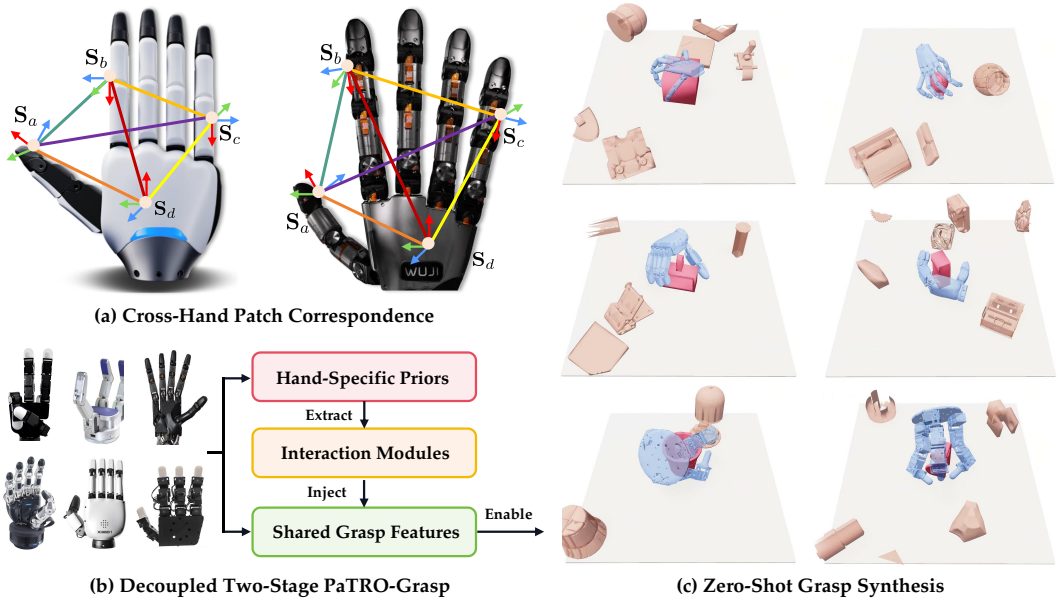
PaTRO-Grasp: Zero-Shot Cross-Embodiment Dexterous Grasping with Surface Patches

Anonymous authors
Paper under double-blind review



(a) Cross-Hand Patch Correspondence

(b) Decoupled Two-Stage PaTRO-Grasp

(c) Zero-Shot Grasp Synthesis

Figure 1: (a) We introduce a surface-patch representation for dexterous hands, establishing cross-hand correspondence in $\mathrm{SE}(3)$ space. (b) Building on surface patch, PaTRO-Grasp disentangles hand-specific priors through pretraining from embodiment-shared grasping knowledge learned by a shared policy. (c) This paradigm enables zero-shot grasp synthesis on novel hand embodiments.

Abstract

The growing diversity of dexterous hands calls for grasping policies that generalize to unseen hands. However, zero-shot transfer remains challenging due to substantial differences in action spaces across embodiments. To tackle this challenge, we propose **PaTRO-Grasp**, a universal policy built on surface-patch representations for arbitrary dexterous hands. PaTRO-Grasp adopts a two-stage training paradigm to establish cross-hand correspondence in a shared high-dimensional feature space, enabling zero-shot policy transfer to novel hand embodiments for grasp synthesis in cluttered environments. Across novel hands spanning three-finger grippers to four- and five-finger dexterous hands, PaTRO-Grasp achieves an average success rate of 84.3%, demonstrating strong zero-shot generalization across diverse hand morphologies. The code and videos are available at the anonymous website https://patrograsp.github.io/.

1 Introduction

Cross-embodiment dexterous grasping is increasingly important as the diversity of dexterous hands continues to grow, while zero-shot transfer to unseen embodiments remains particularly challenging due to substantial differences in action spaces and kinematic structures across hands. One line of work focuses on unified grasp synthesis across multiple known embodiments (Wei et al., 2025;

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.

Fei et al., 2026; Zhang et al., 2024b), enabling a single model to share grasping knowledge among different hands. However, these methods are typically trained and evaluated on the same set of embodiments, limiting their ability to generalize to novel hands. A second line of work explicitly targets unseen-hand transfer through hand-agnostic or canonical intermediate representations, including object-centric contact prediction with hand-specific optimization (Li et al., 2023; Xu et al., 2024b; 2023; Wu et al., 2025), morphology-aligned action representations (Wu et al., 2026), and canonical shared action spaces (Wei et al., 2026; Casas et al., 2026; Yuan et al., 2025). While these approaches broaden zero-shot generalization, they often rely on hand-specific optimization, predefined cross-hand correspondences, or manually designed shared action spaces, which can hinder transfer between hands with substantially different morphologies and kinematic structures.

To address these limitations, we propose PaTRO-Grasp, a two-stage diffusion framework based on surface-patch representations for zero-shot cross-embodiment grasp synthesis. By defining local surface patches on each hand, we map its configuration space into distributions of surface-patch $\mathrm{SE}(3)$ poses, providing a unified representation across dexterous hands. In the first stage, the pretrained branch learns the distribution of valid surface-patch $\mathrm{SE}(3)$ configurations, while learnable embeddings associated with each surface patch capture hand-specific geometric and kinematic information. In the second stage, the main grasping branch learns embodiment-shared grasping knowledge from training hands, while a cross-branch interaction module incorporates pretrained hand-specific features into the main branch. For a novel hand, zero-shot deployment requires only a lightweight optimization of its patch embeddings under the frozen pretrained diffusion model, enabling rapid adaptation without grasp annotations or updates to the grasping model. In conclusion, our contributions are summarized as:

(1) We introduce a hand-structure pretraining module based on surface-patch representations, where learnable patch embeddings encode hand-specific geometric and kinematic information.

(2) We propose PaTRO-Grasp, a dual-branch diffusion framework for zero-shot cross-embodiment grasp synthesis, enabling fast and lightweight deployment to novel hands without additional training.

(3) We conduct extensive experiments in both simulation and the real world, achieving an average success rate of 84.3% for zero-shot grasp synthesis on unseen hands in cluttered scenes.

2 Related Work

2.1 Representations for Dexterous Grasping

Zero-shot dexterous grasping requires to transfer grasp knowledge to unseen hands with substantial differences in geometry, kinematics, and action spaces. Conventional robot-centric representations directly parameterize grasps using wrist poses and joint configurations (Wan et al., 2023; Zhong et al., 2025; Zhang et al., 2024a; Jiang et al., 2021; Xu et al., 2024a), making the grasp representation inherently embodiment-specific and difficult to transfer directly to unseen hands. Object-centric representations describe grasps through hand-agnostic contacts on the object surface (Attarian et al., 2023; Li et al., 2023; Ma et al., 2025; Liu et al., 2023b; Zhao et al., 2024; Xu et al., 2024b), allowing the same contact representation to be shared across different hands. However, each hand still requires a separate optimization process to recover executable joint configurations under its geometry, kinematics, and joint limits, which introduces substantial computational cost and complicates transfer across diverse hand morphologies.

Beyond these representations, interaction-based representations further model the geometric and spatial relationships between hand and object (Yang et al., 2026; Zhang et al., 2026; Wei et al., 2025; Fei et al., 2026; Liu et al., 2023a; Lee et al., 2026). Typically, $\mathcal{D}(\mathcal{R}, \mathcal{O})$ introduces a unified distance-based representation to model hand-object interactions across different embodiments, together with a contrastive pretraining strategy to learn shared hand representations. However, its dense distance features introduce considerable computational overhead. Moreover, distance alone does not explicitly encode link orientations, limiting its ability to fully characterize the hand configuration space. $\mathcal{T}(\mathcal{R}, \mathcal{O})$ instead represents hand-object relations using relative $\mathrm{SE}(3)$ transformations between hand links and the object, providing a compact and structured cross-embodiment representation. However, the link-wise transformations depend on URDF-specific frame definitions, whose origins and axes can vary substantially across hands, making direct zero-shot transfer to unseen embodiments challenging. In contrast, PaTRO-Grasp anchors local $\mathrm{SE}(3)$ frames directly on the

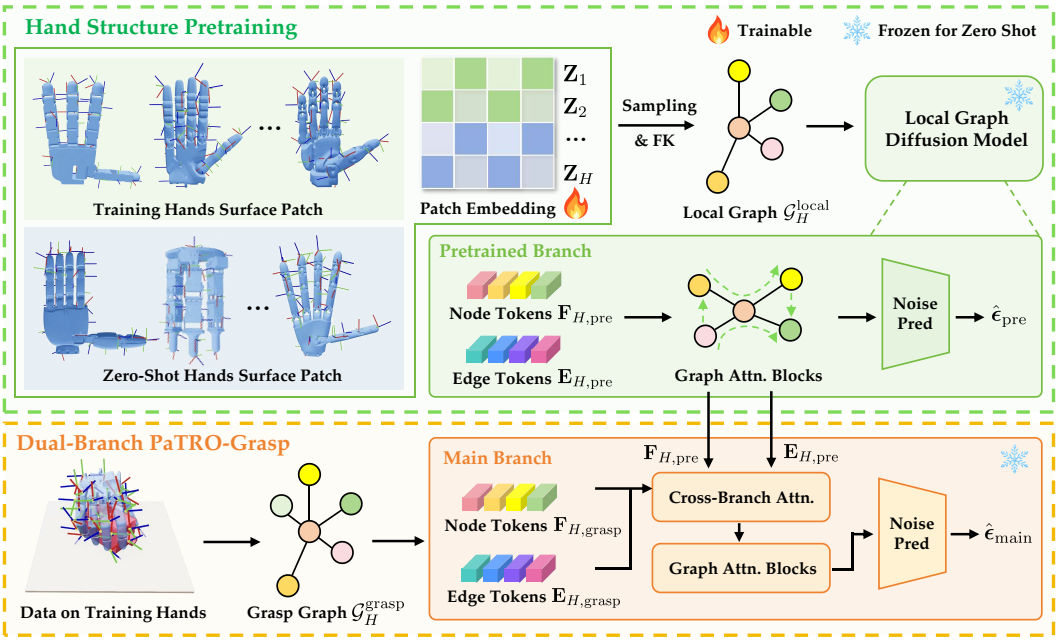


Figure 2: **Overview of PaTRO-Grasp pipeline.** PaTRO-Grasp employs a dual-branch diffusion framework for zero-shot cross-embodiment grasp synthesis, requiring only lightweight patch-embedding optimization for novel hands deployment without additional training.

hand surface, avoiding the ambiguity of URDF-dependent link frames. More importantly, its pre-training stage learns the valid surface-patch configuration space of each hand, providing a complete hand-specific structural prior for subsequent cross-embodiment grasp transfer.

2.2 Cross-Embodiment Grasp Learning

Beyond representation design, cross-embodiment grasping requires transferring learned grasping knowledge across hands with different morphologies and kinematics. Existing approaches achieve this by conditioning grasp generation on hand morphology (Attarian et al., 2023), mapping different hands into universal or canonical action spaces (Yuan et al., 2025; Wei et al., 2026; Wu et al., 2026), or introducing explicit cross-hand alignment and hand-specific structural features (Khargonkar et al., 2025; Wei et al., 2025; Wu et al., 2025). While effective, these approaches often rely on predefined cross-hand correspondences, canonicalized action spaces, or hand-specific retargeting, which can make transfer challenging when the target hand differs substantially in morphology and kinematic constraints. To address these limitations, PaTRO-Grasp directly learns a hand-specific prior over the valid surface-patch $\mathrm{SE}(3)$ configurations induced by the configuration space of each hand, without requiring explicit cross-hand alignment. This formulation supports transfer across substantially different hand morphologies, including embodiments ranging from three-finger grippers to five-finger dexterous hands. For a novel hand, this prior can be efficiently adapted through lightweight optimization, requiring no target-hand grasp annotations or retraining of the shared grasping model.

3 Method

Overview. Zero-shot dexterous grasping requires a unified representation that generalizes across diverse hand embodiments. We represent an arbitrary dexterous hand using a fixed number of patch frames distributed over its surface (Sec. 3.1). Each hand then undergoes an annotation-free pretraining stage to encode its geometry and kinematic structure into implicit embeddings (Sec. 3.2). With the pretrained patch embedding, the grasping model encodes grasp poses across different hands into a shared feature space, enabling zero-shot generalization to arbitrary hands after training (Sec. 3.3). The overall pipeline of PaTRO-Grasp is presented in Figure 2.

3.1 SURFACE PATCH REPRESENTATION

Previous cross-embodiment hand representations mainly rely on hand point clouds (Wei et al., 2025) or link-wise SE(3) frames (Fei et al., 2026). However, these representations can incur high computational cost, provide limited contact information, and remain sensitive to how link frames are defined in the URDF. To overcome these limitations, we propose to represent arbitrary hand skeleton with a fixed number of patch frame attached on the hand surface. For subsequent discussion, we divide the full set of hands into disjoint training and test sets: $\mathcal{H} = \mathcal{H}_{\text{train}} \cup \mathcal{H}_{\text{test}}$.

For each hand $H \in \mathcal{H}$, we sample multiple sets of surface patches over its mesh. Each surface patch is represented by a local frame, where the z-axis follows the surface normal and the x-axis aligns with the principal axis of the corresponding link mesh. We refer to each sampled set of surface patches as a surface-patch distribution, which forms a patch-based structural representation of the hand. In each sampled distribution, we ensure that every link is covered by at least one patch. For each hand embodiment, we generate N patch distributions, with each distribution containing M surface patches. As summarized in Eq. 1, this sampling process produces a diverse set of patch-based hand skeletons for each embodiment, effectively augmenting the limited number of training hands during pretraining.

$$\mathcal{S} = \{\mathcal{S}_H^n\}_{H\in\mathcal{H},\, n=1}^{N}, \quad \mathcal{S}_H^n = \{\mathbf{T}_{H,m}^n\}_{m=1}^{M}, \quad \mathbf{T}_{H,m}^n \in SE(3). \tag{1}$$

Compared with point-cloud representations, surface patches explicitly encode local orientation in addition to position, while requiring substantially fewer features to support prediction in $\mathrm{SE}(3)$ space. Compared with link-wise frame representations, surface patches are independent of URDF-specific frame definitions, making the representation more consistent across embodiments. Moreover, since the patches are anchored directly on the hand surface, they capture hand-object contact geometry more naturally, which is beneficial for transferable grasp learning across different hands.

3.2 HAND STRUCTURE PRETRAINING

At the pretraining stage, we aim to encode hand-specific information, including local geometry and kinematics, into implicit patch embeddings. Specifically, for each hand $H \in \mathcal{H}$ and patch distribution $\mathcal{S}_H^n$, we initialize a random learnable embedding $\mathbf{Z}_H^n \in \mathbb{R}^{M\times D}$. We then sample joint configurations from the hand joint space and apply forward kinematics (FK) to obtain the corresponding patch $\mathrm{SE}(3)$ poses expressed in the palm-local frame. In this way, we generate diverse valid patch poses that cover the feasible hand configuration space for each patch distribution.

Then, we train a graph diffusion model f_{pretrain} to model the distribution of each patch set over the feasible hand configuration space. During pretraining, the patch embeddings are jointly optimized with the denoiser, allowing them to encode hand-specific geometry and kinematics. To capture the structural relations among patches, we adopt a graph architecture inspired by $\mathcal{T}(R, O)$ (Fei et al., 2026), where each node represents a surface patch by concatenating its local pose and learnable embedding, and each edge encodes the relative transformation together with the embeddings of the corresponding patch pair:

$$\mathcal{G}_H^{\text{local}} = (\mathcal{V}_H^{\text{local}}, \mathcal{E}_H^{\text{local}}), \quad \mathcal{V}_H^{\text{local}} = \left\{[\mathbf{T}_{H,i}^{\text{local}}, \mathbf{z}_{H,i}]\right\}_{i=1}^{M}, \quad \mathcal{E}_H^{\text{local}} = \{[\mathbf{T}_{H,ij}, \mathbf{z}_{H,i}, \mathbf{z}_{H,j}]\}_{i\neq j}. \tag{2}$$

Here, $\mathbf{T}_{H,i}^{\text{local}} \in \mathrm{SE}(3)$ and $\mathbf{z}_{H,i} \in \mathbb{R}^D$ denote the local pose and learnable embedding of patch i, respectively, while $\mathbf{T}_{H,ij} \in \mathrm{SE}(3)$ denotes the relative transformation between patches i and j.

The denoiser f_{pretrain} contains L stacked graph-attention blocks (detailed in Appendix A.1) that progressively model the structural relations among patches. We denote the intermediate patch-node and edge features at layer l as $\mathbf{F}_{H,\text{pre}}^{l}$ and $\mathbf{E}_{H,\text{pre}}^{l}$, respectively. At each layer, information is propagated between patch nodes through their connected edges, allowing the network to progressively capture the patch SE3 distribution in the hand configuration space:

$$\left(\mathbf{F}_{H,\text{pre}}^{l+1}, \mathbf{E}_{H,\text{pre}}^{l+1}\right) = f_{\text{pretrain}}^{l}\left(\mathbf{F}_{H,\text{pre}}^{l}, \mathbf{E}_{H,\text{pre}}^{l}\right), \quad l = 0, \ldots, L-1. \tag{3}$$

Since inter-patch relations provide more informative cues about hand structure than absolute patch poses, we perform diffusion on the graph edges by injecting noise $\boldsymbol{\epsilon}_{\text{pre}}$ into the relative patch trans-

formations. The multi-layer node and edge features are aggregated by an MLP to predict the noise, and the model is trained with the standard noise-prediction objective (Song et al., 2020):

$$\hat{\boldsymbol{\epsilon}}_{\text{pre}} = \text{MLP}\left(\{\mathbf{F}^l_{H,\text{pre}}, \mathbf{E}^l_{H,\text{pre}}\}_{l=1}^{L}\right), \qquad \mathcal{L}_{\text{diffusion}} = \mathbb{E}\left[\|\boldsymbol{\epsilon}_{\text{pre}} - \hat{\boldsymbol{\epsilon}}_{\text{pre}}\|_2^2\right]. \tag{4}$$

We optimize this objective differently for training and novel hands. For training hands $H \in \mathcal{H}_{\text{train}}$, we jointly optimize the diffusion model parameters θ and the corresponding patch embeddings, allowing the model to learn shared hand-structure priors while the embeddings capture hand-specific geometry and kinematics. For a novel hand $H \in \mathcal{H}_{\text{test}}$, we freeze the pretrained diffusion model and optimize only its patch embeddings using FK-sampled patch poses, requiring only a lightweight and rapid optimization before zero-shot deployment. Finally, we retain the pretrained diffusion model parameters θ and the patch embeddings of all hands, $\mathcal{Z} = \{\mathbf{Z}_H \in \mathbb{R}^{N\times M\times D}\}_{H\in\mathcal{H}}$, for subsequent dual-branch grasping policy training and inference.

$$\text{Pretraining:} \quad \begin{cases} \min\limits_{\theta,\mathbf{Z}_H} \mathcal{L}_{\text{diffusion}}\left(f_{\text{pretrain}}\left(\mathcal{G}_H^{\text{local}}(\mathbf{Z}_H);\theta\right)\right), & H \in \mathcal{H}_{\text{train}}, \\ \min\limits_{\mathbf{Z}_H} \mathcal{L}_{\text{diffusion}}\left(f_{\text{pretrain}}\left(\mathcal{G}_H^{\text{local}}(\mathbf{Z}_H);\theta\right)\right), \quad \theta \text{ fixed}, & H \in \mathcal{H}_{\text{test}}. \end{cases} \tag{5}$$

3.3 Dual-Branch Policy for Zero-Shot Grasping

After pretraining, we introduce a dual-branch grasping policy for zero-shot grasp synthesis on novel hands. For each training hand $H \in \mathcal{H}_{\text{train}}$, we represent each grasp annotation in clutter as an observation-action pair:

$$(\mathbf{O}_H, \mathbf{A}_H) = (\mathbf{P}, \mathbf{X}_H), \qquad \mathbf{X}_H = \{\mathbf{T}_{H,i}\}_{i=1}^{M}, \tag{6}$$

where $\mathbf{P}$ denotes the observed scene point cloud and $\mathbf{X}_H$ represents the target grasp as the $\text{SE}(3)$ poses in the world frame of its M surface patches.

To enable zero-shot grasp transfer across embodiments, PaTRO-Grasp adopts a dual-branch architecture that disentangles hand-specific structure from embodiment-shared grasping knowledge. Specifically, PaTRO-Grasp consists of a pretrained hand branch f_{pretrain}, a main grasping branch f_{main}, and a cross-branch interaction module f_{interact}. Both f_{pretrain} and f_{main} are composed of multiple graph-attention denoising blocks, while f_{interact} performs layer-wise cross-attention between their intermediate graph features. The pretrained branch f_{pretrain} captures the geometry and kinematics of each hand, whereas f_{main} learns grasping knowledge shared across embodiments and incorporates hand-specific features through f_{interact}. After pretraining, we freeze f_{pretrain} and train only f_{main} and f_{interact} on grasp annotations from $\mathcal{H}_{\text{train}}$. Together, these three modules enable direct zero-shot transfer to novel hand embodiments. We detail the network architecture below.

Main Grasping Branch. The main branch f_{main} is a scene-conditioned graph diffusion model that learns embodiment-shared grasping knowledge. Given the noisy hand state $\mathbf{X}_H$, pretrained patch embeddings $\mathbf{Z}_H$, and scene observation $\mathbf{O}_H$, we construct a grasping graph:

$$\mathcal{G}_H^{\text{grasp}} = (\mathcal{V}_H^{\text{grasp}}, \mathcal{E}_H^{\text{grasp}}), \quad \mathcal{V}_H^{\text{grasp}} = \mathcal{V}_H^{\text{patch}} \cup \mathcal{V}_H^{\text{obs}}, \quad \mathcal{E}_H^{\text{grasp}} = \mathcal{E}_H^{\text{pp}} \cup \mathcal{E}_H^{\text{op}}. \tag{7}$$

Here, $\mathcal{V}_H^{\text{patch}}$ and $\mathcal{V}_H^{\text{obs}}$ denote patch and observation nodes, while $\mathcal{E}_H^{\text{pp}}$ and $\mathcal{E}_H^{\text{op}}$ capture patch–patch and observation–patch relations, respectively. Each observation node corresponds to a sampled scene point, represented by its pose $\mathbf{T}_j^{\text{obs}}$ and feature $\mathbf{f}_j^{\text{obs}}$ extracted by a pretrained PointNet++ (Qi et al., 2017) encoder. With K sampled points, the node and edge features are defined as:

$$\begin{aligned} \mathcal{V}_H^{\text{patch}} &= \{[\mathbf{T}_{H,i}, \mathbf{z}_{H,i}]\}_{i=1}^{M}, & \mathcal{V}_H^{\text{obs}} &= \left\{[\mathbf{T}_j^{\text{obs}}, \mathbf{f}_j^{\text{obs}}]\right\}_{j=1}^{K}, \\ \mathcal{E}_H^{\text{pp}} &= \{[\mathbf{T}_{ij}, \mathbf{z}_{H,i}, \mathbf{z}_{H,j}]\}_{i\neq j}, & \mathcal{E}_H^{\text{op}} &= \left\{\mathbf{T}_{ij}^{\text{obs}}\right\}_{i,j}. \end{aligned} \tag{8}$$

Similar to the pretraining branch defined in Equation 3, the main denoiser f_{main} consists of L stacked graph-attention blocks. We denote its intermediate patch-node and edge features at layer l as $\mathbf{F}^l_{H,\text{main}}$ and $\mathbf{E}^l_{H,\text{main}}$, respectively. Building on the same graph-attention design, each block

propagates observation information to the patch nodes while performing message passing among patches, allowing f_{main} to learn embodiment-shared grasping knowledge from the spatial relations between surface patches and cluttered-scene observations.

Cross-Branch Interaction. To incorporate hand-specific geometry and kinematics into the embodiment-shared grasping branch, we introduce a layer-wise interaction module f_{interact} between the aligned graph denoising layers of f_{pretrain} and f_{main}. At each layer, the main-branch node and edge features separately attend to their pretrained features through multi-head cross-attention:

$$\begin{aligned}\widetilde{\mathbf{F}}^{\,l}_{H,\text{main}} &= \mathbf{F}^{\,l}_{H,\text{main}} + \mathcal{A}^{l}_{\text{node}}\left(\mathbf{F}^{\,l}_{H,\text{main}}, \mathbf{F}^{\,l}_{H,\text{pre}}\right), \\ \widetilde{\mathbf{E}}^{\,l}_{H,\text{main}} &= \mathbf{E}^{\,l}_{H,\text{main}} + \mathcal{A}^{l}_{\text{edge}}\left(\mathbf{E}^{\,l}_{H,\text{main}}, \mathbf{E}^{\,l}_{H,\text{pre}}\right),\end{aligned} \tag{9}$$

where $\mathcal{A}^{l}_{\text{node}}$ and $\mathcal{A}^{l}_{\text{edge}}$ denote the node- and edge-level cross-attention modules, respectively. The interaction-enhanced features are then passed to the corresponding graph-attention block of f_{main}. Overall, the layer-wise dual-branch feature update in PaTRO-Grasp is given by:

$$\left(\mathbf{F}^{\,l+1}_{H,\text{main}}, \mathbf{E}^{\,l+1}_{H,\text{main}}\right) = f^{\,l}_{\text{main}}\left(\widetilde{\mathbf{F}}^{\,l}_{H,\text{main}}, \widetilde{\mathbf{E}}^{\,l}_{H,\text{main}}, \mathcal{V}^{\text{obs}}_{H}\right), \quad l = 0, \ldots, L-1. \tag{10}$$

Finally, the interaction-enhanced features from all layers are aggregated to predict the diffusion noise over the target patch poses as Equation 11. Through this dual-branch architecture, f_{main} learns embodiment-shared, scene-conditioned grasping representations, while the frozen f_{pretrain} provides hand-specific geometric and kinematic priors through layer-wise interaction, enabling zero-shot generalization for grasp synthesis to unseen hand embodiments.

$$\hat{\boldsymbol{\epsilon}}_{\text{main}} = \text{MLP}\left(\{\mathbf{F}^{\,l}_{H,\text{main}}, \mathbf{E}^{\,l}_{H,\text{main}}\}_{l=1}^{L}\right), \qquad \mathcal{L}_{\text{diffusion}} = \mathbb{E}\left[\|\boldsymbol{\epsilon}_{\text{main}} - \hat{\boldsymbol{\epsilon}}_{\text{main}}\|_2^2\right]. \tag{11}$$

Zero-Shot Inference. For a novel hand $H \in \mathcal{H}_{\text{test}}$, the key insight is to adapt only its hand-specific prior while directly reusing the embodiment-shared grasping model. We first construct its surface-patch distributions following Section 3.1 and optimize the corresponding patch embeddings $\mathbf{Z}^{n}_{H}$ with f_{pretrain} frozen, as described in Section 3.2. Given the optimized embeddings and scene observation $\mathbf{O}_H$, PaTRO-Grasp performs reverse DDIM sampling from Gaussian noise, progressively denoising the patch poses into a valid grasp in the scene. Finally, the predicted patch $\text{SE}(3)$ poses are retargeted to executable joint configurations using either Pyroki (Kim et al., 2025) or a pretrained MLP-based retargeting head. Both approaches enable accurate and real-time conversion from patch poses to joint values for robot execution.

4 Experiment

In this section, we perform a series of experiments aimed to address the following questions:

Q1: How successful is zero-shot grasp synthesis based on surface patch representation? (Sec. 4.2)
Q2: Why does the two-stage training enable zero-shot cross-embodiment grasping? (Sec. 4.3)
Q3: How does each network component contribute to zero-shot grasp synthesis? (Sec. 4.4)
Q4: How does PaTRO-Grasp perform in the real world? (Sec. 4.5)

4.1 Experimental Setting

Dataset. We collect 53,675 objects from UniDexGrasp (Xu et al., 2023) and ObjaverseXL (Deitke et al., 2023), using 47,737 for training and 5,938 for testing. Since existing cross-embodiment grasp synthesis methods such as GenDexGrasp (Li et al., 2023), $\mathcal{D}(\mathcal{R}, \mathcal{O})$ (Wei et al., 2025), $\mathcal{T}(\mathcal{R}, \mathcal{O})$ (Fei et al., 2026), and $\mathcal{OHRA}$ Grasp (Wei et al., 2026) are designed primarily for isolated-object settings, we evaluate them on the test objects without surrounding clutter. In contrast, both DexGrasp-Zero (Wu et al., 2026) and PaTRO-Grasp support tabletop grasping, and are therefore evaluated under the same cluttered tabletop setting. Specifically, each target object is placed on a table and surrounded by 5–15 randomly placed obstacle objects, requiring the model to synthesize a collision-free grasp for the designated target.

Method	Training Hands						Zero-shot Hands			
	Allegro	Barrett	Shadow	LEAP	Schunk	Ability	XHand	DClaw	XHand-Pro	Wuji
GenDexGrasp (Li et al., 2023)	45.26%	30.03%	14.23%	-	-	-	12.56%	19.68%	29.16%	8.87%
$\mathcal{D}(\mathcal{R},\mathcal{O})$ Grasp (Wei et al., 2025)	81.04%	77.80%	42.05%	-	-	-	0.76%	20.86%	10.09%	29.17%
$\mathcal{T}(\mathcal{R},\mathcal{O})$ Grasp (Fei et al., 2026)	90.50%	82.74%	70.09%	-	-	-	0.05%	2.48%	2.45%	2.08%
$\mathcal{OHRA}$ Grasp (Wei et al., 2026)	59.56%	83.86%	67.30%	26.73%	-	-	1.39%	9.25%	2.45%	2.08%
DexGrasp-Zero Wu et al. (2026)	17.64%	-	20.63%	-	5.30%	1.17%	1.14%	-	1.14%	0.03%
PaTRO-Grasp	**91.43%**	**83.53%**	**93.59%**	**88.77%**	**93.27%**	**79.98%**	**82.47%**	**89.74%**	**83.36%**	**81.45%**

Table 1: **Zero-shot grasp synthesis performance across different dexterous hands.**

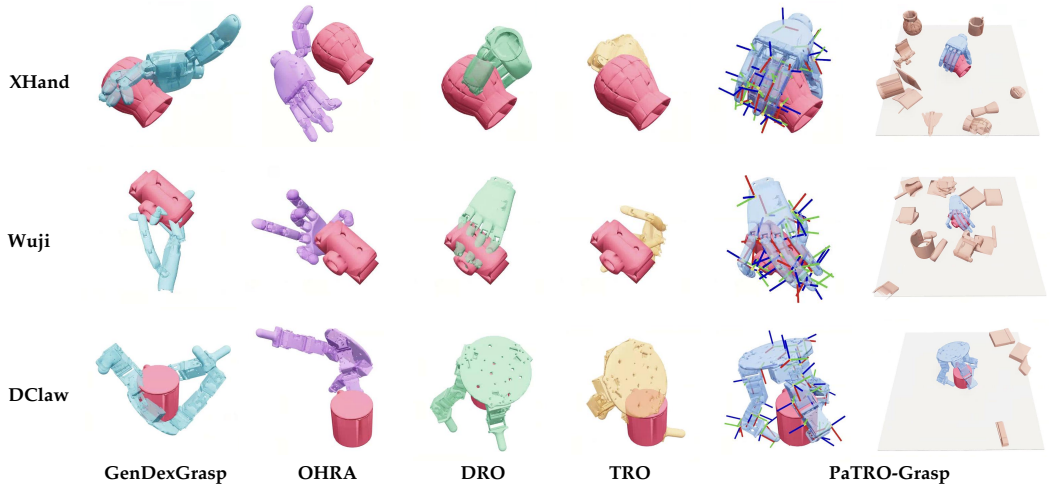


Figure 3: **Zero-shot grasp synthesis of PaTRO-Grasp and baselines.** For PaTRO-Grasp, we visualize the predicted surface-patch frames on the retargeted hand mesh and resulting grasps in clutter.

Evaluation Protocol. In simulation, we evaluate PaTRO-Grasp and all baselines in Isaac Gym. For DexGrasp-Zero, we reimplement the method in Isaac Gym following its released code to ensure a fair comparison. For isolated-object grasp synthesis baselines, we follow the evaluation protocol of $\mathcal{D}(\mathcal{R},\mathcal{O})$ Grasp, where each predicted grasp is executed by a controller and considered successful if it satisfies force closure. For DexGrasp-Zero and PaTRO-Grasp, we measure execution success by whether the predicted grasp can lift the target object above the tabletop and maintain a stable hold.

Implementation Details. For each dexterous hand, we sample $N = 1024$ surface-patch distributions, each containing $M = 32$ patch frames anchored on the hand surface, and define the corresponding patch embeddings as $\mathbf{Z}_H \in \mathbb{R}^{1024\times32\times16}$. Both the pretrained and main grasping branches consist of $L = 6$ graph-attention denoising blocks, with cross-branch interaction applied between the corresponding layers. For both diffusion models, we use a consine noise schedule with 1,000 diffusion steps and 20 DDIM steps for reverse sampling. We train the main grasping model on eight NVIDIA A100 80GB GPUs using a learning rate of 1×10^{-4}. Additional architectural details and hyperparameter settings are provided in the Appendix.

4.2 Cross-Embodiment Zero-Shot Grasp Synthesis

We train PaTRO-Grasp on six dexterous hands spanning three- to five-finger embodiments, enabling the main branch to learn grasping knowledge across a wide range of degrees of freedom. As shown in Table 1, PaTRO-Grasp achieves an average success rate of 88.4% on the training hands and 84.3% on unseen test hands, outperforming all grasp synthesis baselines and demonstrating robust zero-shot generalization across diverse embodiments, from three-finger grippers to high-DoF dexterous hands.

Figure 3 further visualizes the zero-shot predictions of all methods on the same object. GenDexGrasp optimizes hand joints to match the predicted contact map, but it suffers from hand-object penetration due to insufficient geometric constraints. $\mathcal{OHRA}$ Grasp and $\mathcal{D}(\mathcal{R},\mathcal{O})$ Grasp can produce hand poses with slight penetration or insufficient object contact, indicating limited generalization of their learned hand representations to unseen embodiments. For $\mathcal{T}(\mathcal{R},\mathcal{O})$ Grasp, unseen link-frame definitions make it difficult to distinguish the roles of different links, leading to collapse in link-wise

Method	Training Hands							Zero-shot Hands				
	Allegro	Barrett	Shadow	LEAP	Schunk	Ability	**Avg.**	XHand	DClaw	XHand-Pro	Wuji	**Avg.**
Trans. Error (cm)	1.500	1.277	1.002	1.606	0.789	0.620	**1.132**	1.408	1.672	1.150	1.320	**1.388**
Rot. Error (deg)	3.813	4.439	4.198	5.314	3.309	3.113	**4.031**	4.106	4.800	3.968	5.466	**4.585**

Table 2: **Patch pose errors between generated patch poses and FK-recovered poses after IK.**

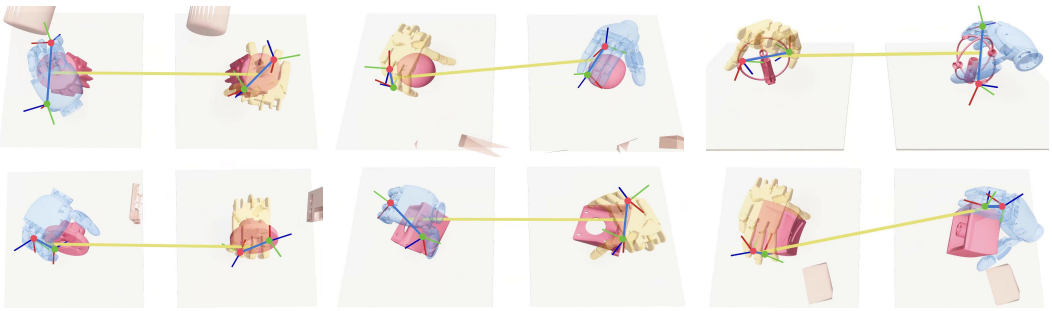


Figure 4: **Patch-edge feature correspondence between Shadow and XHand.** Blue lines denote patch edges, while yellow lines highlight the edge pair with the highest feature similarity across the two hands.

$\mathrm{SE}(3)$ predictions. In contrast, PaTRO-Grasp demonstrates strong zero-shot generalization across diverse dexterous embodiments based on surface patch prediction in $\mathrm{SE}(3)$ space, and additionally supports grasp synthesis in cluttered scenes.

4.3 Two-Stage Training for Zero-Shot Transfer

In this section, we investigate why the surface patch representation and two-stage training framework of PaTRO-Grasp enable zero-shot transfer across hand embodiments. Specifically, we study two key questions: **(1) Can the pretrained diffusion model capture the feasible hand configuration space through surface-patch SE(3) poses?** and **(2) What embodiment-shared features does the model learn to enable zero-shot grasp synthesis on unseen hands?**

Pretrained Surface-Patch Configuration Distribution. To examine whether the pretrained diffusion model captures the valid surface-patch distribution induced by the hand configuration space, we use DDIM sampling to generate surface-patch $\mathrm{SE}(3)$ poses from f_{pretrain}. We then use Pyroki (Kim et al., 2025) to solve IK for the hand joint configuration that best matches the generated patch poses. After obtaining the joint values, we apply FK to recover the corresponding patch poses and compute the per-patch translation and rotation errors from the original generated poses. We report the average errors in Table 2. Consistently low errors indicate that the generated surface-patch configurations can be closely realized by valid hand joint configurations. The comparable errors between training and zero-shot hands further demonstrate that optimizing only the patch embeddings is sufficient to adapt the pretrained model to unseen hand embodiments.

Embodiment-Shared Grasp Features. To enable zero-shot grasp synthesis across hand embodiments, embodiment-shared grasping knowledge should preserve similar relative relationships among surface patches when different hands perform similar grasps. Based on this intuition, we perform inference with different hands in the same scenes and compare their intermediate grasp edge features $\mathbf{E}^{,l}_{H,\text{main}}$. As shown in Figure 4, similar patch relations give rise to similar edge features across both training and unseen hands. This indicates that the learned edge features capture embodiment-shared relational patterns and establish correspondences between surface patches across different hands. Such cross-hand correspondences provide a transferable representation of grasp structure, facilitating the transfer of grasping knowledge to unseen embodiments.

4.4 Ablation

In this section, we conduct ablation studies on the PaTRO-Grasp architecture to analyze how each component contributes to zero-shot grasp transfer across different hand embodiments. On XHand,

Model	SR
Full model	**82.47%**
w/o interaction module f_{interact}	0.03%
w/o patch embedding in cross-attn.	64.51%

Table 3: **Ablation of PaTRO-Grasp.**

Peach	Cabbage	Box	Starfruit	Mango
7/10	7/10	9/10	10/10	7/10
Bear	Bottle	Cookie	Mug	Cuboid
8/10	8/10	9/10	8/10	9/10

Table 4: **Real-world results on XHand in clutter.**

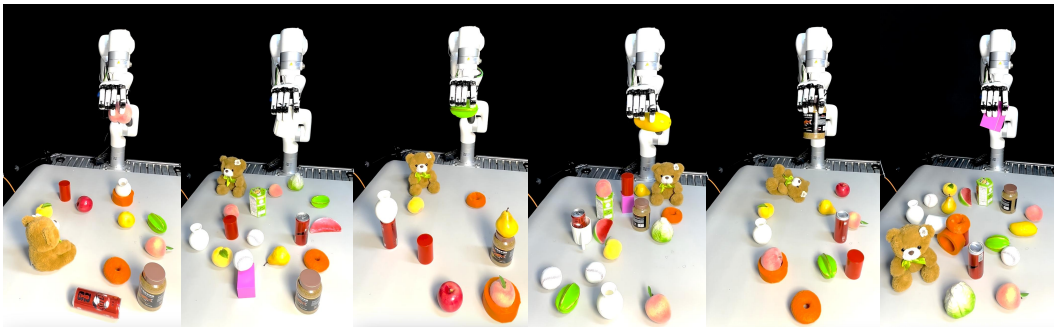


Figure 5: **Zero-shot grasp synthesis on XHand in the real world.**

we evaluate three zero-shot inference settings: (1) the full dual-branch PaTRO-Grasp model; (2) the main grasping branch alone, without f_{interact} to incorporate hand-specific features from the pretrained branch; and (3) the dual-branch model with cross-branch attention preserved but patch embeddings removed from the node and edge features, eliminating explicit patch identity during interaction.

As shown in Table 3, the single-branch model completely fails in zero-shot grasp synthesis. Without hand-specific geometric and kinematic information from the pretrained branch, the main grasping branch cannot form hand-aware representations for the novel embodiment, resulting in unsuccessful grasp synthesis. Meanwhile, removing the pretrained patch embeddings from cross-branch attention leads to a 17.96% performance drop. This indicates that the pretrained embeddings encode patch identity, enabling the interaction module to establish correct cross-branch correspondences and effectively transfer hand-specific geometry and kinematics into the main grasping branch.

4.5 Real-world Experiment

We conduct real-world experiments using a uFactory xArm7 equipped with a right XHand on ten target objects placed in cluttered scenes. As shown in Figure 5, two Intel RealSense D435i RGB-D cameras are used to capture the scene point cloud, which serves as the observation input to the PaTRO-Grasp policy. The policy then performs zero-shot grasp synthesis to predict feasible grasp poses. We use Pyroki (Kim et al., 2025) to compute a collision-free motion plan to the predicted grasp, and execute the grasp using the same controller as in our simulation evaluation protocol. After closing the hand, the robot lifts the target object to evaluate grasp success. In Table 5, we report the per-object success rate over 10 grasp trials for each target object, achieving an average success rate of 82%. Additional qualitative results and videos are available on our anonymous project website https://patrograsp.github.io/.

5 Conclusion

In this paper, we introduced PaTRO-Grasp, a dual-branch diffusion framework based on surface-patch representations for zero-shot cross-embodiment grasp synthesis. By decoupling hand-specific geometry and kinematics from embodiment-shared grasping knowledge across two branches and connecting them through an interaction module, PaTRO-Grasp achieves an average success rate of 84.3% on unseen hand embodiments, demonstrating strong zero-shot cross-embodiment generalization. Furthermore, the surface-patch representation and two-stage framework offer a general approach to zero-shot cross-embodiment transfer. By modeling diverse embodiments in a shared surface-patch $\mathrm{SE}(3)$ space while preserving embodiment-specific priors, this framework could extend beyond dexterous hands to arm-hand and whole-body manipulation.

AI USE STATEMENT

We used generative AI tools for code editing and debugging, manuscript polishing, and improving the presentation of results. All AI-assisted content was independently reviewed and verified by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies dexterous robotic grasping in simulation and real-world robotic systems and does not involve human subjects, personal data, or sensitive information. Real-world experiments were conducted in controlled laboratory environments with appropriate safety precautions.

REPRODUCIBILITY STATEMENT

We provide the model architecture, surface-patch representation, training and inference procedures, datasets, implementation details, evaluation protocols, and additional results in the main paper and appendix. An anonymous project page is also provided with code and additional results.

REFERENCES

Maria Attarian, Muhammad Adil Asif, Jingzhou Liu, Ruthrash Hari, Animesh Garg, Igor Gilitschenski, and Jonathan Tompson. Geometry Matching for Multi-Embodiment Grasping. In *CoRL*, 2023.

Luis Felipe Casas, Robert Teal, Keval Shah, Abhijit Tadepalli, Wanxin Jin, and Yu Xiang. Cross-embodiment robot manipulation via a unified hand action space. *arXiv preprint arXiv:2607.03570*, 2026.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *NeurIPS*, 36:35799–35813, 2023.

Xin Fei, Zhixuan Xu, Huaicong Fang, Tianrui Zhang, and Lin Shao. T(r,o) grasp: Efficient graph diffusion of robot-object spatial transformation for cross-embodiment dexterous grasping. In *ICRA*, 2026.

Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, 2021.

Ninad Khargonkar, Luis Felipe Casas, Balakrishnan Prabhakaran, and Yu Xiang. RobotFingerPrint: Unified Gripper Coordinate Space for Multi-Gripper Grasp Synthesis and Transfer. In *IROS*, 2025.

Chung Min Kim, Brent Yi, Hongsuk Choi, Yi Ma, Ken Goldberg, and Angjoo Kanazawa. Pyroki: A modular toolkit for robot kinematic optimization. In *IROS*, pp. 1312–1319. IEEE, 2025.

Yeonseo Lee, Taeyeop Lee, Hyosup Shin, Guebin Hwang, and Sungho Jo. GraspGraphNet: Graph-Structured Multi-Embodiment Dexterous Grasp Generation. arXiv preprint arXiv:2607.11031, 2026.

Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *ICRA*, 2023.

Qingtao Liu, Yu Cui, Qi Ye, Zhengnan Sun, Haoming Li, Gaofeng Li, Lin Shao, and Jiming Chen. Dexrepnet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations. In *IROS*, pp. 3153–3160. IEEE, 2023a.

Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. ContactGen: Generative Contact Modeling for Grasp Generation. In *CVPR*, 2023b.

Yiyao Ma, Kai Chen, Kexin Zheng, and Qi Dou. Contact Map Transfer with Conditional Diffusion Model for Generalizable Dexterous Grasp Generation. In *NeurIPS*, 2025.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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. In *CVPR*, pp. 3868–3879, 2023.

Zhenyu Wei, Zhixuan Xu, Jingxiang Guo, Yiwen Hou, Chongkai Gao, Zhehao Cai, Jiayu Luo, and Lin Shao. $\mathcal{D}(\mathcal{R}, \mathcal{O})$ grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. In *ICRA*, 2025.

Zhenyu Wei, Yunchao Yao, and Mingyu Ding. One hand to rule them all: Canonical representations for unified dexterous manipulation. In *RSS*, 2026.

Yuliang Wu, Yanhan Lin, WengKit Lao, Yuhao Lin, Yi-Lin Wei, Wei-Shi Zheng, and Ancong Wu. Dexgrasp-zero: A morphology-aligned policy for zero-shot cross-embodiment dexterous grasping. In *RSS*, 2026.

Zhiyuan Wu, Rolandos Alexandros Potamias, Xuyang Zhang, Zhongqun Zhang, Jiankang Deng, and Shan Luo. Cedex: Cross-embodiment dexterous grasp generation at scale from human-like contact representations. *arXiv preprint arXiv:2509.24661*, 2025.

Guo-Hao Xu, Yi-Lin Wei, Dian Zheng, Xiao-Ming Wu, and Wei-Shi Zheng. Dexterous grasp transformer. In *CVPR*, 2024a.

Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *CVPR*, pp. 4737–4746, 2023.

Zhixuan Xu, Chongkai Gao, Zixuan Liu, Gang Yang, Chenrui Tie, Haozhuo Zheng, Haoyu Zhou, Weikun Peng, Debang Wang, Tianrun Hu, Tianyi Chen, Zhouliang Yu, and Lin Shao. ManiFoundation Model for General-Purpose Robotic Manipulation of Contact Synthesis with Arbitrary Objects and Robots. In *IROS*, 2024b.

Xuanze Yang, Yumeng Liu, Haiyang Xin, Changhao Li, Haowei Shen, Kai Xu, Ligang Liu, and Ruizhen Hu. InterMASH: A Unified Geometric Representation for Grasp Synthesis. In *SIGGRAPH Asia 2026 Conference Papers*, 2026.

Haoqi Yuan, Bohan Zhou, Yuhui Fu, and Zongqing Lu. Cross-embodiment dexterous grasping with reinforcement learning. In *ICLR*, volume 2025, pp. 81413–81434, 2025.

Heng Zhang, Kevin Yuchen Ma, Mike Zheng Shou, Weisi Lin, and Yan Wu. MANGO-Grasp: Mahalanobis Fields over Geometry-Oriented 3D Gaussians for Cross-Embodiment Dexterous Grasping. arXiv preprint arXiv:2608.02014, 2026.

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. In *8th Annual Conference on Robot Learning*, 2024a.

Zhengshen Zhang, Lei Zhou, Chenchen Liu, Zhiyang Liu, Chengran Yuan, Sheng Guo, Ruiteng Zhao, Marcelo H Ang Jr, and Francis EH Tay. Dexgrasp-diffusion: Diffusion-based unified functional grasp synthesis method for multi-dexterous robotic hands. *arXiv preprint arXiv:2407.09899*, 2024b.

Fuqiang Zhao, Dzmitry Tsetserukou, and Qian Liu. GrainGrasp: Dexterous Grasp Generation with Fine-grained Contact Guidance. In *ICRA*, 2024.

Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. DexGrasp Anything: Towards Universal Robotic Dexterous Grasping with Physics Awareness. In *CVPR*, 2025.

Appendix

A Implementation Details

A.1 Graph Attention Block

We detail the graph attention blocks used in both the pretrained branch f_{pretrain} and the main grasping branch f_{main}. Given input node tokens $\mathbf{F}_H^l$ and edge tokens $\mathbf{E}_H^l$ at layer l, we first inject the diffusion timestep embedding $\mathbf{t}$ into both node and edge features:

$$\bar{\mathbf{F}}_H^l = \phi_F\left(\mathbf{F}_H^l, \mathbf{t}\right), \qquad \bar{\mathbf{E}}_H^l = \phi_E\left(\mathbf{E}_H^l, \mathbf{t}\right), \tag{12}$$

where ϕ_F and ϕ_E denote learnable time-conditioning layers. For each graph relation, including observation-patch, and patch-patch interactions, we apply an edge-conditioned multi-head attention layer. Given a target node i and a connected source node j, the query and key are projected from their node features, while the value additionally incorporates their pairwise edge feature:

$$\mathbf{q}_i = \mathbf{W}_Q\bar{\mathbf{f}}_i, \qquad \mathbf{k}_j = \mathbf{W}_K\bar{\mathbf{f}}_j, \qquad \mathbf{v}_{ij} = \mathbf{W}_O\sigma\left(\mathbf{W}_s\bar{\mathbf{f}}_j + \mathbf{W}_d\bar{\mathbf{f}}_i + \mathbf{W}_e\bar{\mathbf{e}}_{ij}\right), \tag{13}$$

where $\sigma(\cdot)$ denotes the SiLU activation. Separate projection parameters are used for different graph relations. Then, the target node is updated by multi-head cross attention:

$$\mathbf{f}_i' = \mathrm{MHA}\left(\mathbf{q}_i, \{\mathbf{k}_j\}_j, \{\mathbf{v}_{ij}\}_j\right) + \phi_{\text{self}}\left(\bar{\mathbf{f}}_i\right), \tag{14}$$

where $\mathrm{MHA}(\cdot)$ and $\phi_{\text{self}}(\cdot)$ denote multi-head attention operation and a learnable self-projection layer. In the main grasping branch, the attention is sequentially applied to observation-patch and patch-patch relations, allowing each patch to aggregate scene context from observation tokens and embodiment-specific structural information from other patches.

The edge features are updated in parallel with the node features. For each edge (i, j), we first aggregate the surrounding edge context by averaging the edge features associated with its two endpoints:

$$\mathbf{p}_{ij}^l = \frac{1}{2M}\left(\sum_{k=1}^{M}\mathbf{e}_{kj}^l + \sum_{k=1}^{M}\mathbf{e}_{ik}^l\right), \tag{15}$$

The updated edge feature is then obtained by combining this contextual feature with the edge-conditioned attention value $\mathbf{v}_{ij}^l$:

$$\mathbf{e}_{ij}^{l+1} = \mathbf{W}_o\sigma\left(\mathbf{W}_v\mathbf{v}_{ij}^l + \mathbf{W}_p\mathbf{p}_{ij}^l\right). \tag{16}$$

In this way, each edge feature is refined by combining its local hand–scene relation with the surrounding patch-level relational context.

A.2 Model Architecture Setting

Parameter	Value
Number of patch distributions N	1024
Number of patches M	32
Patch embedding dim.	16
Graph attention layers L	6
Hidden feature dim.	384
Number of attention heads	12
Model parameters of f_{pretrain}	18.75 M
Model parameters of f_{main}	28.30 M
Model parameters of f_{interact}	13.12 M

Table 5: **Model architecture settings.**

Parameter	Value
Learning rate	1×10^{-4}
Diffusion steps	1000
Diffusion scheduler	cosine
Training batch size per GPU	32
Training iterations	100 K
Optimizer	AdamW
DDIM inference steps	20
Pyroki iterations	50
Pyroki trans./rot. weight	10.0/5.0

Table 6: **Training and inference settings.**

We summarize the model architecture, training, and inference configurations of PaTRO-Grasp in Tables 5 and 6. The pretrained and main grasping branches use matched graph-attention architectures, enabling layer-wise interaction between their intermediate features, while the main branch additionally incorporates scene observations for grasp synthesis. The interaction module provides cross-branch feature fusion. During training, we freeze the pretraining branch and optimize both the main branch and interaction module.

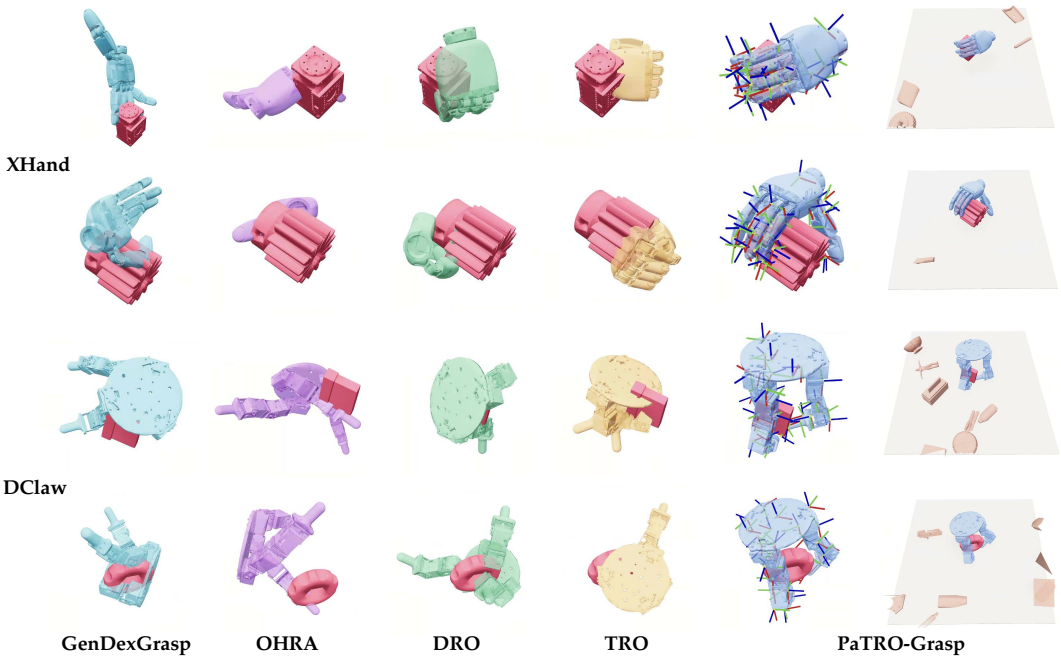


Figure 6: **Additional visualization of zero-shot grasp synthesis of PaTRO-Grasp and baselines.**

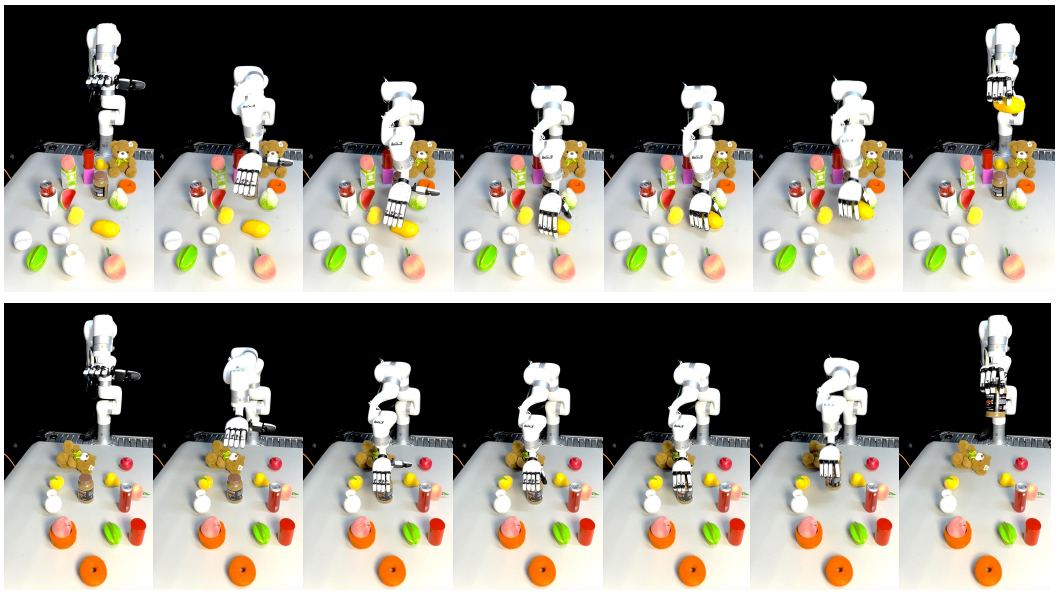


Figure 7: **Additional visualization of zero-shot grasp synthesis of PaTRO-Grasp in the real world.**

B VISUALIZATION

We provide additional visualizations of: (1) zero-shot grasp synthesis in simulation for PaTRO-Grasp and other baselines (Figure 6); and (2) zero-shot real-world grasping with XHand (Figure 7). These results further demonstrate that PaTRO-Grasp demonstrates strong zero-shot generalization across hands in both simulation and real-world cluttered scenes.