Beyond Perception Errors: Semantic Fixation in Large Vision-Language Models

Md Tanvirul Alam
Rochester Institute of Technology
Rochester, NY, USA
ma8235@rit.edu

Abstract

Large vision-language models (VLMs) often rely on familiar semantic priors, but existing evaluations do not cleanly separate perception failures from rule-mapping failures. We study this behavior as *semantic fixation*: preserving a default interpretation even when the prompt specifies an alternative, equally valid mapping. To isolate this effect, we introduce *VLM-Fix*, a controlled benchmark over four abstract strategy games that evaluates identical terminal board states under paired standard and inverse rule formulations. Across 14 open and closed VLMs, accuracy consistently favors standard rules, revealing a robust semantic-fixation gap. Prompt interventions support this mechanism: neutral alias prompts substantially narrow the inverse-rule gap, while semantically loaded aliases reopen it. Post-training is strongly rule-aligned: training on one rule improves same-rule transfer but hurts opposite-rule transfer, while joint-rule training improves broader transfer. To test external validity beyond synthetic games, we evaluate analogous defamiliarization interventions on *VLMBias* and observe the same qualitative pattern. Finally, late-layer activation steering partially recovers degraded performance, indicating that semantic-fixation errors are at least partly editable in late representations. Project page, code, and dataset available at https://maveryn.github.io/vlm-fix/.

1 Introduction

Large vision-language models (VLMs) have recently achieved strong performance across instruction following, visual reasoning, and open-ended image understanding, and benchmark evaluations similarly report high aggregate results on broad multimodal tasks (Hurst et al., 2024; Yang et al., 2023; Liu et al., 2023a; Deitke et al., 2025; Wang et al., 2025; Bai et al., 2025a). However, these gains do not necessarily translate to robust reasoning in controlled settings: a growing body of work shows that VLMs often rely heavily on learned language priors, which can drive systematic bias and hallucination (Agrawal et al., 2018; Hall et al., 2023; Lee et al., 2025; Zhou et al., 2023; Sharma et al., 2024; Fu et al., 2025; Vo et al., 2025). Across such evaluations, model predictions often track familiar semantic patterns rather than task-specified evidence (Ruggeri & Nozza, 2023; Huang et al., 2025a; Vo et al., 2025; Zhou et al., 2023).

These findings motivate a complementary way to study prior dependence. Prior work often probes this behavior through biased prompting or counterfactual visual manipulations, but those settings can still leave ambiguity about whether failures stem from perceptual changes or from rigid semantic expectations. Here we target the latter directly, asking whether VLMs can revise decisions when task goals redefine the same perceptual state. We call this failure mode *semantic fixation*, adapting the cognitive psychology notion of fixation (the Einstellung effect), where people persist with familiar solution patterns despite equally valid alternatives (Luchins, 1942; Bilalić et al., 2008). Related prior-driven behavior has also been observed in modern language models, including anchoring effects (Alavi Naeini et al., 2023; Huang et al., 2025b; Kim et al., 2025). In this work, we examine semantic fixation in VLMs: models may default to familiar meanings even when prompts specify alternative mappings.

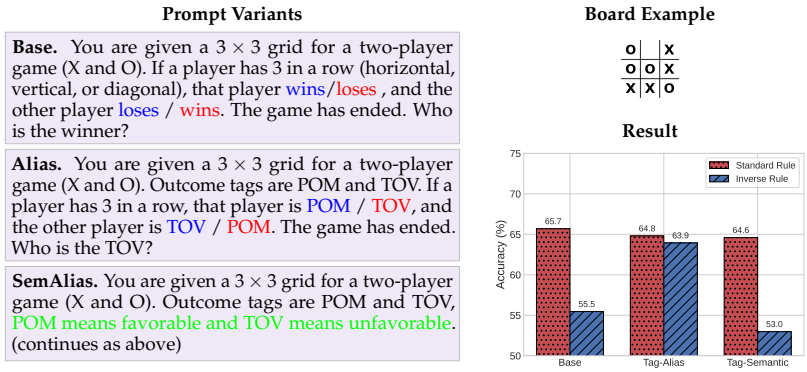

Figure 1: Tic-Tac-Toe example from VLM-Fix. Left: base, tag-alias, and tag-semantic prompt variants. Right-top: terminal board state. Right-bottom: accuracy trend across prompt variants. Blue marks standard-rule assignments, red marks inverse (misère) assignments, and green marks the added semantic definition in the tag-semantic prompt.

Because the same perceptual state can require different decisions under different task goals, robust reasoning depends on semantic remapping, not only visual recognition (Jang et al., 2022; Shi et al., 2025).

To investigate this hypothesis, we build *VLM-Fix*, a synthetic benchmark spanning four abstract strategy games: Tic-Tac-Toe, Connect Four, Reversi, and Dots and Boxes. For each game, we evaluate identical terminal board states under both standard and inverse (misère-style) rules, so only the winner/loser semantics change. This design isolates interpretation from perception and yields a direct measure of semantic fixation. Across both open and closed VLMs, accuracy is consistently higher under standard rules, revealing a robust semantic-fixation gap. Prompt interventions further support this mechanism: replacing winner/loser terms with neutral tags substantially narrows the inverse-rule gap, while reintroducing semantic valence with the same tags restores it (Figure 1).

Beyond these inference-time prompt effects, we examine how post-training reshapes semantic fixation. Post-training with supervised fine-tuning (SFT) and reinforcement learning with verifiable rewards (RLVR) is strongly rule-aligned: training on one rule improves same-rule performance but hurts cross-rule transfer, whereas joint training improves generalization across both rules. Finally, late-layer activation steering recovers degraded performance, indicating that much of the error stems from late semantic readouts rather than failures in visual recognition. We also observe a similar trend on the *VLMBias* benchmark (Vo et al., 2025), specifically on its counterfactual counting task, supporting external validity beyond synthetic games.

Our main findings are:

1. On *VLM-Fix*, inverse rules reduce accuracy despite identical visual states, indicating semantic fixation under rule remapping.
2. Semantic framing strongly modulates this gap: neutral aliases reduce it, while semantically loaded aliases restore it.
3. Post-training is rule-dependent: SFT/RLVR improve same-rule performance but induce negative transfer to the opposite rule, while joint rule training improves generalization across both rules.
4. Late-layer activation steering partially recovers degraded inverse-rule performance, indicating that some of the error is editable at late semantic readout stages.
5. We observe the same qualitative pattern on *VLMBias* (Vo et al., 2025), supporting external validity beyond the synthetic game setting.

2 Related Work

Prior work shows that large vision-language models (VLMs) exhibit systematic bias and reliability limitations beyond aggregate benchmark scores. Social-bias evaluations report stereotypical and representational skew in multimodal outputs (Ruggeri & Nozza, 2023; Huang et al., 2025a), while broader robustness studies find brittle behavior under distribution shift and weak grounding under challenging visual conditions (Yang et al., 2023; Liu et al., 2023b). Counterfactual and minimally perturbed benchmarks further expose this fragility: C-VQA reports large drops on counterfactual questions (Zhang et al., 2023), VisMin shows weaknesses on counting and spatial edits (Awal et al., 2024), and SpatialEval finds reduced visual reliance when redundant text is present (Wang et al., 2024). Hallucination-focused analyses (e.g., POPE, HallusionBench) and representation-level studies also indicate persistent misalignment between visual evidence and model outputs (Li et al., 2023; Guan et al., 2024; Fu et al., 2025).

Fixation-style reasoning has been studied more directly in language models. Prior work links LLM behavior to the cognitive notion of fixation (Einstellung), showing persistent reliance on familiar but suboptimal mappings (Luchins, 1942; Bilalić et al., 2008; Alavi Naeini et al., 2023). Related findings on anchoring and inflexible reasoning similarly suggest that LLM predictions can remain tied to initial or familiar semantic frames (Huang et al., 2025b; Kim et al., 2025). Broader cognitive-bias evaluations report human-like intuitive biases in modern LLMs (Hagendorff et al., 2023). However, these studies are predominantly text-only and do not isolate fixation in multimodal settings where perceptual evidence is held constant and only semantic mapping changes.

Bias-mitigation work in both LLMs and VLMs has explored decoding, prompting, and counterfactual training interventions. In language models, self-debiasing and causal-intervention methods reduce harmful continuations and biased reasoning (Schick et al., 2021; Wu et al., 2024; Sun et al., 2024; Xia et al., 2024). In vision-language models, counterfactual prompt learning and counterfactual data interventions aim to reduce spurious correlations between semantics and visual categories (He et al., 2022; Howard et al., 2024; Vo et al., 2025). Our study complements these lines by providing a controlled rule-switch setting that cleanly separates perception from semantic interpretation, and by showing that neutral alias prompts reduce the inverse-rule gap while semantically loaded aliases restore it.

3 VLM-Fix: Dataset and Evaluation Setup

We design VLM-Fix, a controlled synthetic benchmark for evaluating model behavior under matched visual states but different semantic interpretations. The benchmark comprises four abstract strategy games: Tic-Tac-Toe (3×3), Reversi (5×5), Connect Four (4×4), and Dots and Boxes (6×6). For each game, we generate 300 unique terminal base states, excluding draws and balancing the canonical winner evenly across the two players. Focusing on terminal states fixes the visual evidence at the end of play and avoids confounds from intermediate game trajectories; these same states are reused across all image and prompt settings. Full details of base-state generation are provided in Appendix A.

Rule conditions. Each example is evaluated under one of two rule conditions: **Standard**, corresponding to the canonical objective of the game, and **Inverse**, which reverses the usual winning condition. For each condition, we ask two complementary query types: the identity of the *winner* and the identity of the *loser*. This matched design controls for board complexity, allowing performance differences to be attributed to semantic interpretation rather than board-state difficulty. The inverse rules we consider are valid game variants (Wikipedia contributors, 2026; Keller, 2026), although they are likely to be much less represented in web-scale pretraining data than their canonical counterparts. Table 1 summarizes the rule definitions used in VLM-Fix.

Rendering variants. We use three visual rendering styles. **Base** uses the canonical board appearance and default player symbols. **Checkerboard** changes the board texture while keeping the canonical symbols unchanged. **Glyph** preserves the board layout but replaces

Table 1: Standard and inverse rule conditions in VLM-Fix.

Game	Standard rule	Inverse rule
Tic-Tac-Toe	A player who gets 3-in-a-row wins.	A player who gets 3-in-a-row loses.
Connect Four	A player who gets 4-in-a-row wins.	A player who gets 4-in-a-row loses.
Reversi	The player with more pieces wins.	The player with fewer pieces wins.
Dots and Boxes	The player who claims more boxes wins.	The player who claims fewer boxes wins.

Base: You are given a 5x5 grid for a two-player game. Players are Black and White. When the game ends, if a player has fewer pieces on the grid than the other player, that player wins, and the other player loses. The game has ended. Who is the winner?

Alias: You are given a 5x5 grid for a two-player game. Players are Black and White. Outcome tags: KAP and POM. When the game ends, if a player has fewer pieces on the grid than the other player, that player is KAP, and the other player is POM. The game has ended. Who is the KAP?

SemAlias: You are given a 5x5 grid for a two-player game. Players are Black and White. Outcome tags: KAP and POM. KAP means favorable outcome; POM means unfavorable outcome. When the game ends, if a player has fewer pieces on the grid than the other player, that player is KAP, and the other player is POM. The game has ended. Who is the KAP?

Figure 2: Compact Reversi example from VLM-Fix. Left: the three rendering variants, shown in order as the base, glyph, and checkerboard renderings used in the benchmark. Right: the three prompt variants that preserve the same underlying inverse-rule state while changing only semantic framing.

canonical player symbols with randomly sampled alphabetic glyphs. For each $(\text{game}, \text{state})$ pair, glyph symbols are deterministically sampled from `A..Z`, excluding $\{X, O, A, B\}$ to avoid collisions with canonical game labels.

Prompt variants. We use three prompt families to vary the semantic framing of the same visual input. Across all variants, we avoid naming the game verbatim in the prompt (e.g., "Tic-Tac-Toe" or "Reversi") to reduce lexical cues toward canonical rule interpretations. **Base** states the game rule directly and asks for the winner or loser using canonical outcome terms. **Alias** replaces canonical terms such as *winner* and *loser* with arbitrary tags. **SemAlias** retains these arbitrary tags but explicitly defines their meaning (e.g., *favorable* versus *unfavorable*). This design helps disentangle the effect of lexical familiarity from that of explicit semantic grounding. Figure 2 shows a representative Reversi example across the three rendering variants and three prompt variants; corresponding examples for the other games are shown in Appendix Figure 8.

Dataset size. In the main experiments, we consider five configurations: `Base`, `Glyph`, `Checkerboard`, `Alias`, and `SemAlias`. `Base/Glyph/Checkerboard` share the same prompt template but differ in rendering, whereas `Base/Alias/SemAlias` share the canonical rendering but differ in prompt semantics. For each base state, we evaluate both rule conditions (**Standard** and **Inverse**), both query targets (**winner** and **loser**), and both multimodal orderings (**image-first** and **text-first**), yielding $300 \times 2 \times 2 \times 2 = 2400$ examples per game for each image–prompt configuration. Across the five main configurations, this results in 12,000 evaluated examples per game. We further evaluate each example under both **Direct** answering and **Chain-of-Thought (CoT)** prompting, yielding 24,000 evaluated examples per game. We additionally include descriptive-rule and text-only variants as control settings to isolate instruction-following effects from visual grounding: in the descriptive-rule variant, prompts ask directly about board properties (e.g., who forms the winning pattern or who has more/fewer pieces) rather than asking who is the winner or loser under a named rule, while in the text-only variant, the board is provided as an ASCII rendering without an image. Example prompts for both controls are shown in the Appendix.

Table 2: VLM-Fix results under the baseline setting (`Base`; canonical rendering; direct answering). Each cell reports accuracy (%) under the standard or inverse rule, aggregated over winner/loser question targets and image-first/text-first orders.

Model	Tic-Tac-Toe		Reversi		Connect Four		Dots and Boxes		Average	
	Std	Inv	Std	Inv	Std	Inv	Std	Inv	Std	Inv
GPT-4.1	86.0	76.0	85.7	75.0	69.0	41.3	88.3	60.3	82.2	63.2
GPT-5.2	93.7	86.0	95.7	91.0	69.0	53.0	93.7	75.7	88.0	76.4
Sonnet-4.0	51.0	51.7	76.0	50.3	63.3	65.7	95.0	92.3	71.3	65.0
Sonnet-4.5	50.0	56.0	85.7	67.0	53.0	86.7	78.3	77.0	66.8	71.7
Qwen2.5-VL-3B	54.2	49.9	57.2	50.0	51.7	50.6	51.7	44.2	53.7	48.7
Qwen2.5-VL-7B	66.5	44.7	61.3	41.5	58.9	48.2	65.8	32.6	63.1	41.8
InternVL3.5-4B	60.7	45.9	63.7	34.2	52.2	49.8	70.1	30.5	61.7	40.1
InternVL3.5-8B	73.8	57.6	63.2	45.2	54.9	49.0	72.2	55.6	66.0	51.9
InternVL3.5-14B	66.4	58.9	52.9	55.4	62.9	49.2	53.2	42.9	58.9	51.6
Qwen3-VL-4B	57.4	48.0	64.3	52.1	50.3	47.8	65.0	38.8	59.3	46.7
Qwen3-VL-8B	57.6	47.7	69.8	42.4	51.4	51.6	76.2	36.7	63.8	44.6
Qwen3-VL-32B	70.0	66.4	83.4	56.9	54.6	51.3	73.7	55.9	70.4	57.6
Molmo2-4B	61.3	42.8	65.1	22.2	55.2	48.7	71.7	22.7	63.3	34.1
Molmo2-8B	71.1	44.9	82.2	41.4	55.2	47.4	77.6	34.5	71.5	42.1
Average	65.7	55.5	71.9	51.8	57.3	52.9	73.8	50.0	67.1	52.5

4 Experiment & Results

4.1 Models

We evaluate 14 VLMs spanning five model families and multiple parameter scales: GPT-4.1 and GPT-5.2 (Singh et al., 2025); Sonnet-4.0 and Sonnet-4.5 (Anthropic, 2026); Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen3-VL-4B, Qwen3-VL-8B, and Qwen3-VL-32B (Bai et al., 2025b;a); InternVL3.5-4B, InternVL3.5-8B, and InternVL3.5-14B (Wang et al., 2025); and Molmo2-4B and Molmo2-8B (Clark et al., 2026). We use greedy decoding (temperature 0) with a maximum generation budget of 1,024 tokens. To keep API evaluation cost manageable, the four closed-source models are evaluated on an aligned reduced subset of 300 states per game, whereas the open-weight models are evaluated on the full benchmark expansions.

4.2 Results

Baseline Performance. Table 2 reports baseline results under `Base`, where prompts use canonical winner/loser semantics. Averaged across games and models, standard-rule accuracy is 67.1% versus 52.5% under inverse rules, a 14.6-point gap. Because standard and inverse evaluations reuse identical terminal boards, this drop cannot be attributed to changes in visual evidence; instead, it indicates difficulty with semantic rule remapping. The gap appears in all four games and is largest for Dots and Boxes (73.8% vs. 50.0%, a 23.8-point difference). At the model level, 13 of 14 models show lower inverse-rule accuracy than standard-rule accuracy; the only exception is Sonnet-4.5. Together, these baseline results provide the main empirical signal of semantic fixation: performance is strong when task semantics align with familiar priors, but degrades when the same perceptual state requires an alternative interpretation.

Input Interventions. Table 3 summarizes how the main input interventions affect the average standard/inverse gap across the 14 models. Visual perturbations yield only modest changes: `Glyph` slightly improves inverse accuracy (54.95), while `Checkerboard` slightly lowers overall accuracy. In contrast, semantic framing interventions are much stronger: `Alias` raises inverse accuracy to 63.08 and reduces the standard–inverse gap to 2.29 points, the smallest among all settings. When semantic valence is reintroduced with `SemAlias`, inverse accuracy drops back to 53.51 and the gap reopens. This pattern aligns with our

Table 3: Average effect of input interventions on VLM-Fix direct results, aggregated over the 14 models. Columns report standard and inverse accuracy (%) for each game and the macro-average across games.

Config	Tic-Tac-Toe		Reversi		Connect Four		Dots and Boxes		Average	
	Std	Inv	Std	Inv	Std	Inv	Std	Inv	Std	Inv
Base	**65.7**	55.5	71.9	51.8	57.3	52.9	73.8	50.0	67.1	52.5
Glyph	65.2	54.6	**74.1**	53.3	**61.8**	56.9	70.3	54.9	**67.8**	54.9
Checkerboard	63.7	53.9	67.1	52.6	58.4	52.9	**74.5**	52.3	65.9	52.9
Alias	64.8	**63.9**	68.3	**66.6**	57.5	**57.9**	70.9	**63.9**	65.4	**63.1**
SemAlias	64.6	53.0	71.6	54.7	58.3	53.5	70.3	52.9	66.2	53.5

Table 4: Boundary-condition results on VLM-Fix. Columns report standard and inverse accuracy (%) for each game and the macro-average across games.

Setting	Tic-Tac-Toe		Reversi		Connect Four		Dots and Boxes		Average	
	Std	Inv	Std	Inv	Std	Inv	Std	Inv	Std	Inv
Direct (Base)	65.7	55.5	71.9	51.8	57.3	52.9	73.8	50.0	67.1	52.5
CoT (Base)	**85.6**	76.4	91.9	74.1	83.3	79.2	79.2	66.5	85.0	74.1
CoT (Alias)	84.2	**84.8**	91.3	**90.6**	**84.4**	**83.9**	79.4	77.3	84.8	**84.1**
CoT (SemAlias)	84.0	80.8	**92.3**	77.0	**84.4**	80.3	81.3	71.0	**85.5**	77.3
Descriptive	65.5	64.4	78.0	74.1	59.1	56.0	**81.3**	**80.7**	71.0	68.8
Text-only	69.8	58.5	70.9	59.7	64.9	57.1	71.2	59.2	69.2	58.6

semantic-fixation account: neutral aliases reduce prior-driven bias, whereas semantically loaded tags restore it. Full game-wise breakdowns are reported in the Appendix.

Control Settings. Table 4 reports results for additional control settings. Descriptive prompting is substantially more balanced than the baseline (71.0/68.8), indicating that the standard–inverse gap in the base setup is not simply due to poor instruction following. Text-only evaluation still shows a clear inverse drop (69.2/58.6), indicating that the effect is largely text-driven and persists even without image input.

CoT improves overall accuracy relative to direct prompting, but the same qualitative pattern remains. Under `Base`, a sizable standard–inverse gap persists (85.0 vs. 74.1). `Alias` largely closes this gap (84.8/84.1), whereas `SemAlias` partially reopens it (85.5/77.3), consistent with the direct-answering results.

Input Order. We also compare image-first and text-first orderings for the 10 open-weight models; full marginals are reported in Appendix Tables 16 and 17. Under direct prompting with `Base`, text-first amplifies the standard–inverse gap from 13.2 points (60.2/47.0) to 21.4 points (66.2/44.8). Under CoT, this order effect largely disappears, with similar standard–inverse gaps for image-first and text-first prompting.

5 External Validity on *VLMBias*

VLM-Fix is designed to isolate semantic remapping cleanly rather than to maximize naturalism, which lets us derive a broader intervention hypothesis: if prior-driven failures are partly sustained by familiar image–prompt associations, then defamiliarizing either side of the pair should reduce error even in tasks without an explicit standard/inverse rule switch. We therefore evaluate on *VLMBias* (Vo et al., 2025), a counterfactual-image benchmark designed to expose prior-driven errors, not as a second direct measurement of semantic fixation itself, but as an external transfer test of whether VLM-Fix-motivated defamiliarization interventions carry over to a more natural prior-sensitive setting. We focus on four counting subsets (**Animals**, **Logos**, **Flags**, and **Game Boards**).

Image and Prompt Interventions. We consider three interventions relative to `Base`: (1) `Flip`, which vertically flips the image while keeping the prompt unchanged; (2) `Alias`, which replaces the target object term (e.g., "animal") with a generic `ITEM` token (one ITEM = one counted object) while keeping the image unchanged; and (3) `Flip+Alias`, which combines both. `Flip` creates visually defamiliarized inputs, especially for photo-realistic **Animals** and

Base: Count the legs of this animal. Answer with a number in curly brackets, e.g., {9}.

Alias: Count the visible legs in this image, but report them as ITEMs (one ITEM = one visible leg). How many ITEMs are there? Answer with a number in curly brackets, e.g., {9}.

Figure 3: Representative Animals example from *VLMBias*. Left: the Base and Flip images. Right: the corresponding Base and Alias prompts.

Table 5: Aggregate *VLMBias* results across the four evaluation subsets, averaged over the 14 models using the canonical row-level summary.

Config	Animals		Logos		Flags		Game Boards		Overall	
	Acc	Bias	Acc	Bias	Acc	Bias	Acc	Bias	Acc	Bias
Base	3.6	93.2	13.9	71.4	25.9	55.1	11.4	66.6	11.6	76.7
Flip	5.7	88.9	16.6	60.2	24.2	55.4	13.7	61.1	13.3	70.9
Alias	8.5	83.8	14.9	65.6	30.6	51.8	14.1	58.0	15.0	69.5
Flip+Alias	22.2	67.5	15.7	56.0	29.0	49.8	16.2	51.1	20.7	58.9

Logos, while preserving task-relevant pixels and full scene complexity. Unlike the setup in (Vo et al., 2025), which removes backgrounds (and can reduce perceptual difficulty), `Flip` primarily alters orientation cues. `Alias` is motivated by VLM-Fix findings that reducing semantically loaded priors reduces bias; because counting datasets typically lack explicit rule semantics, it tests whether lexical defamiliarization alone can reduce prior-driven errors. Figure 3 shows a representative Animals example; examples for the other subsets appear in the Appendix.

Results. Table 5 summarizes aggregate *VLMBias* performance across the four subsets, averaged over the 14 evaluated models. Relative to `Base`, both single perturbations help: `Flip` improves overall accuracy from 11.6% to 13.3% while reducing bias from 76.7% to 70.9%, and `Alias` further improves these to 15.0% and 69.5%, respectively. The combined setting `Flip+Alias` is strongest overall, reaching 20.7% accuracy and 58.9% bias, with the largest task-level gain in Animals (3.6% $\rightarrow$ 22.2%). Crucially, this pattern holds without exception: in the pooled summary, `Flip+Alias` improves accuracy and reduces bias for all 14 models. Model-wise and subset-wise results are in Appendix B.6.

6 Training Interventions and Transfer

We explore two post-training strategies, supervised fine-tuning (SFT) and reinforcement learning with verifiable rewards (RLVR) using GRPO (Shao et al., 2024; Guo et al., 2025), across two transfer settings: rule transfer within VLM-Fix and synthetic leg-count transfer to *VLMBias*. These experiments test whether targeted post-training can improve robustness under counterfactual task formulations, and whether such gains transfer across rules, games, and benchmarks.

6.1 VLM-Fix Rule-Transfer Splits

We evaluate SFT and RLVR on three VLM-Fix transfer splits, D1–D3, designed to probe cross-rule and cross-game generalization. D1 trains on standard-rule supervision and its canonical test file contains inverse-rule examples, D2 reverses this direction, and D3 trains on Tic-Tac-Toe and Reversi with both rules before testing on Connect Four and Dots and Boxes with both rules. For D1 and D2, we also evaluate the complementary held-out same-rule split, so comparisons include both same-rule and cross-rule transfer under the original image and base prompt. Full split construction, training hyperparameters, and evaluation details are provided in Appendix C.1.

Results. Figure 4 shows a consistent same-rule versus cross-rule pattern. On D1 and D2, post-training improves held-out accuracy when evaluation matches the training rule

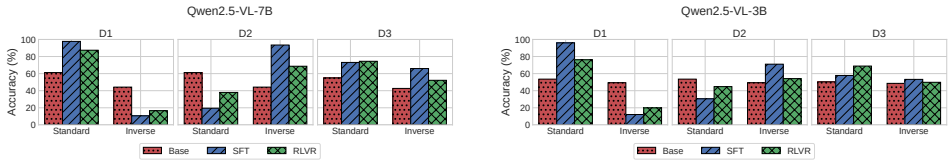

Figure 4: VLM-Fix splits (D1–D3) for Qwen2.5-VL-7B (left) and Qwen2.5-VL-3B (right): D1/D2 report same-rule and opposite-rule evaluation, and D3 reports held-out standard/inverse evaluation on Connect Four and Dots and Boxes.

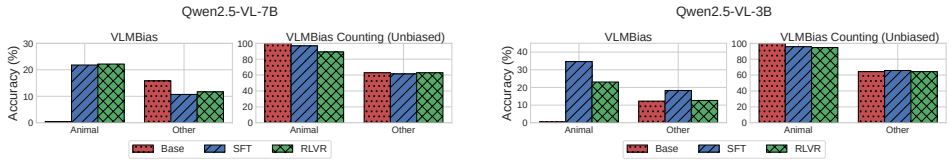

Figure 5: Synthetic leg-count transfer for Qwen2.5-VL-7B (left) and Qwen2.5-VL-3B (right), comparing Base, SFT, and RLVR on *VLMBias* and *VLMBias Counting (Unbiased)* (Animals vs Other).

(with SFT typically strongest), but performance drops below the base model when the rule mapping is flipped; RLVR is less brittle than SFT under this shift. On D3, where training covers both rules on source games, both methods improve transfer to held-out Connect Four and Dots and Boxes, with RLVR stronger on standard-rule evaluation and SFT stronger on inverse-rule evaluation (especially for Qwen2.5-VL-7B). Additional SFT D1–D3 transfer results for Molmo2-4B and InternVL3.5-4B are reported in Appendix Figure 11. Overall, post-training is strongly rule-conditional: same-rule transfer is strong, cross-rule transfer is negative, and cross-game transfer improves when both semantic mappings are seen during training.

6.2 Synthetic Leg-Count Transfer to VLMBias

We next test transfer from a synthetic leg-counting dataset of procedurally rendered bird and quadruped glyphs. We post-train the same Qwen2.5-VL-3B and Qwen2.5-VL-7B backbones with SFT and RLVR, then evaluate on *VLMBias* (Animals vs Other) and *VLMBias Counting (Unbiased)* (Animals vs Other). Example synthetic glyph categories are shown in Appendix Figure 10. Full dataset construction and training details are provided in Appendix C.2.

Results. Figure 5 shows that transfer from synthetic leg-count supervision is concentrated on the challenging *VLMBias* Animals slice. Both Qwen models start at 0% there, then rise to 34.6%/23.1% (SFT/RLVR) for Qwen2.5-VL-3B and 21.8%/22.2% for Qwen2.5-VL-7B. Gains on the *VLMBias* Other slice are smaller and less consistent. By contrast, *VLMBias Counting (Unbiased)* is already easy for the base models (Animals near ceiling, Other around the mid-60s), so post-training yields only minor changes and can slightly reduce accuracy. Additional synthetic-leg-count transfer results for Molmo2-4B and InternVL3.5-4B are reported in Appendix Figure 12. Overall, synthetic leg-count training can debias models on the targeted animal leg-count task, but this benefit does not generalize to other objects.

7 Activation Steering Analysis

7.1 VLM-Fix steering

Setup. We study whether late-layer activation steering can edit rule-sensitive behavior on VLM-Fix without retraining. In the main analysis, donor and target examples come from the same game under canonical rendering, base prompts, and direct answering. We patch one of the final 12 decoder layers at the query token during prompt processing and steer

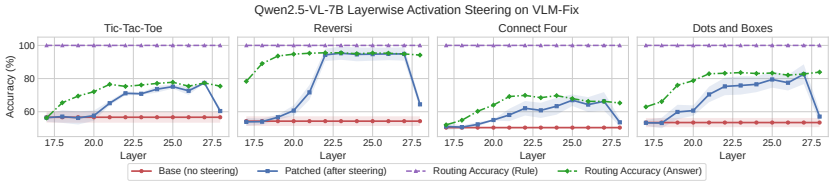

Figure 6: Layerwise activation-steering results on VLM-Fix for Qwen2.5-VL-7B across Tic-Tac-Toe, Reversi, Connect Four, and Dots and Boxes (left to right).

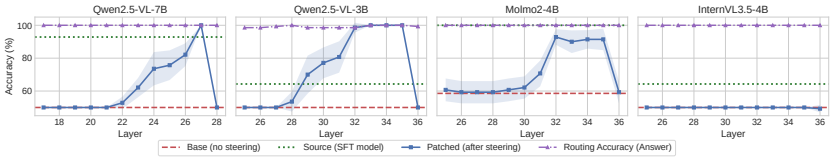

Figure 7: Layerwise SFT→Base activation-steering results on *VLMBias* Animals for Qwen2.5-VL-7B, Qwen2.5-VL-3B, Molmo2-4B, and InternVL3.5-4B (left to right).

the target activation toward a matched donor representation. Matching is determined by a lightweight router that predicts the relevant rule/answer bucket before patching. Full protocol details, split construction, and routing definitions are provided in Appendix D.1.

Results. Figure 6 shows that late-layer steering improves patched accuracy when routing is reliable, with the strongest gains in Reversi and Dots and Boxes. Gains are concentrated in later layers, where rule and answer routing are more stable, and are limited in harder settings such as Connect Four. Additional VLM-Fix steering results for Molmo2 and InternVL models are reported in Appendix D. Overall, the results indicate that rule-dependent representations are editable, but only when donor routing is accurate.

7.2 VLMBias Animals steering

Setup. We run a parallel donor-based steering analysis on the Animals subset of VLMBias in a controlled 2-leg versus 3-leg setting. Because base models are near 0% on the 3-leg slice, they do not provide reliable donor traces; we therefore use the SFT models trained on synthetic leg-count as the donor and steer the corresponding base model across the final 12 decoder layers using held-out train/test splits. As in the VLM-Fix analysis, steering is routed through a simple classifier before the patched generation step. Full dataset construction, split details, and steering protocol are given in Appendix D.2.

Results. Figure 7 shows that donor-based steering can improve patched accuracy, with the clearest late-layer gains in the Qwen models. Additional VLMBias Animals steering results for Molmo2-4B and InternVL3.5-4B are reported in Appendix D. Overall, the results indicate that the relevant counting representation is partly editable, but its accessibility varies across architectures and depends on having reliable donor traces (here provided by SFT).

8 Discussion and Limitations

In this work, we study semantic fixation in a controlled setting where perception is held fixed and only rule semantics change. Across games and models in VLM-Fix, identical boards yield a consistent standard–inverse gap, and prompt interventions strengthen this account: neutral aliasing narrows the gap, while semantically loaded aliases reopen it. Post-training results further show strong rule alignment, with robust same-rule gains and broader transfer primarily when both mappings are included during training. Consistent trends on *VLMBias* suggest that defamiliarizing image–prompt associations can also help in

a related prior-sensitive benchmark. At the same time, these conclusions are bounded by the evaluation settings, since VLM-Fix is synthetic by design and *VLMBias* provides external support in a related but not identical task family. Activation-steering results indicate that the effect is editable in late layers, but developing a full mechanistic account of how semantic representations are formed and used during inference remains an important direction for future work.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4971–4980, 2018.

Saeid Alavi Naeini, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati. Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset. *Advances in Neural Information Processing Systems*, 36:5631–5652, 2023.

Anthropic. Claude Model Overview. docs.anthropic.com/en/docs/about-claude/models, 2026. Accessed: 2026-03-21.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. Vismin: Visual minimal-change understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025b. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Merim Bilalić, Peter McLeod, and Fernand Gobet. Why good thoughts block better ones: The mechanism of the pernicious einstellung (set) effect. *Cognition*, 108(3):652–661, 2008. doi: 10.1016/j.cognition.2008.05.005.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 91–104, 2025.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10):833–838, 2023.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *Advances in Neural Information Processing Systems*, 36:63687–63723, 2023.

Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Wang. CPL: Counterfactual prompt learning for vision and language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3407–3418, 2022.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11975–11985, 2024. doi: 10.1109/CVPR52733.2024.01138.

Jen-tse Huang, Ruijia Wang, Yiqiao Jin, Yang Song, Esin Durmus, Dale Schuurmans, David Blei, Jacob Steinhardt, and Tatsunori Hashimoto. VisBias: A benchmark for measuring explicit and implicit social biases in vision language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 18534–18549, 2025a. doi: 10.18653/v1/2025.emnlp-main.908.

Saining Huang, Yukun Jia, Xing Shi, Siqi Miao, Lu Ma, Yilun Wang, and Chao Miao. An empirical study of anchoring effect in large language models. *arXiv preprint arXiv:2505.15392*, 2025b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *conference on Robot Learning*, pp. 991–1002. PMLR, 2022.

Michael Keller. Reversi. https://www.solitairelaboratory.com/abstract/reversi.html, 2026. Solitaire Laboratory. Accessed: 2026-03-23.

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific reports*, 15(1):39426, 2025.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

Yifan Li, Yixiao Du, Kun Cheng, Xin Zhou, Fanzhang Yu, Gang Song, Yuan Liu, Yichong Wang, Yunzhi Xie, Li Chen, et al. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023a.

Yifan Liu, Yujie Wang, Yixuan Ma, Yiyu Zhang, Yunchao Wei, Yinan Shi, Wenting Zhao, Yu You, Philip S. Yu, Kaipeng Yang, Jilan Sun, Hao Lu, Xuan Han, Yixuan Chen, Zhihai Wang, Jiaqi Dai, and Yu Qiao. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023b.

Abraham S. Luchins. Mechanization in problem solving: The effect of einstellung. *Psychological Monographs*, 54(6):i–95, 1942. doi: 10.1037/h0093502.

Federico Ruggeri and Debora Nozza. A multi-dimensional benchmark for measuring social biases in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7238–7251, 2023. doi: 10.18653/v1/2023.findings-acl.403.

Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. A vision check-up for language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14410–14419, 2024.

Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Qian Sun, Chao Zhang, Yuxin Dong, Kuang Hu, Yu Qiao, Chao Wen, Fei Wang, and Bin Wang. Causal-guided active learning for debiasing large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14374–14388, 2024. doi: 10.18653/v1/2024.acl-long.778.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Wikipedia contributors. Tic-tac-toe variants. Wikipedia, The Free Encyclopedia, 2026. Accessed: 2026-03-23.

Yuhang Wu, Jiaxin Lu, Jie Li, Xuanli He, Kun Zhou, Xinfeng Wang, Xin Wang, and Jianwei Gao. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14030–14058, 2024. doi: 10.18653/v1/2024.acl-long.758.

Mengzhou Xia, Yue Wang, Dan Zhang, and Mohit Bansal. Aligning large language models to improve and interpret social bias mitigation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7311–7335, 2024. doi: 10.18653/v1/2024.naacl-long.262.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Liu, Steven Lin, Zicheng Wang, and Lijuan Li. The dawn of LLMs: Preliminary explorations with GPT-4v(ision). *arXiv preprint arXiv:2309.17421*, 2023.

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. What if the TV was off? examining counterfactual reasoning abilities of multi-modal language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 21853–21862, 2023.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10185–10197, 2023.

A Dataset Details

A.1 Base-State Construction

We construct 300 unique terminal states per game, exclude draws, and balance canonical winners (150/150).

Tic-Tac-Toe (3×3). We sample legal alternating-play terminal boards with an exclusive single-line winner (exactly one winning line). To control terminal-pattern diversity, we enforce win-pattern quotas:

- 100 horizontal-win states,
- 100 vertical-win states,
- 50 main-diagonal-win states,
- 50 anti-diagonal-win states.

Reversi (5×5). We generate terminal states via random legal self-play from the standard center initialization, with pass handling applied whenever a player has no legal move and rollouts continued until a terminal board is reached. Draw states are removed.

Connect Four (4×4). We generate terminal states via random legal drop dynamics (gravity-constrained column drops) and stop each trajectory at the first winning terminal. Boards in which both players have winning lines are removed. A retained terminal may contain multiple winning lines, but only for the same winner. Across the 300 retained states, winning-line orientation counts are 109 vertical, 105 horizontal, 44 main-diagonal, and 42 anti-diagonal.

Dots and Boxes (6×6). Terminal boards are represented as fully assigned claimed-cell outcomes. These are synthetic terminal outcomes (not full edge-by-edge legal game trajectories). We enforce a non-zero margin constraint by sampling winner margins from $\{2, 4, 6, 8, 10, 12\}$.

A.2 Prompt-Output Formatting

For all prompt families, we enforce explicit answer-format instructions by response variant: Representative VLM-Fix rendering and prompt examples are shown in Figure 8.

- **Direct:** "Answer with only `<label1>` or `<label2>`. Do not add any other text."
- **CoT:** "Reason step by step. After that, give the answer inside `\boxed{ }`."

Representative VLMBias examples across Animals, Logos, Flags, and Game Boards are shown in Figure 9.

B Additional Results

B.1 Task-wise accuracy on VLM-Fix

Across the four game-specific direct evaluations (Tables 6, 7, 8, and 9), Reversi and Dots and Boxes show the highest standard accuracy (74.07% and 74.49%), while Connect Four is consistently the hardest (57.26%–61.79%). Inverse-rule accuracy is lower than standard under Base, Glyph, Checkerboard, and SemAlias, but Alias consistently closes this gap across all games (Tic-Tac-Toe: 64.80/63.93, Reversi: 68.32/66.58, Connect Four: 57.48/57.90, Dots and Boxes: 70.88/63.89). Overall, neutral aliasing is the strongest intervention for mitigating rule-inversion sensitivity in VLM-Fix.

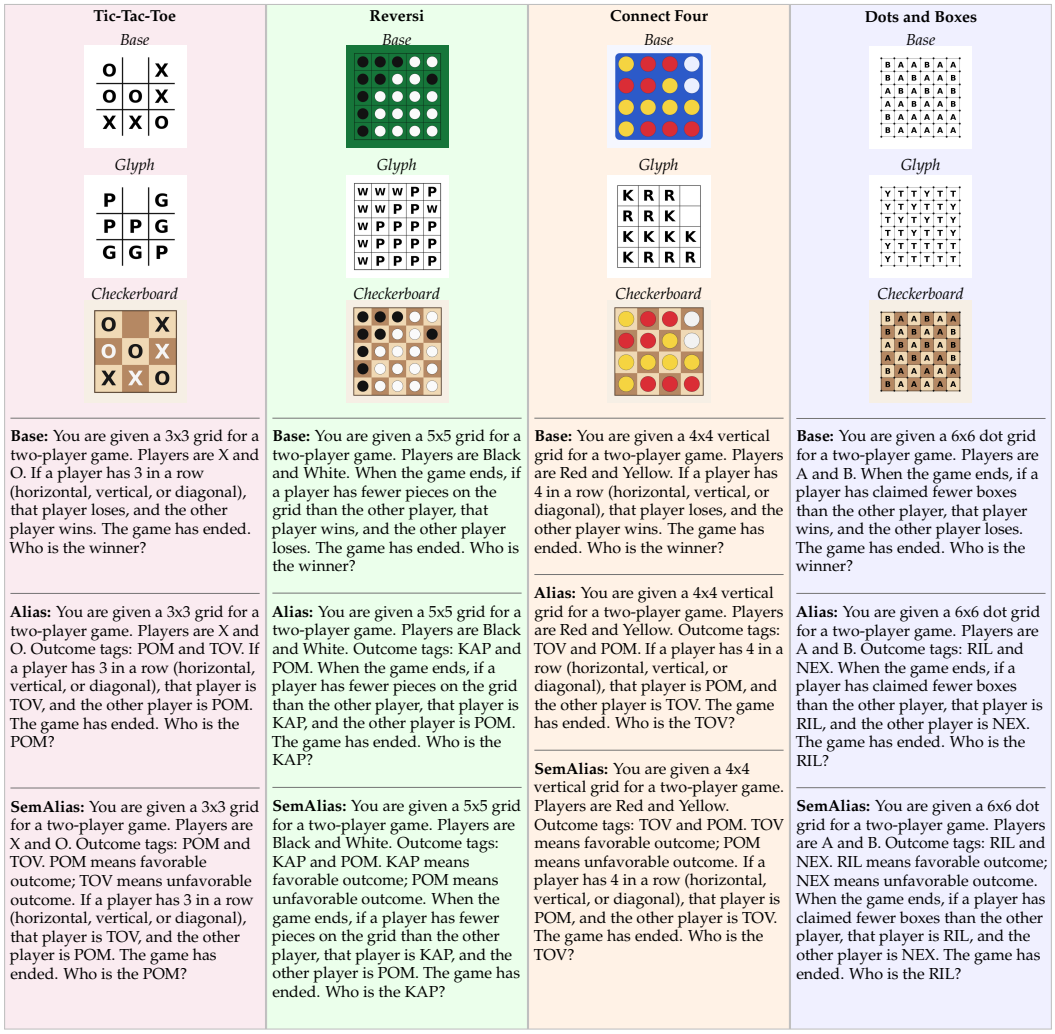

Figure 8: Representative VLM-Fix inputs across four games. In each column, the top three images show rendering variants in the order *Base* → *Glyph* → *Checkerboard*. Below the images, we show the three prompt variants used in this work: *Base*, *Alias*, and *SemAlias*.

B.2 Results with CoT Prompting

Tables 10, 11, 12, and 13 report the CoT results with the same intervention columns used in the direct tables. API models are included in the Base, Alias, and SemAlias CoT conditions; Glyph and Checkerboard CoT entries are only available for the open-weight models. Across games, Alias remains the most reliable way to improve inverse-rule accuracy, while Base and SemAlias retain larger rule gaps on several tasks.

B.3 Results with Descriptive prompting

Under descriptive prompting (Table 14), performance is generally strong and more balanced between standard and inverse settings than in rule-style prompting, with an overall mean of 70.99%/68.79%. GPT-5.2 is the strongest model overall (88.83%/85.25%), followed by GPT-4.1 (84.67%/82.58%); among open models, Qwen3-VL-32B is best (78.90%/78.90%). Across games, Dots and Boxes is the easiest on average (81.29%/80.66%), while Connect Four remains the most challenging (59.15%/56.04%).

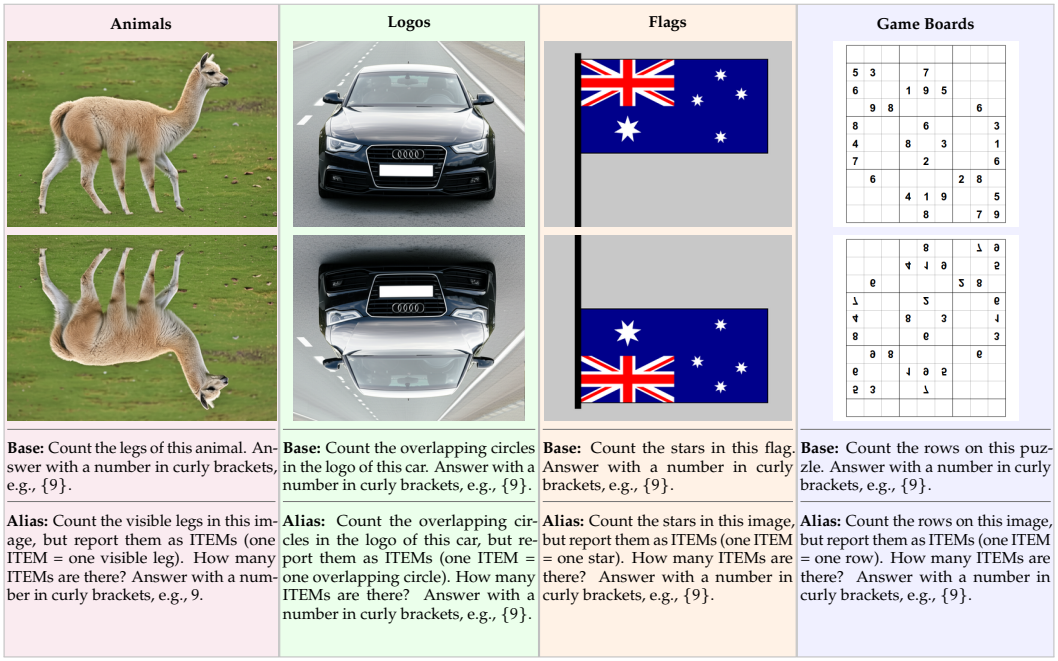

Figure 9: Representative examples from the four *VLMBias* subsets. Top row: Base images. Second row: Flip images. Third row: Base prompts. Fourth row: Alias prompts.

Table 6: VLM-Fix direct results on Tic-Tac-Toe. Cells report **standard/inverse accuracy (%)**; subscripts in non-baseline columns give Holm-adjusted paired McNemar p-values versus Base for the same rule. Within each row, red and blue mark the largest and smallest (Standard − Inverse) gaps.

Model	Base Std/Inv (Acc)	Glyph Std/Inv (Acc)	Checkerboard Std/Inv (Acc)	Alias Std/Inv (Acc)	SemAlias Std/Inv (Acc)
GPT-4.1	86.0/76.0	$84.0_{(\sim 1.00)}/60.3_{(<0.01)}$	$77.3_{(\sim 0.01)}/61.3_{(<0.01)}$	$88.0_{(\sim 0.42)}/85.3_{(<0.01)}$	$87.0_{(\sim 1.00)}/21.3_{(<0.01)}$
GPT-5.2	93.7/86.0	$82.0_{(<0.01)}/70.7_{(<0.01)}$	$89.0_{(\sim 0.20)}/70.0_{(<0.01)}$	$94.0_{(\sim 1.00)}/88.3_{(\sim 1.00)}$	$95.3_{(\sim 1.00)}/86.0_{(\sim 1.00)}$
Sonnet-4.0	51.0/51.7	$62.3_{(<0.01)}/52.3_{(\sim 1.00)}$	$54.3_{(\sim 0.10)}/50.0_{(\sim 0.75)}$	$57.7_{(<0.01)}/57.7_{(\sim 0.66)}$	$59.3_{(<0.01)}/50.0_{(\sim 1.00)}$
Sonnet-4.5	50.0/56.0	$61.0_{(\sim 0.01)}/58.3_{(\sim 1.00)}$	$50.0_{(\sim 1.00)}/52.7_{(\sim 0.26)}$	$54.3_{(<0.01)}/52.0_{(\sim 0.18)}$	$52.3_{(\sim 0.06)}/52.7_{(\sim 0.25)}$
Qwen2.5-VL-3B	54.2/49.9	$51.7_{(\sim 0.60)}/48.9_{(\sim 1.00)}$	$53.6_{(\sim 1.00)}/50.0_{(\sim 1.00)}$	$52.0_{(\sim 0.03)}/51.9_{(\sim 0.37)}$	$52.3_{(\sim 0.08)}/51.2_{(\sim 0.73)}$
Qwen2.5-VL-7B	66.5/44.7	$59.5_{(<0.01)}/49.1_{(\sim 0.16)}$	$65.2_{(\sim 1.00)}/43.9_{(\sim 1.00)}$	$65.8_{(\sim 1.00)}/63.7_{(<0.01)}$	$64.2_{(\sim 0.23)}/51.2_{(<0.01)}$
InternVL3.5-4B	60.7/45.9	$60.2_{(\sim 1.00)}/51.7_{(<0.01)}$	$54.4_{(<0.01)}/48.8_{(\sim 0.03)}$	$60.1_{(\sim 1.00)}/60.6_{(<0.01)}$	$57.2_{(\sim 0.02)}/55.7_{(<0.01)}$
InternVL3.5-8B	73.8/57.6	$69.1_{(<0.01)}/57.4_{(\sim 1.00)}$	$71.2_{(\sim 0.02)}/57.8_{(\sim 1.00)}$	$69.4_{(<0.01)}/66.3_{(<0.01)}$	$71.2_{(<0.01)}/49.8_{(<0.01)}$
InternVL3.5-14B	66.4/58.9	$68.0_{(\sim 1.00)}/58.9_{(\sim 1.00)}$	$65.2_{(\sim 1.00)}/55.4_{(\sim 0.02)}$	$67.2_{(\sim 1.00)}/67.8_{(<0.01)}$	$68.3_{(\sim 0.19)}/59.6_{(\sim 1.00)}$
Qwen3-VL-4B	57.4/48.0	$55.5_{(\sim 1.00)}/48.3_{(\sim 1.00)}$	$58.2_{(\sim 1.00)}/51.4_{(\sim 0.05)}$	$55.8_{(\sim 1.00)}/57.6_{(<0.01)}$	$54.4_{(\sim 0.17)}/52.8_{(<0.01)}$
Qwen3-VL-8B	57.6/47.7	$65.8_{(<0.01)}/49.7_{(\sim 0.68)}$	$55.8_{(\sim 0.09)}/50.0_{(\sim 0.14)}$	$54.7_{(<0.01)}/54.8_{(<0.01)}$	$54.6_{(<0.01)}/45.8_{(\sim 0.68)}$
Qwen3-VL-32B	70.0/66.4	$65.9_{(<0.01)}/61.3_{(<0.01)}$	$66.1_{(<0.01)}/64.9_{(\sim 0.28)}$	$68.2_{(\sim 0.03)}/70.0_{(<0.01)}$	$66.9_{(<0.01)}/62.3_{(\sim 0.02)}$
Molmo2-4B	61.3/42.8	$59.2_{(\sim 0.56)}/48.2_{(<0.01)}$	$62.0_{(\sim 1.00)}/52.3_{(<0.01)}$	$57.6_{(<0.01)}/56.7_{(<0.01)}$	$57.9_{(<0.01)}/51.6_{(<0.01)}$
Molmo2-8B	71.1/44.9	$68.8_{(\sim 0.24)}/49.8_{(<0.01)}$	$68.8_{(\sim 0.14)}/46.6_{(\sim 0.45)}$	$62.4_{(<0.01)}/62.3_{(<0.01)}$	$63.2_{(<0.01)}/51.7_{(<0.01)}$
Average	65.7/55.5	65.2/54.6	63.7/53.9	64.8/63.9	64.6/53.0

Table 7: VLM-Fix direct results on Reversi (same format as Table 6).

Model	Base Std/Inv (Acc)	Glyph Std/Inv (Acc)	Checkerboard Std/Inv (Acc)	Alias Std/Inv (Acc)	SemAlias Std/Inv (Acc)
GPT-4.1	85.7/75.0	87.3(-1.00)/77.7(-0.72)	86.3(-1.00)/79.3(-0.71)	87.3(-1.00)/80.0(-0.04)	86.3(-1.00)/81.3(<0.01)
GPT-5.2	95.7/91.0	92.0(-0.26)/84.3(-0.03)	90.3(-0.03)/86.7(-0.18)	96.3(-1.00)/93.0(-1.00)	95.3(-1.00)/88.7(-1.00)
Sonnet-4.0	76.0/50.3	80.7(-0.32)/55.3(-0.64)	69.3(-0.28)/51.3(-0.91)	84.7(-0.02)/71.3(<0.01)	84.7(-0.02)/78.0(<0.01)
Sonnet-4.5	85.7/67.0	86.7(-1.00)/81.3(<0.01)	79.0(-0.44)/69.7(-0.40)	89.0(-1.00)/82.3(<0.01)	86.7(-1.00)/75.7(<0.01)
Qwen2.5-VL-3B	57.2/50.0	57.7(-1.00)/43.1(<0.01)	56.4(-1.00)/50.0(-1.00)	58.0(-1.00)/55.8(<0.01)	54.5(-0.07)/55.8(<0.01)
Qwen2.5-VL-7B	61.3/41.5	70.1(<0.01)/40.4(-1.00)	58.6(-0.07)/44.6(-0.05)	60.2(-0.66)/63.1(<0.01)	59.8(-0.59)/41.2(-1.00)
InternVL3.5-4B	63.7/34.2	66.5(-0.19)/40.5(<0.01)	49.9(<0.01)/42.7(<0.01)	61.2(-0.09)/56.4(<0.01)	62.0(-0.19)/40.0(<0.01)
InternVL3.5-8B	63.2/45.2	72.9(<0.01)/46.2(-0.98)	54.5(<0.01)/40.2(<0.01)	55.2(<0.01)/51.7(<0.01)	71.3(<0.01)/42.5(-0.09)
InternVL3.5-14B	52.9/55.4	67.4(<0.01)/56.3(-1.00)	51.4(<0.01)/56.0(-1.00)	51.1(<0.01)/55.3(-1.00)	49.8(<0.01)/54.3(-1.00)
Qwen3-VL-4B	64.3/52.1	61.9(-0.29)/51.7(-1.00)	57.9(<0.01)/47.2(<0.01)	55.0(<0.01)/54.9(-0.38)	58.1(<0.01)/42.1(<0.01)
Qwen3-VL-8B	69.8/42.4	71.1(-1.00)/43.3(-1.00)	59.7(<0.01)/42.9(-1.00)	58.9(<0.01)/58.2(<0.01)	69.7(-1.00)/41.3(-1.00)
Qwen3-VL-32B	83.4/56.9	71.7(<0.01)/59.5(-0.24)	74.8(<0.01)/59.0(-0.24)	76.8(<0.01)/90.9(<0.01)	79.2(<0.01)/45.1(<0.01)
Molmo2-4B	65.1/22.2	72.5(<0.01)/25.1(-0.16)	68.7(<0.01)/25.0(<0.01)	48.5(<0.01)/58.7(<0.01)	61.7(-0.01)/38.0(<0.01)
Molmo2-8B	82.2/41.4	78.5(-0.10)/41.2(-1.00)	83.2(-0.84)/42.2(-1.00)	74.3(<0.01)/60.5(<0.01)	83.3(-0.84)/41.3(-1.00)
Average	71.9/51.8	74.1/53.3	67.1/52.6	68.3/66.6	71.6/54.7

Table 8: VLM-Fix direct results on Connect Four (same format as Table 6).

Model	Base Std/Inv (Acc)	Glyph Std/Inv (Acc)	Checkerboard Std/Inv (Acc)	Alias Std/Inv (Acc)	SemAlias Std/Inv (Acc)
GPT-4.1	69.0/41.3	74.0(-1.00)/48.3(-0.33)	73.3(-1.00)/52.0(-0.05)	69.7(-1.00)/62.7(<0.01)	69.0(-1.00)/36.3(-0.35)
GPT-5.2	69.0/53.0	71.0(-1.00)/65.3(<0.01)	69.0(-1.00)/53.0(-1.00)	68.0(-1.00)/56.0(-1.00)	71.7(-1.00)/51.7(-1.00)
Sonnet-4.0	63.3/65.7	56.0(-0.34)/69.7(-1.00)	60.3(-1.00)/68.0(-1.00)	65.3(-1.00)/78.3(<0.01)	68.3(-0.63)/82.7(<0.01)
Sonnet-4.5	53.0/86.7	74.0(<0.01)/87.7(-1.00)	67.0(<0.01)/72.7(<0.01)	60.7(<0.01)/77.3(<0.01)	62.0(<0.01)/92.7(-0.04)
Qwen2.5-VL-3B	51.7/50.6	52.2(-1.00)/49.7(-1.00)	49.4(-0.13)/50.2(-1.00)	50.6(-1.00)/49.6(-1.00)	50.3(-1.00)/47.7(-0.62)
Qwen2.5-VL-7B	58.9/48.2	62.7(-0.23)/50.7(-0.88)	57.0(-0.96)/48.5(-1.00)	60.7(-1.00)/57.8(<0.01)	57.5(-1.00)/47.9(-1.00)
InternVL3.5-4B	52.2/49.8	60.0(<0.01)/50.4(-1.00)	53.4(-0.95)/49.8(-1.00)	50.9(-1.00)/50.6(-1.00)	52.2(-1.00)/49.1(-1.00)
InternVL3.5-8B	54.9/49.0	68.2(<0.01)/55.9(<0.01)	56.2(-0.77)/50.2(-1.00)	54.4(-1.00)/54.1(<0.01)	55.4(-1.00)/48.7(-1.00)
InternVL3.5-14B	62.9/49.2	68.8(<0.01)/59.8(<0.01)	63.2(-1.00)/47.5(-0.07)	58.2(<0.01)/59.6(<0.01)	60.8(-0.34)/46.5(-0.21)
Qwen3-VL-4B	50.3/47.8	52.6(-1.00)/50.2(-0.62)	52.2(-1.00)/48.0(-1.00)	51.0(-1.00)/50.5(-0.33)	51.3(-1.00)/48.4(-1.00)
Qwen3-VL-8B	51.4/51.6	52.7(-1.00)/50.4(-0.93)	51.6(-1.00)/51.0(-0.84)	48.0(-0.37)/48.8(-0.34)	50.2(-1.00)/45.5(<0.01)
Qwen3-VL-32B	54.6/51.3	56.9(-0.35)/56.6(<0.01)	53.7(-0.87)/52.2(-0.88)	57.0(-0.08)/56.7(<0.01)	57.8(-0.04)/49.0(-0.88)
Molmo2-4B	55.2/48.7	54.2(-1.00)/50.2(-1.00)	56.1(-1.00)/48.5(-1.00)	51.7(-0.10)/51.7(-0.99)	51.8(-0.16)/48.2(-1.00)
Molmo2-8B	55.2/47.4	61.7(<0.01)/52.1(-0.02)	55.4(-1.00)/48.8(-0.51)	58.7(-0.03)/57.1(<0.01)	57.5(-0.16)/54.0(<0.01)
Average	57.3/52.9	61.8/56.9	58.4/52.9	57.5/57.9	58.3/53.5

Table 9: VLM-Fix direct results on Dots and Boxes (same format as Table 6).

Model	Base Std/Inv (Acc)	Glyph Std/Inv (Acc)	Checkerboard Std/Inv (Acc)	Alias Std/Inv (Acc)	SemAlias Std/Inv (Acc)
GPT-4.1	88.3/60.3	80.3(-0.02)/60.7(-1.00)	90.7(-1.00)/71.3(<0.01)	89.7(-1.00)/81.0(<0.01)	85.7(-0.91)/72.0(<0.01)
GPT-5.2	93.7/75.7	91.7(-1.00)/74.3(-1.00)	90.7(-1.00)/74.3(-1.00)	94.0(-1.00)/76.3(-1.00)	92.0(-1.00)/73.0(-1.00)
Sonnet-4.0	95.0/92.3	97.3(-1.00)/96.0(-0.21)	98.3(-0.10)/95.7(-0.30)	94.3(-1.00)/96.0(-0.21)	96.0(-1.00)/95.0(-0.30)
Sonnet-4.5	78.3/77.0	97.0(<0.01)/94.7(<0.01)	83.0(-0.78)/67.0(<0.01)	80.0(-1.00)/64.0(<0.01)	77.0(-1.00)/68.0(<0.01)
Qwen2.5-VL-3B	51.7/44.2	51.7(-1.00)/47.8(-0.37)	57.2(<0.01)/41.5(-0.37)	51.2(-1.00)/51.4(<0.01)	51.8(-1.00)/46.0(-0.58)
Qwen2.5-VL-7B	65.8/32.6	64.2(-0.52)/41.5(<0.01)	63.1(<0.01)/38.1(<0.01)	54.9(<0.01)/52.4(<0.01)	53.2(<0.01)/38.4(<0.01)
InternVL3.5-4B	70.1/30.5	58.8(<0.01)/43.3(<0.01)	69.5(-1.00)/39.8(<0.01)	71.8(-0.72)/50.4(<0.01)	70.4(-1.00)/32.4(-0.25)
InternVL3.5-8B	72.2/55.6	56.9(<0.01)/50.6(-0.03)	75.9(<0.01)/56.4(-0.79)	70.2(-0.14)/53.1(-0.38)	63.3(<0.01)/50.3(<0.01)
InternVL3.5-14B	53.2/42.9	59.1(<0.01)/48.4(<0.01)	56.6(<0.01)/42.3(-0.95)	54.8(-0.02)/69.7(<0.01)	53.9(-0.42)/49.8(<0.01)
Qwen3-VL-4B	65.0/38.8	53.5(<0.01)/47.2(<0.01)	60.2(<0.01)/44.0(<0.01)	59.5(<0.01)/48.5(<0.01)	57.5(<0.01)/39.2(-1.00)
Qwen3-VL-8B	76.2/36.7	68.2(<0.01)/36.8(-1.00)	74.9(-0.86)/39.0(-0.02)	75.8(-1.00)/58.7(<0.01)	73.2(-0.20)/39.3(<0.01)
Qwen3-VL-32B	73.7/55.9	70.7(-0.04)/53.5(-0.38)	76.2(<0.01)/55.9(-1.00)	72.6(-0.07)/76.4(<0.01)	71.2(<0.01)/57.6(-0.94)
Molmo2-4B	71.7/22.7	62.5(<0.01)/32.6(<0.01)	72.8(-0.72)/25.4(-0.03)	49.3(<0.01)/59.3(<0.01)	64.0(<0.01)/37.2(<0.01)
Molmo2-8B	77.6/34.5	71.8(<0.01)/41.5(<0.01)	73.7(<0.01)/41.3(<0.01)	74.1(<0.01)/57.2(<0.01)	75.0(<0.01)/42.9(<0.01)
Average	73.8/50.0	70.3/54.9	74.5/52.3	70.9/63.9	70.3/52.9

Table 10: VLM-Fix CoT results on Tic-Tac-Toe. API rows are included for the Base, Alias, and SemAlias CoT conditions; Glyph and Checkerboard CoT entries are unavailable for API models. Cells report **standard/inverse accuracy (%)**; subscripts in non-baseline columns give Holm-adjusted paired McNemar p-values versus Base for the same rule. Within each row, red and blue mark the largest and smallest (Standard − Inverse) gaps.

Model	Base Std/Inv (Acc)	Glyph Std/Inv (Acc)	Checkerboard Std/Inv (Acc)	Alias Std/Inv (Acc)	SemAlias Std/Inv (Acc)
GPT-4.1	97.0/94.0	–	–	$95.7_{(\sim 0.69)}$/$96.7_{(\sim 0.37)}$	$95.0_{(\sim 0.58)}$/$97.0_{(\sim 0.37)}$
GPT-5.2	100.0/99.7	–	–	$100.0_{(\sim 1.00)}$/$100.0_{(\sim 1.00)}$	$100.0_{(\sim 1.00)}$/$100.0_{(\sim 1.00)}$
Sonnet-4.0	95.0/96.7	–	–	$97.7_{(\sim 0.03)}$/$97.0_{(\sim 1.00)}$	$97.7_{(\sim 0.08)}$/$96.7_{(\sim 1.00)}$
Sonnet-4.5	96.7/95.3	–	–	$95.3_{(\sim 0.58)}$/$96.7_{(\sim 1.00)}$	$99.0_{(\sim 0.26)}$/$96.0_{(\sim 1.00)}$
Qwen2.5-VL-3B	59.8/47.6	$57.2_{(\sim 1.00)}$/$49.7_{(\sim 1.00)}$	$59.2_{(\sim 1.00)}$/$46.8_{(\sim 1.00)}$	$59.3_{(\sim 1.00)}$/$57.5_{(<0.01)}$	$58.8_{(\sim 1.00)}$/$54.3_{(<0.01)}$
Qwen2.5-VL-7B	71.3/32.8	$70.0_{(\sim 1.00)}$/$45.2_{(<0.01)}$	$70.9_{(\sim 1.00)}$/$33.6_{(\sim 1.00)}$	$69.0_{(\sim 0.83)}$/$69.3_{(<0.01)}$	$68.8_{(\sim 0.80)}$/$64.6_{(<0.01)}$
InternVL3.5-4B	85.0/58.4	$88.4_{(\sim 0.02)}$/$83.6_{(<0.01)}$	$83.1_{(\sim 0.30)}$/$62.6_{(\sim 0.02)}$	$81.2_{(<0.01)}$/$82.1_{(<0.01)}$	$78.7_{(<0.01)}$/$76.7_{(<0.01)}$
InternVL3.5-8B	90.9/85.2	$93.0_{(\sim 0.24)}$/$90.0_{(<0.01)}$	$89.8_{(\sim 0.69)}$/$82.3_{(\sim 0.12)}$	$89.2_{(\sim 0.29)}$/$89.7_{(<0.01)}$	$85.8_{(<0.01)}$/$83.4_{(\sim 0.40)}$
InternVL3.5-14B	86.7/82.9	$87.4_{(\sim 1.00)}$/$83.8_{(\sim 1.00)}$	$84.7_{(\sim 0.39)}$/$84.3_{(\sim 1.00)}$	$79.8_{(<0.01)}$/$83.3_{(\sim 1.00)}$	$77.3_{(<0.01)}$/$69.0_{(<0.01)}$
Qwen3-VL-4B	78.9/79.7	$73.4_{(<0.01)}$/$73.8_{(<0.01)}$	$75.8_{(\sim 0.21)}$/$74.8_{(\sim 0.01)}$	$79.7_{(\sim 1.00)}$/$80.2_{(\sim 1.00)}$	$80.9_{(\sim 0.59)}$/$81.5_{(\sim 0.79)}$
Qwen3-VL-8B	83.2/82.2	$76.3_{(<0.01)}$/$73.2_{(<0.01)}$	$87.0_{(\sim 0.02)}$/$84.8_{(\sim 0.37)}$	$80.8_{(\sim 0.18)}$/$82.7_{(\sim 1.00)}$	$84.9_{(\sim 0.36)}$/$84.1_{(\sim 0.69)}$
Qwen3-VL-32B	99.4/99.8	$98.3_{(\sim 0.12)}$/$98.0_{(<0.01)}$	$100.0_{(\sim 0.12)}$/$100.0_{(\sim 0.88)}$	$99.6_{(\sim 1.00)}$/$99.4_{(\sim 0.88)}$	$99.4_{(\sim 1.00)}$/$99.3_{(\sim 0.75)}$
Molmo2-4B	75.1/52.8	$72.1_{(\sim 0.29)}$/$59.5_{(<0.01)}$	$80.1_{(<0.01)}$/$55.2_{(\sim 0.33)}$	$77.2_{(\sim 0.29)}$/$78.8_{(<0.01)}$	$78.5_{(\sim 0.14)}$/$71.8_{(<0.01)}$
Molmo2-8B	79.8/63.2	$75.6_{(\sim 0.04)}$/$68.2_{(\sim 0.02)}$	$76.8_{(\sim 0.10)}$/$64.4_{(\sim 1.00)}$	$73.8_{(<0.01)}$/$73.2_{(<0.01)}$	$71.3_{(<0.01)}$/$56.6_{(<0.01)}$
Average	85.6/76.4	79.2/72.5	80.7/68.9	84.2/84.8	84.0/80.8

Table 11: VLM-Fix CoT results on Reversi.

Model	Base Std/Inv (Acc)	Glyph Std/Inv (Acc)	Checkerboard Std/Inv (Acc)	Alias Std/Inv (Acc)	SemAlias Std/Inv (Acc)
GPT-4.1	98.0/96.0	–	–	$96.7_{(\sim 1.00)}$/$96.7_{(\sim 1.00)}$	$97.7_{(\sim 1.00)}$/$97.3_{(\sim 1.00)}$
GPT-5.2	100.0/100.0	–	–	$100.0_{(\sim 1.00)}$/$99.7_{(\sim 1.00)}$	$100.0_{(\sim 1.00)}$/$99.0_{(\sim 1.00)}$
Sonnet-4.0	99.3/99.3	–	–	$99.7_{(\sim 1.00)}$/$99.7_{(\sim 1.00)}$	$100.0_{(\sim 1.00)}$/$99.3_{(\sim 1.00)}$
Sonnet-4.5	100.0/100.0	–	–	$100.0_{(\sim 1.00)}$/$100.0_{(\sim 1.00)}$	$100.0_{(\sim 1.00)}$/$100.0_{(\sim 1.00)}$
Qwen2.5-VL-3B	70.7/33.8	$67.6_{(\sim 0.25)}$/$36.9_{(\sim 0.32)}$	$64.1_{(<0.01)}$/$31.5_{(\sim 0.32)}$	$72.3_{(\sim 0.54)}$/$53.7_{(<0.01)}$	$76.8_{(<0.01)}$/$41.6_{(<0.01)}$
Qwen2.5-VL-7B	85.8/20.7	$87.8_{(\sim 0.36)}$/$29.6_{(<0.01)}$	$76.2_{(<0.01)}$/$29.9_{(<0.01)}$	$86.0_{(\sim 1.00)}$/$84.4_{(<0.01)}$	$85.6_{(\sim 1.00)}$/$29.4_{(<0.01)}$
InternVL3.5-4B	94.0/72.8	$93.9_{(\sim 1.00)}$/$73.5_{(\sim 1.00)}$	$95.2_{(\sim 0.75)}$/$79.6_{(<0.01)}$	$89.5_{(<0.01)}$/$93.8_{(<0.01)}$	$94.2_{(\sim 1.00)}$/$68.0_{(<0.01)}$
InternVL3.5-8B	94.1/71.5	$95.7_{(\sim 0.44)}$/$85.7_{(<0.01)}$	$93.1_{(\sim 0.61)}$/$71.8_{(\sim 1.00)}$	$89.7_{(<0.01)}$/$92.2_{(<0.01)}$	$92.8_{(\sim 0.53)}$/$82.3_{(<0.01)}$
InternVL3.5-14B	76.2/78.5	$85.1_{(<0.01)}$/$77.9_{(\sim 1.00)}$	$73.5_{(\sim 0.07)}$/$76.5_{(\sim 1.00)}$	$81.1_{(<0.01)}$/$88.5_{(<0.01)}$	$81.8_{(<0.01)}$/$78.6_{(\sim 1.00)}$
Qwen3-VL-4B	93.5/85.7	$95.2_{(\sim 0.62)}$/$78.8_{(<0.01)}$	$94.4_{(\sim 1.00)}$/$87.3_{(\sim 0.35)}$	$93.8_{(\sim 1.00)}$/$93.0_{(<0.01)}$	$93.1_{(\sim 1.00)}$/$90.7_{(<0.01)}$
Qwen3-VL-8B	98.0/92.3	$98.3_{(\sim 1.00)}$/$90.2_{(\sim 0.07)}$	$97.8_{(\sim 1.00)}$/$97.7_{(<0.01)}$	$97.7_{(\sim 1.00)}$/$97.8_{(<0.01)}$	$98.6_{(\sim 1.00)}$/$96.5_{(<0.01)}$
Qwen3-VL-32B	100.0/99.8	$100.0_{(\sim 1.00)}$/$99.9_{(\sim 1.00)}$	$100.0_{(\sim 1.00)}$/$99.8_{(\sim 1.00)}$	$99.4_{(\sim 0.12)}$/$99.1_{(\sim 0.07)}$	$99.8_{(\sim 1.00)}$/$97.2_{(<0.01)}$
Molmo2-4B	85.5/45.2	$90.8_{(<0.01)}$/$21.8_{(<0.01)}$	$85.2_{(\sim 1.00)}$/$37.6_{(<0.01)}$	$80.9_{(<0.01)}$/$79.3_{(<0.01)}$	$79.7_{(<0.01)}$/$50.3_{(\sim 0.02)}$
Molmo2-8B	92.0/42.2	$93.6_{(\sim 0.81)}$/$16.3_{(<0.01)}$	$90.4_{(\sim 0.81)}$/$36.3_{(<0.01)}$	$91.8_{(\sim 1.00)}$/$90.3_{(<0.01)}$	$92.4_{(\sim 1.00)}$/$47.6_{(<0.01)}$
Average	91.9/74.1	90.8/61.1	87.0/64.8	91.3/90.6	92.3/77.0

Table 12: VLM-Fix CoT results on Connect Four.

Model	Base Std/Inv (Acc)	Glyph Std/Inv (Acc)	Checkerboard Std/Inv (Acc)	Alias Std/Inv (Acc)	SemAlias Std/Inv (Acc)
GPT-4.1	98.7/99.0	–	–	$98.3_{(\sim 1.00)}$/$99.0_{(\sim 1.00)}$	$99.3_{(\sim 1.00)}$/$94.7_{(\sim 0.02)}$
GPT-5.2	99.0/99.0	–	–	$100.0_{(\sim 1.00)}$/$99.0_{(\sim 1.00)}$	$100.0_{(\sim 1.00)}$/$99.3_{(\sim 1.00)}$
Sonnet-4.0	99.7/99.7	–	–	$100.0_{(\sim 1.00)}$/$100.0_{(\sim 1.00)}$	$100.0_{(\sim 1.00)}$/$99.7_{(\sim 1.00)}$
Sonnet-4.5	100.0/98.7	–	–	$100.0_{(\sim 1.00)}$/$100.0_{(\sim 0.50)}$	$100.0_{(\sim 1.00)}$/$99.0_{(\sim 1.00)}$
Qwen2.5-VL-3B	43.5/42.0	$55.9_{(<0.01)}$/$46.0_{(\sim 0.10)}$	$40.0_{(\sim 0.13)}$/$33.5_{(<0.01)}$	$48.2_{(\sim 0.06)}$/$49.2_{(<0.01)}$	$49.8_{(<0.01)}$/$47.8_{(<0.01)}$
Qwen2.5-VL-7B	69.0/36.6	$67.9_{(\sim 1.00)}$/$44.8_{(<0.01)}$	$69.7_{(\sim 1.00)}$/$37.4_{(\sim 1.00)}$	$67.3_{(\sim 1.00)}$/$67.2_{(<0.01)}$	$63.2_{(<0.01)}$/$60.3_{(<0.01)}$
InternVL3.5-4B	77.8/66.2	$83.9_{(<0.01)}$/$83.0_{(<0.01)}$	$77.2_{(\sim 1.00)}$/$64.7_{(\sim 0.64)}$	$79.8_{(\sim 1.00)}$/$80.2_{(<0.01)}$	$79.4_{(\sim 1.00)}$/$73.5_{(<0.01)}$
InternVL3.5-8B	88.2/87.2	$89.7_{(\sim 1.00)}$/$86.8_{(\sim 1.00)}$	$89.2_{(\sim 1.00)}$/$87.2_{(\sim 1.00)}$	$88.5_{(\sim 1.00)}$/$85.2_{(\sim 0.97)}$	$88.4_{(\sim 1.00)}$/$81.1_{(<0.01)}$
InternVL3.5-14B	82.9/77.4	$84.2_{(\sim 1.00)}$/$83.8_{(<0.01)}$	$80.9_{(\sim 0.79)}$/$75.6_{(\sim 0.31)}$	$84.1_{(\sim 1.00)}$/$82.3_{(<0.01)}$	$82.0_{(\sim 1.00)}$/$71.2_{(<0.01)}$
Qwen3-VL-4B	91.6/93.4	$87.3_{(<0.01)}$/$86.5_{(<0.01)}$	$92.8_{(\sim 0.54)}$/$94.9_{(\sim 0.74)}$	$93.7_{(\sim 0.23)}$/$92.3_{(\sim 0.74)}$	$94.6_{(\sim 0.02)}$/$94.8_{(\sim 0.74)}$
Qwen3-VL-8B	86.3/88.8	$62.7_{(<0.01)}$/$62.2_{(<0.01)}$	$99.1_{(<0.01)}$/$99.2_{(<0.01)}$	$89.2_{(\sim 0.03)}$/$89.2_{(\sim 1.00)}$	$90.7_{(<0.01)}$/$88.8_{(\sim 1.00)}$
Qwen3-VL-32B	99.6/99.7	$99.2_{(\sim 1.00)}$/$99.2_{(\sim 1.00)}$	$99.9_{(\sim 1.00)}$/$99.9_{(\sim 1.00)}$	$99.7_{(\sim 1.00)}$/$99.4_{(\sim 1.00)}$	$99.8_{(\sim 1.00)}$/$99.8_{(\sim 1.00)}$
Molmo2-4B	56.2/56.7	$68.4_{(<0.01)}$/$62.6_{(\sim 0.01)}$	$57.5_{(\sim 0.95)}$/$54.8_{(\sim 0.62)}$	$66.3_{(<0.01)}$/$65.6_{(<0.01)}$	$66.8_{(<0.01)}$/$61.3_{(\sim 0.05)}$
Molmo2-8B	73.3/64.2	$72.6_{(\sim 1.00)}$/$66.4_{(\sim 1.00)}$	$71.8_{(\sim 1.00)}$/$64.8_{(\sim 1.00)}$	$66.8_{(<0.01)}$/$65.7_{(\sim 1.00)}$	$67.8_{(\sim 0.01)}$/$52.4_{(<0.01)}$
Average	83.3/79.2	77.2/72.1	77.8/71.2	84.4/83.9	84.4/80.3

Table 13: VLM-Fix CoT results on Dots & Boxes.

Model	Base Std/Inv (Acc)	Glyph Std/Inv (Acc)	Checkerboard Std/Inv (Acc)	Alias Std/Inv (Acc)	SemAlias Std/Inv (Acc)
GPT-4.1	85.7/81.7	–	–	88.0$_{(-0.76)}$/85.0$_{(-0.59)}$	90.7$_{(-0.13)}$/86.0$_{(-0.44)}$
GPT-5.2	100.0/100.0	–	–	99.7$_{(-1.00)}$/100.0$_{(-1.00)}$	100.0$_{(-1.00)}$/100.0$_{(-1.00)}$
Sonnet-4.0	100.0/99.7	–	–	99.7$_{(-1.00)}$/99.7$_{(-1.00)}$	99.7$_{(-1.00)}$/99.7$_{(-1.00)}$
Sonnet-4.5	100.0/100.0	–	–	100.0$_{(-1.00)}$/100.0$_{(-1.00)}$	100.0$_{(-1.00)}$/100.0$_{(-1.00)}$
Qwen2.5-VL-3B	52.1/49.0	52.5$_{(-1.00)}$/42.0$_{(-0.02)}$	50.6$_{(-1.00)}$/45.8$_{(-0.23)}$	57.8$_{(-0.02)}$/45.5$_{(-0.26)}$	59.8$_{(<0.01)}$/47.8$_{(-1.00)}$
Qwen2.5-VL-7B	63.1/38.5	65.2$_{(-0.56)}$/38.2$_{(-1.00)}$	65.8$_{(-0.45)}$/40.7$_{(-0.88)}$	69.0$_{(<0.01)}$/65.1$_{(<0.01)}$	68.4$_{(-0.01)}$/45.2$_{(<0.01)}$
InternVL3.5-4B	89.2/74.2	86.9$_{(-0.42)}$/67.9$_{(<0.01)}$	93.4$_{(<0.01)}$/84.2$_{(<0.01)}$	89.6$_{(-1.00)}$/86.0$_{(<0.01)}$	90.0$_{(-1.00)}$/66.4$_{(<0.01)}$
InternVL3.5-8B	79.1/81.2	75.9$_{(-0.09)}$/81.0$_{(-1.00)}$	83.6$_{(<0.01)}$/80.4$_{(-1.00)}$	88.5$_{(<0.01)}$/84.8$_{(-0.09)}$	91.5$_{(<0.01)}$/74.2$_{(<0.01)}$
InternVL3.5-14B	64.9/54.6	60.3$_{(-0.08)}$/59.5$_{(-0.08)}$	67.8$_{(-0.24)}$/60.2$_{(-0.04)}$	52.6$_{(<0.01)}$/52.4$_{(-1.00)}$	57.6$_{(<0.01)}$/55.1$_{(-1.00)}$
Qwen3-VL-4B	59.7/58.3	72.2$_{(<0.01)}$/68.8$_{(<0.01)}$	89.1$_{(<0.01)}$/89.0$_{(<0.01)}$	56.2$_{(-0.28)}$/61.8$_{(-0.32)}$	58.8$_{(-1.00)}$/61.3$_{(-0.32)}$
Qwen3-VL-8B	59.1/57.5	53.7$_{(-0.05)}$/56.8$_{(-1.00)}$	71.0$_{(<0.01)}$/68.8$_{(<0.01)}$	57.8$_{(-1.00)}$/62.2$_{(-0.06)}$	63.0$_{(-0.16)}$/65.2$_{(<0.01)}$
Qwen3-VL-32B	93.5/93.6	90.8$_{(-0.05)}$/90.5$_{(-0.02)}$	97.9$_{(<0.01)}$/98.2$_{(<0.01)}$	93.0$_{(-1.00)}$/95.7$_{(-0.04)}$	96.5$_{(<0.01)}$/90.4$_{(-0.02)}$
Molmo2-4B	83.9/13.0	79.0$_{(<0.01)}$/19.0$_{(<0.01)}$	82.3$_{(-1.00)}$/18.8$_{(<0.01)}$	82.7$_{(-1.00)}$/66.2$_{(<0.01)}$	85.0$_{(-1.00)}$/48.8$_{(<0.01)}$
Molmo2-8B	79.0/30.3	78.2$_{(-1.00)}$/33.1$_{(-0.23)}$	67.0$_{(<0.01)}$/41.6$_{(<0.01)}$	76.9$_{(-1.00)}$/77.8$_{(<0.01)}$	76.8$_{(-1.00)}$/53.8$_{(<0.01)}$
Average	79.2/66.5	71.5/55.7	76.8/62.8	79.4/77.3	81.3/71.0

Table 14: VLM-Fix Direct Results under Descriptive Prompting (Canonical Rendering). Each cell reports **Standard/Inverse accuracy (%)** for the corresponding game. The **Average** column is the per-model mean across the four games.

Model	Tic-Tac-Toe	Reversi	Connect Four	Dots and Boxes	Average
GPT-4.1	85.0/87.0	91.0/92.0	72.0/60.3	90.7/91.0	84.7/82.6
GPT-5.2	91.7/91.7	94.3/94.0	73.0/67.0	96.3/88.3	88.8/85.2
Sonnet-4.0	70.0/58.7	70.3/50.7	64.3/53.7	95.7/96.0	75.1/64.8
Sonnet-4.5	61.3/54.7	88.0/52.0	71.0/56.0	99.7/98.3	80.0/65.2
Qwen2.5-VL-3B	52.3/52.3	72.8/72.9	49.7/49.6	48.7/48.8	55.9/55.9
Qwen2.5-VL-7B	62.2/62.3	77.2/77.2	55.9/55.8	85.8/85.7	70.2/70.2
InternVL3.5-4B	54.8/54.8	70.8/70.8	50.2/50.1	79.9/79.9	63.9/63.9
InternVL3.5-8B	72.0/72.0	62.4/62.5	55.0/55.1	89.8/89.8	69.8/69.8
InternVL3.5-14B	69.2/69.3	56.0/56.0	58.8/58.7	66.4/66.5	62.6/62.6
Qwen3-VL-4B	53.6/53.7	79.8/79.8	51.2/51.2	58.6/58.5	60.8/60.8
Qwen3-VL-8B	54.1/54.1	72.0/72.0	53.7/53.7	81.8/81.8	65.4/65.4
Qwen3-VL-32B	73.2/73.2	87.8/87.8	62.8/62.8	91.8/91.8	78.9/78.9
Molmo2-4B	54.8/54.8	80.1/80.1	54.8/54.8	71.6/71.6	65.3/65.3
Molmo2-8B	62.7/62.7	89.6/89.6	55.8/55.8	81.4/81.4	72.4/72.4
Average	65.5/64.4	78.0/74.1	59.1/56.0	81.3/80.7	71.0/68.8

B.4 Text-only Input Results

Prompt format for text-only inputs. Each text-only query concatenates an instruction prompt and a text board with a blank-line separator. The templates below follow the same content as `vlm_fix_text_only_prompts.tex`, using ASCII board rendering for compatibility in this paper.

```
<PROMPT>

Board:
<ASCII_BOARD>
```

```
Answer with only <label1> or <label2>. Do not add any other text.
```

Inverse-rule direct prompt templates.

Tic-Tac-Toe (3x3)

```
You are given a 3x3 grid for a two-player game. Players are X and O. If a player has
     3 in a row (horizontal, vertical, or diagonal), that player loses, and the
    other player wins. The game has ended. Who is the winner? Answer with only X or
     O. Do not add any other text.

Board:
 X | O | X
---+---+---
 O | X | O
---+---+---
 X | | O
```

Reversi (5x5)

```
You are given a 5x5 grid for a two-player game. Players are Black and White. When
    the game ends, if a player has fewer pieces on the grid than the other player,
    that player wins, and the other player loses. The game has ended. Who is the
    winner? Answer with only Black or White. Do not add any other text.

Board:
+---+---+---+---+---+
| B | W | B | W | B |
+---+---+---+---+---+
| W | B | W | B | W |
+---+---+---+---+---+
| B | W | B | W | B |
+---+---+---+---+---+
| W | B | W | B | W |
+---+---+---+---+---+
| B | W | B | W | B |
+---+---+---+---+---+
```

Connect Four (4x4)

```
You are given a 4x4 vertical grid for a two-player game. Players are Red and Yellow.
     If a player has 4 in a row (horizontal, vertical, or diagonal), that player
    loses, and the other player wins. The game has ended. Who is the winner? Answer
     with only Red or Yellow. Do not add any other text.

Board:
```

```
 R | Y | R | Y
---+---+---+---
 Y | R | Y | R
---+---+---+---
 R | Y | R | Y
---+---+---+---
 Y | R | Y | R
```

Dots and Boxes (6x6)

```
You are given a 6x6 dot grid for a two-player game. Players are A and B. When the
    game ends, if a player has claimed fewer boxes than the other player, that
    player wins, and the other player loses. The game has ended. Who is the winner?
     Answer with only A or B. Do not add any other text.

Board:
o---o---o---o---o---o---o
| A | B | A | B | A | B |
o---o---o---o---o---o---o
| B | A | B | A | B | A |
o---o---o---o---o---o---o
| A | B | A | B | A | B |
o---o---o---o---o---o---o
| B | A | B | A | B | A |
o---o---o---o---o---o---o
| A | B | A | B | A | B |
o---o---o---o---o---o---o
| B | A | B | A | B | A |
o---o---o---o---o---o---o
```

In the text-only setting (Table 15), performance is lower than image-based results with CoT or descriptive prompting, and it shows a clear standard–inverse asymmetry (overall 69.22%/58.62%). Dots and Boxes has the highest standard accuracy (71.20%), while Connect Four is lower but more balanced (64.92%/57.07%). GPT-5.2 remains the strongest model (94.83%/90.42%). Most open models show larger inverse degradation.

B.5 Input-Order Marginals for Open Models

Tables 16 and 17 summarize image-first versus text-first marginals for the 10 open-weight models, averaging over the four VLM-Fix games. In the direct setting, text-first ordering tends to amplify the standard–inverse gap for the base, glyph, and checkerboard conditions, whereas alias prompting and descriptive prompting remain much more balanced. Under CoT prompting, the same qualitative pattern persists, but the order effect is generally smaller and alias prompting remains the most balanced condition overall.

B.6 VLMBias Results

Table 18 reports the full model-wise averages across the four *VLMBias* subsets under the four input configurations.

Task-wise accuracy in *VLMBias* (Tables 19, 20, 21, and 22) is strongly subset-dependent and generally improves with alias-based prompting. Averaged across models, Flags shows the highest accuracy overall (25.89, 24.23, 30.57, and 29.02 across the four configurations), while Animals is the most difficult under base settings (3.62/5.70) but rises markedly with aliasing, especially Flip+Alias (22.17). Game Boards and Logos remain in a mid-low range, improving from 11.39 to 16.16 and from 13.85 to 16.65/15.72, respectively, indicating that prompt and image perturbations can help accuracy but do not fully close the gap across subsets.

Table 15: VLM-Fix text-only direct results (canonical text board and standard prompt). API models are aligned to the same reduced 300-state winner-only seed subset used in their direct API runs. Each cell reports **Standard/Inverse accuracy (%)** for the corresponding game. The **Average** column is the per-model mean across the four games.

Model	Tic-Tac-Toe	Reversi	Connect Four	Dots and Boxes	Average
GPT-4.1	96.3/57.7	90.0/91.3	88.3/40.0	93.0/95.7	91.9/71.2
GPT-5.2	98.3/93.7	95.3/95.7	87.7/82.0	97.7/90.3	94.8/90.4
Sonnet-4.0	86.0/71.7	93.0/93.0	82.3/80.3	100.0/98.3	90.3/85.8
Sonnet-4.5	77.7/67.3	99.0/95.0	66.7/71.0	100.0/99.7	85.8/83.2
Qwen2.5-VL-3B	51.5/50.0	50.3/50.0	49.8/50.2	57.2/42.5	52.2/48.2
Qwen2.5-VL-7B	63.0/45.5	70.3/36.7	70.2/49.8	70.5/34.2	68.5/41.5
InternVL3.5-4B	50.2/49.8	51.5/48.8	54.8/46.2	57.8/46.8	53.6/47.9
InternVL3.5-8B	55.0/52.8	73.2/52.2	61.2/58.3	54.5/50.5	61.0/53.5
InternVL3.5-14B	72.0/58.7	59.7/59.7	53.3/52.5	47.2/51.0	58.0/55.5
Qwen3-VL-4B	58.8/52.7	50.7/54.0	54.3/55.3	52.8/38.3	54.2/50.1
Qwen3-VL-8B	65.3/53.2	57.2/36.8	55.8/49.2	57.5/40.7	59.0/45.0
Qwen3-VL-32B	73.0/61.2	67.2/58.2	65.7/58.5	65.5/58.3	67.8/59.0
Molmo2-4B	62.7/48.2	57.5/26.2	54.2/50.5	66.5/41.0	60.2/41.5
Molmo2-8B	67.7/56.5	78.3/38.8	64.5/55.2	76.7/41.3	71.8/48.0
Average	69.8/58.5	70.9/59.7	64.9/57.1	71.2/59.2	69.2/58.6

Table 16: VLM-Fix order marginals for direct prompting, averaged over the four games and reported for the 10 open-weight models only. Each cell reports standard/inverse accuracy (%) for a fixed condition under image-first versus text-first input order.

Model	Base		Glyph		Checkerboard		Alias		SemAlias		Descriptive	
	Img-first	Text-first	Img-first	Text-first	Img-first	Text-first	Img-first	Text-first	Img-first	Text-first	Img-first	Text-first
Qwen2.5-VL-3B	53.5/49.3	53.8/48.1	53.6/47.6	53.0/47.1	54.1/49.2	54.2/46.6	55.0/54.2	50.9/50.2	53.2/52.2	51.3/48.2	56.1/56.2	55.6/55.5
Qwen2.5-VL-7B	61.2/44.2	65.0/39.3	59.6/46.2	68.7/44.7	61.1/43.3	60.8/44.2	60.8/59.4	60.0/59.1	60.4/44.5	57.0/44.9	72.2/72.2	68.3/68.2
InternVL3.5-4B	54.5/42.8	68.9/37.3	56.7/44.8	66.1/48.2	52.5/46.5	61.1/44.1	56.0/56.6	66.0/52.4	56.7/43.2	64.2/45.3	58.5/58.5	69.4/69.3
InternVL3.5-8B	61.8/52.8	70.2/51.0	63.4/50.9	70.1/54.2	60.7/51.5	68.2/50.8	59.2/53.8	65.4/58.8	64.2/49.4	66.5/46.2	67.5/67.5	72.1/72.1
InternVL3.5-14B	56.4/48.1	61.3/55.1	61.3/53.2	70.3/58.5	56.7/48.8	61.5/51.8	54.0/58.8	61.6/67.4	54.4/51.6	62.0/53.5	59.4/59.5	65.8/65.8
Qwen3-VL-4B	56.4/49.0	62.1/44.4	53.2/52.4	58.6/46.3	55.2/49.6	59.1/45.7	55.0/54.0	55.6/51.8	56.8/46.3	53.8/44.9	61.2/61.2	60.4/60.4
Qwen3-VL-8B	64.0/49.2	63.5/40.0	60.9/48.5	68.0/41.6	62.7/50.6	58.2/40.9	58.8/54.2	59.9/56.0	61.8/46.9	62.0/39.1	66.1/66.1	64.7/64.6
Qwen3-VL-32B	69.8/58.2	71.0/57.1	63.4/56.0	69.2/59.5	65.5/58.5	69.9/57.6	67.6/68.8	69.7/78.2	68.5/50.1	69.0/56.9	79.3/79.3	78.5/78.5
Molmo2-4B	57.4/32.2	69.2/36.0	57.7/39.7	66.5/38.3	60.3/35.5	69.5/40.1	51.2/61.0	52.3/52.2	58.5/41.4	59.2/46.1	65.5/65.5	65.2/65.2
Molmo2-8B	66.5/44.2	76.5/39.9	65.4/44.5	75.0/47.9	65.5/46.9	75.0/42.6	62.7/57.0	72.1/61.5	66.2/45.3	73.3/49.7	73.5/73.5	71.3/71.3
Average	60.2/47.0	66.2/44.8	59.5/48.4	66.5/48.6	59.4/48.0	63.8/46.4	58.0/57.8	61.3/58.8	60.1/47.1	61.9/47.5	65.9/66.0	67.1/67.1

C Post-Training

C.1 VLM-Fix Post-Training Details

C.1.1 Game/state generation details

We build balanced terminal-state pools for all four games. Reversi and Dots and Boxes each use 1024 generated terminal states. Connect Four also uses 1024 states, with balanced win-pattern categories (341 horizontal, 341 vertical, 171 main diagonal, 171 anti-diagonal). For Tic-Tac-Toe, we start from 920 exclusive single-line terminal-winner states and add 104 sampled top-up slots to reach 1024 total state slots.

C.1.2 Exact split sizes and composition

- **D1 train (8192):** 4 games $\times$ 1024 states/game $\times$ 2 targets (winner/loser), standard rule only.
- **D2 train (8192):** 4 games $\times$ 1024 states/game $\times$ 2 targets, inverse rule only.
- **D3 train (8192):** 2 games (Tic-Tac-Toe, Reversi) $\times$ 2 rules $\times$ 1024 states/game $\times$ 2 targets.
- **D1 test (2400):** canonical evaluation file for D1-trained models: all 4 games, inverse rule only, 300 benchmark states/game $\times$ 2 targets.

Table 17: VLM-Fix order marginals for CoT prompting, averaged over the four games and reported for the 10 open-weight models only. Each cell reports standard/inverse accuracy (%) for a fixed condition under image-first versus text-first input order.

Model	Base		Glyph		Checkerboard		Alias		SemAlias	
	Img-first	Text-first	Img-first	Text-first	Img-first	Text-first	Img-first	Text-first	Img-first	Text-first
Qwen2.5-VL-3B	57.2/44.3	55.8/41.9	58.5/43.9	58.1/43.4	56.5/42.4	50.5/36.4	59.6/52.0	59.2/51.0	63.0/48.4	59.6/47.3
Qwen2.5-VL-7B	72.2/30.2	72.4/34.1	71.7/36.6	73.8/42.2	70.2/33.3	71.1/37.5	75.0/71.0	70.6/72.0	71.0/50.4	72.0/49.4
InternVL3.5-4B	88.0/65.8	85.0/70.1	88.6/77.2	88.0/76.8	87.8/73.0	86.6/72.6	85.2/84.9	84.9/86.2	87.7/71.1	83.5/71.2
InternVL3.5-8B	88.8/83.5	87.4/79.0	89.2/86.4	88.0/85.3	89.5/82.4	88.4/78.5	88.2/85.4	89.7/90.5	87.4/78.5	91.8/82.0
InternVL3.5-14B	78.6/71.8	76.7/74.9	80.4/77.1	78.2/75.4	77.1/75.6	76.4/72.7	75.2/75.1	73.6/78.2	75.8/72.1	73.6/64.8
Qwen3-VL-4B	80.0/77.9	81.8/80.6	83.9/74.7	80.2/79.2	87.6/87.4	88.4/85.7	79.5/80.3	82.1/83.3	80.3/79.8	83.4/84.3
Qwen3-VL-8B	82.5/78.3	80.8/82.1	73.1/69.2	72.4/71.9	93.6/91.4	83.8/83.9	81.0/83.8	81.7/82.1	84.9/84.0	83.7/83.3
Qwen3-VL-32B	99.5/99.4	96.7/97.0	97.8/97.7	96.3/96.1	100.0/99.8	99.0/99.2	98.8/99.0	97.0/97.8	99.5/97.2	98.3/96.1
Molmo2-4B	74.7/45.1	75.7/38.7	77.9/42.7	77.2/38.8	74.5/43.6	78.0/39.6	77.4/73.5	76.2/71.5	77.0/56.9	78.0/59.2
Molmo2-8B	81.6/48.2	80.5/51.7	79.4/44.8	80.6/47.2	77.6/49.6	75.4/54.0	75.3/76.9	79.4/76.7	76.8/60.9	77.4/44.3
Average	80.3/64.5	79.3/65.0	80.0/65.0	79.3/65.6	81.4/67.8	79.8/66.0	79.5/78.2	79.4/78.9	80.3/69.9	80.1/68.2

Table 18: VLMBias results averaged across the four subsets (Animals, Flags, Game Boards, Logos). Accuracy and Bias are shown side-by-side for four input configurations. Cells report percentages; for non-baseline columns, subscripts show Holm-adjusted paired McNemar p-values versus Base (within the same metric). Within each model row, blue marks the best value and red marks the worst value.

Model	Accuracy ↑				Bias ↓			
	Base	Flip	Alias	Flip+Alias	Base	Flip	Alias	Flip+Alias
GPT-4.1	9.9	10.5$_{(\sim 1.00)}$	12.0$_{(\sim 0.09)}$	10.7$_{(\sim 1.00)}$	81.0	71.7$_{(<0.01)}$	73.0$_{(<0.01)}$	76.1$_{(<0.01)}$
GPT-5.2	20.2	21.0$_{(\sim 1.00)}$	26.2$_{(<0.01)}$	32.0$_{(<0.01)}$	76.8	70.8$_{(<0.01)}$	68.7$_{(<0.01)}$	54.4$_{(<0.01)}$
Sonnet-4.0	6.2	6.5$_{(\sim 1.00)}$	7.2$_{(\sim 0.35)}$	9.8$_{(<0.01)}$	89.3	84.0$_{(<0.01)}$	87.5$_{(\sim 0.08)}$	79.5$_{(<0.01)}$
Sonnet-4.5	6.9	8.6$_{(\sim 0.07)}$	9.3$_{(<0.01)}$	17.3$_{(<0.01)}$	92.3	87.2$_{(<0.01)}$	85.9$_{(<0.01)}$	77.0$_{(<0.01)}$
Qwen2.5-VL-3B	7.4	7.4$_{(\sim 1.00)}$	8.3$_{(\sim 0.42)}$	10.7$_{(<0.01)}$	66.4	63.6$_{(<0.01)}$	59.5$_{(<0.01)}$	53.8$_{(<0.01)}$
Qwen2.5-VL-7B	9.5	8.8$_{(\sim 0.76)}$	15.9$_{(<0.01)}$	18.7$_{(<0.01)}$	75.5	72.4$_{(<0.01)}$	69.8$_{(<0.01)}$	62.7$_{(<0.01)}$
InternVL3.5-4B	6.6	11.1$_{(<0.01)}$	9.4$_{(<0.01)}$	14.0$_{(<0.01)}$	68.9	61.3$_{(<0.01)}$	55.0$_{(<0.01)}$	30.7$_{(<0.01)}$
InternVL3.5-8B	7.2	9.1$_{(\sim 0.04)}$	11.9$_{(<0.01)}$	27.9$_{(<0.01)}$	77.1	65.9$_{(<0.01)}$	67.5$_{(<0.01)}$	40.7$_{(<0.01)}$
InternVL3.5-14B	13.0	14.7$_{(\sim 0.07)}$	15.8$_{(<0.01)}$	31.7$_{(<0.01)}$	65.6	61.5$_{(<0.01)}$	56.4$_{(<0.01)}$	40.8$_{(<0.01)}$
Qwen3-VL-4B	8.2	10.4$_{(<0.01)}$	10.8$_{(<0.01)}$	16.7$_{(<0.01)}$	83.5	77.1$_{(<0.01)}$	74.9$_{(<0.01)}$	65.2$_{(<0.01)}$
Qwen3-VL-8B	5.6	8.6$_{(<0.01)}$	7.6$_{(<0.01)}$	14.2$_{(<0.01)}$	86.5	81.8$_{(<0.01)}$	79.2$_{(<0.01)}$	72.9$_{(<0.01)}$
Qwen3-VL-32B	8.3	11.0$_{(<0.01)}$	12.3$_{(<0.01)}$	16.8$_{(<0.01)}$	88.0	83.6$_{(<0.01)}$	84.2$_{(<0.01)}$	77.6$_{(<0.01)}$
Molmo2-4B	25.2	29.9$_{(<0.01)}$	34.6$_{(<0.01)}$	36.4$_{(<0.01)}$	66.2	57.3$_{(<0.01)}$	54.7$_{(<0.01)}$	44.7$_{(<0.01)}$
Molmo2-8B	27.8	27.9$_{(\sim 1.00)}$	28.3$_{(\sim 1.00)}$	32.7$_{(<0.01)}$	56.5	54.9$_{(\sim 0.63)}$	57.0$_{(\sim 1.00)}$	48.7$_{(<0.01)}$
Average	11.6	13.3	15.0	20.7	76.7	70.9	69.5	58.9

- **D2 test (2400):** canonical evaluation file for D2-trained models: all 4 games, standard rule only, 300 benchmark states/game × 2 targets.
- **D3 test (2400):** Connect Four + Dots and Boxes, both rules, 300 states/game/rule × 2 targets.

For the main-text D1/D2 comparison in Figure 4, we additionally evaluate each trained model on the complementary held-out rule slice drawn from the same benchmark-derived pools. Concretely, D1-trained models are evaluated on both the canonical D1 inverse slice and the held-out standard slice (the D2 test file), while D2-trained models are evaluated on both the canonical D2 standard slice and the held-out inverse slice (the D1 test file). The figure groups these paired evaluations by training split to make same-rule versus cross-rule transfer explicit.

C.1.3 SFT hyperparameters

We train Qwen2.5-VL-3B and Qwen2.5-VL-7B with QLoRA SFT (NF4 4-bit quantization, LoRA rank 8, all-module targeting) for one epoch, using a global batch size of 128, learning rate 1×10^{-4}, cosine decay with warmup ratio 0.03, and bf16 precision.

Table 19: VLMBias results for the Animals subset (same metrics and annotations as the VLMBias tables below).

Model	Accuracy ↑ Base	Flip	Alias	Flip+Alias	Bias ↓ Base	Flip	Alias	Flip+Alias
GPT-4.1	9.9	2.0 (<0.01)	14.1 (~0.07)	5.9 (~0.07)	72.5	65.2 (<0.01)	58.6 (<0.01)	76.6 (~0.17)
GPT-5.2	4.6	9.2 (<0.01)	17.0 (<0.01)	39.9 (<0.01)	94.1	89.6 (<0.01)	79.5 (<0.01)	53.3 (<0.01)
Sonnet-4.0	0.0	0.2 (~1.00)	0.0 (~1.00)	3.7 (<0.01)	99.8	97.4 (<0.01)	98.5 (~0.03)	92.3 (<0.01)
Sonnet-4.5	0.0	0.5 (~0.50)	2.0 (<0.01)	17.8 (<0.01)	100.0	97.6 (<0.01)	95.1 (<0.01)	79.5 (<0.01)
Qwen2.5-VL-3B	0.0	0.2 (~1.00)	0.0 (~1.00)	1.5 (~0.05)	96.7	97.3 (~1.00)	96.7 (~1.00)	90.5 (<0.01)
Qwen2.5-VL-7B	0.0	4.0 (<0.01)	10.3 (<0.01)	20.5 (<0.01)	97.6	93.2 (<0.01)	86.3 (<0.01)	75.1 (<0.01)
InternVL3.5-4B	0.0	4.2 (<0.01)	0.0 (~1.00)	15.0 (<0.01)	98.4	86.8 (<0.01)	72.5 (<0.01)	22.0 (<0.01)
InternVL3.5-8B	0.0	1.5 (~0.03)	1.1 (~0.06)	41.2 (<0.01)	99.3	92.9 (<0.01)	90.7 (<0.01)	50.0 (<0.01)
InternVL3.5-14B	0.0	2.6 (<0.01)	1.8 (<0.01)	43.2 (<0.01)	99.1	94.7 (<0.01)	82.4 (<0.01)	49.8 (<0.01)
Qwen3-VL-4B	0.0	2.0 (<0.01)	3.7 (<0.01)	17.6 (<0.01)	99.8	94.0 (<0.01)	93.6 (<0.01)	76.2 (<0.01)
Qwen3-VL-8B	0.5	6.4 (<0.01)	3.7 (<0.01)	17.2 (<0.01)	99.5	93.6 (<0.01)	90.8 (<0.01)	82.4 (<0.01)
Qwen3-VL-32B	2.7	6.4 (<0.01)	10.3 (<0.01)	16.3 (<0.01)	96.9	93.6 (<0.01)	89.4 (<0.01)	83.7 (<0.01)
Molmo2-4B	12.1	18.7 (<0.01)	27.8 (<0.01)	33.0 (<0.01)	83.3	74.0 (<0.01)	69.6 (<0.01)	55.3 (<0.01)
Molmo2-8B	20.9	22.0 (~1.00)	27.1 (<0.01)	37.7 (<0.01)	68.3	75.5 (<0.01)	70.0 (~0.68)	59.0 (<0.01)
Average	3.6	5.7	8.5	22.2	93.2	88.9	83.8	67.5

Table 20: VLMBias results for the Flags subset (same format as Table 19).

Model	Accuracy ↑ Base	Flip	Alias	Flip+Alias	Bias ↓ Base	Flip	Alias	Flip+Alias
GPT-4.1	13.3	9.2 (~0.03)	13.3 (~1.00)	8.3 (~0.03)	83.8	90.0 (<0.01)	85.0 (~0.75)	91.7 (<0.01)
GPT-5.2	22.9	24.6 (~1.00)	26.7 (~0.20)	31.2 (~0.04)	73.8	64.6 (~0.02)	68.8 (~0.03)	52.5 (<0.01)
Sonnet-4.0	25.0	22.9 (~1.00)	30.0 (~0.21)	28.3 (~1.00)	74.2	74.2 (~1.00)	70.0 (~0.39)	67.1 (~0.07)
Sonnet-4.5	24.2	20.0 (~0.98)	27.1 (~0.98)	26.2 (~1.00)	73.8	74.2 (~1.00)	69.6 (~0.21)	66.7 (~0.16)
Qwen2.5-VL-3B	24.2	22.9 (~1.00)	24.2 (~1.00)	30.8 (~0.04)	35.4	33.3 (~1.00)	32.5 (~0.84)	29.2 (~0.24)
Qwen2.5-VL-7B	27.9	24.6 (~0.86)	31.7 (~0.73)	25.8 (~1.00)	45.0	46.7 (~1.00)	41.7 (~1.00)	42.5 (~1.00)
InternVL3.5-4B	23.8	26.2 (~1.00)	29.2 (<0.01)	26.7 (~1.00)	33.8	41.2 (~0.12)	30.0 (~0.60)	35.0 (~1.00)
InternVL3.5-8B	27.1	27.9 (~1.00)	34.6 (~0.06)	37.1 (~0.06)	47.5	43.8 (~0.65)	42.5 (~0.65)	30.0 (<0.01)
InternVL3.5-14B	35.0	33.3 (~1.00)	40.0 (~0.58)	35.0 (~1.00)	33.8	36.2 (~1.00)	30.8 (~1.00)	31.7 (~1.00)
Qwen3-VL-4B	19.2	16.2 (~0.55)	21.7 (~0.58)	16.7 (~0.58)	61.3	65.4 (~0.52)	61.3 (~1.00)	57.5 (~0.84)
Qwen3-VL-8B	17.9	15.8 (~0.53)	23.3 (~0.01)	23.3 (~0.04)	65.8	65.0 (~1.00)	65.0 (~1.00)	62.5 (~0.91)
Qwen3-VL-32B	20.0	22.1 (~0.85)	28.3 (<0.01)	30.0 (<0.01)	73.3	68.3 (~0.12)	67.5 (<0.01)	62.5 (<0.01)
Molmo2-4B	40.8	38.8 (~1.00)	51.7 (<0.01)	51.7 (<0.01)	37.1	34.6 (~0.65)	27.5 (<0.01)	27.5 (<0.01)
Molmo2-8B	41.2	34.6 (~0.37)	46.2 (~0.35)	35.0 (~0.37)	33.8	37.9 (~0.62)	32.5 (~1.00)	40.8 (~0.16)
Average	25.9	24.2	30.6	29.0	55.1	55.4	51.8	49.8

C.1.4 RLVR hyperparameters

RLVR uses GRPO with a binary exact-match reward and KL regularization ($\beta = 10^{-2}$). We use one epoch, rollout multiplicity $n = 5$, global batch size 128, and AdamW in bf16 with learning rate 1×10^{-6} and weight decay 10^{-2}.

C.1.5 Evaluation protocol

We evaluate Base, SFT-merged, and RLVR-merged checkpoints on the canonical D1/D2/D3 test files and report exact-match accuracy from extracted final labels. For the main-text D1/D2 comparison, we additionally evaluate on the complementary held-out rule slices to make same-rule versus cross-rule transfer explicit under matched input formatting.

C.2 Synthetic Leg-Count Transfer Details

C.2.1 Dataset construction

We build a synthetic counting dataset (`synth-legs-train-8192`) with 8192 image–text pairs, balanced 50/50 between bird glyphs and quadruped-style animal glyphs. Bird leg counts are

Table 21: VLMBias results for the Game Boards subset (same format as Table 19).

Model	Accuracy ↑				Bias ↓			
	Base	Flip	Alias	Flip+Alias	Base	Flip	Alias	Flip+Alias
GPT-4.1	0.0	0.0 (~1.00)	0.0 (~1.00)	0.0 (~1.00)	97.6	86.3 (<0.01)	97.6 (~1.00)	88.1 (<0.01)
GPT-5.2	35.7	41.1 (~0.73)	39.9 (~1.00)	35.1 (~1.00)	56.5	49.4 (~0.20)	47.6 (~0.08)	50.0 (~0.20)
Sonnet-4.0	14.9	16.7 (~1.00)	16.1 (~1.00)	24.4 (~0.08)	75.6	74.4 (~1.00)	74.4 (~1.00)	61.3 (<0.01)
Sonnet-4.5	20.8	28.0 (~0.13)	28.6 (~0.06)	30.4 (~0.10)	78.6	69.0 (~0.04)	70.2 (~0.04)	65.5 (~0.02)
Qwen2.5-VL-3B	6.5	8.9 (~0.91)	17.9 (<0.01)	22.6 (<0.01)	50.0	48.2 (~1.00)	26.2 (<0.01)	19.0 (<0.01)
Qwen2.5-VL-7B	12.5	10.7 (~1.00)	13.1 (~1.00)	14.3 (~1.00)	83.3	81.0 (~0.85)	78.6 (~0.61)	75.0 (~0.17)
InternVL3.5-4B	7.1	22.6 (<0.01)	4.8 (~1.00)	13.1 (~0.49)	35.1	20.8 (<0.01)	28.6 (~0.01)	13.1 (<0.01)
InternVL3.5-8B	14.3	15.5 (~1.00)	22.6 (~0.15)	16.7 (~1.00)	32.7	19.6 (<0.01)	14.9 (<0.01)	6.5 (<0.01)
InternVL3.5-14B	18.5	23.2 (~1.00)	20.2 (~1.00)	20.2 (~1.00)	48.8	36.9 (<0.01)	45.2 (~0.57)	32.1 (<0.01)
Qwen3-VL-4B	4.2	4.2 (~1.00)	9.5 (~0.47)	9.5 (~0.47)	82.1	78.6 (~0.57)	57.1 (<0.01)	60.7 (<0.01)
Qwen3-VL-8B	9.5	3.6 (~0.04)	4.8 (~0.54)	6.0 (~0.61)	82.1	83.3 (~1.00)	78.6 (~1.00)	76.2 (~0.31)
Qwen3-VL-32B	8.9	10.7 (~1.00)	7.1 (~1.00)	14.9 (~0.66)	73.8	72.6 (~1.00)	81.0 (~0.32)	62.5 (~0.19)
Molmo2-4B	6.5	7.1 (~1.00)	10.7 (~1.00)	16.7 (~0.06)	79.2	78.6 (~1.00)	58.3 (<0.01)	54.8 (<0.01)
Molmo2-8B	0.0	0.0 (~1.00)	2.4 (~0.75)	2.4 (~0.75)	57.1	56.0 (~1.00)	53.6 (~0.12)	51.2 (~0.01)
Average	11.4	13.7	14.1	16.2	66.6	61.1	58.0	51.1

Table 22: VLMBias results for the Logos subset (same format as Table 19).

Model	Accuracy ↑				Bias ↓			
	Base	Flip	Alias	Flip+Alias	Base	Flip	Alias	Flip+Alias
GPT-4.1	12.1	26.8 (<0.01)	13.3 (~0.92)	22.7 (<0.01)	83.8	63.8 (<0.01)	74.9 (<0.01)	61.6 (<0.01)
GPT-5.2	33.1	26.3 (~0.03)	32.6 (~1.00)	20.8 (<0.01)	63.8	58.2 (~0.03)	63.0 (~1.00)	58.7 (~0.04)
Sonnet-4.0	0.0	1.2 (~0.38)	0.0 (~1.00)	1.2 (~0.38)	89.6	75.8 (<0.01)	88.4 (~1.00)	77.1 (<0.01)
Sonnet-4.5	0.5	4.6 (<0.01)	0.7 (~1.00)	6.0 (<0.01)	98.3	88.4 (<0.01)	89.6 (<0.01)	84.3 (<0.01)
Qwen2.5-VL-3B	7.7	7.2 (~1.00)	6.3 (~1.00)	6.3 (~1.00)	51.0	43.0 (<0.01)	39.6 (<0.01)	33.8 (<0.01)
Qwen2.5-VL-7B	10.1	5.3 (<0.01)	15.5 (<0.01)	14.0 (<0.01)	60.9	56.5 (<0.01)	60.9 (~1.00)	53.1 (<0.01)
InternVL3.5-4B	5.1	6.8 (~1.00)	12.1 (<0.01)	5.8 (~1.00)	64.3	55.8 (<0.01)	57.0 (<0.01)	46.9 (<0.01)
InternVL3.5-8B	2.4	5.8 (~0.03)	8.7 (<0.01)	9.4 (<0.01)	83.1	62.1 (<0.01)	72.9 (<0.01)	48.6 (<0.01)
InternVL3.5-14B	15.2	16.4 (~0.85)	18.4 (~0.02)	19.3 (<0.01)	46.9	42.3 (<0.01)	41.5 (<0.01)	37.7 (<0.01)
Qwen3-VL-4B	14.3	20.5 (<0.01)	14.5 (~1.00)	18.4 (~0.15)	75.4	61.1 (<0.01)	65.2 (<0.01)	57.0 (<0.01)
Qwen3-VL-8B	3.6	9.2 (<0.01)	4.8 (~0.45)	8.2 (<0.01)	83.1	75.4 (<0.01)	72.5 (<0.01)	65.0 (<0.01)
Qwen3-VL-32B	8.5	10.9 (~0.66)	7.7 (~0.75)	10.6 (~0.66)	90.6	83.6 (<0.01)	88.4 (~0.04)	84.5 (<0.01)
Molmo2-4B	41.1	48.8 (<0.01)	43.5 (~0.35)	40.1 (~1.00)	55.3	39.9 (<0.01)	49.3 (<0.01)	36.7 (<0.01)
Molmo2-8B	40.3	43.2 (~1.00)	30.0 (<0.01)	37.2 (~1.00)	53.9	37.2 (<0.01)	55.6 (~0.24)	38.6 (<0.01)
Average	13.9	16.6	14.9	15.7	71.4	60.2	65.6	56.0

sampled from $\{1, 2, 3\}$, and quadruped counts from $\{3, 4, 5\}$. We use procedural variation in pose, shape, color, and background to reduce template-specific overfitting. Representative synthetic glyph categories are shown in Figure 10.

The training instruction is shared across all rows:

> Count the number of legs in this animal glyph image. Answer with only a number in curly brackets, e.g., {4}. Do not add any other text.

C.2.2 Post-training protocols

We post-train Qwen2.5-VL-3B and Qwen2.5-VL-7B with both SFT and RLVR using the same core settings as above. SFT uses QLoRA (NF4, LoRA rank 8, one epoch, global batch 128, learning rate 1×10^{-4}, bf16). RLVR uses GRPO with binary exact-match reward and KL regularization ($\beta = 10^{-2}$), one epoch, rollout $n = 5$, global batch 128, and AdamW with learning rate 1×10^{-6}. We evaluate Base/SFT/RLVR with the same answer extractor and held-out test suites.

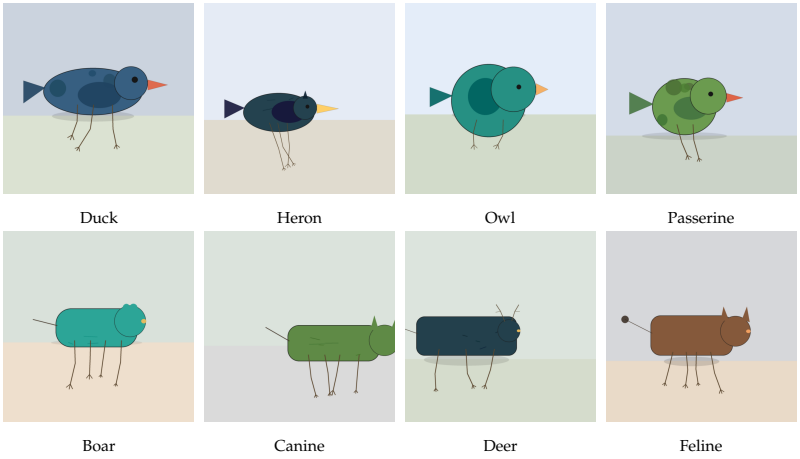

Figure 10: Additional synthetic glyph examples from the procedurally rendered leg-counting training set. Top: bird categories. Bottom: quadruped-style animal categories.

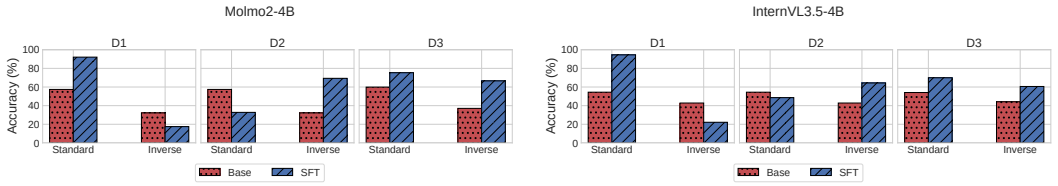

Figure 11: Post-training performance on the three VLM-Fix transfer splits (D1–D3) for additional models with available post-training results: Molmo2-4B (left) and InternVL3.5-4B (right). For D1 and D2, we report both held-out standard-rule and inverse-rule evaluation under the original image and base prompt; for D3, we report held-out standard and inverse evaluation on Connect Four and Dots and Boxes.

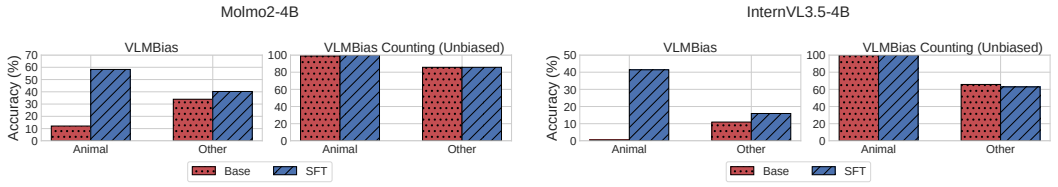

Figure 12: Post-training transfer from the synthetic leg-counting dataset for additional models, comparing Base and SFT for Molmo2-4B (left) and InternVL3.5-4B (right). Each panel reports accuracy on *VLMBias* (Animals vs Other) and *VLMBias Counting (Unbiased)* (Animals vs Other).

C.3 Additional Model Results

C.3.1 VLM-Fix transfer

Additional post-training transfer results for Molmo2-4B and InternVL3.5-4B on D1–D3 are shown in Figure 11.

C.3.2 VLMBias Animals transfer

Additional synthetic-leg-count transfer results for Molmo2-4B and InternVL3.5-4B are shown in Figure 12.

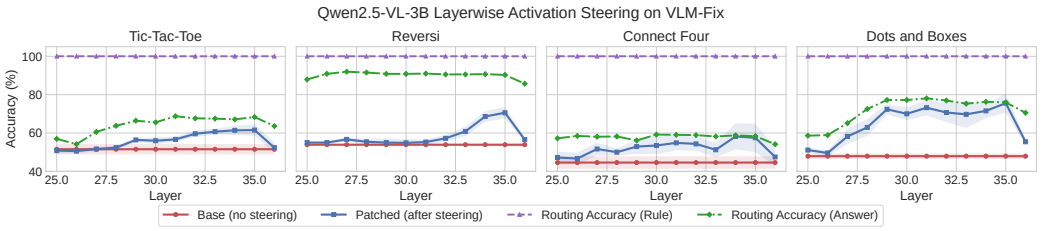

Figure 13: Qwen2.5-VL-3B layerwise activation steering on VLM-Fix across Tic-Tac-Toe, Reversi, Connect Four, and Dots and Boxes.

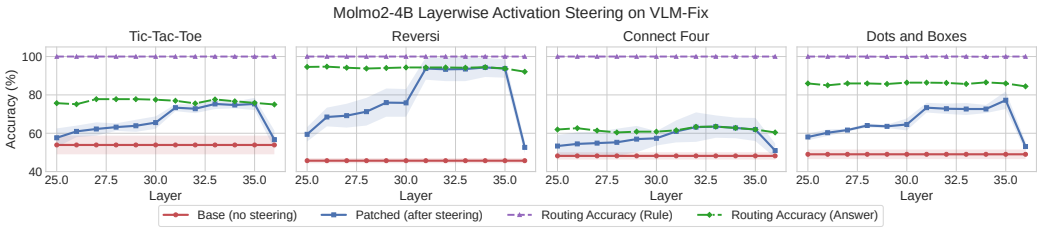

Figure 14: Molmo2-4B layerwise activation steering on VLM-Fix across Tic-Tac-Toe, Reversi, Connect Four, and Dots and Boxes.

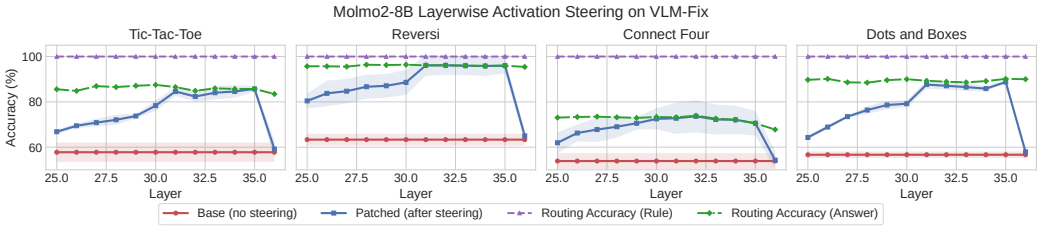

Figure 15: Molmo2-8B layerwise activation steering on VLM-Fix across Tic-Tac-Toe, Reversi, Connect Four, and Dots and Boxes.

Evaluation-coverage note. InternVL3.5-4B and Molmo2-4B use full D1/D2/D3 evaluation files ($n = 2400$ each) in this repository snapshot.

D Activation Steering

D.1 VLM-Fix

For each VLM-Fix game, we use a fixed pool of 100 underlying board states and include all paired variants (standard/inverse rule, winner/loser query, image-first/text-first order). Train/test splitting is done at the *state level* (70/30) and repeated over three random splits to prevent leakage across paired variants.

At each of the final 12 decoder layers, we train a two-stage linear router on the training split (rule classifier, then rule-conditional answer classifier), compute donor centroids by rule–answer bucket (with fallback to all examples in a bucket when donor-correct examples are unavailable), and apply a single query-token patch with intervention scale $\alpha = 1.0$. Reported curves show mean and standard deviation across the three splits.

Additional VLM-Fix steering results for Qwen2.5-VL-3B, Molmo2-4B, Molmo2-8B, and InternVL3.5-4B are shown in Figures 13, 14, 15, and 16.

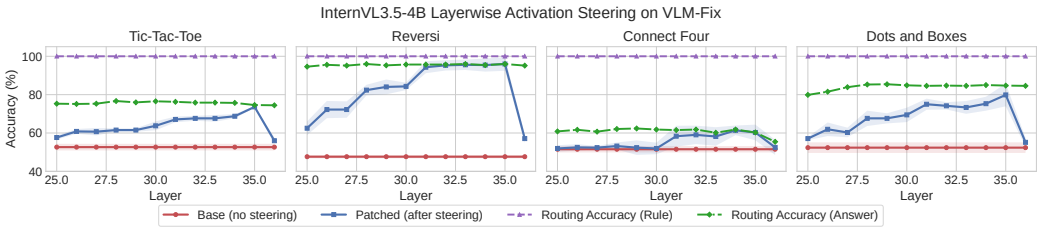

Figure 16: InternVL3.5-4B layerwise activation steering on VLM-Fix across Tic-Tac-Toe, Reversi, Connect Four, and Dots and Boxes.

D.2 VLMBias Animals

Additional details not shown in the main text. The VLMBias steering analysis is run in a controlled 2-leg vs 3-leg regime. We deduplicate by underlying image identity before splitting, then use stratified 70/30 train/test splits across ten random seeds so class balance is preserved.

For each of the final 12 decoder layers, we train a linear router on base-model activations, compute class centroids from SFT-donor activations (with fallback when donor-correct examples are sparse), and patch once at the query token with $\alpha = 1.0$. Layerwise results report base, donor, patched, and routing accuracy as mean and standard deviation over seeds.