Gripper-aware Vision Language Action Models

Anonymous ECCV 2026 Submission

Paper ID #5218

Fig. 1: We tackle the challenge of gripper-specific grasping in robotics, e.g., a suction cup gripper lifts a thin, flat, DVD-case-like box from above, while a parallel-jaw gripper pushes it to the table edge and grasps it from the side. We first introduce MiGA, a large-scale multi-gripper-aware dataset collected by diverse gripper types. Leveraging this dataset, we then develop a Gripper-aware Vision Language Action (GVLA) model that incorporates gripper conditioning through a new multi-gripper tokenization method and a dual Mixture-of-Adapter. Our approach improves task performance over gripper-agnostic baselines and enables cross-gripper adaptation.

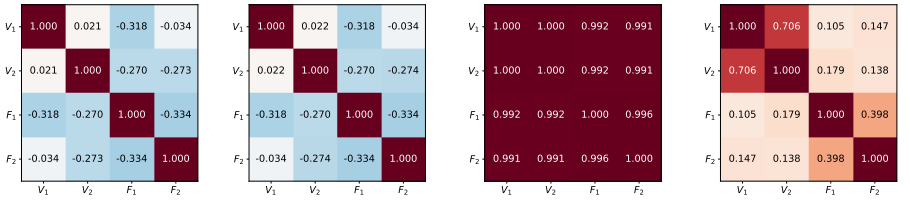
Abstract. Vision language action models (VLAs) have advanced general purpose robotic grasping and manipulation by enabling robots to interpret visual observations and natural language instructions to generate executable action sequences. However, existing VLAs often implicitly assume gripper invariance, despite grasping strategies being inherently embodiment-dependent. Different gripper types, such as parallel-jaw and suction, usually require distinct interaction strategies to achieve the same grasping objective. Moreover, current datasets for VLAs predominantly rely on parallel-jaw grippers, limiting gripper-aware learning. To address this gap, we introduce MiGA, a multi-gripper-aware dataset spanning five distinct gripper types across multiple robots with 103,000 demonstrations, explicitly capturing strategy divergence under shared task objectives. We further propose GVLA, which combines a new multi-gripper tokenizer with adapter-based policy routing. Our new gripper encoding induces structured embedding information that balances parameter sharing and strategy differentiation, while layer-wise probing confirms meaningful gripper-conditioned representations for VLAs. Intensive experiments in both simulation and real-world robots show that our GVLA outperforms the current baselines by a clear margin. Our method also improves zero-shot generalization or few-shot adaptation to new objects or unseen tasks, and enabling cross-gripper transfer.

1 Introduction

General-purpose robotic grasping and manipulation with vision-language-action models (VLAs) have been making a transformative impact on the robotics research community recently [10, 41, 50, 53]. Although achieving promising results, current VLAs implicitly assume gripper invariance across the learning tasks, while gripper configurations and grasping strategies are inherently embodiment-dependent [31, 70, 72]. Different types of grippers necessitate fundamentally distinct manipulation and grasping strategies. For example, as shown in Fig. 1 and our demonstration video, grasping a thin, flat object (*i.e.*, a DVD or a facing-down box) with parallel-jaw grippers may require first repositioning the object by sliding it to the table edge and then grasping it from the side; meanwhile, a suction gripper can directly approach from above and lift it. This example exemplifies a critical insight: achieving the same task objective does not imply a shared strategy space, where successful execution depends on the robot's embodiment and its corresponding feasible interactions. Moreover, incorporating robot hardware configurations into the learning of task space under different task contexts requires VLAs to learn distinct strategies for different gripper embodiments. These expectations for VLAs pose challenges in both dataset collection and the design of learning mechanisms. Herein, we pose the central question: "*Can VLAs learn embodiment-dependence and strategy-level divergence of multiple robotic grippers to tackle diverse, real-world tasks?*"

To enable VLAs training, an important requirement is the availability of large-scale datasets. Current robotic datasets, such as Open X-Embodiment [42], Bridge V2 [54], and DROID [26] dominantly use parallel-jaw grippers, leaving gripper diversity underexplored [60]. This is increasingly limiting as different grippers such as suction cups, multi-finger hands, and soft grippers are also being used in various robotic tasks [31, 70, 72]. Moreover, different gripper types often require their own manipulation strategies, while existing datasets do not explicitly encode such embodiment-dependent strategy divergence, preventing VLAs from learning gripper-aware policies. To address this gap, we introduce **MiGA**, a large-scale multi-gripper-aware dataset comprising 103,000 demonstrations spanning over various tasks with five distinct gripper types. Unlike prior datasets that provide single-solution trajectories with limited gripper types, MiGA explicitly captures how identical task objectives necessitate fundamentally different strategies across different gripper embodiments. Each task includes demonstrations from at least three different gripper types, yielding 132 distinct gripper-strategy pairs with natural language descriptions. Collected across multiple robot platforms in both simulation and real-world environments, MiGA provides the first comprehensive dataset for training and benchmarking gripper-aware policies that can reason about gripper-dependent grasping.

Recently, several works have investigated gripper representation for VLAs in robotic applications [15, 64, 68]. However, existing gripper representations, such as graph-based encoding [24, 64, 67], primarily target grasp pose transfer among morphologically similar grippers rather than learning robust griper tokenization. To investigate whether existing gripper representations might suffice, we visu-



(a) MLP **(b)** VQ-VAE **(c)** Language Prompt **(d)** Ours

Fig. 2: Gripper embedding feature heatmap of different gripper representations. Our new multi-gripper tokenization method (d) produces well-clustered representations organized by gripper type. In contrast, other methods (a-c) fail to capture the underlying semantic structure of different gripper types. V_1 and V_2 denote the vacuum gripper: UR10 Suction Cup [23] and Cobot Pump [48], F_1, and F_2 denote finger-based gripper: Franka Panda Hand [16] and Shadow Hand [4].

alize three gripper tokenization approaches when using the same π_0 [5] backbone: MLP embeddings [47], VQ-VAE tokenization [51], and language-based embeding [29]. Fig. 2(a–c) shows that existing methods struggle to produce meaningful gripper structure: MLP and VQ-VAE produce inconsistent similarities unrelated to morphology, while language prompts collapse to near-identical representations across grippers. To effectively embed gripper information into VLAs, we introduce two new components: (*i*) a new multi-gripper tokenization that encodes gripper information at three levels: domain-level tokens capturing robot characteristics, type-level tokens encoding gripper category constraints, and instance-level tokens preserving individual gripper features. Fig. 2d shows that our multi-gripper tokenization produces structured representations in the embedding space, with same-type grippers clustering while remaining well separated from different types; and (*ii*) a dual Mixture-of-Adapters (MoA) that directly modulates action generation through gripper-specific expert routing, enabling parameter-efficient fine-tuning while maintaining knowledge sharing across grippers. The intensive experiments on both simulation and real-world robots show that by effectively embedding the gripper information, our method outperforms recent baselines by a clear margin. To summarize, our contributions are as follows:

(1) We introduce MiGA, a large-scale multi-gripper-aware dataset featuring diverse gripper types on complex tasks, explicitly capturing how the same task requires different strategies depending on the gripper morphology.
(2) We propose GVLA, a fine-tuning framework with a new multi-gripper tokenizer and MoA to enable strategy-aware manipulation learning.

2 Related Work

Gripper Representation. Several robotic research work has recognized that different end-effectors require distinct approaches [7,9,20,49]. Existing approaches encode gripper morphology through geometric distance fields between robot and object point clouds [24, 59, 64], diffusion-based synthesis [17, 63], graph-based approaches operate on kinematic topology [2, 44, 58], and contact-centric representations decouple hand embodiment from object geometry [15, 30, 68]. Beyond

Table 1: Comparison of existing datasets for VLAs. Our dataset provides a manually collected, gripper-aware data spanning multiple gripper types with gripper-specific manipulation strategies in both simulation and real-world settings.

Dataset	#Traj.	#Gripper Types	#Robots	Gripper-Specific Solution	Depth Image	Sub-step Decomposition	Data Domain	Collection
Open X-Embodiment [42]	1M+	1	22	✗	✗	✗	Real	Aggregation
DROID [26]	76K	1	1	✗	✗	✗	Real	Human
Bridge V2 [54]	60K	1	1	✗	✗	✗	Real	Human
RT-1 [6]	130K	1	1	✗	✗	✗	Real	Human
RH20T [12]	110K	1	7	✗	✓	✗	Real	Human
BridgeDataV2 [12]	60.1K	1	1	✗	✓	✗	Real	84% Human, 16% Scripted
RoboMind [62]	107K	2	4	✗	✓	✓	Real	Human
GraspVLA [10]	1B	1	1	✗	✗	✗	Sim	Synthetic
Libero [34]	4.5K	1	1	✗	✗	✗	Sim	Human
MiGA (Ours)	**103K**	**5**	**5**	✓	✓	✓	**Real+Sim**	**Human**

these, a subset of works further explores cross-gripper transfer via shared eigengrasp spaces [67] and latent action alignment [3]. Despite these advances, existing methods primarily focus on transferring grasp pose representations across grippers, while overlooking strategy-level differences induced by gripper morphology [2, 67]. Incorporating gripper-aware reasoning into VLAs, where strategies span from approach trajectories to final execution, remains underexplored.

VLAs and Datasets. VLAs leverage pretrained vision-language models to enable reasoning and generalization in robotic tasks. Early work like RT series [6, 71, 76] demonstrated this potential, while recent efforts such as OpenVLA [27, 28] and π series [5, 21, 22] have further improved scalability and architectural design. Recent work has demonstrated strong performance in grasping tasks [10, 70, 75], with efforts extending to vacuum grippers [72] and dexterous hands [8, 31, 43, 70]. However, VLA models strongly rely on training data. As shown in Table 1, existing large-scale robot datasets predominantly feature parallel-jaw grippers [6, 12, 26, 42, 54], limiting VLA models' ability to learn morphology-dependent manipulation strategies [7, 69]. To address this limitation, we introduce a multi-gripper dataset across five different gripper types. Unlike prior datasets, our dataset captures trajectory-level strategy variation across grippers, providing the foundation for gripper-aware policy learning.

Soft Prompt Learning. Soft prompting provides a parameter-efficient alternative to full fine-tuning by optimizing continuous prompt embeddings while freezing backbone parameters [32, 36, 37, 73]. In vision-language models (VLMs), prompt learning has shown strong few-shot adaptation performance [18, 73, 74], with extensions to dynamic routing [11], multi-modal prompts [25], and knowledge preservation [65]. Hierarchical prompt tuning [57] further models structured semantic associations across multiple levels via relationship-guided attention,modeling both fine-grained attributes and holistic category semantics. In robotics, prompt-based adaptation has been explored for embodiment generalization; X-VLA [69] introduces per-embodiment soft prompts to absorb hardware variations across platforms. However, prior work focuses on robot-level diversity and overlooks end-effector morphology. We address this gap with gripper-aware soft prompt tokenization that encodes gripper taxonomy, embedding morphological priors directly into prompt space, and combine it with gripper-conditioned mixture-of-adapters [56, 66] to enable structured sharing and specialization.

3 The MiGA Dataset

Although several datasets have been proposed for VLA, most of them are limited to one gripper type, typically parallel-jaw grippers, and therefore overlook alternative grippers and how they influence the grasping strategy (Table 1). To address this limitation, we introduce MiGA, a multi-gripper-aware dataset that explicitly captures the coupling between gripper morphology and grasping strategy. MiGA has three key properties: (*i*) Multi-gripper coverage: demonstrations collected with five common gripper types; (*ii*) Gripper-specific task design: A broad set of tasks intentionally constructed to elicit morphology-dependent solutions, highlighting differences in contact formation, approach planning, and manipulation execution; and (*iii*) Real-to-sim parity: real-world demonstrations accompanied by closely matched simulation environments that enable reproducible benchmarking across models. Beyond trajectories, MiGA provides sub-step decompositions and annotated failure demonstrations, offering rich supervision for learning gripper capabilities and challenging manipulation tasks.

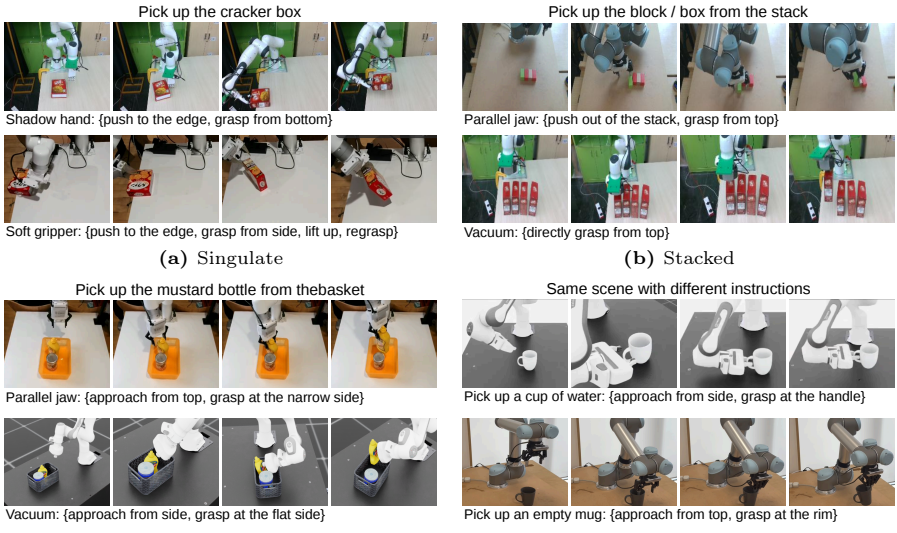
3.1 Data Collection Setup

Gripper Setup. We employ five gripper types for data collection: (*i*) parallel-jaw grippers (Franka Panda [16], Robotiq 2f-85 [45]) that achieve two-finger form-closure; (*ii*) three-finger grippers (Robotiq 3-Finger [46]) that provide multi-point form-closure; (*iii*) our in-house soft two finger gripper that provides compliant adaptation; (*iv*) suction gripper (Cobot Pump [48], UR10 Suction Cup [23]) that adhesion-based grasping with suction; and (*v*) dexterous five fingers hand (Inspire Hand [4]) that provide high degree-of-freedom (DoF) multi-contact. These grippers differ not only in geometry but also in contact mechanics, force transmission, and controllable DoF. As a result, identical manipulation tasks require qualitatively distinct strategies across grippers, including variations in approach direction, contact region selection, and pre-grasp configuration. With several gripper types, our dataset enables learning not only trajectory policies, but the structured relationship between grippers and grasping strategies.

Robot Setup. MiGA includes demonstrations collected in both simulation and real-world settings with five gripper types. Simulation provides a reproducible testbed for algorithm development, while real-world data captures the real physical complexities. For simulation, we use NVIDIA Isaac Lab [38] with Franka Panda and UR10 robots. Real-world demonstrations are collected using UFACTORY xArm7, Franka Panda, and UR5 robots. All setups use multi-view RGB-D observations from wrist-mounted cameras and third-view cameras, providing complementary end-effector and global scene perspectives.

3.2 Task Design

To explore how gripper morphology drives grasping strategy, we design tasks in which identical objectives require different solutions across grippers. The task suite spans four scenarios: (*i*) Singulated tasks vary object geometry, texture, and pose, revealing how contact feasibility depends on morphology, e.g., flat surfaces favor adhesion-based grippers while irregular geometries require compliance or multi-point contact. (*ii*) Stacked scenes introduce occlusion and collision

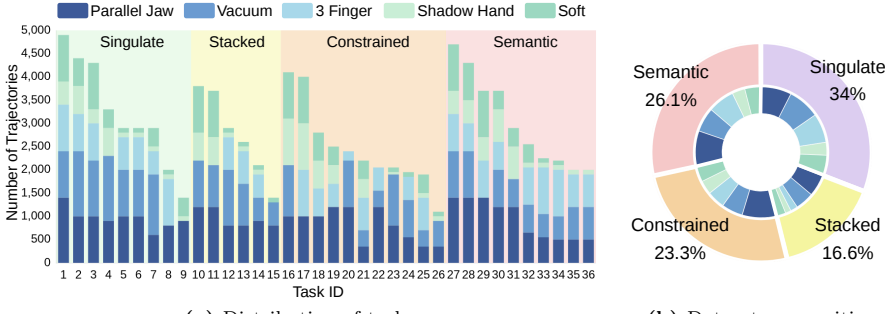


(a) Singulate **(b)** Stacked

(c) Constrained **(d)** Semantic

Fig. 3: Illustration of gripper-specific strategy variations across four task categories.

constraints that demand morphology-specific pre-grasp planning and insertion strategies. (*iii*) Constrained-space tasks impose spatial limitations that amplify trade-offs between gripper size, reachability, and alignment precision. (*iv*) Semantic grasping further requires reasoning about object function and task context, such as selecting safe contact regions when handling a filled container. Our Supplementary Material provides a detailed description of these tasks.



(a) Distribution of tasks **(b)** Dataset composition

Fig. 4: MiGA dataset statistics.

3.3 MiGA Dataset Statistics

Fig. 4 summarizes the statistics of our dataset. MiGA comprises 103, 000 demonstrations spanning 36 tasks and 5 distinct gripper types, with multi-view RGB-D observations, proprioceptive states, and gripper-strategy annotations. Each task includes demonstrations from at least 3 distinct gripper types, ensuring multi-strategy supervision under identical task objectives. In addition, MiGA provides natural language descriptions for each gripper-task pair. The dataset additionally includes failure demonstrations (around 5% of the dataset), enabling analysis of embodiment limitations and failure boundaries.

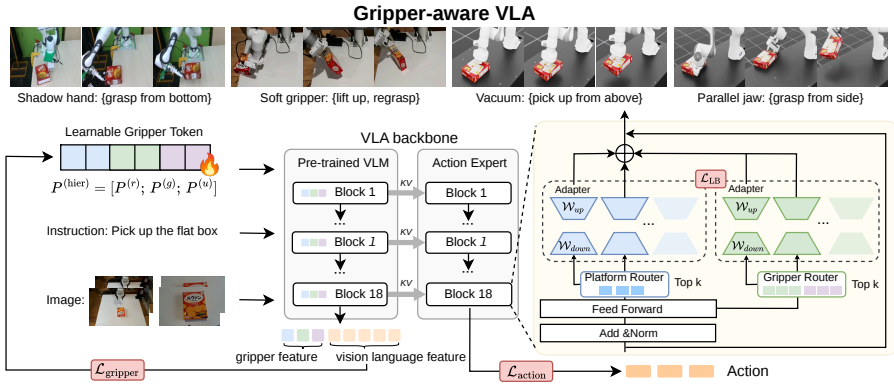


Fig. 5: An overview of GVLA architecture. Our method extends the pretrained VLA backbone with two gripper-aware components. (Left) Multi-gripper tokenizer encodes gripper knowledge through three-level soft prompts: platform-specific, gripper type-specific, and instance-specific learnable tokens. (Right) A dual MoA routes action tokens through type-stratified expert pools using a cascaded domain and a gripper router, applying top-k selected adapters as residual connections to adapt to gripper-specific manipulation strategies.

3.4 How will MiGA be useful to the community?

With its large-scale and multi-gripper nature, MiGA can be used in several research tasks. Here we highlight three key tasks that can be useful from our dataset: (*i*) Complex grasping with trajectory-level planning: Prior work primarily targets grasp pose prediction [13,14,61], overlooking the grasping complexity on the whole grasping process [19,39], MiGA provides full trajectories, sub-action decomposition, enabling learning of both execution policies and high-level planning. (*ii*) Cross-gripper learning: Existing cross-gripper studies focus on grasp pose transfer between multi-finger grippers [3, 17, 24]. MiGA supports strategy comparison across different grippers, enabling morphology-dependent strategy learning beyond geometric adaptation. (*iii*) VLA benchmarking: Most VLAs are trained on parallel-jaw dominant data [6, 42, 55], MiGA enables training VLAs with diverse grippers and provides a benchmark for evaluating cross-gripper generalization and gripper-conditioned policy learning.

4 Gripper-aware Vision Language Action Model

To demonstrate the importance of gripper-aware learning in robotics and the usefulness of our MiGA dataset, we introduce GVLA, a framework that aims to include gripper information into existing VLA models. We first introduce a multi-gripper tokenizer to embed the gripper information, allowing the VLA backbone to align high-level reasoning with gripper morphology and constraints while preserving shared representations across platforms. Second, we introduce execution-level control via a dual MoA, which is routed based on gripper tokens, enabling parameter-efficient fine-tuning. GVLA is designed to be embedded into different existing VLA backbones. In practice, we choose the $\pi_{0.5}$ backbone as it shows the competitive accuracy. Fig. 5 shows an overview of our method.

4.1 Gripper Representation

Given observation $\mathcal{O}_t$, to condition the observation with gripper configuration. we introduce a multi-granularity representation scheme that factorises embodiment conditioning across three levels of granularity. At time t, the observation $\mathcal{O}_t$ is embedded as $X \in \mathbb{R}^{n_{\text{obs}} \times d}$, where n_{obs} denotes the observation token length and d denotes the transformer embedding dimension. Instead of encoding embodiment variation using fixed text prompts or general encoders such as MLPs, we employ soft prompts: learnable embeddings that are randomly initialized and optimized end-to-end through gradient-based training. Jointly optimized with the backbone under gripper prediction and action losses, these prompts learn a latent mapping $\Phi : \mathcal{H} \to \mathbb{R}^{p \times d}$ from gripper hardware configurations to a continuous prompt space, where $P^{(h)} \approx \Phi(h)$ for gripper configuration h.

To represent gripper characteristics, we decompose embodiment information into three independent sets of learnable embeddings at different granularity levels. Platform token $P^{(r)} \in \mathbb{R}^{p_r \times d}$, indexed by robot platform $r \in \mathcal{R}$, capture platform-specific kinematic structure and configuration. Gripper-type token $P^{(g)} \in \mathbb{R}^{p_g \times d}$, indexed by end-effector category $g \in \mathcal{G}$, encodes mechanism-specific manipulation priors shared across all grippers of the same type. Instance-level token $P^{(u)} \in \mathbb{R}^{p_u \times d}$, indexed by gripper instance $u \in \mathcal{U}$, model fine-grained characteristics unique to individual grippers. The prompts are concatenated using a predefined three-level order:

$$P^{(\text{h})} = [P^{(r)};\, P^{(g)};\, P^{(u)}] \in \mathbb{R}^{(p_r+p_g+p_u)\times d}, \tag{1}$$

and prepended to the embedded observation to form the conditioned input:

$$\tilde{X}^{(r,g,u)} = [P^{(\text{h})};\, X] \in \mathbb{R}^{(p_r+p_g+p_u+n_{\text{obs}})\times d}. \tag{2}$$

This multi-granularity factorization encodes gripper information at three levels of specificity, capturing shared knowledge across different gripper configurations at each level, which facilitates efficient adaptation to new gripper instances.

4.2 Gripper-Aware Mixture of Adapters

While our multi-gripper tokenizer provides high-level conditioning, effective execution requires modulating action generation. We introduce a dual MoA mechanism that decomposes gripper-aware adaptation into platform-specific and gripper-specific modulation.

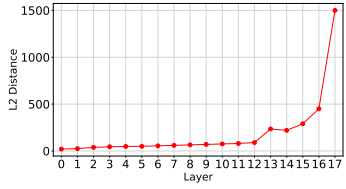


Fig. 6: Layer probing analysis.

Each MoA computes a gating function over the pooled conditioning tokens:

$$G(P) = \text{Softmax}(\text{TopK}(\text{MLP}(\text{mean}(P)))). \tag{3}$$

We instantiate two parallel routers: (i) a platform-aware gate $G^{(p)}$ driven by $\texttt{mean}(P^{(r)})$ and (ii) a gripper-aware gate $G^{(r)}$ driven by $\texttt{mean}([P^{(g)};\, P^{(u)}])$, each returning normalised weights and selected expert indices.

Each expert implements a bottleneck transformation:

$$\mathcal{A} = \mathcal{W}^{\text{up}}(\text{GeLU}(\mathcal{W}^{\text{down}}(\mathbf{x}))), \tag{4}$$

where $\mathcal{W}^{\text{down}} \in \mathbb{R}^{d \times d_b}$, $\mathcal{W}^{\text{up}} \in \mathbb{R}^{d_b \times d}$ with $d_b \ll d$. Given layer activation $\mathbf{x} \in \mathbb{R}^{B \times S \times d}$, the final output combines both platform and gripper adaptations additively:

$$\mathbf{x} \leftarrow \mathbf{x} + \sum_{i=1}^{k} G_i^{(p)} \cdot \mathcal{A}_i^{(p)} + \sum_{j=1}^{k} G_j^{(g)} \cdot \mathcal{A}_j^{(g)}. \tag{5}$$

Where to Insert MoA? To insert MoA into the VLA backbone, we analyze how gripper features influence hidden representations by measuring gripper type sensitivity of the action expert layers in the VLA backbone. This is defined as:

$$S_{\text{type}} = \frac{1}{|\mathcal{G}|^2} \sum_{\substack{g_1, g_2 \in \mathcal{G} \\ g_1 \neq g_2}} \left\| \text{mean}_{u \in \mathcal{U}_{g_1}} h_l^u - \text{mean}_{u \in \mathcal{U}_{g_2}} h_l^u \right\|_2, \tag{6}$$

where h_l^u denotes the action hidden representation at layer l for gripper u. Fig 6 shows that the type sensitivity remains low in several early layers but increases sharply in the final layer. We therefore insert MoA at the final layer to directly modulate action generation via multi-gripper-aware conditioned routing. More studies validating this design are provided in our Supplementary Material.

4.3 Fine-tuning Objective

To jointly learn gripper-aware action generation and gripper-consistent representations, we optimize GVLA with the following objective:

$$\mathcal{L} = \mathcal{L}_{\text{action}} + \lambda_{\text{gripper}} \mathcal{L}_{\text{gripper}} + \lambda_{\text{LB}} \mathcal{L}_{\text{LB}}, \tag{7}$$

Action Loss. Following [5], we supervise action tokens using a conditional flow matching loss [33, 35]:

$$\mathcal{L}_{\text{action}} = \mathbb{E} \left\| v_\theta(A_t^\tau, \mathcal{O}_t) - u(A_t^\tau | A_t) \right\|^2, \tag{8}$$

where the network learns to predict the vector field $u(A_t^\tau | A_t) = \boldsymbol{\epsilon} - A_t$ that transports noisy actions $A_t^\tau = \tau A_t + (1 - \tau)\boldsymbol{\epsilon}$ back to clean actions A_t, with noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathrm{I})$ and flow timestep $\tau \in [0, 1]$.

Gripper Prediction Loss. To encourage gripper-discriminative features in shared vision-language embeddings, we add an auxiliary classification loss. After gripper-conditioned encoding, the pooled visual observation representation $\mathcal{O}_t$ are used to predict the gripper type:

$$\mathcal{L}_{\text{gripper}} = -\log \frac{\exp\left(f_\phi^{(y_g)}(\mathcal{O}_t)\right)}{\sum_{k=1}^{|\mathcal{G}|} \exp\left(f_\phi^{(k)}(\mathcal{O}_t)\right)} \tag{9}$$

where f_ϕ is a linear classifier and $y_g \in \mathcal{G}$ the ground-truth label. This loss integrates gripper-relevant information into the shared embeddings, supporting gripper-sensitive reasoning and stabilizing downstream routing.

Load Balance Loss. To prevent router collapse and ensure uniform adapter utilization across the type-stratified pools, we introduce a load balance regularization. For each adapter pool, we measure the variance of adapter usage distribution and penalize imbalanced activation patterns:

$$\mathcal{L}_{\text{LB}} = \frac{1}{|\mathcal{R}|} \sum_{p \in \mathcal{R}} \text{Var}(\alpha_p) + \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \text{Var}(\alpha_g), \tag{10}$$

where $\alpha_p = [\alpha_p^1, \ldots, \alpha_p^{\mathcal{A}_p}]$ denotes the mean activation frequency of the $\mathcal{A}_p$ adapters in platform pool p, and $\alpha_g = [\alpha_g^1, \ldots, \alpha_g^{\mathcal{A}_g}]$ represents the usage distribution for gripper type pool g. This objective encourages the routers to distribute load evenly within each pool, preventing underutilization of specialized adapters while maintaining the hierarchical routing structure.

5 Experiments

We conduct experiments to investigate: (*i*) Overall Performance: How does gripper-aware conditioning improve manipulation success compared to baselines across task categories? (*ii*) Gripper-Aware Transfer: Does our model enable efficient zero-shot generalization or few-shot adaptation? (*iii*) Ablation Study and Robot Validation: How do our multi-gripper tokenizer and MoA contribute to performance, and how does GVLA perform in real-robot experiments?

Baselines. To validate the effectiveness of our method, we benchmark representative models from both traditional and VLA-based methods: For two-stage open-loop grasping, we include AnyGrasp [13], a state-of-the-art grasp detector trained on large-scale data, and GraspMAS [40], a recent multi-agent approach for language-driven grasp detection. For VLA-based methods, we include GraspVLA [10], which leverages large-scale grasping data to enable zero-shot generalization across diverse scenes, and OpenVLA-OFT [27], an optimized variant of OpenVLA [28]. We also implement two variants based on the π_0 architecture [5]: the original π_0 and the improved $\pi_{0.5}$, and evaluate their performance both as vanilla baselines and with our proposed gripper-aware extensions.

Evaluation Metrics. We use five metrics to evaluate the results: (*i*) Success Rate (SR) is defined as $\text{SR} = \frac{1}{N}\sum_{i=1}^{N} S_i$; (*ii*) The Prediction Error (PE) measures the average deviation between predicted and ground-truth actions over the action horizon H: $\text{PE} = \frac{1}{H}\sum_{h=1}^{H} |a_{\text{pred}}^{(h)} - a_{\text{gt}}^{(h)}|$; (*iii*) To measure behavioral specialization, we use the Counterfactual Action Prediction Divergence (CAPD) [52], which quantifies how much predicted actions change when only the gripper identity is modified while all other inputs remain fixed. Higher CAPD indicates gripper-specific behavior; (*iv*) To understand how gripper information is encoded throughout the network, we measure Linear Probe Accuracy (LPA) [1], which trains a logistic regression classifier $\phi^{(l)}$ on hidden states at each transformer layer l; trained on in-distribution data and evaluated on unseen gripper

Table 2: Baseline comparison across task categories.

Method	Flat	Stacked	Constrained	Semantic	Avg.(%)
AnyGrasp [13]	0.00	0.00	46.00	0.00	11.50
GraspMAS [40]	0.00	0.00	40.00	8.00	12.00
GraspVLA [10]	1.50	0.00	30.00	0.00	7.88
OpenVLA-OFT [27]	41.00	52.50	24.00	50.00	41.88
π_0 [5]	30.00	21.50	36.00	65.00	38.13
$\pi_{0.5}$ [22]	52.50	71.00	57.50	52.50	58.38
GVLA (π_0 backbone) (Ours)	27.50	45.00	50.00	70.00	48.13
GVLA ($\pi_{0.5}$ backbone) (Ours)	**53.00**	**76.00**	**62.50**	**72.50**	**66.00**

(see supplementary for full definitions); and (v) To quantify the contribution of the gripper token, we further introduce the Gripper Contribution Score (GCS): $\text{GCS} = \text{LPA}^{(l)}(\tilde{g} = g_i) - \text{LPA}^{(l)}(\tilde{g} = g_{\text{fixed}})$. A large GCS indicates that the soft prompt contributes to gripper discrimination beyond visual features alone.

5.1 Main Result

Table 2 compares our GVLA, fine-tuned on our MiGA dataset from π_0 and $\pi_{0.5}$ backbones against different baselines across four task categories across different gripper types in simulation. This table shows that traditional two-stage methods (AnyGrasp [13], GraspMAS [40]) collapse entirely on flat and stacked scenes, exposing their limited capacity to generalize beyond simple grasp configurations. Despite large-scale pre-training, GraspVLA [10] encounters similar degradation, highlighting the importance of our dataset design, which encodes diverse gripper-dependent grasping strategies instead of relying on a single downward grasping prior. Among VLA-based baselines, $\pi_{0.5}$ achieves the highest average performance (58.38%), yet remains limited. Leveraging this backbone, GVLA improves performance by 7.62% and surpasses all competing approaches , suggesting that gripper-aware conditioning enables the policy to capture gripper-specific manipulation strategies in scenarios where the robot behaviors differ substantially.

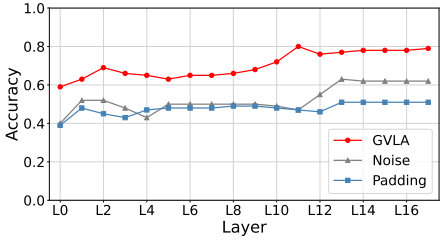


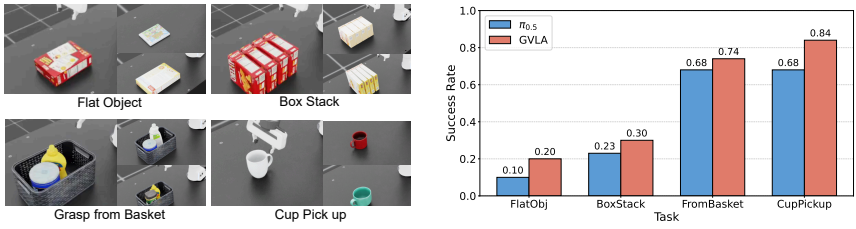
Fig. 7: Linear accuracy result.

Table 3: Tokenizer comparison.

Method	PE ↓	CAPD ↑	GCS ↑
MLP [47]	0.053	0.82	0.014
VQ-VAE [51]	0.051	0.70	0.002
LP [29]	0.053	0.64	0.225
GVLA (Ours)	**0.032**	**1.34**	**0.249**

5.2 Gripper-Aware Analysis

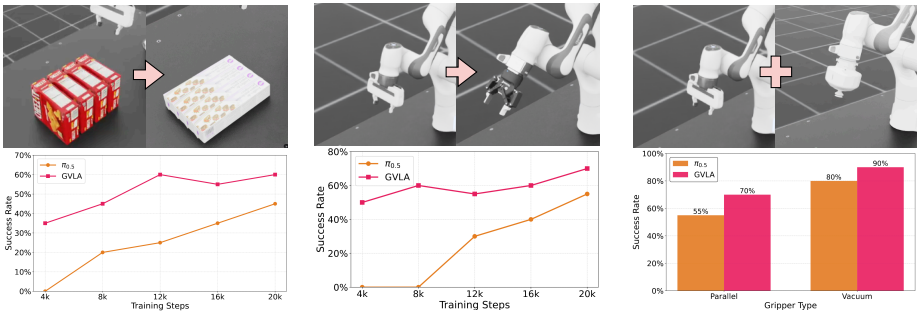
Does Gripper Conditioning Steer Specialized Pathways? Fig. 7 presents linear probe accuracy across network depth for three conditioning strategies. The "Noise" token baseline replaces gripper tokens with noise (hence, in this case, the backbone only relies on the visual and language tokens for learning) to test whether the input tokens suffice for effectively learning the tasks. The "Padding" token baseline uses a fixed, gripper-agnostic token to test whether



(a) Cross-object generalization setup. **(b)** Cross-object generalization success rate.

Fig. 8: Cross-object generalization results.

improvements arise from explicit conditioning rather than implicit visual inference. Both baselines remain low in accuracy, demonstrating that neither arbitrary tokens nor visual cues alone provide sufficient embodiment information. In contrast, our GVLA with multi-gripper tokenizer exhibits consistently elevated accuracy, rising from 59% at layer L0 to 80% at L12, and maintaining high separability in late layers. This persistent gripper-discriminative signal indicates that our multi-gripper tokenization enables the model to encode and propagate gripper-specific features throughout the network through explicit conditioning, rather than just relying on incidental visual patterns or treating the gripper information as auxiliary metadata. Table 3 further supports this observation: GVLA achieves the lowest prediction error and the highest behavioral divergence, indicative of learning effective griper-specific behavior. The combination of low PE and high CAPD demonstrates that our method improves both predictive accuracy and policy specialization, enabling GVLA to discover manipulation strategies aligned with each gripper's affordances.

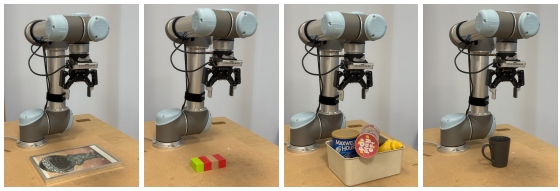


(a) Task adaptation. **(b)** Gripper adaptation **(c)** Mix data adaptation

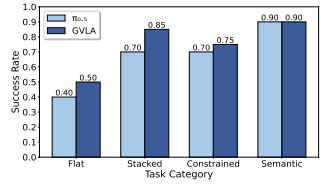
Fig. 9: Adaptation results.

Cross-object Generalization. To evaluate zero-shot generalization, we construct task variants by introducing new objects that share similar functional attributes with the training set and evaluate performance on unseen grippers. This setting measures the model's ability to generalize to new grasping scenarios without any task-specific fine-tuning. Fig. 8a shows the setup of this experiment. As reported in Fig. 8b, GVLA maintains strong performance across unseen tasks and grippers, demonstrating robust zero-shot transfer capability.

Gripper-Aware Adaptation. To evaluate whether our model can effectively disentangle gripper-specific representations and enable knowledge transfer across



(a) Real robot experiment setup.



(b) Real robot experiment results.

Fig. 10: Real-world robotic experiment setups and results.

grippers, we conduct few-shot adaptation experiments under three settings: (*i*) adapting to a new task with the same gripper, (*ii*) adapting to a new task with a previously unseen gripper, and (*iii*) adapting to a new task using a mixture of data from two different grippers, which are jointly used for training. Our Supplementary Material provides the detailed experimental setup. Fig. 9 shows that our model exhibits faster adaptation on new tasks involving the same type of gripper, as these tasks share a common adapter and underlying knowledge representation. When trained with mixed-gripper data, the model achieves further performance improvements, as jointly learning from multiple grippers reduces ambiguity and confusion between different gripper types.

Component Analysis. Table 4 evaluates the contribution of each component in GVLA. We report prediction error (PE) on the pre-trained MiGA dataset and adaptation success rate (Adapt.) when transferring to an unseen gripper (Robotiq 2F-85) after 20K fine-tuning steps.

Table 4: Component analysis.

	Configuration	PE↓	Adapt.↑
Gripper Repr.	w/o $P^{(g)}$	0.037	0.86
	w/o $P^{(p)}$	0.035	0.54
	w/o $P^{(u)}$	0.032	0.90
Dual MoA	w/o MoA	0.037	0.52
	w/o MoA$^{(g)}$	0.033	0.56
	w/o MoA$^{(p)}$	0.034	0.54
	GVLA	**0.032**	**0.92**

The result shows that removing the gripper-type token $P^{(g)}$ degrades both PE and adaptation, indicating that type-level priors contribute to both prediction quality and adaptation. The platform token $P^{(p)}$ is critical for cross-embodiment transfer: without it, adaptation drops sharply despite comparable PE, highlighting the necessity of platform disentanglement. In contrast, removing the instance token $P^{(u)}$ only slightly lowers adaptation, suggesting it mainly provides fine-grained specialization for novel gripper instances. Ablating the dual MoA restores PE to baseline and severely reduces adaptation, demonstrating that representation-level prompting alone is insufficient without computation-level modulation. Removing either MoA$^{(g)}$ or MoA$^{(p)}$ yields similar degradation, implying complementary routing roles. The full GVLA achieves the highest PE and adaptation score, supporting that disentangled prompting and dual routing jointly enable robust cross-gripper learning.

5.3 Real-world Robotic Validation

Robot Results. To validate whether gripper-aware conditioning transfers effectively to real-world settings, we conduct a targeted evaluation on cross-domain task generalization: we evaluate our model on a real UR5 arm equipped with a Robotiq 2f-85 gripper, adapting to tasks unseen during training using only

10 demonstrations and 20k fine-tuning steps. Fig. 10a shows the setup of our robotic experiment. As shown in Fig. 10b, our GVLA outperforms the $\pi_{0.5}$ baseline across all tasks, demonstrating that gripper-aware conditioning facilitates embodiment-specific knowledge transfer, enabling efficient few-shot adaptation to novel tasks across unseen platform-gripper configurations.

Failure Cases. Our experiment also reveals several failure cases that are worth noticing. Fig. 11 shows some of these failure cases. We observe that the failure cases are mostly because of: (*i*) Intra-type misalignment: When adapting to unseen grippers, the model selects a type-consistent strategy but fails to precisely align the end-effector, exposing the lack of fine-grained geometric detail for contact planning. (*ii*) Physical gripper limitation: On some tasks, although the model generates feasible trajectories, the grippers are not able to grasp the object due to the physical constraints of the gripper. (*iii*) Kinematic failure: In complex grasping tasks, the robot occasionally enters kinematically infeasible configurations without recovery, exposing a disconnect between high-level action prediction and low-level kinematic feasibility.

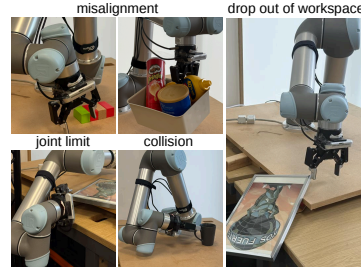


Fig. 11: Failure cases

6 Discussion

Limitations. While our work shows encouraging results, it faces certain limitations. First, the simulation environment of our MiGA dataset is based on NVIDIA Issac Lab, which has certain limitations in modeling accurate gripper motion in complex cases (e.g., soft or multi-DoF grippers). Second, in our GVLA, the gripper information is currently encoded to guide the strategy-level action differences, without explicit kinematic or contact modeling, which limits precise intra-type adaptation. Incorporating richer geometric representations could enable instance-aware contact planning. Furthermore, although our dual MoA encourages gripper-specific specialization, gripper and visual representations remain partially entangled, causing the policy to over-rely on image information under domain shift. Strengthening morphology-aware reasoning and vision disentanglement are therefore critical for robust cross-domain transfer.

Conclusion. To address the limitation of gripper invariance in existing VLAs and the gripper-dependent divergence of grasping strategies across different gripper types, we made two key contributions. First, we introduce MiGA, a multi-gripper dataset collected with five gripper types and 103,000 demonstrations from simulation and real-world robots. MiGA is designed to reveal distinct action strategies across grippers. Second, we propose GVLA, a gripper-aware VLA framework that uses our new multi-gripper tokenizer and a dual MoA design to condition the policy with action and gripper routing information. Intensive experiment results show that GVLA can effectively learns gripper-discriminative representations, outperforming recent methods in complex grasping tasks and enabling stronger few-shot adaptation to new objects, new tasks, or grippers. Our code and dataset will be released to support future research.

References

1. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv:1610.01644 (2016)
2. Attarian, M., Asif, M.A., Liu, J., Hari, R., Garg, A., Gilitschenski, I., Tompson, J.: Geometry matching for multi-embodiment grasping. In: CoRL (2023)
3. Bauer, E., Nava, E., Katzschmann, R.K.: Latent action diffusion for cross-embodiment manipulation. arXiv:2506.14608 (2025)
4. Beijing Inspire Robots Technology: The dexterous hands rh56dftp. https://en.inspire-robots.com/ (2023)
5. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: A vision-language-action flow model for general robot control. arXiv.2410.24164
6. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv:2212.06817 (2022)
7. Casas, L.F., Khargonkar, N., Prabhakaran, B., Xiang, Y.: Multigrippergrasp: A dataset for robotic grasping from parallel jaw grippers to dexterous hands. In: IROS (2024)
8. Cui, Y., Zhang, Y., Tao, L., Li, Y., Yi, X., Li, Z.: End-to-end dexterous arm-hand vla policies via shared autonomy: Vr teleoperation augmented by autonomous hand vla policy for efficient data collection. arXiv:2511.00139 (2025)
9. D'Avella, S., Sundaram, A.M., Friedl, W., Tripicchio, P., Roa, M.A.: Multimodal grasp planner for hybrid grippers in cluttered scenes. IEEE RA-L (2023)
10. Deng, S., Yan, M., Wei, S., Ma, H., Yang, Y., Chen, J., Zhang, Z., Yang, T., Zhang, X., Cui, H., et al.: Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. In: CoRL (2025)
11. Du, Y., Niu, T., Zhao, R.: Mixture of prompts learning for vision-language models. Frontiers in Artificial Intelligence (2025)
12. Fang, H.S., Fang, H., Tang, Z., Liu, J., Wang, J., Zhu, H., Lu, C.: Rh20t: A robotic dataset for learning diverse skills in one-shot. In: RSS Workshop (2023)
13. Fang, H.S., Wang, C., Fang, H., Gou, M., Liu, J., Yan, H., Liu, W., Xie, Y., Lu, C.: Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. IEEE T-RO (2023)
14. Fang, H.S., Wang, C., Gou, M., Lu, C.: Graspnet-1billion: A large-scale benchmark for general object grasping. In: CVPR (2020)
15. Fang, H.S., Yan, H., Tang, Z., Fang, H., Wang, C., Lu, C.: Anydexgrasp: General dexterous grasping for different hands with human-level learning efficiency. arXiv:2502.16420 (2025)
16. Franka Emika GmbH: Franka emika panda robot. https://www.franka.de/ (2017)
17. Freiberg, R., Qualmann, A., Vien, N.A., Neumann, G.: Diffusion for multi-embodiment grasping. IEEE RA-L (2025)
18. Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., Liao, R., Qin, Y., Tresp, V., Torr, P.: A systematic survey of prompt engineering on vision-language foundation models. arXiv:2307.12980 (2023)
19. Han, B., Parakh, M., Geng, D., Defay, J.A., Luyang, G., Deng, J.: Fetchbench: A simulation benchmark for robot fetching. In: CoRL (2025)
20. Hernandez, J., Sunny, M.S.H., Sanjuan, J., Rulik, I., Zarif, M.I.I., Ahamed, S.I., Ahmed, H.U., Rahman, M.H.: Current designs of robotic arm grippers: A comprehensive systematic review. Robotics (2023)

21. Intelligence, P., Amin, A., Aniceto, R., Balakrishna, A., Black, K., Conley, K., Connors, G., Darpinian, J., Dhabalia, K., DiCarlo, J., et al.: $\pi_{0.6}$: a vla that learns from experience. arXiv:2511.14759 (2025)
22. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv:2504.16054 (2025)
23. Isaac Sim: Ur10 short suction gripper. https://docs.isaacsim.omniverse.nvidia.com/4.5.0/assets/usd_assets_robots.html
24. Khargonkar, N., Casas, L.F., Prabhakaran, B., Xiang, Y.: Robotfingerprint: Unified gripper coordinate space for multi-gripper grasp synthesis and transfer. In: IROS (2025)
25. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: CVPR (2023)
26. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., et al.: Droid: A large-scale in-the-wild robot manipulation dataset. In: RSS Workshop (2024)
27. Kim, M.J., Finn, C., Liang, P.: Fine-tuning vision-language-action models: Optimizing speed and success. arXiv:2502.19645 (2025)
28. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Sanketi, P.R., Vuong, Q., et al.: Openvla: An open-source vision-language-action model. In: CoRL (2024)
29. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
30. Li, P., Liu, T., Li, Y., Geng, Y., Zhu, Y., Yang, Y., Huang, S.: Gendexgrasp: Generalizable dexterous grasping. In: ICRA (2023)
31. Li, Q., Deng, Y., Liang, Y., Luo, L., Zhou, L., Yao, C., Zeng, L., Feng, Z., Liang, H., Xu, S., et al.: Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. arXiv:2510.21571 (2025)
32. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv:2101.00190 (2021)
33. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2022)
34. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. NeurIPS (2023)
35. Liu, Q.: Rectified flow: A marginal preserving approach to optimal transport. arXiv:2209.14577 (2022)
36. Liu, X., Sun, T., Huang, X.J., Qiu, X.: Late prompt tuning: A late prompt could be better than many prompts. In: EMNLP (2022)
37. Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., Tang, J.: P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: ACL (2022)
38. Mittal, M., Yu, C., Yu, Q., Liu, J., Rudin, N., Hoeller, D., Yuan, J.L., Singh, R., Guo, Y., Mazhar, H., et al.: Orbit: A unified simulation framework for interactive robot learning environments. IEEE RA-L (2023)
39. Murali, A., Sundaralingam, B., Chao, Y.W., Yuan, W., Yamada, J., Carlson, M., Ramos, F., Birchfield, S., Fox, D., Eppner, C.: Graspgen: A diffusion-based framework for 6-dof grasping with on-generator training. arXiv:2507.13097 (2025)
40. Nguyen, Q., Le, T., Nguyen, H., Vo, T., Ta, T.D., Huang, B., Vu, M.N., Nguyen, A.: Graspmas: Zero-shot language-driven grasp detection with multi-agent system. In: IROS (2025)

41. Nguyen, T., Vu, M.N., Vuong, A., Nguyen, D., Vo, T., Le, N., Nguyen, A.: Open-vocabulary affordance detection in 3d point clouds. In: IROS (2023)
42. O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In: ICRA (2024)
43. Pan, C., Junge, K., Hughes, J.: Vision-language-action model and diffusion policy switching enables dexterous control of an anthropomorphic hand. arXiv:2410.14022 (2024)
44. Patel, A., Song, S.: Get-zero: Graph embodiment transformer for zero-shot embodiment generalization. In: ICRA (2025)
45. Robotiq Inc.: Robotiq 2-finger adaptive robot gripper - 85. https://robotiq.com/ (2021)
46. Robotiq Inc.: Robotiq 3-finger adaptive robot gripper. https://robotiq.com/ (2021)
47. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. Tech. rep. (1985)
48. Schmalz: Schmalz cobot pump. https://www.schmalz.com/
49. Song, S., Zeng, A., Lee, J., Funkhouser, T.: Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. IEEE RA-L (2020)
50. Stone, A., Xiao, T., Lu, Y., Gopalakrishnan, K., Lee, K.H., Vuong, Q., Wohlhart, P., Kirmani, S., Zitkovich, B., Xia, F., et al.: Open-world object manipulation using pre-trained vision-language models. In: CoRL (2023)
51. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. NeurIPS (2017)
52. Veitch, V., D'Amour, A., Yadlowsky, S., Eisenstein, J.: Counterfactual invariance to spurious correlations: Why and how to pass stress tests. arXiv preprint arXiv:2106.00545 (2021)
53. Vuong, A.D., Vu, M.N., Huang, B., Nguyen, N., Le, H., Vo, T., Nguyen, A.: Language-driven grasp detection. In: CVPR (2024)
54. Walke, H.R., Black, K., Zhao, T.Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A.W., Myers, V., Kim, M.J., Du, M., et al.: Bridgedata v2: A dataset for robot learning at scale. In: CoRL (2023)
55. Wang, L., Chen, X., Zhao, J., He, K.: Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. NeurIPS (2024)
56. Wang, Y., Mukherjee, S., Liu, X., Gao, J., Awadallah, A.H., Gao, J.: Adamix: Mixture-of-adapter for parameter-efficient tuning of large language models. arXiv:2205.12410 (2022)
57. Wang, Z., Wang, P., Liu, T., Lin, B., Cao, Y., Sui, Z., Wang, H.: Hpt: Hierarchy-aware prompt tuning for hierarchical text classification. In: EMNLP (2022)
58. Wei, Y., Attarian, M., Gilitschenski, I.: Geomatch++: Morphology conditioned geometry matching for multi-embodiment grasping. In: CoRL Workshop (2024)
59. Wei, Z., Xu, Z., Guo, J., Hou, Y., Gao, C., Zhehao, C., Luo, J., Shao, L.: D (r, o) grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. In: CoRL Workshop (2024)
60. Wen, R., Chen, G., Cui, Z., Du, M., Gou, Y., Han, Z., Huang, L., Lei, M., Li, Y., Li, Z., et al.: Gr-dexter technical report. arXiv:2512.24210 (2025)
61. Wu, C., Chen, J., Cao, Q., Zhang, J., Tai, Y., Sun, L., Jia, K.: Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps. NeurIPS (2020)

62. Wu, K., Hou, C., Liu, J., Che, Z., Ju, X., Yang, Z., Li, M., Zhao, Y., Xu, Z., Yang, G., et al.: Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. arXiv:2412.13877 (2024)
63. Wu, Z., Potamias, R.A., Zhang, X., Zhang, Z., Deng, J., Luo, S.: Cedex: Cross-embodiment dexterous grasp generation at scale from human-like contact representations. arXiv:2509.24661 (2025)
64. Xu, Z., Qi, B., Agrawal, S., Song, S.: Adagrasp: Learning an adaptive gripper-aware grasping policy. In: ICRA (2021)
65. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: CVPR (2023)
66. Yu, J., Zhuge, Y., Zhang, L., Hu, P., Wang, D., Lu, H., He, Y.: Boosting continual learning of vision-language models via mixture-of-experts adapters. In: CVPR (2024)
67. Yuan, H., Zhou, B., Fu, Y., Lu, Z.: Cross-embodiment dexterous grasping with reinforcement learning. In: ICLR (2024)
68. Zhang, H., Ma, K.Y., Shou, M.Z., Lin, W., Wu, Y.: Cross-embodiment dexterous hand articulation generation via morphology-aware learning. arXiv:2510.06068 (2025)
69. Zheng, J., Li, J., Wang, Z., Liu, D., Kang, X., Feng, Y., Zheng, Y., Zou, J., Chen, Y., Zeng, J., et al.: X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. arXiv:2510.10274 (2025)
70. Zhong, Y., Huang, X., Li, R., Zhang, C., Chen, Z., Guan, T., Zeng, F., Lui, K.N., Ye, Y., Liang, Y., et al.: Dexgraspvla: A vision-language-action framework towards general dexterous grasping. arXiv:2502.20900 (2025)
71. Zhou, A.: Rt-2: Vision-language-action models for generalizable robotic control: A comprehensive review. Advances in Engineering Technology Research (2025)
72. Zhou, H., Huang, S., Li, M., Zhang, H., Fan, L., Shi, S.: Vacuumvla: Boosting vla capabilities via a unified suction and gripping tool for complex robotic manipulation. arXiv:2511.21557 (2025)
73. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR (2022)
74. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV (2022)
75. Zhu, J., Sun, X., Zhang, Q., Liu, M.: Vla-grasp: a vision-language-action modeling with cross-modality fusion for task-oriented grasping. Complex & Intelligent Systems (2025)
76. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL (2023)