

OmniDexGrasp: Generalizable Dexterous Grasping via Foundation Model and Force Feedback

Yi-Lin Wei*, Zhexi Luo*, Yuhao Lin, Mu Lin, Zhizhao Liang, Shuoyu Chen, Wei-Shi Zheng†

School of Computer Science and Engineering, Sun Yat-sen University, China

{weiylin5, luozhx25, linyh96, linn67, liangzhzh26, chenshy357}@mail2.sysu.edu.cn wszheng@ieee.org

<https://isee-laboratory.github.io/OmniDexGrasp/>

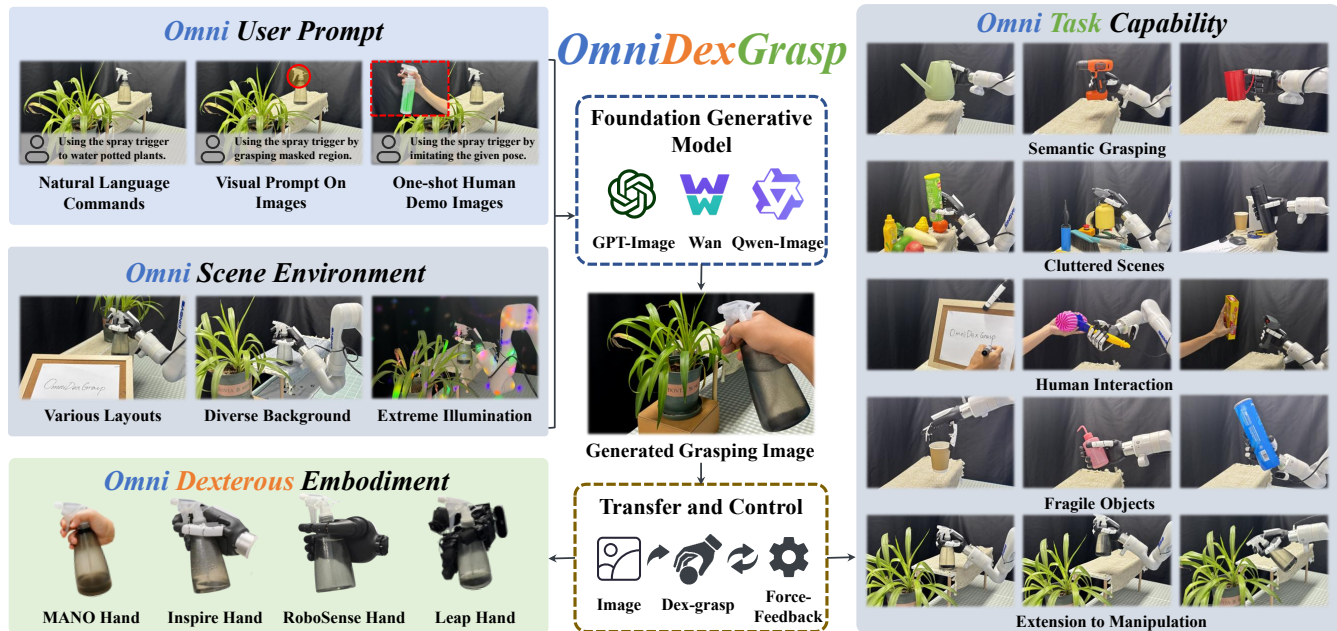


Fig. 1: OmniDexGrasp can achieve generalizable dexterous grasping with omni capabilities in user prompting, dexterous embodiment, scenes, and grasping tasks, by leveraging the foundation model and the propose transfer and grasp strategy.

Abstract—Enabling robots to dexterously grasp and manipulate objects based on human commands is a promising direction in robotics. However, existing approaches are challenging to generalize across diverse objects or tasks due to the limited scale of semantic dexterous grasp datasets. Foundation models offer a new way to enhance generalization, yet directly leveraging them to generate feasible robotic actions remains challenging due to the gap between abstract model knowledge and physical robot execution. To address these challenges, we propose OmniDexGrasp, a generalizable framework that achieves omni-capabilities in user prompting, dexterous embodiment, and grasping tasks by combining foundation models with the transfer and control strategies. OmniDexGrasp integrates three key modules: (i) foundation models are used to enhance generalization by generating human grasp images supporting omni-capability of user prompt and task; (ii) a human-image-to-robot-action transfer strategy converts human demonstrations into executable robot actions, enabling omni dexterous embodiment; (iii) force-aware adaptive

grasp strategy ensures robust and stable grasp execution. Experiments in simulation and on real robots validate the effectiveness of OmniDexGrasp on diverse user prompts, grasp task and dexterous hands, and further results show its extensibility to dexterous manipulation tasks.

I. INTRODUCTION

Robotic dexterous grasping stands as a central and highly challenging problem in robotics [1]. The long-term objective is to establish a generalizable and unified framework that enables different robots to understand and execute human commands across a wide range of objects, tasks, and environments [2], [3], [4], [5], [6], [7]. Achieving such capabilities would significantly enhance the practical utility of robots in industrial, domestic, and real-world applications.

With the development of deep learning, data-driven methods show promising performance on diverse objects in dexterous grasp [8], [9], [10]. These approaches based on conditional generative models leverage visual observation and user commands as input to generate grasping actions [11], [12], trained on datasets containing aligned observation-action pairs [13], [14]. However, the generalization and omni-capabilities of these approaches remain challenging by

*Equal contribution.

†Corresponding author.

the scale and diversity of available datasets [15], [16], making it challenging to reliably handle novel object categories, diverse user commands, and dexterous embodiments.

Recent advances in foundation generative models offer a promising pathway to overcome the generalization challenge [17], [18]. Trained on large-scale datasets, these models can learn the knowledge relevant to grasping and manipulation [19], [3]. However, their application to robotics remains challenging due to the gap between the foundation models' high-level knowledge and the low-level physical constraints of robotic execution [20]. As a result, it is challenging for foundation models to directly generate executable robot actions, and the outputs may lead to physically infeasible interactions due to model hallucinations [21].

To address these challenges, we propose OmniDexGrasp, a framework for generalizable dexterous grasping with omni-capabilities in user prompting, grasping tasks, and dexterous embodiment. The motivation of OmniDexGrasp is to exploit the generalization and multi-modal adaptability of foundation generative models by leveraging diverse user prompts to generate human grasp for various grasping tasks without requiring additional robot data. Subsequently, to address the limitation that foundation generative models cannot directly yield robot actions, we introduce a human-image-to-robot-action transfer strategy that covert human grasp images generated by foundation model into executable robot action. Finally, to address the absence of physical constraints in generative models, we develop a force-aware adaptive grasping module that ensures robust and stable execution with appropriate force by controlling the force applied by fingers.

Extensive experiments are conducted in both simulation and real-world environments to validate the effectiveness of the proposed OmniDexGrasp framework. In the real world, we evaluate six dexterous grasping tasks that involve varying objects, user prompts the task objectives. And we also conduct semantic grasping experiments in simulation across objects of different categories. Additionally, we conduct experiments to show that the framework can be extended to dexterous manipulation and flexibly adapted to different dexterous hands.

The main contributions are summarized as follows:

- 1) We propose the OmniDexGrasp framework, with the insight of leveraging foundation models, to achieve generalizable dexterous grasping with omni-capabilities across tasks, scenes, user prompts, and robot embodiments.
- 2) We introduce a human-image-to-robot-action transfer strategy that accurately converts visual contents generated by foundation models into executable dexterous robot actions.
- 3) We present a force-aware adaptive grasp strategy to achieve stable and reliable grasp execution by controlling the force applied by fingers.

II. RELATED WORK

A. Dexterous Grasp

Dexterous grasping allows robots to manipulate real-world objects, making it a key topic in robotics research. While early studies [8], [9] mainly focused on stability and quality,

recent research [11], [12], [22] emphasizes the semantics of grasping, leveraging human language as a condition to guide the task. However, as most of these methods are data-driven, the lack of large-scale and well-semantically-annotated datasets [13], [23] limits their generalization to diverse real-world objects. Recently, some works [15], [16] based on affordance or retrieval show certain levels of cross-category generalization, but their reliance on similar parts shared across categories makes them struggle to extend to truly novel objects with distinct structures in real-world scenarios. To tackle this challenge, we introduce the OmniDex-Grasp framework, which leverages pretrained foundation models to transfer their learned grasping and manipulation patterns to generalizable robot execution.

B. Foundation Models for Robotics

With the rapid progress of computer vision and multi-modal learning, foundation models, such as multi-modal large language models (MLLMs) [18], image generation models [17], and video generation models [24], which are trained on diverse multimodal datasets, have shown strong capabilities in capturing underlying patterns and generalizing to guide robotic execution [25], [26]. Recently, some works fine-tunes large language models (LLMs) to build vision-language-action (VLA) models by robot action data [13], [27], while others [28] directly train video generation models on robot video datasets. However, these approaches still depend on robot-specific fine-tuning data, inherently constraining their generalization ability. Alternatively, some works [19], [29] leverage human videos generated by pretrained foundation models for robotic manipulation with parallel grippers; yet, extending this paradigm to dexterous robotic hands poses greater challenges, especially as model hallucinations resulting in physically implausible interactions. To address these limitations, we propose a novel framework that combines the learning-free utilization of foundation models with a transfer and force-aware strategy, thereby achieving both strong generalization and high-quality dexterous execution.

III. OMNIDEXGRASP FRAMEWORK

A. Overview

Problem Formulation. Given the scene single-view RGB-D observation (\mathcal{O} for partial point cloud and \mathcal{I}^{obs} for RGB images) and command of user \mathcal{C} as input, our goal is to generate intention-aligned and high-quality dexterous actions $\mathcal{G}^{dex} = (T_{dex}, J_{dex})$, where T_{dex} denotes the 6D pose of arm end-effector, and J_{dex} represents the joint angles of the dexterous hand.

Framework Overview. The OmniDexGrasp framework consists of three main components: (1) human grasp image generation by foundation generative models (Sec. III-B), (2) a transfer strategy that maps human grasp images to executable dexterous robot actions (Sec. III-C), and (3) a force-aware adaptive grasping strategy to achieve stable grasp with appropriate force (Sec. III-D). Furthermore, we extend our framework to manipulation tasks. (Sec. III-E).

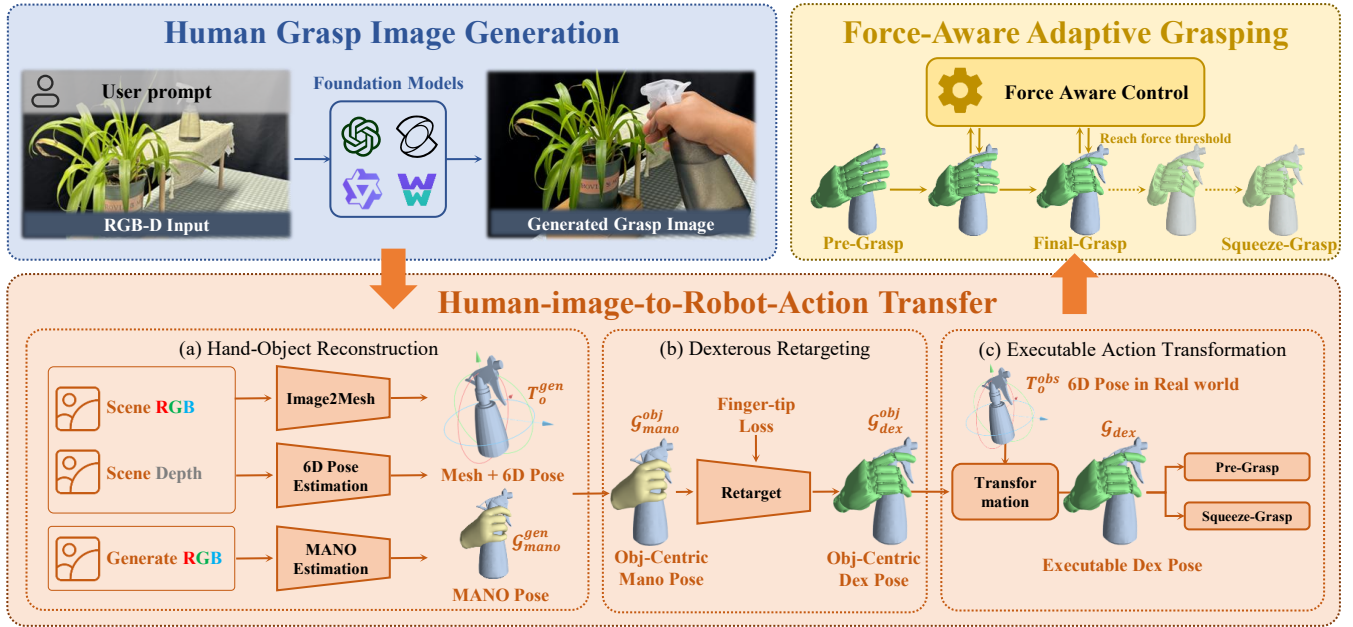


Fig. 2: Overview of OmniDexGrasp: (1) human grasp image generation via foundation generative models, (2) human-to-robot action transfer mapping grasp images to dexterous robot actions, and (3) force-aware adaptive grasping for stable grasps.

B. Grasp Image Generation by Foundation Models

To achieve generalizable grasping, we first employ generative models [18] to generate images of human grasping. The advantages of generating human grasps, rather than directly generating robot grasps, are twofold. First, these models are predominantly trained on datasets rich in human behaviors, which enables them to better exploit prior knowledge of the human dynamics principles. Second, human grasp serves as an effective intermediate representation to bridge diverse heterogeneous dexterous hands, thereby avoiding the need to train a separate model for each specific robot.

Specifically, we feed the RGB observation \mathcal{I}^{obs} and a user prompt into foundation generative models to produce human grasp images \mathcal{I}^{gen} . The prompt is omni-capable, supporting diverse forms such as a language instruction, a grasp point or region mask on the observation, and a demonstration image. To further enhance generation quality, we design a prompt template to generate detailed system prompts, including both positive and negative components. An example prompt of language command is shown below:

Prompt Example

Positive: Object: {name}. Intention: {intent}. Based on the input image and grasp intention, generate a image of a human right hand grasping the object. Camera fixed, hand enters from bottom-right, grasps the object, then stays still. Realistic style, uniform lighting, clear details.

Negative: Overly saturated colors, overexposed, blurry details, grayish tone, worst quality, low quality, artifacts, ugly, incomplete, extra fingers, poorly rendered hands, deformed, disfigured, malformed limbs, fused fingers

Our framework supports different types of foundation generative models, including both image generation and video

generation. For image generation, we directly synthesize grasp images and employ both closed-source models (GPT-Image [18]) and an open-source model (Qwen-Image [17]). For video generation, we extract final frame which is asked to keeping in the grasping action, as the grasp image, utilizing a closed-source model (Kling [30]) and open-source models (WanX [24]). We conduct a comparative study (Sec. IV-F) to evaluate the performance of different models. Unless otherwise specified, GPT-Image is adopted as the default image generator throughout experiments.

C. Human-image-to-Robot-action Transfer

Although generating human grasp images is effective, it also introduces the challenge of how to obtain robot dexterous actions, which are both consistent with the images and physically plausible. As shown in Figure 3, the object’s generated pose may be altered due to model hallucination. Directly executing the corresponding action often results in failure. To solve this challenge, we propose a human-image-to-robot-action transfer strategy, which inputs human grasp image and outputs executable dexterous action \mathcal{G}^{dex} . The pipeline consists of three components: hand-object reconstruction, dexterous retargeting, and executable action transformation.

(1) Hand-object Reconstruction. The aim of this component is to reconstruct the 3D representations of the hand (MANO [31] parameters $\mathcal{G}_{mano}^{obj} = (T_{mano}^{obj}, J_{mano})$: 6D pose T_{mano}^{obj} in object coordinate system and articulated pose J_{mano}) and the object (object mesh M_o and scale s_o) with physically plausible interactions. Specially, for the hand reconstruction, we use HaMeR [32] to obtain articulated pose J_{mano} and wrist 6D pose T_{mano}^{gen} in the camera coordinate system. For object reconstruction, we employ Hyper3D [33], an image-to-mesh 3D generation model, on the unobstructed

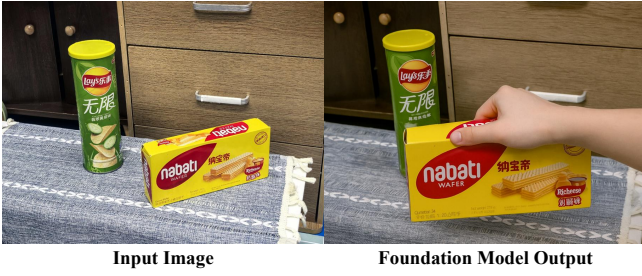


Fig. 3: Visualization of the inputs and outputs of foundation generative models, illustrating the potential change of object pose in the generated image.

original images \mathcal{I}^{obs} to obtain object mesh M_o . The object scale s_o is estimated following Any6D [34] by optimization. The object 6D poses T_o^{gen} are estimated via a MegaPose [35] in the generated image camera coordinate system.

We find that the separate estimation may introduce error in object-hand interaction, which is particularly along the camera’s depth axis. To address this, we optimize the t_z (the translation component of T_{mano}^{gen} along the depth axis) following EasyHOI [36] by aligning hand-object interaction consistency. Finally, we transform the optimized wrist pose into the object coordinate frame by $T_{mano}^{obj} = T_o^{gen-1} T_{mano}^{gen}$.

(2) Dexterous Retargeting. The aim of this component is to obtain dexterous hand poses by retargeting the reconstructed MANO human hand \mathcal{G}_{mano}^{obj} to the dexterous hand \mathcal{G}_{dex}^{obj} . To achieve this, we employ a two-step optimization that initializes and subsequently refines the dexterous poses. In the first step, the dexterous poses are initialized by copying the wrist’s 6-DoF parameters and those of structurally similar joints along the kinematic tree to provide reliable initial values. In the second step, the poses are further optimized in the parameter space by aligning the fingertip positions $p_k^{dex,ft}$ with the corresponding human fingertip positions $p_k^{mano,ft}$. and the optimization objective is formulated as:

$$\min_{\mathcal{G}_{dex}^{obj}=(T_{dex}^{obj}, J_{dex})} \sum_k \left\| p_k^{dex,ft} - p_k^{mano,ft} \right\|_2^2. \quad (1)$$

(3) Executable Action Transformation. The aim of this component is to obtain executable dexterous poses that are both aligned with real objects and feasible for execution on physical robotic platforms. First, we need align the real object pose with object in generated images, as the generated images do not always correspond precisely to the original scene due to the model hallucination as shown in Figure 3. Second, we need to transfer the dexterous pose to robot arm coordinate system to execute action.

Specially, we estimate the 6D pose T_o^{obs} of object in \mathcal{I}^{obs} by FoundationPose [37]. Then, the grasp pose is mapped from the object coordinate system to the real camera coordinate system according to the estimated object pose with transformation $\mathbf{T}_{o \rightarrow c} = T_o^{obs}$. Finally, the grasp is converted from the real camera coordinate system to the robot arm coordinate system via the hand-eye calibration with transformation $\mathbf{T}_{c \rightarrow r}$. The complete transformation can

be written as:

$$\mathcal{G}_{dex} = \mathbf{T}_{c \rightarrow r} \cdot \mathbf{T}_{o \rightarrow c} \cdot \mathcal{G}_{dex}^{obj}. \quad (2)$$

D. Force-Aware Adaptive Grasping

Although the human-to-robot action transfer strategy generates executable grasp poses, it only provides a target hand configuration. Executing this target in an open-loop manner can easily result in unstable grasps or excessive forces that may damage the object. This limitation arises because the predicted target is not always perfect, due to model hallucinations or the accumulation of errors from earlier stages in the framework.

To address this challenge, we propose a force-aware adaptive grasping strategy based on force-constrained position control. The target grasping force F_{target} is predicted by a foundation model (e.g., GPT-4o), while the force feedback is measured by the force sensor of each finger (e.g. the strain gauge-based sensor integrated into the electric cylinder for Inspire Hand). And the target finger configuration is set slightly tighter than the action obtained from the transfer strategy. The grasp is then executed under force-constrained position control, allowing the hand to adaptively adjust its squeezing motion in response to real-time force feedback, thereby ensuring stable and safe dexterous grasping.

Specifically, we first obtain a pre-grasp pose \mathcal{G}_{dex}^{pre} and a squeeze-grasp pose \mathcal{G}_{dex}^{squ} by optimization, where the former provides a collision-free initialization and the latter introduces a more stable grasp contact. The optimization is conducted by adjusting the target contact point $p_k^{mano,ft}$ in Eq. 1. For pre-grasp pose, the contact point is shifted outward by 5 cm along the local surface normal, while inward by 1 cm for the squeeze-grasp pose. During this optimization, the wrist translation and rotation are kept fixed to maintain optimization stability.

Then, the fingers are controlled to move from the pre-grasp pose toward the squeeze-grasp pose under a force-constrained position control strategy. Once the measured contact force reaches the predicted threshold, the current finger position is locked as the new target, preventing further motion to avoid over-squeezing. This strategy enables the hand to adapt to the object’s compliance while ensuring stable and safe grasping. The control output $u(t)$ generated by the PD controller \mathcal{C} is defined as:

$$u(t) = \mathcal{C}(\mathcal{G}^{target} - \mathcal{G}(t)),$$

$$\mathcal{G}^{target} = \begin{cases} \mathcal{G}_{dex}^{squ}, & F(t) < F_{target} \\ \mathcal{G}(t), & F(t) \geq F_{target} \end{cases} \quad (3)$$

where $\mathcal{G}(t)$ and \mathcal{G}^{target} denote the current and reference joint positions of the finger, $F(t)$ is the force measured by the force sensor.

Finally, the overall grasp is executed in two stages. In the first stage, the arm and dexterous hand move to a pre-grasp pose, where the arm is positioned approximately 10 cm away from the object along the wrist axis to ensure a safe approach. In the second stage, the arm and hand advance together toward the object until the final grasp pose is achieved.

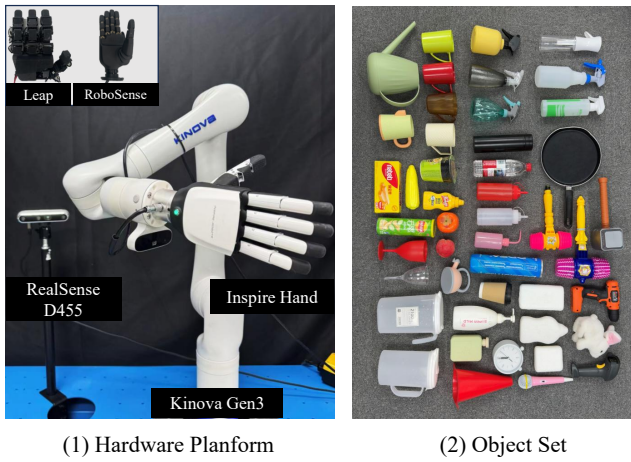


Fig. 4: Visualization of real world hardware platform and the objects used in experiments.

E. Extension to Manipulation

The OmniDexGrasp framework can be naturally extended from dexterous grasping to manipulation tasks, as grasping constitutes a fundamental prerequisite for prehensile manipulation. This extension leverages the generative capabilities of foundation models to reason about object motion post-grasping. Specifically, object motion can be obtained through existing techniques: for instance, by employing an LLM-driven keypoint prediction model [3] to estimate object keypoint movements, or by prompting a generative model to synthesize manipulation videos conditioned on successful grasps and subsequently tracking the object pose in the generated video using FoundationPose. The dexterous manipulation pose is then derived by updating $\mathbf{T}_{o \rightarrow c}$ in Eq. 2 based on the tracked object pose. By coupling these estimated object trajectories with the object-grasp constraint, our framework extends beyond grasping, demonstrating its potential for generalizable embodied intelligence. Qualitative results are illustrated in Figure 4 and the supplementary video.

IV. EXPERIMENTS

A. Experiment settings

We conduct diverse experiments on both real world and simulation experiments, we introduce the experiment settings from 1) robot platform, 2) task description, 3) the details of objects used in experiments.

Robot platform. For real world experiments, as shown in 4, we conduct all quantitative experiments on a 6-DOF Kinova Gen3 arm, a Realsense D455 camera, and an linkage-driven 6-DOF Inspire FTP Hand. Additionally, we conduct qualitative experiment with fully-actuated 16 DOF Leap Hand [38] and the linkage-driven 8-DOF RoboSense Papert Hand. For simulation experiments, we conduct experiments on the tendon-driven 22-DOF Shadow Hand.

Task description. For real world experiments, we evaluate our framework on six different grasping tasks to assess omni-ability across diverse scenarios and user prompts as

shown in Table I. Since no available methods are capable of simultaneously handling such a wide range of tasks, we further conduct extensive and comparison experiments on semantic grasping in [15] with representative baselines in both real world (Table II) and simulation (Table III).

Details of objects. We details the objects used in our experiments. In our real-world experiments, we used more than 40 objects selected to cover a wide range of shapes, sizes, and functional properties. These objects were grouped into eight categories: handles, spray bottles, liquid containers, packaged food items, cylinders, shafts, fragile objects, and mixed-category objects. In our simulation experiments, we employed over 100 objects from 33 distinct categories [15]. The testing objects are divided into three groups: seen, similar, and novel. "Seen" objects belong to categories encountered during model training, "similar" objects come from categories that are related but not identical to the training set, and "novel" objects represent entirely unseen categories. This categorization is designed to enable a more systematic evaluation of training-based grasp models. Importantly, our model has never been exposed to any dexterous robot grasping data during training, highlighting its ability to generalize to unseen objects and categories in simulation.

B. Evaluation Metric

We employ two types of evaluation metrics: grasp success rate and intention-consistency score. For grasp success rate *Suc.* in real-world experiments, a grasp is considered successful if the robotic hand can lift and hold the object while accurately following the given language command. Furthermore, we employ two GPT-assisted score inspired by [12]. Specifically, the grasping image and instruction is provided to GPT-4o or human expert, which evaluates two points on a 5-point scale: 1) *Inten.*: the alignment between the executed action and the command; 2) *Stab.*: grasp stability.

C. Can OmniDexGrasp achieve omni-capabilities across diverse dexterous grasping tasks?

We evaluate OmniDexGrasp on six representative dexterous grasping tasks (Table I, Figure 5). The framework achieves consistently high grasp success rates (*Suc.*, avg. 87.9%) and intention-consistency scores (*Inten.*, avg. 4.14), showing strong adaptability to diverse objects, scenes, and prompt modalities. Semantic-dependent tasks (semantic grasping, cluttered grasping, human-robot handover) benefit from foundation model integration, while visual- and demo-prompt tasks (region grasping, one-shot grasping) validate its flexibility on system inputs. Performance on fragile object grasping further highlights its ability to maintain stability while applying appropriate force. Overall, OmniDexGrasp demonstrates omni-ability across diverse tasks within a unified framework.

To the best of our knowledge, since no available dexterous grasping baseline can realize all these omni-capabilities, we do not include direct comparisons with other methods in Table I. Alternatively, we present extensive comparisons on

Task	Description	Prompt Type	ours		w/o Force		w/o Transfer	
			Suc.	Inten.	Suc.	Inten.	Suc.	Inten.
Semantic Grasping	Grasp guided by Language.	Language	83.3	3.82	60.0	2.90	20.0	0.40
Region / Point Grasping	Grasp guided by grasp point or mask.	Visual prompts	89.0	3.89	30.0	1.67	20.0	0.60
Targeted Grasp in Clutter	Grasp user-specified object in clutter.	Language	87.5	4.00	62.5	3.00	50.0	0.00
Human-Robot Handover	Grasp for handover interaction	Language	100.0	4.80	30.0	1.90	25.0	0.75
One-Shot Grasping	Grasp guided by single demonstration	Demo image	80.0	3.80	60.0	3.11	10.0	0.40
Fragile Object Grasping	Grasp fragile object safely	Language/force	88.0	4.50	56.0	2.67	0.0	0.33

TABLE I: The results of the framework on 6 diverse dexterous tasks and the ablation study. "w/o force" means using the poses obtained from the transfer strategy and executing them with open-loop position control with force feedback. "w/o transfer" means directly using the grasp poses estimated from the generative model.

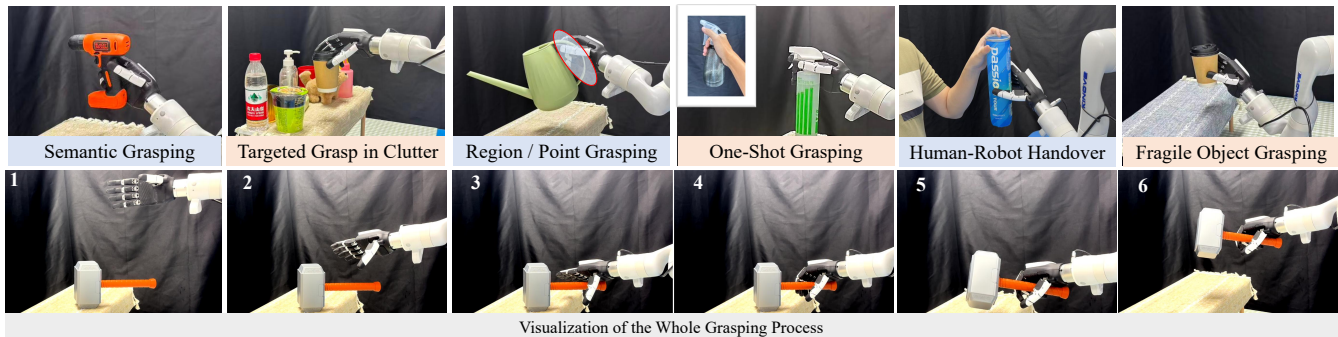


Fig. 5: Visualization of the six tasks (upper) and the whole grasping motion process (lower).

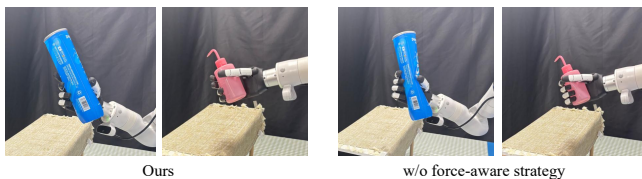


Fig. 6: The visualization of the force-aware strategy ablation shows that grasping without it may damage objects.

semantic grasping tasks (Section IV-E) through real-robot experiments (Table II) and simulation datasets (Table III). And we further conduct ablation studies in these six diverse tasks (Table I) to validate the effectiveness of each component of our method.

D. Are the transfer and force-aware grasping strategies crucial for the framework?

To validate the effectiveness of the two proposed strategies, we conduct an ablation study, as shown in Table I. The column "w/o Transfer" refers to directly using the grasp poses estimated from the generative model without first transferring them to the object coordinate system and then to the real-world coordinate system. The results show a performance decline because the generative model may produce hallucinations, causing the object pose to change during generation, which can lead to grasp failure. The column "w/o Force" refers to using the poses obtained from the transfer strategy and executing them with open-loop position control. Without force-aware adaptive grasping strategy, the grasp may become unstable if the fingers close too loosely, resulting in an approximate 40% reduction in success rate in average, or it may damage the object if the

fingers close too tightly as shown in Table 6. Overall, both strategies contribute to enhancing the robustness and stability of robotic grasping across different tasks.

E. Does OmniDexGrasp outperform existing dexterous grasping methods?

We conduct comparison experiments in both real-world and simulation settings to evaluate semantic grasping tasks and perform comprehensive comparisons. All baseline are trained following [15], where the baseline in Table II is trained on its complete dataset [15] containing over 43,504 samples and 1,536 objects across 33 categories. And the baselines in Table III are trained only on 10 seen categories and evaluated on 11 similar and 12 novel categories for a better evaluation of these generalization.

Table II presents the results in real-world experiments, OmniDexGrasp outperforms prior methods in both grasp stability and intention consistency, owing to the generalization capabilities of foundation models as well as our proposed transfer and force-aware grasping strategies. Table III presents comparisons with additional baseline methods in simulation, while prior methods achieve high performance on seen categories, their performance drops sharply on similar and novel categories, with average scores falling from 4.59 (seen categories) to 2.45 (similar) and 1.98 (novel) respectively, indicating limited generalization ability. In contrast, OmniDexGrasp maintains robust performance across all splits by leveraging foundation models to learn generalizable grasping rules rather than dataset-specific patterns, enabling effective adaptation to unseen scenarios.

	Handle		Spray Bottle		Liquid Container		Packaged Food		Cylinder		Shaft		Fragile Objects		Mixed Category	
	Suc.	Inten.	Suc.	Inten.	Suc.	Inten.	Suc.	Inten.	Suc.	Inten.	Suc.	Inten.	Suc.	Inten.	Suc.	Inten.
AffordDex	20.0	0.80	87.5	3.75	25.0	1.00	25.0	0.71	25.0	1.88	33.3	1.50	55.6	2.44	37.5	1.63
Ours	60.0	3.40	94.4	4.60	80.0	4.20	58.3	2.50	73.3	3.2	77.8	2.67	91.7	4.25	100.0	4.50

TABLE II: Grasping performance on different object categories in real-world experiments. Each category reports success rate (Suc.) and intention-consistency score (Inten.). The AffordDex [15] is trained on the dataset [15] following its paper.

	Seen		Similar		Novel	
	Stab.	Inten.	Stab.	Inten.	Stab.	Inten.
SceneDiffuser [39]	3.59	4.25	2.55	2.05	2.09	1.95
DexGYS [11]	3.95	4.67	2.95	2.22	2.83	1.88
AffordDex [15]	4.01	4.85	3.58	3.08	3.25	2.12
Ours	4.29	3.82	3.96	3.55	4.20	3.88

TABLE III: The results on the simulation dataset. “Seen,” “Similar,” and “Novel” denote categories that are seen during training, unseen but similar to the training categories, and unseen novel objects, respectively.

	Stab.	Inten.
Qwen-Image [17]	3.10	2.51
GPT-Image [18]	4.84	4.25
Wan2.2 [24]	2.52	1.62
Wan2.2* [24]	3.49	2.62
Kling-video [30]	4.83	4.07

TABLE IV: Generation quality across different foundation generative models.

F. Can different foundation generative models meet the requirements for dexterous grasping?

We analyze the performance of various foundation generative models to evaluate their suitability for dexterous grasping (Table IV, Figure 7). Specifically, we provide 20 scene images to different foundation generative models and assess their outputs using GPT-4o. Overall, foundation models demonstrate promising capabilities. We find that closed-source models, such as GPT-Image and Kling-Video, generally outperform open-source models like Owen-Image and Wan2.2. However, we find that the performance of open-source models can improve substantially after fine-tuning on task-specific datasets. For example, Wan2.2* denotes the weights fine-tuned on the hand-object interaction dataset [40], indicating considerable potential for enhancement.

Furthermore, we find that image-based foundation models exhibit stronger instruction-following abilities (higher intention-consistency scores). Conversely, video-based models tend to be better in maintaining scene and object pose consistency, which is important for grasping tasks.

G. Extension to manipulation tasks and cross embodiment

We conduct qualitative experiments to evaluate the applicability of our approach for achieving cross-embodiment on different dexterous hands and extending it to manipulation tasks beyond dexterous grasping. Visualizations of

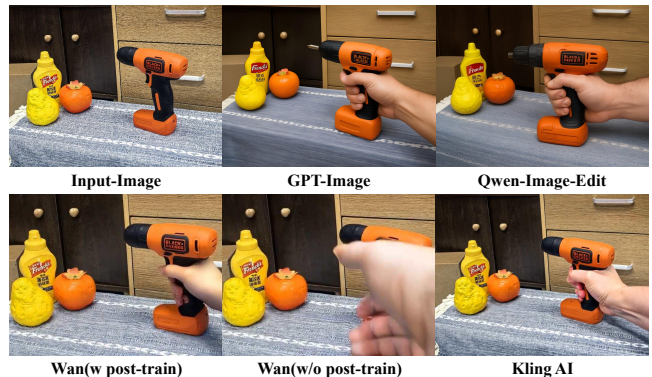


Fig. 7: The visualization of images generated from different foundation generative models.

these experiments are provided in Figure 4 and the supplementary video. Benefiting from our human-image-to-robot-action strategy, the framework can be easily adapted to different dexterous hands. Furthermore, by leveraging video generation models [30] and LLM-driven keypoint prediction models [3], our approach can accomplish several dexterous manipulation tasks. Since grasping forms the fundamental basis of manipulation, we anticipate that our method can be further extended to more complex manipulation scenarios. More visualizations are provided in the supplementary video.

V. CONCLUSIONS

We believe that achieving generalizable dexterous grasping with omni-capabilities—across diverse user prompts, dexterous embodiments, and task scenarios—is a central challenge in robotics. To address this, we propose the OmniDex-Grasp framework, which leverages foundation models to acquire generalizable understanding ability of scenes and user intentions, and generates human grasp images to guide dexterous grasping. We introduce a human-image-to-robot-action transfer strategy that converts these grasp images into feasible robot actions through hand-object reconstruction and dexterous retargeting. In addition, we design a force-aware adaptive grasping mechanism to ensure stable and safe execution with force feedback. Extensive experiments in both simulation and real-world settings demonstrate the effectiveness of our approach: OmniDexGrasp achieves high success rates and intention-consistency scores across six diverse dexterous grasping tasks, and outperforms state-of-the-art methods on unseen object categories. Overall, our framework highlights the potential of foundation-model-driven approaches for generalizable embodied intelligence.

VI. ACKNOWLEDGMENT

We would like to thank RoboSense for providing the Papert hand. This work was supported partially by NSFC (92470202, U21A20471), Guangdong NSF Project (No. 2023B1515040025), Guangdong Key Research and Development Program (No.2024B0101040004, 2025B0909020002).

REFERENCES

- [1] J. Jian, Y.-L. Wei, C. Mou, Y. Lin, X. Zhu, Y. Shen, W.-S. Zheng, and R. Hu, "ZeroDexgrasp: Zero-shot task-oriented dexterous grasp synthesis with prompt-based multi-stage semantic reasoning," *arXiv preprint arXiv:2511.13327*, 2025.
- [2] H.-S. Fang, H. Yan, Z. Tang, H. Fang, C. Wang, and C. Lu, "Anydex-grasp: General dexterous grasping for different hands with human-level learning efficiency," *arXiv preprint arXiv:2502.16420*, 2025.
- [3] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," *arXiv preprint arXiv:2409.01652*, 2024.
- [4] Z. Wei, Z. Xu, J. Guo, Y. Hou, C. Gao, Z. Cai, J. Luo, and L. Shao, "D (r, o) grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping," *arXiv preprint arXiv:2410.01702*, 2024.
- [5] Y. Lin, Y.-L. Wei, H. Liao, M. Lin, C. Xing, H. Li, D. Zhang, M. Cutkosky, and W.-S. Zheng, "Typetele: Releasing dexterity in teleoperation by dexterous manipulation types," *arXiv preprint arXiv:2507.01857*, 2025.
- [6] Z. Zheng, J.-F. Cai, X.-M. Wu, Y.-L. Wei, Y.-M. Tang, A. Wu, and W.-S. Zheng, "imanip: Skill-incremental learning for robotic manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 13 890–13 900.
- [7] Y.-L. Wei, H. Liao, Y. Lin, P. Wang, Z. Liang, G. Liu, and W.-S. Zheng, "Cyclemanip: Enabling cyclic task manipulation via effective historical perception and understanding," *arXiv preprint arXiv:2512.01022*, 2025.
- [8] G.-H. Xu, Y.-L. Wei, D. Zheng, X.-M. Wu, and W.-S. Zheng, "Dexterous grasp transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [9] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," *arXiv preprint arXiv:2210.02697*, 2022.
- [10] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. 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*, 2023, pp. 4737–4746.
- [11] Y.-L. Wei, J.-J. Jiang, C. Xing, X.-T. Tan, X.-M. Wu, H. Li, M. Cutkosky, and W.-S. Zheng, "Grasp as you say: Language-guided dexterous grasp generation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 46 881–46 907, 2024.
- [12] K. Li, J. Wang, L. Yang, C. Lu, and B. Dai, "Semgrasp: Semantic grasp generation via language aligned discretization," in *European Conference on Computer Vision*. Springer, 2024, pp. 109–127.
- [13] J. He, D. Li, X. Yu, Z. Qi, W. Zhang, J. Chen, Z. Zhang, Z. Zhang, L. Yi, and H. Wang, "Dexvlg: Dexterous vision-language-grasp model at scale," *arXiv preprint arXiv:2507.02747*, 2025.
- [14] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu, "Oakink: A large-scale knowledge repository for understanding hand-object interaction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 953–20 962.
- [15] Y.-L. Wei, M. Lin, Y. Lin, J.-J. Jiang, X.-M. Wu, L.-A. Zeng, and W.-S. Zheng, "Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance," *arXiv preprint arXiv:2503.07360*, 2025.
- [16] J. Jian, X. Liu, Z. Chen, M. Li, J. Liu, and R. Hu, "G-dexgrasp: Generalizable dexterous grasping synthesis via part-aware prior retrieval and prior-assisted generation," *arXiv preprint arXiv:2503.19457*, 2025.
- [17] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S.-m. Yin, S. Bai, X. Xu, Y. Chen *et al.*, "Qwen-image technical report," *arXiv preprint arXiv:2508.02324*, 2025.
- [18] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [19] S. Patel, S. Mohan, H. Mai, U. Jain, S. Lazebnik, and Y. Li, "Robotic manipulation by imitating generated videos without physical demonstrations," *arXiv preprint arXiv:2507.00990*, 2025.
- [20] M. Tayyab Khan and A. Waheed, "Foundation model driven robotics: A comprehensive review," *arXiv e-prints*, pp. arXiv–2507, 2025.
- [21] M. Pan, J. Zhang, T. Wu, Y. Zhao, W. Gao, and H. Dong, "Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 359–17 369.
- [22] L. Huang, H. Zhang, Z. Wu, S. Christen, and J. Song, "Fungrasp: Functional grasping for diverse dexterous hands," *IEEE Robotics and Automation Letters*, 2025.
- [23] J. Jian, X. Liu, M. Li, R. Hu, and J. Liu, "Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 713–14 724.
- [24] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.
- [25] K.-Y. Lin, H. Wang, W. Ren, and K. Han, "Panoptic captioning: An equivalence bridge for image and text," in *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025.
- [26] J. Zhou, T. Ma, K. Lin, R. Qiu, Z. Wang, and J. Liang, "Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 22 551–22 561.
- [27] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.
- [28] J. Jang, S. Ye, Z. Lin, J. Xiang, J. Bjorck, Y. Fang, F. Hu, S. Huang, K. Kundalia, Y.-C. Lin *et al.*, "Dreamgen: Unlocking generalization in robot learning through neural trajectories," *arXiv e-prints*, pp. arXiv–2505, 2025.
- [29] J. Liang, R. Liu, E. Ozguroglu, S. Sudhakar, A. Dave, P. Tokmakov, S. Song, and C. Vondrick, "Dreamitate: Real-world visuomotor policy learning via video generation," *arXiv preprint arXiv:2406.16862*, 2024.
- [30] "Kling ai," <https://www.klingai.com/>, 2024, accessed: 2024-02-10.
- [31] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.
- [32] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3d with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9826–9836.
- [33] Hyper3D, "Hyper3d: Ai-powered 3d model generator," 2024. [Online]. Available: <https://hyper3d.ai/>
- [34] T. Lee, B. Wen, M. Kang, G. Kang, I. S. Kweon, and K.-J. Yoon, "Any6d: Model-free 6d pose estimation of novel objects," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 633–11 643.
- [35] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," *arXiv preprint arXiv:2212.06870*, 2022.
- [36] Y. Liu, X. Long, Z. Yang, Y. Liu, M. Habermann, C. Theobalt, Y. Ma, and W. Wang, "Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 7037–7047.
- [37] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 868–17 879.
- [38] K. Shaw, A. Agarwal, and D. Pathak, "Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning," *arXiv preprint arXiv:2309.06440*, 2023.
- [39] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu, "Diffusion-based generation, optimization, and planning in 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 750–16 761.
- [40] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang, "Egodex: Learning dexterous manipulation from large-scale egocentric video," *arXiv preprint arXiv:2505.11709*, 2025.