DEFORMABLE ENERGY LANDSCAPES FOR DEXTEROUS GRASP GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dexterous grasp synthesis is typically formulated as conditional generation: a model samples hand configurations given an object, while separate stages rank, refine, or physically verify the results. In contrast, we learn a conditional energy field, a differentiable scalar over the hand's configuration space whose landscape is deformed by the target object, and synthesize grasps via gradient descent. Operating on a learned manifold of hand articulations, the field guides the hand entirely through its gradient. Since descent can start from any configuration, a single field can serve as a generator from uninformative initializations, as an operator to evaluate and refine the outputs or failures of other methods, and as a teacher for a motion prior. We instantiate this framework across graphics and robotics tasks, including static grasp synthesis and reach-and-grasp motion for the human hand, as well as grasp generation and cross-embodiment retargeting for robot hands. Our approach demonstrates competitive performance on all these distinct tasks.

1 INTRODUCTION

Dexterous grasp synthesis is commonly formulated as conditional generation. Conditional variational autoencoders, diffusion, and flow models are widely used to generate human grasps (Taheri et al., 2020; Jiang et al., 2021; Wu et al., 2025), robot-hand grasps (Xu et al., 2023; Zhong et al., 2025; Huang et al., 2026), and hand-object interactions (Christen et al., 2024; Li et al., 2025; Chen et al., 2026a). Such a model reproduces the distribution of demonstrated hand configurations given an object. However, relying purely on mapping a distribution of successes presents a fundamental limitation: the model captures a configuration's statistical likelihood rather than its physical viability. Because standard generators map from a learned prior or noise to a target distribution, they lack an explicit understanding of failure. Consequently, when a configuration is a near-miss (e.g., plausible but physically unstable) improving it within a generative framework is difficult. It typically requires either drawing a completely new sample or executing a reverse-diffusion process, as the generative model itself does not provide a continuous directional signal to correct the specific geometric or physical flaws of the current state.

We shift from matching a distribution of successes to modeling a continuous landscape over the configuration space. We learn a conditional energy field, a differentiable scalar function trained on both successful grasps and failed endpoints. The properties that make a configuration a grasp such as closure, stability under an external wrench, and agreement with how people hold the object, are aggregated into this scalar. By exposing a value and a gradient at every configuration, the field explicitly and deterministically steers the hand: the value grades the viability of the current state, and the gradient indicates exactly how to move it toward a valid grasp. Synthesis then reduces to descent. Since gradient descent can start from any point, this single mechanism naturally unifies multiple roles. When initialized from an uninformative state, the field acts as a generator. When initialized from the flawed output of another method, an intermediate motion frame, or a source grasp mapped from a different robot embodiment, it functions directly as an operator that evaluates and refines the configuration. This unified, gradient-driven formulation allows a single approach to handle distinct tasks without requiring task-specific transport models or specialized test-time simulators.

Concretely, let $\theta \in \Theta$ be a hand configuration and $c \in \mathcal{C}$ a condition. A conditional energy field is a differentiable map $E_\phi : \Theta \times \mathcal{C} \to \mathbb{R}$. Fixing c defines a landscape over the configuration space, and varying c deforms this landscape: $E_\phi(\cdot \mid c) = \mathcal{D}_c\, E_\phi(\cdot \mid \varnothing)$. Given an initial distribution q_0,

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.

synthesis is the gradient flow

$$\dot{\theta}(t) \;=\; -\nabla_\theta E_\phi(\theta(t) \mid c)\,, \qquad \theta(0) \sim q_0, \tag{1}$$

integrated to a stationary point. While many existing learned fields operate in $\mathbb{R}^3$ or on object surfaces to predict signed distances or contacts (Karunratanakul et al., 2020; Grady et al., 2021; Liu et al., 2023; Morales et al., 2025), these representations require a separate optimization stage to fit the hand. Instead, we follow energy-based inference (LeCun et al., 2006; Du & Mordatch, 2019) and define our scalar directly on the configuration (Lu et al., 2019; Weng et al., 2022; Urain et al., 2023). We differ from prior work in two key choices. First, our descent operates over the wrist pose and a learned articulation latent of a multi-fingered hand, avoiding adversarial minima in raw joint angles. Second, we regress the scalar to a graded target using a corpus that explicitly includes the field's own failed endpoints. This ensures that the fitted landscape can reliably guide the flow within a bounded region, fulfilling its roles without an explicit generator or external refiner.

As the condition only deforms the landscape, our formulation cleanly separates task-specific choices from the core framework. Moving to a new task simply updates three components: the configuration space (e.g., human or robot hand), the condition (e.g., object geometry or motion phase), and the target scalar that defines a valid grasp. The architecture of the field and its training principles remain shared. We apply this formulation to four tasks: static grasp synthesis and reach-and-grasp motion generation for the human hand, and from-scratch grasp generation and cross-embodiment retargeting for robot hands. For human-hand tasks, the scalar is a demonstration-anchored geometric distance. For robot-hand tasks, it is a simulated stability margin. Despite optimizing different target criteria, both classes share the field formulation.

The major contributions of our work include the following:

- A conditional energy field over hand configurations for grasp synthesis, formulated as an unconditioned landscape deformed by its conditions and used through its gradient.
- Four general design principles that characterize when a learned energy field reliably supports synthesis by descent, validate through paired ablations across all tasks.
- A unified framework applied to four distinct tasks: static grasp generation, dynamic reach-and-grasp motion, robot grasp synthesis, and cross-embodiment retargeting. On each task, our single method achieves competitive results against established task-specific baselines.

2 Related Work

Grasp fields in space. Human-hand fields in $\mathbb{R}^3$ or on an object surface predict signed distances (Karunratanakul et al., 2020), contact (Grady et al., 2021; Liu et al., 2023), correspondences (Zhou et al., 2022), or basis-point distances (Morales et al., 2025). These approaches yield an intermediate representation from which a separate optimization stage recovers the hand pose. Conversely, field-free generators sample from a latent space (Taheri et al., 2020; Jiang et al., 2021; Zhao et al., 2024; Wu et al., 2025) or predict a contact map (Li et al., 2024; Chen et al., 2026b). In contrast, our field lives directly on the hand's configuration and outputs a single scalar. Gradient descent moves the configuration itself, removing the need for intermediate representations or separate fitting steps.

Fields on the configuration. We build on the concept proposed by Lu et al. (2019), who ascend a learned success probability over palm pose and preshape. Related fields include NGDF (Weng et al., 2022) and SE(3)-DiffusionFields (Urain et al., 2023) over gripper poses. While prior works deploy energy fields primarily to refine failing initializations (Lu et al., 2019) or optimize motions (Weng et al., 2022; Urain et al., 2023), we extend the field's utility to three roles: generator, operator, and teacher. We differ from them in two fundamental design choices. First, we descend over wrist pose and a learned articulation latent. While restricting search to a learned manifold is standard (Tiwari et al., 2022; He et al., 2024), we find it essential because optimizing in raw joint angles leads our field into low-energy configurations that are not physically valid grasps. Second, we regress a demonstration distance or simulated margin, graded as in NGDF and Get a Grip (Lum et al., 2024). Importantly, our corpus includes the field's own failed endpoints, avoiding overconfidence in unexplored regions. Classifier- or contrastively-trained scalars can have flat or noisy gradients (Urain et al., 2023; Ruan et al., 2025), whereas Ruan et al. (2025) and others employ eikonal or derivative losses (Czarnecki et al., 2017; Atzmon & Lipman, 2020; Ma et al., 2020; Sóti et al., 2025).

We demonstrate that simple, graded value regression sufficiently stabilizes the gradient landscape without requiring complex derivative regularization.

Physics and operators. For robot-hand grasp synthesis, recent methods (Xu et al., 2023; Zhong et al., 2025; Zhu et al., 2025) employ generative models guided by physical energies and achieve impressive performance. Such methods may require reverse-diffusion sampling or heavy test-time adaptation with simulators (Jiang et al., 2021; Lu et al., 2024). Other approaches optimize physical energies directly (Liu et al., 2021; Wang et al., 2022), or refine prior systems' outputs (Grady et al., 2021; Zhou et al., 2022; Weng et al., 2024). Our operator role is a learned, criterion-specific instance of this refinement and selection process, executing directly on the configuration. Our field achieves generation and refinement deterministically via gradient descent.

Cross-embodiment and motion. NeuralGrasps shares a latent space among robot grippers and a human hand (Khargonkar et al., 2023); Grasp2Grasp learns a latent Schrödinger bridge with physics in training-time transport costs and no simulator at generation (Zhong et al., 2026); and Qi et al. (2026) transfer grasps via per-grasp optimization. Our retargeting descends a scalar fitted to the benchmark simulator's margin for the target hand on training objects, starting from the mapped source grasp. For motion generation, kinematic methods sample hand–object sequences (Christen et al., 2024; Li et al., 2025; Chen et al., 2026a) or generate motion toward a sampled final grasp (Taheri et al., 2022; Wang et al., 2026). In contrast, our field governs a kinematic motion system across four stages (conditioning the prior, per-step selection, per-step refinement, and termination below a per-object threshold) without requiring a pre-sampled target grasp.

We unite configuration fields (Lu et al., 2019; Weng et al., 2022), pose manifolds (Tiwari et al., 2022; He et al., 2024), shared latents (Khargonkar et al., 2023), and latent transport (Zhong et al., 2026) into a single formulation. Our contributions lie in establishing a unified framework evaluated on four diverse tasks, identifying four core design principles, and demonstrating that an energy field can serve effectively as an end-to-end generator and versatile operator.

3 Method

3.1 Conditional energy fields and synthesis by descent

We define a hand configuration as $\theta = (g, z)$, where $g \in SE(3)$ is the wrist pose in the object's frame and z is a coordinate on a learned manifold $\mathcal{Z}$ of feasible articulations. Fingers are decoded via $\mathrm{dec}(z)$ rather than optimized as raw joint angles, and restricting descent to $\Theta = SE(3) \times \mathcal{Z}$ is a core design choice. The condition c is compositional: it relies on the object geometry at its base, plus, depending on the task, the identity of the driven hand or the phase of a reaching motion.

We formulate $E(\cdot \mid c)$ as a deformation $\mathcal{D}_c$ of the unconditioned landscape $E(\cdot \mid \varnothing)$. The conditions together with the deformations are layered: with object o and hand h, $E(\cdot \mid o, h) = \mathcal{D}_h \mathcal{D}_o E(\cdot \mid \varnothing)$. A base terrain is reshaped first by the object and then by the embodiment. This formulation enables one field to drive multiple hands that a new hand simply acts as a new deformation operator.

During inference, the field moves the hand via its gradient $-\nabla_\theta E$. We learn a scalar field because its gradient is conservative. The optimization accumulated along any path is path-independent, and the direction's quality is determined by how the scalar is supervised. The endpoints reached from q_0 sample the minima of $E(\cdot \mid c)$ weighted by their basin volume. Therefore, capturing diverse grasps is a property of the landscape and the initialization q_0, independent of explicit sampling heuristics.

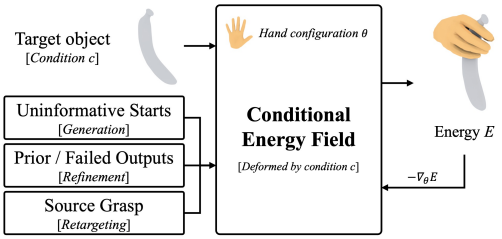


Figure 1: Synthesis by descent. Depending on the initialization q_0, the exact same gradient descent loop over the configuration manifold Θ handles from-scratch *generation*, output *refinement*, and cross-embodiment *retargeting*.

The field's definition of a successful grasp is dictated by the target scalar $d^\star(\theta \mid c)$. We employ two families of targets: a *demonstration-anchored* distance from θ to ground-truth grasps, and a *simulator-anchored* margin under physical perturbation tests. To guide the gradient flow ef-

fectively across the configuration space, we decompose the target scalar into two distinct terms:

$$d^\star(\theta \mid c) \;=\; G(\theta \mid c) \;+\; Q(\theta \mid c), \tag{2}$$

where G provides a guidance slope when the primary criterion is uninformative, and Q grades the grasp quality locally near the object (saturating elsewhere). How these terms are joined depends on the target family: When Q is an analytic penalty defined everywhere, they are simply summed. When Q is only valid in contact regions, G is gated by Q's validity, ensuring the optimization relies on one term at a time. Because all task knowledge sits in $d^\star$, labelling may use meshes, distance grids, or a simulator during training. A fitted field is defined as the pair $(E, \mathcal{S})$, where $\mathcal{S}$ is the support of the labeled distribution. While the flow is defined everywhere on Θ, it is mathematically reliable only on $\mathcal{S}$. Rather than artificially confining descent to a trust region, we construct the training distribution $\mathcal{S}$ to cover the spatial extent the descent trajectory will traverse.

The initial distribution q_0 is highly flexible, allowing the field to serve multiple roles: initialized from an uninformative q_0, the field acts as a generator; initialized from another method's outputs or failures, it serves as an operator for evaluation and refinement. When integrated for a bounded number of steps from an intermediate state of a motion, its displacement provides a signal that a sequence model can learn to reproduce, acting as a teacher. The methodological framework remains identical across these roles, with the source of q_0 varying.

3.2 Design principles

Four core properties dictate whether gradient descent successfully navigates from an uninformative initialization to a viable grasp. These are intrinsic properties of the fitted landscape rather than task-specific constraints, and each establishes a fundamental rule for constructing the training target.

1. Bounded Reliability within Label Support. The fitted gradient $\nabla_\theta E$ only provides a valid descent direction for $d^\star$ within the support $\mathcal{S}$ of the labeled distribution, unconstrained elsewhere. Gradient descent compounds this vulnerability: a single step outside $\mathcal{S}$ exposes the trajectory to arbitrary gradients with no mechanism to return. Consequently, the reliable range of descent is strictly bounded by the extent of $\mathcal{S}$. Rather than restricting the optimizer with a conservative trust region which merely limits refinement, we proactively construct $\mathcal{S}$ to cover the expected descent paths. Specifically, the support includes configurations with the wrist initialized on a spherical shell with uniform $SO(3)$ orientations, multi-scale perturbations of demonstrated grasps, and critically, the relabeled endpoints of the field's own failed descents.

2. Decoupling Guidance and Quality Objectives. The two terms in Equation 2 have distinct roles: G supplies a non-vanishing gradient where the primary criterion is uninformative, drawing the start distribution into the basin of attraction where Q becomes active. Q then determines the precise location and ordering of the minima within that region. Neither term succeeds alone: A target consisting only of demonstration proximity is informative everywhere, but its minimizer is the exact demonstrated configuration, reducing generation to retrieval. A target consisting only of grasp quality has a vanishing gradient far from the object. Note that the decomposition itself does not prevent the terms from interfering. If left active, G competes with Q precisely where physical contact matters most. Thus, G must yield to Q in the contact region, either via a gating mechanism or dynamic weighting. The reliance on G scales with the initialization distance: the farther q_0 is from the object, the more G dictates the outcome.

3. Continuous Regression for Informative Gradients. As inference uses E exclusively through $\nabla_\theta E$, the supervision signal must determine informative gradients rather than just decision boundaries. The target must therefore be graded across the configuration space, including regions where the criterion rejects the grasp. A binary accept/reject label is piecewise constant: its gradient vanishes almost everywhere, providing no direction for a rejected configuration to recover. Consequently, margin and contrastive objectives suffer a similar flaw: they constrain E only near the boundary between accepted and rejected sets, creating a barrier at the edge rather than a smooth basin inside it. Thus, the descended energy is fitted purely by regression to $d^\star$.

This is also critical for selection: a fitted energy $E = E^\star + \varepsilon$ contains approximation error ε, globally selecting the minimum across disparate candidates risks minimizing ε instead of $E^\star$, known as the optimizer's curse (Smith & Winkler, 2006). However, continuous gradient descent evaluates only local differences along a path where ε largely cancels out, making local gradients significantly more robust for guidance than the global scalar value.

4. Coordinate Restriction over Energy Penalization. A fitted energy inevitably contains local minima outside $\mathcal{S}$ where it was never constrained. Descent in unrestricted joint coordinates frequently discovers these regions, yielding configurations that minimize E without being valid grasps. To address this, we restrict the optimizer's coordinate space, leaving E unchanged. The hand articulation moves exclusively through z, the latent manifold fitted to the training grasps.

4 EXPERIMENTS

Tasks. We evaluate our approach on four tasks across two hand types and two initialization sources for q_0. On the human hand, the target is demonstration-anchored, whereas on the robot hands, the quality term is a simulated stability margin. When q_0 is uninformative (e.g., random initializations on a spherical shell), the field acts as a generator for static MANO (Romero et al., 2017) grasps and Shadow-hand grasps. For the Shadow hand, the conceptual terms G and Q from Equation 2 are decoupled into two sequential fields. When q_0 is supplied by another process, the field descends from a pre-determined configuration. In the reach-and-grasp task, initial states come from a learned prior over motion primitives. Here, the field serves two distinct roles. First, it acts as a teacher providing displacement targets that the prior learns to reproduce. Second, it functions as an operator that selects among the prior's proposals and refines the final frame of each primitive. In cross-embodiment retargeting, q_0 is a source grasp mapped through a shared latent space, which the field refines as an operator. For evaluations, we directly report baseline metrics from their original publications if available for all tasks. Where a specific evaluation is absent, we execute the official open-source implementations of selected state-of-the-art methods.

Implementation. All fields share a common architecture with a hidden dimension of 512. A two-level PointNet++ (Qi et al., 2017) set-abstraction encoder processes the object's point cloud into 128 tokens and a global feature. For human-hand tasks, the articulation manifold $\mathcal{Z}$ is the latent space of a pre-trained variational autoencoder (VAE) (Kingma & Welling, 2013). The hand is represented as keypoints computed via forward kinematics from $(g, \mathrm{dec}(z))$. We extract 16 keypoints for MANO, and 28, 21, or 12 keypoints for the Shadow, Allegro, or Barrett hands, respectively. These keypoints attend to the object tokens through cross-attention blocks. A residual MLP modulated by the global feature vector maps the pooled keypoint features to the scalar energy E. Fields are trained using AdamW with learning rate 3×10^{-4}, and a smooth-ℓ_1 loss against standardized labels.

Ablations. Each ablation isolates one component, where G and Q are the two terms of Equation 2.

- **w/o far**. We discard configurations with a keypoint distance exceeding 30 mm from a reference grasp (human demonstrations for human-hand tasks and successful grasps in the dataset for robot-hand tasks). For each object, we retain the 512 nearest configurations and retrain the model.
- **w/o guidance**. We remove G from the target and retrain. Specifically, G is the distance to demonstrations for the human hand, the gated fingertip reach for retargeting, and the stage-1 geometric field for the Shadow hand.
- **w/o quality**. We remove Q from the target and retrain. Q is the penetration and fingertip-reach penalties for the human hand, and the simulated graded margin for retargeting. For the Shadow hand, Q represents the stage-2 field, so grasps are generated exclusively from the geometric field.
- **w/o selection**. At inference, we zero the energy term of the selection rule. The Shadow hand keeps farthest-point sampling, and the reach-and-grasp pipeline keeps its affordance term.
- **w/o latent**. At inference, we optimize raw finger joint angles instead of the VAE latent.

4.1 HUMAN-HAND TASKS

For both human-hand tasks, we report an optional condition labeled "+ pen.", which adds a point-cloud penetration penalty to the descent objective to reduce hand–object interpenetration. This penalty is computed directly from the input point cloud, requiring no learned parameters. It penalizes each hand vertex lying inside the object, scaling the cost by its area times its depth.

Static grasp synthesis. Given an object's point cloud, the goal is to generate static MANO grasps. A final grasp is the lowest-energy endpoint of descent evaluated across multiple initializations. During inference, we augment the demonstration-anchored field with a simple geometric hinge loss on the distance from the finger-pad centroids to the object's input cloud. However, because the pad centroid

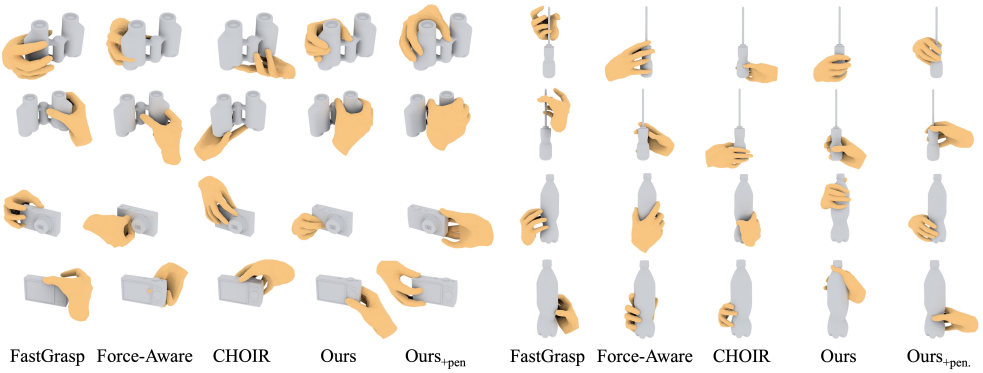


Figure 2: Qualitative results of static MANO (Romero et al., 2017) grasps synthesis.

Table 1: Static grasp synthesis, FastGrasp (Wu et al., 2025) protocol.

Method	OakInk PenVol ↓	OakInk SimDisp ↓	OakInk Contact ↑	OakInk Entropy ↑	OakInk ClusSz ↑	GRAB PenVol ↓	GRAB SimDisp ↓	GRAB Contact ↑	GRAB Entropy ↑	GRAB ClusSz ↑
GraspTTA (Jiang et al., 2021)	3.66	3.65	81.4	2.91	3.74	3.71	3.02	80.0	2.80	3.36
GrabNet (Yang et al., 2022)	6.60	**1.21**	94	1.68	1.22	1.72	3.65	96	2.72	1.93
ContactGen (Liu et al., 2023)	4.85	2.01	94	2.88	4.07	2.16	2.72	96	2.88	4.11
ClickDiff (Li et al., 2024)	12.05	1.75	93.4	2.89	4.10	10.37	**0.59**	**100**	2.89	1.55
CHOIR (Morales et al., 2025)	9.17	2.36	96.0	2.93	5.17	4.93	1.86	99.2	2.84	**4.74**
FastGrasp (Wu et al., 2025)	4.37	1.45	94	2.92	4.96	**1.25**	1.67	**100**	**2.93**	1.87
Force-Aware (Chen et al., 2026b)	5.50	1.86	99.7	2.90	**5.39**	5.47	0.74	**100**	2.80	4.68
Ours	9.13	1.42	**100**	**2.96**	4.90	6.69	1.09	**100**	2.86	4.50
Ours$_{+pen.}$	**3.05**	1.88	98.7	2.95	4.98	3.21	1.71	98.3	2.78	4.47
Ours *w/o far*	9.54	1.45	97.8	2.93	4.96	6.97	1.29	99.2	2.84	3.80
Ours *w/o guidance*	8.63	3.16	92.7	2.91	4.95	6.13	2.37	99.2	2.89	4.58
Ours *w/o quality*	9.90	1.61	100	2.93	4.91	8.99	1.13	100	2.75	4.56
Ours *w/o selection*	11.66	1.62	100	2.93	4.88	11.18	1.37	100	2.76	4.67
Ours *w/o latent*	8.39	1.72	100	2.87	4.78	6.72	1.23	100	2.87	4.65

lies slightly beneath the skin, the hinge implicitly pulls the hand into the object, which is one source of the penetration reported below. We evaluate on OakInk (Yang et al., 2022) and GRAB (Taheri et al., 2020) following the protocol of Wu et al. (2025). The evaluated metrics include penetration volume, simulated displacement (the object's displacement under gravity with the hand held fixed), contact ratio (the share of grasps with non-zero penetration volume), and the entropy and cluster size of a k-means partition to measure grasp diversity.

Quantitative results are reported in Table 1 and qualitative results are shown in Figure 2. On both benchmarks, our base formulation establishes strong contact closure and grasp diversity. However, because the unconstrained field pulls the hand tightly toward the object to minimize the demonstration-anchored distance, it inherently incurs high penetration volumes. Applying the inference-time penetration penalty directly mitigates this interpenetration, driving the penetration volume down significantly. The field accommodates this zero-shot constraint without compromising simulated displacement or significantly sacrificing the contact ratio.

For the ablations, *w/o far* worsens stability, corroborating Principle 1 that the descent trajectory is only reliable within the explicit support of the labeled distribution. Removing the guidance term drastically degrades physical stability, causing SimDisp to spike to 3.16 on OakInk. This confirms Principle 2 that the guidance gradient is essential for pulling uninformative initializations into the local basin where the quality term is decisive. Finally, *w/o latent* lowers penetration but noticeably reduces grasp diversity, validating Principle 4 that the articulation manifold protects the flow from adversarial minima outside the data distribution.

Reach-and-grasp motion. Given an object, the objective is to generate a right-hand motion that starts from an open hand away from a static object and ends in a grasp. This follows the setting of Wang et al. (2026), our main comparison. While their method samples a final grasp, generates motion towards it, and optimizes contact against a mesh at test time, our method requires no given target grasp. Instead, we generate motion auto-regressively, one primitive at a time. A primitive comprises an eight-frame window of finger poses and wrist increments, expressed in the wrist frame

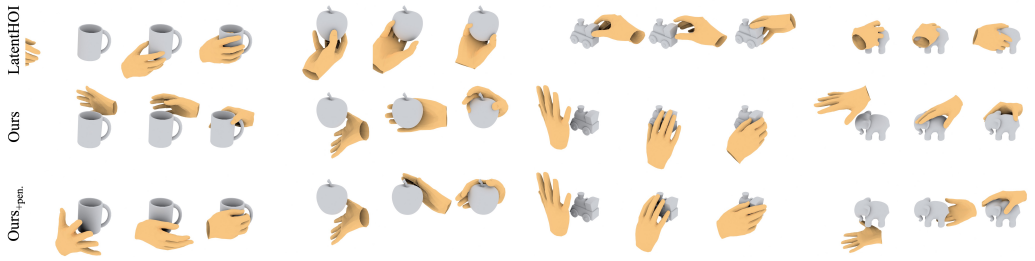


Figure 3: Qualitative comparison with LatentHOI (Li et al., 2025) on reach-and-grasp motion.

Table 2: Reach-and-grasp under two evaluation protocols.

(Wang et al., 2026) Protocol

Method	Depth ↓	Ratio ↓	Score ↑	Valid ↑
GOAL (Taheri et al., 2022)	0.024	0.103	0.072	0.538
SAGA (Wu et al., 2022)	0.012	0.052	0.069	0.532
(Wang et al., 2026)	**0.008**	0.043	0.091	0.551
Ours	0.022	0.050	**0.097**	**0.808**
Ours$_{+pen.}$	0.012	**0.036**	0.077	0.758
Ours *w/o far*	0.031	0.050	0.079	0.767
Ours *w/o guidance*	0.029	0.060	0.063	0.733
Ours *w/o quality*	0.035	0.063	0.113	0.950
Ours *w/o selection*	0.026	0.052	0.094	0.758
Ours *w/o latent*	0.022	0.049	0.076	0.692

LatentHOI (Li et al., 2025) Protocol

Method	IV ↓	ID ↓	CR ↑	IVU ↓
MLD (Chen et al., 2023)	9.62	1.06	10.23	0.24
LatentHOI (Li et al., 2025)	6.38	0.77	11.94	0.10
HO-Flow (Chen et al., 2026a)	5.31	**0.61**	11.38	**0.07**
Ours	7.93	1.09	**14.13**	0.14
Ours$_{+pen.}$	**4.90**	0.62	10.94	0.10
Ours *w/o far*	9.39	1.34	10.90	0.22
Ours *w/o guidance*	5.93	1.01	10.11	0.16
Ours *w/o quality*	10.77	1.13	13.46	0.20
Ours *w/o selection*	9.04	1.14	11.71	0.21
Ours *w/o latent*	7.72	1.06	13.97	0.16

at the start of the window and conditioned on the two preceding frames. A transformer VAE encodes this primitive into a 64-dimensional latent space, and a latent diffusion model conditioned on the motion history serves as the kinematic prior. A FiLM (Perez et al., 2018) adapter on the frozen prior conditions it on the displacement of a 30-step field descent from the current hand configuration. This adapter is distilled from the field's selections over the frozen prior's samples on training objects.

At each step, the adapted prior proposes a set of candidate primitives. They are ranked by the field's energy plus an affordance term that rewards fingers landing near demonstrated contact regions. The field then refines the last frame of the chosen primitive via a fixed-length descent and blends the correction smoothly into the primitive's final frames. The episode ends once the energy falls below a per-object threshold, which is calibrated via a descent guided by the object's signed distances.

Following the protocol of Wang et al. (2026), our evaluation considers interpenetration depth and ratio, the Ferrari–Canny force-closure score, and the fraction of grasps with a positive score. We additional evaluate against the protocol of Li et al. (2025). Since their original task differs (text-conditioned, dynamic object, bimanual interaction), we adapt their metrics to our setting: evaluating the right hand on a static object without text guidance, scoring sequences from the first contact frame to the final grasp. We measure interpenetration volume (IV) and depth (ID), contact ratio (CR), and Interpenetration Volume per contact Unit (IVU).

Under the protocol of Wang et al. (2026), our field acts as an iterative teacher and operator (handling both trajectory selection and final-state refinement), outperforming prior kinematic methods without requiring a sampled target grasp. Our descent achieves a superior Ferrari–Canny score and higher grasp validity. Integrating the penetration penalty constrains interpenetration depth and ratio with a minor trade-off in force-closure. Under the grasping metrics of Li et al. (2025), our formulation also confirms its geometric robustness, yielding the highest contact ratio, and the lowest interpenetration volume when penalized.

The ablations reinforce the necessity of our structured descent throughout the motion generation process. *w/o selection* depresses validity, proving that the energy landscape is critical for dynamically filtering trajectory proposals at every step, rather than merely refining the final state. Removing the latent constraint (*w/o latent*) collapses validity, confirming that unconstrained joint descent mid-motion frequently strands the hand in physically non-viable kinematics.

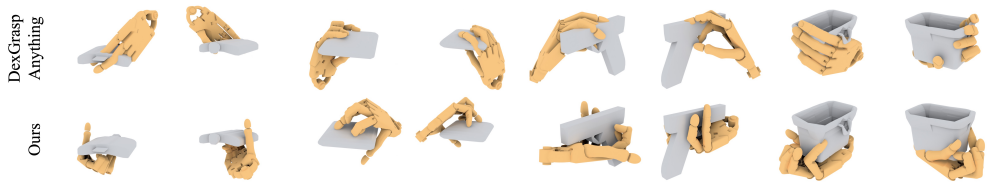


Figure 4: Qualitative comparison with DexGrasp Anything (Zhong et al., 2025) on Shadow grasp.

Table 3: Shadow-hand grasp generation, DexGrasp Anything (Zhong et al., 2025) protocol.

Method	Suc.6 ↑	Suc.1 ↑	Pen ↓	Div ↑
UniDexGrasp (Xu et al., 2023)	33.9	70.1	31.9	0.14
UGG (Lu et al., 2024)	46.9	79.0	25.2	0.14
DexGrasp Any. (Zhong et al., 2025)	53.6	<u>90.4</u>	21.5	<u>0.22</u>
EvolvingGrasp (Zhu et al., 2025)	<u>65.4</u>	**92.3**	**15.9**	–
Ours	**65.6**	84.2	<u>20.2</u>	**0.25**
DexGrasp Any.∗	50.7	84.6	21.2	0.15
DexGrasp Any. +*Ours*	66.3	87.7	17.2	0.26
UGG +*Ours*	62.4	84.4	21.0	0.18
Ours *w/o far*	41.8	58.8	21.8	0.22
Ours *w/o guidance*	48.1	64.6	19.7	0.08
Ours *w/o quality*	13.8	36.9	29.9	0.05
Ours *w/o selection*	52.8	69.3	20.3	0.23
Ours *w/o latent*	64.3	82.0	19.8	0.25

Table 4: Success rate of Cross-embodiment retargeting, Grasp2Grasp (Zhong et al., 2026) protocol. H/A/S/B: human, Allegro, Shadow, Barrett.

Method	H→A	H→S	S→A	H→B
CrossDex (Yuan et al., 2025)	36.5	9.0	35.1	-
Dex-Retargeting (Qin et al., 2023)	56.9	16.2	56.4	-
RFP (Khargonkar et al., 2025)	66.8	33.9	70.3	55.9
Grasp2Grasp (Zhong et al., 2026)	**77.7**	<u>42.6</u>	<u>78.7</u>	<u>63.1</u>
Ours	<u>75.3</u>	**48.6**	**84.5**	**64.8**
Ours *w/o far*	68.9	44.3	66.7	64.0
Ours *w/o guidance*	76.0	48.5	78.2	62.1
Ours *w/o quality*	61.1	30.3	60.0	52.9
Ours *w/o selection*	71.2	46.1	80.2	61.4
Ours *w/o latent*	67.3	44.6	66.9	58.2

4.2 Robot-hand tasks

In contrast to the human-hand tasks, both robot-hand tasks employ a simulator-anchored field. The regression target is the physical stability margin measured under a six-direction perturbation test in IssacGym simulator. Because this training formulation intentionally mirrors the benchmark's own scoring metric, the supervision labels and the evaluation physics are naturally aligned.

Shadow grasp synthesis. The objective is to generate Shadow-hand grasps per object given its point cloud. Our approach realizes Principle 2 (§3.2) by deploying two fields in sequence. First, a geometric field (acting as the guidance term G) descends to convergence from multiple initializations on a spherical shell around the object. Then, a physical field (acting as the quality term Q, fitted to a simulated stability margin) descends from each of these endpoints. Grasps are kept by energy margin and farthest-point sampling. Inference requires no simulator, learned ranker, or demonstration. This sequential design ensures that the first stage supplies global coverage and the second enforces physical quality. Qualitative visualizations of the synthesized grasps are provided in Figure 4.

We follow the evaluation protocol of Zhong et al. (2025) on the DexGraspNet (Wang et al., 2022) with the IsaacGym simulator. Suc.6 is the share of grasps passing all six directions and Suc.1 is the share passing at least one. Pen is the largest depth of hand surface points inside the object, averaged over passing grasps; Div is the per-joint standard deviation of the passing grasps' joint angles, averaged over joints and objects. Due to computational costs and observed performance saturation, we limit our physical field's training data to approximately 200k simulated labels, comprising both successful and failed endpoints. To establish a fair comparison, we train a data-aligned baseline (DexGrasp Any.∗ in Table 3), restricted to 200k successful grasps sampled from DexGraspNet.

Our sequential generator formulation achieves a 65.6% success rate under the six-direction perturbation test (Suc.6) and a kinematic diversity of 0.25. Furthermore, to demonstrate the field's capacity as a plug-and-play operator, we target and refine the failed generations of existing baselines (specifically, grasps failing the Suc.6 criterion). By replacing these failures with our gradient-refined configurations (DexGrasp Any. +*Ours* and UGG +*Ours*), our method improves overall Suc.6 performance with a possible marginal drop in Suc.1.

For the ablations, removing the geometric guidance stage (*w/o guidance*) drops Suc.6 to 48.1%, showing that the physical field alone cannot attract uninformative initializations. Dropping the physical refinement (*w/o quality*) collapses Suc.6 to 13.8%, confirming that geometric proximity alone is insufficient for physical stability. Omitting the spatial label threshold (*w/o far*) degrades performance to 41.8%, suggesting that descent loses reliability outside the explicitly labeled train-

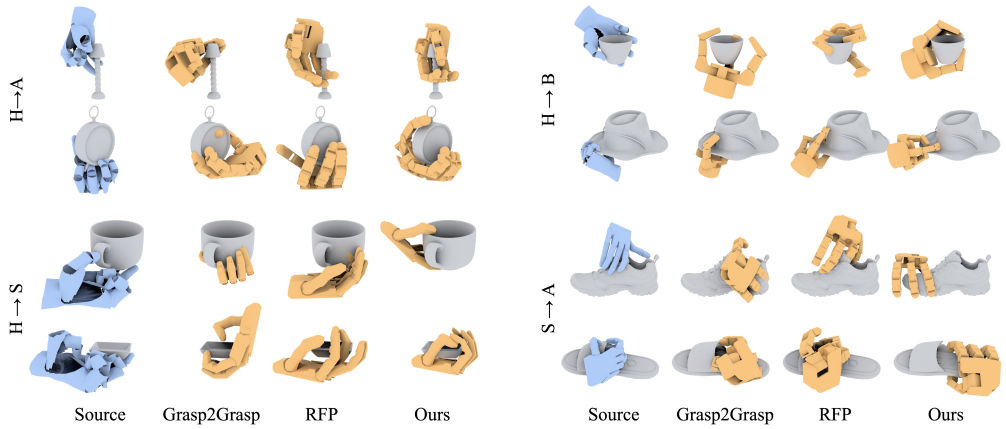


Figure 5: Qualitative results of cross-embodiment retargeting.

ing support (Principle 1). *w/o selection* reduces Suc.6 to 52.8%, highlighting the field's essential role in dynamically filtering diverse candidates.

Cross-embodiment retargeting. Given a source grasp, the objective is to generate a corresponding grasp for a target hand. We evaluate four transfer directions: human to Allegro, human to Shadow, Shadow to Allegro, and an additional human to Barrett not originally included in the benchmark. The source configuration is first mapped through a shared latent space to initialize the target hand. The target hand's field is fitted to the simulator's graded margin, supplemented by a fingertip-reach term where the margin is zero. We perform gradient descent in the shared latent space starting from this mapped initialization and parallel local perturbations, selecting the lowest-energy endpoint.

We follow the evaluation protocol of Zhong et al. (2026). A transferred grasp succeeds if, without gravity, the object moves less than 2 cm under each of six directional pushes. As shown in Table 4, our approach achieves leading performance on three of the four directions while remaining competitive on Human→Allegro. This demonstrates that a well-fitted landscape can adapt a cross-domain initialization to a new embodiment without dedicated transport models or test-time simulation.

The ablations highlight the mechanics driving this transfer. Removing the simulated margin target (*w/o quality*) severely impairs physical stability, confirming that a geometric mapping alone cannot guarantee a secure hold. Omitting the spatial label filtering (*w/o far*) causes broad declines across all transfers. Descending in unconstrained joint angles (*w/o latent*) universally degrades performance.

5 Discussion and Limitations

In this work, we cast dexterous grasp synthesis as gradient descent over a conditionally deformed energy landscape. By framing synthesis as a continuous optimization problem on a learned articulation manifold, a single architectural framework acts as a generator, an operator, and a teacher. Evaluated across static generation, kinematic motion, and cross-embodiment retargeting, our unified approach remains highly competitive with task-specific models.

While the human-hand fields are fitted to a demonstration-anchored distance and the robot-hand fields to a simulator's margin, they share a unified construction and recipe. Because these regression targets represent fundamentally different criteria, improvements found under one criterion rarely transfer to the other. We claim one methodology and one recipe, not one universal field. Within its support, the field functions as a smooth proximity to the labeled set. It supplies local descent directions, not a global ranking, and it is not a probability density.

Hoever, the reliance on a fitted landscape introduces several limitations. As established in our design principles, the field's reliability is bounded by the explicit support of its training data. Thus, extreme initializations falling far outside will trap the optimizer. Also, the field inherently absorbs the biases of its target scalar. The human-hand field's geometric target intrinsically pulls the hand into the object, thus we utilize the penetration penalty at test time. Similarly, the robot-hand field acts purely as a neural proxy for its training simulator, meaning any inherent sim-to-real gap in the physics engine is baked into the synthesized grasps.

AI Disclosure Statement

In this work, we used generative AI tools to aid and polish the writing to improve readability. For the literature survey, AI tools were utilized for retrieval and discovery, specifically to find related work. We also used AI tools to assist in research execution, including module and function implementation, environment setup, and training supervision.

We explicitly have not used generative AI tools to draft complete sections of the paper from scratch, generate synthetic datasets, or prove mathematical claims. The remaining required disclosure tasks outlined in the ICLR 2027 policy are not applicable to this work.

We have reviewed all AI-assisted work. Specifically, all AI-generated code, environment setups, and training scripts were executed and verified for correctness by the authors. AI-suggested literature was manually validated through direct reading of the source material to ensure relevance and prevent hallucinated citations. Furthermore, all AI-assisted text was carefully reviewed and revised to ensure scientific accuracy and alignment with the authors' intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Matan Atzmon and Yaron Lipman. Sald: Sign agnostic learning with derivatives. *arXiv preprint arXiv:2006.05400*, 2020.

Gavin Barill, Neil G Dickson, Ryan Schmidt, David IW Levin, and Alec Jacobson. Fast winding numbers for soups and clouds. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010. IEEE, 2023.

Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, Jiankang Deng, Cordelia Schmid, and Stefanos Zafeiriou. Ho-flow: Generalizable hand-object interaction generation with latent flow matching. *arXiv preprint arXiv:2604.10836*, 2026a.

Zhuo Chen, Zhongqun Zhang, Yihua Cheng, Aleš Leonardis, and Hyung Jin Chang. Force-aware 3d contact modeling for stable grasp generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3219–3227, 2026b.

Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Wojciech M Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. *Advances in neural information processing systems*, 30, 2017.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1471–1481. IEEE, 2021.

Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1661–1671. IEEE, 2024.

Yuansen Huang, Jiayi Chen, Haoran Liu, Yubin Ke, Bing Han, Jiangran Lyu, Mi Yan, Li Yi, and He Wang. Kpgrasp: Scalable keypoint flow matching for dexterous grasp generation. *arXiv preprint arXiv:2606.09314*, 2026.

Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11087–11096. IEEE, 2021.

Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pp. 333–344. IEEE, 2020.

Ninad Khargonkar, Neil Song, Zesheng Xu, Balakrishnan Prabhakaran, and Yu Xiang. Neuralgrasps: Learning implicit representations for grasps of multiple robotic hands. In *Conference on robot learning*, pp. 516–526. PMLR, 2023.

Ninad Khargonkar, Luis Felipe Casas, Balakrishnan Prabhakaran, and Yu Xiang. Robotfingerprint: Unified gripper coordinate space for multi-gripper grasp synthesis and transfer. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11381–11388. IEEE, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17416–17425. IEEE, 2025.

Peiming Li, Ziyi Wang, Mengyuan Liu, Hong Liu, and Chen Chen. Clickdiff: click to induce semantic contact map for controllable grasp generation with diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 273–281, 2024.

Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20552–20563. IEEE, 2023.

Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. 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.

Jiaxin Lu, Hao Kang, Haoxiang Li, Bo Liu, Yiding Yang, Qixing Huang, and Gang Hua. Ugg: Unified generative grasping. In *European Conference on Computer Vision*, pp. 414–433. Springer, 2024.

Qingkai Lu, Kautilya Chenna, Balakumar Sundaralingam, and Tucker Hermans. Planning multi-fingered grasps as probabilistic inference in a learned deep network. In *Robotics Research: The 18th International Symposium ISRR*, pp. 455–472. Springer, 2019.

Tyler Ga Wei Lum, Albert H Li, Preston Culbertson, Krishnan Srinivasan, Aaron D Ames, Mac Schwager, and Jeannette Bohg. Get a grip: Multi-finger grasp evaluation at scale enables robust sim-to-real transfer. *arXiv preprint arXiv:2410.23701*, 2024.

Baorui Ma, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Neural-pull: Learning signed distance functions from point clouds by learning to pull space onto surfaces. *arXiv preprint arXiv:2011.13495*, 2020.

Théo Morales, Omid Taheri, and Gerard Lacey. A versatile and differentiable hand-object interaction representation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 23–33. IEEE, 2025.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

Liyuan Qi, Olaoluwa Popoola, Muhammad Ali Imran, and Wasim Ahmad. Genhand: generalised human grasp kinematic retargeting. *npj Robotics*, 4(1):19, 2026.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. *arXiv preprint arXiv:2307.04577*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

Quanyuan Ruan, Jiabao Lei, Wenhao Yuan, Yanglin Zhang, Dekun Lu, Guiliang Liu, and Kui Jia. Prof. robot: Differentiable robot rendering without static and self-collisions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22562–22572. IEEE, 2025.

James E Smith and Robert L Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

Gergely Sóti, Xi Huang, Christian Wurll, and Björn Hein. dgrasp: Nerf-informed implicit grasp policies with supervised optimization slopes. *Robotics and Autonomous Systems*, 186:104921, 2025.

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pp. 581–600. Springer, 2020.

Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13253–13263. IEEE, 2022.

Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision*, pp. 572–589. Springer, 2022.

Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 5923–5930. IEEE, 2023.

Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *arXiv preprint arXiv:2210.02697*, 2022.

X Wang, Y Li, C Wang, and M Christie. Grasping motion generation through latent diffusion models. In *Computer Graphics Forum*, pp. e70423. Wiley Online Library, 2026.

Thomas Weng, David Held, Franziska Meier, and Mustafa Mukadam. Neural grasp distance fields for robot manipulation. *arXiv preprint arXiv:2211.02647*, 2022.

Zehang Weng, Haofei Lu, Danica Kragic, and Jens Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models. *IEEE Robotics and Automation Letters*, 9(12):11834–11840, 2024.

Xiaofei Wu, Tao Liu, Caoji Li, Yuexin Ma, Yujiao Shi, and Xuming He. Fastgrasp: Efficient grasp synthesis with diffusion. In *2025 International Conference on 3D Vision (3DV)*, pp. 735–747. IEEE, 2025.

Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision*, pp. 257–274. Springer, 2022.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4737–4746. IEEE, 2023.

Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20921–20930. IEEE, 2022.

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

Zhe Zhao, Mengshi Qi, and Huadong Ma. Decomposed vector-quantized variational autoencoder for human grasp generation. In *European Conference on Computer Vision*, pp. 447–463. Springer, 2024.

Tao Zhong, Jonah Buchanan, and Christine Allen-Blanchette. Grasp2grasp: Vision-based dexterous grasp translation via schrödinger bridges. *Advances in Neural Information Processing Systems*, 38:17044–17072, 2026.

Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22584–22594. IEEE, 2025.

Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2022.

Yufei Zhu, Yiming Zhong, Zemin Yang, Peishan Cong, Jingyi Yu, Xinge Zhu, and Yuexin Ma. Evolvinggrasp: Evolutionary grasp generation via efficient preference alignment. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11665–11674. IEEE, 2025.

A Representations and Architectures

A.1 Configuration Manifolds

To avoid adversarial minima in raw joint angles (Principle 4), gradient descent strictly operates on $\Theta = SE(3) \times \mathcal{Z}$. The manifold $\mathcal{Z}$ is defined by a small variational autoencoder (VAE) (Kingma & Welling, 2013) trained specific to each hand type. These manifolds serve solely to decode fingers during descent and to parameterize motion displacements; they do not act as density priors over grasps.

- **MANO:** Compresses 45 finger parameters into a 24-dimensional latent space using a hidden width of 512. This identical manifold serves both the static and reach-and-grasp tasks.
- **Shadow Hand:** Compresses 24 joint angles into a 16-dimensional latent space with a hidden width of 256 and a KL-divergence weight of 10^{-3}.
- **Retargeting:** Employs an object-agnostic, 32-dimensional shared latent space equipped with hand-specific encoders and decoders. It is trained on human demonstrations paired with optimization-based robot retargeting, utilizing contrastive alignment and dimensionless geometric fingertip descriptors to functionally align diverse embodiments with the human hand acting as a universal hub. During inference, a source grasp is mapped through this shared latent to initialize the target hand, and subsequent field descent optimizes within this compact space.

A.2 Energy Field Architecture and Training

The hand enters the field as 3D keypoints in the object's coordinate frame, computed via forward kinematics from $(g, \mathrm{dec}(z))$. We extract 16 skeleton joints for MANO, 28 keypoints for the Shadow hand, and the respective link frames for target hands in retargeting. Each keypoint is projected via a lightweight MLP and summed with a learned per-keypoint positional embedding.

A two-level PointNet++ (Qi et al., 2017) over the object's oriented point cloud reduces $2{,}048$ points to 128 tokens and one global code. Preliminary experiments indicated that utilizing a pre-trained point-cloud encoder benefited geometric criteria but degraded physical stability. Therefore, the encoder is initialized randomly and trained end-to-end jointly with the field on every task.

Keypoint tokens attend to the object tokens through cross-attention blocks and are modulated by the global object code via FiLM (Perez et al., 2018). A residual MLP trunk maps this pooled representation to the scalar energy. The human-hand field uses a hidden width of 512, depth of 10, and 4 cross-attention blocks. The Shadow-hand and retargeting fields use a width of 512, depth of 8, and 3 blocks.

Models are trained using the AdamW optimizer with a batch size of 512 using a Smooth-ℓ_1 regression loss against the standardized targets. The human-hand field trains for $16{,}000$ steps at a learning rate of 5×10^{-4}, followed by a $4{,}000$-step fine-tuning phase at 1.5×10^{-4} after the hard-negative mining round. Retargeting fields are trained for $60{,}000$ steps at 3×10^{-4}.

B Target Formulations and Corpus Construction

B.1 Target Formulation per Task

All task knowledge is injected through the scalar $d^\star = G + Q$ from Equation 2. All distances are in millimeters, and every target is divided by 10 to ensure fields regress quantities of order one. Let $D_g(\theta)$ denote the mean Euclidean distance between the forward-kinematic keypoints of θ and those of the nearest reference grasp. Let maxpen represent the deepest hand penetration into the object mesh, and $\mathrm{tipreach}$ the mean surface distance of the three closest fingertips beyond a $4\,\mathrm{mm}$ contact tolerance (yielding zero once three fingertips touch).

Human Hand (Static and Reach-and-Grasp). The target is demonstration-anchored and fully analytic. To quantify hand-object interpenetration (maxpen), we evaluate the signed distance from hand vertices to the object surface. For a vertex x near the object ($< 4\,\mathrm{mm}$), we project to the tangent plane of the nearest point-cloud neighbor. For distant points, we use the generalized winding

number (Barill et al., 2018) to avoid misleading normals on thin geometries. Building upon this, the volume-based penetration penalty is defined as:

$$P \;=\; \sum_{v} a_v \;\min\big(\,\max(-s(x_v) - \delta,\, 0),\; 5\,\mathrm{mm}\big) \quad [\mathrm{m}^3],$$

where a_v is the rest-pose area of the MANO vertex and δ is a margin tolerance ($\delta = 0$ for static grasps, $\delta = 1.5\,\mathrm{mm}$ for motion). At training, the target $d^\star$ integrates these criteria:

$$d^\star = D_g + 2\,\mathrm{maxpen} + 1.5\,\mathrm{tipreach}. \tag{3}$$

Here, D_g serves as the guidance term, while the geometric penalties act as the quality term. Because this explicit geometric penalty inherently conflicts with the physics-based penetration tolerances of IsaacGym, it is applied exclusively to the human-hand tasks.

Shadow Hand. The components of Equation 2 are unrolled into two sequential fields. The geometric field regresses the hand target (using DexGraspNet (Wang et al., 2022) grasps as references), compressed via $y = \mathrm{asinh}(d^\star/10)$ to prevent distant placements from dominating the loss. The physical field regresses a simulator-anchored margin. For each grasp, we apply a six-direction pull test in the simulator, recording s (the smallest breaking force in multiples of the object's weight). The target becomes:

$$d^\star = -\,s \;+\; \lambda\,\frac{\mathrm{tipreach}}{10}\,\mathbb{1}[s \le 0] \;+\; w\,\frac{\max(0,\; \mathrm{maxpen} - 30)}{10}, \qquad \lambda = w = 1. \tag{4}$$

The guidance term (tipreach) activates only where the physical margin carries no information ($s \le 0$). Penetration is penalized only beyond a generous 30 mm ceiling to prevent artificial barrier constraints from overriding the simulator's own preference for firm grips.

Retargeting. The margin is graded identically to the benchmark simulator. The object displacement δ_d under each of the six pushes is recorded, and the quality term isolates the worst direction:

$$y = \max\big(-4,\; 1 - \max_{d}\,\delta_d/20\,\mathrm{mm}\big), \qquad d^\star = -\,y + \lambda\,\frac{\mathrm{tipreach}}{10}\,\mathbb{1}[y \le 0], \qquad \lambda = 1. \tag{5}$$

A value of $y = 1$ indicates a perfectly rigid hold, $y = 0$ is the pass threshold, and -4 is the saturating floor enforcing bounded reliability (Principle 3).

B.2 Corpus Construction and Mining

To ensure the fitted landscape reliably guides descent (Principle 1), the training distribution explicitly covers the expected descent trajectories.

Each reference grasp is perturbed in the descent coordinates (wrist pose and manifold latent z) with a scale drawn log-uniformly between 0.01 and 1.2 standard deviations, ensuring coverage from sub-millimeter offsets to the far side of the object. For robot hands, wrist and finger scales are perturbed independently to decouple proximity from physical viability. To provide a global basin gradient, 25–30% of the corpus comprises far placements with random orientations on a spherical shell. The robot-hand corpora add 10% interpolated rows between references and far placements to explicitly label the descent path, and the Shadow-hand corpus adds 15% cross-object rows (grasps from other objects mapped as hard negatives).

The training corpora are constructed by sampling configurations across the state space of the target objects. Specifically, the human-hand fields are trained on approximately 30M configurations sampled around OakInk (Yang et al., 2022) objects and 10M around GRAB (Taheri et al., 2020) objects. For the Shadow hand, we generate 3.6M configurations to train the geometric field. The physical field is trained on 400k configurations sampled around DexGraspNet objects, comprising 200k simulated in-contact states and 200k free-space states. For retargeting, we construct a corpus by sampling approximately 3.5M configurations per target hand.

After the initial fit, we descend the field from random initializations on training objects. The resulting endpoints are jittered, re-labeled by the object SDF or simulator, and injected back into the corpus as hard negatives.

C Inference and Task Systems

C.1 Selection Strategy

In our work, candidate energies are only compared locally within a single query (Principle 4) to prevent optimizing for approximation errors across disjoint basins.

- **Static Human Synthesis:** The query evaluates 8 random initializations and simply returns the single lowest-energy endpoint.
- **Shadow Hand (Two-Stage):** The geometric field descends to convergence from random points on the spherical shell. Its 48 endpoints are then handed over to the physical field, triggering a fresh Adam optimizer run from each geometry-refined state. The top 16 endpoints within an energy margin of 0.2 from the global minimum are selected using farthest-point sampling in key-point space, ensuring the returned grasps span diverse kinematic families rather than collapsing into a single basin.
- **Retargeting:** Each source configuration yields 8 perturbed initialization endpoints, and the single lowest-energy refined state is returned.

C.2 Reach-and-Grasp Motion System

In our reach-and-grasp formulation, motion is generated auto-regressively as a sequence of discrete kinematic primitives. A single frame comprises 45 MANO axis-angle parameters and a 9-dimensional wrist pose variation relative to the window's start. Each motion primitive consists of $F = 8$ future frames conditioned on $H = 2$ history frames.

To map these sequences into a compact latent space, we employ a transformer VAE matching the architecture of MLD (Chen et al., 2023) with a hidden dimension of 256. A six-layer encoder processes distribution tokens alongside the history and future sequences to output a 64-dimensional latent. A six-layer decoder then reconstructs the future trajectory. To model temporal distribution, we employ a latent diffusion model (Rombach et al., 2022) as our kinematic prior (eight layers, hidden dimension of 512, sampled via 20 DDIM steps). Notably, this standalone prior processes no geometric information regarding the target object.

Object awareness is dynamically injected into the prior via the energy field. At each step, the field executes a 30-step Adam descent from the current hand configuration. The resulting 51-dimensional displacement vector is processed by FiLM layers integrated into the frozen prior. During training, the field selects the lowest-energy candidate out of 48 prior proposals that satisfies a geometric non-penetration gate. The FiLM adapter is then trained via distillation to reproduce this selected trajectory using only the displacement vector, making the field the sole conduit of spatial reasoning.

During inference, the auto-regressive generation employs a robust propose-and-refine mechanism. At each step, the adapted prior proposes a set of 8 candidate primitives. These candidates are evaluated and ranked using a combined scoring function: the field's energy evaluated at the boundary frame, plus a geometric affordance term that explicitly rewards fingertips landing near demonstrated contact regions. Once the optimal candidate is selected, the field operates directly as a refiner, executing a fixed-length gradient descent to optimize the final frame of this chosen primitive. To prevent kinematic jitter, this spatial correction is then smoothly blended backward into the primitive's preceding frames, ensuring a seamless trajectory. The episode successfully terminates when the refined field energy at a boundary frame falls below a strict per-object threshold (calibrated independently via the field's own descents on that specific object).

D Evaluation Metrics

Static Grasps, following the FastGrasp (Wu et al., 2025):

- **PenVol** (cm^3, $\downarrow$): The object mesh is voxelized into a $1\,\mathrm{mm}$ hollow surface shell. PenVol is the total volume of shell voxels whose centers fall within the hand mesh, averaged over grasps. Because the shell is hollow, this metric scales with the enclosed surface area rather than true intersection volume.

- **SimDisp** ($\mathrm{cm}, \downarrow$): The hand is held fixed while the object (represented as a 1 kg convex decomposition) is simulated under gravity in PyBullet for 35 steps ($1/240$ s per step). SimDisp is the mean object displacement. For reference, an ungrasped object in free fall drops 10.7 cm over this horizon.
- **Contact** ($\%, \uparrow$): The percentage of grasps with $\mathrm{PenVol} > 0$ (i.e., at least one shell voxel center lies inside the hand mesh).
- **Entropy** and **Cluster Size** ($\uparrow$): Grasp diversity metrics. The 21 hand keypoints are transformed into a canonical wrist frame, flattened to 63 dimensions, whitened, and clustered via k-means ($k = 20$). Entropy measures the cluster occupancy distribution (maximum $\ln 20 \approx 3.0$); Cluster Size is the mean distance from a grasp to its assigned centroid. Both metrics evaluate the pooled set of an entire protocol block.

Reach-and-Grasp, following LatentHOI (Li et al., 2025):

- **IV** ($\mathrm{cm}^3, \downarrow$): Interpenetration Volume. Calculated identically to PenVol using 5 mm hollow shell voxels, evaluated strictly on penetrating frames.
- **ID** ($\mathrm{cm}, \downarrow$): Interpenetration Depth. The maximum distance from an interpenetrating hand vertex to the object surface, averaged across penetrating frames.
- **CR** ($\%, \uparrow$): Contact Ratio. The fraction of hand vertices within an signed 0.45 mm distance of the object surface, averaged over the 50 frames.
- **IVU** ($\mathrm{cm}, \downarrow$): The ratio of IV to the total contact area per sequence, where contact area is CR scaled by the approximate MANO surface area ($407\,\mathrm{cm}^2$). Sequences with $\mathrm{CR} = 0$ are excluded.

Reach-and-Grasp, following Wang et al. (2026):

- **Depth** ($\mathrm{cm}, \downarrow$): The mean penetration depth across all 778 hand vertices. Vertices outside the object are counted as zero.
- **Ratio** ($\downarrow$): The fraction of hand vertices located inside the object mesh.
- **Score** ($\uparrow$): The Ferrari–Canny ϵ force-closure score of the final grasp. Contacts are defined as hand vertices within 5 mm of the surface, projected to the nearest surface point with its normal, and downsampled to a maximum of 64 via farthest-point sampling. Using an eight-edge friction cone ($\mu = 0.5$) and scaling torques by the object's largest radius, ϵ is the radius of the largest origin-centered ball inside the convex hull of the contact wrenches.
- **Valid** ($\%, \uparrow$): The fraction of episodes where $\mathrm{Score} > 0$.

Shadow-Hand Grasps, following Zhong et al. (2025). Simulation uses a convex-decomposed object, no gravity, friction of 2.0, and a fixed hand base. Following a 200-step joint-closing phase, the object is sequentially pushed along all six Cartesian directions (50 steps per direction, using a force $500\times$ the object volume) without state resets between directions.

- **Suc.6** and **Suc.1** ($\%, \uparrow$): A push is successful if the object displaces less than 2 cm. Suc.6 is the fraction of grasps passing all six directions; Suc.1 requires passing at least one direction.
- **Pen** ($\mathrm{mm}, \downarrow$): A point-cloud interpenetration estimate computed by comparing 64 surface points per hand link (excluding the forearm) against the object point cloud. It records the maximum inward projection along the nearest object point's normal, averaged over successful grasps.
- **Div** ($\mathrm{rad}, \uparrow$): The standard deviation of the 24 joint angles across successful grasps, averaged over all joints and valid objects.

Retargeting, following Zhong et al. (2026). Simulation uses no gravity, an object density of 10^4, and friction 10. Fingers are closed via a fixed 0.2 rad offset (Allegro, Barrett) or normal-force closure (Shadow), followed by six 50-step pushes at $5000\times$ the object volume.

- **Suc.** ($\%, \uparrow$): The percentage of grasps where the object displaces less than 2 cm under all six pushes.
- **Div.** ($\mathrm{rad}, \uparrow$): The standard deviation of joint angles over all generated grasps for a given direction. Because grasps are pooled *across* objects, this measures global dataset spread rather than per-object diversity, rendering raw values incomparable across different hand architectures.

E SIMULATION-LABEL NOISE

Simulator-anchored targets are inherently noisy; a rigid-body simulation of a grasp is not a perfectly smooth function of its initial configuration. Sub-millimeter changes in pose alter which contacts form and the solver's resolution order. This section quantifies this variance and outlines its methodological implications.

On the Shadow-hand corpus, we simulated 200k contact rows twice: once as a single settle, and once as the mean of five settles of jittered copies (2 mm, 0.01 rad). The single-settle margin exhibits a standard deviation of 0.135, while the five-settle mean drops to 0.100. The difference (0.084) indicates that roughly 38% of the variance of a single label is not reproducible by a local average. Two independent runs of IsaacGym on 100k rows agree on the pass/fail label only 77% of the time. For generated grasps resting deep in their basins, the simulator reproduces its own label on 81–85% of samples when the batch composition changes, causing aggregate success rates to fluctuate by up to 2 pp. Consequently, we treat success-rate differences below 3 pp as statistical ties.

This variance is concentrated exactly where precision matters most: near the minimum of a reference grasp. Far from the object, the label is zero or geometric, hence deterministic. However, inside the labeled contact bowl (Principle 2), the simulation "floor" is rough. Neighboring configurations exhibit label differences of the same magnitude as the actual gradient we aim to learn.

A field fitted to single-settle labels naturally regresses toward the local mean, smoothing this roughness, but it inevitably absorbs an approximation error proportional to the noise. As warned in Principle 4, globally ranking disparate candidates based on such a field falls prey to the optimizer's curse, selecting for noise rather than true physical superiority. Therefore, we compare energies only locally among multiple candidates. Furthermore, a field fitted to these labels cannot mathematically exceed the simulator's own self-agreement rate. Since the simulator only reproduces passing grasps roughly nine times out of ten, a portion of the gap between our reported success rates and 100% is an artifact of the physics engine, not a limitation of the energy field. In contrast, the demonstration-anchored targets used for human-hand tasks are deterministic and unaffected by this phenomenon.