Restoring Linguistic Grounding in VLA Models via Train-Free Attention Recalibration

Ninghao Zhang[1*], Bin Zhu[2*†], Shijie Zhou[3,4], and Jingjing Chen[3,4]

[1] Tsinghua University [2] Singapore Management University
[3] Institute of Trustworthy Embodied AI, Fudan University
[4] Shanghai Key Laboratory of Multimodal Embodied AI
Project page: `https://ray-nh.github.io/igar/`

Abstract. Vision-Language-Action (VLA) models enable robots to perform manipulation tasks directly from natural language instructions and are increasingly viewed as a foundation for generalist robotic policies. However, their reliability under Out-Of-Distribution (OOD) instructions remains underexplored. In this paper, we reveal a critical failure mode in which VLA policies continue executing visually plausible actions even when the language instruction contradicts the scene. We refer to this phenomenon as **linguistic blindness**, where VLA policies prioritize visual priors over instruction semantics during action generation. To systematically analyze this issue, we introduce **ICBench**, a diagnostic benchmark constructed from the LIBERO dataset that probes language–action coupling by injecting controlled OOD instruction contradictions while keeping the visual environment unchanged. Evaluations on three representative VLA architectures, including π_0, $\pi_{0.5}$, and OpenVLA-OFT, show that these models frequently succeed at tasks despite logically impossible instructions, revealing a strong visual bias in action generation. To mitigate this issue, we propose **Instruction-Guided Attention Recalibration (IGAR)**, a train-free inference-time mechanism that rebalances attention distributions to restore the influence of language instructions. IGAR operates without retraining or architectural modification and can be directly applied to existing VLA models. Experiments across 30 LIBERO tasks demonstrate that IGAR substantially reduces erroneous execution under OOD contradictory instructions while preserving baseline task performance. We additionally validate the approach on a real Franka robotic arm, where IGAR effectively prevents manipulation triggered by inconsistent instructions.

Keywords: Vision-Language-Action Models · Robotic Manipulation · Attention Recalibration

1 Introduction

Vision-Language-Action (VLA) models are emerging as a promising paradigm for building generalist robotic policies. By integrating large-scale vision-language

[1] * Equal contribution.
[2] † Corresponding author and project lead. Email: binzhu@smu.edu.sg

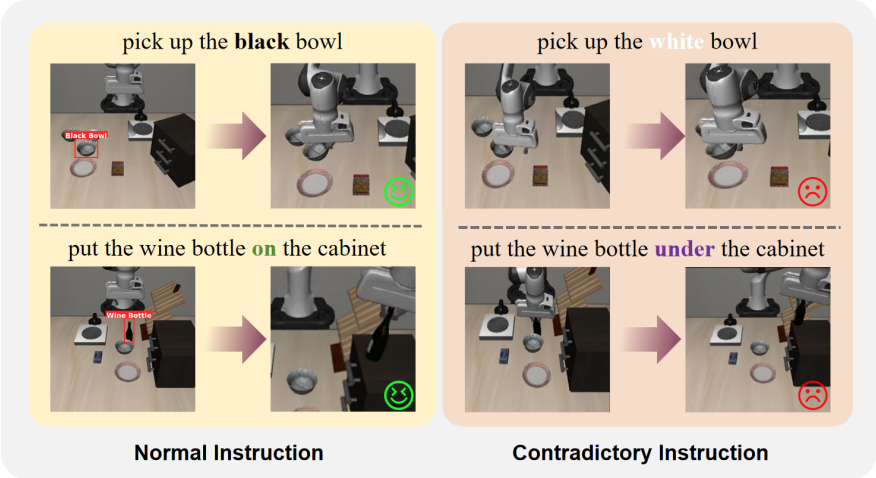

Fig. 1: Linguistic blindness in Vision-Language-Action (VLA) models. Under normal instructions (left), the robot completes the task correctly. Under contradictory instructions (right), a structured form of OOD linguistic input, the robot often follows the same visually plausible trajectory while ignoring the instruction.

models [1, 7, 25, 28] with action generation modules, these systems enable robots to execute complex manipulation tasks directly from natural language instructions across diverse environments [3, 4, 15, 16, 22, 26, 33, 37].

Despite these advances, the reliability of VLA models in real-world deployment remains a critical concern. In safety-critical environments, robots must strictly obey linguistic constraints provided by users. However, we observe that modern VLA models frequently execute visually plausible actions even when the instruction is semantically inconsistent with the scene. As illustrated in Fig. 1, a robot can successfully execute tasks such as *"pick up the black bowl"* or *"put the wine bottle on the cabinet"* once a VLA model is well trained. However, when the instruction is modified to contradict the scene, for example, *"pick up the white bowl"* when no white bowl is present, or *"put the wine bottle under the cabinet"* when this is physically impossible, the robot often continues to execute the original visually plausible trajectory while ignoring the instruction. These contradictory instructions represent a structured form of out-of-distribution (OOD) linguistic input, revealing a fundamental failure mode that we term **linguistic blindness**: the tendency of VLA policies to prioritize visual priors over instruction semantics during action generation. This vulnerability poses serious risks for real-world robotics. Unlike conversational AI systems, errors in robotic control directly translate into physical actions that may damage objects, violate safety constraints, or cause hazardous behavior [11, 18, 27, 31]. Ensuring that robots remain sensitive to language instructions is therefore essential for trustworthy embodied intelligence [32, 34].

However, diagnosing linguistic grounding failures in VLA models is challenging. Existing evaluations primarily measure task success under valid instructions [14, 17, 20], which cannot distinguish whether successful execution arises from genuine language grounding or from purely visual heuristics. To address this limitation, we introduce **ICBench**, a controlled diagnostic benchmark built upon the LIBERO dataset [20]. ICBench isolates the coupling between language and action by injecting structured instruction contradictions while keeping the visual scene unchanged. Specifically, we modify task instructions by altering object attributes or spatial relations of the target location, as shown in Fig. 1. These modifications create semantically inconsistent instructions that function as controlled OOD linguistic perturbations. Under such contradictions, a language-grounded policy should detect the inconsistency and fail to execute the task, whereas a linguistically insensitive policy may still succeed by relying on visual priors. Using ICBench, we conduct a systematic analysis of three representative VLA architectures: π_0 [4], $\pi_{0.5}$ [3], and OpenVLA-OFT [15]. Our evaluation reveals a striking phenomenon: across multiple task suites, VLA models often maintain high success rates even when instructions are logically impossible or inconsistent. This demonstrates that action generation is frequently dominated by visual cues, with language playing only a limited role in guiding behavior.

To address this issue, we propose **I**nstruction-**G**uided **A**ttention **R**ecalibration (**IGAR**), a train-free intervention that restores linguistic grounding by rebalancing attention distributions during the forward pass. IGAR identifies attention sink [13,29] tokens through hidden-state spike analysis, selects grounding-critical cross-modal heads, and redistributes attention mass toward under-weighted instruction tokens. Importantly, IGAR requires no gradient updates, no additional training data, and no architectural modification, making it a lightweight plug-and-play module for deployed robotic VLA policies. Extensive experiments on 30 tasks from the LIBERO benchmark demonstrate the effectiveness of IGAR. Under OOD contradictory instructions, IGAR significantly reduces erroneous task execution and substantially increases the Linguistic Grounding Score (LGS), indicating stronger reliance on instruction semantics. At the same time, IGAR preserves VLA baseline task performance under normal instructions. We further validate our approach on a real Franka robotic arm, where IGAR successfully interrupts the manipulation tasks when contradictory instructions are applied.

2 Related Work

2.1 Vision-Language-Action Models

The integration of Vision Language Models (VLMs) with robotic control has accelerated the transition from task-specific policies to generalist Vision-Language-Action (VLA) models. Early milestones like RT-1 [5] and RT-2 [37] formulated motor control as sequence-to-sequence generation, a paradigm rapidly scaled by massive robotic datasets [22]. Subsequent models further improve action generation through stronger vision-language backbones and advanced control paradigms. For example, OpenVLA [16] leverages large pretrained VLMs, while

architectures like Octo [9], π_0 [4], OpenVLA-OFT [15], RDT [21], GR00T [2], and $\pi_{0.5}$ [3] refine continuous control through diffusion and flow-matching paradigms. Despite these advances, the internal mechanisms fusing visual and textual modalities during action generation remain poorly understood, particularly regarding how language instructions influence control decisions.

2.2 Linguistic Grounding and Modality Bias

Multimodal foundation models are known to exhibit strong modality biases that can undermine reliable language grounding. Multimodal foundation models are vulnerable to modality collapse and visual hallucination. Evaluations [19] show Vision-Language Models (VLMs) frequently hallucinate absent objects due to strong textual priors. Conversely, VLMs often exhibit bag-of-words behavior or disregard order and composition information in text [30]. Recent works on Gaslighting Negation Attack [24,35,36] further demonstrate that negation-based instructions can systematically mislead multimodal large models, revealing their limited ability to faithfully ground contradictory language. Existing works have also identified attention imbalance in LLMs and VLMs, where certain visual tokens attract disproportionately high attention regardless of their relevance to the instruction [13,29]. This vulnerability becomes critical in VLA architectures: unlike conversational agents, ignoring linguistically grounded constraints in robotic control can lead to catastrophic physical failures [11].

Recognizing this risk, initial efforts have sought to address linguistic vulnerabilities in embodied control. To mitigate text-following biases, data-centric strategies like CAST [10] and CounterfactualVLA [23] expand training corpora with synthesized counterfactual scenarios, particularly within navigation and autonomous driving. SayCan [6] grounds language in robotic affordances by combining large language models with value functions of low-level skills for real-world action selection. In addition, the community has also started probing policy resilience, such as LIBERO-PLUS [8], which evaluates robustness against perturbations such as camera viewpoints and instructions. However, most existing works focus on task success under valid instructions and do not explicitly disentangle whether success arises from genuine language grounding or from visually driven heuristics. In contrast, this paper introduces a controlled diagnostic setting that isolates language–action coupling by injecting structured contradictory instructions while keeping the visual scene unchanged. Unlike existing attention-sink mitigation methods [12, 13, 29] developed for language generation in LLMs or MLLMs, IGAR is the first inference-time intervention designed for VLA policies, selectively recalibrating grounding-critical cross-modal attention to recover instruction influence during robot action generation.

3 ICBench: A Controlled Instruction Contradiction Benchmark

To rigorously evaluate linguistic grounding in Vision-Language-Action (VLA) models, we introduce ICBench, a controlled benchmark that injects out-of-distribution

(OOD) semantic contradictions into task instructions while keeping the visual scene and environment dynamics unchanged. The objective of ICBench is diagnostic rather than adversarial: it probes whether language meaningfully influences action generation. If a model genuinely integrates linguistic constraints into motor planning, contradictory instructions should prevent successful execution. Conversely, if the policy predominantly relies on visual priors, it may continue executing visually valid behaviors despite semantic inconsistencies.

3.1 Instruction Contradiction Construction

Let a standard task be defined by instruction ℓ and environment observation $\boldsymbol{o}_t$. ICBench constructs a modified instruction $\tilde{\ell}$ such that: (1) The environment observation $\boldsymbol{o}_t$ remains unchanged. (2) $\tilde{\ell}$ is semantically incompatible with the scene configuration. By holding the environment fixed, ICBench enables controlled measurement of language–action coupling without confounding perception or control quality.

Under this setting, the contradictory instruction $\tilde{\ell}$ can be viewed as an OOD linguistic perturbation relative to the original task instruction. A language-grounded policy should fail or abstain from execution, while a linguistically insensitive policy may still succeed by exploiting visual priors. Importantly, within ICBench, high task success under contradictory instructions indicates weak linguistic grounding. Therefore, task success rate (SR) serves as a proxy for instruction adherence rather than manipulation capability.

3.2 Contradiction Taxonomy and Design Principles

Contradiction Taxonomy. ICBench introduces four structured contradiction types. Each perturbation modifies only a minimal subset of instruction tokens while preserving grammatical plausibility and fluency. These modifications introduce controlled OOD linguistic inputs without altering the visual environment.

V1: Operand Attribute Substitution. The manipulated object's descriptive attribute (e.g., colour) is replaced with a non-existent alternative. For instance, "pick up the black bowl" → "pick up the white bowl". This tests grounding of object-level attribute semantics.

V2: Target Attribute Augmentation. A contradictory attribute is inserted into the target location description. For example, "place it on the plate" → "place it on the black plate." This evaluates grounding of relational destination constraints.

V3: Dual Attribute Perturbation. Both operand and target attributes are contradicted simultaneously. For example, "put black bowl on the plate" → "put white bowl on black plate". This increases semantic inconsistency while maintaining syntactic validity.

V4: Spatial Relation Substitution. The spatial preposition is replaced with a contradictory alternative. For instance, "put the block on the table" → "put the block under the table". Unlike attribute substitutions, spatial contradictions directly alter trajectory planning, thereby testing deeper relational grounding.

Design Principles. ICBench is designed with three key principles: (1) *Minimal Surface Perturbation.* Only a small number of tokens are modified to ensure the instruction remains natural and close to the training distribution. (2) *Deterministic Unsatisfiability.* Each contradictory instruction is guaranteed to be incompatible with the fixed scene configuration, eliminating ambiguous evaluation cases. (3) *Different Semantic Complexity.* Perturbations range from local attribute mismatches (V1–V3) to spatial inconsistencies (V4), enabling analysis of shallow versus deep linguistic grounding.

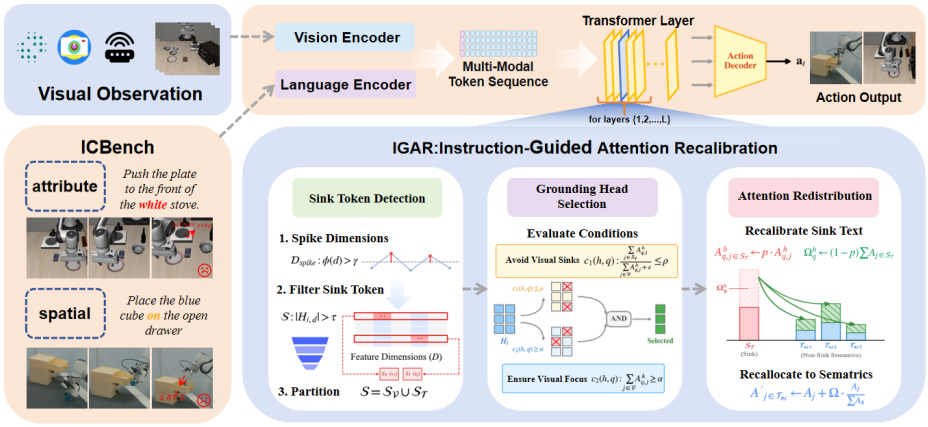

Fig. 2: Overview of the IGAR framework. IGAR is a train-free and plug-and-play intervention that restores linguistic grounding in VLA models via three stages: (1) detecting attention sink tokens through hidden-state spike analysis, (2) selecting grounding heads that exhibit cross-modal imbalance, and (3) redistributing attention from sink tokens to instruction tokens.

4 IGAR: Instruction-Guided Attention Recalibration

As illustrated in Fig. 2, we propose **I**nstruction-**G**uided **A**ttention **R**ecalibration (**IGAR**), a train-free inference-time mechanism that restores linguistic grounding in Vision-Language-Action (VLA) models by correcting vision-dominant attention sink imbalance during action prediction. We observe that, in modern VLA architectures, action-query tokens disproportionately attend to visually salient tokens, forming vision-dominant attention sinks that suppress instruction tokens even when the instruction semantically constrains the action. We attribute linguistic blindness to a structural cross-modal competition imbalance induced by attention sink dynamics. IGAR addresses this imbalance by detecting attention sinks through hidden-state spike analysis, selectively targeting cross-modal fusion heads, and recalibrating attention mass toward instruction tokens during inference. Our method operates entirely within the forward pass, requires no gradient updates or retraining, does not modify model parameters, and can be

seamlessly integrated into transformer-based VLA backbones without architectural changes, making it suitable as a plug-and-play module for deployed robotic policies.

4.1 Problem Formulation

Instruction Contradiction Setting. A VLA model f_θ maps observation $\boldsymbol{o}_t$ and a language instruction ℓ to action $\boldsymbol{a}_t = f_\theta(\boldsymbol{o}_t, \ell)$. Under ICBench, we evaluate the model using a semantically contradictory instruction $\tilde{\ell}$, while keeping the observation unchanged. By construction, $\tilde{\ell}$ is incompatible with the visual scene and cannot be satisfied through valid task execution. This setup isolates the role of language instruction grounding in action generation. The diagnostic evaluation is particularly important for safety-critical robotic deployment, where blindly executing an impossible instruction may lead to unsafe behaviors. A policy that genuinely integrates instruction semantics should fail or abstain under contradiction. A policy that relies primarily on visual priors may still succeed despite the semantic inconsistency.

Linguistic Grounding Score. Let $\mathrm{SR}(f_\theta, \ell)$ denote the task success rate of policy f_θ when executing under standard instruction ℓ.

We define the Linguistic Grounding Score (LGS) as:

$$\mathrm{LGS}(\tilde{\ell}) = \mathrm{SR}(f_\theta, \ell) - \mathrm{SR}(f_\theta, \tilde{\ell}), \tag{1}$$

where $\mathrm{SR}(f_\theta, \tilde{\ell})$ is the success rate respect to the standard instruction when executing the task using contradictory instruction $\tilde{\ell}$. As a result, a perfect grounded model should fail under contradiction, yielding $\mathrm{LGS}(\tilde{\ell}) = \mathrm{SR}(f_\theta, \ell)$. A linguistically insensitive model that ignores instruction semantics will succeed regardless of contradiction, yielding LGS$\approx$0.

4.2 Instruction-Guided Attention Recalibration

IGAR operates within the model's forward pass by selectively recalibrating attention distributions in the transformer layers. Rather than globally modifying attention weights, IGAR intervenes only in structurally imbalanced heads identified through attention sink analysis. The procedure consists of three stages: (1) attention sink token detection, (2) grounding head selection, and (3) attention redistribution.

Attention Sink Token Detection. Let $\boldsymbol{H} \in \mathbb{R}^{N \times D}$ denote the hidden states from an intermediate transformer layer, where N is the sequence length and D the hidden dimension.

We first compute the RMS norm per token:

$$r_i = \sqrt{\frac{1}{D}\sum_{d=1}^{D} H_{i,d}^2}, \quad i = 1, \ldots, N. \tag{2}$$

However, large token norms alone do not necessarily indicate structural dominance. To detect localized extreme activations, which is characteristic of attention sinks, we introduce a spike ratio per feature dimension:

$$\phi(d) = \frac{\max_i |H_{i,d}|}{\text{mean}_i |H_{i,d}| + \epsilon}, \tag{3}$$

where ϵ is a small constant for numerical stability. We select the top-k dimensions (k=5) from the set $\mathcal{D}_{\text{spike}} = \{d : \phi(d) > \gamma\}$ with γ=3.0. This spike criterion ensures that only dimensions with *localized* extreme activations (few tokens with high values, not all tokens uniformly high) are selected, avoiding false-positive sink identification. A token i is then classified as a **sink token** if:

$$\mathcal{S} = \left\{ i : \max_{d \in \mathcal{D}_{\text{spike}}} |H_{i,d}| > \tau \right\}, \quad \tau = 20. \tag{4}$$

We partition sinks into visual sinks $\mathcal{S}_\mathcal{V} = \mathcal{S} \cap \mathcal{V}$ and text sinks $\mathcal{S}_\mathcal{T} = \mathcal{S} \cap \mathcal{T}$.

Grounding Head Selection. Not all attention heads are equally susceptible to sink distortion. For each head h in layer l and each query position q beyond the image region, we evaluate two conditions:

$$c_1(h, q) : \frac{\sum_{j \in \mathcal{S}_\mathcal{V}} A^h_{q,j}}{\sum_{j \in \mathcal{V}} A^h_{q,j} + \epsilon} \leq \rho, \tag{5}$$

$$c_2(h, q) : \sum_{j \in \mathcal{V}} A^h_{q,j} \geq \alpha. \tag{6}$$

A head-query pair (h, q) is selected for reallocation if both conditions hold, with ρ=0.4 and α=0.01. Condition c_1 ensures the head is not already dominated by visual sinks (structural attention rather than semantic), while c_2 ensures the head allocates meaningful attention to visual tokens (filtering out text-only heads). When no visual sinks are detected (*i.e.*, $\mathcal{S}_\mathcal{V}$=$\emptyset$), c_1 is trivially satisfied and selection proceeds on c_2 alone.

Attention Redistribution. For each selected head-query pair (h, q), IGAR redistributes attention from sink tokens to non-sink tokens by first scaling down the attention of sink tokens by a factor $p = 0.6$. This operation frees a total budget

$$\Omega^h_q = (1 - p) \sum_{j \in \mathcal{S}_\mathcal{T}} A^h_{q,j} \tag{7}$$

which is then reallocated to non-sink tokens proportionally to their original attention weights:

$$A'^h_{q,j} = A^h_{q,j} + \Omega^h_q \cdot \frac{A^h_{q,j}}{\sum_{j' \in \mathcal{T}_{\text{ns}}} A^h_{q,j'} + \epsilon}, \quad \forall\, j \in \mathcal{T}_{\text{ns}} \tag{8}$$

where $\mathcal{T}_{\text{ns}} = \mathcal{T} \setminus \mathcal{S}_\mathcal{T}$ denotes the set of non-sink text tokens.

5 Experiments

We evaluate the proposed Instruction-Guided Attention Recalibration (IGAR) under the controlled OOD instruction contradiction setting introduced by ICBench to answer three key questions:

- **Q1:** Do current Vision-Language-Action (VLA) models exhibit linguistic blindness under contradictory instructions?
- **Q2:** Can IGAR restore language grounding during action generation?
- **Q3:** Does IGAR preserve performance on valid in-distribution tasks?

5.1 Experimental Setup

VLAs and Benchmark. We evaluate three representative VLA architectures (π_0 [4], $\pi_{0.5}$ [3], and OpenVLA-OFT [15]) across 30 simulated robot manipulation tasks from the LIBERO benchmark [20], covering *Spatial*, *Object*, and *Goal* suites. For each task variant, we perform 50 independent rollouts. These environments cover diverse robotic manipulation settings, including spatial reasoning, object manipulation, and goal-conditioned tasks. Our ICBench is constructed by introducing controlled contradictory instructions into the LIBERO tasks.

Evaluation Metrics. We report two primary evaluation metrics. Task Success Rate (SR) measures the percentage of successful task completions. Under standard instructions, a high SR indicates correct task execution and is therefore desirable. However, under the ICBench contradiction setting, a high SR becomes undesirable, as it implies that the policy completes the task despite contradictory instructions, suggesting that the model relies primarily on visual priors rather than instruction semantics. To quantify the influence of language instructions on action generation, we adopt the Linguistic Grounding Score (LGS) defined in Sec. 4.1. LGS measures the performance gap between executions under valid instructions and contradictory instructions, with higher values indicating stronger reliance on linguistic instructions.

Implementation Details. IGAR is applied uniformly at inference time without any task-specific tuning. For attention-based decoders ($\pi_{0.5}$ and OpenVLA-OFT), interventions span the initial 16 layers using unified constraints for visual-sink portion boundaries (ρ=0.4) and text-sink decay (p=0.6), alongside fixed detection thresholds (τ=20, γ=3.0). For π_0's continuous flow-matching diffusion, we deploy a discrete contrastive formulation that restricts divergence triggers (τ_{detect}=0.15) and interpolation boundaries ($\tau_{\max}$=0.50). This standardized configuration isolates the inherent robustness of IGAR across fundamentally disparate generation paradigms.

5.2 Diagnosing Linguistic Blindness

Table 1 presents the VLA baseline evaluation under ICBench, comparing performance under normal instructions and contradictory instructions. Across all three VLAs, we observe a consistent pattern: modern VLA models frequently ignore

contradictory language inputs and continue executing visually plausible actions, revealing a systemic form of *linguistic blindness*. First, models often maintain extremely high success rates even when the instruction is semantically invalid. Across the *Spatial* and *Object* suites, success rates frequently exceed 90% under contradictory instructions, particularly for $\pi_{0.5}$ and OpenVLA-OFT. This indicates that the policy continues to complete the task by relying on visual cues rather than instruction semantics. As a consequence, the resulting LGS remain very low, suggesting that language contributes minimally to action generation in these scenarios. Second, the severity of linguistic blindness varies across architectures. π_0 shows relatively higher LGS values compared to $\pi_{0.5}$ and OpenVLA-OFT, indicating slightly stronger language grounding. In contrast, $\pi_{0.5}$ often exhibits low LGS values, implying that contradictory instructions have almost no influence on the executed actions. These results demonstrate that despite their multimodal design, current VLA models often rely on vision-dominant execution strategies. Language instructions exert limited influence, though they directly constrain the physical feasibility of the action, highlighting a critical vulnerability for deploying VLA systems in real-world settings.

Table 1: Linguistic Blindness Diagnosis on ICBench. Success Rate (SR, %) and Linguistic Grounding Score (LGS) under normal and contradictory instructions. High SR and low LGS under contradiction (V1-V4) indicate that the policy ignores instruction semantics and follows visual priors. Highlighted cells denote SR $\geq 90\%$, revealing severe linguistic blindness.

		π_0		$\pi_{0.5}$		**OpenVLA-OFT**	
Suite	**Instruction**	SR	LGS	SR	LGS	SR	LGS
Spatial	Normal	96.8	–	97.4	–	97.6	–
	V1 (operand attribute)	90.4	6.4	96.2	1.2	97.8	-0.2
	V2 (target attribute)	96.2	0.6	97.8	-0.4	96.4	1.2
	V3 (dual attribute)	89.8	7.0	96.4	1.0	96.2	1.4
	V4 (spatial relation)	92.4	4.4	97.6	-0.2	92.4	5.2
Object	Normal	98.8	–	98.4	–	98.4	–
	V1 (operand attribute)	93.8	5.0	96.2	2.2	97.8	0.6
	V2 (target attribute)	90.2	8.6	90.4	8.0	96.2	2.2
	V3 (dual attribute)	98.2	0.6	96.4	2.0	99.8	-1.4
	V4 (spatial relation)	91.6	7.2	88.2	10.2	99.6	-1.2
Goal	Normal	95.8	–	97.6	–	98.0	–
	V1 (operand attribute)	90.2	5.6	93.8	3.8	97.8	0.2
	V2 (target attribute)	84.4	11.4	90.4	7.2	98.2	-0.2
	V3 (dual attribute)	88.2	7.6	94.2	3.4	98.4	-0.4
	V4 (spatial relation)	76.4	19.4	93.6	4.0	90.2	7.8

5.3 IGAR Restores Linguistic Grounding

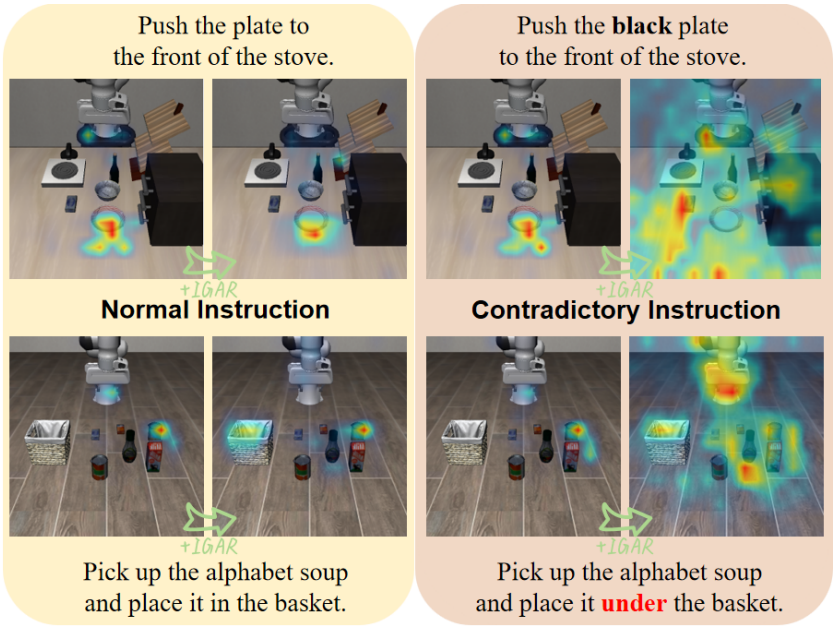

Fig. 3: Attention visualization of OpenVLA-OFT with and without IGAR. We visualize cross-modal attention maps under normal and contradictory instructions. The baseline policy attends primarily to salient regions regardless of instruction semantics, while IGAR redistributes attention toward instruction-relevant objects and spatial regions, mitigating visual attention sinks and improving linguistic grounding.

Table 2 reports the performance of VLA models under contradictory instructions after applying IGAR. Compared with the baseline results in Table 1, IGAR consistently reduces erroneous task execution while substantially increasing the LGS, indicating that action generation becomes more sensitive to instruction semantics. First, IGAR effectively suppresses visually driven execution under contradictory instructions. Across all suites, SR drops significantly compared to the baseline setting, particularly for the π_0 and OpenVLA-OFT. For example, in the *Goal* suite, SR decreases to as low as 36.4% under spatial contradictions (V4), while LGS increases dramatically to 59.4. This indicates that the policy refrains from executing actions that violate instruction semantics, demonstrating restored language grounding. Second, the improvement varies across model architectures and suites. The largest gains are observed for π_0, where LGS consistently exceeds 40 and reaches up to 59.4 in goal suites. OpenVLA-OFT also shows substantial improvements, achieving LGS values above 30 in several goal tasks. In contrast, the $\pi_{0.5}$ model exhibits more limited gains, showing strong reliability on visual cues. Overall, the results demonstrate that IGAR effectively restores

the influence of language instructions during action generation. By redistributing attention toward instruction tokens, the method mitigates vision-dominant biases and enables VLA policies to better detect and respond to semantically inconsistent instructions.

Table 2: IGAR mitigation under ICBench. We evaluate VLA models under contradictory instructions after applying IGAR. SR measures task completion despite contradiction (lower is better), while LGS measures the influence of language instructions (higher is better). Green cells indicates LGS ≥ 10.

Suite.	Contradiction	π_0		$\pi_{0.5}$		OpenVLA-OFT	
		SR	LGS	SR	LGS	SR	LGS
Spatial	V1	76.4	20.4	95.8	1.6	86.4	11.2
	V2	84.2	12.6	93.6	3.8	86.2	11.4
	V3	80.4	16.4	94.8	2.6	90.2	7.4
	V4	76.2	20.6	99.6	-2.2	88.4	9.2
Object	V1	88.2	10.6	94.2	4.2	88.4	10.0
	V2	86.4	12.4	90.4	8.0	84.2	14.2
	V3	93.6	5.2	87.6	10.8	88.2	10.2
	V4	90.4	8.4	88.4	10.0	82.4	16.0
Goal	V1	46.4	49.4	90.2	7.4	66.4	31.6
	V2	40.2	55.6	82.8	14.8	70.2	27.8
	V3	46.2	49.6	92.4	5.2	64.2	33.8
	V4	36.4	59.4	96.2	1.4	58.4	39.6

5.4 Baseline Preservation Under IGAR

Table 3 evaluates whether IGAR affects performance under normal (non-contradictory) instructions. We compare the VLA baseline success rates with results obtained when IGAR is applied to the same tasks. Overall, IGAR preserves baseline performance with only marginal impact. Across all suites, the π_0 model shows negligible performance changes, with an average decrease of only 0.4%. Similarly, OpenVLA-OFT maintains almost identical performance, with an average change of +0.5%. These results indicate that the attention recalibration introduced by IGAR does not interfere with correct instruction following when the language input is consistent with the visual scene. Furthermore, we measure per-step inference latency on a single NVIDIA RTX 5090 GPU over 50 trials. For OpenVLA-OFT, latency increases from 98.7 ± 2.0 ms to 105.6 ± 0.5 ms, corresponding to a modest +7.0 ms overhead.

5.5 Ablation Study and Hyperparameter Sensitivity

We systematically analyze the influence of key hyperparameters governing IGAR's intervention constraints using OpenVLA-OFT on the `libero_goal` benchmark.

Table 3: Baseline preservation under IGAR. Success Rate (SR) comparison under normal (non-contradictory) instructions for baseline and baseline with our IGAR. "OFT" refers to "OpenVLA-OFT".

Suites	$\boldsymbol{\pi_0}$	$\boldsymbol{\pi_0}$+IGAR	$\boldsymbol{\pi_{0.5}}$	$\boldsymbol{\pi_{0.5}}$+IGAR	**OFT**	**OFT**+IGAR
Spatial	96.8	96.4 (−0.4)	97.4	98.2 (+0.8)	97.6	97.6 (+0.0)
Object	98.8	98.2 (−0.6)	98.4	94.4 (−4.0)	98.4	99.8 (+1.4)
Goal	95.8	95.6 (−0.2)	97.6	96.4 (−1.2)	98.0	98.2 (+0.2)
Average	97.1	96.7 (−0.4)	97.8	96.3 (−1.5)	98.0	98.5 (+0.5)

As detailed in Figure 4, we evaluate the LGS to measure the effectiveness of language-action grounding. Applying an overly aggressive text-sink decay factor ($p < 0.3$) or an insufficient one ($p > 0.8$) disrupts the optimal reinforcement of instructional tokens, directly reducing the overall LGS. Similarly, excessively loose visual-sink bounds ($\rho > 0.6$) fail to adequately intercept structural visual heads responsible for attention sinks, while overly strict boundaries interfere with normal processing. Furthermore, intervening across too many transformer layers ($L > 24$) tends to corrupt deep representations that are highly specialized for continuous action regression, whereas intervening on too few layers fails to provide a strong enough corrective signal. Our default configurations ($p = 0.6$, $\rho = 0.4$, $L = 16$) strategically target the mid-level layers where cross-modal semantic fusion primarily occurs, fully maximizing the LGS and effectively mitigating linguistic blindness.

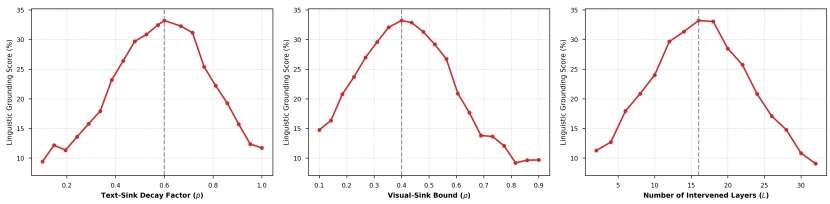

Fig. 4: Hyperparameter Sensitivity. Impact of the text-sink decay factor (p), head selection bound (ρ), and number of intervened layers (L) in terms of linguistic grounding performance (LGS). Results are reported using the OpenVLA-OFT architecture on the libero_goal benchmark suite. The dashed lines indicate the selected values.

We further ablate the three stages of IGAR using $\pi_{0.5}$ on LIBERO-Goal under ICBench V1. As shown in Table 4, full IGAR increases attention-to-text from 0.08 to 0.15 and improves LGS from 3.8 to 7.4. Removing sink detection or head selection weakens the effect, while detection alone leaves both attention-to-text and LGS unchanged. These results verify that sink detection, grounding-head selection, and attention redistribution are complementary rather than redundant.

Table 4: Component ablation of IGAR. Results are reported using $\pi_{0.5}$ on LIBERO-Goal under ICBench V1.

Config	Sink Det.	Head Sel.	Redist.	Attn-to-Text (↑)	LGS (↑)
Baseline ($\pi_{0.5}$)	–	–	–	0.08	3.8
Full IGAR	✓	✓	✓	**0.15**	**7.4**
w/o Sink Detection	–	✓	✓	0.11	4.6
w/o Head Selection	✓	–	✓	0.10	4.2
Detection only	✓	–	–	0.08	3.8

5.6 Real-World Evaluation

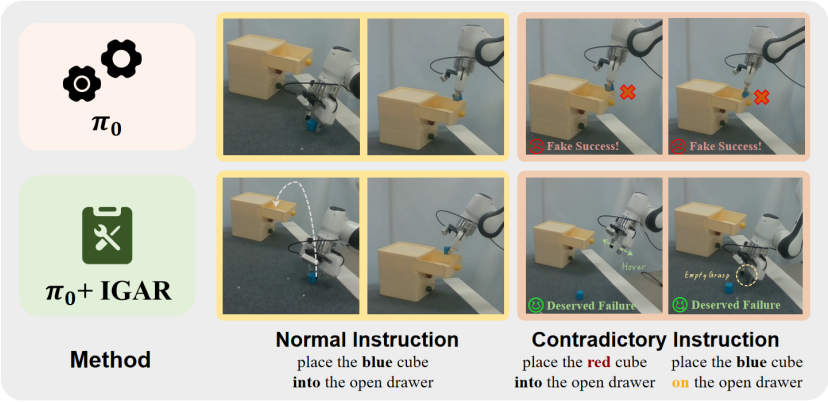

Fig. 5: Real-world experiments. We test the π_0 policy with and without IGAR on the task "placing the blue cube into the open drawer". Under normal instructions (left), both policies successfully place the blue cube into the open drawer. Under contradictory instructions (right), the π_0 policy still executes a visually plausible trajectory and produces a *fake success*. In contrast, IGAR restores linguistic grounding and prevents incorrect task execution, resulting in a *deserved failure*.

We further validate IGAR on a real robotic platform to examine whether the linguistic grounding improvements observed in simulation transfer to physical manipulation. As shown in Fig. 5, experiments are conducted on a Franka Research 3 robotic arm equipped with two Intel RealSense D435 cameras, one mounted on the wrist and another providing a third-person viewpoint. We evaluate the π_0 policy on a cube placement task: *"place the blue cube into the open drawer"*. Under normal instructions, both the π_0 policy and π_0+IGAR reliably complete the task, confirming that IGAR preserves the original policy behavior when the instruction is consistent with the scene. However, when the instruction becomes contradictory (e.g., requesting a non-existent object attribute or an impossible spatial relation), the baseline policy frequently continues executing the

visually plausible trajectory and completes the manipulation despite the semantic inconsistency. This behavior results in *fake success*, where the robot performs the correct physical action but violates the instruction semantics. In contrast, IGAR restores the influence of language during action generation. Given contradictory instructions, the recalibrated policy refrains from completing the task and instead produces safe behaviors such as hovering or empty grasp attempts, effectively interrupting the manipulation. These outcomes represent *deserved failures*, indicating that the policy correctly detects the instruction inconsistency.

6 Conclusion

In this paper, we have revealed a critical reliability issue in Vision-Language-Action (VLA) models that we term *linguistic blindness*, where policies prioritize visual priors over instruction semantics during action generation. To systematically diagnose this failure mode, we introduce ICBench, a controlled benchmark that evaluates language–action grounding under structured contradictory instructions. We further propose Instruction-Guided Attention Recalibration (IGAR), a train-free inference-time intervention that restores linguistic grounding by redistributing attention from sink tokens to instruction tokens. Experiments on three representative VLA architectures demonstrate that modern models often execute visually plausible actions even when instructions are logically inconsistent with the scene, while IGAR effectively mitigates such failures without degrading baseline task performance.

Acknowledgment

This project is supported by the Ministry of Education (MOE), Singapore, under its Academic Research Fund (AcRF) Tier 2 (Proposal ID: T2EP20125-0048). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore. This project was also supported by NSFC Project (No. 62522206) and Tsinghua University Initiative Scientific Research Program (Student Academic Research Advancement Program).

References

1. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
2. Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734 (2025)
3. Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M.R., Finn, C., Fusai, N., Galliker, M.Y., Ghosh, D., Groom, L., Hausman, K., brian ichter, Jakubczak, S., Jones, T., Ke, L., LeBlanc, D., Levine, S., Li-Bell, A.,

Mothukuri, M., Nair, S., Pertsch, K., Ren, A.Z., Shi, L.X., Smith, L., Springenberg, J.T., Stachowicz, K., Tanner, J., Vuong, Q., Walke, H., Walling, A., Wang, H., Yu, L., Zhilinsky, U.: Pi0.5: a vision-language-action model with open-world generalization. In: 9th Annual Conference on Robot Learning (2025)
4. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: pi0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
5. 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 preprint arXiv:2212.06817 (2022)
6. Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R., et al.: Do as i can, not as i say: Grounding language in robotic affordances. In: Conference on robot learning. pp. 287–318. PMLR (2023)
7. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
8. Fei, S., Wang, S., Shi, J., Dai, Z., Cai, J., Qian, P., Ji, L., He, X., Zhang, S., Fei, Z., et al.: Libero-plus: In-depth robustness analysis of vision-language-action models. arXiv preprint arXiv:2510.13626 (2025)
9. Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024)
10. Glossop, C., Chen, W., Bhorkar, A., Shah, D., Levine, S.: Cast: Counterfactual labels improve instruction following in vision-language-action models. arXiv preprint arXiv:2508.13446 (2025)
11. Guiochet, J., Machin, M., Waeselynck, H.: Safety-critical advanced robots: A survey. Robotics and Autonomous Systems **94**, 43–52 (2017)
12. Jiao, P., Zhu, B., Chen, J., Ngo, C.W., Jiang, Y.G.: Don't deceive me: Mitigating gaslighting through attention reallocation in lmms. arXiv preprint arXiv:2504.09456 (2025)
13. Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models. In: The Thirteenth International Conference on Learning Representations (2025)
14. 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. arXiv preprint arXiv:2403.12945 (2024)
15. Kim, M.J., Finn, C., Liang, P.: Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint arXiv:2502.19645 (2025)
16. 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: 8th Annual Conference on Robot Learning (2024)
17. Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., Wang, C., Levine, G., Lingelbach, M., Sun, J., et al.: Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In: Conference on Robot Learning. pp. 80–93. PMLR (2023)
18. Li, H., Wang, J., Mei, Z., Majumdar, A., Chen, J., Zhu, B.: Robotrustbench: Benchmarking the trustworthiness of video world models for robotic manipulation. arXiv preprint arXiv:2606.01600 (2026)
19. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)

20. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. Advances in Neural Information Processing Systems **36**, 44776–44791 (2023)
21. Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H., Zhu, J.: Rdt-1b: a diffusion foundation model for bimanual manipulation. In: The Thirteenth International Conference on Learning Representations (2025)
22. 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. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6892–6903. IEEE (2024)
23. Peng, Z., Ding, W., You, Y., Chen, Y., Luo, W., Tian, T., Cao, Y., Sharma, A., Xu, D., Ivanovic, B., et al.: Counterfactual vla: Self-reflective vision-language-action model with adaptive reasoning. arXiv preprint arXiv:2512.24426 (2025)
24. Tang, Z., Jiao, P., Zhu, B., Qi, H., Chen, J., Jiang, Y.G.: Spatiotemporal sycophancy: Negation-based gaslighting in video large language models. In: Findings of the Association for Computational Linguistics: ACL 2026. pp. 14836–14852 (2026)
25. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
26. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al.: Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 (2025)
27. Team, G.R., Choromanski, K., Devin, C., Du, Y., Dwibedi, D., Gao, R., Jindal, A., Kipf, T., Kirmani, S., Leal, I., et al.: Evaluating gemini robotics policies in a veo world simulator. arXiv preprint arXiv:2512.10675 (2025)
28. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
29. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: The Twelfth International Conference on Learning Representations (2024)
30. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations (2023)
31. Zacharaki, A., Kostavelis, I., Gasteratos, A., Dokas, I.: Safety bounds in human robot interaction: A survey. Safety science **127**, 104667 (2020)
32. Zhang, B., Zhang, Y., Ji, J., Lei, Y., Dai, J., Chen, Y., Yang, Y.: Safevla: Towards safety alignment of vision-language-action model via constrained learning. arXiv preprint arXiv:2503.03480 (2025)
33. Zhang, S., Xu, Z., Liu, P., Yu, X., Li, Y., Gao, Q., Fei, Z., Yin, Z., Wu, Z., Jiang, Y.G., et al.: Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11142–11152 (2025)
34. Zhou, S., Zhu, B., Yang, J., Zhao, X., Chen, J., Jiang, Y.G.: Rc-nf: Robot-conditioned normalizing flow for real-time anomaly detection in robotic manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 43050–43060 (2026)
35. Zhu, B., Gui, Y., Qi, H., Chen, J., Ngo, C.W., Lim, E.P.: Benchmarking gaslighting negation attacks against multimodal large language models. In: Proceedings of the 2026 International Conference on Multimedia Retrieval. pp. 2041–2049 (2026)

36. Zhu, B., Yin, H., Chen, J., Jiang, Y.G.: Benchmarking gaslighting negation attacks against reasoning models. In: International Conference on Multimedia Modeling. pp. 188–202. Springer (2026)
37. 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: Conference on Robot Learning. pp. 2165–2183 (2023)