Martinez-Gonzalez, David Mulero-Perez, Sergiu Oprea, Manuel Benavent-Lledo, Sergio Orts-Escolano, and Jose Garcia-Rodriguez. Synthetic contact maps to predict grasp regions on objects. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2022. doi: 10.1109/IJCNN55064.2022.9892548.

[16] Chandradeep Pokhariya, Ishaan N Shah, Angela Xing, Zekun Li, Kefan Chen, Avinash Sharma, and Srinath Sridhar. Manus: Markerless grasp capture using articulated 3d gaussians, 2024. URL https://arxiv.org/abs/2312.02137.

[17] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

[18] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

[19] Ritvik Singh, Karl Van Wyk, Pieter Abbeel, Jitendra Malik, Nathan Ratliff, and Ankur Handa. End-to-end rl improves dexterous grasping policies, 2025. URL https://arxiv.org/abs/2509.16434.

[20] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning, 2023. URL https://arxiv.org/abs/2304.00464.

[21] Keyu Wang, Bingcong Lu, Zhengxue Cheng, Hengdi Zhang, and Li Song. D3grasp: Diverse and deformable dexterous grasping for general objects, 2025. URL https://arxiv.org/abs/2509.19892.

[22] Zifan Wang, Junyu Chen, Ziqing Chen, Pengwei Xie, Rui Chen, and Li Yi. Genh2r: Learning generalizable human-to-robot handover via scalable simulation, demonstration, and imitation, 2024. URL https://arxiv.org/abs/2401.00929.

[23] Zhenyu Wei, Zhixuan Xu, Jingxiang Guo, Yiwen Hou, Chongkai Gao, Zhehao Cai, Jiayu Luo, and Lin Shao. D (r, o) grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. *arXiv preprint arXiv:2410.01702*, 2024.

[24] Zehang Weng, Haofei Lu, Danica Kragic, and Jens Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models, 2024. URL https://arxiv.org/abs/2402.02989.

[25] Tianhao Wu, Jinzhou Li, Jiyao Zhang, Mingdong Wu, and Hao Dong. Canonical representation and force-based pretraining of 3d tactile for dexterous visuo-tactile policy

learning, 2025. URL https://arxiv.org/abs/2409.17549.

[26] Yueh-Hua Wu, Jiashun Wang, and Xiaolong Wang. Learning generalizable dexterous manipulation from human grasp affordance. In *Conference on Robot Learning*, pages 618–629. PMLR, 2023.

[27] Wei Xu, Yanchao Zhao, Weichao Guo, and Xinjun Sheng. Hierarchical reinforcement learning for articulated tool manipulation with multifingered hand, 2025. URL https://arxiv.org/abs/2507.06822.

[28] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, Tengyu Liu, Li Yi, and He Wang. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy, 2023. URL https://arxiv.org/abs/2303.00938.

[29] Jianglong Ye, Keyi Wang, Chengjing Yuan, Ruihan Yang, Yiquan Li, Jiyue Zhu, Yuzhe Qin, Xueyan Zou, and Xiaolong Wang. Dex1b: Learning with 1b demonstrations for dexterous manipulation. In *Robotics: Science and Systems (RSS)*, 2025.

[30] Jessica Yin, Haozhi Qi, Jitendra Malik, James Pikul, Mark Yim, and Tess Hellebrekers. Learning in-hand translation using tactile skin with shear and normal force sensing, 2025. URL https://arxiv.org/abs/2407.07885.

[31] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch, 2023. URL https://arxiv.org/abs/2303.10880.

[32] Hui Zhang, Jianzhi Lyu, Chuangchuang Zhou, Hongzhuo Liang, Yuyang Tu, Fuchun Sun, and Jianwei Zhang. Adg-net: A sim2real multimodal learning framework for adaptive dexterous grasping. *IEEE Transactions on Cybernetics*, 55(2):840–853, 2025. doi: 10.1109/TCYB.2024.3518975.

[33] Hui Zhang, Zijian Wu, Linyi Huang, Sammy Christen, and Jie Song. Robustdexgrasp: Robust dexterous grasping of general objects, 2025. URL https://arxiv.org/abs/2504.05287.

[34] Jialiang Zhang, Haoran Liu, Danshi Li, Xinqiang Yu, Haoran Geng, Yufei Ding, Jiayi Chen, and He Wang. Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes, 2024. URL https://arxiv.org/abs/2410.23004.

[35] Lei Zhang, Kaixin Bai, Guowen Huang, Zhenshan Bing, Zhaopeng Chen, Alois Knoll, and Jianwei Zhang. Contactdexnet: Multi-fingered robotic hand grasping in cluttered environments through hand-object contact semantic mapping, 2025. URL https://arxiv.org/abs/2404.08844.

[36] Yifan Zhong, Xuchuan Huang, Ruochong Li, Ceyao Zhang, Zhang Chen, Tianrui Guan, Fanlian Zeng, Ka Num Lui, Yuyao Ye, Yitao Liang, Yaodong Yang, and Yuanpei Chen. Dexgraspvla: A vision-language-action framework towards general dexterous grasping, 2025. URL https://arxiv.org/abs/2502.20900.

[37] Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness, 2025. URL https://arxiv.org/abs/2503.08257.

APPENDIX A
NETWORK ARCHITECTURE

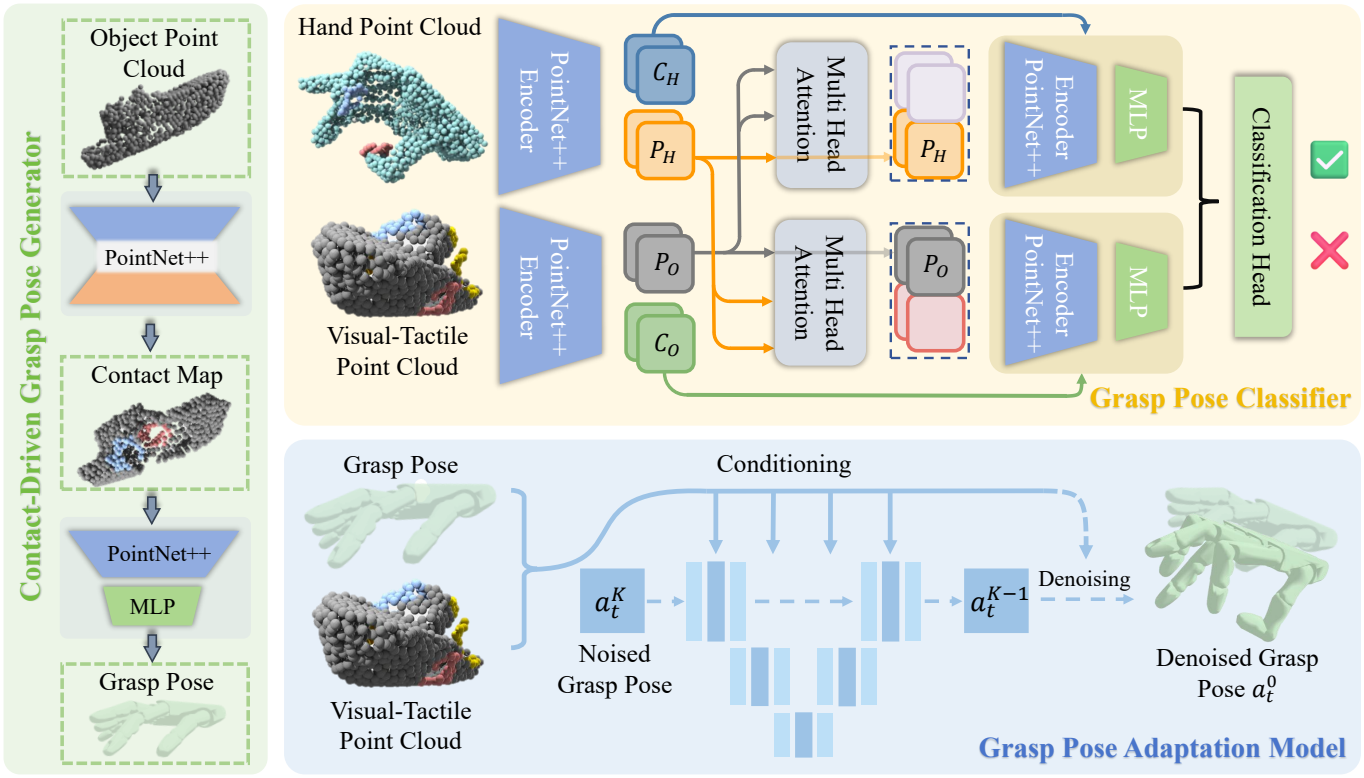

Fig. 4: **Architecture of the grasping framework. Green:** the object point cloud is processed by PointNet++ to predict a visuo–tactile contact map and an initial grasp pose. **Yellow:** a dual-encoder PointNet++ encodes the hand and visuo–tactile point clouds, followed by multi-head attention for grasp success prediction. **Blue:** if failure is predicted, a diffusion-based module iteratively refines the grasp pose conditioned on visuo–tactile features.

Fig. 4 details the architecture summarized in the main text: a PointNet++ backbone predicts the contact map and initial pose, a dual-encoder branch fused by multi-head attention predicts grasp success, and a diffusion module refines the pose on predicted failure.

APPENDIX B
ADDITIONAL EXPERIMENTAL DETAILS

A. Simulation Setup

The hand is based on ShadowHand: five fingers with 22 DoF plus a 6-DoF root pose (28 DoF total), moved in simulation by directly setting the root pose. Objects are from UniDexGrasp++ [20]: 50 objects across 6 categories, 20 for training and 30 unseen-category objects for testing. A grasp succeeds if the object stays stable without slippage for at least 1 s after lift-off. We report grasp success rate on *Seen Objects*, *Unseen Objects* (same categories, unseen instances), and *Unseen Categories*.

B. Baselines

The eight baselines in Table I are: **DexGraspAnything** [37] and **DexDiffuser** [24] (visual grasp generators), **UniDexGrasp++** [20] (diffusion without tactile feedback), **UniDexGrasp(PPO)** [28] (PPO on our visuo-tactile inputs), **Robot Synesthesia** (tactile distillation), **Intuitive Closed-Loop** (a "move closer on no contact" heuristic), **ContactDexNet** [35] (contact-map grasps), and **DexGraspVLA** [36] (an RGB VLA model).

C. Ablation Settings

The ablations in Table II remove semantic contact IDs in generation or in adaptation; remove the adaptation step; use the object point cloud only; drop all tactile features from the hand–object point cloud; and replace point clouds with tactile-labeled images.

D. Real-World Setup

Real-world experiments use a **Psibot** hand with high-resolution tactile sensors (Fig. 5) on more than 20 objects of distinct geometry and material. A grasp succeeds if the object is held stably for at least three seconds after lifting. This setting tests direct sim-to-real transfer of the visuo–tactile representation without any per-object tuning, exercising both the contact-driven generator and the tactile-guided adaptation module on real hardware whose sensor noise and contact characteristics differ markedly from those in simulation.

E. Representation Choice

We explored richer tactile modalities (force directions, shear forces) but found they need much more data, whereas our tactile-mapped point cloud already handles most scenarios (88% vs. 81% accuracy in small-data tests) and transfers better (85% vs. 72% in the real world). We therefore adopt the simpler representation for its efficiency and simulation–real alignment.

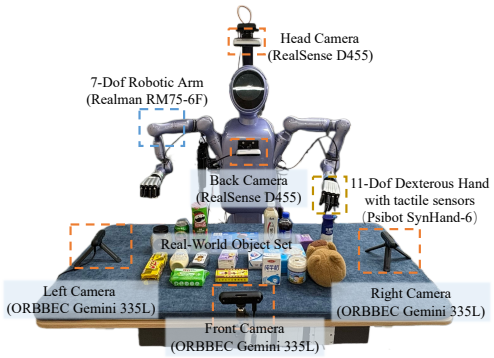

Fig. 5: **Real-world setup.** A 7-DoF arm carries an 11-DoF Psibot dexterous hand with fingertip/palm tactile sensors, observed by head, side, and front RGB-D cameras.

F. Analysis of Tactile Signals

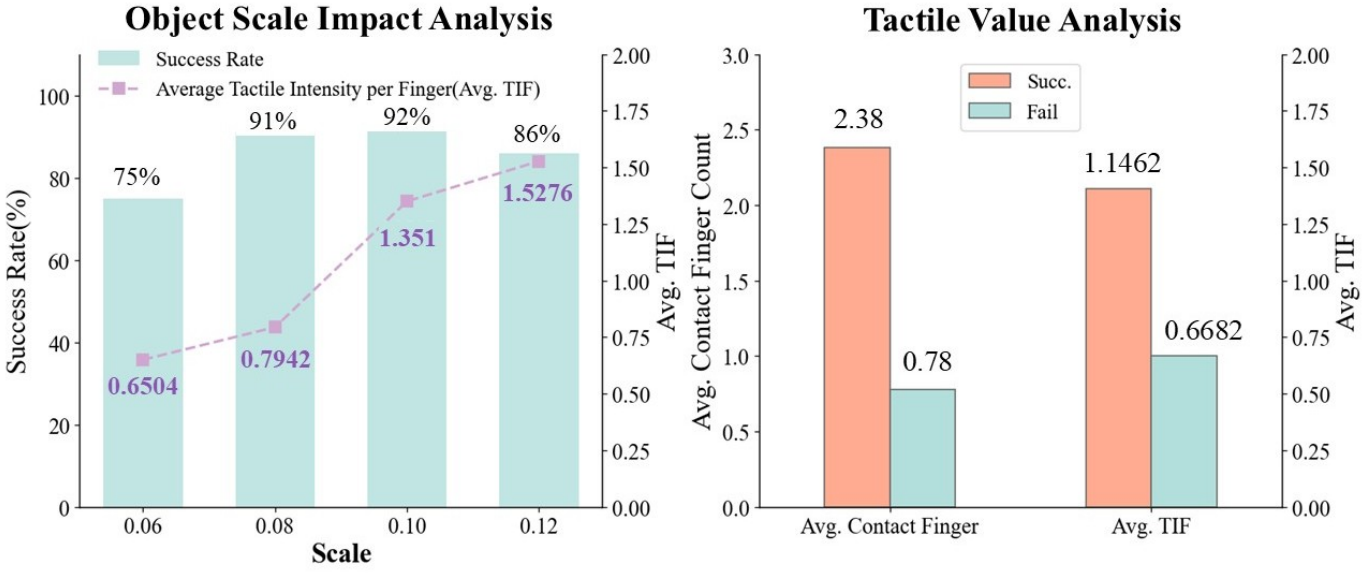

Fig. 6: **(Left)** As object size increases, average tactile intensity (Avg. TIF) rises while success rate first increases then slightly drops. **(Right)** Avg. TIF and contact-finger count are higher for successful than for failed grasps.

Tactile signals are normalized to $[0, 1]$. As Fig. 6 shows, larger objects yield higher average intensity, and successful grasps show higher intensity and more contact fingers than failed ones, confirming the learned contact semantics track physical stability.

APPENDIX C
ADDITIONAL QUALITATIVE RESULTS: GENERATION

The contact-driven generator produces stable poses for complex, irregular objects in simulation (Fig. 7) and the real world (Fig. 8). Real-world results use the robot's right arm, versus the left in the main text, showing the method is agnostic to mounting position and kinematic configuration.

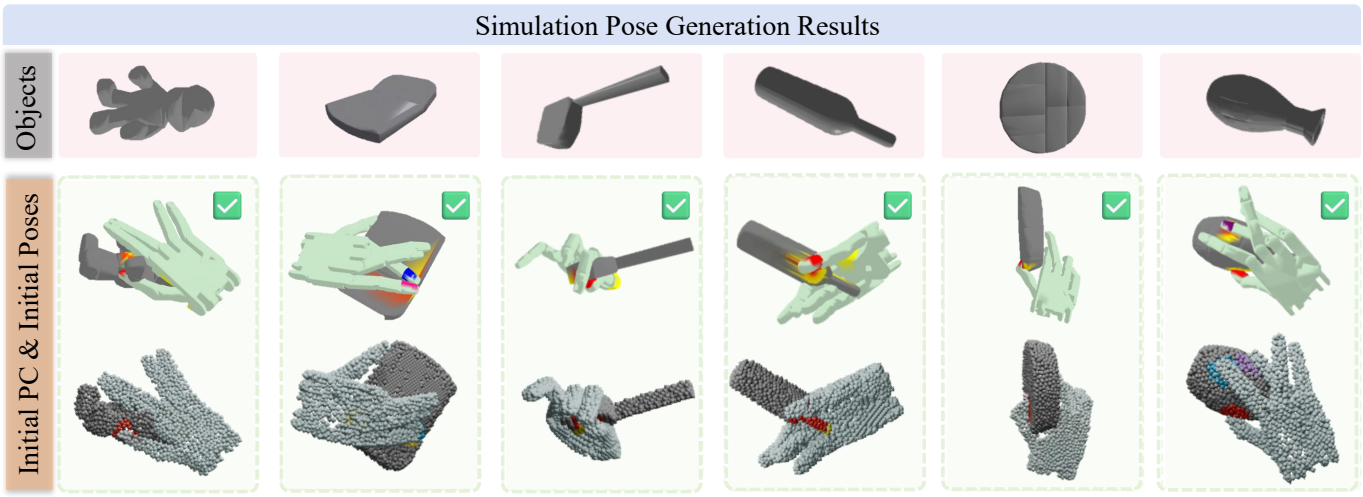

Fig. 7: Successful grasp pose generation on diverse simulation objects.

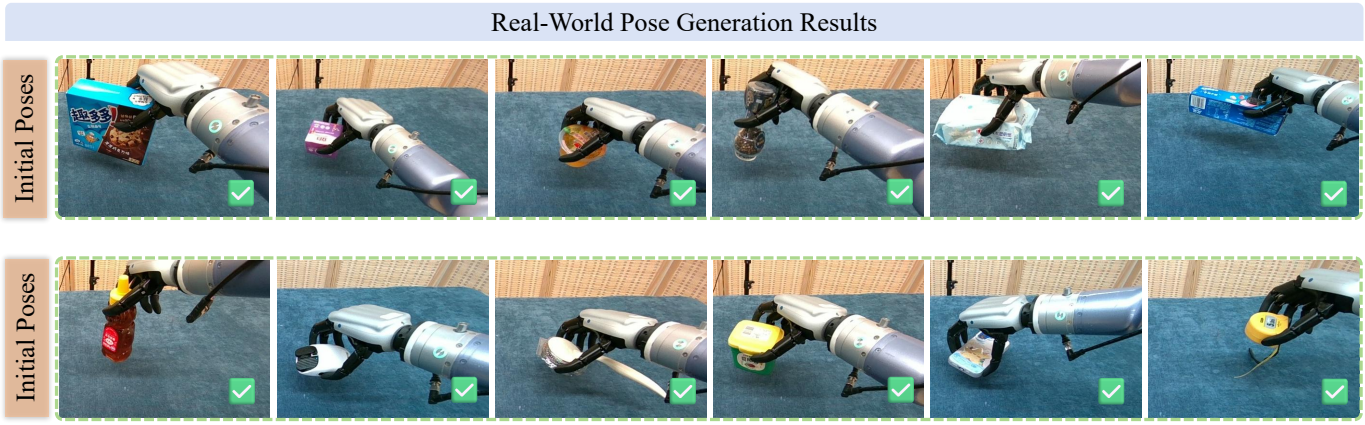

Fig. 8: Successful grasp pose generation on real-world objects.

APPENDIX D
TRAINING AND INFERENCE DETAILS

All four models train on one NVIDIA RTX 4090 (24 GB): the CMap and pose generators (batch 32, lr 1×10^{-3}) in about 3 hours each, the classifier (batch 128, lr 3×10^{-4}) in about 10 hours, and the adapter (batch 256, lr 1×10^{-4}) in about three days. Inference uses about 16 GB in IsaacGym and 2 GB on the real robot. Closed-loop refinement uses $\tau_{succ} = 0.5$ and $K = 5$; it triggers in roughly 30% of trials, over 90% of which succeed within 3 iterations. The diffusion backbone runs at 20–25 Hz, fast enough for real-time refinement when the object moves.

APPENDIX E
ADDITIONAL QUALITATIVE RESULTS: ADAPTATION

Initial poses may fail from insufficient contact points, limited contact area, or geometric mismatch; the classifier flags these and the adaptation module corrects them (Fig. 9), with further simulation and real-world examples in Fig. 10 and Fig. 11.

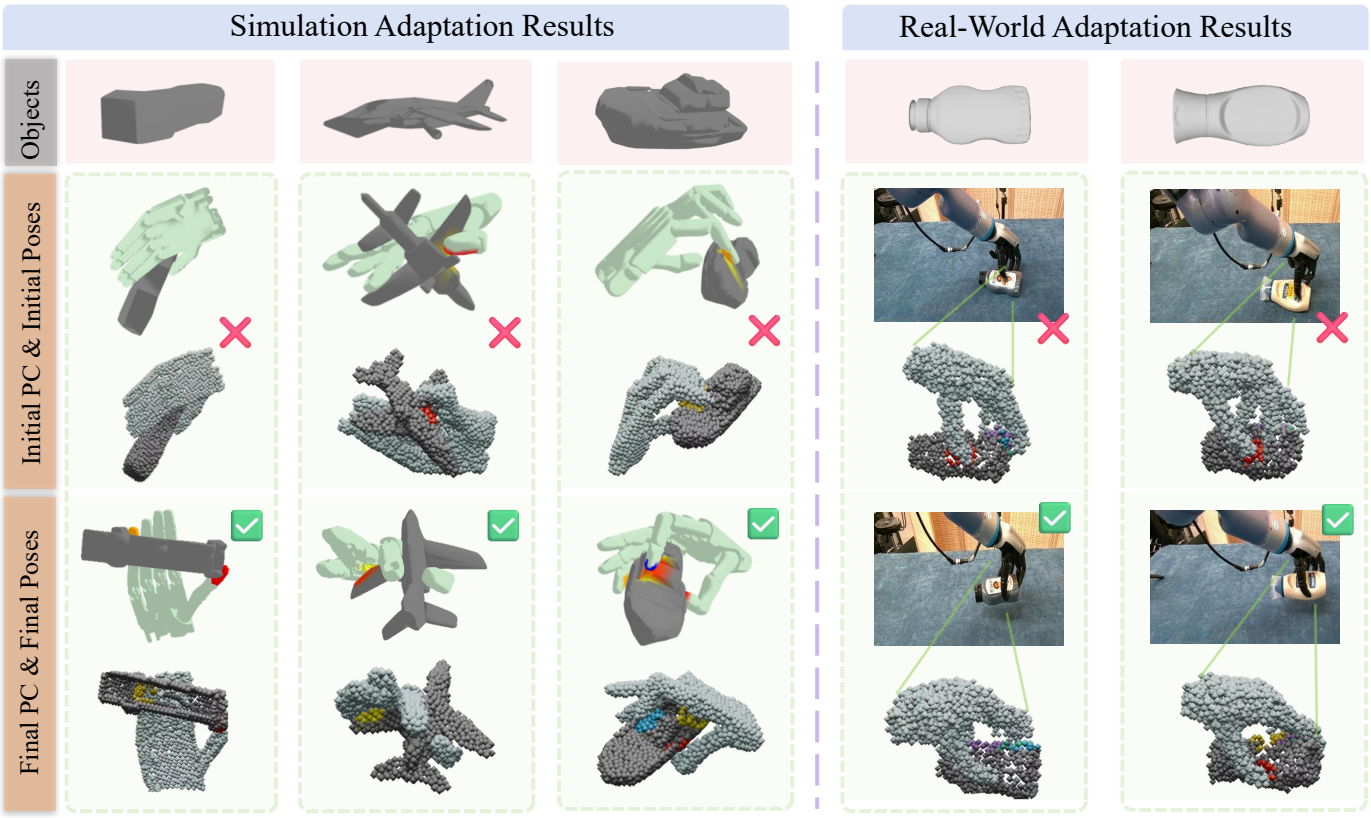

Fig. 9: **Grasp pose adaptation.** Row 1: object samples. Rows 2–3: unstable initial grasps with sparse or weak contact. Rows 4–5: refined poses with denser contact and larger support after adaptation.

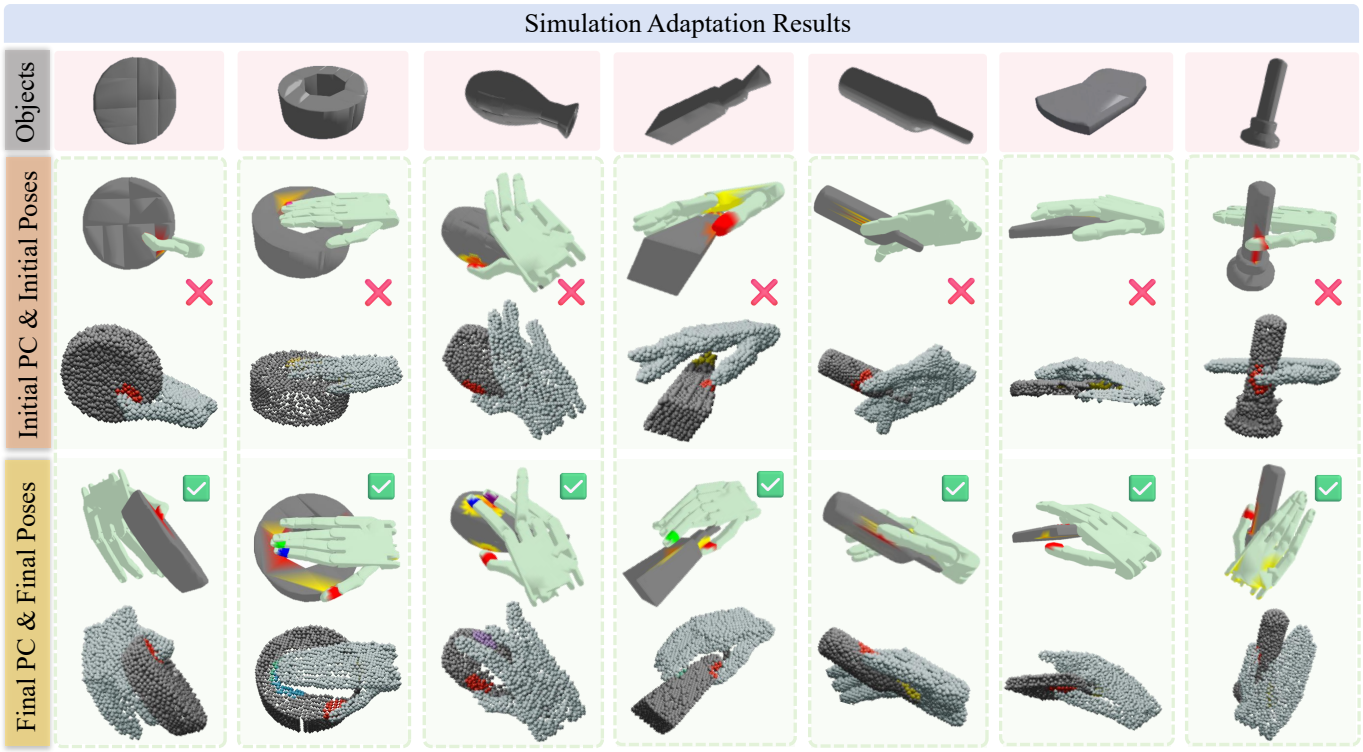

Fig. 10: Grasp pose adaptation on diverse simulation objects.

Fig. 11: Grasp pose adaptation on real-world objects.

APPENDIX F
LIMITATIONS AND FUTURE WORK

The tactile module encodes only contact occurrence, not force or pressure, which may cause excessive force on fragile objects, and we focus on single-object grasping; cluttered and functional grasping remain open. Future work will improve simulation fidelity, extend the tasks, and explore richer visuo-tactile fusion.