English Harm Probes Fail Across Languages in Vision-Language-Action Models

Anonymous Author(s)

Abstract: Harmfulness representations in language models can transfer across languages even when refusal fails. Does this property hold in vision-language-action models (VLAs)? We evaluate four robot policies and three vision-language model (VLM) baselines across six languages and two hazard designs. Every probe is fitted and tested within the same frozen model. On scene-grounded instructions, English-trained VLM probes retain 0.81–0.93 AUROC in Chinese. OpenVLA's probe instead scores 0.96 in English and 0.51 in Chinese; MiniVLA Bridge scores 0.95 and 0.57. Topic-matched negatives reproduce weak VLA transfer alongside stronger VLM baselines. Target-language refitting recovers substantial discrimination, and the language gap persists inside the action expert. Strong English harm detection can therefore coexist with poor ranking of translated hazards in VLAs.

Keywords: Vision-language-action models, Harm detection, Multilingual safety

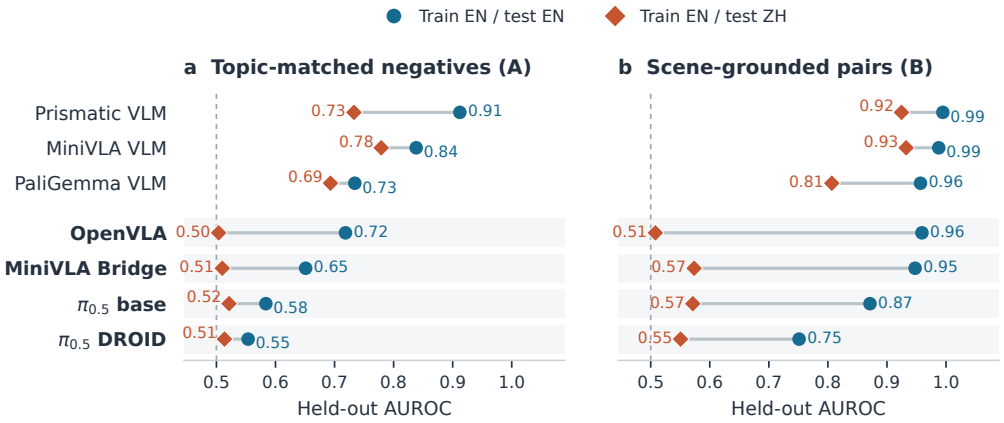


Figure 1: **VLM baselines retain cross-language detection; VLA probes lose it.** Each row fits and tests a separate probe within one frozen model. Circles and diamonds score that model's English-trained probe in English and Chinese. Unshaded rows are VLM baselines; shaded rows are VLAs. Points average five grouped folds and fixed deep layers; 0.5 denotes chance.

1 Introduction

A robot may receive the same hazardous request in several languages. For VLAs, whose outputs are actions, an internal harm detector can supply a signal to a separate safety response. SAFE-Dict already uses concept representations for inference-time VLA safety [1]. Such readouts must remain discriminative under translation.

Language-model results give a concrete reason to expect transfer. Zhao et al. [2] identify harmfulness and refusal as separate representations. Across 23 languages, Aziz et al. [3] find that a high-resource harmfulness direction remains discriminative even as refusal deteriorates. Wang et al.

Submitted to the 10th Conference on Robot Learning (CoRL 2026). Do not distribute.

[4] also demonstrate cross-language transfer of refusal directions. We test whether this premise carries over to robot policies.

Our contribution is an evaluation of cross-language harm-probe validity in four robot policies, with three VLM baselines, two hazard designs, and target-language refits. Each probe operates within a single model. Figure 1 captures the finding: strong cross-language detection in VLM baselines coexists with weak VLA transfer despite high English scores.

2 Measuring cross-language harm detection

Two hazard designs, six languages. Set A contains 760 instructions across eight categories: human injury, fire, electrical hazards, water damage, falling/breaking objects, chemicals, children/pets, and crushing/pinching. Each contributes 40 hazardous, 40 topic-matched safe, and 15 neutral instructions; safe and neutral instructions share the negative label. Images are assigned independently of labels and held fixed across translations. The resulting 4,560 inputs cover English (EN), Chinese (ZH), Japanese (JA), Korean (KO), Spanish (ES), and German (DE).

Set B contains 240 RoboShackles frames [5], each with two safe and two hazardous instructions authored from visible content, yielding 5,760 inputs across six languages. Safe and hazardous wordings differ more lexically, providing an easier detection task alongside A's topic matching. Labels encode authored instruction hazardousness; RoboShackles frames depict edited robot scenes.

Instructions and translations were authored and audited by a language-model assistant. Across 8,720 non-English sentences in the hazard sets and supporting battery, the audit identified 22 meaning issues with hazard labels preserved. Chinese hazard texts had zero flagged meaning errors. Every checkpoint uses the audited original version, sharing any translation artifacts across comparisons.

Frozen models and independently fitted probes. The four policies are OpenVLA-7B [6], MiniVLA Bridge [7], and $\pi_{0.5}$ base and DROID [8]. VLM baselines are Prismatic [9], the Qwen2.5-0.5B-based MiniVLA VLM [7], and PaliGemma-3B [10]. Every result trains and tests a probe in one model's hidden-state space. EN→ZH denotes changing the test language within that model. Translated items share their image and fold; corresponding model families receive identical token and pixel tensors.

We read the final prompt position in Prismatic, OpenVLA, and MiniVLA. PaliGemma and $\pi_{0.5}$ use the final prefix slot after `Action:`, with a fixed synthetic zero robot state. We separately probe $\pi_{0.5}$'s flow-matching expert at its first action position at $t = 1$, using identical Gaussian noise across examples; the expert attends to the full prefix.

Fit once in the source language, test in the target. For a fixed model's hidden state $h_k(I, x)$ at index k, a source-language probe assigns

$$s_{a,k}(I, x) = w_{a,k}^{\mathsf{T}} \widetilde{h}_k(I, x) + b_{a,k}, \qquad T_{a\to b,k} = \mathrm{AUROC}_{(I,x,y)\in D_b^{\mathrm{test}}}(s_{a,k}(I, x), y). \tag{1}$$

Here $\widetilde{h}$ is standardized using the source training split. L2 logistic regression uses $C = 0.01$ and five folds grouped by instruction item in A or frame group in B. Translations share their group's fold. The same held-out groups support EN→EN validation, EN→target transfer, and target→target refits.

We average fold AUROCs, then fixed hidden-state indices 8–32 for Prismatic and OpenVLA, 6–24 for MiniVLA, and 5–19 for PaliGemma and $\pi_{0.5}$, including the final normalized state. Scores are descriptive means; 0.5 is chance ranking. A second estimator uses difference-of-means directions, $d_{a,k} = \mathbb{E}[h_k \mid y = 1, a] - \mathbb{E}[h_k \mid y = 0, a]$, scored by $h_k^{\mathsf{T}} d_{a,k}$. Controls use token counts, mean input-token embeddings (Table 1), and held-out hazard categories.

Table 1: **Text baselines and target-language refits.** "Text EN/ZH" is an English-fitted difference-of-means direction on mean input-token embeddings, tested in English/Chinese. "ZH refit" is the deep logistic probe trained and tested in Chinese. All entries are held-out AUROCs.

	Set A			Set B		
	Text EN	Text ZH	ZH refit	Text EN	Text ZH	ZH refit
Prismatic VLM	0.73	0.49	0.82	0.92	0.67	0.99
MiniVLA VLM	0.68	0.61	0.85	0.92	0.76	0.98
PaliGemma VLM	0.63	0.53	0.75	0.93	0.81	0.94
OpenVLA	0.73	0.52	0.55	0.93	0.66	0.69
MiniVLA Bridge	0.70	0.59	0.59	0.92	0.71	0.85
$\pi_{0.5}$ base	0.61	0.49	0.56	0.81	0.55	0.76
$\pi_{0.5}$ DROID	0.62	0.48	0.53	0.72	0.47	0.66

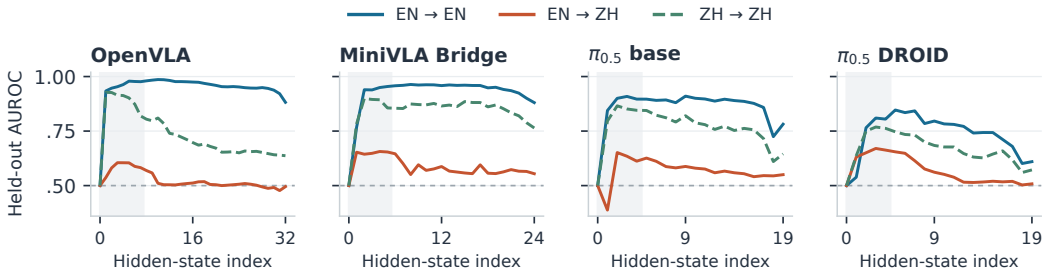


Figure 2: **Language-specific detection emerges across depth.** Set B logistic probes within each VLA: English fit/test (blue), English fit/Chinese test (orange), and Chinese fit/test (green). All curves use the same held-out groups. Gray regions mark early indices excluded from summary means. The gap between transferred and in-language readouts extends across the deep layers.

3 Results

English detection does not predict transfer. On Set B, independently fitted English probes in Prismatic, MiniVLA VLM, and PaliGemma retain 0.92, 0.93, and 0.81 AUROC on Chinese inputs to their respective models. OpenVLA's probe scores 0.96 in English but 0.51 in Chinese; MiniVLA Bridge scores 0.95 and 0.57. The $\pi_{0.5}$ base probe scores 0.87 and 0.57; DROID scores 0.75 and 0.55. Each VLA's English-fitted readout is unchanged across test languages: strong English detection gives little assurance of Chinese discrimination.

Set A reproduces weak VLA transfer under topic matching: Chinese AUROCs are 0.50, 0.51, 0.52, and 0.51, versus 0.73, 0.78, and 0.69 in the three VLM baselines. English detection also weakens, to 0.72 in OpenVLA and 0.65 in Bridge. The result spans both easier and topic-matched negatives.

The pattern spans languages and layers. Across policies, Set B Japanese and Korean transfer both range from 0.51 to 0.56; VLM baselines range from 0.76–0.87 and 0.72–0.83, respectively. Spanish retains partial transfer in OpenVLA (0.71) and Bridge (0.66); both $\pi_{0.5}$ policies score 0.58. German scores 0.54–0.58 in the VLAs. Figure 2 shows the English–Chinese gap across deep VLA layers.

Lexical and length controls preserve the gap. Set B's text-only English direction achieves approximately 0.92 in Prismatic, OpenVLA, and both MiniVLA models (Table 1), revealing substantial lexical separability. Deep English-to-Chinese probes fall below even the embedding baselines in OpenVLA (0.51 versus 0.66) and Bridge (0.57 versus 0.71).

Length matching restricts hazardous–safe test pairs to token counts differing by at most one. Set B difference-of-means Chinese transfer is 0.85/0.85 in Prismatic/MiniVLA VLM and 0.50/0.59 in

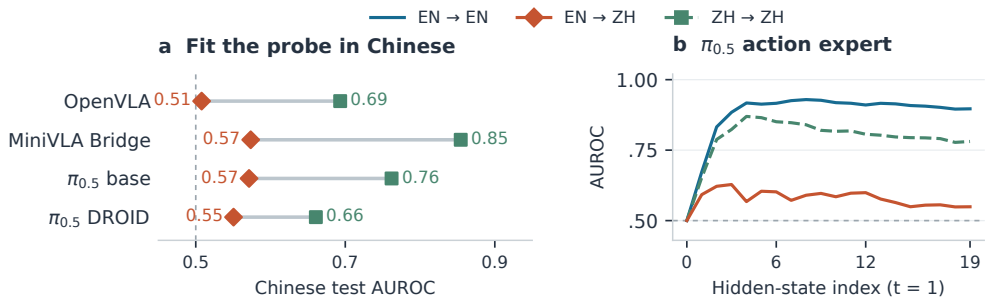


Figure 3: **Language-specific detection survives inside the policy.** Set B, logistic probes. (a) Chinese-trained probes recover part of the discrimination lost by English-trained probes on the same Chinese test groups. (b) In the $\pi_{0.5}$ base action expert, within-language detection becomes strong while English-to-Chinese transfer stays much weaker. All examples share the same initial noise at $t = 1$; the curves include the suffix input, 18 blocks, and final normalization.

OpenVLA/Bridge. Holding out a hazard category yields 0.85/0.88 in those VLMs and 0.49/0.54 in the VLAs. Both controls preserve stronger VLM transfer.

A Chinese fit recovers signal that the English fit misses. Chinese refitting raises Set B OpenVLA from 0.51 to 0.69, Bridge from 0.57 to 0.85, and $\pi_{0.5}$ base from 0.57 to 0.76 (Fig. 3a). Set A refits are weaker (Table 1). Recoverability depends on instruction design: the results support both reduced separability and poorer cross-language alignment at these readouts.

The action expert preserves the language gap. Inside $\pi_{0.5}$ base, Set B English detection reaches 0.91 and Chinese refitting reaches 0.81, while English-to-Chinese transfer is 0.58 (Fig. 3b). The DROID expert shows the same ordering (0.81, 0.74, 0.60). With identical initial noise, information arrives through prefix conditioning but remains poorly accessible to the expert's English-trained detector.

Controlled contrasts expose language-specific directions. A supporting battery crosses 12 instruction distinctions, six languages, and 100 shared frames, using two half-frame folds. Deep-layer English–Chinese direction cosine averages 0.02 in OpenVLA and 0.03 in the $\pi_{0.5}$ prefix. English synonym transfer over ten distinctions reaches 0.92 in OpenVLA: generalization across English wordings coexists with nearly orthogonal directions across languages.

4 Implications for multilingual robot safety

Multilingual VLA studies establish language-dependent task performance and representation shifts [11, 12]; RoboShackles supplies multilingual safety scenarios [5]. Our results connect these findings to activation-based safety: the validity of an internal harm detector itself depends on language.

The positive LLM transfer results [3, 4] are conditional: Balani and Panda [13] show dependence on negative-example difficulty. Here, independently fitted VLM probes retain stronger cross-language discrimination under both hazard designs, while VLA probes lose *ranking* quality across languages. Changing a decision threshold leaves AUROC unchanged; target-language refits recover discrimination on the same hazards.

Robot safety evaluations should therefore report source-language validation, cross-language transfer, and target-language refitting on the same held-out groups. Within each fixed VLA, these measurements distinguish an effective English detector from one that generalizes across languages. The gap appears at policy readouts and inside the action expert. **A harm probe's language of validation is part of its safety specification.**

References

[1] S. Wen, S. Yang, S. Fu, J. Zhang, L. Hu, and D. Wang. Concept-based dictionary learning for inference-time safety in vision language action models. *arXiv preprint arXiv:2602.01834*, 2026. URL https://arxiv.org/abs/2602.01834.

[2] J. Zhao, J. Huang, Z. Wu, D. Bau, and W. Shi. LLMs encode harmfulness and refusal separately. *arXiv preprint arXiv:2507.11878*, 2025. URL https://arxiv.org/abs/2507.11878.

[3] R. Aziz, I. A. Hanif, and F. Koto. Low-resource safety failures are action failures, not representation failures. *arXiv preprint arXiv:2606.01196*, 2026. URL https://arxiv.org/abs/2606.01196.

[4] X. Wang, M. Wang, Y. Liu, H. Schütze, and B. Plank. Refusal direction is universal across safety-aligned languages. *arXiv preprint arXiv:2505.17306*, 2025. URL https://arxiv.org/abs/2505.17306.

[5] Z. Yin, C. Liu, W. Yang, R. Li, and Y. Xue. ROBOSHACKLES: A safety dataset for human-injury prevention in embodied foundation models. *arXiv preprint arXiv:2606.18632*, 2026. URL https://arxiv.org/abs/2606.18632.

[6] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. URL https://arxiv.org/abs/2406.09246.

[7] Stanford ILIAD. MiniVLA: OpenVLA-mini code and checkpoints. GitHub repository, 2026. URL https://github.com/Stanford-ILIAD/openvla-mini. Accessed September 24, 2026.

[8] Physical Intelligence, K. Black, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL https://arxiv.org/abs/2504.16054.

[9] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh. Prismatic VLMs: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*, 2024. URL https://arxiv.org/abs/2402.07865.

[10] L. Beyer et al. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024. URL https://arxiv.org/abs/2407.07726.

[11] H. Chen, H. Li, J. Cao, Y. Li, Y. Jiang, H. Wen, K. Huang, S. Guo, and H. Wan. Beyond english: Uncovering the multilingual gap in vision-language-action models. *arXiv preprint arXiv:2606.15714*, 2026. URL https://arxiv.org/abs/2606.15714.

[12] X. Dong, Z. Han, T. Niu, Q. Zhu, and W. Che. When does language matter? multilingual instructions reveal step-wise language sensitivity in vision-language-action models. *arXiv preprint arXiv:2606.11906*, 2026. URL https://arxiv.org/abs/2606.11906.

[13] P. Balani and S. Panda. Hard negatives reveal what easy negatives hide: Cross-lingual harmfulness representations degrade with resource tier under hard negatives. *arXiv preprint arXiv:2609.27758*, 2026. URL https://arxiv.org/abs/2609.27758.