R2G-VLA: RISK- AND ROLE-GUIDED GROUNDING CORRECTION FOR ROBUST VLA POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) policies have advanced language-conditioned manipulation, yet their closed-loop robustness remains limited by visual grounding shifts. In real scenes, task semantics may stay unchanged while target positions, receiver locations, object relations, or viewpoints change, inducing visual-token-level distribution shifts in frozen policies. Such shifts can mislead action prediction through distractors or perturbed relational cues, making inference-time grounding correction important for robust VLA execution during deployment. We propose R2G-VLA, a risk- and role-guided grounding correction framework for robust VLA policies under physical scene perturbations. R2G-VLA uses action-token uncertainty as a self-diagnostic signal and propagates it to visual tokens to estimate token-level grounding risk. Instead of intervening on scattered high-response tokens, it aggregates risky tokens into spatially coherent regions that better match object-level perturbations in manipulation scenes. To avoid role-agnostic suppression, R2G-VLA further performs semantic role-aware token correction over target, receiver, support, background, and distractor regions, preserving task-relevant evidence while reducing misleading visual cues. The corrected visual representation is re-injected into the same policy for a second forward pass, and a safety gate accepts the correction only when uncertainty and action-shift constraints are satisfied. Experiments on LIBERO-Spatial show that R2G-VLA improves robustness across five physical scene perturbations, highlighting the effectiveness of risk- and role-guided internal grounding correction.

1 INTRODUCTION

Vision-language-action policies have become a central paradigm for language-conditioned robotic manipulation, enabling closed-loop action prediction from visual observations and natural-language instructions. Despite recent progress, their robustness remains limited under task-preserving scene perturbations. In deployment, the instruction may remain semantically valid while the scene configuration changes through target displacement, receiver motion, relational layout variation, similar distractors, or viewpoint shifts. These perturbations differ from standard image corruptions or language variations, as they modify the task-relevant grounding evidence in the physical workspace. As a result, VLA policies can suffer from visual grounding drift, where action prediction is biased by perturbed object cues, distracting regions, or misaligned relational evidence.

Addressing such grounding drift is especially challenging when foundation-scale VLA policies are deployed with frozen weights or inaccessible parameters. Re-collecting demonstrations and fine-tuning for every scene variation is often impractical, motivating run-time or zero-shot adaptation methods (Kim et al., 2024; Hancock et al., 2025; Heo et al., 2026). Recent robustness studies further show that VLA policies can degrade under physical, visual, and multimodal perturbations, indicating that these failures go beyond low-level observation noise (Hancock et al., 2025; Liu et al., 2025; Guo et al., 2025). These findings motivate inference-time correction mechanisms that preserve the pretrained policy while adjusting the visual evidence used for action prediction. Such correction, however, cannot be reduced to simply removing sensitive visual tokens. Under scene-level perturbations, high-risk tokens may arise from a shifted target, a moved receiver, a support surface, background clutter, or a distractor. Although these tokens can all affect action uncertainty, they play different roles in manipulation and should not be treated as equivalent evidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing VLA robustness methods address different parts of this problem, but they do not fully resolve this risk-role mismatch. Run-time observation intervention methods edit the input observation to reduce task-irrelevant visual influence before policy inference (Hancock et al., 2025). Token pruning methods move the intervention into the transformer representation space by removing distracting visual tokens (Li et al., 2026). Object-centric grounding methods improve clutter resistance through geometry-aware grounding and policy adaptation (Vo et al., 2025). Robust training and perturbation-aware optimization improve policy behavior under multimodal disturbances, but require additional optimization or adapted policies (Guo et al., 2025). These approaches are effective in their respective settings, yet they either modify the observation, require adaptation, or primarily treat perturbation-sensitive evidence as content to edit, prune, or suppress. Consequently, robust frozen-policy inference requires internal grounding correction that jointly considers token-level risk and manipulation role, rather than role-agnostic suppression of sensitive visual evidence.

Following this perspective, we propose R2G-VLA, a risk- and role-guided grounding correction framework for frozen VLA policies under physical scene perturbations. R2G-VLA leaves both the input observation and policy parameters unchanged, and instead performs model-update-free correction on intermediate visual tokens during inference. It first uses action-token uncertainty as a self-diagnostic signal and propagates it back to visual tokens, producing a grounding-risk map that identifies visual evidence associated with unstable action prediction. To better match the object-level structure of physical perturbations, high-risk tokens are aggregated into spatially connected regions rather than treated as isolated token cells. Within these regions, R2G-VLA incorporates task-level role information to distinguish target, receiver, support, background, and distractor tokens, and applies role-dependent correction strengths to preserve task-critical grounding cues while attenuating misleading evidence. The corrected visual representation is then re-injected into the same frozen policy for reinference. A conservative safety gate accepts the corrected action only when uncertainty and action-shift constraints indicate reliable intervention, otherwise retaining the original policy output. By combining uncertainty-driven risk localization, spatially coherent region selection, and role-conditioned token correction, R2G-VLA improves robustness without input editing, model updates, or role-agnostic token suppression.

In summary, our key contributions are as follows:

- We formulate robustness to physical scene perturbations as a policy-internal grounding correction problem for frozen VLA policies, and propose R2G-VLA, a risk- and role-guided framework that improves robust action prediction without editing the input observation or updating model parameters.
- We introduce an action-uncertainty grounding risk map with connected-region aggregation. It propagates action uncertainty to visual tokens and identifies spatially coherent high-risk evidence that destabilizes action prediction.
- We develop semantic role-guided token correction with conservative safety-gated reinference. Instead of uniformly suppressing perturbation-sensitive tokens, R2G-VLA modulates risky visual evidence by manipulation role and rejects unreliable corrections to preserve stable frozen-policy execution.

2 Related Work

2.1 Robust VLA Policies under Physical Perturbations

Vision-language-action (VLA) policies have become a central direction for language-conditioned robotic manipulation, with representative systems such as RT-1 (Brohan et al., 2022), RT-2 (Zitkovich et al., 2023), OpenVLA (Kim et al., 2024), Octo (Team et al., 2024), and π_0 (Black et al., 2024). These models show strong potential as general-purpose robot policies, but recent studies also reveal their vulnerability to visual clutter, scene variation, and physical perturbations (Cheng et al., 2024; Liu et al., 2025). BYOVLA performs run-time observation intervention by editing task-irrelevant image regions before action prediction (Hancock et al., 2025). RobustVLA and STRONG-VLA improve robustness through perturbation-aware training or post-training optimization (Guo et al., 2025; Xie et al., 2026). Other methods study viewpoint adaptation or generative robustness for visual manipulation (Heo et al., 2026; Zhang et al., 2025). While these works improve VLA robustness from different perspectives, they often operate at the input level, rely on additional training

or adaptation, or do not explicitly correct internal grounding evidence. R2G-VLA instead focuses on correcting risky visual-token evidence inside a frozen VLA policy during inference.

2.2 Grounding Risk Localization and Region Aggregation

Attribution and interpretability methods estimate which visual regions or intermediate features contribute to model predictions using perturbation, gradient, or local surrogate signals (Ribeiro et al., 2016; Selvaraju et al., 2017). Transformer-specific analyses further study token-level information flow through attention rollout or relevance propagation (Abnar & Zuidema, 2020; Chefer et al., 2021). Token selection and pruning methods reduce computation by selecting, merging, or reorganizing informative visual tokens (Dosovitskiy et al., 2020; Rao et al., 2021; Bolya et al., 2022; Liang et al., 2022). More closely related to VLA robustness, DTP prunes distracting image tokens inside VLA models (Li et al., 2026). However, these methods mainly address post-hoc explanation, computational efficiency, or distracting-token removal. They do not directly localize visual tokens that increase action uncertainty under physical perturbations. In contrast, R2G-VLA constructs an action-uncertainty-driven grounding risk map and aggregates high-risk tokens into connected regions, matching the spatially coherent structure of object-level perturbations.

2.3 Semantic Role-Aware Grounding Correction

Grounding is essential for language-conditioned manipulation, where policies must align instructions with task-relevant visual evidence (Ahn et al., 2022). Transporter Networks, CLIPort, and PerAct ground manipulation actions in spatial, semantic, or voxelized scene representations (Zeng et al., 2021; Shridhar et al., 2021; 2023), while broader manipulation policies, visual representations, and benchmarks study action grounding across diverse tasks and embodiments (James et al., 2020; Chi et al., 2025). Recent VLA robustness methods move closer to grounding correction: OBEYED-VLA improves clutter resistance with object-centric and geometry-aware grounding (Vo et al., 2025), and DTP suppresses distracting visual tokens (Li et al., 2026). However, existing methods usually focus on finding task-relevant objects or suppressing distracting evidence, but do not explicitly decide how risky evidence should be corrected according to its manipulation role. R2G-VLA assigns risky tokens to target, receiver, support, background, and distractor roles, enabling task-relevant evidence to be preserved while misleading background or distractor cues are more strongly corrected.

3 Method

3.1 Problem Formulation

We consider an instruction-following manipulation setting with a frozen vision-language-action (VLA) policy π_θ. At time step t, the policy receives an RGB observation I_t and a language instruction L, and predicts an action $a_t = \pi_\theta(I_t, L)$. The policy parameters θ remain fixed throughout inference, and no additional training objective or parameter update is introduced.

Internally, the VLA policy encodes the visual observation into a set of visual tokens $V = \{v_i\}_{i=1}^{N}$ arranged on a spatial token grid. Each token corresponds to a local visual region and provides part of the grounding evidence used for action prediction. Under physical scene perturbations, the task semantics may remain unchanged while the observed workspace configuration changes, such as target positions, receiver locations, object relations, distractors, or camera viewpoints. These changes can shift the visual-token-level grounding evidence used by the frozen policy and lead to unstable action predictions.

Given the perturbed observation and instruction, the frozen policy first produces an initial action

$$a_0 = \pi_\theta(I_t, L). \tag{1}$$

Our goal is to obtain a robust action a^* through an inference-time correction procedure that operates on intermediate visual tokens:

$$a^* = \mathcal{C}_\phi(\pi_\theta, I_t, L). \tag{2}$$

Here, $\mathcal{C}_\phi$ denotes a non-trainable correction procedure controlled only by inference hyperparameters ϕ. It does not modify the input image or update the VLA policy. Instead, R2G-VLA corrects visual-token-level grounding evidence inside the frozen policy and returns the corrected action only when conservative safety constraints are satisfied.

3.2 OVERVIEW OF R2G-VLA

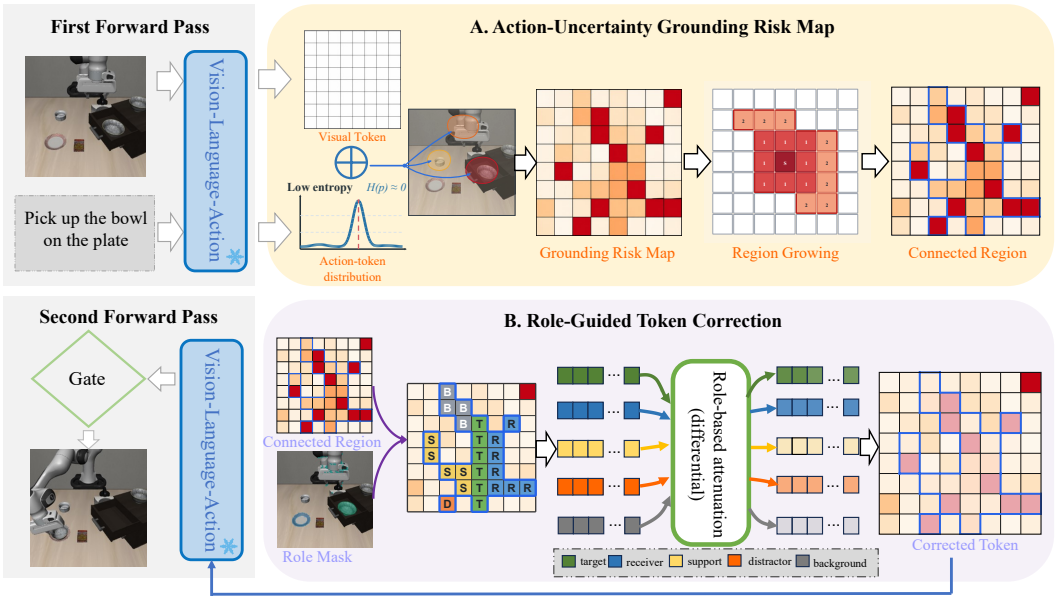

Figure 1: Overview of R2G-VLA. R2G-VLA performs model-update-free grounding correction inside a frozen VLA policy. It uses action uncertainty to localize high-risk visual-token regions, applies role-guided token correction, and accepts the corrected action through a conservative safety gate.

R2G-VLA performs two-pass inference within the same frozen VLA policy. The first pass processes (I_t, L) and produces the initial action a_0, action-token distributions, and visual tokens V. We back-propagate action-token entropy to V to estimate grounding risk, then aggregate high-risk tokens into spatially connected regions.

Selected tokens are assigned target, receiver, support, background, or distractor roles and reweighted accordingly. The corrected tokens V' are re-injected for a second pass, yielding action a_1 and uncertainty U_1. A safety gate accepts a_1 when uncertainty and action-shift constraints are satisfied; otherwise, it retains a_0. The complete algorithm, OpenVLA hook location, and implementation safeguards are provided in the supplementary material.

3.3 ACTION-UNCERTAINTY GROUNDING RISK MAP

The first core module of R2G-VLA identifies which visual tokens introduce grounding risk for the frozen policy. Given the first forward pass, let $\{p_m^0\}_{m=1}^M$ denote the action-token distributions predicted by the VLA policy, where M is the number of action tokens. We measure the initial action uncertainty by the mean entropy over these distributions:

$$U_0 = \frac{1}{M}\sum_{m=1}^{M} H(p_m^0), \qquad H(p_m^0) = -\sum_{c} p_m^0(c)\log p_m^0(c). \tag{3}$$

Here, U_0 is used as an internal self-diagnostic signal. A high uncertainty indicates that the frozen policy is less confident about its action prediction under the current visual observation and language instruction.

To locate the visual evidence responsible for this uncertainty, we backpropagate U_0 to the intermediate visual tokens $V = \{v_i\}_{i=1}^N$ while keeping all policy parameters fixed. For each visual token,

we compute the gradient magnitude:

$$g_i = \left\| \frac{\partial U_0}{\partial v_i} \right\|_2 . \tag{4}$$

A larger g_i indicates that small changes to token v_i would more strongly affect the action uncertainty. Therefore, tokens with high gradient magnitude are treated as riskier grounding evidence for the current action prediction. This gradient computation does not require environment gradients, labels, or parameter updates; it only probes the sensitivity of the frozen policy's action uncertainty to its own visual tokens.

We normalize the token scores over the visual-token grid to obtain the grounding risk map:

$$\tilde{g}_i = \frac{g_i - \min_j g_j}{\max_j g_j - \min_j g_j + \epsilon}, \tag{5}$$

where ϵ ensures numerical stability. The normalized map measures each token's local influence on action uncertainty and serves as a proxy for grounding risk, rather than generic visual saliency. Details of gradient capture, token extraction, numerical handling, and the visual-token hook are provided in the supplementary material.

3.4 Connected-Region Risk Aggregation

After obtaining the grounding risk map, a simple strategy would be to select the top-k tokens with the largest risk scores. However, physical perturbations in manipulation scenes usually affect spatially coherent object regions rather than isolated token cells. For example, a shifted target, a displaced receiver, or an added distractor often occupies a contiguous area on the visual-token grid. Selecting scattered top-k tokens may therefore produce fragmented evidence and lead to unstable correction.

R2G-VLA instead aggregates high-risk tokens into connected regions. Given the normalized risk score $\tilde{g}_i$, we first identify high-confidence seed tokens and candidate grow tokens:

$$S = \{i \mid \tilde{g}_i \geq \tau_s\}, \qquad B = \{i \mid \tilde{g}_i \geq \tau_g\}, \tag{6}$$

where τ_s and τ_g are the seed and grow thresholds, with $\tau_s > \tau_g$. Starting from seed tokens in S, each region is expanded by adding neighboring tokens from B according to spatial adjacency on the visual-token grid:

$$R = \mathrm{Grow}(S, B). \tag{7}$$

The selected set R contains one or more connected high-risk regions used for the subsequent correction step.

This connected-region aggregation encourages the selected tokens to follow the spatial footprint of perturbed objects or relational evidence. It also reduces sensitivity to isolated noisy gradients by requiring risky evidence to be spatially supported by neighboring tokens. As a result, R2G-VLA performs correction on coherent grounding regions rather than disconnected token responses.

3.5 Role-Guided Token Correction

The grounding risk map identifies where the frozen policy is sensitive, but high-risk tokens should not all be treated in the same way. In manipulation tasks, different visual regions play different semantic roles. Target and receiver regions usually contain task-critical evidence that should be preserved, support regions provide useful workspace context, while background and distractor regions are more likely to introduce misleading grounding cues. Therefore, R2G-VLA performs semantic role-guided correction rather than uniformly suppressing all risky tokens.

For each token in the selected connected region R, we assign a semantic label $r(i)$ from five predefined roles: target, receiver, support, background, and distractor. The role assignment is provided by an explicit role mask rather than learned as an additional module. In simulation, the role mask can be obtained from task metadata and object-level annotations, such as the target object, receiver object, support surface, and distractor objects. These masks are pooled onto the visual-token grid to determine the role of each selected token. The deterministic role-resolution interface, counterfactual mask construction, overlap handling, and validity checks are detailed in the supplementary material.

Given the role assignment, R2G-VLA applies role-dependent token reweighting:

$$v_i' = \begin{cases} \alpha_{r(i)} v_i, & i \in R, \\ v_i, & i \notin R. \end{cases} \tag{8}$$

where $\alpha_{r(i)}$ is the correction coefficient associated with the role of token i. We use milder attenuation for task-relevant roles and stronger attenuation for less reliable or misleading roles. In our implementation, we set $\alpha_{\text{target}} = 0.85$, $\alpha_{\text{receiver}} = 0.90$, and $\alpha_{\text{support}} = 0.80$ for task-relevant roles, and $\alpha_{\text{background}} = 0.60$ and $\alpha_{\text{distractor}} = 0.40$ for background and distractor regions, respectively.

Unlike role-agnostic suppression, this design retains more evidence from high-risk task-relevant tokens while more strongly attenuating background and distractor cues.

3.6 Second Forward Pass and Safety Gate

The corrected token set $V' = \{v_i'\}_{i=1}^N$ is re-injected into the frozen VLA policy for a second forward pass. This produces a corrected action a_1 and corrected uncertainty U_1. Since token correction may occasionally over-adjust the policy output, especially when the risk map is diffuse or the role assignment is uncertain, R2G-VLA uses a conservative safety gate before returning the final action:

$$\mathcal{G}(U_0, U_1, a_0, a_1) = \mathbf{1}\begin{bmatrix} U_1 \le \tau_U, \\ \|a_1 - a_0\|_2 \le \tau_a \end{bmatrix}. \tag{9}$$

$$a^* = \begin{cases} a_1, & \mathcal{G}(U_0, U_1, a_0, a_1) = 1, \\ a_0, & \mathcal{G}(U_0, U_1, a_0, a_1) = 0. \end{cases} \tag{10}$$

The gate in Eq. (9) accepts a_1 only when the corrected uncertainty remains below an absolute ceiling and the action displacement from a_0 to a_1 stays within a safe range; Eq. (10) otherwise falls back to the original action a_0. In the main experiments, we use $\tau_U = 2.05$ and $\tau_a = 1.25$. This gate preserves the benefits of test-time correction while limiting harmful changes under ambiguous observations.

4 Experiments

4.1 Experimental Setup

We evaluate R2G-VLA on the LIBERO-Spatial benchmark (Liu et al., 2023) using tasks 0-9. Each task is evaluated under one clean condition and five task-preserving physical scene perturbations, including target position shift, receiver position shift, relational layout shift, camera pose shift, and similar distractor. The language instruction and task success predicate are kept unchanged across clean and perturbed rollouts, so performance changes reflect robustness to physical grounding shifts rather than instruction mismatch.

Success rate is used as the primary metric. An episode is counted as successful if the simulator reports task completion within the rollout horizon. We compare R2G-VLA with frozen OpenVLA and a RobustVLA-style OpenVLA baseline trained with perturbation-aligned physical-proxy visual augmentations. The RobustVLA-style baseline is not an exact reproduction of the full RobustVLA training pipeline, but a perturbation-aligned physical-proxy training baseline built on OpenVLA. Its training configuration, proxy augmentations, and evaluation parity with the other methods are provided in the supplementary material. All methods are evaluated on 10 tasks with 5 seeds per condition, resulting in 50 episodes for each clean or perturbed condition. Unless otherwise specified, all R2G-VLA variants keep the VLA policy frozen and use the same inference hyperparameters across tasks, seeds, and perturbation types.

The five perturbations are designed to cover complementary grounding failure modes in language-conditioned manipulation. Target and receiver shifts evaluate object-level grounding under changed workspace configurations. Relational layout shift stresses spatial relation grounding. Camera pose shift tests viewpoint sensitivity without changing task semantics. Similar distractor evaluates whether the policy can separate task-relevant evidence from plausible competing objects. All perturbations are applied before rollout and kept fixed within each episode. We provide the full set of fixed inference hyperparameters, fallback rules, ablation configurations, exact perturbation construction, seed pairing, perturbation audit, and computing environment in the supplementary material.

4.2 MAIN RESULTS

Table 1: Success rate (%) on LIBERO-Spatial. R2G-VLA is compared with frozen OpenVLA and a RobustVLA-style physical-proxy training baseline under clean and physically perturbed scenes.

Condition	OpenVLA	RobustVLA-style	R2G-VLA
Clean	70.0	72.0	**78.0**
Target position	38.0	46.0	**56.0**
Receiver position	60.0	48.0	**68.0**
Relational layout	52.0	**54.0**	**54.0**
Camera pose	46.0	**56.0**	54.0
Similar distractor	38.0	**54.0**	50.0
Perturbed Avg.	46.8	51.6	**56.4**
All	50.7	55.0	**60.0**

Table 1 reports the success rates of OpenVLA, the RobustVLA-style physical-proxy training baseline, and R2G-VLA on LIBERO-Spatial. Compared with frozen OpenVLA, R2G-VLA improves the overall success rate from 50.7% to 60.0% and the perturbed average from 46.8% to 56.4%. Since R2G-VLA does not edit the input observation or update policy parameters, these results show that policy-internal grounding correction can improve deployment-time robustness while preserving the frozen-policy setting. Complete episode counts and reproducibility details are reported in the supplementary material.

R2G-VLA also outperforms the RobustVLA-style baseline, improving the overall success rate from 55.0% to 60.0% and the perturbed average from 51.6% to 56.4%. This comparison suggests that inference-time grounding correction provides a complementary robustness mechanism to training-time physical-proxy augmentation. Rather than learning robustness through additional perturbation-aligned training, R2G-VLA corrects the internal visual grounding evidence used for action prediction at test time.

Across individual perturbations, R2G-VLA achieves clear gains over OpenVLA on all five perturbed settings. The improvement is especially pronounced under target position shift, where the success rate increases from 38.0% to 56.0%, and receiver position shift, where it increases from 60.0% to 68.0%. These perturbations directly change the object or placement region that determines the correct manipulation action. The results therefore align with the design of R2G-VLA, where action-uncertainty risk mapping identifies unstable grounding evidence and role-guided correction modulates this evidence according to its manipulation role.

R2G-VLA also improves over OpenVLA under relational layout, camera pose, and similar distractor perturbations. The gain under relational layout is relatively modest, increasing from 52.0% to 54.0%, suggesting that relation-level shifts may involve more distributed scene evidence than localized object displacement. Under camera pose and similar distractor perturbations, the RobustVLA-style baseline remains competitive, achieving 56.0% and 54.0%, respectively. This indicates that perturbation-aligned visual augmentation can help with viewpoint changes and distractor appearances. Nevertheless, R2G-VLA achieves the best overall and perturbed-average performance without additional policy training, showing the advantage of model-update-free internal grounding correction.

Figure 2 visualizes the robustness improvement of R2G-VLA over OpenVLA across the five physical scene perturbations. The consistent improvement across perturbation types indicates that R2G-VLA does not only benefit a single failure, but provides broader robustness gains under task-preserving scene changes. The larger improvements on target, receiver, and distractor-related perturbations further support the motivation that physical scene changes primarily affect the visual grounding evidence used by the frozen VLA policy.

The clean performance also improves from 70.0% to 78.0%, indicating that R2G-VLA does not trade clean-scene performance for perturbation robustness. This behavior is partly attributed to the conservative safety-gated reinference design, which accepts the corrected action only when the correction satisfies uncertainty and action-shift constraints. Overall, these results support that physical

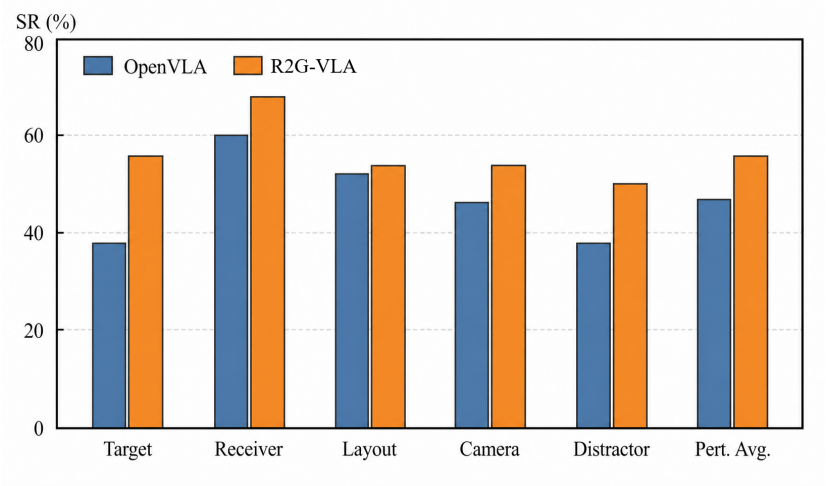

Figure 2: Success rates of OpenVLA and R2G-VLA across five physical perturbation families and their perturbed average.

scene perturbations induce internal grounding shifts in frozen VLA policies, and correcting high-risk visual evidence according to both action uncertainty and manipulation role improves robust execution without model updates.

4.3 Ablation Study

Table 2: Ablation study of R2G-VLA. The variants evaluate the effects of the grounding risk map, connected-region aggregation, role-guided correction, and safety-gated reinference.

Variant	Clean	Perturbed Avg.	All
R2G-VLA	**78.0**	**56.4**	**60.0**
w/o Risk Map	70.0	47.6	51.3
w/o Conn. Aggregation	76.0	46.8	51.7
w/o Role-Guided Corr.	74.0	48.0	52.3
w/o Action-Shift Gate	**78.0**	47.6	52.7

Table 2 evaluates each component under the same protocol as the full model; variant implementations are detailed in the supplementary material.

Grounding risk map. Removing the action-uncertainty grounding risk map reduces overall success from 60.0% to 51.3% and perturbed-average success from 56.4% to 47.6%. This supports using uncertainty-driven risk estimates to select visual tokens for correction.

Connected-region aggregation. Replacing connected-region aggregation with top-k token selection reduces overall success to 51.7% and perturbed-average success to 46.8%. These results favor spatially coherent region selection over scattered high-risk tokens under physical scene perturbations.

Role-guided correction. Removing role-guided correction reduces overall success to 52.3% and perturbed-average success to 48.0%. The decline supports distinguishing task-relevant target and receiver evidence from background and distractor cues when correcting risky tokens.

Action-shift gate. Disabling the action-shift constraint reduces overall success to 52.7% and perturbed-average success to 47.6%, while clean success remains 78.0%. This supports retaining the action-shift check alongside uncertainty-based acceptance to limit unreliable corrections under perturbations.

Together, the ablations support the contribution of risk localization, spatial aggregation, role-aware correction, and conservative action acceptance to robust frozen-policy execution.

4.4 Qualitative Grounding Analysis

Figure 3 visualizes the intermediate outputs of R2G-VLA under representative physical scene perturbations. The action-uncertainty grounding risk map highlights visual-token regions associated with unstable action prediction. The connected-region panel further shows that R2G-VLA converts high-response token cells into spatially coherent evidence regions, which better match object footprints, receiver areas, and local distractor regions in manipulation scenes.

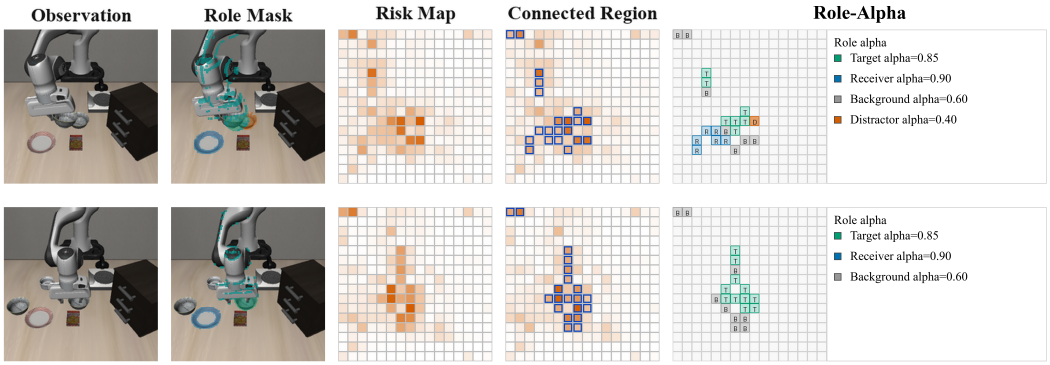

Figure 3: Qualitative visualization of R2G-VLA under representative physical scene perturbations. From left to right, we show the RGB observation, explicit role mask, action-uncertainty grounding risk map, connected high-risk region, and role-guided token correction.

The role-guided correction visualization shows that R2G-VLA does not simply suppress all selected high-risk tokens. Instead, selected tokens are assigned to manipulation-related roles and corrected with role-dependent strengths. Task-relevant target and receiver tokens receive milder correction, while background and distractor tokens are attenuated more strongly. This provides qualitative support for the claim that high-risk visual evidence should be corrected according to manipulation role rather than uniformly removed or suppressed. Across the two examples, the selected regions shift with the affected scene elements rather than remaining fixed in image coordinates. The figure therefore shows where R2G-VLA applies token correction and how the assigned semantic roles determine its strength.

5 Conclusion

We presented R2G-VLA, a risk- and role-guided grounding correction framework for robust frozen VLA execution under physical scene perturbations. R2G-VLA departs from input editing, robustness finetuning, and role-agnostic token suppression by correcting internal visual grounding evidence within the frozen policy. Its key idea is to decompose test-time grounding correction into two coupled questions. Action-uncertainty risk mapping identifies which visual tokens destabilize action prediction, while role-guided correction determines how these tokens should be modulated according to their manipulation roles. Connected-region aggregation further aligns token intervention with object-level scene changes, and safety-gated reinference prevents unreliable corrections from being executed. Experiments on LIBERO-Spatial show that R2G-VLA improves both overall and perturbed performance over frozen OpenVLA and a RobustVLA-style baseline. Ablations validate the importance of the grounding risk map, connected-region aggregation, role-guided correction, and conservative action acceptance. These results demonstrate that risk- and role-guided internal grounding correction is an effective model-update-free direction for reliable VLA execution under deployment-time scene changes.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and LaTeX format conversion. The authors reviewed the resulting revisions and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The supplementary material provides the complete inference algorithm, implementation details, and fixed inference parameters in Section A; benchmark settings, perturbation construction, evaluation protocol, and computing environment in Section B; baseline and ablation configurations in Section C; and the hyperparameter development and parameter-freezing procedure in Section D.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 782–791, 2021.

Hao Cheng, Erjia Xiao, Yichi Wang, Chengyuan Yu, Mengshu Sun, Qiang Zhang, Jiahang Cao, Yijie Guo, Ning Liu, Kaidi Xu, et al. Manipulation facing threats: Evaluating physical vulnerabilities in end-to-end vision language action models. *arXiv preprint arXiv:2409.13174*, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jianing Guo, Zhenhong Wu, Chang Tu, Yiyao Ma, Xiangqi Kong, Zhiqian Liu, Jiaming Ji, Shuning Zhang, Yuanpei Chen, Kai Chen, et al. On robustness of vision-language-action model against multi-modal perturbations. *arXiv preprint arXiv:2510.00037*, 2025.

Asher J Hancock, Allen Z Ren, and Anirudha Majumdar. Run-time observation interventions make vision-language-action models more visually robust. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9499–9506. IEEE, 2025.

Hyeongjun Heo, Seungyeon Woo, Sang Min Kim, Junho Kim, Junho Lee, Yonghyeon Lee, and Young Min Kim. Anycamvla: Zero-shot camera adaptation for viewpoint robust vision-language-action models. *arXiv preprint arXiv:2603.05868*, 2026.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Chenyang Li, Jieyuan Liu, Bin Li, Bo Gao, Yilin Yuan, Yangfan He, Yuchen Li, and Jingqun Tang. Dtp: A simple yet effective distracting token pruning framework for vision-language action models. *arXiv preprint arXiv:2601.16065*, 2026.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Hanqing Liu, Shouwei Ruan, Jiahuan Long, Junqi Wu, Jiacheng Hou, Huili Tang, Tingsong Jiang, Weien Zhou, and Wen Yao. Eva-vla: Evaluating vision-language-action models' robustness under real-world physical variations. *arXiv preprint arXiv:2509.18953*, 2025.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. *arXiv preprint arXiv:2109.12098*, 2021.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Khoa Vo, Taisei Hanyu, Yuki Ikebe, Trong Thang Pham, Nhat Chung, Minh Nhat Vu, Duy Nguyen Ho Minh, Anh Nguyen, Anthony Gunderman, Chase Rainwater, et al. Clutter-resistant vision-language-action models through object-centric and geometry grounding. *arXiv preprint arXiv:2512.22519*, 2025.

Yuhan Xie, Yuping Yan, Yunqi Zhao, Handing Wang, and Yaochu Jin. Strong-vla: Decoupled robustness learning for vision-language-action models under multimodal perturbations. *arXiv preprint arXiv:2604.10055*, 2026.

Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pp. 726–747. PMLR, 2021.

Hongyin Zhang, Pengxiang Ding, Shangke Lyu, Ying Peng, and Donglin Wang. Gevrm: Goal-expressive video generation model for robust visual manipulation. *arXiv preprint arXiv:2502.09268*, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.