SECOND-GRASP: SEMANTIC CONTACT-GUIDED DEXTEROUS GRASPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Achieving reliable robotic manipulation, such as dexterous grasping, requires a synergy between physically stable interactions and semantic task guidance, yet these objectives are often treated as separate, disjoint goals. In this paper, we investigate how to integrate physically stable dexterous grasping with language-guided grasp generation to enable robotic grasping that is both physically reliable and semantically aligned with task intent. To this end, we propose **SECOND-Grasp** (SEmantic CONtact-guided Dexterous Grasping), a unified framework that enables robotic hands to dynamically adjust grasping strategies based on semantic reasoning while ensuring physical feasibility. We begin by obtaining coarse contact proposals through vision-language reasoning to infer where contacts should occur based on object properties, followed by segmentation to localize these regions across views. To further ensure consistency across multiple viewpoints, we introduce Semantic-Geometric Consistency Refinement (SGCR), which refines initial contact predictions by enforcing semantic consistency across views and removing geometrically invalid regions, yielding reliable 3D contact maps. Lastly, we derive a feasible hand pose for each contact map via inverse kinematics, generating a supervision signal for policy learning. Our approach, trained on DexGraspNet, consistently outperforms baselines in lifting success rate on both seen and unseen categories, achieving 98.2% and 97.7%, respectively, while also improving intent-aware grasping by 12.8% and 26.2%. We further show that SECOND-Grasp realizes semantic intent through distinct, physically executable grasping strategies across diverse objects and robotic hand embodiments.

1 INTRODUCTION

Dexterous grasping has been widely studied as a fundamental capability for robotic manipulation, with the primary goal of achieving physically stable object interactions (Wang et al., 2022; Xu et al., 2023; Wan et al., 2023; Xu et al., 2024; Zhong et al., 2025). Recent advances have focused on generating multi-fingered hand configurations (Xu et al., 2023; 2024; Ma et al., 2025) that satisfy force closure, collision avoidance, and robust object lifting. However, these approaches largely operate in a pose-centric manner, with limited consideration of high-level task intent or semantic guidance. In contrast, vision-language reasoning-based manipulation for multi-fingered hands (He et al., 2025; Zhong et al., 2026; Lee et al., 2026) has emerged as a promising direction for leveraging semantic cues in action generation, but it often focuses on producing plausible hand poses without explicitly ensuring physical feasibility (Lee et al., 2026; Li et al., 2024b).

This disconnection raises two fundamental challenges: without semantic grounding, grasp strategies cannot adapt to task-specific requirements (Li et al., 2024c; Wei et al., 2024; Lee et al., 2026), and without explicit contact modeling, they fail to ensure physically stable execution in practice (Liu et al., 2023; 2021; Wang et al., 2022). Existing methods address these as separate problems, yet bridging the two remains largely unexplored. We argue that truly capable dexterous manipulation requires bridging this gap through explicit contact-level reasoning that jointly grounds semantic intent and physical feasibility.

To address this, we propose **SECOND-Grasp** (SEmantic CONtact-guided Dexterous Grasping), a unified framework that explicitly models contact-level interactions to couple semantic understanding with physical feasibility in grasp planning. These contact-level semantics correspond to object parts

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.

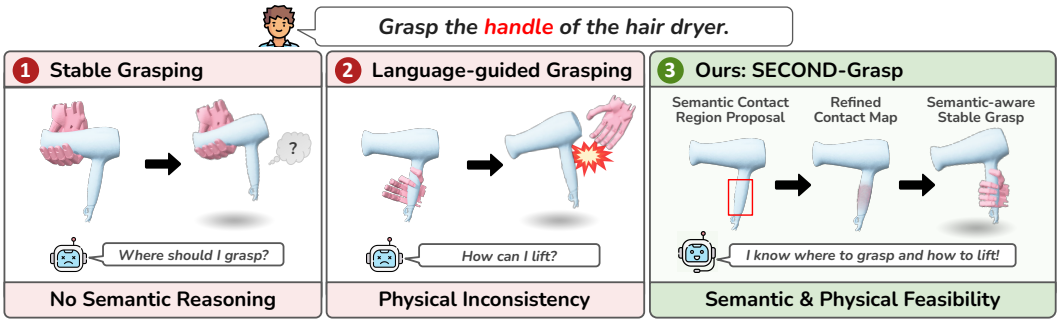

Figure 1: Contact-based grasping ensures physical feasibility but lacks semantic reasoning, while language-based grasping captures intent but often fails to lift. SECOND-Grasp bridges both, enabling grasps that are semantically meaningful and physically stable.

(e.g., handle, body, base) that can be spatially grounded to guide intent-aware grasp planning. Rather than directly predicting grasp regions from observations, **SECOND-Grasp** operates in a procedural manner. First, it infers meaningful contact regions for each target object by leveraging internet-scale semantic priors of vision-language models. Then, it grounds these proposals in multi-view observation space using a segmentation model, providing semantical and geometrical guidance for downstream policy learning.

However, despite the strong priors provided by vision-language and segmentation models, their predictions may still suffer from inconsistencies between semantic correspondence and geometric alignment across views. We therefore introduce Semantic-Geometric Consistency Refinement (SGCR), which enforces cross-view consistency through two sequential stages: (1) a Semantic Refinement Stage that assigns higher confidence to regions consistently predicted across multiple views; and (2) a Geometric Refinement Stage that enforces geometric consistency of the contact map using a local convexity-based criterion, which resolves surface ambiguity by verifying whether predictions lie on the same surface region. Together, these stages produce contact maps that are both semantically grounded and geometrically coherent, allowing the model to capture diverse interaction intents across object parts and surface geometries.

To use these contact intentions for policy learning, they must be translated into hand-level guidance that supports physical execution. A common approach is to recover grasp poses from contact maps using learned grasp generators (Weng et al., 2024; Lu et al., 2024; Wu et al., 2025). However, because these models are trained on fixed hand-object datasets, they inherit hand- and object-specific priors that may limit generalization to unseen embodiments or objects. We instead adopt contact-based optimization methods (Liu et al., 2023; Zhao et al., 2024; Zhang et al., 2026), which directly convert contact representations into feasible hand configurations. Unlike prior pipelines that treat the optimized configuration as the final grasp, we further use analytic inverse kinematics to derive pseudo pose priors from the refined contact maps for policy learning. Together, the refined contact maps and pseudo pose priors guide the policy toward semantically aligned and physically stable grasping across diverse objects and robotic hands.

Experimental results demonstrate that our approach, trained on DexGraspNet dataset (Zhang et al., 2024a), outperforms existing grasping methods (Xu et al., 2023; Wan et al., 2023; Wang et al., 2025; Yuan et al., 2025; Mao et al., 2025) in object lifting success rate, achieving 98.2% and 97.7% on seen and unseen categories, respectively. Beyond general grasping performance, we evaluate intention alignment success rate, observing improvements of 12.8% and 26.2% for seen and unseen categories. We further validate the transferability of our framework to different hand embodiments using the Shadow Hand (Shadow Robot Company, 2024) and Allegro Hand (Wonik Robotics, 2024), and evaluate its generalizability across diverse object datasets.

Our contributions are summarized as follows:

- **SECOND-Grasp**, a unified grasping framework that bridges semantic intent and physical execution by grounding language-guided interaction goals into semantically meaningful contact regions that guide physically stable grasping.

- **Semantic-Geometric Consistency Refinement (SGCR)**, a two-stage module that improves cross-view consistency through semantic confidence refinement and local convexity-based geometric validation, yielding semantically and geometrically coherent contact maps.
- A policy learning pipeline that derives pseudo pose priors from refined contact maps, enabling semantically aligned and physically stable grasping across diverse objects and robotic hands.

2 Related Work

2.1 Dexterous Grasping

Dexterous grasping aims to achieve stable object grasping and manipulation with multi-fingered robotic hands. To enable scalable learning of dexterous grasping, prior methods have constructed large-scale datasets that pair objects with physically dexterous hand poses (Wang et al., 2022; Li et al., 2023; Liu et al., 2024; Zhong et al., 2025; Taheri et al., 2020). Building on large-scale grasp datasets, policy-learning methods train dexterous hands for object lifting across diverse objects and hand embodiments (Xu et al., 2023; Wan et al., 2023; Huang et al., 2025b; Yuan et al., 2025; Mao et al., 2025). However, these methods primarily focus on physical feasibility or lifting success, without explicitly considering whether the resulting contacts align with semantically meaningful object regions. Recent language-guided dexterous grasping methods incorporate task intent or object usage into grasp generation (Wei et al., 2024; 2025; Lee et al., 2026; Jian et al., 2025), but remain focused on grasp pose synthesis and do not address physically executable lifting. Our work bridges these two directions by grounding dexterous policy learning in contact regions inferred through 2D VLM reasoning, enabling grasps that are semantically meaningful and physically executable.

2.2 2D-to-3D Affordance Grounding

Affordance grounding aims to identify object regions that enable specific interactions (Gibson, 1979; Do et al., 2018; Qian et al., 2024; Li et al., 2024a; Zhu et al., 2025; Moon et al., 2025). Understanding affordances in 3D is essential for embodied interaction and robotic manipulation, as it enables agents to reason about how objects can be used in the physical world (Deng et al., 2021; Qian et al., 2024). However, object affordance is not always determined by geometry alone, as functional regions often depend on how an object is used. This has motivated approaches to incorporate 2D interaction or semantic cues into 3D affordance grounding. IAGNet (Yang et al., 2023) and MIFAG (Gao et al., 2025) use 2D human-object interaction images to guide 3D affordance prediction, while CAST (Huang et al., 2025a) transfers semantic features from 2D vision foundation models into 3D representations. Together, these methods demonstrate that 2D semantic knowledge provides effective priors for localizing functional regions in 3D. Our work builds on this insight, applying 2D VLM reasoning to infer object-specific contact regions and refining them into 3D contact maps for dexterous grasping.

2.3 Grasp Pose Generation

Dexterous grasp generation methods commonly synthesize hand configurations from object observations using learned generative models or analytical solvers. Generative methods learn grasp distributions conditioned on the features of the object via diffusion-based modeling (Weng et al., 2024; Lu et al., 2024; Wu et al., 2025). Although effective in generating diverse and plausible grasps, these models often learn hand-specific priors tied to training embodiments and object distributions, which can limit transferability to novel hands or objects. Analytical methods instead recover hand configurations from geometric or kinematic constraints without learning a grasp distribution (Turpin et al., 2022; 2023; Liu et al., 2021; Zurbrügg et al., 2025; Murray et al., 2017; Siciliano et al., 2009; Buss, 2004). Within analytical approaches, contact-based optimization methods use predicted contacts to recover hand configurations, demonstrating contact as an effective intermediate representation for grasp synthesis (Liu et al., 2023; Zhao et al., 2024; Zhang et al., 2026). Our work builds on these perspectives, but rather than treating the recovered configuration as the final grasp, we use it as a pseudo-pose prior to guide policy learning toward executable grasping behavior.

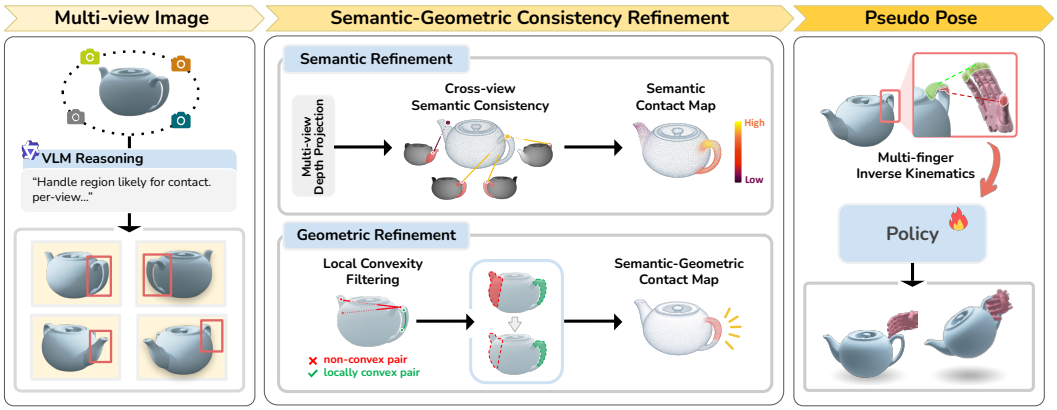

Figure 2: **Overview.** Given multi-view observations, **SECOND-Grasp** first infers semantic contact regions using vision-language reasoning (Section 4.1). These 2D proposals are then projected into 3D and refined through Semantic-Geometric Consistency Refinement, which enforces cross-view semantic consistency and geometric consistency based on local convexity (Section 4.2). The refined semantic-geometric contact map provides inverse-kinematics-based pseudo pose guidance (Section 4.3) for policy learning (Section 4.4). See individual sections for technical details.

3 Preliminaries

We formulate dexterous grasping as a Partially Observable Markov Decision Process (POMDP) with timesteps $t = 0, \ldots, T$. For each timestep, the environment evolves according to state:

$$s_t = (x_t^{hand}, x_t^{obj}, x^{geo}), \tag{1}$$

where x_t^{hand} denotes robot proprioceptive states (e.g., joint configurations and fingertip states), x_t^{obj} denotes object pose and velocities, and x^{geo} encodes object geometric information. Due to partial observability, the agent instead receives an observation:

$$o_t = (x_t^{hand}, v_t^{img}, v_t^{point}), \tag{2}$$

generated according to $P(o_t|s_t)$. Here, v_t^{img} denotes object-centric images from a third-person view and v_t^{point} denotes partial point cloud observations. The policy input additionally includes a pseudo-pose condition c that guides grasping toward the selected contact region (Appendix C.1). Using full state information in simulation, we first train a privileged state-based policy $\pi_{\text{state}}(a_t \mid s_t, c)$. We then distill it into a vision-based policy $\pi_{\text{vision}}(a_t \mid o_t, c)$ that could operate under partial observability in real-world settings. For both policies, the action is parameterized by:

$$a_t = (\Delta w_t, \Delta\phi_t, \Delta\theta_t), \tag{3}$$

where $\Delta w_t \in \mathbb{R}^3$ and $\Delta\phi_t \in \mathbb{R}^3$ denote the translational and rotational increases in the wrist, respectively, and $\Delta\theta_t \in \mathbb{R}^{d_\theta}$ denotes the incremental joint configurations for the actuated hand joints. Here, d_θ depends on the robotic hand embodiment (see Appendix D.2). Both policies are trained to maximize the expected discounted return $\mathbb{E}[\sum_{t=0}^{T-1} \gamma^t r_t]$, where $\gamma \in (0, 1)$ is the discount factor.

4 Method

Our SECOND-Grasp enables robotic hands to grasp objects semantically and physically in a plausible manner. We first generate a 3D contact map of object-specific graspable regions and derive a hand pose via inverse kinematics. We then use the contact map and IK-derived pose prior to guide policy learning toward semantically aligned and stable grasping. The overview is illustrated in Figure 2.

4.1 Semantic Contact Region Proposal

Identifying semantic contact regions solely from 3D geometry is ambiguous, as graspable parts are defined not just by shape but by functional intent—how an object is meant to be used. This motivates us to leverage 2D vision-language reasoning as a semantic prior to guide contact region proposal.

Given multi-view RGB images of an object $\{\mathcal{I}_i\}_{i=1}^{I}$ from $I = 4$ fixed viewpoints (front, back, left, right), we use a pretrained large vision-language model (VLM) (Bai et al., 2025) to extract an initial set of K graspable regions $\{g_k\}_{k=1}^{K}$ (e.g., handle, body, base) and their 2D bounding boxes $\{b_k\}_{k=1}^{K}$. Additionally, each bounding box is associated with a confidence score $c_i^{(k)} \in [0, 1]$, which indicates the VLM's estimated likelihood that the bounding box contains the k-th graspable region. To achieve a precise pixel-level localization, we apply a segmentation model (Ravi et al., 2024) $\mathcal{F}_{\text{seg}}$ within each bounding box and obtain a binary mask $M_i^{(k)} = \mathcal{F}_{\text{seg}}(I_i, b_i^{(k)})$. The mask is further filtered by valid object regions from the rendered depth map to yield $\tilde{M}_i^{(k)}$.

4.2 Semantic-Geometric Consistency Refinement

While the 2D VLM provides high-level semantic contact proposals, accurate robotic grasping requires finer 3D refinement. We therefore introduce Semantic-Geometric Consistency Refinement (SGCR) to produce precise contact regions that are consistent with both semantic and geometric cues.

4.2.1 Contact Confidence Initialization

The confidence score $c_n^{(k)}$ assigned by the VLM reflects the model's semantic confidence in the presence of the k-th graspable region, but do not account for precise localization quality. To address this, we re-weight the raw confidence $c_i^{(k)}$ by the valid object region ratio $\rho_i^{(k)}$, which is obtained by the area of $\tilde{M}_i^{(k)}$ over the predicted bounding box:

$$\bar{c}_i^{(k)} = \sigma(\rho_i^{(k)})\, c_i^{(k)}, \tag{4}$$

where $\sigma(\cdot)$ is a sigmoid function for down-weighting predictions with low overlap. The initial per-view confidence map is then constructed as $S_i^{(k,0)}(p) = \bar{c}_i^{(k)} \cdot \tilde{M}_i^{(k)}(p)$, where p denotes a pixel location, assigning the calibrated score to each valid mask pixel.

4.2.2 Cross-View Semantic Refinement

Regions that are consistently predicted across multiple views better correspond to semantically expressive parts than those predicted from a single view. We therefore assign higher confidence for such cross-view consistent predictions.

Let $\mathcal{N}(i)$ denote the set of neighboring views of source view i. For each pixel p in $\tilde{M}_i^{(k)}$, we back-project p into 3D space and re-project it onto each neighboring view $j \in \mathcal{N}(i)$. We verify whether the re-projected point is geometrically consistent and satisfies a depth consistency, which is measured by the discrepancy between the actual and the re-projected depth:

$$\delta_{i\to j}(p) = \left| D_{i\to j}(p) - z_{i\to j}(p) \right|, \tag{5}$$

where $D_{i\to j}(p)$ is the actual depth of the reference pixel point p in view j, and $z_{i\to j}(p)$ is the re-projected depth of pixel point p in view j. We consider cross-view predictions to be consistent if $\delta_{i\to j}(p)$ is less than a predefined threshold τ and the re-projected point lies within $\tilde{M}_j^{(k)}$. We then accumulate this cross-view support to refine the source-view confidence:

$$S_i^{(k,\text{ref})}(p) = S_i^{(k,0)}(p) + \sum_{j\in\mathcal{N}(i)} \alpha\, \bar{c}_j^{(k)}\, \mathbf{1}\big(\delta_{i\to j}(p) < \tau \ \wedge\ p_{i\to j} \in \tilde{M}_j^{(k)}\big), \tag{6}$$

where α is a scaling factor to moderate the contribution of cross-view agreement, preventing it from overwhelming initial confidence. The refined maps are then normalized globally across all views by their maximum value to yield $S_i^{(k,\text{norm})}$, which serves as semantic guidance for the subsequent 3D geometric refinement step. We denote the normalized confidence maps as $\mathcal{S}^{(k,\text{norm})} = \{S_i^{(k,\text{norm})}\}_{i=1}^{I}$.

4.2.3 3D Geometric Refinement via Local Convexity

Refined maps $S^{(k,\text{norm})}$ capture semantic consistency in all views, but do not enforce geometric coherence at the surface level of the object. As a result, predicted regions may be overly dispersed

or fragmented into multiple disconnected components, resulting in ambiguous contact targets for the robotic hand. To address this, we propose a novel local convex-pair refinement module that explicitly enforces surface-level geometric coherence, consolidating fragmented predictions into spatially consistent and physically meaningful contact regions.

We first lift the multi-view maps $S^{(k,\mathrm{norm})}$ into 3D by back-projecting each view into object space using the corresponding depth maps, resulting in a set of 3D points $\mathcal{Q}^{(k)} \subset \mathcal{O}$, where $\mathcal{O}$ denotes the object surface. Here, each point is associated with a confidence score inherited from $S^{(k,\mathrm{norm})}$. Then, we define a high-confidence seed region $\mathcal{Q}^{(k)}_{\mathrm{seed}}$ as the top 10% of points in $\mathcal{Q}^{(k)}$ by confidence scores.

Starting from the seed region $\mathcal{Q}^{(k)}_{\mathrm{seed}}$, we identify points on the same object surface via local-convexity consistency. A pair of surface points $(q_1, q_2) \in \mathcal{O} \times \mathcal{O}$ is defined as a locally convex pair if the line segment connecting them lies entirely within the volume of the object:

$$\mathcal{C}(\mathcal{O}) = \{(q_1, q_2) \in \mathcal{O} \times \mathcal{O} \mid (1-\lambda)q_1 + \lambda q_2 \in \mathrm{Vol}(\mathcal{O}),\ \forall \lambda \in [0,1]\}\,. \tag{7}$$

where $\mathrm{Vol}(\mathcal{O})$ denotes the volume enclosed by the object surface $\mathcal{O}$. Then, we retain a point $q \in \mathcal{Q}^{(k)}$ in the contact map $\mathcal{Q}^{(k)}_{\mathrm{final}}$ if it satisfies local-convexity consistency with any seed point $q_s \in \mathcal{Q}^{(k)}_{\mathrm{seed}}$:

$$\mathcal{Q}^{(k)}_{\mathrm{final}} = \mathcal{Q}^{(k)}_{\mathrm{seed}} \cup \left\{ q \in \mathcal{Q}^{(k)} \;\middle|\; \exists\, q_s \in \mathcal{Q}^{(k)}_{\mathrm{seed}}, (q, q_s) \in \mathcal{C}(\mathcal{O}) \right\}. \tag{8}$$

Because candidate points are validated directly against the seeds, rather than through iterative propagation or global expansion, this formulation defines local convexity in a seed-centered, pairwise manner. A candidate point is retained only if it forms a locally convex pair with at least one seed, which naturally restricts the accepted region around each seed to a compact local neighborhood, such as a cylinder-like subset of a possibly curved cup handle. By taking the union of these seed-centered neighborhoods, the semantically grounded seed region is expanded into a geometrically coherent contact map without requiring the entire semantic part to be globally convex.

4.3 Hand Pose Prior via Inverse Kinematics

To bridge the refined contact map with policy learning, we derive a contact-aligned pseudo hand pose through inverse kinematics (IK). Given the refined contact map $\mathcal{Q}^{(k)}_{\mathrm{final}}$, we estimate its principal axis using PCA (Wold et al., 1987) and partition the contact points into a thumb region $\mathcal{Q}^{(k)}_{\mathrm{thumb}}$ and a non-thumb region $\mathcal{Q}^{(k)}_{\mathrm{non\text{-}thumb}}$, encouraging an oppositional grasp structure. Each robot finger $f \in F$ is then assigned a target contact region $\mathcal{R}^{(k)}_f$, where F denotes the set of fingers of the target hand and may vary across embodiments.

Let $h = (w, \phi, \theta)$ denote the hand configuration, where $w \in \mathbb{R}^3$ and $\phi \in \mathbb{R}^3$ are the wrist position and orientation, and $\theta \in \mathbb{R}^{d_\theta}$ is the joint configuration. For each finger f, let $\xi_f(h)$ denote the fingertip position induced by h. We obtain the pseudo hand pose by minimizing the fingertip-to-region distance:

$$\min_h \sum_{f \in F} \min_{y \in \mathcal{R}^{(k)}_f} \|\xi_f(h) - y\|_2^2\,. \tag{9}$$

We solve this objective with 12 Jacobian-based damped least-squares IK updates (Buss, 2004) to obtain an efficient hand-level prior. Solver details are provided in Appendix C.1.

4.4 Policy Learning

The hand configuration $h^\star = (w^\star, \phi^\star, \theta^\star)$ extracted from the inverse kinematics (IK) solution serves as a pose prior for guiding policy learning. Together with the refined contact map, this prior shapes the reward to encourage semantically aligned and physically stable grasping behaviors. The contact map and pseudo hand pose are used as reward guidance, allowing semantic contact intent to be distilled into the learned policy. We train π_{state} using reinforcement learning with:

$$r = r_{\mathrm{contact}} + r_{\mathrm{pose}} + r_{\mathrm{task}}, \tag{10}$$

where r_{contact} encourages fingertips to approach the refined contact region, r_{pose} penalizes deviation from the IK-derived pseudo hand pose $h^{\star}$, and r_{task} follows standard lifting and goal-distance objectives (Xu et al., 2023). For standard grasping, no test-time pose prior is required, while intent-aware grasping provides the corresponding pseudo hand pose as an observation condition without retraining the policy.

5 EXPERIMENTS

Our experiments aim to evaluate four aspects of the proposed method: (1) grasp success and generalization to unseen object instances and categories, (2) intent-aware grasping, including contact alignment and grasp-style diversity, (3) the effectiveness of the contact-map refinement procedure and other key components, and (4) robustness under cross-dataset object distribution shifts.

5.1 EXPERIMENTAL SETTINGS

Train & Evaluation Settings We train and evaluate all simulation experiments in Isaac Gym (Makoviychuk et al., 2021). For the main benchmark, we use DexGraspNet (Wang et al., 2022) object assets, training on 3,200 objects and evaluating on 141 unseen objects from seen categories and 100 unseen objects from unseen categories, following prior protocols (Xu et al., 2023; Wan et al., 2023; Huang et al., 2025b). Importantly, we use only the object assets and do not rely on ground-truth hand poses. Unless otherwise stated, we use the Shadow Hand in a tabletop lifting task, where success is defined as lifting the object off the table and maintaining a stable hold.

To obtain a vision-based policy under partial observability, we adopt a teacher–student framework following (Wan et al., 2023). A privileged state-based policy serves as the teacher, and we distill it into a vision-based student policy using DAgger (Ross et al., 2011). Methods marked with † denote results reproduced in our work. Additional implementation details are provided in Appendix A.1.

Evaluation Metrics We evaluate grasping performance using several complementary metrics. Grasp Success Rate (GSR, ↑) measures success rate of grasp-and-lift trials under a predefined stability criterion. Mean Success Affordance Distance (mSAD, ↓) evaluates how closely the realized contact aligns with the intended region. Style Diversity (SD, ↑) measures the diversity of successful grasp configurations, where higher values indicate more diverse strategies, following prior work (Mao et al., 2025). We further introduce Intent Success Rate (ISR, ↑), which counts successful grasps with fingertip contacts within a 4 cm threshold of the intended contact region, providing a stricter measure of intent-consistent lifting. Additional implementation details are provided in the Appendix A.3.

Table 1: Grasp Success Rates (GSR, ↑) on DexGraspNet (Wang et al., 2022) with the Shadow Hand. Our method achieves the highest performance across all settings.

Method	State-Based Setting			Vision-Based Setting		
	Train	Test		Train	Test	
		Seen Cat.	Unseen Cat.		Seen Cat.	Unseen Cat.
UniDexGrasp (Xu et al., 2023)	79.4	74.3	70.8	73.7	68.6	65.1
UniDexGrasp++ (Wan et al., 2023)	87.9	84.3	83.1	85.4	79.6	76.7
UniGraspTransformer (Wang et al., 2025)	91.2	89.2	88.3	88.9	87.3	86.8
DemoGrasp (Yuan et al., 2025)	95.2	95.5	94.4	92.2	92.3	90.1
DemoFunGrasp† (Mao et al., 2025)	90.0	91.9	89.9	91.4	85.8	84.5
SECOND-Grasp (Ours)	**97.9**	**98.2**	**97.7**	**95.7**	**96.5**	**94.9**

5.2 RESULTS ON GRASPING PERFORMANCE

Table 1 compares GSR against existing dexterous grasping methods on DexGraspNet Wang et al. (2022). Our method achieves the highest success rates across all settings, reaching 98.2% / 97.7% on seen / unseen categories in the state-based setting and 96.5% / 94.9% in the more challenging vision-based setting. DemoGrasp Yuan et al. (2025) and DemoFunGrasp Mao et al. (2025) adapt fixed demonstrations to diverse objects, while ResDex Huang et al. (2025b) uses geometry-unaware policies without explicitly modeling local contact geometry. In contrast, our method derives a pseudo target pose from the refined contact map via inverse kinematics, grounding the hand configuration in local object geometry. This geometry-aware supervision encourages consistent hand configurations for similar contact regions, supporting generalization to unseen objects.

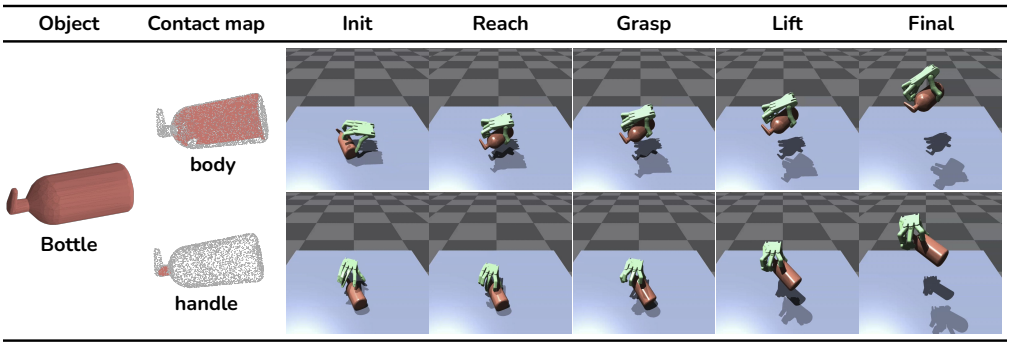

Figure 3: Qualitative results of intent-aware grasping. Given different intents on the same object, SECOND-Grasp adapts its grasp behavior to the specified contact region.

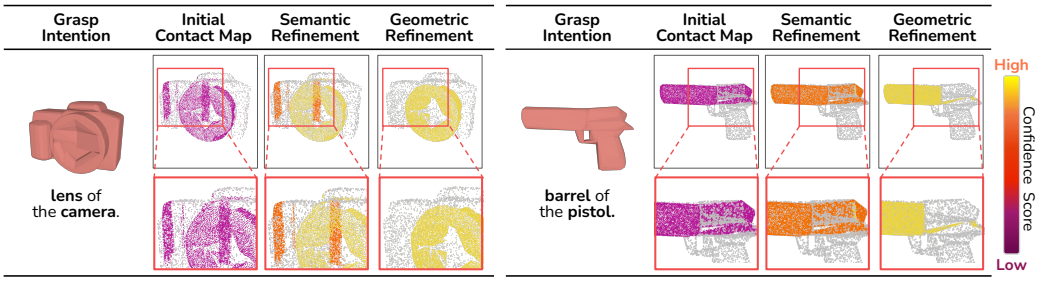

Figure 4: Qualitative visualization of SGCR. Given a grasp intention, SGCR refines coarse contact proposals into localized and geometrically coherent contact maps. Warmer colors indicate higher contact confidence.

5.3 Results on Intent-Aware Grasping

To evaluate whether successful grasps align with the intended contact regions, we report GSR, mSAD, and ISR. We reproduce DemoGrasp (Yuan et al., 2025) and DemoFunGrasp (Mao et al., 2025) under the same evaluation setting using 128 environments. As shown in Table 2, our method consistently outperforms prior work across both seen and unseen categories. In particular, our method achieves higher GSR and ISR while obtaining the lowest mSAD of 2.42 cm and 2.29 cm on seen and unseen categories, respectively, improving over DemoFunGrasp by 0.50 cm and 0.94 cm. This improved contact accuracy leads to higher ISR, indicating more accurate alignment with the intended contact regions, while the higher GSR reflects improved overall grasp success. Unlike prior methods that rely on a single contact point, our semantically and geometrically grounded contact map provides spatial guidance, enabling more precise localization of the intended region during execution. Additional qualitative visualizations are provided in Appendix D.7.

Table 2: Comparison of GSR, mSAD, and ISR on (Mao et al., 2025). Our method achieves higher success rates while better aligning with intended contact regions.

Method	Seen Cat.			Unseen Cat.		
	GSR↑	mSAD↓	ISR↑	GSR↑	mSAD↓	ISR↑
DemoGrasp† (Yuan et al., 2025)	95.50	4.10	52.57	94.40	4.27	33.71
DemoFunGrasp† (Mao et al., 2025)	91.92	2.92	73.74	89.90	3.23	62.63
SECOND-Grasp (Ours)	**98.24**	**2.42**	**86.52**	**97.65**	**2.29**	**88.87**

5.4 Qualitative Results

We present qualitative results of the proposed method. Figure 7 visualizes the refinement process of Semantic-Geometric Consistency Refinement (SGCR), where coarse semantic contact proposals are refined into more localized and geometrically consistent contact regions on the object surface. Figure 3 shows grasp executions conditioned on different contact maps. For the same object, specified regions lead to distinct grasp behaviors, such as grasping the bottle body when highlighted and grasping the handle when specified. These results show that our method follows the given contact intent rather than collapsing to a fixed grasp strategy. See Appendix D for more examples.

5.5 Ablation Study

We conduct an ablation study to analyze each component, including the pseudo pose prior, contact map, and Semantic-Geometric Consistency Refinement(SGCR). The pseudo pose prior serves as an alternative to dataset-provided joint configurations, maintaining performance while alleviating limited diversity of predefined grasp poses. However, without the contact map, the pseudo pose prior alone cannot generate grasps that reflect object-specific semantics or geometry. The contact map provides spatial cues that guide where to make contact, enabling semantically diverse grasp generation. Finally, SGCR refines the contact map through semantically grounded geometric consistency, improving grasp accuracy while preserving diversity. SGCR refines the contact map and is therefore evaluated with it.

Table 3: Ablation study on pseudo pose, contact map, and SGCR, evaluated using GSR and SD. Each component consistently improves grasp success and diversity.

Pseudo Pose	Contact Map	SGCR	Seen Cat.		Unseen Cat.	
			GSR↑	SD↑	GSR↑	SD↑
✗	✗	✗	90.82	0.90	88.96	0.89
✓	✗	✗	93.77	1.27	92.77	1.21
✓	✓	✗	97.85	1.54	96.88	1.68
✓	✓	✓	**98.24**	**1.71**	**97.65**	**1.80**

5.6 Cross-Dataset Generalization

Table 1 shows that our method achieves strong performance on DexGraspNet (Wang et al., 2022). To assess robustness under distribution shift, we further evaluate cross-dataset performance on several object datasets, including DexGraspAnything(DGA) (Zhong et al., 2025), EGAD (Morrison et al., 2020), Omni6DPose (Zhang et al., 2024b), ModelNet40 (Wu et al., 2015), and VisualDexterity (Chen et al., 2023)(Table 4). Our method consistently achieves the highest success rates across datasets. Notably, even compared to DemoGrasp (Yuan et al., 2025), which additionally trains on YCB (Calli et al., 2015), our method achieves substantial improvements on DGA (+18.81%) and ModelNet40 (+16.01%), indicating stronger robustness under distribution shift. We attribute this to our geometrically grounded contact maps, derived from object geometry and providing stable guidance across objects. We randomly rescale object sizes to 6-15cm following prior work (Yuan et al., 2025).

Table 4: Cross-dataset generalization (GSR, ↑) on five object datasets. Our method demonstrates improved robustness to distribution shifts.

Method	Train \ Test	DGA	EGAD	Omni6DPose	ModelNet40	VisualDexterity
UniGraspTransformer† (Wang et al., 2025)	DexGraspNet	44.53	69.80	47.95	40.45	68.63
DemoGrasp(Yuan et al., 2025)	DexGraspNet + YCB	60.99	94.58	74.84	71.88	93.67
DemoFunGrasp†(Mao et al., 2025)	DexGraspNet + YCB	50.00	83.84	57.58	50.50	81.82
SECOND-Grasp (Ours)	DexGraspNet	**79.80**	**96.68**	**77.84**	**87.89**	**94.14**

6 Conclusion

We presented SECOND-Grasp, a framework for dexterous grasping that bridges semantic intent and physical feasibility through contact-level reasoning. By combining vision-language reasoning, multi-view segmentation, and Semantic-Geometric Consistency Refinement, our method produces reliable 3D contact maps. From these contact maps, we derive pseudo target poses via inverse kinematics, providing pose priors for object-specific and physically feasible policy learning. SECOND-Grasp further derives contact maps and pose priors without training, enabling generalization to unseen objects. Extensive experiments demonstrate that our method improves grasping performance while capturing semantic intent and ensuring stable lifting. Results across multiple cross-dataset benchmarks confirm generalization, and comprehensive ablations validate the effectiveness of our approach.

Limitations and Future work. Our main contribution lies in generating and executing semantically meaningful contacts with a high-DoF dexterous hand. Following the scope commonly considered in existing dexterous grasping baselines, our current framework focuses on hand-level grasping and does not yet consider an arm–hand system. Extending the framework to arm–hand systems would further broaden its applicability to tasks involving reaching and grasping, which we leave for future work.

AI USE STATEMENT

We used an AI in a limited manner for grammar and style polishing only. It did not contribute to research ideation, experimental design, analysis. The authors take full responsibility for all contents.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Sam Buss. Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods. 2004.

Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pp. 510–517. IEEE, 2015.

Tao Chen, Megha Tippur, Siyang Wu, Vikash Kumar, Edward Adelson, and Pulkit Agrawal. Visual dexterity: In-hand reorientation of novel and complex object shapes. *Science Robotics*, 8(84): eadc9244, 2023.

Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1778–1787, 2021.

Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 5882–5889. IEEE, 2018.

Xianqiang Gao, Pingrui Zhang, Delin Qu, Dong Wang, Zhigang Wang, Yan Ding, and Bin Zhao. Learning 2d invariant affordance knowledge for 3d affordance grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3095–3103, 2025.

James J Gibson. The theory of affordances. the ecological approach to visual perception. *The people, place and, space reader*, pp. 56–60, 1979.

Jiawei He, Danshi Li, Xinqiang Yu, Zekun Qi, Wenyao Zhang, Jiayi Chen, Zhaoxiang Zhang, Zhizheng Zhang, Li Yi, and He Wang. Dexvlg: Dexterous vision-language-grasp model at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14248–14258, 2025.

Yu Huang, Zelin Peng, Changsong Wen, Xiaokang Yang, and Wei Shen. Unlocking 3d affordance segmentation with 2d semantic knowledge. *arXiv preprint arXiv:2510.08316*, 2025a.

Ziye Huang, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Efficient residual learning with mixture-of-experts for universal dexterous grasping. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Juntao Jian, Xiuping Liu, Zixuan Chen, Manyi Li, Jian Liu, and Ruizhen Hu. G-dexgrasp: Generalizable dexterous grasping synthesis via part-aware prior retrieval and prior-assisted generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11447–11457, 2025.

Junha Lee, Eunha Park, and Minsu Cho. Dexter: Language-driven dexterous grasp generation with embodied reasoning. *arXiv preprint arXiv:2601.16046*, 2026.

Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3086–3096, 2024a.

Haosheng Li, Weixin Mao, Weipeng Deng, Chenyu Meng, Haoqiang Fan, Tiancai Wang, Yoshie Osamu, Ping Tan, Hongan Wang, and Xiaoming Deng. Multi-grasplm: A multimodal llm for multi-hand semantic guided grasp generation. *arXiv preprint arXiv:2412.08468*, 2024b.

Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 8068–8074. IEEE, 2023.

Samuel Li, Sarthak Bhagat, Joseph Campbell, Yaqi Xie, Woojun Kim, Katia Sycara, and Simon Stepputtis. Shapegrasp: Zero-shot task-oriented grasping with large language models through geometric decomposition. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10527–10534. IEEE, 2024c.

Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20609–20620, 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.

Yumeng Liu, Yaxun Yang, Youzhuo Wang, Xiaofei Wu, Jiamin Wang, Yichen Yao, Sören Schwertfeger, Sibei Yang, Wenping Wang, Jingyi Yu, et al. Realdex: Towards human-like grasping for robotic dexterous hand. *arXiv preprint arXiv:2402.13853*, 2024.

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.

Yiyao Ma, Kai Chen, Kexin Zheng, and Qi Dou. Contact map transfer with conditional diffusion model for generalizable dexterous grasp generation. *arXiv preprint arXiv:2511.01276*, 2025.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Chuan Mao, Haoqi Yuan, Ziye Huang, Chaoyi Xu, Kai Ma, and Zongqing Lu. Universal dexterous functional grasping via demonstration-editing reinforcement learning. *arXiv preprint arXiv:2512.13380*, 2025.

WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Selective contrastive learning for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5210–5220, 2025.

Douglas Morrison, Peter Corke, and Jürgen Leitner. Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. *IEEE Robotics and Automation Letters*, 5(3): 4368–4375, 2020.

Richard M Murray, Zexiang Li, and S Shankar Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 2017.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7587–7597, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Shadow Robot Company. Dexterous hand. `https://www.shadowrobot.com/dexterous-hand-series/`, 2024.

Bruno Siciliano, Lorenzo Sciavicco, Luigi Villani, and Giuseppe Oriolo. *Robotics: modelling, planning and control*. Springer, 2009.

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.

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. In *European Conference on Computer Vision*, pp. 201–221. Springer, 2022.

Dylan Turpin, Tao Zhong, Shutong Zhang, Guanglei Zhu, Jingzhou Liu, Ritvik Singh, Eric Heiden, Miles Macklin, Stavros Tsogkas, Sven Dickinson, et al. Fast-grasp'd: Dexterous multi-finger grasp generation through differentiable simulation. *arXiv preprint arXiv:2306.08132*, 2023.

Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3891–3902, 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.

Wenbo Wang, Fangyun Wei, Lei Zhou, Xi Chen, Lin Luo, Xiaohan Yi, Yizhong Zhang, Yaobo Liang, Chang Xu, Yan Lu, et al. Unigrasptransformer: Simplified policy distillation for scalable dexterous robotic grasping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12199–12208, 2025.

Yi-Lin Wei, Jian-Jian Jiang, Chengyi Xing, Xian-Tuo Tan, Xiao-Ming Wu, Hao Li, Mark Cutkosky, and Wei-Shi Zheng. Grasp as you say: Language-guided dexterous grasp generation. *Advances in Neural Information Processing Systems*, 37:46881–46907, 2024.

Yi-Lin Wei, Mu Lin, Yuhao Lin, Jian-Jian Jiang, Xiao-Ming Wu, Ling-An Zeng, and Wei-Shi Zheng. Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11818–11828, 2025.

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.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Wonik Robotics. Allegro hand. `https://www.wonikrobotics.com/research-robot-hand`, 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.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22378–22389. IEEE, 2024.

Guo-Hao Xu, Yi-Lin Wei, Dian Zheng, Xiao-Ming Wu, and Wei-Shi Zheng. Dexterous grasp transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17933–17942, 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4737–4746, 2023.

Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10905–10915, 2023.

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

Haoqi Yuan, Ziye Huang, Ye Wang, Chuan Mao, Chaoyi Xu, and Zongqing Lu. Demograsp: Universal dexterous grasping from a single demonstration. *arXiv preprint arXiv:2509.22149*, 2025.

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.

Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pp. 199–216. Springer, 2024b.

Jiyao Zhang, Zhiyuan Ma, Tianhao Wu, Zeyuan Chen, and Hao Dong. Cadgrasp: Learning contact and collision aware general dexterous grasping in cluttered scenes. *arXiv preprint arXiv:2601.15039*, 2026.

Fuqiang Zhao, Dzmitry Tsetserukou, and Qian Liu. Graingrasp: Dexterous grasp generation with fine-grained contact guidance. In *2024 IEEE international conference on robotics and automation (ICRA)*, pp. 6470–6476. IEEE, 2024.

Yifan Zhong, Xuchuan Huang, Ruochong Li, Ceyao Zhang, Zhang Chen, Tianrui Guan, Fanlian Zeng, Ka Nam Lui, Yuyao Ye, Yitao Liang, et al. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18836–18844, 2026.

Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22584–22594, 2025.

He Zhu, Quyu Kong, Kechun Xu, Xunlong Xia, Bing Deng, Jieping Ye, Rong Xiong, and Yue Wang. Grounding 3d object affordance with language instructions, visual observations and interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17337–17346, 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*, pp. 2583–2602. PMLR, 2025.

A Additional Implementation Details

A.1 Simulation and Training Setup

We conduct our experiments in Isaac Gym (Makoviychuk et al., 2021) using a dexterous hand in a tabletop manipulation environment. The table size is set to $1.0 \times 1.0 \times 0.6$ m, and objects are initialized at a height of 0.7 m, i.e., 0.1 m above the tabletop. For the experiments, we use DexGraspNet (Wang et al., 2022) objects with scale factors of $\{0.06, 0.08, 0.10, 0.12, 0.15\}$. The task goal is to lift an object placed on the table to 0.3 m above the tabletop. Each training run takes approximately 16 hours on a single NVIDIA GeForce RTX 4090 GPU. The training hyperparameters used in simulation are reported in Table 5.

Table 5: Hyperparameters of our experimental setup.

Hyperparameter	**Value**
Common	
Episode length	200
Training epochs per iteration	5
Action moving average coefficient	1.0
Training iterations	20,000
Default random seed	8
State-based policy	
Number of environments	11,000
Parallel rollout steps per iteration	8
Number of minibatches per epoch	8
Vision-based policy	
Number of environments	4,096
Parallel rollout steps per iteration	1
Number of minibatches per epoch	4

A.2 Reward Terms

Here, we provide additional implementation details for the reward terms in Eq. 10. In implementation, each reward term is computed at every simulation timestep t. The policy reward consists of three components: a pose-guidance reward, a contact-conditioned pose reward, and the task reward. The first two terms encourage the hand to follow the reference hand state given by the IK-derived pseudo pose $h^\star$ and to make contact with the target region $\mathcal{Q}^{(k)}_{\text{final}}$, while the task reward drives the actual grasp-and-lift behavior. We first define a pose-tracking score that measures how closely the current hand state matches the reference hand state:

$$r_t^{\text{track}} = \exp(-\lambda_w \|\Delta w_t\|_2 - \lambda_\phi \|\Delta \phi_t\|_2 - \lambda_\theta \|\Delta \theta_t\|_1)\,, \tag{11}$$

where Δw_t, $\Delta \phi_t$, and $\Delta \theta_t$ denote the wrist-position, wrist-orientation, and joint-configuration errors with respect to the reference hand state, respectively. We set $\lambda_w = 7.0$, $\lambda_\phi = 2.0$, and $\lambda_\theta = 0.12$. The pose-guidance reward is then defined as

$$r_t^{\text{pose}} = \beta\, \kappa_t\, r_t^{\text{track}}, \tag{12}$$

where $\beta = 0.55$. The weight κ_t decays over the first 80 time steps with a minimum value of 0.15. This emphasizes reference-pose guidance early in the episode and gradually reduces its effect over time.

To encourage the fingertips to reach the intended contact region $\mathcal{Q}^{(k)}_{\text{final}}$, we add a contact-conditioned pose reward:

$$r_t^{\text{contact}} = \beta_c\, b_t\, r_t^{\text{track}}, \tag{13}$$

Here, $b_t \in \{0, 1\}$ denotes a binary indicator of fingertip contact with $\mathcal{Q}^{(k)}_{\text{final}}$, and we set $\beta_c = 0.25$. In addition to the pose-guidance terms, we use a task reward r_t^{task} based on the grasp-and-lift reward from the baseline (Huang et al., 2025b). It consists of approach, lift, goal-reaching, and bonus terms:

$$r_t^{\text{task}} = r_t^{\text{approach}} + r_t^{\text{lift}} + r_t^{\text{goal}} + r_t^{\text{bonus}}. \tag{14}$$

The approach term guides the hand toward the object. The lift term encourages the object to be raised from the table, and the goal-reaching term encourages the lifted object to move toward the target position. The bonus term provides an additional reward when the object is sufficiently close to the target. The total reward at timestep t is given by

$$r_t = r_t^{\text{pose}} + r_t^{\text{contact}} + r_t^{\text{task}}. \tag{15}$$

A.3 Evaluation Metrics

We evaluate each method using the following four metrics.

- **Grasp Success Rate (GSR)** GSR measures the fraction of episodes in which the object is lifted and held near the target position. An episode is considered successful if the object remains within 5 cm of the lifting goal for 20 consecutive time steps.
- **Mean Success Affordance Distance (mSAD)** mSAD measures intent alignment using the Euclidean distance between fingertip contacts and the intended contact region. It is averaged over successful episodes only, with lower values indicating better alignment.
- **Style Diversity (SD)** SD measures the diversity of hand postures among successful episodes. We collect the final hand joint configuration from each successful episode, compute pairwise Euclidean distances among them, and report the average distance.
- **Intent Success Rate (ISR)** ISR measures the fraction of episodes that satisfy both grasp success and intent-consistent contact. For episode n, intent success is defined as $\mathrm{GSR}_n = 1$ and $\mathrm{SAD}_n < 0.04\,\mathrm{m}$. The reported ISR is averaged over all evaluation episodes.

B Contact Map Generation Details

B.1 VLM-based Grasp Intent Proposal

For each object, we render four side-view RGB images (e.g., front, left, back, and right) and provide them jointly to Qwen3-VL-8B-Instruct (Bai et al., 2025). The VLM predicts between two and four grasp intents, where each intent corresponds to a distinct object-centric graspable part.

For each intent, the VLM outputs a part name, a short description, per-view visibility, a 2D bounding box for each visible view, and a confidence score. The prompt discourages overly complex object-part names, enforces consistent naming across views, and suppresses bounding-box predictions for views where the part is not visible. The VLM output is required to be valid JSON, which is parsed into a structured set of per-view part proposals. The full prompt, including a compact description of the required JSON output format, is shown in Fig. 5.

We decode with temperature 0.2, repetition penalty 1.1, and a maximum of 3000 new tokens. During post-processing, invalid bounding boxes are removed and confidence scores are clipped to $[0, 1]$. If the visibility flag is missing or unreliable, we determine visibility based on whether a valid bounding box exists. These predictions are used as coarse semantic proposals; the final 3D contact maps are obtained through mask extraction, depth projection, and Semantic-Geometric Consistency Refinement.

B.2 Semantic-Geometric Contact Map Generation

This section provides additional implementation details for generating the semantic-geometric contact maps used in our experiments. Starting from the VLM-predicted (Bai et al., 2025) per-view bounding boxes, we first apply SAM2 (Ravi et al., 2024) within each box to obtain a binary mask for the corresponding graspable part. The extracted mask is intersected with the valid object silhouette obtained from the rendered depth map, which removes mask leakage outside the object region.

For each grasp intention, we then construct a per-view confidence map by assigning the VLM confidence score to pixels inside the filtered mask and zero elsewhere. These confidence maps are refined using neighboring-view support: mask pixels are lifted to 3D using the rendered depth and camera parameters, reprojected to adjacent views, and considered supported only when the reprojected location is depth-consistent and lies inside the corresponding part Eq. 5. Supported

Prompt Details for Contact Map Generation

You are given 4 RGB images of the `object`, captured sequentially while moving counter- clockwise around the object from an arbitrary starting angle. All images are side views of the object.

Task:
Analyze the four views jointly, predict between `min_part` and `max_part` distinct object-centric graspable parts, and localize each predicted part consistently across all visible views using tight 2D bounding boxes.

Rules:
- Each intention must correspond to one distinct physical 3D part that is plausible for robotic grasping.
- Use simple, broad, human-readable part names, and keep the name for the same physical part across views.
- Intention names must be simple, human-readable object-part names in plain English.
- Output a bbox only when the target part is visible; otherwise set visible to false and bbox to null.
- Bounding boxes must tightly cover only the visible region of the target part.
- Confidence score must be in [0, 1].

You are an intelligent assistant for multi-view robotic grasp-part prediction. First predict multiple graspable intentions from all four views, then localize each intention consistently across the four views.

Output format:
- The model outputs a JSON object with the object name and grasp intentions.
- Each intention contains a part name and per-view bbox, visibility, and confidence score for image_1–image_4.

Figure 5: Prompt used for VLM-based grasp intent proposal, including the JSON output format.

pixels receive additional confidence from the neighboring view, and the resulting maps are globally normalized across the four views.

The normalized 2D confidence maps are finally lifted to the object surface to obtain a scored 3D point cloud for each grasp intention. We aggregate points from all visible views and associate each point with its refined confidence score. To remove fragmented or geometrically inconsistent regions, we select a high-confidence seed component and expand it only to nearby candidate points that satisfy the local convexity criterion in Eq. 7. The remaining points form the final semantic-geometric contact map used for inverse-kinematics-based pose initialization and policy learning. Figure 7 visually illustrates the overall semantic-geometric contact map generation process. Robustness to noisy and incomplete geometric observations and sensitivity to the local-convexity threshold are evaluated in Appendix B.3.

B.3 Robustness of Semantic-Geometric Consistency Refinement

Sensing variations SGCR refines imperfect initial contact maps to improve contact localization. To evaluate SGCR under sensing variations, we generate more imperfect contact maps by introducing depth noise and reduced camera coverage. Specifically, we add Gaussian depth noise with $\sigma \in \{2, 5, 10\}$ mm and reduce angular camera coverage from $360°$ to $270°$, $180°$, and $90°$. As shown in Table 6, GSR decreases only moderately as the depth noise increases, remaining above 95% for both seen and unseen categories even at $\sigma = 10$ mm. The performance is more affected by reduced camera coverage, particularly under the $90°$ setting. This is expected, as SGCR relies on cross-view semantic refinement to improve the initial contacts, and fewer available views provide less cross-view evidence for this refinement.

Table 6: Robustness to depth noise and reduced angular camera coverage. GSR is reported in percent.

Metric	**Clean**	**Depth Noise**			**Camera Coverage**		
		$\sigma = 2$ mm	$\sigma = 5$ mm	$\sigma = 10$ mm	$270°$	$180°$	$90°$
Seen cat. GSR ↑	98.2	97.3	96.7	95.5	95.9	92.9	89.8
Unseen cat. GSR ↑	97.7	95.9	95.8	95.2	93.6	93.4	86.7

Sensitivity to the local-convexity threshold We further analyze the effect of the local-convexity threshold used in geometric refinement. A threshold of 0.0 corresponds to the baseline without geometric refinement, while positive thresholds progressively filter candidate contact points based

Table 7: Sensitivity to the local-convexity threshold. Refinement Ratio denotes the percentage of the initial contact region removed by geometric refinement.

Metric	w/o SGCR	Local-Convexity Threshold				
	0.0	0.2	0.4	**0.6 (Ours)**	0.8	1.0
Refinement Ratio (%)	0	20.4	24.7	28.5	33.9	44.7
Seen cat. GSR ↑	97.9	98.0	98.0	**98.2**	98.2	96.1
Unseen cat. GSR ↑	96.9	97.1	97.3	**97.7**	96.1	95.7

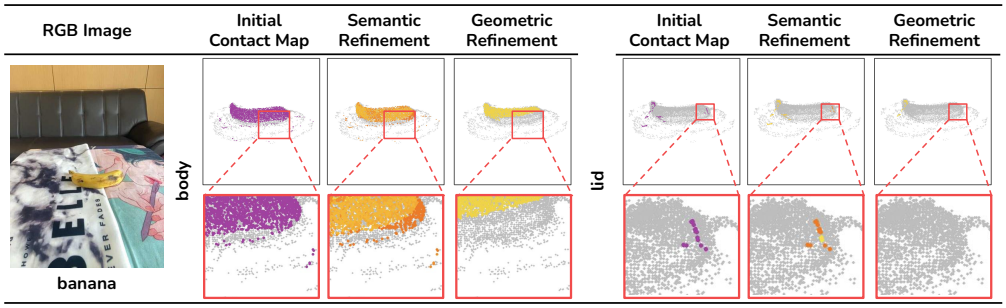

Figure 6: Qualitative SGCR result on a real RGB-D sample from WildRGB-D (Xia et al., 2024).

on local-convexity consistency. As shown in Table 7, even a threshold of 0.2 improves GSR over the baseline, indicating the benefit of introducing geometric refinement. Performance remains stable over a broad range of thresholds, and 0.6 achieves the highest success rate while matching the best seen-category performance. We therefore use 0.6 as the default threshold throughout our experiments.

Upstream error propagation Localization errors in VLM bounding boxes and SAM2 masks can propagate through the contact map to the IK-derived pose guidance. SGCR addresses cross-view disagreement by increasing confidence for depth-consistent points supported by the corresponding part masks, followed by local-convexity filtering of geometrically inconsistent candidates. This refinement reduces the influence of inconsistent proposals before they are used to construct hand-level guidance.

Real-scene contact maps We further apply SGCR to WildRGB-D (Xia et al., 2024), which is captured in the wild with diverse backgrounds and includes imperfect depth observations. Even on such real-world data, SGCR effectively refines the initial contacts, demonstrating its robustness to in-the-wild RGB-D observations. Figure 6 shows a representative example on real RGB-D data from WildRGB-D, where SGCR progressively sharpens the contact maps for the body and tip regions while suppressing geometrically inconsistent candidates. Additional qualitative visualizations are provided in the supplementary material (`index.html`, *Real-Scene Contact Maps*).

C Inverse Kinematics

C.1 Jacobian-Based Inverse Kinematics

This appendix details the local damped least-squares solver (Buss, 2004) used to optimize Eq. 9. Starting from an initial hand configuration $h_0 = (w_0, \phi_0, \theta_0)$, we perform 12 iterative updates.

PCA-based finger-region assignment In the object-local coordinate frame, we center the refined contact points by their mean and estimate the principal axis using PCA (Wold et al., 1987). To resolve the sign ambiguity, we orient the axis so that its object-local z component is non-negative. We then project the centered points onto this axis and use the median projection as the partition threshold. Points whose projection values do not exceed the median form the thumb region, while the remaining points form the non-thumb region.

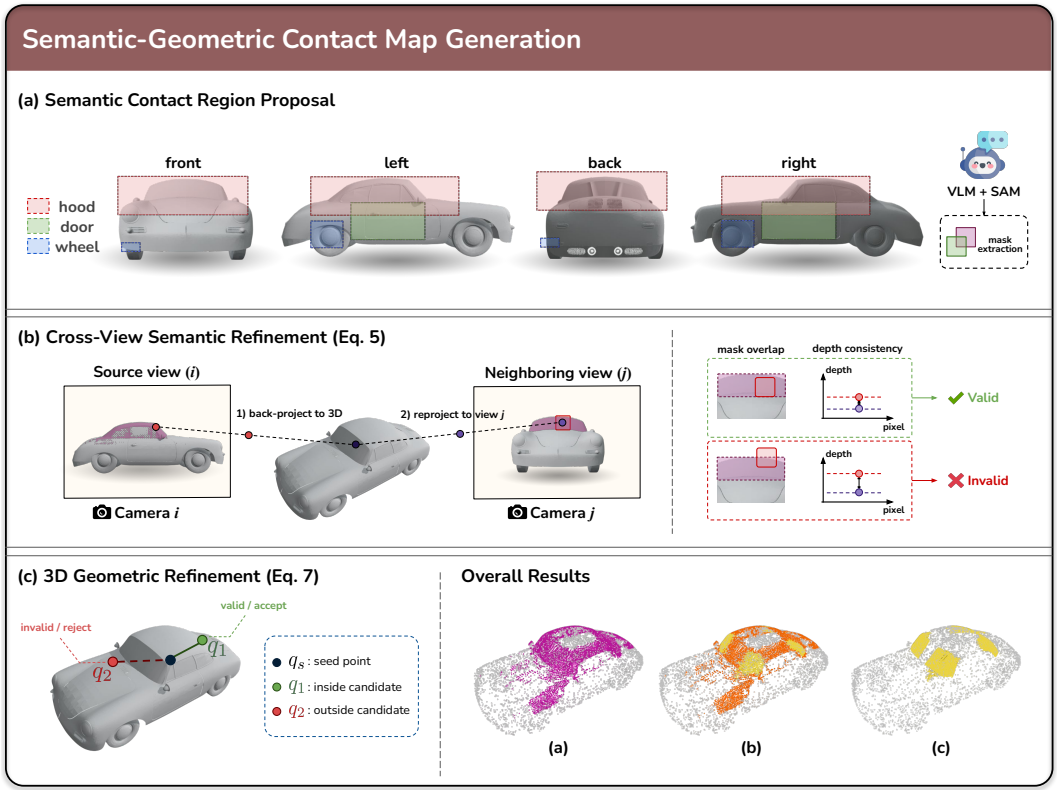

Figure 7: Overview of Semantic Contact Map Generation. Visual illustration of the geometric refinement steps in Semantic-Geometric Consistency Refinement.

For hands with multiple non-thumb fingers, we sort the non-thumb points along the object-local lateral direction and divide them into contiguous groups with approximately equal numbers of points. These groups are assigned to the fingers in anatomical order to define their target contact regions $\mathcal{R}_f^{(k)}$. The regions are fixed before IK and remain unchanged throughout its iterations.

IK updates At each update, given the current configuration h, we first instantiate the closest target point for each fingertip within its assigned contact region:

$$y_f^\star = \arg\min_{y \in \mathcal{R}_f^{(k)}} \|\xi_f(h) - y\|_2 . \tag{16}$$

The fingertip-position residuals are then concatenated as

$$e = \left[\xi_f(h) - y_f^\star\right]_{f \in F} . \tag{17}$$

Let $J = \partial e / \partial h$ denote the stacked fingertip Jacobian evaluated at the current configuration. We compute the damped least-squares update as

$$\Delta h = -J^\top \left(JJ^\top + \lambda_{\text{dls}}^2 I\right)^{-1} e, \qquad h \leftarrow h + \eta \Delta h, \tag{18}$$

where λ_{dls} is the damping coefficient and η is the step size. After each update, the nearest target points are recomputed from the assigned contact regions, yielding a local alternating procedure between target-point selection and IK update.

Let the final IK state be $h^\star = (w^\star, \phi^\star, \theta^\star)$, which serves as the reference hand configuration for pose guidance during policy learning. The pseudo pose is also included in the policy observation as the intent condition described in Section 4.4.

Pose guidance and intent conditioning During training, the IK-derived pseudo pose is included in the policy observation as an intent condition and, together with the refined contact map, is used to shape the reward as described in Appendix A.2. For standard grasping evaluation, no test-time contact map, pseudo pose, or IK is required. For intent-aware evaluation, the requested contact region is refined by SGCR and converted by IK into a pseudo pose, which is supplied through the same observation interface used during training. Both settings use the same trained policy without retraining.

C.2 Analysis of IK-based Pseudo Pose Guidance

The role of the pseudo pose in SECOND-Grasp is not to produce a final executable grasp, but to provide an intermediate hand-level prior for policy learning. Given a refined semantic-geometric contact map, we only need a hand configuration that roughly aligns the fingertips with the intended contact region while respecting the kinematic structure of the target hand. Thus, we prioritize efficient contact-aligned guidance over exhaustive pose optimization.

To determine the number of IK iterations used in our pseudo-pose generation pipeline, we evaluate the efficiency–alignment trade-off by running IK with 1, 3, 12, and 50 iterations and comparing it with baseline pose-generation methods on the same refined contact maps $\mathcal{Q}_{\text{final}}^{(k)}$. We sample 100 objects from DexGraspNet (Wang et al., 2022) and report inference time and Contact Coverage (Cov.). Inference time is measured in seconds as the average wall-clock time required to generate one pseudo pose from the same object and refined contact map input. Following prior works (Wei et al., 2024; Ma et al., 2025), Cov.@τ measures the percentage of hand surface points whose signed distance to the object surface lies within $[-\tau, \tau]$, where $\tau \in \{2\text{ mm}, 5\text{ mm}\}$.

As shown in Table 8, increasing the number of IK iterations improves contact alignment. The 12-iteration setting achieves the best Cov.@2mm and matches the best Cov.@5mm among IK settings, while increasing to 50 iterations does not improve coverage and increases inference time by more than $4\times$. Based on this trade-off, we adopt 12 IK iterations throughout our method. Compared with optimization- and generative-based pose-generation methods, this setting provides competitive contact alignment with substantially lower computational cost. These results support our choice of IK-based pseudo poses as efficient intermediate guidance for policy learning rather than final grasp synthesis.

Table 8: Efficiency and contact-alignment comparison for pseudo-pose generation. IK with 12 iterations provides a practical trade-off between contact coverage and inference time, making it suitable as intermediate guidance for policy learning.

Metrics	Inverse Kinematics				Optimization-based Method	Generative-based Method	
	Iter. $= 1$	Iter. $= 3$	Iter. $= 12$	Iter. $= 50$	GenDexGrasp (Li et al., 2023)	UGG (Lu et al., 2024)	DGA (Zhong et al., 2025)
Inf. Time (s) ↓	0.004	0.037	0.187	0.836	11.329	9.222	0.622
Cov.@2mm (%) ↑	2.220	3.330	3.780	3.110	2.622	0.001	1.206
Cov.@5mm (%) ↑	12.000	15.110	15.110	14.440	16.585	2.547	8.839

C.3 Effect of IK-derived Pseudo Pose on Policy Learning

We further analyze how the IK-derived pseudo pose affects policy learning. Learning dexterous grasping from task reward alone often requires extensive exploration, as the policy must discover effective contact locations and finger coordination by itself. The IK-derived pseudo pose converts the refined contact map into a hand-level target $h^\star$ that aligns the fingertips with the intended contact regions.

To isolate this effect, we train two policy variants under the same task reward, object split, simulation setting, and hyperparameters: one with auxiliary supervision toward the pseudo pose $h^\star$, and one without it. For each variant, we plot the training time episode success rate and normalized reward in Figure 8. Without pseudo-pose supervision, the policy improves only after a long delay, indicating that contact discovery from task feedback alone is inefficient. In contrast, the policy trained with $h^\star$ achieves high reward and success much earlier, showing that the pseudo pose serves as an intermediate hand-pose prior that accelerates early contact formation and grasp stabilization. The final performance gap is smaller than the early training gap, suggesting that the main benefit of pseudo-pose supervision is not to prescribe the final grasp, but to guide the policy toward contact-rich states during learning.

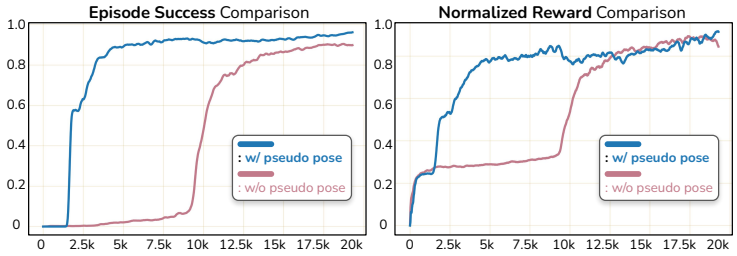

Figure 8: Effect of IK-derived Pseudo-Pose Supervision on Policy Learning.

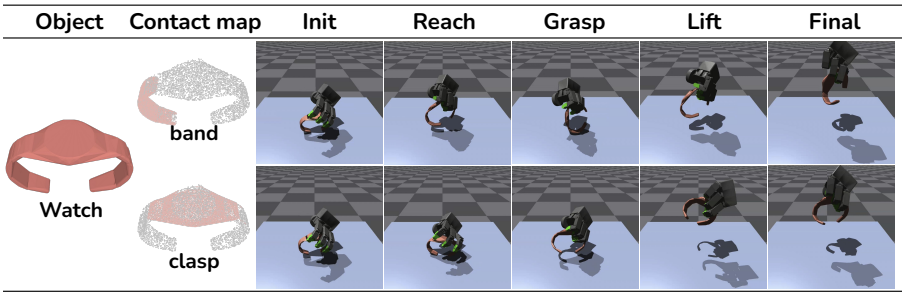

Figure 9: Additional qualitative results of intent-aware grasping on the Allegro Hand.

D Additional Results

D.1 Grasp Style Diversity

Since each object has multiple semantic intentions, the policy should generate distinct grasp strategies rather than collapsing to a single strategy. A policy that fails to account for such variation would produce similar grasp configurations across different situations. To evaluate whether the policy avoids such collapse, we measure grasp style diversity (SD) on the DexGraspNet (Wang et al., 2022), which quantifies the diversity of successful grasping strategies by computing the average Euclidean distance between all pairs of successful hand joint configurations. As shown in Table 9, our method achieves higher SD than prior works (Xu et al., 2023; Yuan et al., 2025; Mao et al., 2025). Importantly, this improvement comes alongside higher grasp success rates, indicating that the gains are not due to trivial diversification but reflect effective grasp strategies.

Table 9: Grasp Style Diversity (SD, ↑) on (Wang et al., 2022). Our method produces more diverse successful grasp strategies.

Method	Seen Cat.	Unseen Cat.
DemoGrasp† (Yuan et al., 2025)	1.66	1.65
DemoFunGrasp (Mao et al., 2025)	1.48	1.44
Ours	**1.71**	**1.80**

D.2 Transfer to Allegro Hand

We further test whether the same contact-map representation can support grasping with a different robot hand, the Allegro Hand. This transfer is not a direct pose copy because the two hands have different finger layouts and joint spaces. Shadow Hand commonly has five digits and 20 actuated DoFs, whereas Allegro Hand has four fingers and 16 actuated DoFs. Therefore, a joint pose defined for Shadow Hand cannot be directly applied to Allegro Hand. Prior methods (Yuan et al., 2024) often transfer grasps through hand-specific pose retargeting, which may require additional optimization or training. Our method instead keeps the contact map on the object and recomputes the joint target using IK with the Allegro Hand. Thus, the intended object region stays the same, while the hand pose is adapted to the new hand embodiment.

Table 10: Comparison of pseudo-pose efficiency across different IK iterations and baseline pose-generation methods. All methods are evaluated on the same sampled DexGraspNet training objects.

Hand / Dataset	Metrics	Inverse Kinematics				Generative-based Method
		Iter. $= 1$	Iter. $= 3$	Iter. $= 12$	Iter. $= 50$	Dexdiffuser (Weng et al., 2024)
Allegrohand (DexGraspNet)	Inf. Time (s) ↓	0.028	0.078	0.302	0.468	5.086
	Cov.@2mm ↑	0.005	0.026	0.038	0.040	0.003
	Cov.@5mm ↑	0.056	0.087	0.158	0.173	0.033

Figure 9 shows that the Allegro Hand grasps and lifts the object from the intended region, such as the watch band or clasp. Table 10 shows that 12 iterations of IK provide a good balance between runtime and contact coverage. Compared with the pose-generation baseline, our result is much faster ($0.302\,\mathrm{s}$ vs. $5.086\,\mathrm{s}$) and achieves higher Cov.@5mm (0.158 vs. 0.033). These results show that our method provides reliable contact map and effective initial poses even for a different hand embodiment, demonstrating that the intent-aware representation can transfer beyond the original hand embodiment.

D.3 Variability Across Training Seeds

We train DemoGrasp (Yuan et al., 2025) and our method using the same five independent random seeds to assess training stability. As shown in Table 11, our method achieves higher mean GSR, ISR, and SD and lower mean mSAD on both seen and unseen categories. The smaller standard deviation in GSR further supports consistent grasping performance across independent training runs.

Table 11: Performance across five independent training seeds, reported as mean ± standard deviation. GSR and ISR are in percent, and mSAD is in centimeters.

Method	GSR ↑	ISR ↑	mSAD ↓	SD ↑
Seen categories				
DemoGrasp (Yuan et al., 2025)	94.2 ± 1.9	48.2 ± 2.7	4.27 ± 0.30	1.50 ± 0.17
Ours	97.9 ± 0.7	85.4 ± 2.6	1.88 ± 0.33	1.86 ± 0.18
Unseen categories				
DemoGrasp (Yuan et al., 2025)	92.9 ± 1.8	30.3 ± 3.5	4.21 ± 0.22	1.52 ± 0.17
Ours	97.3 ± 0.4	85.4 ± 2.2	1.79 ± 0.31	1.80 ± 0.15

D.4 Sensitivity to the ISR Threshold

ISR is introduced by DemoFunGrasp (Mao et al., 2025), which uses a contact-distance threshold of 4 cm as its default setting. For a fair comparison, we follow the same 4 cm threshold in our main evaluation. To examine the sensitivity of ISR to this threshold, we further evaluate performance over thresholds ranging from 1 to 5 cm. A smaller threshold corresponds to a stricter constraint for intended grasping, while a larger threshold corresponds to a more relaxed constraint. As shown in Table 12, ISR consistently increases as the threshold is relaxed from 1 to 5 cm, as expected, while our method outperforms DemoGrasp (Yuan et al., 2025) across all tested thresholds, including the stricter 1–3 cm settings.

Table 12: ISR sensitivity to the contact-distance threshold. All values are percentages. ISR requires both grasp success and alignment within the specified threshold.

Method	GSR ↑	ISR ↑				
		1 cm	2 cm	3 cm	**4 cm (default)**	5 cm
DemoGrasp (Yuan et al., 2025)	94.4	8.3	16.8	22.1	33.7	34.2
Ours	97.7	17.2	47.3	72.3	88.9	90.0

D.5 Robustness under Realistic Execution Conditions

As mentioned in UniDexGrasp++ (Wan et al., 2023), vision-based policies provide a setting closer to real-world deployment by avoiding privileged object-state information. Following this setting,

Table 13: Robustness of the fixed vision-based policy to execution perturbations on seen categories. Nominal denotes the default setting without added perturbations. GSR and ISR are reported in %, and mSAD in cm. Friction multipliers are relative to the default simulator value.

Method	Metric	Nominal (Clean)	Action Execution Noise			Control Latency			Hand–Object Friction	
						1 step	2 steps	3 steps		
			$\sigma = 0.02$	$\sigma = 0.05$	$\sigma = 0.10$	(16.6 ms)	(33.2 ms)	(49.8 ms)	$0.5\times$	$1.5\times$
Ours	GSR↑	96.5	93.3	92.9	93.1	90.1	83.7	73.0	91.0	97.0
	ISR↑	86.2	82.4	80.7	80.8	82.7	78.5	65.6	81.4	79.4
	mSAD↓	2.31	2.97	2.71	2.71	2.83	2.79	3.36	2.49	2.76

we further introduce action noise, control latency, and friction variations to better reflect execution uncertainties encountered on real robots. We evaluate the fixed vision-based policy on objects from seen categories under these perturbations without retraining. As shown in Table 13, performance remains relatively stable across the action-noise levels and friction settings, while larger control latency leads to a more noticeable decrease in GSR and ISR.

D.6 Qualitative Results of Contact-Map Refinement

Figure 10 shows additional qualitative examples of Semantic-Geometric Consistency Refinement (SGCR) on objects from datasets beyond DexGraspNet (Wang et al., 2022). Starting from coarse contact proposals, SGCR removes noisy regions and refines the contact map around the intent-specified object part. Across diverse datasets, including DexGraspAnything (DGA) (Zhong et al., 2025), EGAD (Morrison et al., 2020), Omni6DPose (Zhang et al., 2024b), ModelNet40 (Wu et al., 2015), and VisualDexterity (Chen et al., 2023), the refined maps become more compact and geometrically consistent while preserving the intended contact semantics. These results provide additional evidence that SGCR remains effective under cross-dataset object distribution shifts.

D.7 Qualitative Results of Intent-aware Grasping

Figure 11 presents additional qualitative examples of intent-aware grasping across diverse object categories and semantic part intents. Rather than producing a single stereotyped grasp, our method adapts the refined contact map and the resulting grasp behavior to the requested intent. For instance, on the same microscope object, different intents such as head, arm, and base lead to distinct contact maps concentrated on the corresponding semantic regions. The learned policy then follows these intent-specific contact cues to approach, grasp, and lift the object from the desired part, resulting in clearly different hand placements. Across objects and semantic parts, these examples highlight that our method performs intent-conditioned grasping instead of relying on a fixed grasping pattern.

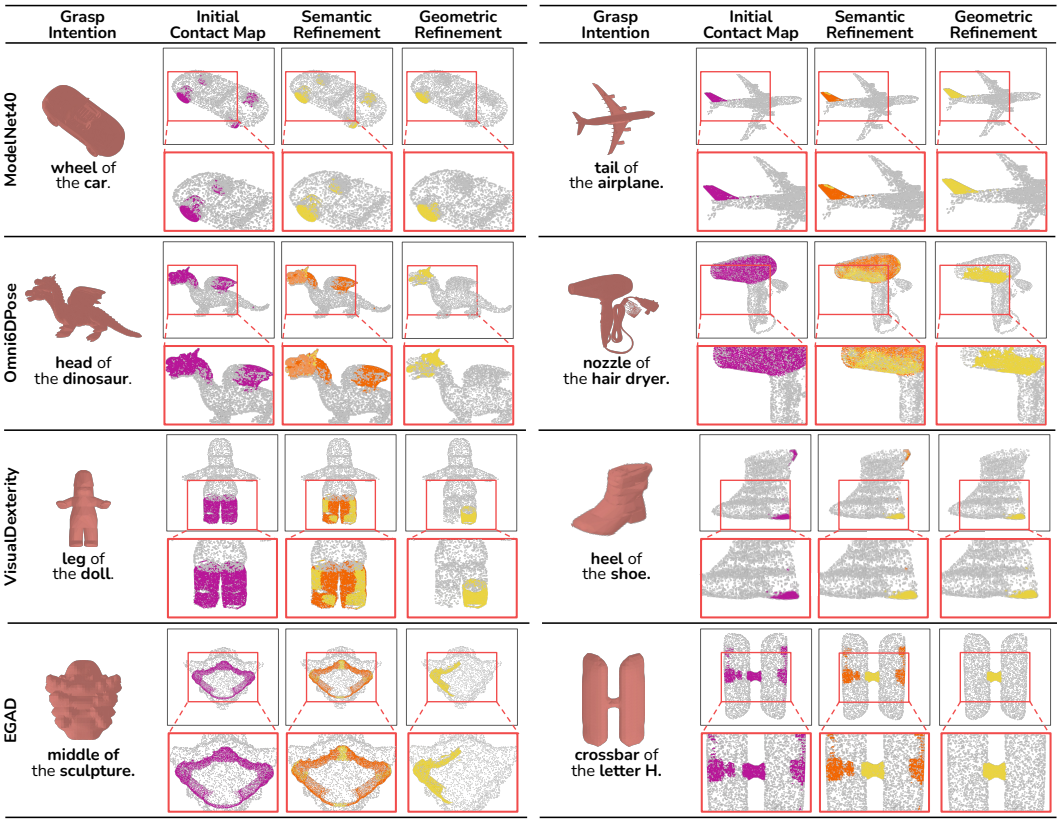

Figure 10: Cross-dataset qualitative results of SGCR contact-map refinement. Coarse contact proposals are refined into geometrically coherent contact maps on objects from diverse datasets.



Figure 11: Additional qualitative results of intent-aware grasping. SECOND-Grasp adapts the refined contact map and resulting grasp behavior according to the specified semantic part intent.