Scalable Cross-embodiment Dexterous Grasping via Morphology-Prior Diffusion

Sihang Li[1], Zheming Zhou[2], Omid Ghasemalizadeh[2], Luca Carlone[2,3], Min Sun[2], Chen Feng[1], and Cheng-Hao Kuo[2]

[1] New York University, Brooklyn, NY 11201, USA
{sihangli,cfeng}@nyu.edu
[2] Amazon Lab126, Sunnyvale, CA 94089, USA
{zhemiz,ghasemal,xlucacar,minnsun,chkuo}@amazon.com
[3] Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract. This paper presents SOMO, a scalable framework for cross-embodiment grasp synthesis that transfers to novel robot hands using only their hand description (i.e., a Unified Robot Description Format (URDF) file), without requiring any hand–object interaction annotations. Unlike prior approaches that rely on hand-specific models or annotated grasp data for each embodiment, SOMO introduces a shared Morphology-Prior Diffusion model applicable across heterogeneous hands through three key designs. First, grasping is formulated as a 3D assembly problem by predicting per-link $SE(3)$ poses, enabling geometry-driven generation independent of hand-specific kinematics, with feasibility enforced through a post joint optimization stage. Second, a classifier-free training strategy learns a morphology prior from physically valid hand configurations generated through forward-kinematic exploration without object–grasp annotations, enabling grasp synthesis for unseen robot hands given only their URDF files. Third, a 3D shape-aware VAE encodes link and object geometry into a shared embedding, enabling consistent reasoning about hand–object complementarity across embodiments. Experiments show that SOMO achieves state-of-the-art grasp synthesis across six robot hands and demonstrates strong annotation-free generalization to previously unseen hands. Code and models will be released soon.

Keywords: Dexterous Grasping · 3D Assembly · Diffusion Model

1 Introduction

Dexterous grasping is a fundamental capability for achieving precise, human-level manipulation. While individual hands can be addressed by dedicated models, the growing diversity of robotic morphologies demands a cross-embodiment solution. However, efficiently generating high-quality grasps for any robot hand, from a two-fingered gripper to an anthropomorphic hand, remains an open problem.

Existing approaches to cross-embodiment dexterous grasping can be broadly categorized into two paradigms: 1) *Object-oriented methods* [1,5,13,29,37–39] describe grasps in terms of object-surface features such as contact maps [1,35,38,39]

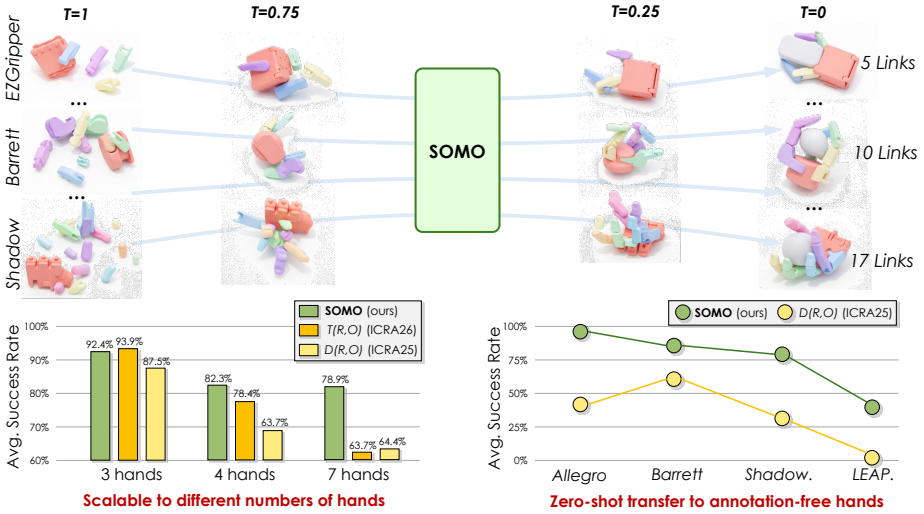

Fig. 1: SOMO formulates cross-embodiment dexterous grasping as **3D assembly**, predicting per-link poses within a unified framework. It scales gracefully to diverse hand morphologies and enables zero-shot transfer to annotation-free hands.

or affordance fields [20, 34]. These approaches naturally promote transferability across different hands and object shapes. However, the reliance on inverse kinematics optimization under penetration-free and joint-limit constraints makes them time-consuming. Moreover, they will suffer from partial point cloud observations in real-world scenarios, and frequently fail to converge for objects with complex geometries, hindering scalability. 2)*Interaction-oriented methods* move beyond pure object-surface descriptions by jointly modeling the spatial relationship between the hand and the object. $\mathcal{D}(\mathcal{R}, \mathcal{O})$ [33] pioneered this direction, constructing a distance matrix between points sampled on the hand and the object surface to embed both entities in a shared interaction space. This formulation retains embodiment awareness while capturing transferable object-level features, facilitating generalization across both hands and objects. However, memory-intensive point-to-point maps in $\mathcal{D}(\mathcal{R}, \mathcal{O})$ hinder scalability. Furthermore, the method remains sensitive to initialization due to its reliance on specific initial configurations. $\mathcal{T}(\mathcal{R}, \mathcal{O})$ [6] addresses these limitations with a heterogeneous graph between robot links and object patches, coupled with a graph diffusion model. While achieving strong results on specific hands, $\mathcal{T}(\mathcal{R}, \mathcal{O})$ is constrained by a fixed node count in its graph representation and relies on hand-crafted features [23] due to weighted-sampled link point clouds. These design choices not only lead to a loss of fine-grained geometric detail for stable grasping but also prevent the model from scaling to morphologically diverse hands.

Considering these limitations, we argue that a truly scalable framework for cross-embodiment dexterous grasping should employ a *unified representation* that naturally accommodates hands with heterogeneous kinematic structures—

varying numbers of joints, link topologies, and coupling constraints—within a single generative model, without any per-hand architectural components. We present SOMO, a Morphology-Prior Diffusion model that realizes scalable cross-embodiment dexterous grasping through three synergistic designs: (i) Grasping as 3D assembly. Inspired by general 3D reassembly task [10,14], we reformulate cross-embodiment dexterous grasping as a conditional 3D assembly task, where SOMO predicts $SE(3)$ poses for a robot's constituent rigid links rather than hand-specific joint angles. This formulation treats a grasp as a collection of $SE(3)$ tokens, allowing a single transformer-based denoiser to handle variable-length inputs, which natively supports diverse morphologies—from simple parallel jaw grippers to complex 21-link hands, as shown in Fig. 1. (ii) Shared 3D shape-aware encoding. A frozen pretrained 3D VAE [15] encodes both object point clouds and individual link meshes into a common geometric latent space using the same encoder weights. This shared space lets the denoiser reason about how a link's geometry complements the local object surface, and maps novel link geometries into the same embedding space as known hands. (iii) Classifier-free training with morphology prior. We extend the classifier-free guidance paradigm [8] beyond its original label-dropping formulation: the unconditional path is trained on a distinct *morphology-prior distribution* that captures physically valid hand poses through forward-kinematic self-exploration, without any object-grasp annotations. At inference, both signals are composed via guided denoising, enabling grasp synthesis for annotation-free hands that participated only in the morphology prior path. In summary, our contributions are:

1. We reformulate cross-embodiment grasp generation as a 3D reassembly problem, predicting per-link $SE(3)$ poses in a unified framework across robot hands with diverse morphologies, while ensuring feasibility and robustness through post-joint optimization.
2. A classifier-free training paradigm learns a morphology prior from forward-kinematic self-exploration of valid hand configurations, enabling annotation-free transfer to novel robots given only a robot hand URDF.
3. A shared 3D shape-aware VAE encodes robot links and objects into a common geometric embedding. Integrated with the proposed formulation and training scheme, SOMO achieves *state-of-the-art* grasp synthesis across six morphologically diverse hands and strong generalization to unseen hands.

2 Related Work

Cross-embodiment Dexterous Grasping. Dexterous grasping requires coordinating multiple fingers to achieve stable object manipulation, yet generalizing across robot hands with diverse kinematics and morphologies remains an open challenge. Object-oriented methods encode grasps through contact representations on object surfaces. Contact map approaches [29, 38] predict dense contact regions, while point-based methods [13, 16, 37] sample discrete interaction locations. Recent works leverage affordance fields [5, 39], geometric primitives [1], and hybrid learning-optimization schemes [9, 20] for cross-embodiment

transfer. However, these methods rely on inverse kinematics optimization under penetration-free and joint-limit constraints, which scales poorly with complex object geometry and is time-consuming. Partial point cloud observations further degrade performance, as occluded contact surfaces lead to unreliable grasp inference. Interaction-oriented methods directly model hand-object spatial relationships, bypassing explicit contact reasoning. $\mathcal{D}(\mathcal{R}, \mathcal{O})$ [33] pioneered this paradigm through point-to-point distance matrices, enabling direct grasp synthesis but incurring quadratic memory costs and initialization sensitivity. $\mathcal{T}(\mathcal{R}, \mathcal{O})$ [6] improves efficiency via graph diffusion with relative $SE(3)$ transformations, yet relies on fixed node structures and hand-crafted BPS features [23]. Critically, both approaches entangle morphology-specific geometry with interaction patterns, necessitating complete retraining for novel kinematic structures—preventing scalable cross-embodiment deployment. A complementary line of work, UniMorphGrasp [36], maps every hand into a single canonical (24-DoF ShadowHand) joint space and uses a per-hand binary mask to deactivate non-existent joints, which requires a manual correspondence between each target hand and the canonical slots and full grasp supervision on every hand. `SOMO` transcends these limitations by reformulating grasping as 3D assembly, operating directly on native URDF-parsed links as variable-length per-link $SE(3)$ tokens, so a new morphology only adds tokens—with no canonical mapping or mask—and can be served zero-shot through the morphology-prior path.Free from fixed graph topologies, `SOMO` scales naturally to hands with arbitrary morphologies.

3D Assembly. The 3D assembly problem seeks to predict transformations that arrange constituent parts into target configurations. Compared with the previous corresponding-based methods [12, 18], recent generative methods employ diffusion / flow matching models on the $SE(3)$ manifold. PuzzleFusion++ [32] and DiffAssemble [28] leverage diffusion with refined noise schedules. GARF [14] explores the flow matching model with specific-designed feature extractor. CRAG [10] proposes a unified model to combine generation model with 3D assembly model. Notably, RPF [30] and Assembler [40] follow a paradigm of first predicting fragment shapes then recovering valid poses through SVD post-optimization. Cross-embodiment grasping shares this essence: valid hand configurations must be recovered under kinematic constraints via inverse kinematics. Both problems fundamentally require constructing plausible transformations on a constrained manifold. Inspired by this connection, `SOMO` casts grasping as assembly, treating robot links parsed from URDFs as parts to arrange around objects. This yields an embodiment-agnostic architecture where each robot's links become $SE(3)$ pose tokens processed identically regardless of their source morphology, followed by an inverse-kinematics optimization [11] for valid configuration.

Classifier-Free Guidance. Classifier-free guidance (CFG) [8] enables conditional diffusion models to learn both conditional and unconditional distributions by randomly dropping conditioning during training. It has been widely adopted in text-to-image synthesis [24, 26, 27], 3D generation [22], human motion synthesis [31], and robot manipulation [25], where the unconditional path models underlying constraints. `SOMO` reinterprets CFG by replacing the data-derived

unconditional path with a *morphology prior* learned through forward-kinematic self-exploration. By sampling random joint configurations within URDF limits and learning to denoise the resulting link poses, our unconditional branch captures what valid hand configurations look like—without any grasp annotations. This transforms CFG into a mechanism for zero-shot cross-embodiment transfer: new hands participate only in morphology learning during training, yet benefit from object-conditional knowledge learned by other embodiments at inference.

3 SOMO Framework

Given an object point cloud and a robot embodiment specified by its URDF, SOMO generates per-link $SE(3)$ grasping poses through a diffusion model shared across all hands. We describe our approach in three parts: Sec. 3.1 introduces the overall architecture that reformulates grasping as 3D assembly of rigid links; Sec. 3.2 presents a shared 3D VAE that encodes both object surfaces and link morphologies into a common geometric space; Sec. 3.3 details the classifier-free training paradigm that learns a morphology prior from valid hand configurations, enabling zero-shot transfer to novel robots.

3.1 Morphology-Prior Diffusion

Overview. A key barrier to cross-embodiment grasp generation is that different hands have different kinematic structures: varying numbers of joints, link lengths, and coupling constraints. We sidestep this entirely by reformulating grasping as a 3D assembly problem. Given a robot hand with embodiment $e \in \mathcal{E}$, we parse its URDF into L_e rigid links and seek to predict an SE(3) pose $T_l \in \mathbb{R}^6$ for every link $l = 1, \ldots, L_e$, where each $T_l = [\mathbf{t}_l, \mathbf{r}_l]$ consists of a translation $\mathbf{t}_l \in \mathbb{R}^3$ and an axis-angle $\mathbf{r}_l \in \mathbb{R}^3$ representing rotation. A valid grasp is then an arrangement of these rigid pieces around the object—akin to assembling a 3D puzzle—and the hand-specific kinematics is only consulted after generation, via inverse kinematics, to recover joint angles. Because the generative model never sees joint angles, it operates in a representation that is shared across all embodiments by construction.

Link tokenization. A key limitation of $\mathcal{T}(\mathcal{R}, \mathcal{O})$ [6] is its weighted point cloud sampling strategy, which allocates points proportionally to link surface area. This under-represents geometrically small but functionally critical links such as fingertips, yielding suboptimal features for fine-grained grasping. SOMO instead adopts *uniform sampling*: for each link l in the URDF, we sample K points with normals from its mesh, ensuring every link—regardless of size—receives equal representational capacity. Each link's point cloud is then passed through a frozen pretrained 3D feature extractor Φ (Sec. 3.2) to obtain a shape-aware latent, which a learnable two-layer MLP ϕ_{link} projects into the link embedding $\mathbf{e}_l \in \mathbb{R}^{d_{\text{link}}}$. Since link meshes are fixed for a given robot, all $\mathbf{e}_l$ are pre-computed once at initialization and cached, incurring no overhead during training or inference. At

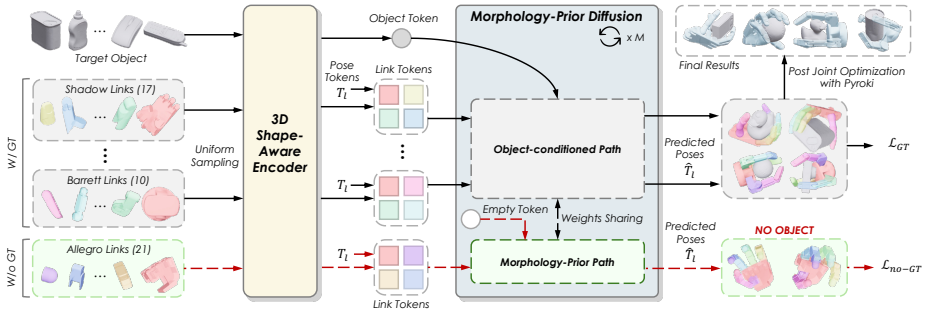

Fig. 2: SOMO framework overview. The Morphology-Prior Diffusion consists of object-conditioned path and Morphology-Prior path, the latter of which (red dashed line) is only activated in the training phase, while can enable zero-shot transfer.

each denoising step m, the robot token for link l is constructed as

$$\mathbf{V}_R^{(m)}[l] \;=\; \big[\,\tilde{\mathbf{t}}_l^{(m)},\; \tilde{\mathbf{r}}_l^{(m)},\; \mathbf{e}_l\,\big], \tag{1}$$

where $\tilde{\mathbf{t}}_l^{(m)}$ and $\tilde{\mathbf{r}}_l^{(m)}$ are the noisy translation and axis-angle rotation at step m, encoding the current spatial estimate of link l from which the denoiser predicts noise for the next step. The translation is normalized to the object-centric frame via $\tilde{\mathbf{t}}_l = (\mathbf{t}_l - \mathbf{c}_O)/s_O$ (with centroid $\mathbf{c}_O$ and scale s_O). The morphology embedding $\mathbf{e}_l$ remains fixed across all steps, serving as a persistent shape prior.

Object tokenization. The same frozen feature extractor Φ encodes the object point cloud $\mathcal{O} = \{\mathbf{x}_i\}_{i=1}^{N}$ with surface normals into P patch tokens that capture local surface geometry to get the $\mathbf{V}_O \in \mathbb{R}^{P\times d_O}$. (details in Sec. 3.2).

Denoiser. The denoiser f_θ first projects the raw tokens $\mathbf{V}_R^{(m)}$ and $\mathbf{V}_O$ into a shared feature dimension d via learned linear layers, then passes them through N_{layer} transformer layers. We overload $\mathbf{V}_R^{(i)}$ to also denote the robot features at layer i (with $\mathbf{V}_R^{(0)}$ being the projected input). Each layer is modulated by AdaLN-Zero [21] conditioning on the diffusion timestep m and computes:

$$\mathbf{V}'_R \;=\; \mathbf{V}_R^{(i)} \;+\; \mathbf{g}_1 \odot \mathrm{SelfAttn}\big(\mathrm{AdaLN}(\mathbf{V}_R^{(i)},\, \mathbf{c}_m)\big), \tag{2}$$

$$\mathbf{V}''_R \;=\; \mathbf{V}'_R \;+\; \mathbf{g}_2 \odot \mathrm{CrossAttn}\big(\mathrm{AdaLN}(\mathbf{V}'_R,\, \mathbf{c}_m),\; \mathbf{V}_O\big), \tag{3}$$

$$\mathbf{V}_R^{(i+1)} \;=\; \mathbf{V}''_R \;+\; \mathbf{g}_3 \odot \mathrm{FFN}\big(\mathrm{AdaLN}(\mathbf{V}''_R,\, \mathbf{c}_m)\big), \tag{4}$$

where $\mathbf{c}_m$ is the timestep embedding and $\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3$ are learned gates produced by AdaLN-Zero from $\mathbf{c}_m$. Self-attention (Eq. 2) enables inter-link spatial reasoning, allowing each link to attend to all other links and capture their coordinated spatial relationships. Cross-attention (Eq. 3) is the sole pathway through which object information enters the denoiser: robot tokens query the object features to ground generation in the target geometry. Crucially, the cross-attention sub-block can be selectively *disabled* by zeroing $\mathbf{V}_O$ and skipping Eq. 3, enabling classifier-free guidance over the object condition (Sec. 3.3). Two separate MLP

heads decode the final aggregated features into translation noise $\hat{\boldsymbol{\epsilon}}^{\mathbf{t}}_{\theta,l}$ and rotation noise $\hat{\boldsymbol{\epsilon}}^{\mathbf{r}}_{\theta,l}$, reflecting the distinct statistics of translational and rotational perturbations.

Forward diffusion. We apply a standard DDPM forward process [7] only to the pose components $\mathbf{p}_l = [\tilde{\mathbf{t}}_l, \tilde{\mathbf{r}}_l] \in \mathbb{R}^6$; the morphology embedding $\mathbf{e}_l$ is concatenated unchanged at every step, serving as a persistent shape prior. Given the clean pose $\mathbf{p}_l^{(0)}$ from ground-truth grasp data, the noisy pose at step m is

$$\mathbf{p}_l^{(m)} \;=\; \sqrt{\bar{\alpha}_m}\,\mathbf{p}_l^{(0)} \;+\; \sqrt{1-\bar{\alpha}_m}\,\boldsymbol{\epsilon}_l, \qquad \boldsymbol{\epsilon}_l \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

where $\bar{\alpha}_m = \prod_{j=1}^{m} \alpha_j$ is the cumulative noise schedule. The noisy pose $\mathbf{p}_l^{(m)}$ is then concatenated with $\mathbf{e}_l$ to form the robot token $\mathbf{V}_R^{(m)}[l] = [\mathbf{p}_l^{(m)}, \mathbf{e}_l]$ (Eq. 1), which is fed into the denoiser f_θ to predict the added noise $\hat{\boldsymbol{\epsilon}}_{\theta,l} = [\hat{\boldsymbol{\epsilon}}^{\mathbf{t}}_{\theta,l}, \hat{\boldsymbol{\epsilon}}^{\mathbf{r}}_{\theta,l}]$.

Training objective. Since different embodiments have different numbers of links, the loss is computed only over valid links. For hands with ground-truth grasp annotations, the object-conditioned loss is:

$$\mathcal{L}_{\text{GT}} \;=\; \mathbb{E}_{m,\boldsymbol{\epsilon}}\left[\frac{1}{L_e}\sum_{l=1}^{L_e}\left\|\boldsymbol{\epsilon}_l - \hat{\boldsymbol{\epsilon}}_{\theta,l}\right\|^2\right], \tag{6}$$

where L_e is the number of links for embodiment e. This loss trains the model to predict grasp-relevant noise conditioned on the object. The complete training objective, which additionally incorporates a morphology prior loss $\mathcal{L}_{\text{no-GT}}$ for annotation-free embodiments, is detailed in Sec. 3.3.

3.2 3D Shape-Aware Encoding

The assembly formulation (Sec. 3.1) treats grasping as arranging robot links around an object, analogous to assembling parts around a fixed anchor part [14, 30]. For the denoiser's cross-attention to meaningfully relate link geometry to the object surface, both must reside in a *shared geometric feature space*. This requirement is addressed by employing a single frozen encoder for both domains.

Concretely, the encoder Φ from the pretrained 3D VAE of TripoSG [15] is adopted, a large-scale shape generation model trained on millions of 3D assets. TripoSG's VAE encodes point clouds with surface normals into compact latent tokens via frequency-based positional embedding and farthest-point-sampled attention, capturing rich geometric priors. The object point cloud $\mathcal{O}$ with normals is first normalized to the object-centric frame, then encoded by Φ into P patch tokens. The encoded latent is used directly rather than decoded features, as we find it provides better alignment with link embeddings in the shared space. The normalization scale s_O is concatenated to preserve absolute size, yielding $\mathbf{V}_O \in \mathbb{R}^{P\times d_O}$. Each link l's mesh point cloud (K uniformly sampled points with normals) is normalized to a unit ball and encoded by the *same* Φ into a single-token latent, which a learnable MLP ϕ_{link} projects into $\mathbf{e}_l \in \mathbb{R}^{d_{\text{link}}}$. This design reasons not only about where to place each link but about how that link's shape complements the local object surface. The only learned parameters are those of ϕ_{link}, keeping parameter-efficient while inheriting large-scale 3D priors.

3.3 Classifier-Free Training Paradigm

Existing cross-embodiment grasp methods assume access to paired object–grasp annotations for every hand, a requirement that fundamentally limits scalability. Curating such datasets demands extensive simulation or teleoperation for each new embodiment, creating a bottleneck that grows linearly with the number of supported hands. This raises the question: can a model generate grasps for a novel robot given only its URDF and link meshes, without observing any grasp demonstrations? This work addresses this question by introducing a classifier-free training paradigm [8] that decomposes grasp generation into two complementary learning signals: an object-conditioned grasp distribution trained on annotated data, and a morphology prior learned purely from kinematic self-exploration. Both signals are unified within a single diffusion model, enabling `SOMO` to synthesize grasps for unseen robot hands at inference time.

Object-conditioned path. For hands with available grasp annotations, training follows the conditional diffusion objective. Given a paired tuple $(\mathcal{O}, \mathbf{p}^{(0)}, e)$ of object point cloud, ground-truth link poses, and embodiment, the model learns the conditional distribution $p(\mathbf{p} \mid \mathcal{O}, e)$ by predicting the noise added to $\mathbf{p}^{(0)}$ (Eq. 6). Object tokens $\mathbf{V}_O$ are encoded from $\mathcal{O}$ and attend to robot tokens through cross-attention (Eq. 3). To support classifier-free guidance, we stochastically drop the object condition with probability p_{uncond}: the object tokens are replaced with zeros ($\mathbf{V}_O \leftarrow \mathbf{0}$) and cross-attention is skipped, exposing the model to the unconditional setting even for annotated hands.

Morphology-prior path. For hands without any grasp annotations, we construct training data by sampling from the hand's configuration manifold. We first sample a single open/close blend factor $\alpha \sim (0.1, 0.9)$ shared across all joints to form a coordinated base hand pose $\mathbf{q}_{\text{base}}$, then inject per-joint variability by Gaussian sampling clamped within the URDF joint limits $[\mathbf{q}_{\min}, \mathbf{q}_{\max}]$. Combined with a random root orientation in $[-\pi, \pi]^3$, FK then yields the per-link $SE(3)$ targets used in the Morphology-prior path. Validity is enforced via per-joint URDF limit clamping and the coordinated blend, without explicit self-collision filtering; the prior thus captures kinematically reachable configurations, while object conditioning and post-hoc IK resolve residual penetrations.Since no object is involved, object tokens are zeroed ($\mathbf{V}_O = \mathbf{0}$) and cross-attention is disabled throughout. The diffusion objective remains identical, predicting noise added to valid link poses, but the learning signal is fundamentally different. Rather than learning where to place links relative to an object, the model learns what physically realizable configurations look like for a given morphology. The self-attention layers capture kinematic coupling between links, i.e., the coordinated spatial relationships that distinguish valid hand poses from arbitrary link arrangements. In effect, this path teaches the model hand morphology.

Unified training. Both paths share the same noise prediction loss and are trained jointly. $\mathcal{L}_{\text{GT}}$ (Eq. 6) denotes the loss over annotated data with stochastic condition dropout, and $\mathcal{L}_{\text{no-GT}}$ the loss over FK-sampled configurations (computed identically but without object conditioning). The overall objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GT}} + \lambda \mathcal{L}_{\text{no-GT}}, \tag{7}$$

Table 1: Number of grasp demonstrations per embodiment.

	Allegro	Barrett	Shadow	Leaphand	Ezgripper	Robotiq_3finger	XHand
Base	4,812	7,834	5,181	7,800	6,073	864	37,545
Extended	31,084	198,041	27,098	7,800	49,818	6,727	37,545

where λ controls the relative weight of the morphology prior. Because the denoiser architecture, diffusion schedule, and loss form are shared, `SOMO` builds a unified representation, where annotated and unannotated embodiments coexist.

Classifier-free guided inference. At test time, the learned morphology prior is composed with object conditioning via two forward passes of f_θ per step:

$$\hat{\boldsymbol{\epsilon}}_{\text{cfg}} = (1-s)\,\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{0},\, \mathbf{p}^{(m)},\, m) \;+\; s\cdot\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{V}_O,\, \mathbf{p}^{(m)},\, m), \tag{8}$$

where s is the guidance scale, and the two arguments denote the object condition and the current noisy poses respectively. The unconditional pass provides the morphology prior—what valid configurations look like—while the conditional pass steers toward the specific object. When $s > 1$, guidance amplifies the object-specific component, producing tighter grasps while remaining anchored to physically valid poses.

Zero-shot transfer. For a novel robot unseen during annotated training, the classifier-free formulation provides a natural generalization mechanism. The new hand participates only in the morphology prior path, learning valid configurations through random FK sampling. At inference, the cross-attention mechanism—trained across annotated embodiments to map object geometry to link placement—generalizes to the new hand via three factors: (i) the shared encoder Φ maps novel link geometries into the same embedding space as known hands; (ii) self-attention layers, trained across diverse topologies, have learned embodiment-agnostic inter-link coordination; (iii) cross-attention layers operate over the unified geometric space and transfer the learned object-to-link mapping. The morphology prior ensures generated poses respect the new hand's kinematic constraints, while object conditioning produces functional grasps—achieving synthesis without a single annotated demonstration.

Post joint optimization. Since the denoiser operates in SE(3) pose space, we recover executable joint angles by solving the same IK formulation as $\mathcal{T}(\mathcal{R}, \mathcal{O})$ [6]: $\mathbf{q}^* = \arg\min_{\mathbf{q}} \sum_l \|\text{FK}_l(\mathbf{q}) - \hat{T}_l\|^2$ subject to URDF joint limits with Pyroki [11].

4 Experiments

4.1 Implementation Details

The denoiser is an 8-layer transformer (N_{layer}=8) with hidden dimension d=384, conditioned on sinusoidal timestep embedding via AdaLN-Zero. The diffusion process uses 1000 linear-scheduled noise steps ($\beta \in [10^{-4}, 0.02]$) for training; inference runs 100-step DDIM with η=1.0. Link embeddings use the frozen TripoSG encoder with d_{vae}=64 projected to d_{link}=128 via a two-layer MLP, with

Fig. 3: Qualitative results on unseen validation object. The first two rows show hands trained with full grasp annotations (rendered in gray); the third row shows annotation-free hands trained solely through the morphology prior path (rendered in green).

K=512 uniformly sampled points per link. The classifier-free dropout probability is p_{uncond}=0.1, and the guidance scale at inference is s=1.25. Training uses Adam (lr = 10^{-4}, StepLR decay γ=0.8 every 20 epochs) with gradient clipping at 1.0. For the object-conditioned baseline (without the morphology prior path), we train for 400 epochs with batch size 64. For the full classifier-free model, we train for 100k global steps. All training is conducted on 8× NVIDIA L40S GPUs.

4.2 Experiment Setup

Dataset. We build our training data upon the CMapDataset [13], a collection of simulated dexterous grasps across diverse hand embodiments. After quality filtering, the base dataset spans seven robot hands and is further augmented following $\mathcal{D}(\mathcal{R}, \mathcal{O})$ [33]. Table 1 summarizes the per-embodiment grasp counts before and after augmentation.

All grasps are distributed across 48 training objects sourced from ContactDB [2] and YCB [3]; evaluation is performed on 10 held-out objects unseen during training. For each object, we pre-sample a large number of surface points from the mesh and randomly draw 512 points with normals per training iteration. At test time, 100 grasps are generated per object for each embodiment.

Evaluation metrics. We follow the same evaluation protocol as $\mathcal{D}(\mathcal{R}, \mathcal{O})$ [33]: (i) Success rate: each generated grasp is executed in Isaac Gym [19]; a grasp is deemed successful if the object displacement remains below 2 cm after six directional perturbation forces are applied. (ii) Efficiency: wall-clock time per grasp, measured end-to-end including denoising and post joint optimization. (iii) Grasp diversity: standard deviation of joint values across successful grasps.

Baselines. Four methods spanning two paradigms are used for comparison. DFC [17] and GenDexGrasp [13] recover grasps via contact map; we provide their evaluations in the supplementary material as their efficiency and success rates are not on par with recent generative benchmarks. $\mathcal{T}(\mathcal{R}, \mathcal{O})$ [6] and $\mathcal{D}(\mathcal{R}, \mathcal{O})$ [33] are the most relevant baselines as they directly synthesize grasps via learned gen-

Table 2: Cross-embodiment grasp synthesis on 10 unseen objects. Suc. (%) denotes success rate and Eff. (sec./grasp) denotes wall-clock time per grasp. Each group progressively adds hands to the training pool. Best results per group are in **bold**, second best underlined.

Method	Allegro		Barrett		Shadow		Avg.	
	Suc.↑	Eff.↓	Suc.↑	Eff.↓	Suc.↑	Eff.↓	Suc.↑	Eff.↓
$\mathcal{D}(\mathcal{R},\mathcal{O})$ [33]	92.10	0.19	87.50	0.42	83.00	0.78	87.53	0.46
†CEDex [35]	88.1	0.08	93.1	0.08	85.00	0.07	88.73	0.08
*$\mathcal{T}(\mathcal{R},\mathcal{O})$ [6]	93.50	0.03	**91.90**	0.03	**94.49**	0.03	**93.90**	0.03
SOMO (ours)	**94.80**	**0.02**	88.80	**0.02**	93.50	**0.01**	92.37	**0.02**

Method	Allegro		Barrett		Shadow		Leap		Avg.	
	Suc.↑	Eff.↓	Suc.↑	Eff.↓	Suc.↑	Eff.↓	Suc.↑	Eff.↓	Suc.↑	Eff.↓
$\mathcal{D}(\mathcal{R},\mathcal{O})$ [33]	89.90	0.20	85.00	0.40	83.70	0.73	20.10	0.34	69.69	0.42
$\mathcal{T}(\mathcal{R},\mathcal{O})$ [6]	**97.20**	0.03	86.20	0.03	**94.00**	0.03	36.30	0.03	78.42	0.03
SOMO (ours)	94.30	**0.02**	**90.00**	**0.01**	91.30	**0.02**	**55.40**	**0.02**	**82.75**	**0.02**

Method	Allegro			Barrett			Shadow			Leap			Robotiq.			EZGrip.			Avg.		
	Suc.↑	Div.↑	Eff.↓	Suc.↑	Div.↑	Eff.↓	Suc.↑	Div.↑	Eff.↓	Suc.↑	Div.↑	Eff.↓	Suc.↑	Div.↑	Eff.↓	Suc.↑	Div.↑	Eff.↓	Suc.↑	Div.↑	Eff.↓
$\mathcal{D}(\mathcal{R},\mathcal{O})$ [33]	92.70	**0.41**	0.21	87.20	0.49	0.41	73.10	**0.42**	0.72	42.00	0.45	0.34	12.90	**0.54**	0.30	78.80	0.59	0.32	64.45	**0.48**	0.39
$\mathcal{T}(\mathcal{R},\mathcal{O})$ [6]	94.40	0.37	0.03	91.30	0.52	0.03	89.70	0.29	0.03	59.00	0.39	0.03	8.60	0.44	0.04	39.30	0.56	0.04	63.72	0.43	0.03
SOMO (ours)	**97.60**	0.39	**0.02**	**92.80**	**0.52**	**0.01**	**91.00**	0.31	**0.02**	**65.30**	**0.46**	**0.02**	**29.70**	0.51	**0.01**	**97.20**	**0.63**	**0.01**	**78.93**	0.47	**0.02**
$\mathcal{D}(\mathcal{R},\mathcal{O})$ [33]	89.30	0.38	0.22	85.50	0.48	0.40	78.30	**0.41**	0.73	40.77	0.45	0.34	10.10	0.49	0.30	79.50	0.59	0.31	63.91	0.47	0.38
$\mathcal{T}(\mathcal{R},\mathcal{O})$ [6]	**96.00**	0.37	0.03	84.60	0.58	0.03	88.00	0.31	0.03	62.00	0.40	0.03	10.60	0.47	0.04	54.00	0.60	0.03	65.87	0.45	0.03
SOMO (ours)	91.20	**0.38**	**0.02**	**90.80**	**0.58**	**0.01**	**94.80**	0.31	**0.02**	**67.70**	**0.47**	**0.02**	**44.20**	**0.49**	**0.02**	**97.40**	**0.63**	**0.01**	**81.02**	**0.48**	**0.02**

*$\mathcal{T}(\mathcal{R},\mathcal{O})$ reproduced with official code; avg. Suc. within 1% of the originally reported 94.83%.
†CEDex's code was released recently; at the time of writing, we report results on the three hands.

erative models. $\mathcal{D}(\mathcal{R},\mathcal{O})$ includes a Configuration-Invariant Pretraining (CIP) stage that learns embodiment-agnostic features from hand point clouds without requiring object-grasp annotations. The CIP stage is fully adopted in the reproduction. For fair comparison on novel (annotation-free) hands, both $\mathcal{D}(\mathcal{R},\mathcal{O})$ and SOMO are given access to the same information. $\mathcal{D}(\mathcal{R},\mathcal{O})$ includes the novel hand in its CIP pretraining, while SOMO includes it in the morphology-prior path; neither method observes any grasp annotations for the novel embodiment.

4.3 Evaluation

We conduct extensive experiments to address three main questions of our work:
Q1: Does SOMO scale favorably as the set of co-trained embodiments diversifies in morphology and the volume of training data grows?
Q2: Can the morphology prior in classifier-free guidance learning enable effective grasp synthesis for a hand trained without any grasp annotations?
Q3: Does encoding link and object features into a shared geometric space via the 3D shape-aware encoder improve generalization in SOMO?

Scaling with Embodiments and Data. Table 2 reports grasp success rate, diversity, and efficiency across four progressive settings. The uncolored three groups are trained on the base dataset; the shaded group is trained on the extended dataset (including XHand; see Table 1 for per-embodiment statistics). XHand is included in training to enrich morphological diversity but omitted from the main quantitative comparison because of the low quality of its grasp annotations: of its 37,545 GT grasps, only 284 (∼0.7%) pass an Isaac Gym stability replay. Trained on this filtered valid subset jointly with the other six hands, SOMO attains 15.80% success / 51.2 diversity on XHand with no architecture or IK change—its difficulty stems from annotation quality rather than underactuated kinematics.

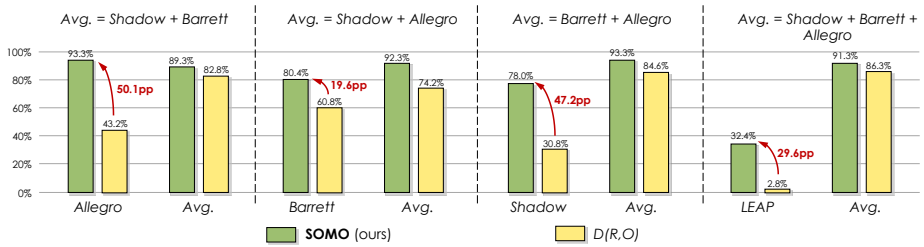

Fig. 4: Classifier-free morphology prior evaluation. Each group holds out one hand from grasp annotations, training it only through the morphology prior path. Avg. is computed over the remaining hands with GT.

Scaling with embodiments. The three groups progressively add hands from 3 to 4 to 7, revealing how each method copes with increasing morphological diversity. With 3 hands, all generative methods achieve strong results; SOMO reaches 92.37% average success, marginally behind $\mathcal{T}(\mathcal{R}, \mathcal{O})$ (93.90%). As the pool grows to 4 hands, baselines degrade sharply ($\mathcal{D}(\mathcal{R}, \mathcal{O})$: 87.53% → 69.69%; $\mathcal{T}(\mathcal{R}, \mathcal{O})$: 93.90% → 78.42%) while SOMO sustains 82.75%, with a much higher LeapHand success of 55.40%. Scaling to 7 hands widens the gap: both baselines fall below 65% (64.45% and 63.72%), whereas SOMO maintains 78.93%—over 15 pp above the best baseline, with 97.20% on EZGripper and 65.30% on LeapHand. In total, SOMO's average success declines by only 13.4 pp from 3 to 7 hands, more than 2× more resilient than the baselines (23.1 pp and 30.2 pp), confirming that it shares grasping knowledge across morphologies rather than suffering inter-embodiment interference.

Scaling with training data. The shaded group trains all 7 hands on the extended dataset, which augments training data via the MultiDex pipeline and additionally incorporates XHand demonstrations (Table 1). This lifts SOMO from 78.93% to 81.02%, with notable gains on hard hands like Robotiq (29.70% → 44.20%), while the baselines remain largely unchanged (64.45% → 63.91% and 63.72% → 65.87%). SOMO thus absorbs additional data and converts it into improved grasp quality—a prerequisite for continued scaling—and generates a grasp in 0.02 s, 10–40× faster than $\mathcal{D}(\mathcal{R}, \mathcal{O})$ (0.19–0.78 s).

A1: *More hands and more data consistently help SOMO, turning cross-embodiment diversity into a scaling advantage rather than an interference.*

Zero-shot Transfer with Morphology-Prior. To isolate the effect of the morphology prior, we conduct a leave-one-out experiment: in each group of Fig. 4, one hand is withheld from all grasp annotations and trained exclusively through the morphology prior path, while the remaining hands retain full supervision. For $\mathcal{D}(\mathcal{R}, \mathcal{O})$, the held-out hand participates in CIP paradigm under the same fair-comparison protocol described in Sec.4.2.

Table 3: Ablation studies. (a) Architectural components. All baseline and variant models are trained on base dataset. Avg. over the 4 supervised hands (Allegro, Barrett, ShadowHand and LEAPHand). (b) Guidance scale s at inference; Barrett (green) has no grasp annotations; Avg. over ShadowHand and Allegro.

(a)

Setup	Trainable Params	Var. Len.	Link Enc.	Obj. Enc.	Avg. Suc.↑	Avg. Div.↑	Avg. Eff.↓
$\mathcal{T}(\mathcal{R},\mathcal{O})$	28.06M	×	BPS	VQVAE	78.42	0.41	0.03
I	23.80M	×	VAE	VAE	80.58	0.41	0.01
II	23.80M	✓	BPS	VQVAE	79.38	0.40	0.02
III	23.80M	✓	VAE	VQVAE	81.63	0.42	0.01
SOMO	23.80M	✓	VAE	VAE	**82.75**	**0.42**	**0.01**

(b)

s	Barrett Suc.↑	Barrett Div.↑	Avg. Suc.↑	Avg. Div.↑
0.75	80.8	0.56	92.4	0.32
1.0	81.4	0.54	**94.2**	0.33
1.25	**83.1**	0.54	93.6	0.32
1.5	81.7	0.55	92.8	0.33
2.0	79.9	0.56	89.9	0.34

On annotation-free hands , SOMO consistently outperforms $\mathcal{D}(\mathcal{R},\mathcal{O})$ by a large margin: 93.30% vs. 43.20% on Allegro, 80.40% vs. 60.80% on Barrett, 78.00% vs. 30.80% on ShadowHand, and 32.40% vs. 2.80% on LeapHand—an average improvement of over 36 pp.Notably, SOMO's annotation-free results on Allegro (93.30%) and Barrett (80.40%) approach their fully supervised counterparts in Table 2 (94.80% and 88.80%), showing that the morphology prior captures enough kinematic structure for cross-attention to generalize object-to-link mappings, while supervised hands within each group stay strong (avg. Suc. >88%). Qualitative results are in Fig. 3.

A2: *The morphology prior enables annotation-free hands to achieve stable grasps, substantially reducing the need for per-embodiment annotations.*

Ablation Study. Table 3(a) progressively introduces SOMO's design choices starting from $\mathcal{T}(\mathcal{R},\mathcal{O})$. Replacing the fixed-topology graph with variable-length tokens yields a modest gain (+0.96 pp). Another critical factor is the shared 3D encoder: progressively replacing BPS link encoding with VAE (II → III, +2.25 pp) and then VQVAE object encoding with VAE (III → SOMO, +1.12 pp) yields a cumulative +3.37 pp, confirming that unifying both domains in a shared geometric space is important. The full SOMO reaches 82.75% at 0.01 s per grasp with only 23.80M parameters, surpassing $\mathcal{T}(\mathcal{R},\mathcal{O})$ in success rate, efficiency, and parameter count. Table 3(b) varies the guidance scale s at inference on a single trained model (Allegro and ShadowHand as GT, Barrett as annotation-free). GT hands peak at s=1.0 (94.2%) and degrade with stronger guidance, as their conditional predictions are already accurate. The annotation-free Barrett peaks at s=1.25 (83.1%): Its conditional signal, transferred from other hands, is directionally correct but less confident, so moderate amplification helps while the morphology prior anchors physical validity. Beyond s=1.5, both groups degrade.

Table 4: IK failure breakdown per embodiment (1000 samples/hand). **PF**: Pyroki failure (the 64-iteration Levenberg–Marquardt budget is exhausted); **DF**: diffusion failure (IK converges within tolerance but the grasp fails the Isaac Gym test). PF/(PF+DF) is the share of failures attributable to IK.

Hand	Allegro	Barrett	Shadow	LEAP	Robotiq	EZGripper	XHand
PF	1	0	0	38	5	0	43
DF	44	90	76	347	788	26	842
PF/(PF+DF)	2.2%	0.0%	0.0%	10.9%	0.6%	0.0%	5.1%

Table 5: Robustness to partial observations. Testing use one-side 50% partial point clouds for all objects. Best results are in **bold**, second best underlined.

Method	Allegro		Barrett		Shadow		Leap		Robotiq.		EZGrip.		Avg.	
	Suc.↑	Eff.↓	Suc.↑	Eff.↓	Suc.↑	Eff.↓	Suc.↑	Eff.↓	Suc.↑	Eff.↓	Suc.↑	Eff.↓	Suc.↑	Eff.↓
$\mathcal{D}(\mathcal{R},\mathcal{O})$ [33]	81.30	0.03	82.50	0.03	71.70	0.03	50.00	0.03	9.30	0.04	37.80	0.04	55.43	0.03
$\mathcal{T}(\mathcal{R},\mathcal{O})$ [6]	86.50	0.21	84.80	0.40	63.30	0.72	30.80	0.34	4.60	0.30	65.10	0.31	55.85	0.38
SOMO (ours)	**89.50**	**0.02**	**88.20**	**0.01**	**84.20**	**0.02**	**55.90**	**0.02**	**20.30**	**0.02**	**86.10**	**0.02**	**70.70**	**0.02**

A3: *The shared 3D encoder consistently improves generalization. The morphology prior integrates seamlessly with object-conditioned training without mutual interference, enabling zero-shot transfer for cross-embodiments.*

Robustness on Partial Observation. Real-world depth sensors produce incomplete point clouds due to self-occlusion and limited viewpoints. Testing all methods on single-view 50% partial point clouds (Table 5), SOMO achieves 70.70% average success, a lead of roughly 15 pp over both baselines.

IK failure Analysis. Every reported success rate is measured through the full generation → Pyroki IK → Isaac pipeline, with *all* IK failures included. Pyroki runs up to 64 Levenberg–Marquardt iterations with default jaxls tolerances (cost 10^{-5}, gradient 10^{-4}, parameter 10^{-6}). Rerunning the full evaluation (1000 samples/hand), we label failures that exhausted the 64-LM budget as a **Pyroki failure (PF)** and those that converged but failed the Sim test as a **diffusion failure (DF)** (Table 4). Pyroki contributes very few failures across all seven hands; the slight elevation on LEAP (10.9%) stems from dense intra-finger coupling but remains a minority of its failures. IK is therefore *not* the bottleneck—major failures come from the diffusion emitting kinematically realizable but physically suboptimal $SE(3)$ targets.

Real-World Demonstrations. To assess whether SOMO grasps transfer to the real world, we deploy on a UR10e arm with a LeapHand and a fixed-view Intel RealSense D435i depth camera (Fig. 5).Given an RGB-D observation, SOMO

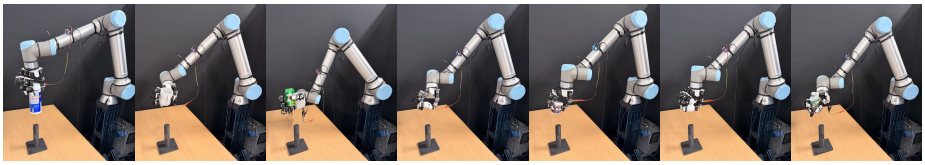

Fig. 5: Real-world grasping with SOMO on a UR10e + LEAPHand platform.

synthesizes a grasp pose relative to the object, which is transformed into the arm base frame; a collision-free trajectory is then planned via MoveIt [4] to the pre-grasp pose, after which the hand closes to execute the grasp.

5 Limitations and Conclusion

While SOMO demonstrates strong scalability and zero-shot transfer, several limitations remain. First, our evaluation focuses on isolated objects; cluttered scenes with occlusions and inter-object contacts pose additional challenges for pose estimation and collision-free planning. Second, the post-optimization IK step can fail for highly underactuated hands, where the $SE(3)$-pose-to-joint mapping becomes ill-conditioned, which may require tighter coupling between grasp generation and kinematic feasibility. Third, performance is sensitive to native URDF link granularity: finer decomposition makes the denoising target harder to fit and adds assignment ambiguity for interchangeable links (*e.g.*, identical fingertips) [30], and adds IK residual terms.

We presented SOMO, which casts cross-embodiment dexterous grasping as 3D assembly of per-link SE(3) poses, pairing a frozen TripoSG encoder with classifier-free guidance so annotation-free hands acquire a morphology prior from forward-kinematic self-exploration alone. It degrades by only 13.4 pp from 3 to 7 hands—whereas the strongest baseline ($\mathcal{T}(\mathcal{R}, \mathcal{O})$) drops over 30 pp—and transfers zero-shot to unseen hands. Real-world trials (80% over 10 objects) give preliminary evidence of sim-to-real feasibility, with comprehensive evaluation left to future work.

Acknowledgements. Chen Feng was supported by his NSF grant No. 2238968.

References

1. Attarian, M., Asif, M.A., Liu, J., Hari, R., Garg, A., Gilitschenski, I., Tompson, J.: Geometry matching for multi-embodiment grasping. In: Conference on Robot Learning. pp. 1242–1256. PMLR (2023) i, iii
2. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: Contactdb: Analyzing and predicting grasp contact via thermal imaging. In: CVPR. pp. 8709–8719 (2019) x
3. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: 2015 international conference on advanced robotics (ICAR). pp. 510–517. IEEE (2015) x

4. Chitta, S., Sucan, I., Cousins, S.: Moveit![ros topics]. IEEE robotics & automation magazine **19**(1), 18–19 (2012) xv, i
5. 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. arXiv preprint arXiv:2502.16420 (2025) i, iii
6. Fei, X., Xu, Z., Fang, H., Zhang, T., Shao, L.: T (r, o) grasp: Efficient graph diffusion of robot-object spatial transformation for cross-embodiment dexterous grasping. arXiv preprint arXiv:2510.12724 (2025) ii, iv, v, ix, x, xi, xiv
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020) vii
8. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) iii, iv, viii
9. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: ICCV (2021) iii
10. Jiang, Z., Li, S., Tan, S., Xu, C., Zhang, J., Galway-Witham, J., Wang, X., Williams, S.A., Iovita, R., Feng, C., et al.: Crag: Can 3d generative models help 3d assembly? arXiv preprint arXiv:2602.22629 (2026) iii, iv
11. Kim*, C.M., Yi*, B., Choi, H., Ma, Y., Goldberg, K., Kanazawa, A.: Pyroki: A modular toolkit for robot kinematic optimization (2025) iv, ix
12. Lee, N., Min, J., Lee, J., Park, C., Cho, M.: Combinative matching for geometric shape assembly. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9540–9549 (2025) iv
13. Li, P., Liu, T., Li, Y., Geng, Y., Zhu, Y., Yang, Y., Huang, S.: Gendexgrasp: Generalizable dexterous grasping. In: ICRA. pp. 8068–8074. IEEE (2023) i, iii, x
14. Li, S., Jiang, Z., Chen, G., Xu, C., Tan, S., Wang, X., Fang, I., Zyskowski, K., McPherron, S.P., Iovita, R., Feng, C., Zhang, J.: Garf: Learning generalizable 3d reassembly for real-world fractures. In: International Conference on Computer Vision (ICCV) (2025) iii, iv, vii
15. Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W., et al.: Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. arXiv preprint arXiv:2502.06608 (2025) iii, vii
16. Liu, S., Zhou, Y., Yang, J., Gupta, S., Wang, S.: Contactgen: Generative contact modeling for grasp generation. In: ICCV. pp. 20609–20620 (2023) iii
17. Liu, T., Liu, Z., Jiao, Z., Zhu, Y., Zhu, S.C.: Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. IEEE Robotics and Automation Letters **7**(1), 470–477 (2021) x
18. Lu, J., Sun, Y., Huang, Q.: Jigsaw: Learning to assemble multiple fractured objects. In: Advances in Neural Information Processing Systems. vol. 36, pp. 14969–14986 (2023) iv
19. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv preprint arXiv:2108.10470 (2021) x
20. 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) ii, iii
21. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) vi
22. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) iv

23. Prokudin, S., Lassner, C., Romero, J.: Efficient learning on point clouds with basis point sets. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4332–4341 (2019) ii, iv
24. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) iv
25. Reuss, M., Li, M., Jia, X., Lioutikov, R.: Goal-conditioned imitation learning using score-based diffusion policies. arXiv preprint arXiv:2304.02532 (2023) iv
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) iv
27. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022) iv
28. Scarpellini, G., Fiorini, S., Giuliari, F., Moreiro, P., Del Bue, A.: Diffassemble: A unified graph-diffusion model for 2d and 3d reassembly. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28098–28108 (2024) iv
29. Shao, L., Ferreira, F., Jorda, M., Nambiar, V., Luo, J., Solowjow, E., Ojea, J.A., Khatib, O., Bohg, J.: Unigrasp: Learning a unified model to grasp with multi-fingered robotic hands. IEEE Robotics and Automation Letters **5**(2), 2286–2293 (2020) i, iii
30. Sun, T., Zhu, L., Huang, S., Song, S., Armeni, I.: Rectified point flow: Generic point cloud pose estimation. In: Advances in Neural Information Processing Systems (NeurIPS) (2025) iv, vii, xv
31. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022) iv
32. Wang, Z., Chen, J., Furukawa, Y.: Puzzlefusion++: Auto-agglomerative 3d fracture assembly by denoise and verify. In: ICLR (2025) iv
33. 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) ii, iv, x, xi, xiv
34. 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) ii
35. Wu, Z., Potamias, R.A., Zhang, X., Zhang, Z., Deng, J., Luo, S.: Cedex: Cross-embodiment dexterous grasp generation at scale from human-like contact representations. arXiv preprint arXiv:2509.24661 (2025) i, xi
36. Wu, Z., Zhang, X., Chen, Z., Deng, J., Potamias, R.A., Luo, S.: Unimorphgrasp: Diffusion model with morphology-awareness for cross-embodiment dexterous grasp generation. arXiv preprint arXiv:2602.00915 (2026) iv
37. Xu, Y., Wan, W., Zhang, J., Liu, H., Shan, Z., Shen, H., Wang, R., Geng, H., Weng, Y., Chen, J., et al.: Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In: CVPR. pp. 4737–4746 (2023) i, iii
38. Xu, Z., Gao, C., Liu, Z., Yang, G., Tie, C., Zheng, H., Zhou, H., Peng, W., Wang, D., Chen, T., Yu, Z., Shao, L.: Manifoundation model for general-purpose robotic manipulation of contact synthesis with arbitrary objects and robots (2024) i, iii
39. Zhao, F., Tsetserukou, D., Liu, Q.: Graingrasp: Dexterous grasp generation with fine-grained contact guidance. In: ICRA. pp. 6470–6476. IEEE (2024) i, iii

40. Zhao, W., Cao, Y.P., Xu, J., Dong, Y., Shan, Y.: Assembler: Scalable 3d part assembly via anchor point diffusion. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers (2025) iv