STABLEGRASP: RECONSTRUCTING PHYSICALLY STABLE HUMAN HAND GRASPS FROM SINGLE IMAGES

Anonymous authors
Paper under double-blind review

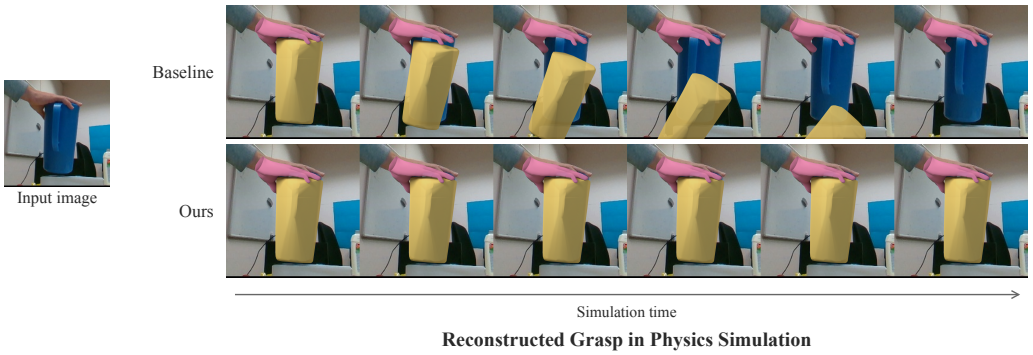

Figure 1: Given a single image of a human hand grasping an object, our method reconstructs the object, the pose of the human hand, and how the hand physically applies forces. Our reconstructed grasps are not only faithful to the images, but also physically stable.

ABSTRACT

Reconstructing a physically stable human grasp from a single RGB image is challenging because physically modeling grasps is itself difficult, and the problem requires estimating not only a visually constrained hand pose but also a control target that stabilizes the grasp. Existing methods either model only visual hand geometry without considering physics, or rely on less plausible physical modeling, which limits the physical validity of the resulting grasps. In this paper, we present StableGrasp, a differentiable simulation-based optimization framework that explicitly separates the visual hand pose from the control target that determines the grasping forces. Our method jointly optimizes hand geometry and control by minimizing the kinetic energy of the grasp in a differentiable simulator, while regularizing the hand geometry to preserve visual consistency and geometric plausibility. The reconstructed grasps are substantially more stable under rigorous physical simulation, while remaining visually consistent with the input images and geometrically plausible. Experiments show that our approach produces far more stable grasps than alternative hand-control strategies, benefiting visual-only grasp reconstruction pipelines by turning their outputs into physically stable grasps.

1 INTRODUCTION

Understanding how humans grasp objects is a fundamental problem at the intersection of embodied intelligence, computer vision, and computer graphics. Accurate reconstruction of hand-object grasps from images is needed for many applications including digital human modeling, AR/VR, robot learning from human demonstrations, and general understanding of dexterous manipulation.

Existing hand-object reconstruction methods have achieved impressive progress in recovering visually plausible grasps from images. Both optimization-based approaches (Liu et al., 2024) and feed-forward models (Chen et al., 2025; Pavlakos et al., 2024; Potamias et al., 2025; Wang et al., 2025) can now produce high-quality 3D hand poses with reasonable hand-object spatial relationships

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.

that are consistent with the input image. However, most of these methods stop at visual fidelity: the poses are evaluated in terms of reconstruction accuracy, pose consistency, and/or geometric plausibility (i.e. the hand is in contact with, but does not penetrate, the object). A grasp may look convincing when rendered in 3D while still being impossible to execute or physically unstable in reality.

More recent work (surveyed below in Section 2) recognizes the importance of physics in grasp reconstruction and synthesis. A common approach is to analyze stability of the grasped object in terms of force closure, which characterizes whether frictional contact forces at a static set of contact points could, in principle, hold the object in place against gravity. Such analysis does not address how the hand should actually be posed and actuated, and the corresponding grasp may still be impossible if the hand cannot produce and maintain the required contact forces.

We argue that stable grasp reconstruction is fundamentally a control problem, and cannot be treated as a visual or geometric optimization task divorced from inference of joint torques that hold the object stably in place over time. For example, a hand may appear to pinch a cup, yet if the pinch force is insufficient or unevenly distributed, the cup will slip or fall rather than remain firmly put. We therefore present StableGrasp, a method for monocular stable human grasp reconstruction that evaluates stability from the viewpoint of the coupled dynamics of the hand and object under simulation. StableGrasp also achieves strong visual stability: the reconstructed pose closely matches the image observations, and during simulation, the reconstructed grasp remains nearly motionless. To the best of our knowledge, StableGrasp is the first method to pursue this combination of rigorous physical stability and visual fidelity.

We evaluate our method on several established datasets (**DexYCB** (Chao et al., 2021), **HO3D** (Hampali et al., 2020), and **CORe50** (Lomonaco & Maltoni, 2017)) against baselines such as heuristic (Zhong et al., 2025) and learning-based (Christen et al., 2022) hand control. Our results show that compared to these baseline methods, StableGrasp substantially improves physical stability while producing reconstructed grasps that are visually consistent with the images and geometrically plausible.

In summary:

- we formulate monocular stable human grasp reconstruction with a notion of physical stability defined by forward simulation of the full hand-object system;
- we propose a two-phase physics-based refinement framework that introduces a kinetic-energy-driven optimization for stable grasp reconstruction, while preserving the visual and geometric fidelity of the reconstructed grasp;
- we use this framework to build a complete monocular reconstruction pipeline that recovers both the visual grasp configuration and a control target whose induced joint torques stabilize the grasp in physical simulation.

2 Related Work

Human Hand Grasp Reconstruction Recovering the 3D pose of a human hand and the grasped object from image observations has been extensively studied in computer vision. Methods for hand-only reconstruction estimate the pose and shape of the 3D hand (Boukhayma et al., 2019; Ge et al., 2019; Lin et al., 2021b;a; Pavlakos et al., 2024; Potamias et al., 2025). Methods for hand-object reconstruction also estimate the geometry of the object and the spatial relationship between hand and object. Methods attacking this problem have proposed using interaction constraints (Tekin et al., 2019; Hasson et al., 2019), contact-aware refinement (Grady et al., 2021), implicit hand-object representations (Chen et al., 2022; 2023b; Qi et al., 2024), unified hand-object estimation (Wang et al., 2025; Liu et al., 2024), or hand-guided reconstruction of held object geometry (Ye et al., 2022; Zhang et al., 2023; Li et al., 2023b; Fu et al., 2024; Aytekin et al., 2026; Chen et al., 2025; Wu et al., 2024). Many of these methods represent the hand using MANO (Romero et al., 2017), a parametric model that maps the pose and shape of the hand to a skinned mesh.

However, these methods only aim to recover visually-plausible hand and object poses that are consistent with the input observations, while StableGrasp's reconstruction is physically stable, verified by forward simulation of the coupled hand-object system.

Physics-aware Grasping Assuming Static Hands The stricter problem of recovering a grasp that is stable under gravity has also received significant attention. "Physical stability" has been modeled in several different ways. One simplified approach is to consider only *one-way* hand→ object coupling, and to analyze stability of the object in terms of contact forces exerted at a set of contact points where the object touches a static hand.

Classical grasping theory in robotics studies this problem through force closure and grasp wrench space analysis (Ferrari & Canny, 1992; Bicchi & Kumar, 2001), which evaluates whether contact forces can resist external object wrenches. This and similar quasi-static criteria have been widely used for dexterous grasp synthesis and evaluation (Liu et al., 2022; Zurbrügg et al., 2025; Wang et al., 2022; Zhang et al., 2024a; Chen et al., 2023a; Li et al., 2023a). Along similar lines, works like Grasp'D (Turpin et al., 2022) and Fast-Grasp'D (Turpin et al., 2023) optimize hand configurations with simulation-based stability objectives, while DeepSimHO (Wang et al., 2023) and UGG (Lu et al., 2024) use learned simulation surrogates to refine reconstructed interactions. Other methods introduce explicit contact forces at fingertips as optimization variables (Hu et al., 2022; Luo et al., 2024) or enforce physics-inspired priors on sliding and penetration at the contact points (Zhang et al., 2024b).

With a static hand, object stability does not establish whether an actuated hand can generate the required contact forces or maintain the grasp over time. In contrast, our approach models the full two-way coupling between an actuated hand and the object. We reconstruct not only a contact-consistent hand pose but also a control target whose induced joint torques realize those contacts and keep the grasp stable under forward simulation.

Physics-aware Grasping with Two-way Coupling A visually or geometrically plausible hand pose is not a stable grasp unless it is paired with an appropriate control strategy that holds the hand and object in place over time. As seen in a line of works in robotics (Chen et al., 2024; Jiang et al., 2024; Xu et al., 2023; Yuan et al., 2024; Lum et al., 2024), stable grasping is more fundamentally a control problem, not only a geometric optimization or pose estimation problem.

Several works incorporate hand control yet do not rigorously model the two-way coupling. Zhao et al. (2022) optimize a monocular reconstruction under a simulated stability criterion, holding the hand at the reconstructed pose with heuristic control that is never optimized, limiting the stability of the result. HOIC (Hu et al., 2024) imitates hand-object interaction in simulation, but it allows extra compensation force and torque applied directly to the object, so the object is not driven by two-way coupling physics alone. In contrast, D-Grasp (Christen et al., 2022) and Grasp-and-Lift (Choi et al., 2026) model strict two-way coupling: D-Grasp learns a policy that controls a hand through PD targets, and Grasp-and-Lift refines vision-aligned hand-object trajectories with physics-in-the-loop optimization. However, both focus on **dynamic** settings in which control can vary throughout time to maintain the grasp. Our method operates in a **static** setting: the unobserved control is represented by **one fixed PD target**. This imposes a distinct constraint on grasp reconstruction, since stability must be explained without freely changing the visible grasp or adapting the control target over a dynamic trajectory.

3 METHODOLOGY

Our goal is to reconstruct, from a single RGB image of a human hand holding an object in place, (a) the geometry of the object, (b) the shape and pose of the hand, and (c) a representation of the joint torques and contact forces that stabilizes the hand-object system in simulation.

To be considered successful, the reconstruction must be visually consistent with the hand and object observed in the image; but in addition the grasp must be *stable*: the reconstructed hand and joint torques must hold the object in place under frictional contact without noticeable motion of **either the hand or object**, as verified through physical simulation.

Representing the Hand and the Object We use MANO (Romero et al., 2017) as our visual hand model. A MANO hand mesh $\mathcal{M}_h^v = \mathcal{M}_{\text{MANO}}(\mathbf{q}_h^v, \boldsymbol{\beta}_h)$ is described by both hand shape parameters (e.g. finger length) $\boldsymbol{\beta}_h$ and hand pose $\mathbf{q}_h^v = (\mathbf{t}_h^v, \boldsymbol{\theta}_h^v) \in \mathbb{R}^{51}$, where $\mathbf{t}_h^v \in \mathbb{R}^3$ denotes the global translation of the hand and $\boldsymbol{\theta}_h^v \in \mathbb{R}^{48}$ consists of one 3D wrist rotation and fifteen 3D finger-joint

rotations (three joints on each of the five fingers). For the object, we aim to reconstruct a watertight manifold triangle mesh $\mathcal{M}_o$ representing its boundary.

The above representations are purely visual and geometric. They are not equipped with physical properties and are thus required to be made simulation-ready. For the hand, we use a proxy $\mathcal{H}(\mathbf{q}_h^v; \boldsymbol{\beta}_h)$: this proxy is an articulated rigid body parameterized by the same hand pose and shape parameters as the MANO hand. For the object, we assume it is a single rigid body with known uniform density ρ_o.

What is a Stable Grasp? A pair of simulation-ready hand and object alone is insufficient to evaluate whether a grasp is stable. The hand pose determines the points of contact with the object, but the contact forces at those points depend on the hand joint torques (i.e., how the hand is gripping the object) which are imperceptible in the input image. Reconstructing a stable grasp thus requires inferring these unknown joint torques as well.

Given the object's acceleration due to gravity $\mathbf{g} \in \mathbb{R}^3$ (which we assume is known), conceptually, a grasp is stable if the contact forces and torques exactly balance gravity (for the object) and the joint torques (for the hand), so that neither moves over time. There are however several complications when formalizing this conceptual notion of grasp stability:

First, simulating the grasp requires a choice of contact model, time integrator, and other simulation parameters. We choose the state-of-the-art MuJoCo (Todorov et al., 2012) library as our simulation engine. The full hand-object configuration at time step t is $\mathbf{x}_t = (\mathbf{q}_{h,t}, \mathbf{q}_{o,t})$, where $\mathbf{q}_{o,t} \in \mathbb{R}^6$ represents the translation and rotation of the object. Starting from the initial configuration $\mathbf{x}_0 = (\mathbf{q}_h^v, \mathbf{0})$ and zero initial configurational velocity, we evolve the system over time using MuJoCo for a horizon of T time steps, and declare that the grasp is stable if $\mathbf{x}_t \approx \mathbf{x}_0$ for $t \leq T$.

Second, even if a set of joint torques and their induced contact forces exactly balance gravity during the first simulation time step, applying those same torques at every time step throughout the time horizon T is not guaranteed to yield a stationary simulated trajectory $\mathbf{x}_t = \mathbf{x}_0$. Any slight force or torque residual (for example, due to floating-point or time-integration error) will start to accelerate the hand and object, so that $\mathbf{x}_t$ deviates increasingly from $\mathbf{x}_0$ over time. Following standard practice in robotics (Andrychowicz et al., 2020), instead of directly reconstructing fixed joint torques, we reconstruct a *target hand pose* $\mathbf{q}_h^{\text{target}} \in \mathbb{R}^{51}$. The joint torques are given by a PD (proportional-derivative) controller at each simulation timestep. Intuitively, it is a separate hand pose (typically intersecting the object) that the hand attempts to reach via joint actuation.

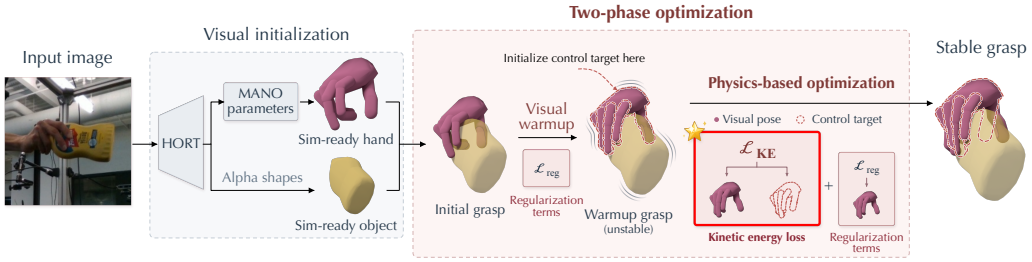

Figure 2: Pipeline overview. We first obtain a visual reconstruction of the hand and the object and make them simulation-ready. Then we refine the grasp with two-phase optimization. Starting from the second phase, a physics-based kinetic energy (KE) loss jointly optimizes the visual hand pose and the control target. The visual hand pose is also constrained by regularization terms.

3.1 Pipeline

Given a single RGB image I together with object density ρ_o and gravity $\mathbf{g}$, our pipeline reconstructs: a mesh $\mathcal{M}_o$ of the object; MANO hand shape parameters $\boldsymbol{\beta}_h$; and visual hand pose parameters $\mathbf{q}_h^v$, all of which are consistent with the image; and a target hand pose $\mathbf{q}_h^{\text{target}}$ such that the object and hand remain stationary after T time steps of physical simulation.

As shown in Fig. 2, our pipeline consists of two steps. *Visual initialization* estimates $\mathcal{M}_o$, $\boldsymbol{\beta}_h$, and $\mathbf{q}_h^v$ from the image without considering physics. *Differentiable physics-based optimization* (described below in Section 3.2) refines the estimated pose. Here we freeze $\mathcal{M}_o$ and $\boldsymbol{\beta}_h$ and seek a target pose $\mathbf{q}_h^{\text{target}}$ that explains grasp stability. The visual pose $\mathbf{q}_h^v$ is jointly optimized at the same time.

We use HORT (Chen et al., 2025) to perform the initial hand and object reconstruction. The hand branch of HORT predicts MANO hand pose $\mathbf{q}_h^v$ and shape parameters $\boldsymbol{\beta}_h$ from the input image; its object branch predicts a 3D object point cloud. We then reconstruct a closed object surface from the point cloud using alpha-shape meshing (Edelsbrunner & Mucke, 1994), and convexly decompose the mesh using CoACD (Wei et al., 2022) to make it simulation-ready in MuJoCo.

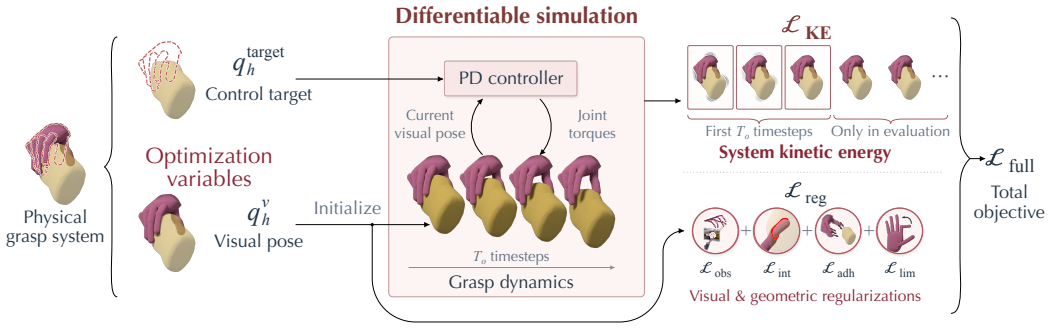

Figure 3: Differentiable physics-based optimization. The visual hand pose $\mathbf{q}_h^v$ initializes the hand–object system, while the fixed control target $\mathbf{q}_h^{\text{target}}$ drives a PD controller during simulation. During physics-based refinement, both poses are jointly optimized using the total hand–object kinetic energy accumulated over the first T_o timesteps and visual and geometric regularization of the visual pose.

3.2 Differentiable Physics-based Optimization

As shown in Fig. 3, the second step of our pipeline jointly refines the visual hand pose $\mathbf{q}_h^v$ and computes the target hand pose $\mathbf{q}_h^{\text{target}}$. We do so by optimizing an objective function that directly measures stability, along with visual and geometric regularizers that keep the hand pose consistent with the image and free of penetrations into the object volume.

Kinetic Energy Objective Recall that a visual pose and target pose pair $\left(\mathbf{q}_h^v, \mathbf{q}_h^{\text{target}}\right)$ is stable if the generalized configuration changes negligibly between time steps: $\mathbf{x}_{t+1} \approx \mathbf{x}_t \approx \mathbf{x}_0$. Since the natural, coordinate-free metric on configurational displacements is kinetic energy, we use the total kinetic energy of the simulated hand-object system, integrated over T_o steps of forward simulation (where $T_o \ll T$), to quantify stability. We use a much shorter horizon because instability emerges within the first few timesteps. This also makes efficient backpropagation possible in practice.

Let $\dot{\mathbf{x}}_t$ denote the system's configurational velocity, and $\mathbf{M}(\mathbf{x}_t)$ denote the generalized mass matrix (which includes both mass and inertia tensors for the hand and the object). We then accumulate kinetic energy over the first T_o simulation steps and define the physical stability term as

$$\mathcal{L}_{\text{KE}} = \sum_{t=1}^{T_o} \frac{1}{2} \dot{\mathbf{x}}_t^\top \mathbf{M}(\mathbf{x}_t) \dot{\mathbf{x}}_t.$$

Visual and Geometric Objectives We regularize the visual hand pose $\mathbf{q}_h^v$ with four terms that keep it consistent with the image and free of penetration into the object. A 2D reprojection term $\mathcal{L}_{\text{obs}}$ penalizes the L_1 distance between the projected 3D hand joints and the 2D joint locations estimated by HORT. With N points sampled on the hand proxy $\mathcal{H}$ and computing their signed distances to the object surface, an interpenetration term $\mathcal{L}_{\text{int}}$ penalizes points lying inside the object, while an adhesive term $\mathcal{L}_{\text{adh}}$ (following prior visual grasp optimization methods like (Liu et al., 2024)) pulls exterior non-palm points toward the surface to encourage contact formation. Palm points are excluded because the rigid proxy palm does not contribute to stability the way a deformable human

palm does. A joint-limit term $\mathcal{L}_{\text{lim}}$ discourages implausible configurations while leaving feasible poses unaffected. We combine the four terms with associated weights into the regularization loss $\mathcal{L}_{\text{reg}}$. Exact formulae are given in the supplementary material.

Two-phase Optimization We use the visual hand pose provided by HORT in the Visual Initialization step as the initial guess for $\mathbf{q}_h^v$. We then optimize in two phases. In the first phase, we perform a visual warm-up in which only $\mathcal{L}_{\text{reg}}$ is minimized. This stage reduces severe penetrations and establishes a visual hand pose with plausible hand-object contact before simulation-based refinement. During this phase, $\mathbf{q}_h^v$ is updated to $\mathbf{q}_h^{\text{warmup}}$. In the second phase, we initialize the target pose $\mathbf{q}_h^{\text{target}}$ to be $\mathbf{q}_h^{\text{warmup}}$, activate the kinetic-energy term, and optimize the full objective

$$\mathcal{L}_{\text{full}} = \lambda_{\text{KE}} \frac{\mathcal{L}_{\text{KE}}(\mathbf{q}_h^v, \mathbf{q}_h^{\text{target}})}{\mathcal{L}_{\text{KE}}^{(0)}} + \lambda_{\text{reg}} \frac{\mathcal{L}_{\text{reg}}(\mathbf{q}_h^v)}{\mathcal{L}_{\text{reg}}^{(0)}}.$$

Here, $\mathcal{L}_{\text{KE}}^{(0)}$ and $\mathcal{L}_{\text{reg}}^{(0)}$ are the values of the two loss bundles evaluated once at the beginning of the second phase, i.e., from the warm-up result. This normalization compensates for any scale mismatch between the two loss terms, so that their relative influence is controlled by λ_{reg} and λ_{KE} rather than by raw loss magnitude alone.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate on DexYCB (Chao et al., 2021), HO3D (Hampali et al., 2020), and CORe50 (Lomonaco & Maltoni, 2017), using the HORT (Chen et al., 2025) test splits for DexYCB and HO3D. Our benchmark contains 242 images from 81 DexYCB sequences, 100 frames from 10 HO3D sequences, and 20 manually selected non-laboratory CORe50 images with diverse objects and hand poses. DexYCB images are final frames from selected camera views with clear hand–object visibility. HO3D frames are temporally subsampled near sequence ends and can depict ongoing manipulation. Details on dataset and how we sample test cases are provided in the supplementary material.

Object densities follow YCB statistics (Calli et al., 2015) for DexYCB and HO3D. For CORe50, which provides neither 3D annotations nor object masses, we use $1000\,\text{kg}/\text{m}^3$. Gravity is approximated by the camera-down direction with magnitude $9.8\,\text{m}/\text{s}^2$.

Implementation Details We use MuJoCo XLA (MJX) (Todorov et al., 2012; Bradbury et al., 2018) and automatically generate each hand proxy from the HORT-estimated MANO shape using a fixed template, with a box palm (Christen et al., 2022) and capsule finger links (Bogo et al., 2016). Each refinement iteration differentiates through $T_o = 20$ simulation steps at 2 ms. Physics-based optimization early-stops after 200 iterations without sufficient loss improvement. Evaluation uses 0.2 ms steps over $T = 5$ s to reduce integration errors, with unified PD gains. PPO's detailed setup and remaining settings are specified in the supplementary material.

4.2 Metrics

Physical Stability We measure hand motion from initial to final states by 3D MPJPE (Ionescu et al., 2013) between the initial and final joint positions of the simulation, together with net object center-of-mass translation and rotation angle. Success Rate is the fraction of grasps whose object center remains within 1 cm of its initial position throughout simulation. Drop Time is the mean time to first exceed this threshold among failed grasps.

Visual Stability We compare initial and final image projections using object-mask IoU and hand PCK (Yang & Ramanan, 2012). PCK@τ is the fraction of the 21 projected joints displaced by less than τ in normalized image coordinates.

Reconstruction Quality For DexYCB and HO3D, we compare $\mathbf{q}_h^v$ with their annotations using 3D MPJPE and 2D PCK. These MANO fits to RGB-D observations (Chao et al., 2021; Hampali

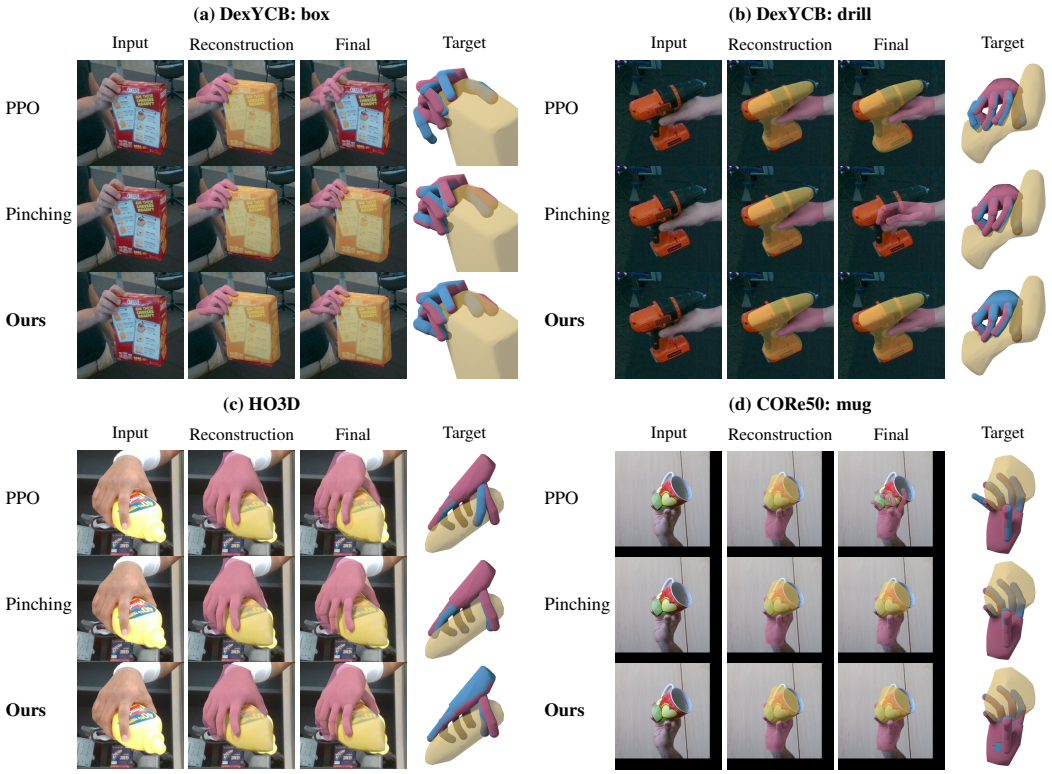

Table 1: Qualitative comparisons. Examples from all three datasets show object drops or visible hand–object motion with PPO and Pinching, while our method preserves stable grasps. Even when the baselines retain the object, their larger shifts are readily noticeable, especially in the supplementary video. The target visualizations (control targets shown in blue) show how our method flexibly adjusts the control target to achieve stability, compared with the prescribed inward displacement of Pinching.

et al., 2020) do not account for grasp dynamics and can contain interpenetration (Wang et al., 2023; Hampali et al., 2021). Contact depends on fine surface geometry and tissue deformation (Grady et al., 2021); stable contact with our rigid proxy and reconstructed object can therefore require pose adjustments relative to these visual references.

4.3 Baseline Comparisons

We compare StableGrasp with alternative approaches for stabilizing reconstructed grasps through hand control: heuristic pinching, and a learned instance-specific PPO baseline. Each baseline specifies how the hand is controlled, thus enabling comparison of grasp stability under forward simulation of the coupled hand–object system. For all baselines, we initialize the hand pose to that computed by our visual warm-up optimization.

Heuristic *Pinching* A commonly-used (Wang et al., 2022; Liu et al., 2024) heuristic for stabilizing a grasp without physics-driven optimization is to specify a target hand pose $\mathbf{q}_h^{\text{target}}$ where each finger has been displaced some fixed distance into the object along the contact normal direction. A PD controller is then used to compute joint torques from the target pose. We adapt the approach of DexGraspAnything (Zhong et al., 2025) as a representative method in this category.

PPO We study learned grasp stabilization with an instance-specific PPO baseline adapted from D-Grasp (Christen et al., 2022). Each policy starts from the warm-up grasp, is individually trained and is allowed to update finger PD targets online over time in the evaluation simulation. In our experiments, we drop out 5 examples where policy training fails with numerical errors; Appendix D details the adaptation, training, and evaluation.

Table 2: Physical and visual stability (mean ± standard deviation) on 357 common examples (DexYCB: 239; HO3D: 99; CORe50: 19), excluding five PPO training failures. All methods use a 1 cm displacement threshold. Motion, IoU, and PCK compare initial and final saved states of successful grasps; Success Rate uses all examples. Our method achieves substantially higher success rates while maintaining much more stable hand and object poses.

Dataset	Method	MPJPE (mm) ↓	Obj. trans. (mm) ↓	Obj. rot. (°) ↓	Object IoU ↑	PCK ↑ @0.025	PCK ↑ @0.05	Success rate ↑	Drop time (s) ↑
DexYCB	PPO	4.6 ± 2.5	4.5 ± 2.8	6.40 ± 4.57	0.88 ± 0.09	0.86 ± 0.10	0.93 ± 0.07	39.33%	0.28 ± 0.70
	Pinching	2.7 ± 0.8	5.9 ± 1.9	4.66 ± 2.79	0.87 ± 0.08	0.95 ± 0.06	**1.00** ± 0.01	64.02%	0.52 ± 0.91
	Ours	**0.2** ± 0.2	**1.2** ± 1.9	**1.07** ± 1.57	**0.97** ± 0.05	**1.00** ± 0.00	**1.00** ± 0.00	**87.45%**	**0.80** ± 1.09
HO3D	PPO	4.2 ± 1.5	5.8 ± 2.2	9.87 ± 5.18	0.84 ± 0.05	0.81 ± 0.10	0.92 ± 0.08	32.32%	**0.59** ± 0.94
	Pinching	2.6 ± 0.6	5.3 ± 1.8	3.40 ± 2.47	0.89 ± 0.05	0.93 ± 0.07	**1.00** ± 0.00	63.64%	0.53 ± 1.06
	Ours	**0.2** ± 0.2	**1.2** ± 1.4	**0.88** ± 1.03	**0.97** ± 0.03	**1.00** ± 0.00	**1.00** ± 0.00	**88.89%**	0.36 ± 0.64
CORe50	PPO	5.2 ± 1.7	3.5 ± 1.2	7.52 ± 7.12	0.91 ± 0.04	0.83 ± 0.06	0.90 ± 0.04	47.37%	0.51 ± 0.98
	Pinching	2.3 ± 0.8	6.1 ± 1.4	4.03 ± 1.76	0.86 ± 0.06	0.98 ± 0.04	**1.00** ± 0.00	73.68%	0.28 ± 0.38
	Ours	**0.3** ± 0.5	**1.4** ± 2.1	**2.27** ± 4.30	**0.96** ± 0.07	**1.00** ± 0.00	**1.00** ± 0.00	**78.95%**	**1.27** ± 1.95

Table 3: Hand reconstruction quality before optimization (HORT), after visual warm-up, and after the full pipeline (mean ± standard deviation).

Dataset	Stage	MPJPE-3D (mm) ↓	PCK@0.025 ↑	PCK@0.05 ↑
DexYCB	HORT	7.7 ± 3.2	0.68 ± 0.20	0.92 ± 0.11
	After warm-up	11.0 ± 3.8	0.62 ± 0.20	0.91 ± 0.12
	Ours	11.3 ± 3.8	0.60 ± 0.20	0.89 ± 0.12
HO3D	HORT	5.6 ± 3.9	0.72 ± 0.24	0.94 ± 0.11
	After warm-up	10.7 ± 4.5	0.60 ± 0.22	0.88 ± 0.14
	Ours	11.0 ± 4.4	0.59 ± 0.22	0.87 ± 0.15

4.3.1 Qualitative Results

Table 1 compares the input images, reconstructed grasps, final simulated states, and a visualization of the optimized target pose. Not dropping an object alone does not certify a stable grasp, because a hand that drifts noticeably is not the stable hand in the image. StableGrasp substantially reduces motion of *both the hand and object* compared with the baselines. We strongly encourage readers to watch the supplementary video: small shifts that are difficult to discern in static images become readily apparent during playback, making the improvement in grasp stability particularly clear.

4.3.2 Quantitative Results

Table 2 shows that StableGrasp achieves higher success rates and lower hand and object motion than both baselines on all three datasets, together with better image-space stability. Note that Success Rate does not tell the full story of grasp stability: an object can significantly shift over time away from its pose as depicted in the input image without dropping completely. The metrics in Table 2 demonstrate that StableGrasp successfully reconstructs grasps that remain stable under rigorous forward physics simulation.

Table 3 quantifies how the hand pose changes along the pipeline. Most of the changes occur in the visual warm-up, while the physics-based refinement moves the pose marginally. This is due to the differences among our simulation-ready hand proxy, the reconstructed object, and the dataset annotations used as the reference, so a physically valid grasp need not coincide exactly with the annotated pose. This is a reasonable consequence of pursuing geometric and physical plausibility and stability, while the qualitative results show that our reconstructions remain visually faithful to the input images.

4.4 Ablation Studies

We ablate warm-up on 86 examples (61 DexYCB, 25 HO3D), with equal maximum budgets: 600 warm-up plus 2000 physics-based iterations, or 2600 physics-based iterations without warm-up. Both use early stopping as in our experiments in Section 4.3. Removing warm-up reduces success from 90.70% to 76.74% and increases hand motion on both datasets (Table 4), demonstrating the benefit of contact initialization.

Table 4: Visual warm-up ablation on 86 test cases. Metrics and aggregation follow Table 2; values are mean ± standard deviation.

Dataset	Method	MPJPE (mm) ↓	Obj. trans. (mm) ↓	Obj. rot. (°) ↓	Object IoU ↑	PCK ↑ @0.025	Success rate ↑	Drop time (s) ↑
DexYCB	No warm-up	0.4 ± 0.6	**1.1** ± 1.7	1.15 ± 2.17	**0.98** ± 0.04	**1.00** ± 0.00	78.69%	**0.07** ± 0.05
	Warm-up	**0.2** ± 0.2	**1.1** ± 2.1	**0.94** ± 1.56	0.97 ± 0.04	**1.00** ± 0.00	**91.80%**	0.06 ± 0.01
HO3D	No warm-up	0.3 ± 0.3	**1.1** ± 1.9	**0.88** ± 1.47	**0.98** ± 0.05	**1.00** ± 0.00	72.00%	0.04 ± 0.01
	Warm-up	**0.2** ± 0.3	1.3 ± 1.4	0.93 ± 1.02	0.97 ± 0.04	**1.00** ± 0.00	**88.00%**	**0.05** ± 0.00

Figure 4 further shows how each warm-up regularizer contributes to a plausible initial grasp.

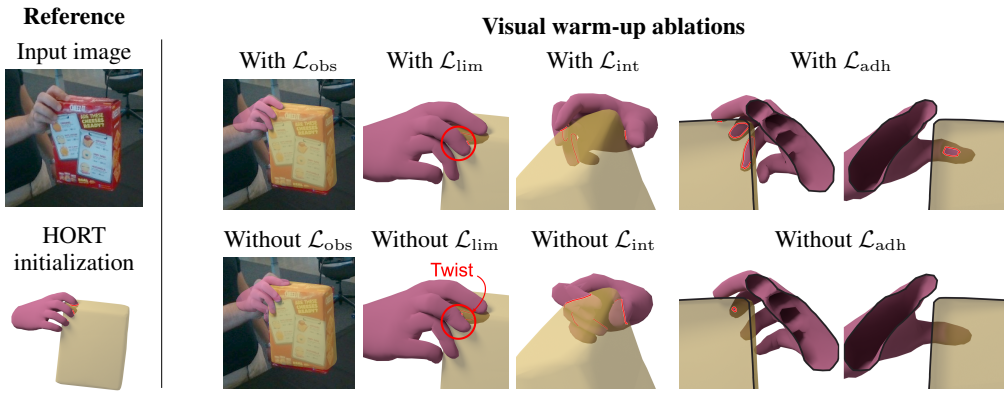

Figure 4: Visual warm-up ablations. Left: input and HORT initialization as references. Right: complete warm-up (top) and individual regularizer ablations (bottom), with matching views in each column. Without $\mathcal{L}_{\text{obs}}$, the reconstructed grasp drifts from the input image. Without $\mathcal{L}_{\text{lim}}$, the middle finger's distal joint exhibits severe, biomechanically implausible twisting. Without $\mathcal{L}_{\text{int}}$, the hand penetrates deeply into the object. Without $\mathcal{L}_{\text{adh}}$, the fingers make insufficient contact with the object.

5 Conclusion

We presented StableGrasp, a method for reconstructing stable human grasps from single RGB images. By jointly optimizing the visual hand pose and a PD control target, it minimizes the kinetic energy of the coupled hand–object system while enforcing visual and geometric constraints. Experiments demonstrate improved grasp stability over baselines while preserving agreement with the input image.

Limitations and Future Work Our local pose optimization keeps object geometry fixed and therefore depends on the initial reconstruction. Alpha-shape meshing can distort thin objects and fill concave or hollow features. Object reconstruction errors account for 27 out of all our 48 failure cases (on the 362 examples we tested). The rigid hand proxy also omits local tissue deformation, which affects contact geometry and force transmission. Future work could combine improved object reconstruction with deformable hand models.

AI USE STATEMENT

In this work, we used generative AI tools to help brainstorm and refine research ideas and to help implement parts of our method's code. In both cases, the authors reviewed and tested the outputs, and the research direction, technical contributions, and final implementation were decided and verified by the authors. We did not use generative AI tools to generate synthetic datasets, to develop theoretical results, or to formulate or prove any mathematical claims.

We additionally used generative AI tools to search for and summarize related work, to draft portions of the supplementary material, and to polish the wording of individual sentences. The main text was written by the authors, with generative AI used only for sentence-level polishing that was subsequently checked; all related-work findings were reviewed and summarized by the authors, and all supplementary content was reviewed and verified by the authors. We have reviewed all AI-assisted work and take full responsibility for the final content of this paper, including its text, claims, and artifacts.

REFERENCES

OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *Int. J. Rob. Res.*, 39(1):3–20, January 2020. ISSN 0278-3649. doi: 10.1177/0278364919887447. URL https://doi.org/10.1177/0278364919887447.

Ayce Idil Aytekin, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Follow my hold: Hand-object interaction reconstruction through geometric guidance. *International Conference on 3D Vision*, 2026.

Antonio Bicchi and Vijay R. Kumar. Robotic grasping and manipulation. 2001. URL https://api.semanticscholar.org/CorpusID:14282573.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 561–578, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46454-1.

Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10843–10852, 2019.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics and Automation Magazine*, 22(3):36–52, 2015. ISSN 1070-9932. doi: 10.1109/mra.2015.2448951. URL http://dx.doi.org/10.1109/MRA.2015.2448951.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jiayi Chen, Yuxing Chen, Jialiang Zhang, and He Wang. Task-oriented dexterous grasp synthesis via differentiable grasp wrench boundary estimator. *arXiv preprint arXiv:2309.13586*, 2023a.

Sirui Chen, Jeannette Bohg, and C. Karen Liu. Springgrasp: An optimization pipeline for robust and compliant dexterous pre-grasp synthesis, 2024.

Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. AlignSDF: Pose-Aligned signed distance fields for hand-object reconstruction. In *ECCV*, 2022.

Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gSDF: Geometry-Driven signed distance functions for 3D hand-object reconstruction. In *CVPR*, 2023b.

Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. HORT: Monocular hand-held objects reconstruction with transformers. In *ICCV*, 2025.

Byeonggyeol Choi, Woojin Oh, and Jongwoo Lim. Grasp-and-lift: Executable 3d hand-object interaction reconstruction via physics-in-the-loop optimization, 2026. URL https://arxiv.org/abs/2601.18121.

Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Herbert Edelsbrunner and Ernst Mucke. Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13, 10 1994. doi: 10.1145/147130.147153.

Carlo Ferrari and John F. Canny. Planning optimal grasps. *Proceedings 1992 IEEE International Conference on Robotics and Automation*, pp. 2290–2295 vol.3, 1992. URL https://api.semanticscholar.org/CorpusID:32592111.

Bowen Fu, Gu Wang, Chenyangguang Zhang, Yan Di, Ziqin Huang, Zhiying Leng, Fabian Manhardt, Xiangyang Ji, and Federico Tombari. D-sco: Dual-stream conditional diffusion for monocular hand-held object reconstruction, 2024. URL https://arxiv.org/abs/2311.14189.

Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020.

Shreyas Hampali, Sayan Deb Sarkar, and Vincent Lepetit. Ho-3d_v3: Improving the accuracy of hand-object annotations of the HO-3D dataset. *CoRR*, abs/2107.00887, 2021. URL https://arxiv.org/abs/2107.00887.

Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

Haoyu Hu, Xinyu Yi, Hao Zhang, Jun-Hai Yong, and Feng Xu. Physical interaction: Reconstructing hand-object interactions with physics. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22 Conference Papers, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394703. doi: 10.1145/3550469.3555421. URL https://doi.org/10.1145/3550469.3555421.

Haoyu Hu, Xinyu Yi, Zhe Cao, Junhai Yong, and Feng Xu. Hand-object interaction controller (hoic): Deep reinforcement learning for reconstructing interactions with physics. *ACM SIGGRAPH 2024 Conference Papers*, 2024. URL https://api.semanticscholar.org/CorpusID:269604838.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

Yongpeng Jiang, Mingrui Yu, Xinghao Zhu, Masayoshi Tomizuka, and Xiang Li. Contact-implicit model predictive control for dexterous in-hand manipulation: A long-horizon and robust approach, 2024. URL https://arxiv.org/abs/2402.18897.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Albert H. Li, Preston Culbertson, Joel W. Burdick, and Aaron D. Ames. Frogger: Fast robust grasp generation via the min-weight metric. *arxiv:2302.13687*, February 2023a.

Kailin Li, Lixin Yang, Haoyu Zhen, Zenan Lin, Xinyu Zhan, Licheng Zhong, Jian Xu, Kejian Wu, and Cewu Lu. Chord: Category-level hand-held object reconstruction via shape deformation, 2023b. URL https://arxiv.org/abs/2308.10574.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021a.

Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021b.

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, January 2022. ISSN 2377-3774. doi: 10.1109/lra.2021.3129138. URL http://dx.doi.org/10.1109/LRA.2021.3129138.

Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. *arXiv preprint arXiv:2411.14280*, 2024.

Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78, pp. 17–26, 2017.

Jiaxin Lu, Hao Kang, Haoxiang Li, Bo Liu, Yiding Yang, Qixing Huang, and Gang Hua. Ugg: Unified generative grasping. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXVII*, pp. 414–433, Berlin, Heidelberg, 2024. Springer-Verlag.

Tyler Ga Wei Lum, Martin Matak, Viktor Makoviychuk, Ankur Handa, Arthur Allshire, Tucker Hermans, Nathan D. Ratliff, and Karl Van Wyk. DextrAH-g: Pixels-to-action dexterous arm-hand grasping with geometric fabrics. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=S2Jwb0i7HN.

Haowen Luo, Yunze Liu, and Li Yi. Physics-aware hand-object interaction denoising, 2024. URL https://arxiv.org/abs/2405.11481.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. 2025.

Haozhe Qi, Chen Zhao, Mathieu Salzmann, and Alexander Mathis. Hoisdf: Constraining 3d hand-object pose estimation with global signed distance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10392–10402, 2024.

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.

Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions, 2019. URL https://arxiv.org/abs/1904.05349.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands, 2022. URL https://arxiv.org/abs/2208.12250.

Dylan Turpin, Tao Zhong, Shutong Zhang, Guanglei Zhu, Eric Heiden, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Fast-grasp'd: Dexterous multi-finger grasp generation through differentiable simulation. In *ICRA*, 2023.

Rong Wang, Wei Mao, and Hongdong Li. DeepsimHO: Stable pose estimation for hand-object interaction via physics simulation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=SxVHyYavHy.

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.

Yinqiao Wang, Hao Xu, Pheng-Ann Heng, and Chi-Wing Fu. Unihope: A unified approach for hand-only and hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12231–12241, June 2025.

Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics (TOG)*, 41(4):1–18, 2022.

Jane Wu, Georgios Pavlakos, Georgia Gkioxari, and Jitendra Malik. Reconstructing hand-held objects in 3d. *arXiv preprint arXiv:2404.06507*, 2024.

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. *arXiv preprint arXiv:2303.00938*, 2023.

Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012.

Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. 2022.

Haoqi Yuan, Bohan Zhou, Yuhui Fu, and Zongqing Lu. Cross-embodiment dexterous grasping with reinforcement learning. *arXiv preprint arXiv:2410.02479*, 2024.

Chenyangguang Zhang, Yan Di, Ruida Zhang, Guangyao Zhai, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ddf-ho: Hand-held object reconstruction via conditional directed distance field, 2023. URL https://arxiv.org/abs/2308.08231.

Jialiang Zhang, Haoran Liu, Danshi Li, XinQiang Yu, Haoran Geng, Yufei Ding, Jiayi Chen, and He Wang. Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In *8th Annual Conference on Robot Learning*, 2024a.

Shutong Zhang, Yi-Ling Qiao, Guanglei Zhu, Eric Heiden, Dylan Turpin, Jingzhou Liu, Ming Lin, Miles Macklin, and Animesh Garg. Handypriors: Physically consistent perception of hand-object interactions with differentiable priors. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024b. doi: 10.1109/ICRA57147.2024.10610748.

Zimeng Zhao, Binghui Zuo, Wei Xie, and Yangang Wang. Stability-driven Contact Reconstruction From Monocular Color Images . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1633–1643, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.00169. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00169.

Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. *arXiv preprint arXiv:2503.08257*, 2025.

René Zurbrügg, Andrei Cramariuc, and Marco Hutter. Graspqp: Differentiable optimization of force closure for diverse and robust dexterous grasping. In *Conference on Robot Learning (CoRL)*, 2025. URL https://graspqp.github.io/.

Appendix

A PD CONTROLLER

We simulate the hand as an articulated rigid-body system with 51 actuated coordinates: three wrist translations, three wrist rotations, and 45 finger rotations. MANO represents each local rotation by an axis-angle vector. Before simulation, we convert these rotations to the corresponding ordered hinge rotations of the articulated hand. The controller operates in these simulation coordinates. Let $\widetilde{\mathbf{q}}_{h,t}$ denote the current hand coordinates and $\widetilde{\mathbf{q}}_h^{\text{target}}$ the control target. At every simulation step, we apply

$$\boldsymbol{\tau}_t^{\text{raw}} = \mathbf{K}_p\big(\widetilde{\mathbf{q}}_h^{\text{target}} - \widetilde{\mathbf{q}}_{h,t}\big) - \mathbf{K}_d\dot{\widetilde{\mathbf{q}}}_{h,t}, \tag{1}$$

$$\boldsymbol{\tau}_t = \text{clip}_{[-\boldsymbol{\tau}_{\max},\boldsymbol{\tau}_{\max}]}(\boldsymbol{\tau}_t^{\text{raw}}). \tag{2}$$

Here, $\mathbf{K}_p$ and $\mathbf{K}_d$ are diagonal, and the desired velocity is zero. The translational entries of $\boldsymbol{\tau}_t$ are forces; the rotational entries are torques. Table 5 gives the gains and actuation limits used in the MuJoCo simulator.

Table 5: PD gains and actuation limits for the MuJoCo/MJX hand.

Coordinates	K_p	K_d	Actuation limit
Wrist translation	5×10^4	5×10^3	100 N
Wrist rotation	5×10^2	5×10^1	15 N m
Index, middle, ring, and little fingers	10	1	2 N m
Thumb	10	1	7 N m

The visual pose initializes the physical hand, with zero hand and object velocities. The target pose remains fixed throughout each simulation.

B VISUAL AND GEOMETRIC REGULARIZATION

All four regularization terms act on the visual pose $\mathbf{q}_h^v$. Hand shape and object geometry remain fixed. Let $\mathbf{p}_i(\mathbf{q}_h^v)$, $i = 1, \ldots, N$, be surface samples on the articulated hand proxy, and let d_i be their signed distances to the object surface, measured in meters and negative inside the object. We write $[z]_+ = \max(z, 0)$. The combined regularizer is

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{obs}}\mathcal{L}_{\text{obs}} + \lambda_{\text{int}}\mathcal{L}_{\text{int}} + \lambda_{\text{adh}}\mathcal{L}_{\text{adh}} + \lambda_{\text{lim}}\mathcal{L}_{\text{lim}}. \tag{3}$$

The formulas below retain the reductions and constant factors used in our implementation, so that they correspond to the weights reported in Table 8.

B.1 REPROJECTION TERM $\mathcal{L}_{\text{obs}}$

Let $\mathbf{u}_j$ be the 2D position of joint j estimated by HORT, $\mathbf{J}_j(\mathbf{q}_h^v)$ the corresponding 3D hand joint, and π the camera projection into the input crop. We use

$$\mathcal{L}_{\text{obs}} = \sum_{j=1}^{21} w_j \left\| \pi\big(\mathbf{J}_j(\mathbf{q}_h^v)\big) - \mathbf{u}_j \right\|_1, \qquad w_j = 10. \tag{4}$$

Coordinates are measured in pixels. The same joint weights are used for all 21 joints. These estimated image observations are fixed during both optimization phases.

B.2 INTERPENETRATION TERM $\mathcal{L}_{\text{int}}$

We penalize the squared depth of samples inside the object:

$$\mathcal{L}_{\text{int}} = \frac{1}{N}\sum_{i=1}^{N}[-d_i]_+^2. \tag{5}$$

This term includes both palm and finger samples. Exterior samples contribute zero, and the averaging is over all N samples.

B.3 Adhesion Term $\mathcal{L}_{\text{adh}}$

To bring separated fingers toward the object, we penalize their exterior distance:

$$\mathcal{L}_{\text{adh}} = c_{\text{adh}} \sum_{i=1}^{N} a_i [d_i]_+, \qquad c_{\text{adh}} = 1.25, \qquad a_i = \begin{cases} 1, & \mathbf{p}_i \text{ lies on a finger link,} \\ 0, & \mathbf{p}_i \text{ lies on the palm.} \end{cases} \tag{6}$$

B.4 Joint-limit Term $\mathcal{L}_{\text{lim}}$

Let θ_k be an axis-angle component of the MANO pose, with lower and upper bounds ℓ_k and u_k. We use

$$\mathcal{L}_{\text{lim}} = \sum_k \left([\ell_k - \theta_k]_+^2 + [\theta_k - u_k]_+^2 \right). \tag{7}$$

Angles are in radians when computing this loss. Table 6 lists the bounds in degrees for readability. They are applied componentwise in MANO's local axis-angle representation. Global wrist rotation and translation is excluded.

Table 6: Local axis-angle bounds for the joint-limit regularizer, in degrees. MCP, PIP, and DIP denote the metacarpophalangeal, proximal interphalangeal, and distal interphalangeal joints.

Joint	Lower bound (x, y, z)	Upper bound (x, y, z)
Non-thumb MCP	$(-30, -30, -45)$	$(30, 30, 90)$
Non-thumb PIP	$(0, 0, 0)$	$(0, 0, 90)$
Non-thumb DIP	$(0, 0, -45)$	$(0, 0, 90)$
Thumb base	$(-90, -45, 0)$	$(90, 45, 90)$
Thumb middle	$(-45, -30, 0)$	$(45, 30, 90)$
Thumb distal	$(0, 0, -45)$	$(0, 0, 90)$

C Implementation Details

C.1 Geometry and Physical Parameters

HORT provides the initial MANO shape and pose and an object point cloud. We reconstruct a closed surface using alpha shapes (Edelsbrunner & Mucke, 1994) and decompose it with CoACD (Wei et al., 2022). We use a decomposition threshold of 0.08, preprocessing resolution 30, and sampling resolution 1000. The search uses 20 nodes, 100 iterations, and depth 3, with merging enabled, at most 96 vertices per convex hull, and seed 0. The convex pieces belong to one rigid object; they have no independent motion.

The articulated hand is generated automatically from a fixed MJCF template and the reconstructed MANO hand. Its collision geometry consists of a rigid palm mesh and capsule finger links. Link placement and capsule geometry are fitted to the reconstructed hand, with no manual adjustment for individual examples. The hand geometry is then fixed throughout optimization and evaluation. Hand material density is 1100 kg/m^3.

For DexYCB and HO3D, we obtain the object's mass from the corresponding YCB object and divide it by the reconstructed volume to set a uniform density. For CORe50, the density is 1000 kg/m^3. We compute object mass, center of mass, and inertia from the reconstructed geometry under this uniform-density assumption. Gravity has magnitude 9.8 m/s^2. For DexYCB, the downward direction in the reference camera is transformed to each selected camera using its extrinsics. HO3D and CORe50 use the downward image direction in camera coordinates.

The MuJoCo/MJX contact model includes normal contact, sliding friction, and torsional friction. Sliding and torsional coefficients are 1.0 and 0.01, respectively; rolling friction is zero. Contact margin and gap are zero. Contact reference parameters are $(0.02, 1)$ and impedance parameters are

$(0, 0.99, 0.001, 0.5, 2)$. We use joint armature 10^{-3} and joint friction loss 10^{-4}. Passive damping is 10 for wrist translations, 3 for wrist rotations, and 0.1 for finger joints. These physical parameters and the PD gains are fixed across examples.

C.2 Differentiable Simulation and Numerical Settings

We use MuJoCo-XLA (MJX) with JAX (Todorov et al., 2012; Bradbury et al., 2018) for differentiable simulation. Each objective evaluation initializes the hand at the current visual pose, the object at its reconstructed pose, and all velocities at zero. It then advances the coupled hand–object system for $T_o = 20$ steps of 2 ms each. The kinetic-energy objective sums the total system energy after each step, giving an optimization horizon of 40 ms. The energy includes the articulated hand and rigid object through the generalized mass matrix; it has no additional object-only weighting. Velocities are the simulator's state variables updated by its dynamics.

We differentiate this simulation with forward-mode automatic differentiation, with 64-bit arithmetic enabled in JAX. The visual pose influences the initial physical state, and the target pose influences the PD actuation throughout the simulated trajectory. The regularizer is computed from the visual pose. We optimize their combined objective without any additional force-closure constraint.

Both optimization and evaluation run in MJX and share the numerical settings in Table 7, together with the same rigid geometry, inertial properties, friction coefficients, and controller gains. They differ in only two respects: the timestep and the friction-to-normal impedance ratio. Optimization uses a 2 ms timestep, whereas evaluation uses a finer 0.2 ms timestep to reduce time-integration error. The impedance ratio (MuJoCo's `impratio`) sets how stiffly tangential friction is enforced relative to normal contact: a low ratio lets the soft-constraint solver introduce non-physical slipping, whereas a high ratio enforces friction closer to the ideal Coulomb model and suppresses this numerical slip.[1] We therefore use a high ratio (2000) at evaluation for physical fidelity, and a low ratio (2) during optimization, where the stiffer friction constraints would otherwise cause gradient issues in our physics-based optimization.

Table 7: Numerical settings shared by StableGrasp and Pinching in MJX.

Setting	Value
Integrator	Implicit-fast
Constraint solver	Newton
Maximum solver iterations	100
Maximum line-search iterations	40
Solver tolerance	10^{-6}
Line-search tolerance	10^{-3}
Friction cone	Pyramidal
No-slip iterations	0
Jacobian	Dense

C.3 Optimization and Stopping Criterion

Both phases use Adam (Kingma & Ba, 2014). Visual warm-up updates the visual pose for 600 iterations using only $\mathcal{L}_{\text{reg}}$. Physics-based refinement initializes both the visual pose and target pose from this warm-up result, then updates them jointly. Object geometry, object initial pose, and hand shape remain fixed. Table 8 lists the optimization settings.

At the start of refinement, we evaluate the regularization and kinetic-energy losses once and divide each by its initial value, bounded below by 10^{-12}. These normalization factors remain fixed during the run. We clip the joint gradient of the two pose variables to norm 10^4. After Adam computes an update, the wrist-translation and wrist-rotation updates are scaled by 10^{-3} and 10^{-2}, respectively, for both pose variables; finger updates retain unit scale. The refinement learning rate remains constant.

[1] See the MuJoCo documentation, `https://mujoco.readthedocs.io/en/stable/computation/index.html`.

Table 8: StableGrasp optimization settings. Learning rates refer to the parameterization of each phase.

Setting	Value
Warm-up iterations	600
Warm-up learning rate: global translation	5×10^{-5}
Warm-up learning rate: rotations	5×10^{-3}
$\lambda_{\mathrm{int}}, \lambda_{\mathrm{adh}}, \lambda_{\mathrm{obs}}, \lambda_{\mathrm{lim}}$	$5 \times 10^{10},\ 500,\ 10,\ 1000$
Maximum refinement iterations	2000
Refinement learning rate	10^{-3}
$\lambda_{\mathrm{reg}}, \lambda_{\mathrm{KE}}$	$0.1,\ 1$
Simulation steps per refinement iteration	20
Adam $(\beta_1, \beta_2, \epsilon)$	$(0.9, 0.999, 10^{-8})$
Refinement gradient-norm limit	10^4
Wrist update scales: translation / rotation	10^{-3} / 10^{-2}
Early-stopping patience	200 iterations
Improvement tolerance: relative / absolute	10^{-3} / 10^{-8}

We retain the lowest-loss visual/target pair. A new loss is considered an improvement when its reduction from the best loss exceeds $\max(10^{-8}, 10^{-3}|\mathcal{L}_{\mathrm{best}}|)$. Refinement terminates when no such improvement occurs for 200 iterations, or when the maximum iteration count is reached. This early-stopping rule is used throughout the experiments.

C.4 Pinching Baseline

Pinching constructs a fixed target from the warm-up grasp. Our Pinching baseline adapts the normal-based grasp-closure procedure used in DexGrasp Anything's evaluation code to our articulated hand proxy. For each hand-surface sample $\mathbf{p}_i$, let r_i denote its distance to the object and $\mathbf{n}_i$ its unit surface normal. We construct reference points as

$$\mathbf{p}_i^* = \begin{cases} \mathbf{p}_i + \delta \mathbf{n}_i, & r_i < 5\,\mathrm{mm}, \\ \mathbf{p}_i, & \text{otherwise}, \end{cases} \qquad \delta = 5\,\mathrm{mm}.$$

We fit only the finger coordinates to these reference points by minimizing $\sum_i \|\mathbf{p}_i(\boldsymbol{\theta}) - \mathbf{p}_i^*\|_2$, with one Adam update at learning rate 0.05. The wrist pose stays fixed during this fit. The resulting hand configuration is the fixed PD target, while the visual warm-up grasp supplies the initial physical state. Pinching uses the same object, shape-dependent hand proxy, physical parameters, PD gains, and 5 s evaluation as our method.

C.5 Computation Time

Experiments run on a workstation with seven NVIDIA Quadro RTX 6000 GPUs and two Intel Xeon Gold 6230 CPUs. Each StableGrasp example uses one GPU and six CPU threads. Table 9 reports per-example mean and standard deviation. XLA/JIT compilation is included in the reported runtimes. The main cost is forward-mode differentiation through the 20-step MJX simulation.

Table 9: Per-example computation cost of StableGrasp, reported as mean $\pm$ standard deviation. Compilation is a component of the total time.

Quantity	Mean $\pm$ standard deviation
Refinement iterations with early stopping	1213.6 ± 609.4
Visual warm-up and physics-based refinement	26.05 ± 13.17 min
Full pipeline	26.77 ± 13.20 min
XLA/JIT compilation	3.29 ± 0.94 min

D INSTANCE-SPECIFIC PPO BASELINE

This experiment aims to study whether policy learning can stabilize an individual reconstructed grasp. We build on the grasping policy of D-Grasp (Christen et al., 2022) and train one PPO policy for each input example. The policy observes the simulated hand–object state and continually adjusts the hand's PD targets. The task is to preserve the reconstructed configuration of both the hand and object under gravity.

D.1 TASK AND REFERENCE GRASP

D-Grasp learns a shared policy for establishing reference grasps and combines it with a separate motion-synthesis module for object transport. Its static grasping evaluation also tests the ability to hold objects after grasp acquisition. Our instance-specific experiment starts directly from the visual warm-up result, $\mathbf{q}_h^{\text{warmup}}$, with zero hand and object velocities. The reconstructed object remains the positional reference throughout the episode. We omit the approach and transport stages, remove the support table, and disable wrist-approach guidance and additional object-weight compensation at the wrist.

We infer the desired contact pattern from the reconstructed geometry. For each of the 21 hand joints, we compute its distance to the nearest object mesh vertex. Joints within 15 mm are mapped to the 16 hand contact regions used by D-Grasp; the distal thumb region is always included. The reference joint positions are recomputed by forward kinematics of the simulation hand at the initial pose. This makes the positional reference consistent with the hand used for policy training.

D.2 PHYSICS AND CONTROL

Both PPO training and policy evaluation run in RaiSim, using an articulated hand with mean MANO shape, a mesh palm, and capsule finger links. We import the reconstructed object mesh and compute its mass, center of mass, and inertia from the mesh and assigned density. The object input, density, gravity, and PD gains are shared with our method. RaiSim uses its mesh collision representation and a friction coefficient of 1.0. StableGrasp uses a shape-dependent hand proxy and convex object decomposition in MuJoCo/MJX; each engine retains its own contact solver.

The policy observes joint positions and velocities, hand–object relative pose and motion, current and desired contacts, and deviations from the reference grasp. Its 51-dimensional output specifies offsets for wrist translation (3), wrist rotation (3), and finger rotations (45). During training, these offsets are added to the initial hand configuration:

$$\mathbf{q}_t^{\text{PD}} = \text{clip}_{[\mathbf{q}_{\min}, \mathbf{q}_{\max}]}\big(\mathbf{q}_{\text{init}} + \mathbf{S}\mathbf{a}_t\big).$$

The diagonal action scales in $\mathbf{S}$ are 0.01 m for translation, 0.01 rad for wrist rotation, and 0.015 rad for finger rotations. They scale the policy output; joint limits are enforced separately by clipping the resulting targets. All target velocities are zero. The proportional/damping gains are $(5\times10^4, 5\times10^3)$ for wrist translation, $(5 \times 10^2, 5 \times 10^1)$ for wrist rotation, and $(10, 1)$ for finger joints.

At evaluation, we use the policy's deterministic mean action at each control step. Following D-Grasp's static holding protocol, the wrist PD target is held fixed, while the policy continues to update finger targets. We enable this holding mode from the initial state. The wrist itself remains part of the simulated articulated system.

D.3 REWARD

We retain D-Grasp's pose, contact, and motion rewards, activate its object position term, and add a penalty on changes in the control target. Table 10 lists all nonzero terms and their effective coefficients. Their weighted sum is clipped from below at -2 at each control step.

The positional error is measured in the object frame. Its coordinate weights are four times larger at fingertips than at other joints and normalized to have mean one before weighting and squaring the residuals. The negative joint-position, joint-rotation, and object-position terms are individually bounded below by -10. The impulse sum is capped at $5m_o$, where m_o is the object mass, following the original implementation. The target-variation term uses consecutive joint-limit-clipped targets and excludes wrist translation; it is zero at the first control step.

Table 10: Reward terms for the instance-specific PPO policy.

Term	Quantity before weighting	Weight
Joint positions	Negative weighted squared joint-position error	2.0
Joint rotations	Negative norm of angular pose error	0.1
Contact coverage	Fraction of desired contact regions in contact	1.0
Contact impulse	Capped sum of impulses at desired contact regions	2.0
Object motion	Squared object linear speed	−1.0
Wrist translation	Squared wrist linear speed	−0.5
Wrist rotation	Squared wrist angular speed	−0.5
Object position	Negative distance from the reference object position	1.0
Target variation	Squared change in wrist-rotation and finger targets	−5.0

D.4 Training

Each example is trained independently from scratch, with ten parallel copies of the same scene. Every copy resets to the same reconstructed state without position or joint-angle noise. Exploration comes from a diagonal Gaussian action distribution with learned standard deviations, initialized to one before action scaling. Observations are passed to the networks without running-mean normalization. Actor and critic are separate networks with two hidden layers of 128 units and LeakyReLU activations.

Table 11: PPO training settings used for each example.

Setting	Value
Collection/update iterations	100
Parallel scene copies	10
Control decisions per copy per iteration	135
Total collected transitions	135,000
PPO epochs / minibatches per epoch	4 / 4
Discount factor / GAE parameter	0.996 / 0.95
Policy clipping parameter	0.2
Value-loss coefficient / entropy coefficient	0.5 / 0
Optimizer / initial learning rate	Adam / 5×10^{-4}
Learning-rate schedule / target KL	Adaptive / 0.01
Maximum gradient norm	0.5
Physics timestep / nominal control interval	2.22 ms / 30 ms
CPU threads per training process	4

The PPO update uses clipped value loss and four minibatches in their collected order. The learning rate decreases by a factor of 1.2 when the measured KL divergence exceeds 0.02, and increases by 1.2 when it lies between zero and 0.005, within bounds $[10^{-5}, 10^{-2}]$. Training runs for the full iteration budget without reward-based early stopping. After each iteration, we save the updated policy when that iteration's collected training return exceeds the previous best by more than 10^{-4}. This checkpoint is used for one deterministic evaluation trajectory. There is one training run per example.

D.5 Evaluation and Included Examples

PPO generates trajectories through closed-loop control in RaiSim. At each control step, it reads the current simulated state, computes new finger targets, and advances the simulation under PD control. We then convert the saved poses to the common hand representation and camera coordinates for kinematic measurements, image projection, and visualization. This conversion uses forward kinematics in MuJoCo; the trajectory dynamics are generated by RaiSim. In Table 1, reconstructions use the MANO surface and final states use the common articulated display proxy, under the same camera and rendering settings for all methods.

Each saved PPO trajectory contains the initial state and 166 control updates. Its stored timestamps are spaced by 30 ms and end at 4.98 s. We evaluate the available trajectories and use their stored

time axis for Drop Time. Pinching and StableGrasp are evaluated for 5 s in MuJoCo. The qualitative comparison shows the initial and final available states of each trajectory.

For all three methods, a grasp fails if the object center of mass moves more than 1 cm from its initial position at any saved state. Motion, IoU, and PCK compare initial and final states and are averaged over each method's successful grasps. Drop Time is the first threshold crossing, averaged over trajectories that cross the threshold. Numerical terminations count as failures in Success Rate and are omitted from the Drop Time average. Standard deviations are computed across examples.

We attempted PPO training for all 362 benchmark examples. Five runs terminated numerically after the released object developed divergent rotational motion, producing non-finite observations or value-loss calculations. Table 2 compares all three methods on the same remaining 357 examples: 239 DexYCB, 99 HO3D, and 19 CORe50. Every included PPO policy completed the 100-iteration budget. These exclusions apply to the baseline comparison; the reconstruction-quality evaluation retains its original examples.

Some aspects of our comparison cannot be placed on identical footing: PPO is trained and rolled out in RaiSim, whereas Pinching and StableGrasp are evaluated in MuJoCo. However, we resolved key choices in favor of the baseline rather than StableGrasp. Most importantly, the PPO policy observes the simulated state and updates its finger PD targets at every control step, exercising closed-loop feedback throughout the episode, whereas StableGrasp commits to a single fixed target with no such adaptation. Its scored horizon is also slightly shorter than the full 5 s used for Pinching and StableGrasp, leaving PPO marginally less time in which the grasp could fail.

E Evaluation and Ablation Protocols

E.1 Examples and Aggregation

The benchmark contains 362 input images: 242 from 81 DexYCB sequences, 100 from ten HO3D sequences, and 20 manually selected CORe50 images. The DexYCB views contribute 81, 80, and 81 images from cameras 1, 3, and 4, respectively. DexYCB and HO3D examples come from the HORT test split. We select frames near the ends of the sequences, where the object is generally held stably in the hand; the selected HO3D frames can still contain ongoing motion. All selected examples are retained for evaluation of our method.

The baseline comparison in Table 2 uses the 357 common examples described in Section D.5. Reconstruction accuracy uses the original 242 DexYCB and 100 HO3D examples with 3D hand annotations. CORe50 contributes to stability evaluation, but has no corresponding 3D hand annotation in this benchmark.

E.2 Physical and Image-space Stability

Let $\mathbf{c}_t$ and $\mathbf{R}_t$ be the object's center of mass and orientation, and let $\mathbf{J}_{j,t}$ be hand joint j, all expressed in the fixed evaluation coordinates. Subscripts 0 and f denote the initial and final saved states. We report hand displacement

$$D_h = \frac{1}{21} \sum_{j=1}^{21} \|\mathbf{J}_{j,f} - \mathbf{J}_{j,0}\|_2, \tag{8}$$

object translation $\|\mathbf{c}_f - \mathbf{c}_0\|_2$, and the rotation angle of $\mathbf{R}_f \mathbf{R}_0^\top$. These quantities measure change from the reconstructed initial state, without fitting an alignment between the two endpoints. Translation and hand displacement are reported in millimeters, and rotation in degrees.

For image-space stability, object IoU is the intersection-over-union of initial and final projected object masks. Hand PCK uses the 21 projected joints. For crop width W and height H, we normalize a pixel coordinate (u, v) to $(u/W, v/H)$ and count a joint as correct when its initial-to-final Euclidean displacement in these normalized coordinates is at most τ. We report $\tau = 0.025$ and 0.05. This crop normalization applies to the reported PCK values throughout the paper.

An example succeeds when the object center of mass stays within 1 cm of its initial position and the simulation remains finite. Drop Time records the first saved state exceeding this threshold. Success Rate is computed over all included examples. Displacement, IoU, and PCK are averaged over

each method's successful examples; Drop Time is averaged over threshold-crossing failures. Consequently, each method's motion statistics describe its own surviving grasps. All reported standard deviations use population normalization across examples.

E.3 Reconstruction Accuracy

Reconstruction accuracy compares the visual pose with dataset annotations, whereas the stability metrics above compare a grasp with its own initial state. For 3D joint error, the evaluation fits a uniform scale and translation from the annotation joints to the reconstructed joints, without rotational alignment, and reports their mean Euclidean distance. The alignment minimizes

$$\min_{s,\mathbf{b}} \sum_{j=1}^{21} \|s\mathbf{J}_j^{\mathrm{GT}} + \mathbf{b} - \mathbf{J}_j^{v}\|_2^2. \tag{9}$$

For 2D PCK, projected reconstruction joints are compared with the annotated image joints using the crop normalization above.

The annotation fits and the simulated grasp use different hand representations and objectives. Image and depth observations constrain the MANO annotation (Chao et al., 2021; Hampali et al., 2020), while our refinement also adjusts contact in an articulated rigid-body proxy under actuation and gravity. Contact locations, surface shape, and compliance affect the forces supported by a grasp (Grady et al., 2021). Thus, matching an annotated pose and maintaining the simulated grasp measure different properties of the reconstruction. We report annotation agreement together with hand and object motion to show both effects of refinement.

E.4 Equal-budget Visual Warm-up Ablation

We use 86 examples: 61 DexYCB and 25 HO3D. The subset is selected systematically from sorted example identifiers, taking every fourth example with offsets 0, 1, and 2 in DexYCB cameras 1, 3, and 4, respectively, and offset 0 in HO3D. This gives 21, 20, 20, and 25 examples from the four groups.

The full method uses 600 warm-up iterations followed by at most 2000 physics-based refinement iterations. The ablation initializes both the visual and target poses directly from the visual reconstruction and runs at most 2600 physics-based iterations. The refinement hyperparameters, 20-step simulation horizon, stopping rule, and evaluation settings are otherwise unchanged. Both variants therefore have a maximum of 2600 optimization iterations. Early stopping determines the executed count.

With warm-up, 64 of 86 refinement runs stop early; the mean number of refinement iterations is 1207.99 (range 207–2000), in addition to 600 warm-up iterations. Without warm-up, 56 of 86 runs stop early, with a mean of 1587.66 physics-based iterations (range 201–2600). The full method succeeds on 78 of 86 examples (90.70%), compared with 66 of 86 (76.74%) without warm-up. Hand displacement is also lower with warm-up on both datasets. The complete per-dataset measurements are reported in Table 4. These results show the benefit of forming contact before optimizing the dynamical objective.

E.5 Reading the Qualitative Results

The qualitative comparisons show the input, reconstruction, and final available simulated state. The target column compares the initial visual hand with its PD target under a common camera. StableGrasp and Pinching use fixed targets. For PPO, this column shows the actual target at the initial control step; subsequent targets are computed by the policy from the evolving state.

The supplementary video shows the temporal behavior between the displayed endpoints. Small finger shifts, wrist oscillations, and object slip can be visible in motion even when two endpoint renderings look similar. Both hand and object motion should be considered when inspecting these sequences.

F FAILURE CASES AND LIMITATIONS

Our analysis of the original 362-example benchmark finds 48 failed grasps, 27 of which involve inaccurate object reconstruction. Thin objects such as scissors, clamps, and markers are sensitive to small meshing errors: a narrow part can become too thick, disappear, or provide contact at an incorrect location. Concave or hollow structures, including bowls and cup handles, can be filled during surface reconstruction. These changes alter the collision geometry and the forces the simulated grasp can support.

Physics-based refinement keeps the object mesh and hand shape fixed and locally optimizes the visual pose and control target. It can adjust nearby contacts, but cannot recover missing object structure or correct severe initial reconstruction errors. The hand proxy also represents fingers and palm as rigid bodies. Real contact includes soft-tissue deformation and distributed pressure, which are approximated by rigid geometry and the simulator's compliant contact model. Object density, gravity, friction, and controller parameters specify the physical setting in which a grasp is optimized and evaluated.