The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

Towards Affordance-Aware Robotic Dexterous Grasping with Human-like Priors

Haoyu Zhao [*1, 2], **Linghao Zhuang** [*1], **Xingyue Zhao** [*2], **Cheng Zeng**[5],
Haoran Xu[4], **Yuming Jiang**[2], **Jun Cen**[2, 3, 4], **Kexiang Wang**[2], **Jiayan Guo**[2],
Siteng Huang[†2, 3, 4], **Xin Li**[2, 3], **Deli Zhao**[2, 3], **Hua Zou**[†1]

[1] Wuhan University
[2] DAMO Academy, Alibaba Group
[3] Hupan Lab
[4] Zhejiang University
[5] Tsinghua University
haoyu.zhao@whu.edu.cn, siteng.huang@gmail.com

Abstract

A dexterous hand capable of generalizable grasping objects is fundamental for the development of general-purpose embodied AI. However, previous methods focus narrowly on low-level grasp stability metrics, neglecting affordance-aware positioning and human-like poses which are crucial for downstream manipulation. To address these limitations, we propose **AffordDex**, a novel framework with two-stage training that learns a universal grasping policy with an inherent understanding of both motion priors and object affordances. In the first stage, a trajectory imitator is pre-trained on a large corpus of human hand motions to instill a strong prior for natural movement. In the second stage, a residual module is trained to adapt these general human-like motions to specific object instances. This refinement is critically guided by two components: our Negative Affordance-aware Segmentation (NAA) module, which identifies functionally inappropriate contact regions, and a privileged teacher-student distillation process that ensures the final vision-based policy is highly successful. Extensive experiments demonstrate that AffordDex not only achieves universal dexterous grasping but also remains remarkably human-like in posture and functionally appropriate in contact location. As a result, AffordDex significantly outperforms state-of-the-art baselines across seen objects, unseen instances, and even entirely novel categories.

Project Page — https://afforddex.github.io/

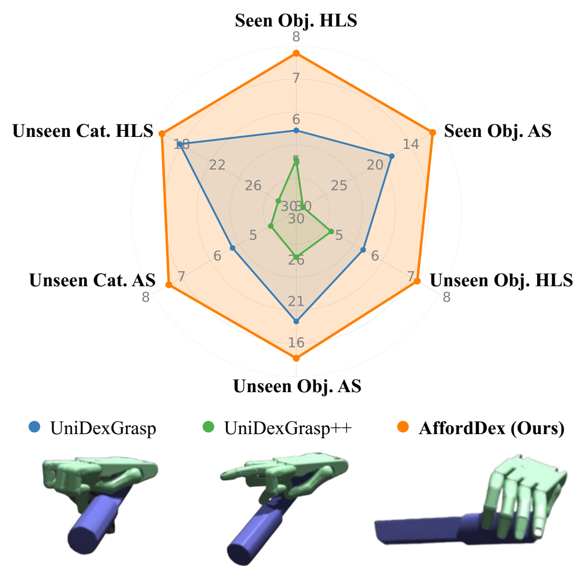

Figure 1: Performance comparision among UniDexGrasp (Xu et al. 2023), UniDexGrasp++ (Wan et al. 2023), and our AffordDex, on the vision-based setting. we report human-likeness score (HLS) and affordance score (AS) across seen objects, unseen objects, and unseen categories. We also present a qualitative comparison, where AffordDex performs natural and safe grasping by avoiding the blade.

1 Introduction

Dexterous grasping, as a foundational capability for robotic manipulation, has garnered significant attention from both academia and industry (Zhao et al. 2024, 2025b). Compared to simpler end-effectors (e.g., parallel jaws, vacuum grippers), dexterous hands closely resemble human hand structure, providing substantially enhanced flexibility, precision, and task adaptability (Zhong et al. 2025). Furthermore, anthropomorphic robots expedite the collection of rich human demonstration data via teleoperation (Li et al. 2025a). Consequently, this synergy has fueled rapid progress, with recent algorithms achieving high success rates in generalizing grasps to novel objects (Fang et al. 2022, 2020; Gou et al. 2021; Wang et al. 2021; Xu et al. 2023; Wan et al. 2023).

Due to the high degrees of freedom (DOFs) of dexterous hands, traditional motion planning-based methods (Andrews and Kry 2013; Bai and Liu 2014) struggle to handle such complex hand joint movements. Recent advancements in reinforcement learning (RL) (Wan et al. 2023; Mandikal and Grauman 2022; Christen et al. 2022; Nagabandi et al. 2020; Mandikal and Grauman 2021) have shown promising results

[*]Equal contributions.
[†]Corresponding Author.

in complex dexterous manipulation. However, the goal of grasping is not merely to lift an object. It involves alignment with human intent and preparation for subsequent manipulation tasks, such as avoiding the blade of a knife or preparing to open a bottle cap. Existing methods, while focused on low-level grasp stability metrics, largely overlook this crucial synthesis of affordance-aware positioning and human-like kinematics, limiting their utility in real-world, multi-step manipulation scenarios.

In this work, we focus on the critical aspect of safety and functional correctness by modeling *negative* affordances—regions to be avoided, which provide clear, unambiguous negative constraints and thus simplify the learning problem. We propose **AffordDex**, a novel framework that learns a universal grasping policy that is both human-like in its motion and functionally aware of object affordances. We achieve this through a structured, two-stage training paradigm. In the first stage, we pre-train a base policy on a large corpus of human hand motions to instill a strong prior for natural movement. In the second stage, a residual module is trained to adapts the general human-like motions from the pre-trained policy to specific objects. This refinement is critically guided by our proposed Negative Affordance-aware Segmentation (NAA) module, which provides explicit visual-geometric constraints on functionally inappropriate contact regions. Moreover, the training is enhanced with a teacher-student distillation framework, which leverages ground-truth state information to ensure the final vision-based policy is highly effective and robust.

As illustrated in Fig. 1, AffordDex produces successful grasps that functionally correct and human-like grasps, such as safely grasping a knife by its handle. Extensive experiments validate that our method significantly outperforms existing approaches on both seen and unseen objects. In summary, AffordDex makes the following contributions:

- We propose **AffordDex**, a two-stage framework that synergistically and effectively integrates human motion priors with functional affordance constraints to achieve generalizable and anthropomorphic dexterous grasping.
- We introduce a Negative Affordance-aware Segmentation (NAA) module that, by reformulating segmentation as a VLM-guided classification problem, provides explicit geometric constraints to prevent functionally improper grasps.
- Extensive experiments demonstrate AffordDex achieves SOTA success rates across multiple levels of generalization while producing grasps that are qualitatively superior in human-likeness and functional appropriateness.

2 Related Work

2.1 Dexterous Grasping

Robotic grasping (Fang et al. 2022, 2020; Gou et al. 2021; Wang et al. 2021) has been a longstanding research, aiming to enable robots to interact with objects reliably and adaptively. While significant advances have been made with simple parallel-jaw grippers (Fang et al. 2020; Mahler et al. 2019), their limited dexterity restricts adaptability to objects with intricate geometries. Dexterous, multi-fingered hands (Xu et al. 2023; Wan et al. 2023) offer a solution but pose a severe control challenge for traditional analytical methods (Bai and Liu 2014; Liu et al. 2021), motivating the shift towards learning-based approaches.

One paradigm decouples the grasping process into static grasp pose generation followed by a dynamic grasping through trajectory planning or goal-conditioned reinforcement learning (RL) (Wan et al. 2023; Christen et al. 2022; Wang et al. 2025). For example, UniDexGrasp++ proposes geometry-aware curriculum learning and leverages the geometry feature for RL. However, these RL-based methods may produce physically unrealistic joint configurations. An alternative paradigm directly learns the entire grasping trajectory through expert demonstrations from humans or reinforcement learning agents (Xu et al. 2023; Liu et al. 2024; Huang et al. 2023; Lu et al. 2024; Zhang et al. 2024b,a). These approaches tend to achieve more natural motions but suffer from poor generalization to novel objects due to the limited diversity of demonstrations and inherent policy constraints. To address these failures, our AffordDex combines strong, human-derived motion priors to ensure natural movement with affordance-based guidance to achieve robust generalization, resulting in a policy that is both natural and functionally effective across a wide range of objects.

2.2 Affordance Prediction

Affordance defines the action possibilities an object offers to an agent (Gibson 2014). In robotics, this translates to identifying object regions suitable for specific interactions, such as grasping, pushing, or lifting. Predicting such affordances is therefore critical for advanced visual understanding and robotic manipulation, as evidenced by extensive research (Li et al. 2019; Cao et al. 2020; Corona et al. 2020; Jiang et al. 2021; Lu et al. 2025; Shao et al. 2025). While pioneering works like GanHand (Corona et al. 2020) introduced generative models for multi-object on-table grasps, and GEAL (Lu et al. 2025) pioneered a dual-branch architecture for cross-modal (3D point cloud to 2D) representation learning, their learned affordances are often task- or category-specific. This inherent specialization limits their ability to generalize to novel objects or adapt to different downstream manipulation requirements. By contrast, humans exhibit exceptional proficiency in inferring universal affordances from visual cues (Zhao et al. 2025a). Inspired by this capability, AffordDex learns to infer functional affordances directly from multi-view rendered images of 3D objects, enabling a generalizable grasping policy that is not constrained to specific object categories or predefined tasks.

3 Methodology

To generate grasps with affordance-aware positioning and human-like kinematics, crucial for facilitating downstream manipulation, we propose a novel two-stage framework. The first stage establishes a strong human motion prior by pre-training a base policy π^H, on a large-scale human motion dataset (Zhan et al. 2024) via imitation learning. This con-

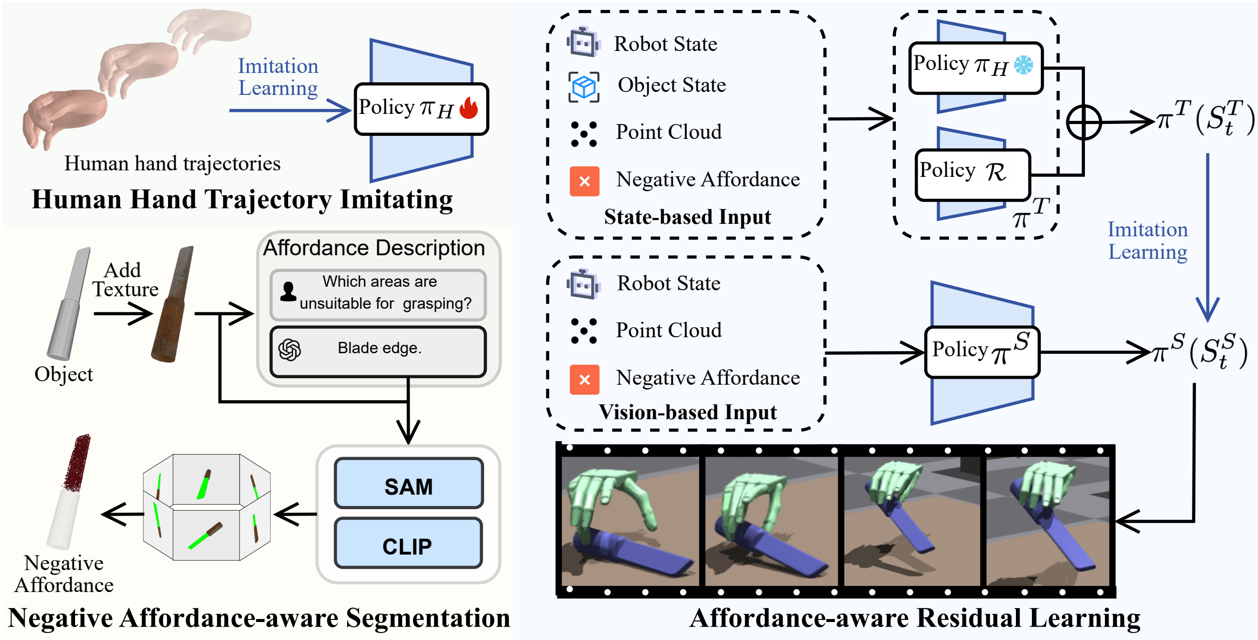

Figure 2: **Pipeline of AffordDex**. To generate grasps with affordance-aware positioning and human-like kinematics, crucial for facilitating downstream manipulation, we propose a novel two-stage framework. The first stage establishes a strong human motion prior by training a base policy π^H, on a human motion dataset via imitation learning. This constrains the policy to a space of natural, human-like movements. Subsequently, the second stage employs reinforcement learning (RL) to refine this coarse policy π^H for precise, functional interaction. We fine-tune π^H with a residual module that is guided by our Negative Affordance-aware Segmentation (NAA) module, which provides explicit constraints on where not to touch the object. The entire learning pipeline is further enhanced by a teacher-student distillation framework, leveraging privileged inputs to significantly boost the final grasping performance.

strains the policy to a manifold of natural, human-like movements. In the second stage, we freeze the weights of π^H and train a lightweight residual module via reinforcement learning (RL) to adapt these general motions to specific object interactions. This RL refinement stage is critically guided by two components: our Negative Affordance-aware Segmentation (NAA) module, which provides explicit constraints on where not to touch an object, and a teacher-student distillation framework that leverages privileged state information to significantly boost the final policy's performance. An overview of our method is illustrated in Fig. 2.

3.1 Human Hand Trajectory Imitating

In this stage, our objective is to learn a base policy π^H, that captures the kinematic priors of natural human hand motions. We formulate this task as a reinforcement learning (RL) problem where the policy $\pi^H(a_t|S_t^H)$ learns to generate dexterous hand action a_t based on the current state S_t^H at time t. To facilitate the following fine-tuning stage, the state consists of robot state R_t, object state O_t, and point cloud representation of object P_t, i.e., $S_t^H = \{R_t, O_t, P_t\}$.

Reward function. We design a reward function r^H to promote both precise imitation of human hand trajectories and the motion stability. It is composed of two terms: a finger imitation reward r^H_{finger} and a smoothness reward r^H_{smooth}.

The finger imitation reward r^H_{finger} encourages the dexterous hand to closely track the reference finger poses from human hand dataset. Following (Li et al. 2025b), we define this reward based on the distance between the corresponding keypoints F on the robot dexterous hand and the MANO hand. The reward at time t is formulated as:

$$r^H_{\text{finger}} = \sum_{f=1}^{F} w_f \cdot \exp\left(-\lambda_f \left\|\mathbf{j}_{d,f} - \mathbf{j}_{h,f}\right\|_2^2\right), \quad (1)$$

where $\mathbf{j}_{d,f}$ is the position of the f-th keypoint on the dexterous hand, $\mathbf{j}_{h,f}$ is its corresponding target position from the reference trajectory, w_f is weight and λ_f is the decay rate.

The smoothness reward r^H_{smooth} encourages energy-efficient movements by penalizing excessive power consumption. This is computed as the element-wise product of joint velocities and applied torques. A detailed formulation of our reward function is available in Supp. Mat.

3.2 Negative Affordance-aware Segmentation

A significant limitation of prior work in grasp synthesis (Xu et al. 2023; Wan et al. 2023; Zhong et al. 2025), is its neglect of the semantic and functional context of the interaction. A classic example is a knife: while its blade is geometrically stable for grasping, any grasp is functionally incorrect and unsafe. To address limitation, we introduce the Negative

Affordance-aware Segmentation (NAA) module to incorporate negative affordances—reasoning about which parts of an object should not be touched. The proposed NAA has the ability to operate in an open-vocabulary manner by harnessing the rich world knowledge embedded in Vision-Language Models (VLMs) (Radford et al. 2021; Achiam et al. 2023). This ensures that the generated grasps are not only geometrically stable but also semantically coherent and task-aware.

VLMs struggle to interpret non-textured 3D meshes, as these models primarily rely on rich visual cues learned from images. To address this, we first apply procedural texturing to raw meshes by (Zhang et al. 2024c), which generates semantically plausible textures based on geometric analysis. Next, we render the textured object from six cardinal directions to create a multi-view image set I as a holistic visual representation. While this may not capture all concavities in highly complex objects, we found it provides a sufficient basis for affordance prediction, representing a practical trade-off between coverage and computational cost. We then query GPT-4V (Achiam et al. 2023) to elicit a detailed description of the object's negative affordances.

VLMs (Radford et al. 2021) and Multimodal Large Language Models (MLLMs) (Achiam et al. 2023) excel at image-level understanding but struggle with the fine-grained spatial localization required for segmentation. To solve this, instead of asking CLIP (Radford et al. 2021) to find "blade part" from the image, we turn the segmentation task into a much simpler classification task. We generate a set of precise object-part masks M_i and use them as a visual prompt to let CLIP identify which mask in M_i has the highest semantic similarity to the textual description "*blade part*".

Specifically, for each image $I_i \in I$, we prompt Segment Anything Model (SAM) (Kirillov et al. 2023) with a dense grid of points G overlaid on I_i, which prompts SAM to perform an exhaustive segmentation, identifying all potential objects and parts. The resulting collection of masks is then refined using Non-Maximum Suppression (NMS) to eliminate duplicates, yielding a clean candidate masks set M_i:

$$M_i = \text{NMS}(\text{SAM}(I_i, G_i)). \tag{2}$$

For each mask $M_i^j \in M_i$, we generate a visually prompted image I_i^j by blurring regions outside the mask with Gaussian filter following (Yang et al. 2023). The prompted image set $\{I_i^j\}$ is then passed to CLIP along with the text query to compute a similarity score for each image-text pair. The mask with the highest similarity score is selected as the final segmentation mask. The mask is then projected into 3D space to segment the corresponding regions of the object's point cloud to get the negative affordance N_t, as shown in Fig. 3. Our NAA is an offline, one-time process. About 160 seconds per object on an RTX 4090 is a reasonable one-time trade-off.

3.3 Affordance-aware Residual Learning

The negative affordance predicted from proposed NAA, we use a residual module $\mathcal{R}$ to refine the pre-trained policy π^H. Since visual pose estimation is inherently less precise than using privileged state information, directly training an effective vision-based policy can be challenging. Therefore, we

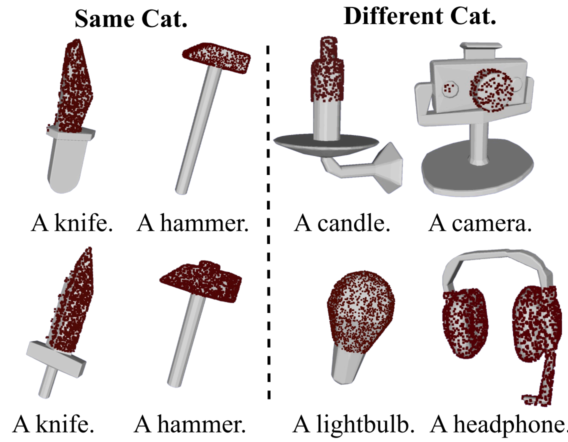

Figure 3: **Visualization** of Negative Affordances Predicted by our NAA. The point cloud, highlighted in red, represents the negative affordances identified on various objects. These points denote regions that are functionally unsafe or inappropriate for grasping, such as a knife's blade.

first train a state-based teacher policy π^T which can access the ground-truth states of the environment, such as object states, to learn residual actions to refine the initial actions predicted by π^H. Once the teacher policy π^T finishes training, we use an imitation learning algorithm, DAgger (Ross, Gordon, and Bagnell 2011), to distill π^T to a vision-based student policy π^S that can access oracle information and let policy help and ease the vision-based policy learning.

State-based teacher policy. In this stage, the inputs are robot state R_t, object state O_t, the scene point clouds P_t, and predicted negative affordance N_t. Here the scene point cloud is fused by multi-view depth cameras. Our goal is to learn residual actions $\Delta a_t = \pi^T(S_t^T)$ with predicted negative affordance by PPO (Schulman et al. 2017). The final action is computed with an element-wise addition:

$$a_t = \pi^H(S_t^T) + \pi^T(S_t^T). \tag{3}$$

Reward function. The reward function r^T is defined as:

$$r^T = -r_d^T - r_g^T + r_s^T - r_n^T \tag{4}$$

where the grasp reward r_d^T penalizes the distance between the dexterous hand and the object, encouraging the hand to maintain contact with the object surface for a secure grasp. The goal reward r_g^T penalizes the distance between the object and the target goal, and the success reward r_s^T provides a bonus when the object successfully reaches the goal. Also the negative affordance reward r_n^T penalizes the dexterous hand to approach the predicted negative affordance. The formal definitions of all rewards are available in our Supp. Mat.

Vision-based student policy. For vision-based policy, we only allow it to access information available in the real world, including robot state R_t, the scene points clouds P_t, and predicted negative affordance N_t. Then, we distill the

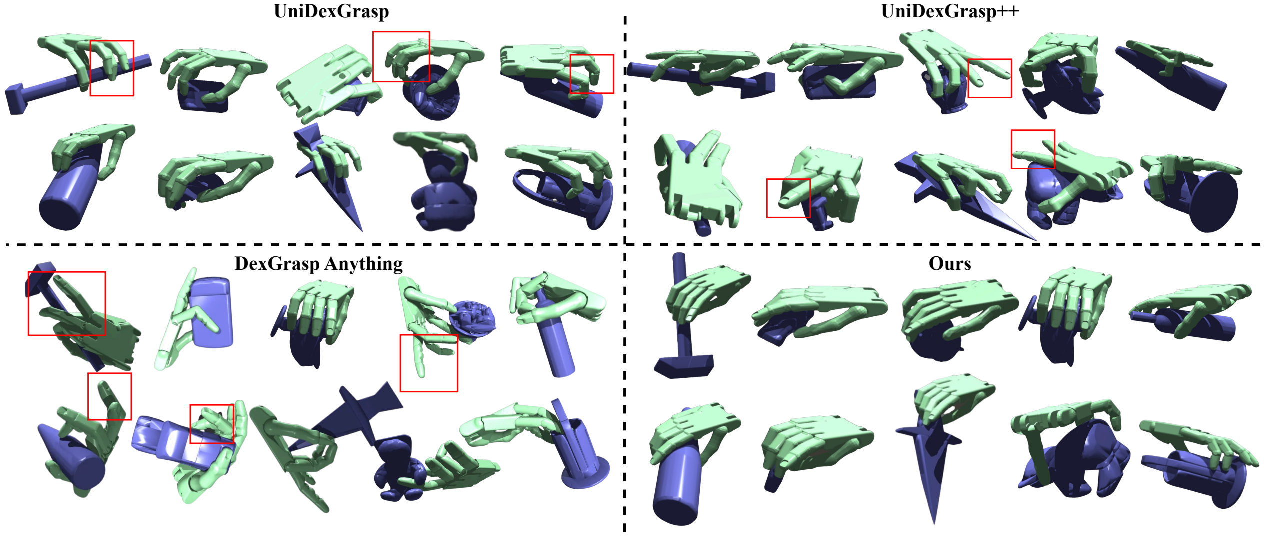

Figure 4: **Qualitative Comparison** on UniDexGrasp (Xu et al. 2023) and OakInk2 (Zhan et al. 2024). A comparison of grasps generated by our AffordDex with several baselines, including UniDexGrasp (Xu et al. 2023), UniDexGrasp++ (Wan et al. 2023), and DexGrasp Anything (Zhong et al. 2025).

teacher policy π^T into vision-based student policy π^S using DAgger (Ross, Gordon, and Bagnell 2011), i.e.,

$$\pi^S = \arg\min_{\pi^S} \|\pi^T(S_t^T) - \pi^S(S_t^S)\|, \tag{5}$$

where the state for the teacher policy $S_t^T = \{R_t, O_t, P_t, N_t\}$, and the state for the student policy $S_t^S = \{R_t, P_t, N_t\}$.

4 Experiments

4.1 Datasets

UniDexGrasp (Xu et al. 2023). This dataset contains 3165 different object instances spanning 133 categories. Evaluation is conducted on these 3,200 seen objects, as well as on 140 unseen objects from seen categories and 100 unseen objects from unseen categories. Each environment is randomly initialized with one object and its initial pose, and the environment consists of a panoramic 3D point cloud P_t captured from the fixed cameras for vision-based policy learning.

OakInk2 (Zhan et al. 2024). This dataset record the manipularion processes with pose and shape of the human upper-body and objects. We pre-train our π^H using about 2,200 right hand manipulation sequences in this dataset. Also we employ objects in OakInk2 to evaluate the generalization capabilities for grasping.

4.2 Metrics

Following previous works (Xu et al. 2023; Wan et al. 2023; Wang et al. 2025), each object is randomly rotated and dropped onto the table to enhance the diversity of its initial poses. We report the **success rate of grasp** $Succ$, **Human-likeness Score** HLS, and **Affordance Score** AS across all objects and grasp attempts. A grasp is considered successful if the object reaches the target goal within 200 steps in simulator. The Human-likeness Score HLS assesses the anthropomorphic quality of the grasp, which is obtained by prompting the Gemini 2.5 Pro (Comanici et al. 2025) to analyze a visual sequence of the grasp execution. This metric is specifically to rate the resemblance of the dexterous hand's motion to that of a typical human, yielding a quantitative measure of naturalness. The Affordance Score AS, in contrast, evaluates the functional correctness of the grasp by penalizing contact with inappropriate object parts. This metric is calculated using a point cloud of 100 "negative affordance" points sampled from our NAA. Specifically, the score is incremented by one for each fingertip that maintains a distance greater than 2cm from any point in this negative set, thus rewarding functionally sound grasps.

4.3 Implementation Details

We conduct our experiments in IssacGym (Makoviychuk et al. 2021) simulator. During training, 4096 environments are simulated in parallel on an NVIDIA RTX 4090 GPU. For the network architecture, we use MLP with 4 hidden layers (1024,1024,512,512) for the policy network and value network in the state-based setting, and an additional PointNet+Transformer (Mu et al. 2021) to encode the 3D scene point cloud input in the vision-based setting. Other detailed hyperparameters are shown in our Supp. Mat.

Dexterous hand configuration. We use the Shadow Hand, which features 24 active degrees of freedom (DOFs). The wrist has 6 DOFs controlled by force and torque, while the fingers have 18 active DOFs controlled by joint angles. Specifically, the thumb has 5 DOFs, the little finger has 4, and the remaining three fingers each have 3. Additionally,

Method	Seen Obj.			Unseen Obj. Seen Cat.			Unseen Obj. Unseen Cat.			OakInk2		
	$Succ\uparrow$	$HLS\uparrow$	$AS\downarrow$	$Succ\uparrow$	$HLS\uparrow$	$AS\downarrow$	$Succ\uparrow$	$HLS\uparrow$	$AS\downarrow$	$Succ\uparrow$	$HLS\uparrow$	$AS\downarrow$
State-Based Setting												
PPO (Schulman et al. 2017)	24.3	-	-	20.9	-	-	17.2	-	-	-	-	-
DAPG (Rajeswaran et al. 2017)	20.8	-	-	15.3	-	-	11.1	-	-	-	-	-
GSL (Jia et al. 2022)	57.3	-	-	54.1	-	-	50.9	-	-	-	-	-
ILAD (Wu, Wang, and Wang 2023)	31.9	-	-	26.4	-	-	23.1	-	-	-	-	-
UniDexGrasp (Xu et al. 2023)	79.4	6.9	12	74.3	6.4	15	70.8	6.3	18	68.4	5.9	18
UniDexGrasp++ (Wan et al. 2023)	87.9	5.4	28	84.3	5.2	26	83.1	5.0	27	79.6	4.9	28
DexGrasp Anything (Zhong et al. 2025)	71.2	-	20	69.1	-	18	67.3	-	22	65.9	-	24
AffordDex	**89.2**	**8.6**	**4**	**87.7**	**8.5**	**7**	**85.2**	**8.1**	**9**	**82.2**	**8.2**	**10**
Vision-Based Setting												
PPO (Schulman et al. 2017)	20.6	-	-	17.2	-	-	15.0	-	-	-	-	-
DAPG (Rajeswaran et al. 2017)	20.8	-	-	15.3	-	-	11.1	-	-	-	-	-
GSL (Jia et al. 2022)	54.1	-	-	50.2	-	-	44.8	-	-	-	-	-
ILAD (Wu, Wang, and Wang 2023)	27.6	-	-	23.2	-	-	20.0	-	-	-	-	-
UniDexGrasp (Xu et al. 2023)	73.7	6.2	16	68.6	6.1	18	65.1	6.0	17	62.8	5.6	20
UniDexGrasp++ (Wan et al. 2023)	85.4	5.4	29	79.6	5.1	25	76.7	4.8	28	74.4	4.7	29
AffordDex	**87.0**	**8.3**	**10**	**82.8**	**7.8**	**14**	**79.2**	**8.0**	**15**	**77.3**	**7.8**	**13**

Table 1: **Quantitative comparisons** on UniDexGrasp (Xu et al. 2023) and OakInk2 (Zhan et al. 2024). HLS denotes Human-likeness Score, while AS is Affordance Score.

HTI	NAA	Distillation	$Succ\uparrow$	$HLS\uparrow$	$AS\downarrow$
State-Based Setting					
			85.4	5.2	27
✓			87.9	8.2	22
✓	✓		**89.2**	**8.6**	**4**
Vision-Based Setting					
			70.1	5.0	27
		✓	84.9	5.6	28
	✓	✓	85.8	7.2	13
✓		✓	86.9	8.1	20
✓	✓	✓	**87.0**	**8.3**	**10**

Table 2: **Ablation Study** on UniDexGrasp (Xu et al. 2023) in Seen Object.

Method	$Succ\uparrow$	$HLS\uparrow$	$AS\downarrow$
UniDexGrasp++ (Wan et al. 2023)	87.9	5.4	28
UniDexGrasp++ + HTI	88.2	7.8	23
UniDexGrasp++ + NAA	88.0	5.9	19
UniDexGrasp++ + HTI + NAA	**88.8**	**8.0**	**12**

Table 3: Results on UniDexGrasp (Xu et al. 2023) in Seen Object in state-based setting.

each finger, excluding the thumb, includes a passive, non-controlled DOF.

4.4 Comparison with SOTA Methods

We evaluate AffordDex by first training a state-based policy and then distilling it into a vision-based one. For comparison, we compare AffordDex with several state-of-the-art methods. These include RL and imitation learning algorithms like PPO (Schulman et al. 2017), DAPG (Rajeswaran et al. 2017), and ILAD (Wu, Wang, and Wang 2023). We also compare against methods with advanced learning paradigms such as GSL (Jia et al. 2022) (generalist-specialist), UniDexGrasp++ (Wan et al. 2023) (geometry-aware curriculum learning), UniDexGrasp++ (Wan et al. 2023) which proposes a geometry-aware curriculumn learning and generalist-specialist learning, and DexGrasp Anything (Zhong et al. 2025), diffusion-based dexterous grasp generation models.

Since DexGrasp Anything (Zhong et al. 2025) generates only the final static grasp pose rather than a grasping motion, the HLS is not applicable in this setting. To evaluate grasp robustness, we apply a random external force ranging from 0 to 200N to the object to simulate object's gravity.

Tab. 1 compares our AffordDex with these SOTA methods using a universal model for dexterous robotic grasping across both state-based, vision-based and even across dataset settings. Our method achieves highest scores in grasping success rate, outperforming other state-of-the-art methods. This improvement stems from our proposed Human Hand Trajectory Imitating, where the policy learns to generate correct and stable grasp poses from human hand motion. This not only greatly enhances the grasp success rate but also leads to a significant improvement in our method's human-likeness score (HLS). The significantly lower Affordance Score AS validates the effectiveness of our Negative Affordance-aware Segmentation module. This result indicates that the module successfully guides the policy away from functionally inappropriate regions, leading to grasps at the most suitable locations.

As illustrated in Fig. 4, our method generates a diverse set of grasps. Crucially, it consistently identifies functionally appropriate grasp locations and forms natural hand postures. This combination of functional awareness and naturalness makes the generated poses highly effective for direct application in downstream manipulation tasks.

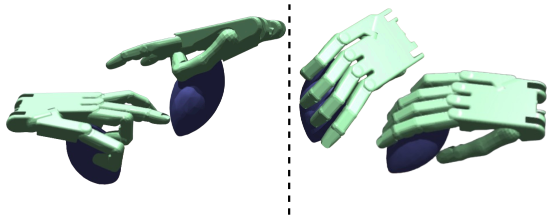

Figure 5: **Ablation Study** on Human Hand Trajectory Imitating (HTI). Without the human motion prior, the policy converges to a solution that, while potentially successful, is kinematically awkward and non-humanlike.

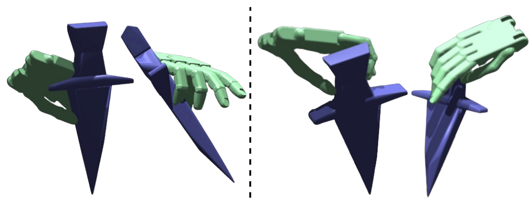

Figure 6: **Ablation Study** on proposed NAA, which guides the policy to a correct and safte position. The higher Affordance Score (AS) for the NAA-guided grasp confirms its superior functional quality.

4.5 Ablation study

Unless otherwise specified, the ablation studies are conducted on the seen objects under the state-based setting.

Human Hand Trajectory Imitating. The results in Tab. 2 and Fig. 5 show the critical role of pre-training on human trajectories (HTI). When this imitation stage is omitted, the policy, while still capable of finding geometrically stable grasps, produces motions that are kinematically unnatural. This is quantitatively reflected in a sharp increase in the Human-Likeness Score (HLS). Such configurations are not merely an aesthetic issue, they can be inefficient, unpredictable, and detrimental to downstream tasks that require fluid, human-centric interaction.

NAA. As shown in Fig. 7, a naive approach combining an MLLM (Achiam et al. 2023) with SAM (Kirillov et al. 2023) (denoted as GPT+SAM) proves ineffective for this task. This baseline first uses the MLLM's coarse localization ability to provide prompts to SAM. However, because MLLMs like GPT-4V (Achiam et al. 2023) excel at image-level understanding but struggle with the fine-grained spatial localization required for segmentation, this process often results in the segmentation of the entire object. In contrast, our NAA module solves this by converting the segmentation task into a simpler classification problem. By first using SAM to generate accurate mask proposals and then using CLIP to select the one with the highest semantic similarity to the negative affordance description, NAA achieves precise segmentation.

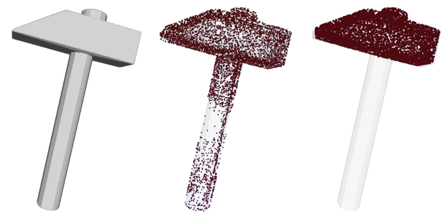

Figure 7: **Ablation Study** on proposed NAA, which has capability to segment fine-grained negative affordance.

As shown in Fig. 6 and Tab. 2, the guidance from NAA results in a significant decrease in the Affordance Score AS, which indicates that the policy successfully learns to make contact at more rational and safer locations on the object. By generating functionally sound grasps, AffordDex greatly improves the feasibility of performing downstream tasks.

Teacher-student distillation. Without the teacher-student distillion (Distillation) grasping accuracy decreases significantly. This is primarily due to the lack of privileged information guidance, which makes it challenging for the single-stage RL policy to learn the position to grasp. As shown in Tab. 2, the policy without teacher-student distillation demonstrates lower grasping success rate.

4.6 Extension to Other Grasping Methods

Notably, our proposed modules demonstrate strong generalizability by significantly enhancing other RL-based methods, such as UniDexGrasp++ (Wan et al. 2023). Specifically, the Human Hand Trajectory Imitating (HTI) module markedly improves the naturalness and human-likeness of its generated poses. Simultaneously, the Affordance-aware Residual Learning, guided by negative affordances from our Negative Affordance-aware Segmentation (NAA), substantially boosts the semantic appropriateness of its grasp locations on the object, as shown in Tab. 3.

5 Conclusion

In this paper, we present AffordDex, a novel framework for generating dexterous grasps that are not only successful but also human-like and functionally correct. Our key insight is that the challenges of naturalness and functional correctness can be effectively decoupled and then synergized: a strong motion prior learned from human data constrains the policy to a manifold of natural poses, while a visual understanding of negative affordances guides the policy to safe and appropriate contact regions. Extensive experiments validate that this approach significantly outperforms state-of-the-art baselines in success rate, pose naturalness, and contact appropriateness. We believe this work lays a crucial foundation for more general-purpose embodied agents and opens new avenues for research in dexterous manipulation.

Acknowledgements

This work was supported by Central Guidance for Local Science and Technology Development Fund (ZYYD2025QY19), an Bingtuan Science and Technology Program (2022DB005). This work was supported by DAMO Academy through Academy Research Intern Program.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Andrews, S.; and Kry, P. G. 2013. Goal directed multi-finger manipulation: Control policies and analysis. *Computers & Graphics*, 37(7): 830–839.

Bai, Y.; and Liu, C. K. 2014. Dexterous manipulation using both palm and fingers. In *International Conference on Robotics and Automation*, 1560–1565. IEEE.

Cao, Z.; Gao, H.; Mangalam, K.; Cai, Q.-Z.; Vo, M.; and Malik, J. 2020. Long-term human motion prediction with scene context. In *Proc. of European Conf. on Computer Vision*, 387–404. Springer.

Christen, S.; Kocabas, M.; Aksan, E.; Hwangbo, J.; Song, J.; and Hilliges, O. 2022. D-Grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 20577–20586.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Corona, E.; Pumarola, A.; Alenya, G.; Moreno-Noguer, F.; and Rogez, G. 2020. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 5031–5041.

Fang, H.; Fang, H.-S.; Xu, S.; and Lu, C. 2022. Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. *Robotics and Automation Letters*, 7(3): 7383–7390.

Fang, H.-S.; Wang, C.; Gou, M.; and Lu, C. 2020. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 11444–11453.

Gibson, J. J. 2014. The theory of affordances:(1979). In *The people, place, and space reader*, 56–60. Routledge.

Gou, M.; Fang, H.-S.; Zhu, Z.; Xu, S.; Wang, C.; and Lu, C. 2021. RGB matters: Learning 7-dof grasp poses on monocular rgbd images. In *International Conference on Robotics and Automation*, 13459–13466. IEEE.

Huang, S.; Wang, Z.; Li, P.; Jia, B.; Liu, T.; Zhu, Y.; Liang, W.; and Zhu, S.-C. 2023. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 16750–16761.

Jia, Z.; Li, X.; Ling, Z.; Liu, S.; Wu, Y.; and Su, H. 2022. Improving policy optimization with generalist-specialist learning. In *Proc. of Intl. Conf. on Machine Learning*, 10104–10119. PMLR.

Jiang, H.; Liu, S.; Wang, J.; and Wang, X. 2021. Hand-object contact consistency reasoning for human grasps generation. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 11107–11116.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 4015–4026.

Li, H.; Zhao, Q.; Xu, H.; Jiang, X.; Ben, Q.; Jia, F.; Zhao, H.; Xu, L.; Zeng, J.; Wang, H.; et al. 2025a. TeleOpBench: A Simulator-Centric Benchmark for Dual-Arm Dexterous Teleoperation. *arXiv preprint arXiv:2505.12748*.

Li, K.; Li, P.; Liu, T.; Li, Y.; and Huang, S. 2025b. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 6991–7003.

Li, X.; Liu, S.; Kim, K.; Wang, X.; Yang, M.-H.; and Kautz, J. 2019. Putting humans in a scene: Learning affordance in 3d indoor environments. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 12368–12376.

Liu, T.; Liu, Z.; Jiao, Z.; Zhu, Y.; and Zhu, S.-C. 2021. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *Robotics and Automation Letters*, 7(1): 470–477.

Liu, Y.; Yang, Y.; Wang, Y.; Wu, X.; Wang, J.; Yao, Y.; Schwertfeger, S.; Yang, S.; Wang, W.; Yu, J.; et al. 2024. RealDex: Towards human-like grasping for robotic dexterous hand. *arXiv preprint arXiv:2402.13853*.

Lu, D.; Kong, L.; Huang, T.; and Lee, G. H. 2025. Geal: Generalizable 3d affordance learning with cross-modal consistency. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 1680–1690.

Lu, J.; Kang, H.; Li, H.; Liu, B.; Yang, Y.; Huang, Q.; and Hua, G. 2024. UGG: Unified generative grasping. In *Proc. of European Conf. on Computer Vision*, 414–433. Springer.

Mahler, J.; Matl, M.; Satish, V.; Danielczuk, M.; DeRose, B.; McKinley, S.; and Goldberg, K. 2019. Learning ambidextrous robot grasping policies. *Science Robotics*, 4(26): eaau4984.

Makoviychuk, V.; Wawrzyniak, L.; Guo, Y.; Lu, M.; Storey, K.; Macklin, M.; Hoeller, D.; Rudin, N.; Allshire, A.; Handa, A.; et al. 2021. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*.

Mandikal, P.; and Grauman, K. 2021. Learning dexterous grasping with object-centric visual affordances. In *international conference on robotics and automation*, 6169–6176. IEEE.

Mandikal, P.; and Grauman, K. 2022. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, 651–661. PMLR.

Mu, T.; Ling, Z.; Xiang, F.; Yang, D.; Li, X.; Tao, S.; Huang, Z.; Jia, Z.; and Su, H. 2021. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. *arXiv preprint arXiv:2107.14483*.

Nagabandi, A.; Konolige, K.; Levine, S.; and Kumar, V. 2020. Deep dynamics models for learning dexterous manipulation. In *Conference on Robot Learning*, 1101–1112. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of Intl. Conf. on Machine Learning*, 8748–8763.

Rajeswaran, A.; Kumar, V.; Gupta, A.; Vezzani, G.; Schulman, J.; Todorov, E.; and Levine, S. 2017. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*.

Ross, S.; Gordon, G.; and Bagnell, D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, 627–635.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Y.; Zhai, W.; Yang, Y.; Luo, H.; Cao, Y.; and Zha, Z.-J. 2025. GREAT: Geometry-Intention Collaborative Inference for Open-Vocabulary 3D Object Affordance Grounding. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 17326–17336.

Wan, W.; Geng, H.; Liu, Y.; Shan, Z.; Yang, Y.; Yi, L.; and Wang, H. 2023. UniDexGrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 3891–3902.

Wang, C.; Fang, H.-S.; Gou, M.; Fang, H.; Gao, J.; and Lu, C. 2021. Graspness discovery in clutters for fast and accurate grasp detection. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 15964–15973.

Wang, W.; Wei, F.; Zhou, L.; Chen, X.; Luo, L.; Yi, X.; Zhang, Y.; Liang, Y.; Xu, C.; Lu, Y.; et al. 2025. UniGrasp-Transformer: Simplified policy distillation for scalable dexterous robotic grasping. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 12199–12208.

Wu, Y.-H.; Wang, J.; and Wang, X. 2023. Learning generalizable dexterous manipulation from human grasp affordance. In *Conference on Robot Learning*, 618–629. PMLR.

Xu, Y.; Wan, W.; Zhang, J.; Liu, H.; Shan, Z.; Shen, H.; Wang, R.; Geng, H.; Weng, Y.; Chen, J.; et al. 2023. UniDexGrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 4737–4746.

Yang, L.; Wang, Y.; Li, X.; Wang, X.; and Yang, J. 2023. Fine-grained visual prompting. *Proc. of Advances in Neural Information Processing Systems*, 36: 24993–25006.

Zhan, X.; Yang, L.; Zhao, Y.; Mao, K.; Xu, H.; Lin, Z.; Li, K.; and Lu, C. 2024. OakInk2: A Dataset of Bimanual Hands-Object Manipulation in Complex Task Completion. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 445–456.

Zhang, H.; Christen, S.; Fan, Z.; Hilliges, O.; and Song, J. 2024a. Graspxl: Generating grasping motions for diverse objects at scale. In *Proc. of European Conf. on Computer Vision*, 386–403. Springer.

Zhang, H.; Christen, S.; Fan, Z.; Zheng, L.; Hwangbo, J.; Song, J.; and Hilliges, O. 2024b. ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *International Conference on 3D Vision*, 235–246. IEEE.

Zhang, H.; Pan, Z.; Zhang, C.; Zhu, L.; and Gao, X. 2024c. TexPainter: Generative Mesh Texturing with Multi-view Consistency. In *ACM SIGGRAPH Conference*, 1–11.

Zhao, H.; Wang, H.; Zhao, X.; Fei, H.; Wang, H.; Long, C.; and Zou, H. 2025a. PhysSplat: Efficient Physics Simulation for 3D Scenes via MLLM-Guided Gaussian Splatting. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 5242–5252.

Zhao, H.; Yang, C.; Wang, H.; Zhao, X.; and Shen, W. 2024. SG-GS: Photo-realistic animatable human avatars with semantically-guided gaussian splatting. *arXiv preprint arXiv:2408.09665*.

Zhao, H.; Zeng, C.; Zhuang, L.; Zhao, Y.; Xue, S.; Wang, H.; Zhao, X.; Li, Z.; Li, K.; Huang, S.; et al. 2025b. High-Fidelity Simulated Data Generation for Real-World Zero-Shot Robotic Manipulation Learning with Gaussian Splatting. *arXiv preprint arXiv:2510.10637*.

Zhong, Y.; Jiang, Q.; Yu, J.; and Ma, Y. 2025. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 22584–22594.