IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

A Taxonomy for Evaluating Generalist Robot Manipulation Policies

Jensen Gao, Suneel Belkhale, Sudeep Dasari, Ashwin Balakrishna, Dhruv Shah, and Dorsa Sadigh, *Member, IEEE*

***Abstract*—Machine learning for robot manipulation promises to unlock generalization to novel tasks and environments. But how should we measure the progress of these policies towards generalization? Evaluating and quantifying generalization is the Wild West of modern robotics, with each work proposing and measuring different types of generalization in their own, often difficult to reproduce settings. In this work, our goal is (1) to outline the forms of generalization we believe are important for robot manipulation in a comprehensive and fine-grained manner, and (2) to provide reproducible guidelines for measuring these notions of generalization. We first propose ★-Gen, a taxonomy of generalization for robot manipulation structured around visual, semantic, and behavioral generalization. Next, we instantiate ★-Gen with two case studies on real-world benchmarking: one based on open-source models and the Bridge V2 dataset, and another based on the bimanual ALOHA 2 platform that covers more dexterous and longer horizon tasks. Our case studies reveal many interesting insights: for example, we observe that open-source vision-language-action models often struggle with semantic generalization, despite pre-training on internet-scale language datasets.**

***Index Terms*—Big Data in robotics and automation, deep learning in grasping and manipulation.**

I. Introduction

LEARNING-based robotics comes with the promise of broad generalization. As an example, an ambitious goal is to train a laundry-folding robot on diverse household data that can fold laundry in new homes. If trained effectively, the robot should be able to fold unseen clothing items in new settings using its extensive prior experience. However, we have yet to reach a point in robot manipulation where policies can reliably generalize in this manner. In pursuit of this vision, recent work has focused on scaling up data collection [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] and developing more expressive models [6], [13], [14], [15], [16], following the successes of other machine learning domains.

While these advances have led to more capable policies, it is often unclear how generalist these policies truly are. Although prior work has shown various forms of generalization, such as visual robustness to distractors, or understanding novel language instructions, there is often a lack of consistency across different evaluations. Each work proposes their own forms of generalization and evaluation conditions, usually with little transparency into how they were decided upon. As a result, it has become difficult to measure progress toward real-world deployability of these policies, which has remained largely elusive, despite promising results in the literature.

To work towards more comprehensive and systematic evaluations, we propose ★-Gen (STAR-Gen) – a **S**ystematic **T**axonomy of the **A**xes of **R**obot **Gen**eralization. We observe that policies require generalization when there are *perturbations* to the policy's *inputs* or required *outputs*. Therefore, to ground our taxonomy, we structure ★-Gen based on the input and output modalities of visuo-lingual control policies: vision, language, and actions. We categorize perturbations as **visual**, **semantic**, and/or **behavioral** based on how they affect these modalities. For each combination of these labels (e.g., **visual** only, or **visual + behavioral**), we define more granular generalization *axes*. For instance, we include *Object Properties* as a **semantic** axis, which involves generalizing from "put carrot on plate" to "put the orange object on the plate".

To demonstrate the practical utility of ★-Gen, we present two real-world case studies on benchmarking generalization. First, we develop BridgeV2-★, a benchmark based on the Bridge V2 dataset [9], that is intended to provide a blueprint for designing generalization benchmarks using a reproducible and open-source platform. We outline the design choices of BridgeV2-★based on our taxonomy, and use it to evaluate state-of-the-art open-source generalist manipulation policies.

Next, we use ★-Gen to develop an additional case study based on the bimanual ALOHA 2 platform [17] that considers more dexterous, varied, and longer-horizon tasks, supported by a large-scale real-world dataset with thousands of hours of demonstrations. This further demonstrates the broad applicability of our taxonomy to a wide range of settings.

In summary, we present ★-Gen, a taxonomy of generalization structured around three modalities – vision, language, and actions – that span the space of generalization for visuo-lingual

Received 21 August 2025; accepted 7 January 2026. Date of publication 22 January 2026; date of current version 30 January 2026. This article was recommended for publication by Associate Editor B. Calli and Editor M. Vincze upon evaluation of the reviewers' comments. This work was supported in part by ONR under Grant N00014-22-1-2293, in part by DARPA under Grant W911NF2210214, and in part by NSF under Grant 1941722. *(Jensen Gao and Suneel Belkhale contributed equally to this work.) (Corresponding author: Jensen Gao.)*

Jensen Gao, Suneel Belkhale, and Dorsa Sadigh are with the Stanford University, Stanford, CA 94305 USA (e-mail: jenseng@stanford.edu; belkhale@stanford.edu; dorsa@cs.stanford.edu).

Sudeep Dasari and Ashwin Balakrishna are with Google DeepMind, Santa Clara, CA 95051 USA.

Dhruv Shah is with Google DeepMind, Santa Clara, CA 95051 USA, and also with the Princeton University, Princeton, NJ 08544 USA (e-mail: shahd@princeton.edu).

We provide videos and other supplementary material at our website stargen-taxonomy.github.io.

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2026.3656785, provided by the authors.

Digital Object Identifier 10.1109/LRA.2026.3656785

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

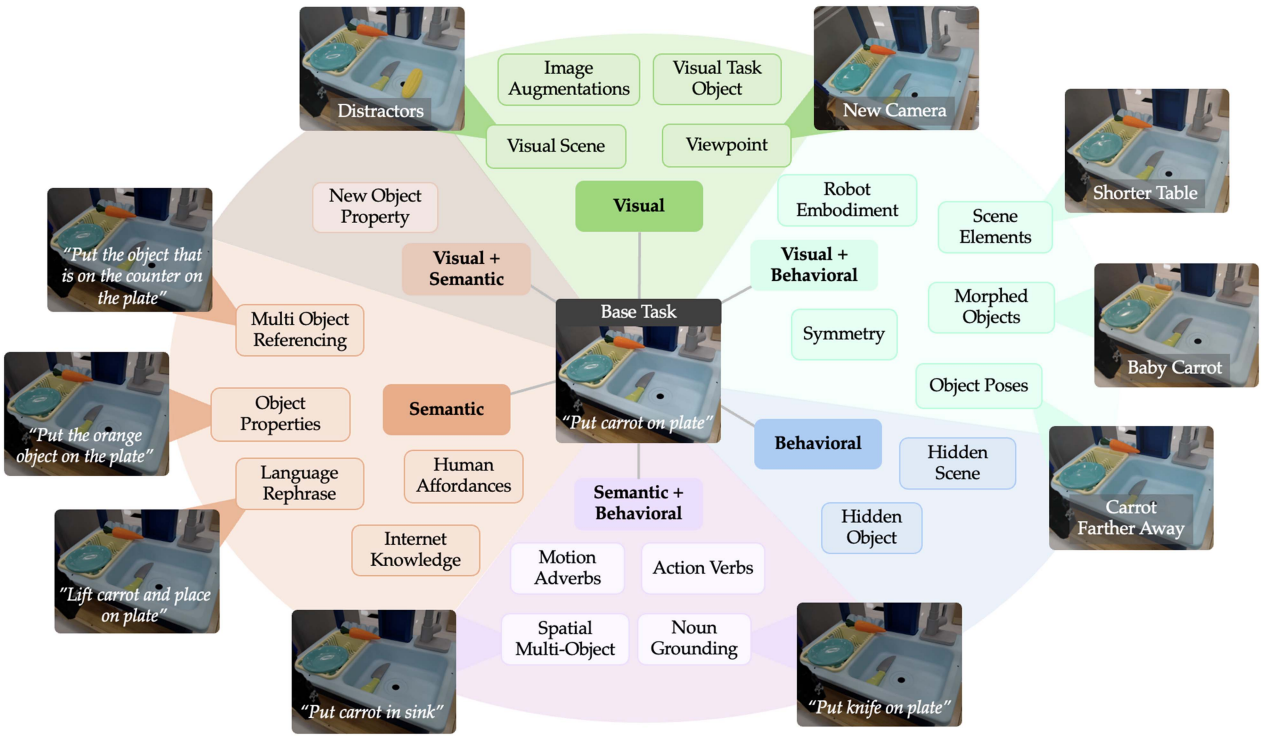


Fig. 1. Visualization of ★-Gen for the example base task "put carrot on plate". ★-Gen is structured around perturbations to the modalities of visuo-lingual policies (**visual**, **semantic**, **behavioral**), with consideration for each of their combinations, which we refer to as *categories*. For each category (colored sectors), we further group perturbations into *axes* (light colored boxes). We provide some example perturbations.

manipulation policies. We instantiate ★-Gen as two real-world case studies that consist of 1600+ robot trials across 14 axes of generalization, which generate more detailed findings on state-of-the-art generalist policies and model design decisions. We hope that by guiding policy training and evaluation efforts, ★-Gen can help advance progress in robot manipulation.

II. Related Work

To achieve broad generalization in robotics, much prior work has focused on scaling up real-world data collection. These efforts typically aim to capture diversity in both environmental conditions and task behavior [1], [2], [3], [9], [10], [11], [18]. While these datasets are usually collected with diversity in mind, it is often unclear what forms of diversity matter, or how this diversity should be achieved. Recent works have investigated best practices for generating diverse robot data [19], [20], [21]. Although these works consider various notions of diversity during data collection and corresponding axes of generalization during evaluation, these axes are neither exhaustive nor standardized.

Nevertheless, recent works have attempted to leverage these datasets for learning generalist robot policies. These works involve training large-scale, visuo-lingual policies on this data, with the goal of generalizing to a wide variety of scenarios [5], [6], [8], [13], [14], [15], [16], [18]. However, each work designs their own evaluations that often focus on a relatively narrow selection of generalization, making it challenging to assess how models make progress towards different forms of generalization.

A compelling alternative is to benchmark generalization in simulation. There has been extensive work in simulated robot manipulation platforms that support task and scene diversity [22], [23], [24], [25], [26], [27]. However, these largely do not come with benchmarks that measure precise notions of generalization. While some works have studied specific distribution shifts in simulation [28], [29], [30], [31], [32], the generalization axes considered are usually inconsistent across works, similar to real-world efforts.

To help unify and provide structure to notions of generalization in robot manipulation, we propose ★-Gen, which considers a superset of generalization axes from prior works. We hope this taxonomy can be useful for developing better datasets and models that make progress towards generalization, and developing better benchmarks to capture this progress.

III. What is Generalization?

In this section, we outline our preliminaries, provide a formal characterization of generalization for robot policies, and list some additional assumptions, which we will use later in Section IV-A to design our taxonomy ★-Gen.

A. Preliminaries

Environment: We define an environment as the tuple $E = (\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{L}, f_o, f_t)$, where $\mathcal{S}$ is the state space, $\mathcal{O}$ is the observation space derived from $\mathcal{S}$ through observation function $f_o : \mathcal{S} \to \mathcal{O}$, $\mathcal{A}$ is the action space, and $f_t : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition function. We assume $\mathcal{O}$ consists of third-person images of a scene, and $\mathcal{A}$ consists of robot actions.

Task: For an environment E, we define a task space T. A task $\tau \in T$ is defined as $\tau = (p_\tau(s_0), l_\tau, R_\tau)$, where $p_\tau(s_0)$ is an

TABLE I
★-GEN: AXES OF GENERALIZATION

Axis	Name	Description	Example Factors
		Visual	
Image Augmentations	V-AUG	Realistic generic augmentations in image space.	lighting, image blur, image contrast
Visual Scene	V-SC	Visual changes to scene elements that do not affect behavior.	surface color/texture, distractor object appearance/placement
Visual Task Object	V-OBJ	Visual changes to task-relevant objects that do not affect behavior.	manipulated object color, other task-relevant object color (e.g., container an object is placed in)
Viewpoint	V-VIEW	Changes to camera viewpoints.	camera pose, partial occlusion
		Semantic	
Object Properties	S-PROP	Changes to instruction that require additional knowledge about physical properties of a task-relevant object.	referencing objects based on color, mass, size
Language Rephrase	S-LANG	Simple rephrasing of the instruction that does not affect underlying behavior.	verb synonyms, removing articles (e.g., "pick up the carrot" → "pick up carrot")
Multi-Object Referencing	S-MO	Changes to instruction that involve references to spatial relationships between multiple objects without changing behavior.	understanding "left", "right", "in" an object (e.g., "pick up carrot" → "pick up object in sink")
Human Affordances	S-AFF	Changes to instruction that require knowledge of human affordances, or how humans interact with an object.	understanding human comfort, object use cases (e.g., "hand me something I can use to clean up this mess")
Internet Knowledge	S-INT	Changes to instruction that require external knowledge that can be found on the internet, and do not fall under the other **semantic** axes.	famous nouns (e.g., celebrities), common knowledge (e.g., tennis balls are green), typos
		Behavioral	
Hidden Object	B-HOBJ	Unobserved changes to task-relevant objects that affect behavior.	task-relevant object mass, friction, fragility
Hidden Scene	B-HSC	Unobserved changes to scene elements that affect behavior.	surface friction, temperature
		Visual + Behavioral	
Object Poses	VB-POSE	Changes to task-relevant object poses in the scene.	manipulated object pose, other object pose
Interacting Scene	VB-ISC	Changes to scene elements that affect behavior.	clutter, surface height
Morphed Objects	VB-MOBJ	Changes to task-relevant objects that affect their geometry.	manipulated object size, shape
Robot Embodiment	VB-ROB	Changes to the robot embodiment that affect behavior.	new robot arm, new gripper or hand
Symmetry	VB-SYM	Specific to bimanual embodiments, symmetry captures changes that require the robot to mirror behavior across arms.	using different arm to perform same absolute motion, flipped absolute motion
		Semantic + Behavioral	
Motion Adverbs	SB-ADV	Changes to instruction motion descriptors that affect behavior.	speed (e.g., "quickly" or "slowly")
Spatial Multi-Object	SB-SMO	Changes to instruction that involve references to spatial relationships between multiple objects, which change the task specification to involve new behavior.	changing goal location for object (e.g., "put carrot on plate" → "put carrot in sink")
Noun Grounding	SB-NOUN	Changes to task-relevant nouns in the instruction to other nouns already present in the scene.	changing manipulated object to another in scene (e.g., "pick carrot" → "pick knife" when both in scene)
Action Verbs	SB-VRB	Changes to action verbs in the instruction that require new behavior.	new action on a manipulated object (e.g., "pick bottle" → "rotate bottle")
		Visual + Semantic	
New Object Property	VS-PROP	Changes to task-relevant object properties that affect object appearance and language instruction, but not behavior.	new object color when base language instruction refers to the object color
		Visual + Semantic + Behavioral	
New Object	VSB-NOBJ	Changes to task-relevant objects to new objects with different appearances, semantic descriptions, and physical characteristics.	new manipulated object (e.g., carrot → zucchini)

initial state distribution for E, $l_\tau \in \mathcal{L}$ is a language instruction, and $R_\tau : (\mathcal{S} \times \mathcal{A})^* \to \{0, 1\}$ is a success function that maps a state/action sequence to a success indicator. $p_\tau(s_0)$ defines an initial observation distribution $p_\tau(o_0)$ induced by f_o.

Policy: A policy $\pi(a \mid o^n, l)$ takes in $n \geq 1$ observations, a language instruction, and outputs an action distribution. We define an **expert policy** $\pi_E(a \mid o^n, l)$ that produces successful episodes (where success is defined by R_τ) for a given task τ.

B. Defining Generalization for Visuo-Lingual Policies

Generalization in robotics is often considered as the performance of a policy π on a task τ' outside its training distribution. To deploy policies in diverse settings, there is a vast space of potential tasks τ' to consider. Furthermore, it can be challenging to characterize how tasks represent generalization from large datasets. To address these challenges and provide a theoretically grounded framework, we propose structuring our taxonomy of generalization around *perturbations* of a given base task, and how they affect the core input and output modalities of a robot policy, which we formalize as follows:

Base Task: A base task τ_B is a task where an end application desires a policy to perform the task (e.g., chopping a specific onion) and perturbations of it (e.g., chopping other onions).

Perturbations: We define a perturbation function as a transformation $P : T \to T$, that applies a task delta to a base task τ to produce a new task τ_P. We categorize perturbations induced by a perturbation function based on how the inputs and outputs of a policy $\pi(a \mid o^n, l)$ are impacted:

- *Visual:* τ_P is a visual perturbation of τ if $p_\tau(o_0) \neq p_{\tau_P}(o_0)$ (the initial distribution of image observations has changed.)
- *Semantic:* τ_P is a semantic perturbation of τ if $l_\tau \neq l_{\tau_P}$ (the language instruction has changed.)
- *Behavioral:* τ_P is a behavioral perturbation of τ if the expert policy π_E changes its action distribution for the task (the required optimal behavior changes.)

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE II
WE PRESENT EXISTING GENERALIZATION BENCHMARKS/DATASETS AND GENERALIST POLICY LEARNING WORKS THROUGH THE LENS OF ★-GEN, INCLUDING BRIDGEV2-★. COLUMNS ARE DIFFERENT AXES IN ★-GEN (A SUBSET OF 17/22 AXES IN TABLE I). A CHECKMARK INDICATES A GIVEN AXIS IS CONSIDERED

		Visual				Semantic					B	VB				SB		VSB
		AUG	SC	OBJ	VIEW	PROP	LANG	MO	AFF	INT	HOBJ	POSE	ISC	MOBJ	ROB	SMO	NOUN	NOBJ
Simulation Data	FactorWorld [30]	✓	✓	✓	✓							✓	✓	✓				
	KitchenShift [28]	✓		✓	✓							✓		✓	✓			
	Colosseum [31]	✓	✓	✓	✓						✓			✓				
	Eff-Comp [19]		✓	✓	✓							✓						
	MimicLabs [21]		✓	✓	✓							✓						
	CALVIN [29]		✓	✓		✓	✓					✓	✓	✓				
	VLABench [34]					✓	✓	✓	✓	✓						✓		✓
Real Data	Scaling [20]			✓								✓	✓	✓				
	BridgeV2 [9]	✓	✓	✓							✓	✓	✓	✓	✓			✓
	DROID [11]		✓	✓	✓									✓	✓			
Policy	BC-Z [4]		✓									✓		✓	✓			✓
	RT-Series [5, 6]		✓			✓		✓		✓		✓	✓	✓	✓	✓		✓
	MT-ACT [8]	✓	✓									✓	✓	✓	✓			✓
	π_0 [16]		✓	✓								✓	✓	✓	✓			✓
	OpenVLA [13]		✓			✓	✓	✓		✓		✓		✓	✓		✓	✓
	BridgeV2-★		✓	✓	✓	✓	✓	✓		✓		✓	✓	✓		✓	✓	✓

This categorization is not mutually exclusive, meaning a perturbation can fall under more than one category. This can also be extended to other policy modalities, e.g., if a policy uses tactile information or sound [33], we can further categorize perturbations based on changes to these modalities.

C. Additional Assumptions

Atomic Perturbations: We focus on *atomic* perturbations, which we loosely define as involving a single change (e.g., "pick up plate" → "push the cup" is not *atomic* because it involves both changing "plate" to "cup" and "pick" to "push").

Short-Horizon Tasks: There are forms of generalization specific to long-horizon manipulation, such as reordering sub-tasks in a sequence. We do not consider this, and instead focus on perturbations that are broadly applicable to short-horizon tasks. However, in our case study on bimanual manipulation (Section VI), we evaluate on longer-horizon tasks to demonstrate that ★-Gen can also be applied to such settings.

IV. AXES OF GENERALIZATION

A. ★-Gen: A Taxonomy of Generalization

Here we define ★-Gen, our taxonomy of generalization. We aim to organize perturbations in a human-interpretable manner to guide policy evaluation. To this end, we define *factors*, *axes*, and *categories*, which represent different levels of hierarchy in our taxonomy, in decreasing order of granularity.

Factors: We define a factor as a human-interpretable, fine-grained grouping of perturbations that affect a task in a common way. For example, if the lighting in a scene is changed in multiple ways, each would represent a separate perturbation. We can then group all such perturbations under the factor "Lighting". Factors can be categorized as **visual**, **semantic**, and/or **behavioral**, based on their constituent perturbations.

Axes: We define an axis as a human-interpretable grouping of similar factors that affect a common set of policy modalities. For example, our taxonomy defines *Image Augmentations* as an axis of **visual** factors that can be varied using simple image transforms, such as "Lighting" or "Image Blur". The axes in our taxonomy are designed to be a practically comprehensive set of the most salient challenges identified in the literature.

Categories: We define a category as a grouping of all axes that affect the same combination of policy modalities. For example, the category **visual** captures all axes that only affect the initial image observations of a task, including *Image Augmentations*. There are seven possible categories (the number of combinations of policy modalities). By capturing all combinations of how policy modalities can be affected by a given perturbation, we intend for these categories to provide a complete framing of generalization conditions for robot manipulation.

We outline the axes in ★-Gen in Table I. For each axis, we provide a description and examples of constituent factors. While these axes are intended to be applicable to a broad range of tasks, some tasks will not always have meaningful instantiations of an axis. In Fig. 1, we show example perturbations of the base task "put carrot on plate" for several axes.

B. Prior Notions of Generalization

In Table II, we list prior works that measure generalization and the axes of ★-Gen they consider, in comparison with BridgeV2-★, a benchmark we designed using ★-Gen (Section V). As shown, ★-Gen aims to be a comprehensive superset of prior efforts to measure generalization. There are other nuances between prior works and ★-Gen that are observable in Table II, some of which we describe here.

Prior works are often not as fine-grained as ★-Gen in their categorization of generalization. For example, RT-2 [6] simply groups many of our **visual + behavior** axes under the blanket category "Behavior Generalization". Prior works also often categorize perturbations in ways that are only applicable to certain tasks. For example, Colosseum [31] considers "Receiver Object" (RO) perturbations (e.g., the "rack" in "put wine in rack"), which does not apply to tasks without such objects. In ★-Gen, we address this by considering different levels of hierarchy, where our high-level categorization is amenable to all tasks, while our lower levels may be more task-specific.

Lastly, prior works differ in how their evaluation protocols consider perturbations, often in less practical ways. For example,

	Axis	Factors
Visual	V-SC	distractors, surface color
	V-OBJ	other object color
	V-VIEW	camera pose
Semantic	S-PROP	referencing color
	S-LANG	changing verbs
	S-MO	understanding "in" and "on"
	S-INT	common obj properties, typos
VB	VB-POSE	manipulated object pose
	VB-ISC	surface height
	VB-MOBJ	manipulated object size, shape
SB	SB-SMO	understanding "in"
	SB-VRB	new action on object
VSB	VSB-NOBJ	new manipulated object

Fig. 2. Axes and factors in BridgeV2-★.

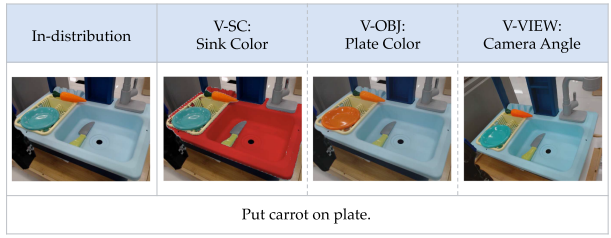


Fig. 3. Examples of **visual** perturbations in BridgeV2-★. Left: in-distribution base task scene. From left to right: we vary sink color (V-SC), plate color (V-OBJ), and camera angle (V-VIEW).

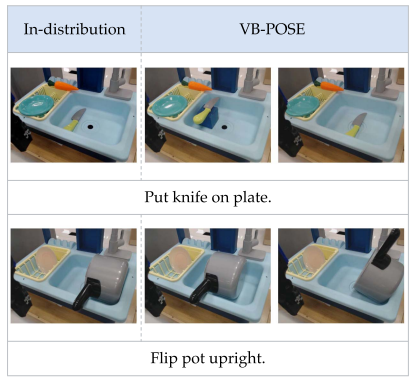


Fig. 4. Examples of object poses (VB-POSE) in BridgeV2-★.

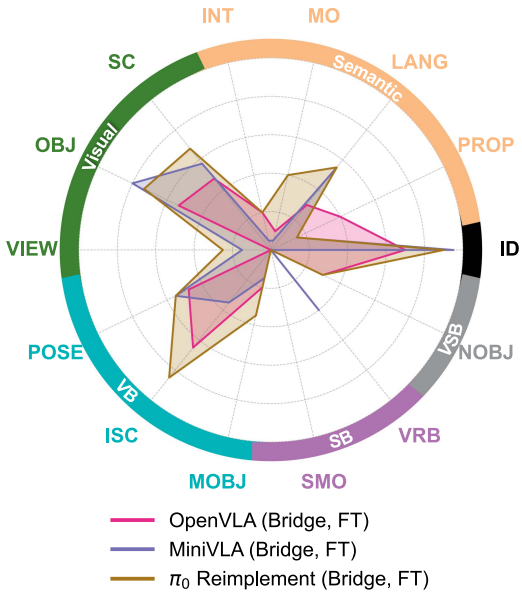


Fig. 5. BridgeV2-★main results. We report aggregated success rates for each model and axis, including in-distribution (ID).

OpenVLA [13] considers perturbations with respect to a portion of training data from a large-scale mixture (OXE), for a scene that their evaluation setting aims to emulate. However, it can be difficult to replicate scenes from an outside data source, possibly leading to unintended perturbations. In BridgeV2-★, we define ★-Gen perturbations with respect to in-domain data from the evaluation scene, to more easily control for this.

V. Case Study 1: Bridge V2

★-Gen provides a broad framework for generalization in robot manipulation, but how should it be used to evaluate policies? In this section, we use ★-Gen to instantiate BridgeV2-★, a real-world benchmark based on Bridge V2 [9]. We first describe our benchmark and our rationale for its design. Then, we use it to evaluate several state-of-the-art open-source models and variations. Our goal is for BridgeV2-★to demonstrate how ★-Gen can be used to design generalization benchmarks using a reproducible and open-source platform.

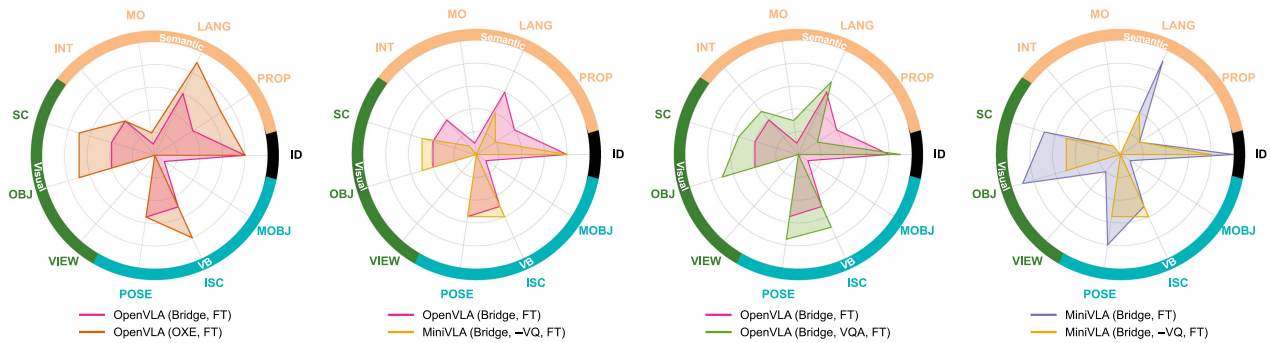
A. Instantiating ★-Gen on Bridge V2

Dataset: We use Bridge V2 [9] as our pre-training dataset and platform, since it has been used in multiple prior works to study generalization [9], [13], [35], and its training environments have been reliably reproduced for evaluation [13], [35], [36], [37].

Base Tasks: We consider the base tasks "put carrot on plate", "put knife on plate", "flip pot upright", and "put plate in sink". We choose these base tasks based on the support of the pre-training data. In particular, they are instantiated in a replication of a sink environment from Bridge V2 that was used to evaluate generalization in prior work [13]. We choose these specific tasks to cover different levels of alignment with the original tasks from Bridge V2 for this sink environment. *Evaluation Conditions:* We evaluate 4 in-distribution base tasks and 55 perturbations that cover 13/22 axes in ★-Gen. We do not cover some axes due to incompatibility with our base tasks. We list our evaluated axes and factors in Table 2, and visualize some base tasks and their perturbations in Figs. 3 and 4 . We further detail our evaluation conditions in our Appendix (can be found on our stargen-taxonomy.github.io website).

Policies: We focus our evaluation on state-of-the-art open-source imitation learning policies that have demonstrated generalization for Bridge V2 tasks in prior work. Specifically, we analyze three vision-language-action (VLA) models that fine-tune foundation models on robot data: OpenVLA [13], MiniVLA [38], and a third-party reimplementation of π_0 [16], [39]. These models cover a range of design decisions that reflect the state of generalist manipulation policies. *Evaluation Procedure:* We evaluate our models using a co-fine-tuning procedure. First, we pre-train each model only on Bridge V2. Next, we collect base task demonstrations from our evaluation

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Scaling robot data (b) Scaling LLM backbones (c) VQA co-training (d) VQ action chunking

Fig. 6. We investigate VLA design decisions. (a) Scaling robot datasets can help. (b) Larger LLMs can provide a modest benefit to **semantic** axes. (c) VQA co-training can help, but has a mixed effect on **semantic** axes. (d) VQ action chunking can help.

TABLE III
AVERAGE PEARSON CORRELATIONS OF PERFORMANCE FOR AXES WITHIN THE SAME CATEGORY (LEFT) AND ACROSS CATEGORIES. (RIGHT)

Visual	Semantic	Visual + Behavioral	Semantic + Behavioral	Across Categories
0.84	-0.07	0.36	0.87	0.02

environment, and co-fine-tune on this with Bridge V2. We denote co-fine-tuned models with (FT).

For the "put carrot" and "put knife" base tasks, we collect 10 demonstrations per base task. For the "flip pot" and "put plate" base tasks, we collect 50 demonstrations per base task. We execute policies until the robot succeeds, reaches a dangerous/irrecoverable state, or terminates after 100 timesteps. We perform five trials per condition for each model.

B. Main Results

In Fig. 5, we report our main results on BridgeV2-★, which consist of 885 trials. We find that existing generalist policies tend to struggle on most axes. In particular, **semantic** generalization is mostly weak, despite the use of language model backbones. This has interesting implications: e.g., instead of relying only on language model initialization to improve **semantic** generalization, perhaps other mechanisms are needed, such as improving robot language annotations [40].

Each model tends to have similar strengths and weaknesses. However, there are some notable differences that the fine-grained nature of our benchmark helps reveal. For example, OpenVLA is noticeably worse at **visual** generalization, while MiniVLA struggles more with **visual + behavioral**. OpenVLA is the best at understanding object properties, but still struggles with other **semantic** axes. π_0 generally performs the best, possibly due to a more capable VLM backbone (PaliGemma [41]), and/or better architecture (flow-based action chunking).

Axes Correlations: To further motivate our high-level categorization based on policy modalities, we investigate performance correlations across models for axes within the same category, compared to across categories. In Table III, we find that correlations are higher within categories, except for **semantic**. We hypothesize this is because our **semantic** axes can require much different forms of reasoning (e.g., understanding object properties is much different than language rephrasing).

Prioritizing Axes: From these results, we provide some general guidelines on axes to prioritize in future work.

- *Visual*-only axes generally exhibit stronger generalization than **behavioral** axes (with the exception of *Viewpoint*). We hypothesize this is because VLA vision-language pre-training is more likely to convey visual robustness than generalization to new behavior. Therefore, future work on generalist manipulation should de-prioritize **visual**-only axes (except *Viewpoint*) in favor of **behavioral** axes.
- While **semantic** generalization is weak, whether to prioritize these axes depends on how the policy is deployed. If the policy is used with open-ended language, then these axes are important. However, if language is more constrained (e.g., a system where a separate model provides a limited set of instructions), they can be de-prioritized.

C. Investigating VLA Design Decisions

While our main results provide insights on model capabilities, it is difficult to disentangle what contributes to generalization. To better understand this, we conduct additional targeted evaluations on model design choices, with t-tests to assess statistical significance.

Scaling Robot Data: In Fig. 6(a) we compare our Bridge-only OpenVLA with a version trained on a significantly larger, cross-embodiment OXE mixture [18]. Consistent with prior work [13], [18], we find that larger and more diverse datasets can significantly improve forms of generalization, such as for **visual + behavioral** axes ($M = 0.22$ vs. $M = 0.48), t(7) = -2.76, p = 0.028$. However, the axes on which the Bridge-only model struggled the most (*Viewpoint*, *Morphed Objects*, *Multi-Object Referencing*) do not improve significantly.

Scaling LLM Backbones: In Fig. 6(b) we compare VLA policies that differ only in the large language model (LLM) backbone. Specifically, we compare OpenVLA (Bridge, FT), using Llama 2 7B [42], and MiniVLA (Bridge, –VQ, FT), using Qwen2.5 0.5B [43]. The only major difference between these two models is their LLM backbone. We find that while the larger LLM improves **semantic** axes, it is not by a significant amount ($M = 0.18$ vs. $M = 0.35), t(7) = -1.87, p = 0.104$. Absolute performance for these and other axes also remain low, suggesting that scaling LLMs only has limited benefits.

VQA Co-training: In Fig. 6(c), we investigate co-training with visual-question answering (VQA) data, which prior work has shown to improve generalization [6]. We find this can

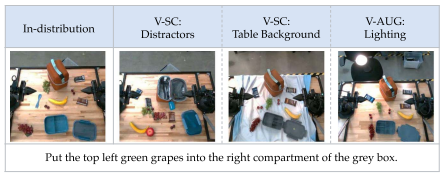


Fig. 7. Examples of **visual** perturbations in our bimanual case study. Left: in-distribution base task. From left to right: we vary distractor objects (V-SC), table background (V-SC), and lighting (V-AUG).

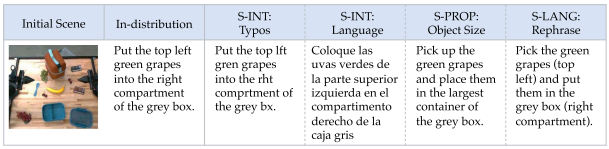


Fig. 8. Examples of **semantic** perturbations in our bimanual case study. Left: in-distribution base task instruction. From left to right: we test robustness to typos (S-INT), language (S-INT), understanding object size (S-PROP), and rephrasing (S-LANG).

help, such as for **visual** axes $(M = 0.30$ vs. $M = 0.45), t(7) = -2.39, p = 0.048$. However, there is surprisingly a mixed effect for **semantic** axes $(M = 0.38$ vs. $M = 0.42), t(7) = -0.51, p = 0.626$, improving three of them, but hurting *Object Properties*. This indicates room for improvement, possibly by using data targeted for embodied reasoning [44].

VQ Action Chunking. In Fig. 6(d), we investigate using binning-based tokenization instead of vector quantized action chunking with MiniVLA. We find that VQ action chunking helps nearly all axes, including **visual** axes by a significant amount $(M = 0.38$ vs. $M = 0.62), t(7) = -2.38, p = 0.049$. This highlights the importance of action chunking and tokenization methods, as also suggested by prior work [45], [46].

VI. Case Study 2: Bimanual Manipulation

Next, we use ★-Gento develop an additional case study based on the bimanual ALOHA 2 platform [17].

A. Experimental Setup

We use a proprietary robot dataset of thousands of hours of teleoperated demonstrations collected on a fleet of ALOHA 2 robots. This allows us to assess generalization for tasks with more variety, dexterity, and horizon length than those considered in BridgeV2-★, such as tightening a water bottle and folding a dress. We evaluate on 17 in-distribution base tasks with 68 perturbations that cover 7 axes in ★-Gen. These are the same conditions used to evaluate generalization for Gemini Robotics [47] (see Appendix C.1.3 in the technical report), but recategorized according to ★-Gen axes. We visualize examples of conditions for **visual** axes in Fig. 7, **semantic** axes in Fig. 8, and **visual + behavioral** axes in Fig. 9.

We evaluate 3 models: a multi-task diffusion policy [48], a reimplementation of π_0 [16], and Gemini Robotics On-Device (GRoD) [49], a proprietary VLA. We report policy task progress. We refer to [47] for more details on our evaluation conditions, protocol, and the diffusion and π_0 policies.

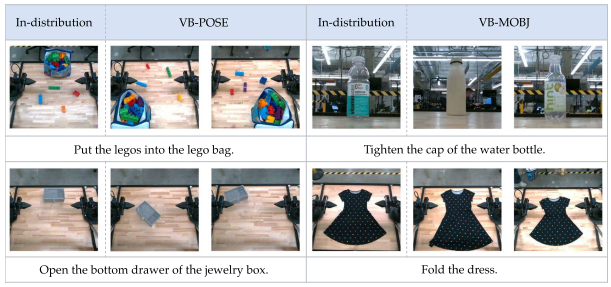


Fig. 9. Examples of **visual + behavioral** perturbations in our bimanual case study. Left: changes to object pose (VB-POSE) from in-distribution. Right: changes to object geometry (VB-MOBJ).

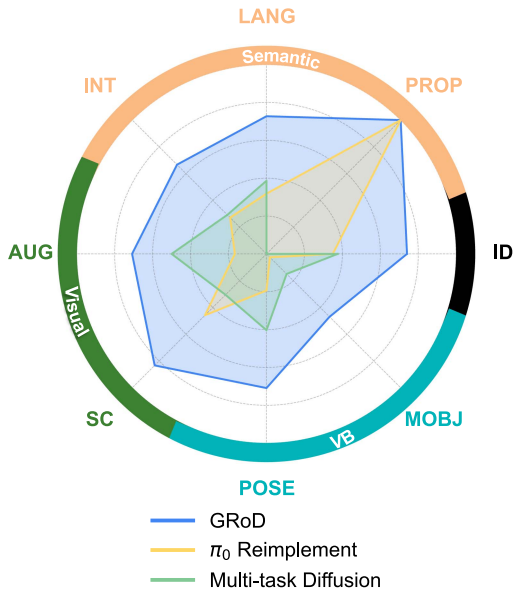


Fig. 10. Results from our bimanual manipulation case study. We report aggregated task progress for each model and axis.

B. Results

We report our results in Fig. 10, which consist of 390 trials. We find that GRoD consistently outperforms the other models, which we speculate is due to its strong VLM backbone and other architectural enhancements. In particular, ★-Gen helps identify perturbations that GRoD is robust to while the other models fail almost entirely, such as translating the instruction to a new language (example in Fig. 8). We use this case study to demonstrate the broad applicability of ★-Gen with more diverse, dexterous, and long-horizon tasks.

VII. Discussion

We present ★-Gen, a taxonomy of generalization for robot manipulation. Our taxonomy not only thoroughly considers the space of visuo-lingual policy generalization, but is also straightforward to instantiate. We demonstrate the considerations and design process for instantiating ★-Gen on a real-world, reproducible benchmark, eliciting key insights about generalist policy capabilities design choices. We further demonstrate using ★-Gen to study generalization for a wider set of tasks and policies on the ALOHA 2 platform. We hope that ★-Gen can help

improve the comprehensiveness of generalization benchmark design to generate better insights.

Limitations and Future Work: Due to constraints imposed by real-world evaluation time (1600+ trials), we only evaluate a subset of axes and factors that we believe most effectively demonstrate the utility of ★-Gen. We hope that ★-Gen can guide future benchmarking efforts that expand the scope considered in our evaluations, such as using simulation to more efficiently and comprehensively measure policy generalization.

While we consider ★-Gen to be a strong starting point, we believe that future work can revise and expand our taxonomy based on the needs of robotics practitioners. Specifically, while we have argued for the completeness of our high-level categories, our set of fine-grained axes is empirically derived. Future work may identify new generalization challenges that could be incorporated as new axes within our framework.

REFERENCES

[1] A. Mandlekar et al., "RoboTurk: A crowdsourcing platform for robotic skill learning through imitation," in *Proc. Conf. Robot Learn.*, 2018, pp. 879–893.

[2] S. Dasari et al., "RoboNet: Large-scale multi-robot learning," in *Proc. Conf. Robot Learn.*, 2019, pp. 885–897.

[3] F. Ebert et al., "Bridge Data: Boosting generalization of robotic skills with cross-domain datasets," in *Proc. Robot., Sci. Syst.*, 2022, p. 63.

[4] E. Jang et al., "BC-Z: Zero-shot task generalization with robotic imitation learning," in *Proc. Conf. Robot Learn.*, 2022, pp. 991–1002.

[5] A. Brohan et al., "RT-1: Robotics transformer for real-world control at scale," in *Proc. Robot. Sci. Syst.*, 2023, p. 25.

[6] A. Brohan et al., "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proc. Conf. Robot Learn.*, 2023, pp. 2165–2183.

[7] N. M. M. Shafiullah et al., "On bringing robots home," 2023, *arXiv:2311.16098*.

[8] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "RoboAgent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," in *Proc. Int. Conf. Robot. Automat.*, 2024, pp. 4788–4795.

[9] H. R. Walke et al., "BridgeData V2: A dataset for robot learning at scale," in *Proc. Conf. Robot Learn.*, 2023, pp. 1723–1736.

[10] H.-S. Fang et al., "RH20 T: A comprehensive robotic dataset for learning diverse skills in one-shot," in *Proc. Int. Conf. Robot. Automat.*, 2024, pp. 653–660.

[11] A. Khazatsky et al., "DROID: A large-scale in-the-wild robot manipulation dataset," in *Proc. Robot., Sci. Syst.*, 2024, p. 120.

[12] S. Mirchandani et al., "RoboCrowd: Scaling robot data collection through crowdsourcing," in *Proc. Int. Conf. Robot. Automat.*, 2025, pp. 1392–1399.

[13] M. Kim et al., "OpenVLA: An open-source vision-language-action model," in *Proc. Conf. Robot Learn.*, 2024, pp. 2679–2713.

[14] L. Wang, X. Chen, J. Zhao, and K. He, "Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 124420–124450.

[15] S. Liu et al., "RDT-1B: A diffusion foundation model for bimanual manipulation," in *Proc. Int. Conf. Learn. Representations*, 2025.

[16] K. Black et al., "π_0: A vision-language-action flow model for general robot control," in *Proc. Robot., Sci. Syst.*, 2025, p. 10.

[17] J. Aldaco et al., "ALOHA 2: An enhanced low-cost hardware for bimanual teleoperation," 2024, *arXiv:2405.02292*.

[18] O. Collaboration et al., "Open X-Embodiment: Robotic learning datasets and RT-X models," in *Proc. Int. Conf. Robot. Automat.*, 2024, pp. 6892–6903.

[19] J. Gao, A. Xie, T. Xiao, C. Finn, and D. Sadigh, "Efficient data collection for robotic manipulation via compositional generalization," in *Proc. Robot., Sci. Syst.*, 2024, p. 13.

[20] F. Lin, Y. Hu, P. Sheng, C. Wen, J. You, and Y. Gao, "Data scaling laws in imitation learning for robotic manipulation," in *Proc. Int. Conf. Learn. Representations*, 2025.

[21] V. Saxena et al., "What matters in learning from large-scale datasets for robot manipulation," in *Proc. Int. Conf. Learn. Representations*, 2025.

[22] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "RLBench: The robot learning benchmark & learning environment," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 3019–3026, Apr. 2020.

[23] T. Mu et al., "ManiSkill: Generalizable manipulation skill benchmark with large-scale demonstrations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021.

[24] A. Szot et al., "Habitat 2.0: Training home assistants to rearrange their habitat," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 251–266.

[25] K. Ehsani et al., "ManipulaTHOR: A framework for visual object manipulation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4497–4506.

[26] C. Li et al., "BEHAVIOR-1 K: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation," in *Proc. Conf. Robot Learn.*, 2023, pp. 80–93.

[27] S. Nasiriany et al., "RoboCasa: Large-scale simulation of everyday tasks for generalist robots," in *Proc. Robot., Sci. Syst.*, 2024, p. 50.

[28] E. Xing, A. Gupta, S. Powers, and V. Dean, "KitchenShift: Evaluating zero-shot generalization of imitation-based policy learning under domain shifts," in *Proc. Workshop Distrib. Shifts, Connecting Methods Appl.*, 2021.

[29] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 7327–7334, Jul. 2022.

[30] A. Xie, L. Lee, T. Xiao, and C. Finn, "Decomposing the generalization gap in imitation learning for visual robotic manipulation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 3153–3160.

[31] W. Pumacay et al., "THE COLOSSEUM: A benchmark for evaluating generalization for robotic manipulation," in *Proc. Robot., Sci. Syst.*, 2024, pp. 133.

[32] X. Li et al., "Evaluating real-world robot manipulation policies in simulation," in *Proc. Conf. Robot Learn.*, 2024, pp. 3705–3728.

[33] M. A. Lee et al., "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *Proc. Int. Conf. Robot. Automat.*, 2018, pp. 8943–8950.

[34] S. Zhang et al., "VLABench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks," in *Proc. Int. Conf. Comput. Vis.*, 2025, pp. 11142–11152.

[35] O. M. Team et al., "Octo: An open-source generalist robot policy," in *Proc. Robot., Sci. Syst.*, 2024, p. 90.

[36] J. Yang, M. S. Mark, B. Vu, A. Sharma, J. Bohg, and C. Finn, "Robot fine-tuning made easy: Pre-training rewards and policies for autonomous real-world reinforcement learning," in *Proc. Int. Conf. Robot. Automat.*, 2024, pp. 4804–4811.

[37] J. Hejna, C. Bhateja, Y. Jian, K. Pertsch, and D. Sadigh, "Re-mix: Optimizing data mixtures for large scale imitation learning," in *Proc. Conf. Robot Learn.*, 2024, pp. 145–164.

[38] S. Belkhale and D. Sadigh, "MiniVLA: A better VLA with a smaller footprint," 2024. [Online]. Available: https://github.com/Stanford-ILIAD/openvla-mini

[39] A. Ren, "Open pi-zero," 2024. [Online]. Available: https://github.com/allenzren/open-pi-zero

[40] L. Smith et al., "STEER: Flexible robotic manipulation via dense language grounding," in *Proc. Int. Conf. Robot. Automat.*, 2025, pp. 16517–16524.

[41] L. Beyer et al., "PaliGemma: A versatile 3B VLM for transfer," 2024, *arXiv:2407.07726*.

[42] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[43] A. Yang et al., "Qwen2.5 Technical Report," 2024, *arXiv:2412.15115*.

[44] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," in *Proc. Conf. Robot Learn.*, 2024, pp. 3157–3181.

[45] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proc. Robot., Sci. Syst.*, 2023, p. 16.

[46] K. Pertsch et al., "FAST: Efficient action tokenization for vision-language-action models," in *Proc. Robot., Sci. Syst.*, 2025, p. 12.

[47] G. R. Team et al., "Gemini robotics: Bringing ai into the physical world," 2025, *arXiv:2503.20020*.

[48] C. Chi et al., "Diffusion policy: Visuomotor policy learning via action diffusion," *Int. J. Robot. Res.*, 2023, pp. 1684–1704.

[49] Google DeepMind, "Gemini robotics on-device," 2025. [Online]. Available: https://deepmind.google/models/gemini-robotics/gemini-robotics-on-device/