DexGraspVLA: A Vision-Language-Action Framework Towards General Dexterous Grasping

Yifan Zhong[1,2*], **Xuchuan Huang**[1,2*], **Ruochong Li**[2,3], **Ceyao Zhang**[1,2], **Zhang Chen**[1,2], **Tianrui Guan**[1,2]
Fanlian Zeng[2,4], **Ka Nam Lui**[1,2], **Yuyao Ye**[1,2], **Yitao Liang**[1,2], **Yaodong Yang**[1,2†], **Yuanpei Chen**[1,2†]

[1]Institute for Artificial Intelligence, Peking University.
[2]PKU-PsiBot Joint Lab.
[3]Hong Kong University of Science and Technology (Guangzhou).
[4]University of Pennsylvania.

Abstract

Dexterous grasping remains a fundamental yet challenging problem in robotics. A general-purpose robot must be capable of grasping diverse objects in arbitrary scenarios. However, existing research typically relies on restrictive assumptions, such as single-object settings or limited environments, showing constrained *generalization*. We present **DexGraspVLA**, a hierarchical framework for robust generalization in language-guided general **dex**terous **grasp**ing and beyond. It utilizes a pre-trained **V**ision-**L**anguage model as the high-level planner and learns a diffusion-based low-level **A**ction controller. The key insight to achieve *generalization* lies in iteratively transforming diverse language and visual inputs into domain-invariant representations via foundation models, where imitation learning can be effectively applied due to the alleviation of domain shift. Notably, our method achieves a *90+%* dexterous grasping success rate under *thousands of* challenging unseen cluttered scenes. Empirical analysis confirms the *consistency* of internal model behavior across environmental *variations*, validating our design. DexGraspVLA also, for the first time, simultaneously demonstrates free-form long-horizon prompt execution, robustness to adversarial objects and human disturbance, and failure recovery. Extended application to nonprehensile grasping further proves its generality.

Website — https://dexgraspvla.github.io

1 Introduction

Dexterous multi-fingered hands, as versatile robotic end-effectors, have demonstrated remarkable capabilities across various manipulation tasks (Qi et al. 2023; Huang et al. 2023a; Lin et al. 2024a; Chen et al. 2022; Zakka et al. 2023; Chen et al. 2023). Among these, grasping serves as the most fundamental prerequisite, yet remains one of the most challenging problems. Existing dexterous grasping approaches primarily consider isolated objects or simplified settings. Nevertheless, real-world applications demand more general grasping capabilities that can function reliably in diverse unseen scenarios, which presents multifaceted challenges. At the object level, the policy must generalize across diverse physical properties including geometries, masses, textures, and orientations. Beyond object characteristics, the system must also demonstrate robustness to various environmental factors, such as lighting conditions, background complexities, and potential disturbances. Compounding these challenges, cluttered scenarios further demand sophisticated reasoning capabilities, as planning the optimal sequence to grasp all objects becomes a crucial cognitive task that extends beyond simple grasp execution.

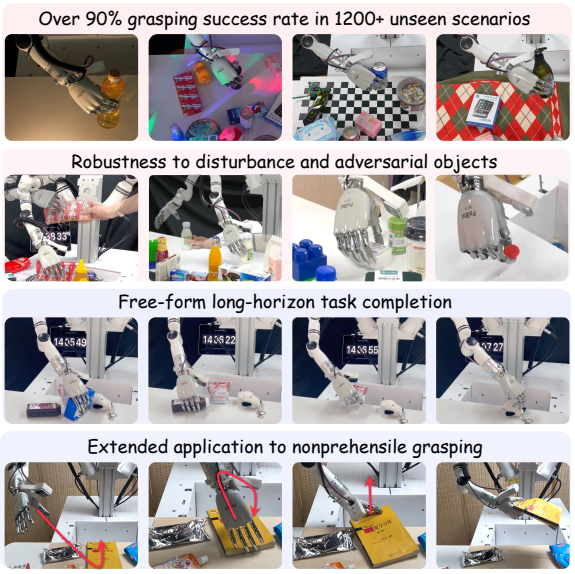

Figure 1: We propose **DexGraspVLA**, a hierarchical VLA framework that reaches a 90+% dexterous grasping success rate under thousands of unseen cluttered scenes in a "zero-shot" environment. It robustly handles adversarial objects, human disturbance, failure recovery, and free-form long-horizon prompts. Extended application to nonprehensile grasping further proves its generality.

One line of research adopts a two-stage pipeline: first predicting a grasp pose from single-frame perception, then executing open-loop motion planning to reach the pose (Chen, Bohg, and Liu 2024; Turpin et al. 2023, 2022). However, these methods rely heavily on precise calibration and me-

*These authors contributed equally.
†Corresponding author emails: yuanpei.chen312@gmail.com, yaodong.yang@pku.edu.cn.

chanical accuracy. By contrast, end-to-end paradigms, such as imitation and reinforcement learning, enable closed-loop control by continuously adjusting actions based on real-time feedback, offering more robust and adaptive solutions. Reinforcement learning has achieved notable successes in simulation (Akkaya et al. 2019; Yang et al. 2024; Pitz et al. 2023; Handa et al. 2023), but simulating real-world physical complexity remains challenging, resulting in an inevitable sim-to-real gap. Imitation learning learns directly from human demonstrations and avoids this gap, but often struggles to generalize beyond the training data. This issue is further compounded by the impracticality of collecting expert trajectories across the full spectrum of objects and environmental variations required for general grasping. As a result, a key challenge is how to effectively leverage limited expert data to achieve broad *generalization*.

The rapid emergence of vision and language foundation models (Oquab et al. 2023; Radford et al. 2021; Hurst et al. 2024; Kirillov et al. 2023) presents promising opportunities for robotic manipulation. Pretrained on internet-scale data, these models exhibit remarkable world knowledge and generalization over visual and linguistic inputs. To harness these capabilities for decision making, researchers have integrated them into action generation, leading to the development of vision-language-action (VLA) models (Zhong et al. 2025). One straightforward approach directly trains vision-language models (VLMs) end-to-end on robot data (Kim et al. 2024; Black et al. 2024). However, this paradigm demands massive manually collected demonstrations (O'Neill et al. 2023) in an attempt to encompass real-world diversity and complexity. Even so, these models exhibit markedly reduced performance on unseen scenarios and still require fine-tuning to handle new conditions. Alternatively, modular frameworks use frozen foundation models to infer task affordances more robustly across environments (Huang et al. 2024, 2023b; Stone et al. 2023), but their low-level policies are typically open-loop or lack generalization. Achieving generalizable closed-loop policies with foundation models remains an open challenge.

In this paper, we present **DexGraspVLA**, a hierarchical VLA framework for robust generalization in language-guided dexterous grasping and beyond, by integrating the complementary strengths of foundation models and imitation learning. The key idea is to leverage foundation models to iteratively transform *diverse* visual and linguistic inputs into *domain-invariant* representations, upon which imitation learning can be efficiently and effectively applied thanks to the alleviation of *domain shift*. As a result, novel scenarios no longer induce failure, as foundation models translate them into representations resembling those encountered during training—thus remaining within the learned policy's domain. Following this principle, DexGraspVLA employs a pre-trained VLM as a high-level planner to plan the overall task and generate *domain-invariant* affordance signals. Guided by these signals, a low-level controller further refines multimodal inputs into *domain-invariant* representations using vision foundation models, and generates closed-loop action through a diffusion-based action head learned via imitation. This design combines the extensive world knowledge and generalization ability of foundation models with action modeling capacity of imitation learning, enabling strong performance in real-world scenarios.

Notably, DexGraspVLA achieves an unprecedented 90.8% success rate for grasping in cluttered scenes spanning 1,287 unseen object, lighting, and background combinations, all tested in a "zero-shot" environment. Its generalization performance significantly surpasses that of existing baselines. Moreover, DexGraspVLA robustly handles adversarial objects, human disturbances, and failure recovery. On a single-object benchmark, it achieves 98.6% success, outperforming ablated variants whose controller learns directly from raw visual inputs by at least 48%. Further analysis reveals *consistent* internal model behaviours across *varying* environments, validating our design and explaining its robustness. Beyond single-step tasks, DexGraspVLA executes free-form, long-horizon instructions with embodied reasoning, reaching 89.6% success rate. We further extend DexGraspVLA to nonprehensile object grasping (Zhou and Held 2023), a challenging task that often requires dexterous pre-grasp maneuvers difficult for parallel grippers. DexGraspVLA achieves strong performance using only a modest number of demonstrations, further highlighting its generality across diverse manipulation scenarios. These results establish DexGraspVLA as a general, instruction-driven framework that learns from limited demonstrations and generalizes reliably to real-world settings, marking a promising step toward general dexterous grasping and beyond.

2 Related Work

Dexterous Grasping. Dexterous grasping methods are typically divided into two-stage and end-to-end approaches. Two-stage methods first generate a grasp pose—via sampling (Zhang et al. 2024b; Fang et al. 2025), optimization (Wang et al. 2023b; Chen, Bohg, and Liu 2024), or regression (Li et al. 2022; Liu et al. 2020)—and reach it with motion planning. Though they benefit from modularity and synthetic data, their open-loop nature makes them vulnerable to disturbances and calibration errors. End-to-end methods learn grasping policies via reinforcement learning in massively parallel simulation (Wan et al. 2023; Zhang et al. 2024a; Singh et al. 2024), which efficiently acquire emergent dexterity but suffer from sim-to-real gaps. In this work, we explore imitation learning from human demonstrations, which has shown promise on complex tasks (Qin, Su, and Wang 2022; Guzey et al. 2024; Lin et al. 2024b). Our core contribution is to address the central challenge of generalization in imitation learning (Intelligence 2025). We show that performing imitation learning on domain-invariant representations derived from foundation models enables strong generalization to unseen scenarios.

Foundation Models for Generalizable Robotic Policies. Vision and language foundation models pre-trained on web-scale data have shown impressive world knowledge and generalization (Kirillov et al. 2023; Oquab et al. 2023; Team 2025), making them promising for robotics. A common approach, as seen in OpenVLA (Kim et al. 2024) and π_0 (Black et al. 2024), directly fine-tunes VLMs on

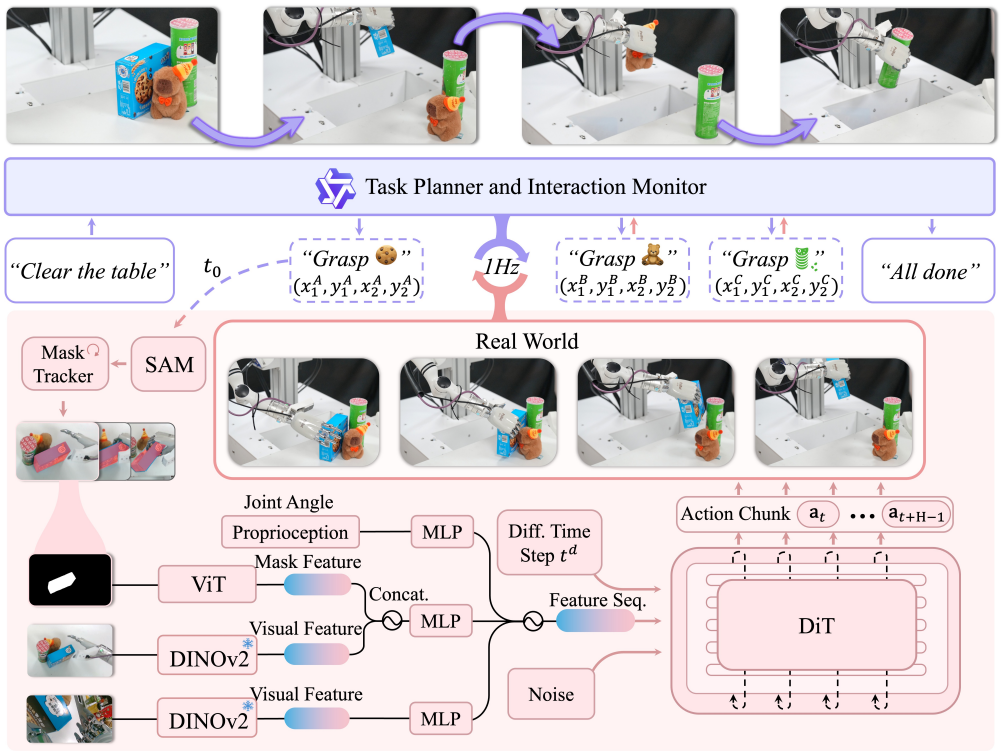

Figure 2: **Overview of DexGraspVLA**. A pre-trained VLM-based high-level planner (purple) decomposes prompts into object-level grasping instructions with bounding boxes. The diffusion-based low-level controller (pink) tracks the target mask, encodes multimodal observations (RGB images, mask, proprioception), and predicts an action chunk via a DiT model. The planner monitors execution and continually proposes new instructions based on the updated scene until the task is fully completed.

robot data in the hope of transferring vision-language knowledge to the policy for broad generalization. However, this typically requires a massive amount of diverse demonstrations (O'Neill et al. 2023), yet still struggles with unseen scenarios and catastrophic forgetting. A more related line of work to ours instead leverages frozen foundation models to robustly infer task affordances—i.e., where and how to manipulate—in novel environments, guiding either motion planning (Huang et al. 2024, 2023b; Pan et al. 2025) or a learned action head (Stone et al. 2023). However, the former often depends heavily on accurate calibration, involves considerable human design, or lacks robustness due to open-loop control. The latter still maps *raw* visual inputs directly to actions, making it vulnerable to domain shift. In contrast, to achieve generalization across diverse real-world domains, our framework employs foundation models to iteratively transform free-form language prompts and diverse visual perceptions into domain-invariant representations. These representations enable imitation learning to be applied efficiently and effectively, collectively leading to robust generalization.

3 Problem Formulation

Our goal is to develop a vision-based control policy for language-guided dexterous grasping, formulated as a sequential decision-making problem. Initially, a language instruction l is given, *e.g.* "*grasp the toy*", to directly specify the target object. At each timestep t, the policy π receives a first-view image $\mathbf{I}_t^{\mathrm{w}} \in \mathbb{R}^{H\times W\times 3}$ from the wrist camera (H and W denote the height and width of the image), a third-view image $\mathbf{I}_t^{\mathrm{h}} \in \mathbb{R}^{H\times W\times 3}$ from the head camera, and the robot proprioception $\mathbf{s}_t \in \mathbb{R}^{13}$ consisting of arm and hand joint angles $\mathbf{s}_t^{\mathrm{arm}} \in \mathbb{R}^7, \mathbf{s}_t^{\mathrm{hand}} \in \mathbb{R}^6$. Conditioned on these observations, the robot produces an action $\mathbf{a}_t = (\mathbf{a}_t^{\mathrm{arm}}, \mathbf{a}_t^{\mathrm{hand}}) \in \mathbb{R}^{13}$, where $\mathbf{a}_t^{\mathrm{arm}} \in \mathbb{R}^7$ and $\mathbf{a}_t^{\mathrm{hand}} \in \mathbb{R}^6$ denote the target joint angles for arm and hand respectively, by sampling from the action distribution $\pi(\cdot|\{\mathbf{I}_j^{\mathrm{w}}\}_{j=0}^t, \{\mathbf{I}_j^{\mathrm{h}}\}_{j=0}^t, \{\mathbf{s}_j\}_{j=0}^t, l)$. This process continues until a termination condition is reached. The robot receives a binary reward $r \in \{0, 1\}$ indicating whether it has completed the instruction l successfully. The goal of the policy π is to maximize the expected reward $\mathbb{E}_{l,\{(\mathbf{I}_j^{\mathrm{w}},\mathbf{I}_j^{\mathrm{h}},\mathbf{s}_j,\mathbf{a}_j)\}_{j=0}^T}[r]$.

More generally, we consider cases where the user prompt p may be a long-horizon task involving multiple grasping steps, such as "*clear the table*". This requires the policy π to reason about the prompt, decompose it into individual grasping instructions $\{l_i\}$, and complete them sequentially.

4 Methods

This section introduces DexGraspVLA, the first hierarchical VLA framework for dexterous grasping. We will first elaborate DexGraspVLA framework (Section 4.1) and then detail our data collection procedure (Section 4.2), which together enable the training of a dexterous grasping policy.

4.1 DexGraspVLA Framework

As illustrated in Figure 2, DexGraspVLA adopts a hierarchical and modularized architecture composed of a planner and a controller. Below we explain how each part is designed.

Planner. We recognize that to achieve general dexterous grasping, the model must handle multimodal inputs, perform visual grounding, and conduct reasoning about user prompts. Building upon recent advances, we adopt an off-the-shelf pre-trained Qwen VLM (Bai et al. 2023; Team 2025) as a high-level planner to dynamically plan and monitor the dexterous grasping workflow. Given a user prompt p (e.g., "clear the table"), the planner proposes a grasping instruction l (e.g., "grasp the cookie") as the first step.

For each l, the planner guides the low-level controller by marking the target object bounding box (x_1, y_1, x_2, y_2) as task affordance in the head camera image $\mathbf{I}^{\mathrm{h}}_{t_0}$ at the initial timestep t_0. While the phrasing and content of language instruction can be diverse and flexible for different users and cases, *i.e.*, showing *domain-variance*, the bounding box is a consistent format for object localization regardless of the changes in language and visual inputs, *i.e.*, achieving *domain-invariance*. Thus, this transformation alleviates the learning challenge for the controller.

On issuing the bounding box, the planner monitors controller execution, resets robot after each grasp attempt, and proposes updated instruction l until prompt p is completed.

Controller. Based on the bounding box (x_1, y_1, x_2, y_2), the controller aims to grasp the intended object in cluttered environments. We feed this bounding box as input to SAM (Kirillov et al. 2023) to obtain an initial binary mask $\mathbf{m}_0 \in \{0,1\}^{H\times W\times 1}$ of the target object and then use Cutie (Cheng et al. 2024) to continuously track the mask over time, producing $\mathbf{m}_t$ at each timestep t. This ensures accurate identification in cluttered scenes throughout the process. The problem is to learn the policy π that effectively models the action distribution $\pi(\cdot|\mathbf{I}^{\mathrm{w}}_t, \mathbf{I}^{\mathrm{h}}_t, \mathbf{s}_t, \mathbf{m}_t)$.

To achieve general-purpose dexterous grasping, the system must generalize effectively across diverse real-world scenarios. However, the high variability in raw visual inputs $\mathbf{I}^{\mathrm{w}}_t, \mathbf{I}^{\mathrm{h}}_t$ poses a fundamental challenge to learning task-critical representations. Traditional imitation learning approaches often fail catastrophically even under minor variations in objects or environmental conditions. To address this issue, our solution is again to convert potentially *domain-varying* inputs into *domain-invariant* representations suitable for imitation learning. We recognize that *while pixel-level perception vary widely, the fine-grained semantic features extracted by foundation models tend to be more robust and consistent* (Tang et al. 2023; Wang et al. 2023a). Thus, we utilize a feature extractor ϕ, DINOv2 (Oquab et al. 2023), to obtain features from raw images. At timestep t, we obtain head camera image features $\mathbf{z}^{\mathrm{h}}_t = \phi^{\mathrm{h}}(\mathbf{I}^{\mathrm{h}}_t) \in \mathbb{R}^{L^{\mathrm{h}}\times D^{\mathrm{h}}}$, and wrist camera image features $\mathbf{z}^{\mathrm{w}}_t = \phi^{\mathrm{w}}(\mathbf{I}^{\mathrm{w}}_t) \in \mathbb{R}^{L^{\mathrm{w}}\times D^{\mathrm{w}}}$, where $L^{\mathrm{h}}, D^{\mathrm{h}}, L^{\mathrm{w}}, D^{\mathrm{w}}$ denote length and hidden dimension of the feature sequences for head and wrist respectively. As we show in Section 5.5, these extracted features remain comparatively invariant to distracting visual factors.

Up to now, raw language and vision inputs, including instruction l and images $\mathbf{I}^{\mathrm{w}}_t, \mathbf{I}^{\mathrm{h}}_t$, have been iteratively transformed into domain-invariant representations, including mask $\mathbf{m}_t$ and features $\mathbf{z}^{\mathrm{h}}_t, \mathbf{z}^{\mathrm{w}}_t$, by leveraging foundation models. This lays the stage for imitation learning. We now learn the policy π that predicts an action chunk of horizon H conditioning on these representations.

To fuse the object mask with head camera features, we project $\mathbf{m}_t$ into the head image feature space using a randomly initialized ViT, producing $\mathbf{z}^{\mathrm{m}}_t \in \mathbb{R}^{L^{\mathrm{h}}\times D^{\mathrm{h}}}$, and concatenate it with $\mathbf{z}^{\mathrm{h}}_t$ patch-wise to obtain $\bar{\mathbf{z}}^{\mathrm{h}}_t \in \mathbb{R}^{L^{\mathrm{h}}\times 2D^{\mathrm{h}}}$. Subsequently, we map $\bar{\mathbf{z}}^{\mathrm{h}}_t$, wrist-camera features $\mathbf{z}^{\mathrm{w}}_t$, and robot state $\mathbf{s}_t$ into a common embedding space with separate MLPs, yielding $\tilde{\mathbf{z}}^{\mathrm{h}}_t$, $\tilde{\mathbf{z}}^{\mathrm{w}}_t$, and $\tilde{\mathbf{z}}^{\mathrm{s}}_t$. These embeddings are then concatenated to form the full observation feature sequence $\tilde{\mathbf{z}}^{\mathrm{obs}}_t \in \mathbb{R}^{(1+L^{\mathrm{h}}+L^{\mathrm{w}})\times D}$.

For action prediction, we employ a DiT (Peebles and Xie 2023) to generate multi-step actions, following the diffusion policy paradigm (Chi et al. 2023; Liu et al. 2024). At each timestep t, we bundle the next H actions into a chunk $\mathbf{A}_t = \mathbf{a}_{t:t+H} = [\mathbf{a}_t, \mathbf{a}_{t+1}, \ldots, \mathbf{a}_{t+H-1}]$. During training, a random diffusion step $t^d = k$ is sampled, and Gaussian noise $\boldsymbol{\epsilon}$ is added to $\mathbf{A}_t$, yielding the noised action tokens $\mathbf{x}_k = \alpha_k \mathbf{A}_t + \sigma_k \boldsymbol{\epsilon}$, where α_k and σ_k are DDPM coefficients. We then feed $\mathbf{x}_k$ into the DiT alongside the observation feature sequence $\tilde{\mathbf{z}}^{\mathrm{obs}}_t$. Each DiT layer performs bidirectional self-attention over the action tokens, cross-attention to $\tilde{\mathbf{z}}^{\mathrm{obs}}_t$, and MLP transformations, ultimately predicting the original noise $\boldsymbol{\epsilon}$. By minimizing the noise prediction error, the model learns to reconstruct the ground-truth action chunk $\mathbf{A}_t$. At inference time, iterative denoising steps recover the intended multi-step action sequence from the learned distribution, enabling imitation of multimodal behaviors. We also employ the receding horizon control strategy that only executes the first H_a actions before generating a new action chunk prediction, enhancing responsiveness.

Overall, DexGraspVLA performs imitation learning on *domain-invariant* representations derived from *domain-varying* inputs via foundation models. This approach leverages the world knowledge and generalization capabilities of foundation models while effectively capturing the mapping from these abstracted representations to action output.

4.2 Data Collection

To train our dexterous grasping policy, we manually collect a dataset consisting of 2,094 successful demonstrations in cluttered scenes using 36 household objects varying in size, weight, geometry, texture, material, and category. Each episode $\tau = \{(\mathbf{I}^{\mathrm{h}}_t, \mathbf{I}^{\mathrm{w}}_t, \mathbf{s}_t, \mathbf{m}_t, \mathbf{a}_t)\}^T_{t=0}$ records raw camera images $\mathbf{I}^{\mathrm{h}}_t, \mathbf{I}^{\mathrm{w}}_t$, robot proprioception $\mathbf{s}_t$, object mask $\mathbf{m}_t$, and action $\mathbf{a}_t$ at each timestep t. The mask $\mathbf{m}_t$ is labeled in the same way as in the controller. For each object, we place it randomly and collect multiple grasping demonstrations, with the surrounding objects randomized between episodes. These demonstrations are performed at typical human motion speeds, taking about 3.5 s each. They undergo rigorous inspection to ensure quality. The DexGraspVLA controller is trained on this dataset with imitation learning.

5 Experiments

In this section, we extensively evaluate DexGraspVLA. All experiments are conducted in a different environment from the demonstration setup, ensuring a "*zero-shot*" setting to rigorously assess generalization to novel real-world scenarios. Our experiments seek to address the following questions: (1) **Large-scale Generalization** (Section 5.2): Can DexGraspVLA generalize to thousands of unseen object, lighting, and background combinations? (2) **Baseline Comparison** (Section 5.3): How does its performance compare to baselines? (3) **Ablation Study** (Section 5.4): How much does imitation learning on domain-invariant representations improve generalization? (4) **Mechanism Analysis** (Section 5.5): Are its internal model behaviors consistent under varying environments? (5) **Long-horizon Task** (Section 5.6): How effectively does DexGraspVLA handle free-form, long-horizon instructions? (6) **Extension to Nonprehensile Grasping** (Section 5.7): Can it be extended to other dexterous manipulation skills beyond grasping?

5.1 Experiment Setups

Hardware Platform. As shown in Figure 3, our setup includes a 7-DoF RealMan RM75-6F arm and a 6-DoF PsiBot G0-R hand. A wrist-mounted RealSense D405C camera provides a first-person view, and a head-mounted D435 camera captures a third-person view. Objects are placed on a table in front, and the control frequency is 20 Hz.

Baselines. We compare DexGraspVLA (Ours) with several state-of-the-art (SOTA) VLA baselines fine-tuned on our dataset, including full-parameter (Full FT) and LoRA fine-tuned variants of π_0 (Black et al. 2024), RDT (Liu et al. 2024), OpenVLA (Kim et al. 2024), and OpenVLA-OFT (Kim, Finn, and Liang 2025). We also evaluate two ablated versions of our method: 1) DINOv2-train: Identical to DexGraspVLA but with trainable DINOv2 encoders. 2) ViT-small: Identical to DexGraspVLA but replaces DINOv2 with smaller, trainable ViTs. Empirically, the ViT-small variant represents an enhanced version of Diffusion Policy (Chi et al. 2023), a SOTA imitation learning baseline. For all experiments, the high-level planner is based on Qwen-VL-Chat (Bai et al. 2023), except in the long-horizon task (Section 5.6), where we use Qwen2.5-VL-72B-Instruct (Team 2025). Implementation details are in the appendix. To account for inference randomness, we report Ours@k ($k = 1, 2, 3$) in Section 5.2, where up to k attempts are allowed per test. Ours@1 is equivalent to Ours. Re-grasps performed by the policy after an initial failure within a single attempt are allowed and not counted separately.

5.2 Large-Scale Generalization Evaluation

Tasks. We curate 360 unseen objects, 6 unseen backgrounds, and 3 unseen lighting conditions. The objects span diverse sizes, weights, geometries, textures, materials, and categories, while remaining graspable by our dexterous hand. Backgrounds and lighting conditions are selected to be visually distinct. We evaluate generalization through three grasping tasks in cluttered scenes (around 6 objects per scene): (1) *Unseen Objects*: Each of the 360 objects is grasped once in a random scene on a white table under white light (360 tests). (2) *Unseen Backgrounds*: A subset of 103 objects $\mathcal{S}$ forms 103 scenes per background under white light, totaling 618 tests. (3) *Unseen Lightings*: The same $\mathcal{S}$ forms 103 scenes per lighting condition on a white table (309 tests). Details can be found in the appendix.

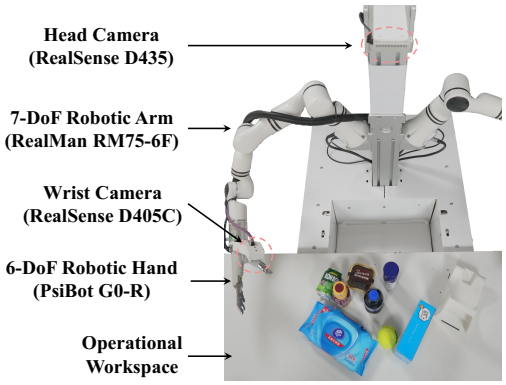

Figure 3: Our hardware platform.

Metric. A grasp is successful if the object is held 10 cm above the table for 20 s. Success rate is the ratio of successes to total tests; aggregated performance is a weighted average by task proportion.

Results. We present the quantitative results in Table 1c. From the 1st row ("Ours@1"), DexGraspVLA achieves a 91.1% single-attempt success rate on 360 unseen objects, 90.5% on 6 unseen backgrounds, and 90.9% under 3 unseen lighting conditions, yielding a 90.8% aggregated success rate. These results demonstrate robust and accurate control of the dexterous hand to grasp specified objects from clutter in diverse unseen conditions, without domain-specific fine-tuning. This highlights strong generalization and suggests that our framework substantially alleviates the overfitting challenge in imitation learning. We further analyze the source of this generalization in Section 5.5 and extend its application in Section 5.7.

Qualitatively, DexGraspVLA robustly handles challenging cases involving transparent, deformable, reflective, or background-camouflaged objects. It also dexterously adapts to diverse geometries and poses — e.g., grasping a bottle from the side, picking up a small earbud case from the top, or retrieving an awkwardly placed box. The closed-loop policy enables re-grasping after failed attempts and tolerates human-induced perturbations by tracking object motion. Such robustness stems from three factors: first, foundation-model-based perception ensures semantic consistency under appearance variation; second, imitation learning avoids the need for explicit object modeling; and third, diffusion-based action head captures multi-modal action distributions.

From the 2nd and 3rd rows ("Ours@2" and "Ours@3"), we observe that allowing up to three attempts further boosts performance to 96.9%, indicating the capacity to reach even higher success rates. Finally, our model takes around 6 s to grasp an object, which is close to that of humans and ensures practical usability in real-world scenarios.

	Seen Objects	Unseen Objects	Unseen Bgs.	Unseen Lights	Aggr.
OpenVLA (LoRA)	33.3%	16.7%	14.6%	4.2%	12.9%
OpenVLA-OFT (LoRA)	25.0%	29.2%	31.3%	31.3%	30.3%
RDT (Full FT)	25.0%	25.0%	31.3%	35.4%	31.1%
π_0 (LoRA)	58.3%	45.8%	14.6%	10.4%	22.7%
π_0 (Full FT)	75.0%	45.8%	20.8%	20.8%	30.3%
Ours	**91.7**%	**91.7**%	**89.6**%	**93.8**%	**91.7**%

(a) Dexterous grasping in cluttered scenes.

	Clear Table	Grasp Green	Grasp Bottles	Grasp Food	Aggr.
Task Success Rate	95.8%	87.5%	91.7%	83.3%	89.6%
Avg. Attempts per Grasp	1.09	1.14	1.09	1.19	1.12
Planner: Instruction Proposal	100.0%	92.6%	94.3%	88.1%	94.3%
Planner: BBox Accuracy	98.7%	98.2%	98.1%	98.3%	98.4%
Controller: Grasping	91.0%	92.6%	92.5%	91.5%	92.2%
Planner: Completion Check	98.7%	94.4%	96.2%	94.9%	96.3%

(b) Long-horizon task performance of DexGraspVLA.

	Unseen Objects	Unseen Bgs.	Unseen Lights	Aggr.
Ours@1	91.1%	90.5%	90.9%	90.8%
Ours@2	95.3%	94.2%	95.1%	94.7%
Ours@3	96.7%	96.7%	97.4%	96.9%

(c) Large-scale generalization evaluation of DexGraspVLA on dexterous grasping.

	Seen Objects	Unseen Objects	Aggr.
ViT-small	60.0%	35.0%	50.5%
DINOv2-train	30.0%	43.5%	34.8%
Ours	**98.5**%	**98.8**%	**98.6**%

(d) Ablation results on single-object grasping.

	Unseen Objects	Unseen Bgs.	Unseen Lights	Aggr.
ViT-small	61.1%	37.5%	22.2%	39.6%
DINOv2-train	66.7%	70.8%	55.6%	66.0%
Ours	**88.9**%	**86.1**%	**77.8**%	**84.7**%

(e) Nonprehensile grasping performance.

Table 1: Comprehensive evaluation of DexGraspVLA and baselines across tasks. Bgs.: Backgrounds; Aggr.: Aggregated.

5.3 Baseline Comparison

Tasks & Metrics. We adopt the same setup as Section 5.2 but on a smaller scale for baseline comparison. The tasks involve 24 unseen objects, 2 unseen backgrounds, and 2 unseen lighting conditions. We also include 12 seen objects under white background and lighting (*Seen Objects*). Metrics remain unchanged; details are in the appendix.

Results. As shown in Table 1a, DexGraspVLA consistently achieves 90+% success across all settings, significantly outperforming fine-tuned VLA models. While π_0 (Full FT) reaches 75% on seen objects, its performance drops sharply under visual variations. Similar declines are observed for π_0 (LoRA) and OpenVLA (LoRA), suggesting overfitting to training language and visual domains. Notably, RDT also uses frozen vision and language foundation models like ours and shows more consistent performance, but still falls short. This suggests that bounding boxes offer stronger grounding than language encoding, and that DINOv2 better preserves visual details than SigLIP (Zhai et al. 2023). Overall, these results validate the design of DexGraspVLA and its superior generalization performance.

5.4 Ablation Study

Tasks & Metrics. To compare DexGraspVLA with ablated variants that learn directly from raw visual inputs without frozen vision encoders, we conduct single-object grasping experiments using 13 seen and 8 unseen objects. Each object is tested at five table locations with two trials per location, yielding 210 tests under white tabletop and lighting. Success rates are computed as in Section 5.2.

Results. Table 1d shows that DexGraspVLA (Ours) consistently achieves over 98% success on both seen and unseen objects, significantly outperforming DINOv2-train and ViT-small variants. Its near-perfect performance in a zero-shot setting indicates strong robustness to domain shift. Interestingly, performance on unseen objects slightly exceeds that on seen ones, suggesting that the model learns the grasping task itself rather than overfitting to training data. In contrast, baselines that map raw inputs to actions fail to generalize, as perceptual changes easily push them out of distribution.

5.5 Internal Model Behavior Analysis

To further validate our design, we examine whether internal model behavior remains consistent under varying visual conditions, as shown in Figure 4. We test DexGraspVLA on the same cluttered scene (9 objects, target: "grasp the blue yogurt in the middle") across four environments: a white table, a calibration board, a colorful tablecloth, and the same tablecloth under disco lighting. For clarity, we display only the tabletop region; full images are in the appendix. While the head images (1st row) appear to be markedly diverse, the DINOv2 features (2nd row) look rather consistent. These features are visualized by mapping principal components to RGB channels as done in Oquab et al. (2023). Across environments, the object properties are robustly maintained and matched, which fundamentally allows DexGraspVLA trained on a single data domain to generalize. The third row shows that Cutie accurately tracks the object, providing the correct guidance to the controller. Based on the domain-invariant mask and the DINOv2 features, the DiT action head now predicts the subsequent actions. In the fourth row, we average and normalize all cross-attentions to the head image from DiT. We find that all attention maps exhibit the same behavior of focusing on the target object instead of being distracted by environments. The fifth row overlays the attention map on the raw image to confirm the reasonable attention pattern. All visualization details are provided in the appendix. Therefore, we substantiate that DexGraspVLA indeed transforms perceptually diverse raw inputs into invariant representations, on which it effectively applies imitation

learning to model the data distribution, explaining its superior generalization performance.

5.6 Long-Horizon Task Evaluation

Tasks. We evaluate DexGraspVLA on long-horizon grasping tasks. We use four types of prompts—"*Clear the table*", "*Grasp all bottles*", "*Grasp all green objects*", and "*Grasp all food*"—which require commonsense and physical reasoning to identify targets sequentially. Each prompt is evaluated in 24 randomly configured cluttered scenes. "*Clear the table*" scenes include three unseen objects; others involve 3–4 unseen objects, with two being relevant.

Metric. For each task, we report the task success rate as the proportion of tests that fully complete all required stages. We further report the average grasping attempts per object in the successful tests, along with success rates for instruction proposal, bounding box prediction, completion check of the planner, and grasp execution of the controller.

Results. Table 1b shows that DexGraspVLA achieves an 89.6% aggregated task success rate across four long-horizon prompts, with each target object attempted slightly more than once. The high-level planner grounds prompt semantics on the observation and proposes correct instructions with a 94.3% average success rate. Its bounding box prediction accuracy is consistently above 98%, which we further substantiate with evaluations in distraction conditions in the appendix. The low-level controller, leveraging its robust and generalizable grasping policy, executes individual grasps with over 91% success, enabling reliable multi-step completion. Additionally, the planner detects task completion with over 94% accuracy, preventing redundant actions. These results highlight the synergy between the high-level and low-level modules in DexGraspVLA, showcasing the effectiveness of its hierarchical framework for long-horizon tasks. An example can be found in the appendix.

5.7 Extension to Nonprehensile Grasping

Tasks & Metric. To show applicability beyond dexterous grasping, we apply DexGraspVLA to a nonprehensile grasping task (Figure 1 last row). We curate 32 flat, wide-surface objects (e.g., plates, boxes, and books) that are difficult to grasp directly and collect 1,029 human demonstrations in cluttered scenes. In these demos, the robot first performs a pre-grasp manipulation by pushing the object toward the table edge, creating an accessible pose, and then executes a final grasp. We keep the DexGraspVLA planner unchanged and train the controller on this dataset; details are provided in the appendix. To evaluate generalization, we curate 18 unseen nonprehensile objects and design three types of tasks: (1) *Unseen Objects* (36 tests): Each object is placed in two cluttered scenes with varying poses on a white table under white light. (2) *Unseen Lighting* (36 tests): Same protocol under disco light. (3) *Unseen Backgrounds* (72 tests): Same protocol on a wooden tabletop or a yellow tablecloth. Success rates are reported as in Section 5.2.

Results. As shown in Table 1e, DexGraspVLA achieves an aggregated generalization performance of 84.7% in the nonprehensile grasping task, showing strong robustness to unseen object appearances, shapes, physical properties, as well as novel backgrounds and lightings—significantly outperforming ablated variants. We observe that DexGraspVLA reliably adapts to object poses, pushing until it extends sufficiently over the edge, followed by a stable grasp. This task is particularly challenging for parallel-jaw grippers, highlighting the dexterity we exhibit. Moreover, DexGraspVLA seamlessly extends to this new task without architectural changes, reflecting three key aspects of generality: (1) the high-level planner's grounding and reasoning ability; (2) the use of bounding boxes as affordance guidance; and (3) applying imitation learning on domain-invariant representations iteratively obtained from foundation models.

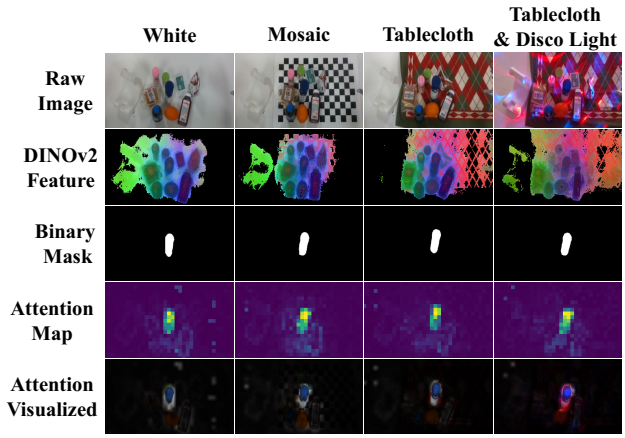

Figure 4: **DexGraspVLA is robust to environmental variations.** The same cluttered scene (1st row) is arranged in four visually different environments (four columns). DINOv2 features (2nd row), masks (3rd row), and attention maps (4th row) are consistent across variations. The 5th row confirms DexGraspVLA is attending to the correct object.

6 Limitation and Conclusion

This paper presents DexGraspVLA, a hierarchical VLA framework aiming for robust generalization in language-guided dexterous grasping and beyond. By leveraging a pre-trained VLM as the high-level planner and vision foundation models in the low-level controller, the system transforms multimodal inputs into domain-invariant representations and learns robust closed-loop policies via imitation learning. Our large-scale evaluations show over 90% grasping success across thousands of unseen cluttered scenes in a zero-shot setting, with empirical evidence of consistent internal behavior. DexGraspVLA also handles free-form long-horizon prompts, recovers from failures, and extends to nonprehensile grasping, demonstrating broad applicability. While effective, it does not yet address functional grasping and subsequent manipulation, nor does it incorporate tactile sensing. In future work, we aim to extend the high-level planner to generate more fine-grained affordance and learn a task-oriented manipulation controller that also integrates tactile feedback, further broadening the scope of DexGraspVLA.

References

Akkaya, I.; Andrychowicz, M.; Chociej, M.; Litwin, M.; McGrew, B.; Petron, A.; Paino, A.; Plappert, M.; Powell, G.; Ribas, R.; et al. 2019. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *arXiv preprint arXiv:2308.12966*.

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*.

Chen, S.; Bohg, J.; and Liu, C. K. 2024. SpringGrasp: An optimization pipeline for robust and compliant dexterous pre-grasp synthesis. *arXiv preprint arXiv:2404.13532*.

Chen, Y.; Wang, C.; Fei-Fei, L.; and Liu, C. K. 2023. Sequential dexterity: Chaining dexterous policies for long-horizon manipulation. In *Conference on Robot Learning*, 3809–3829.

Chen, Y.; Wu, T.; Wang, S.; Feng, X.; Jiang, J.; Lu, Z.; McAleer, S.; Dong, H.; Zhu, S.-C.; and Yang, Y. 2022. Towards human-level bimanual dexterous manipulation with reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 5150–5163.

Cheng, H. K.; Oh, S. W.; Price, B.; Lee, J.-Y.; and Schwing, A. 2024. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3151–3161.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*.

Fang, H.-S.; Yan, H.; Tang, Z.; Fang, H.; Wang, C.; and Lu, C. 2025. AnyDexGrasp: General Dexterous Grasping for Different Hands with Human-level Learning Efficiency. arXiv:2502.16420.

Guzey, I.; Dai, Y.; Evans, B.; Chintala, S.; and Pinto, L. 2024. See to touch: Learning tactile dexterity through visual incentives. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 13825–13832. IEEE.

Handa, A.; Allshire, A.; Makoviychuk, V.; Petrenko, A.; Singh, R.; Liu, J.; Makoviichuk, D.; Van Wyk, K.; Zhurkevich, A.; Sundaralingam, B.; et al. 2023. DeXtreme: Transfer of Agile In-hand Manipulation from Simulation to Reality. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 5977–5984.

Huang, B.; Chen, Y.; Wang, T.; Qin, Y.; Yang, Y.; Atanasov, N.; and Wang, X. 2023a. Dynamic handover: Throw and catch with bimanual hands. In *7th Annual Conference on Robot Learning*.

Huang, W.; Wang, C.; Li, Y.; Zhang, R.; and Fei-Fei, L. 2024. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *8th Annual Conference on Robot Learning*.

Huang, W.; Wang, C.; Zhang, R.; Li, Y.; Wu, J.; and Fei-Fei, L. 2023b. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning*, 540–562.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Intelligence, P. 2025. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. In *9th Annual Conference on Robot Learning*.

Kim, M. J.; Finn, C.; and Liang, P. 2025. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, Y.; Wei, W.; Li, D.; Wang, P.; Li, W.; and Zhong, J. 2022. HGC-Net: Deep anthropomorphic hand grasping in clutter. In *2022 International Conference on Robotics and Automation (ICRA)*, 714–720. IEEE.

Lin, T.; Yin, Z.-H.; Qi, H.; Abbeel, P.; and Malik, J. 2024a. Twisting lids off with two hands. *arXiv preprint arXiv:2403.02338*.

Lin, T.; Zhang, Y.; Li, Q.; Qi, H.; Yi, B.; Levine, S.; and Malik, J. 2024b. Learning Visuotactile Skills with Two Multifingered Hands. *arXiv preprint arXiv:2404.16823*.

Liu, M.; Pan, Z.; Xu, K.; Ganguly, K.; and Manocha, D. 2020. Deep differentiable grasp planner for high-dof grippers. In *Robotics: Science and Systems*.

Liu, S.; Wu, L.; Li, B.; Tan, H.; Chen, H.; Wang, Z.; Xu, K.; Su, H.; and Zhu, J. 2024. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*.

O'Neill, A.; Rehman, A.; Gupta, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandlekar, A.; et al. 2023. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Pan, M.; Zhang, J.; Wu, T.; Zhao, Y.; Gao, W.; and Dong, H. 2025. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17359–17369.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Pitz, J.; Röstel, L.; Sievers, L.; and Bäuml, B. 2023. Dextrous tactile in-hand manipulation using a modular reinforcement learning architecture. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 1852–1858. IEEE.

Qi, H.; Yi, B.; Suresh, S.; Lambeta, M.; Ma, Y.; Calandra, R.; and Malik, J. 2023. General in-hand object rotation with vision and touch. In *Conference on Robot Learning*, 2549–2564. PMLR.

Qin, Y.; Su, H.; and Wang, X. 2022. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters*, 7(4): 10873–10881.

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 *International conference on machine learning*, 8748–8763. PMLR.

Singh, R.; Allshire, A.; Handa, A.; Ratliff, N.; and Van Wyk, K. 2024. DextrAH-RGB: Visuomotor Policies to Grasp Anything with Dexterous Hands. *arXiv preprint arXiv:2412.01791*.

Stone, A.; Xiao, T.; Lu, Y.; Gopalakrishnan, K.; Lee, K.-H.; Vuong, Q.; Wohlhart, P.; Kirmani, S.; Zitkovich, B.; Xia, F.; et al. 2023. Open-World Object Manipulation using Pre-Trained Vision-Language Models. In *7th Annual Conference on Robot Learning*.

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent Correspondence from Image Diffusion. arXiv:2306.03881.

Team, Q. 2025. Qwen2.5-VL.

Turpin, D.; Wang, L.; Heiden, E.; Chen, Y.-C.; Macklin, M.; Tsogkas, S.; Dickinson, S.; and Garg, A. 2022. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *European Conference on Computer Vision*, 201–221. Springer.

Turpin, D.; Zhong, T.; Zhang, S.; Zhu, G.; Heiden, E.; Macklin, M.; Tsogkas, S.; Dickinson, S.; and Garg, A. 2023. Fast-Grasp'D: Dexterous Multi-finger Grasp Generation Through Differentiable Simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 8082–8089. IEEE.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3891–3902.

Wang, Q.; Zhang, H.; Deng, C.; You, Y.; Dong, H.; Zhu, Y.; and Guibas, L. 2023a. Sparsedff: Sparse-view feature distillation for one-shot dexterous manipulation. In *The Twelfth International Conference on Learning Representations*.

Wang, R.; Zhang, J.; Chen, J.; Xu, Y.; Li, P.; Liu, T.; and Wang, H. 2023b. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 11359–11366. IEEE.

Yang, M.; Lu, C.; Church, A.; Lin, Y.; Ford, C.; Li, H.; Psomopoulou, E.; Barton, D. A.; and Lepora, N. F. 2024. AnyRotate: Gravity-Invariant In-Hand Object Rotation with Sim-to-Real Touch. *arXiv preprint arXiv:2405.07391*.

Zakka, K.; Smith, L.; Gileadi, N.; Howell, T.; Peng, X. B.; Singh, S.; Tassa, Y.; Florence, P.; Zeng, A.; and Abbeel, P. 2023. RoboPianist: A Benchmark for High-Dimensional Robot Control. *arXiv preprint arXiv:2304.04150*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

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

Zhang, J.; Liu, H.; Li, D.; Yu, X.; Geng, H.; Ding, Y.; Chen, J.; and Wang, H. 2024b. DexGraspNet 2.0: Learning Generative Dexterous Grasping in Large-scale Synthetic Cluttered Scenes. In *8th Annual Conference on Robot Learning*.

Zhong, Y.; Bai, F.; Cai, S.; Huang, X.; Chen, Z.; Zhang, X.; Wang, Y.; Guo, S.; Guan, T.; Lui, K. N.; et al. 2025. A Survey on Vision-Language-Action Models: An Action Tokenization Perspective. *arXiv preprint arXiv:2507.01925*.

Zhou, W.; and Held, D. 2023. Learning to grasp the ungraspable with emergent extrinsic dexterity. In *Conference on Robot Learning*, 150–160. PMLR.