IMPROVING GENERALIZATION ROBUSTNESS OF MULTIMODAL RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) makes Multimodal Large Language Models (MLLMs) more accurate, but the gains are brittle: simply rewording the instruction or changing the answer template can erase them, which challenges reliable deployment in high-stakes scenarios like medical VQA. The base model barely reacts to such changes, so the brittleness is created by RL post-training itself. We trace it to two issues of the standard RLVR objective. First, the binary verifier conflates format with content, so the reward signal cannot tell a wrong answer apart from a misformatted one. Second, the training prompts cover only a thin slice of the semantically equivalent prompts met at deployment, so a policy that is optimal on the training template can behave differently on unseen ones. Both failures call for an objective over the full class of semantically equivalent prompts, and we identify two conditions that help achieve it: separating format from semantics in the reward, and enforcing answer invariance across equivalent prompts. We therefore propose **Prompt-Invariant RLVR (PIRL)**, consisting of a dynamic trinary reward and a consistency regularizer driven by an embedding-space adversary. Under template stress testing, PIRL's mean accuracy drops by only 0.1–0.4%, whereas GRPO drops 1.3–2.9. On dynamic evaluation, PIRL also achieves the smallest drop among trained methods.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) drives substantial gains on visual question-answering (VQA) benchmarks for large language models and Multimodal Large Language Models (MLLMs) (Shao et al., 2024; Zhang et al., 2025; Meng et al., 2025; Yue et al., 2025; Zhou et al., 2025), optimizing policies against deterministic signals such as symbolic verification or template matching. The prevailing assumption is that such rule-based reward functions incentivize generalizable reasoning (DeepSeek-AI, 2025; Zhou et al., 2026).

However, it fails under prompt perturbation. Figure 1 shows that an RL-tuned MLLM beats its base model on the standard prompt but degrades sharply when the same question is asked with a different template. We call this the *prompt-robustness gap*. The gap matters most in high-stakes domains such as medical and legal VQA, where a model whose answer flips under rewording is unreliable regardless of its accuracy. The open question is whether RLVR can learn prompt-invariant reasoning.

The effect is amplified in the multimodal setting, where the visual side is high-dimensional and continuous while the format side is discrete and sparse. Optimization takes the easier route: it exploits the textual format and leaves visual understanding underdeveloped. We attribute this to two issues.

Reward Entanglement. The widely used rule-based binary rewards (e.g., regex matching) conflate *semantic correctness* with *format compliance* via a multiplicative verifier $V = V_f \cdot V_c$. A zero reward cannot distinguish a multimodal reasoning failure from format non-compliance. We formalize the resulting loss of semantic reward variation in Lemma B.1.

Prompt Dependency. Standard RLVR optimizes over a fixed distribution of training prompts, while deployment prompts may follow different formats. We formalize this as a quantitative non-identifiability result (Proposition B.6): the training objective can assign the same reward to policies that differ on unseen formats, allowing a test-time gap proportional to the test mass on those formats.

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

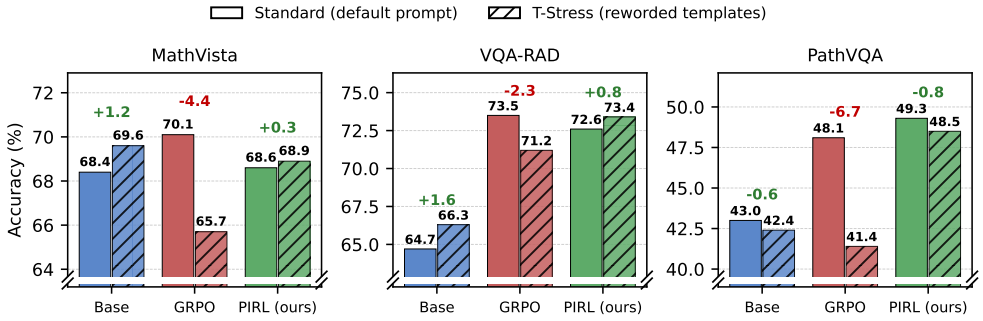


Figure 1: The prompt-robustness gap on Qwen2.5-VL-7B (values from Table 1). GRPO (red) improves over the base model (blue) with the standard template but loses much of the gain when the same questions are asked with reworded, semantically equivalent templates (hatched). **PIRL** (**green**) has lower accuracy reduction under perturbations.

Both failures share a single root: the training objective sees only P_{train} and cannot certify behavior on unobserved but semantics-preserving prompts. We therefore use a worst-case objective over the *semantic equivalence class* $[\![\boldsymbol{x}]\!]$ (Section 2.4). For this objective we identify two design conditions, **(C1) Reward Decomposition** and **(C2) Invariance Control**, with explicit population-level consequences. Our proposed **Prompt-Invariant RLVR (PIRL)** instantiates (C1) with a Dynamic Trinary Reward and uses an embedding-space adversary plus consistency regularization as a practical mechanism for (C2). Unlike prior work that measures prompt-format sensitivity in pretrained LLMs (Sclar et al., 2024; Mizrahi et al., 2024) or uses latent adversarial training for paraphrase robustness in supervised fine-tuning (Fu & Barez, 2025; Zhu et al., 2020; Jiang et al., 2020), we show that RLVR *creates* the gap and close it inside the RL loop.

Our contributions are:

- **Diagnosis.** We identify the RLVR-induced prompt-robustness gap in MLLMs and attribute it to reward entanglement and to objective-level non-identifiability under restricted prompt coverage.
- **Formulation.** We define robustness over semantic equivalence classes and derive two design conditions, *Reward Decomposition* and *Invariance Control*, with population-level guarantees.
- **Method.** PIRL combines a trinary format-aware reward, a semantic score with decoupled advantage normalization, and an embedding-space instruction adversary training framework with a clean-anchored consistency loss.
- **Evidence.** On exam, medical, and document VQA with Qwen2.5-VL-7B and Qwen3-VL-8B, PIRL reduces the mean template-stress (T-Stress) degradation from $-2.9/-1.3$ (GRPO) and $-1.2/-0.9$ (multi-template GRPO on the *same* template pool) to $-0.1/-0.4$. It also attains the highest mean standard, T-Stress, and dynamic-test (D-Test) accuracy on both backbones.

2 Theoretical Analysis of the Robustness Gap

2.1 Preliminary

An MLLM policy $\pi_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ generates a response $\boldsymbol{y}$ conditioned on a prompt $\boldsymbol{x} = (\boldsymbol{v}, \boldsymbol{q}, \mathcal{I})$, where $\boldsymbol{v}$ is a visual input, $\boldsymbol{q}$ is a question, and $\mathcal{I}$ is an instruction specifying the output format via a delimiter such as `\boxed{}`. We define $\mathcal{F}(\mathcal{I})$ as the format constraint induced by the instruction $\mathcal{I}$ (e.g., the regular expression that the verifier matches against the response). Each supervised task instance carries a ground-truth answer $a^* = a^*(\boldsymbol{x})$; semantics-preserving transformations of $\boldsymbol{x}$ inherit the same answer label. When the answer label is clear from context we suppress it in the reward notation. A verifier V returns a binary reward $r \in \{0, 1\}$ indicating whether $\boldsymbol{y}$ satisfies the task. $R(\boldsymbol{y}, \boldsymbol{x}; a^*) \in [R_{\min}, R_{\max}]$ is defined as a more general reward, with $R = V$ in the standard RLVR setting. Let $\mathcal{D}$ denote the distribution over task instances (visual-question pairs together with the fixed training-time instruction and answer labels). The standard RLVR objective (e.g., GRPO (Shao et al., 2024)) is

$$\max_\theta \; \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\, \mathbb{E}_{\boldsymbol{y}\sim\pi_\theta(\cdot|\boldsymbol{x})}\big[V(\boldsymbol{y}, \boldsymbol{x})\big]. \tag{1}$$

GRPO optimizes Eq. (1) via group-relative advantages on G rollouts per prompt; we define $A^{(i)}$ as the group-normalized advantage of rollout i (full form in Appendix F). $\Delta(S)$ denotes the probability measures on S. Multi-template RLVR partially addresses prompt drift by drawing $\boldsymbol{x}$ together with a template transformation $\tau \sim P_{\text{train}}$, where $P_{\text{train}} \in \Delta(\mathcal{T}_{\text{train}})$ is the training-time distribution over the training transformation set $\mathcal{T}_{\text{train}}$ of semantics-preserving instruction templates. Together with $\mathcal{D}$, this induces the joint training distribution. In GRPO and its variants the verifier of Eq. (1) factorizes as a product of a format check and a correctness check,

$$V(\boldsymbol{y}, \boldsymbol{x}) = V_f(\boldsymbol{y}, \mathcal{I})\, V_c(\boldsymbol{y}, \boldsymbol{q}, \boldsymbol{v}), \quad V_f, V_c \in \{0, 1\}, \tag{2}$$

where $V_f = 1$ iff $\boldsymbol{y}$ matches the format constraint $\mathcal{F}(\mathcal{I})$ and $V_c = 1$ iff $\boldsymbol{y}$ encodes the correct answer. Three of the four outcomes $(V_f, V_c) \in \{0, 1\}^2$ map to $V = 0$ and only $(1, 1)$ maps to $V = 1$, so under group-relative advantage normalization (Appendix F) the three zero-reward rollouts share the same scalar advantage; the reward/advantage signal does not distinguish a format failure from a semantic failure or a joint failure. This causes two issues: a loss of within-format advantage signal (Section 2.2, Lemma B.1) and an objective-level non-identifiability (Section 2.3, Proposition B.6).

2.2 Reward Entanglement

On rollouts that fail the format check the binary reward is constant, so the GRPO advantage carries content information only among formatted rollouts. Its within-format share of variance, $\Phi_{\text{wf}} := \mathbb{E}[\text{Var}(V \mid V_f)]/\text{Var}(V)$, equals π_W / p_f, the formatted share of the zero-reward rollouts divided by the format-valid rate (Lemma B.1). It is small on prompts where the policy knows the answer but misses the requested format, and template augmentation and the adversary create such prompts (Proposition B.2): the binary reward discards content signal where robustness is trained. The decomposed reward of Section 2.4, normalized per component (DAN), keeps a within-format share bounded below independently of the format-valid rate (Theorem B.10).

2.3 Prompt Dependency

Let $\mathcal{T}$ be the family of semantics-preserving prompt transformations, and for a task instance $\boldsymbol{x} = (\boldsymbol{v}, \boldsymbol{q}, \mathcal{I})$ let $[\![\boldsymbol{x}]\!] := \{\tau(\boldsymbol{x}) : \tau \in \mathcal{T}\}$ be the *semantic equivalence class*. In the fixed-$\boldsymbol{x}$ analysis below, distributions such as P_{train}, P_{test}, and $\rho_{\boldsymbol{x}}$ are distributions over transformations τ; their push-forward through $\tau \mapsto \tau(\boldsymbol{x})$ is the corresponding prompt distribution on $[\![\boldsymbol{x}]\!]$. For a policy π and transformation distribution ρ, define

$$\mathcal{R}(\pi, \rho; \boldsymbol{x}) := \mathbb{E}_{\tau \sim \rho}\, \mathbb{E}_{\boldsymbol{y} \sim \pi(\cdot \mid \tau(\boldsymbol{x}))} \big[R(\boldsymbol{y}, \tau(\boldsymbol{x})) \big], \tag{3}$$

and $\mathcal{R}(\pi, \tau; \boldsymbol{x})$ for the point mass $\rho = \delta_\tau$. Consider restricted training coverage (P_{train} supports only $\mathcal{T}_{\text{train}} \subsetneq \mathcal{T}$) with a separated unseen format τ_B whose verifier constraint is disjoint from the training formats. Under a τ-dependent verifier, two policies that coincide away from τ_B—one emitting a training format on τ_B, the other the requested format—are both P_{train}-optimal, yet differ by $m_B \Delta_R$ on any test distribution placing mass m_B on τ_B ($\Delta_R := R_{\max} - R_{\min}$). The objective is thus blind at τ_B: its optimum set has P_{test}-diameter at least $m_B \Delta_R$ (Proposition B.6, Appendix B.2).

Restricted template coverage leaves behavior on unseen instructions unconstrained by the training objective (Appendices B.2–B.4); prompt invariance therefore needs (C2).

2.4 Conditions for Prompt Invariance

We take the worst-case prompt distribution on $[\![\boldsymbol{x}]\!]$ as the training target:

$$\max_{\theta}\ \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \Big[\inf_{Q \in \Delta(\mathcal{T})}\ \mathbb{E}_{\tau \sim Q} \mathbb{E}_{\boldsymbol{y} \sim \pi_\theta(\cdot \mid \tau(\boldsymbol{x}))} \big[R(\boldsymbol{y}, \tau(\boldsymbol{x})) \big] \Big]. \tag{4}$$

Eq. (4) is a group-DRO objective (Sagawa et al., 2020) whose groups are generated by prompt transformations (paraphrases, format swaps) rather than given as a partition. $\mathcal{C}(\boldsymbol{x}) \subseteq \mathcal{T}$ is a finite pool of semantics-preserving transformations; Section 3.2 relaxes the infimum to an embedding ball (Lemma B.17).

Two design conditions (formal statements in Appendix A). **(C1) Reward Decomposition** (Definition A.1): a format-aware $R_t \in \{-1, 0, +1\}$ and a bounded, delimiter-free semantic score $R_s \in [0, 1]$

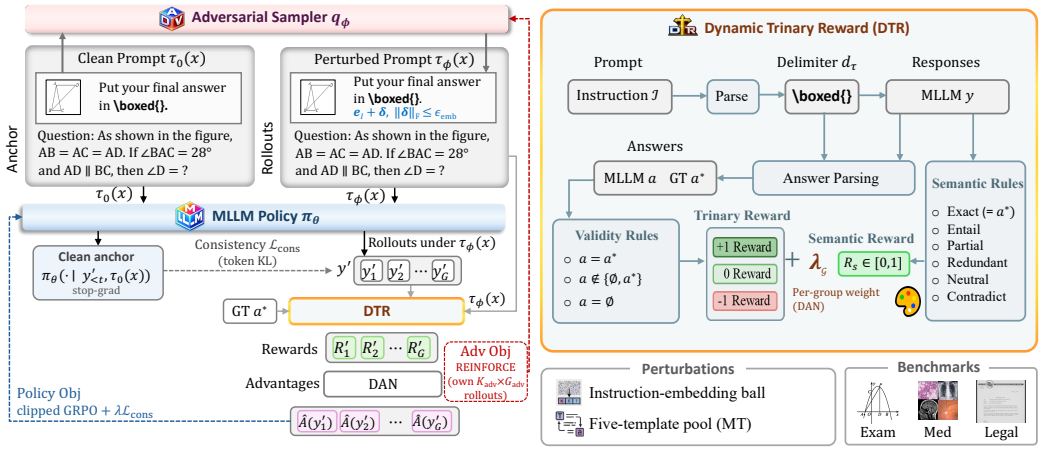


Figure 2: **PIRL.** Left: q_ϕ perturbs only the instruction embedding of a pool prompt $\tau_0(\boldsymbol{x})$ ($\|\boldsymbol{\delta}\|_F \le \epsilon_{\text{emb}}$; text, delimiter, and image unchanged) and ascends $-R_{\text{DTR}}$ by REINFORCE on its own $K_{\text{adv}} \times G_{\text{adv}}$ rollouts (Algorithm 1). DTR scores the G policy rollouts under the final $\tau_\phi(\boldsymbol{x})$, whose DAN advantages drive the clipped GRPO objective. $\tau_0(\boldsymbol{x})$ is only the teacher-forced, stop-gradient anchor of the token-level consistency KL. Right: DTR ($R_t \in \{-1, 0, +1\}$, six-category R_s), combined per group by DAN with weight $\lambda_{\mathcal{G}}$.

with within-bucket dispersion $\mathbb{E}[\text{Var}(R_s \mid V_f, \widetilde{V}_c)] \ge m > 0$ ($\widetilde{V}_c$: extracted answer correct), normalized separately in the advantage. **(C2) Invariance Control** (Definition A.2): for a non-degenerate coverage $\rho_{\boldsymbol{x}} \in \Delta(\mathcal{C}(\boldsymbol{x}))$, $\mathbb{E}_{\tau_1,\tau_2 \sim \rho_{\boldsymbol{x}}}[D(\pi_\theta^{\mathcal{A}_{\tau_1}}(\cdot \mid \tau_1(\boldsymbol{x})) \,\|\, \pi_\theta^{\mathcal{A}_{\tau_2}}(\cdot \mid \tau_2(\boldsymbol{x})))] \le \varepsilon$, where $\pi_\theta^{\mathcal{A}_\tau}$ is the answer marginal under the format-consistent extractor $\mathcal{A}_\tau$ and D an f-divergence with Pinsker constant κ_D. (C2) asks for invariance of the *answer*, not of the surface format.

Consequences. DAN ranks the rollouts of a group by $R_t + \lambda_{\mathcal{G}} R_s$ with a group-calibrated weight $\lambda_{\mathcal{G}}$, so a format failure does not erase correct content (Proposition B.13). PIRL enforces (C2) with two ingredients that act on different template pairs (Proposition B.14). Across the delimiters of the pool, the trinary reward makes fixing one format strictly suboptimal, and near-optimality on the pool implies (C2) in total variation; for same-delimiter embedding neighbors of a pool prompt, the consistency loss bounds the answer-level KL. **Coverage transfer.** For unseen templates graded like the pool, the reward shift from the pool is at most $\min\{\Delta_R \,\text{TV},\ L_\pi W_1\}$, with W_1 the Wasserstein-1 distance of instruction embeddings and L_π the reward's Lipschitz modulus there, so it stays informative for templates disjoint from the pool (Proposition B.16). A bound ω on the adversary's ϵ_{emb}-drop along paths from the pool gives the same bound with $\omega/\epsilon_{\text{emb}}$ in place of L_π, up to an additive ω (Corollary B.18). Delimiter families outside the pool remain the blind spot of Proposition B.6.

3 Method: Prompt-Invariant RLVR

PIRL instantiates (C1) through a Dynamic Trinary Reward (DTR, Section 3.1) whose two components enter GRPO through decoupled advantage normalization (DAN), and enforces (C2) through the trinary reward on the template pool and a consistency regularizer at adversarial embedding neighbors (Section 3.2; Proposition B.14). In the full configuration, we draw a training template $\tau_0 \sim P_{\text{train}}$ from the same five-template pool used by GRPO-MT and perturb the instruction embeddings of $\tau_0(\boldsymbol{x})$; the no-MT ablation fixes τ_0 to a single template. We absorb this template draw into the prompt notation below.

3.1 Dynamic Trinary Reward (DTR)

DTR rewards each rollout with a format-aware trinary verifier $R_t \in \{-1, 0, +1\}$ and a format-independent semantic score $R_s \in [0, 1]$. Let $a_\tau(\boldsymbol{y}) := \mathcal{A}(\boldsymbol{y}; d_\tau)$ be the answer extracted at the delimiter d_τ specified by the current instruction, and let a^* be the ground-truth answer. The trinary

verifier is

$$R_t = \begin{cases} +1, & a_\tau(\boldsymbol{y}) = a^*, \\ 0, & a_\tau(\boldsymbol{y}) \notin \{\emptyset, a^*\}, \\ -1, & a_\tau(\boldsymbol{y}) = \emptyset, \end{cases} \tag{5}$$

which distinguishes correct answer ($+1$), wrong answer in correct format (0), and format failure (-1). With the extraction-based correctness $\widetilde{V}_c := \mathbf{1}\{a_\tau(\boldsymbol{y}) = a^*\}$ (which forces $V_f = 1$), $R_t = V_f + \widetilde{V}_c - 1$ *is* the equal-weight additive format-plus-accuracy reward, up to a shift that group normalization removes; what DTR adds is the within-bucket score R_s and its separate normalization (DAN). R_s (Figure 2, right) collects answer candidates from every recognized answer form and assigns each non-empty response one of six categories (an empty response gets $R_s = 0$): *Exact* (every explicit candidate matches a^*; $R_s = 1$), *Redundant* (a^* appears among conflicting final answers; $R_s = 0.5$), *Partial* (a numeric answer within 9% relative error or a normalized substring match; graded credit), and otherwise *Entail*, *Neutral*, or *Contradict*, decided by a frozen DeBERTa-large-MNLI model (He et al., 2021; Williams et al., 2018) with premise "The answer is ⟨candidate⟩." and hypothesis "The answer is a^*.", where R_s is the entailment probability (Appendix F.2). Numeric candidates are graded by the tolerance ladder first, so the NLI model sees only candidates without rule credit. R_s is computed without the delimiter, so it also scores, and separates, responses whose format fails. Appendix J traces one rollout group through the reward. The total DTR reward is

$$R_{\mathrm{DTR}}(\boldsymbol{y}, \boldsymbol{x}; a^*) = R_t + \alpha_{\mathrm{rew}} R_s, \tag{6}$$

with a fixed weight $\alpha_{\mathrm{rew}} = 0.01 < 1$, so R_{DTR} orders rollouts lexicographically: correct $>$ wrong-but-formatted $>$ format failure, then by R_s. R_{DTR} is the reward the adversary descends (Eq. (9)). For a transformed prompt $\tau(\boldsymbol{x})$ the label is that of the underlying instance, and we write $R_{\mathrm{DTR}}(\boldsymbol{y}, \tau(\boldsymbol{x}))$ for $R_{\mathrm{DTR}}(\boldsymbol{y}, \tau(\boldsymbol{x}); a^*)$.

Decoupled advantage normalization (DAN). Single group normalization of $R_t + \alpha_{\mathrm{rew}} R_s$ lets the small-scale semantic term vanish next to R_t. We z-score R_t and R_s separately within each rollout group $\mathcal{G}$, combine, and re-normalize over all rollouts $\mathcal{B}$ of the refresh batch:

$$\begin{aligned} Z_i &= \mathrm{Norm}_{\mathcal{G}}(R_t^{(i)}) + \alpha_{\mathrm{adv}} \, \mathrm{Norm}_{\mathcal{G}}(R_s^{(i)}), \\ \hat{A}_i &= \mathrm{clip}(\mathrm{Norm}_{\mathcal{B}}(Z_i), \ -5, 5)\,, \end{aligned} \tag{7}$$

where $\mathrm{Norm}_S(z) = (z - \bar{z})/(\sigma_z + \epsilon_{\mathrm{norm}})$ uses the numerical-stability $\epsilon_{\mathrm{norm}} > 0$ of standard GRPO, and $\alpha_{\mathrm{adv}} > 0$ ($= 1$ in all experiments) is the advantage-mixing coefficient; α_{rew} does not enter Eq. (7). DAN is group normalization of $R_t + \lambda_{\mathcal{G}} R_s$ with $\lambda_{\mathcal{G}} := \alpha_{\mathrm{adv}} \sigma_t / \sigma_s$ (the group standard deviations of R_t and R_s), up to a group-level scale; $\lambda_{\mathcal{G}}$ is the dynamic weight of DTR (Proposition B.13). At equal R_s the trinary order is strict, and a format failure overtakes a formatted wrong answer only when its semantic lead exceeds $1/\lambda_{\mathcal{G}}$, so a format failure never erases correct content. Under (C1), the within-format share of the DAN signal is bounded below independently of the format-valid rate p_f and of α_{rew} (Theorem B.10).

3.2 Adversarial sampler and consistency

We approximate the worst case of Eq. (4) with an adversarial sampler q_ϕ that perturbs the prompt in the embedding space (q_ϕ is the point mass at the final perturbation $\phi = \boldsymbol{\delta}$). Given an instance $\boldsymbol{x} = (\boldsymbol{v}, \boldsymbol{q}, \mathcal{I})$, let $\boldsymbol{e}_{\mathcal{I}}(\boldsymbol{x})$ denote the instruction-token embeddings of $\tau_0(\boldsymbol{x})$. The adversary adds a bounded perturbation to this block, $\tilde{\boldsymbol{e}}_{\mathcal{I}}(\boldsymbol{x}) = \boldsymbol{e}_{\mathcal{I}}(\boldsymbol{x}) + \boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_F \le \epsilon_{\mathrm{emb}}$, and we write $\tau_\phi(\boldsymbol{x})$ for the perturbed prompt. The perturbation searches the ϵ_{emb}-ball around the instruction representation for a low expected DTR reward, a local model of reworded instructions (nearest-token decodings of the perturbed instructions keep the instruction; Appendix L); the image, question, answer target, and the delimiter read by the extractor are unchanged ($d_{\tau_\phi} = d_{\tau_0}$). Discrete template transformations use their own requested delimiter d_τ.

Consistency regularization. The adversary finds low-reward perturbations; the consistency term pulls the policy under them back to the clean prompt. On the stored policy rollouts $\boldsymbol{y} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid \tau_\phi(\boldsymbol{x}))$ (teacher-forced), let $p_t^0 = \pi_\theta(\cdot \mid \boldsymbol{y}_{<t}, \tau_0(\boldsymbol{x}))$ and $p_t^\phi = \pi_\theta(\cdot \mid \boldsymbol{y}_{<t}, \tau_\phi(\boldsymbol{x}))$. We use

$$\mathcal{L}_{\mathrm{cons}}(\theta; \boldsymbol{x}, \phi) = \mathbb{E}_{\boldsymbol{y}}\Big[\frac{1}{|\boldsymbol{y}|} \sum_{t=1}^{|\boldsymbol{y}|} D_{\mathrm{KL}}\Big(p_t^\phi \,\Big\|\, \mathrm{sg}(p_t^0)\Big)\Big], \tag{8}$$

Table 1: Accuracy (%) under Standard (default prompt), T-Stress (reworded semantics-preserving templates), and D-Test (operator-based input mutations) evaluation (Appendix E); means over 3 seeds. Parentheses give the gap to Standard (ΔT, ΔD; negative = degradation, red if ≤ -1.5); Mean-column gaps are computed from unrounded means. **Bold**: best accuracy per backbone; **green** (overrides red): smallest degradation among trained methods. The last row gives seed standard deviations of the gaps, pooled over methods per benchmark. GRPO-MT and PIRL share the same five-template training pool.

			Exam VQA			Medical VQA			Legal	
			ID		OOD	ID		OOD	OOD	
Backbone	**Method**		MMK12	MathVista	Olym-Phys	RAD	Path	GMAI	DocVQA	**Mean**
Qwen2.5 VL-7B	Base	Std.	53.1	68.4	8.5	64.7	43.0	**44.8**	66.8	49.9
		T-Stress	54.1 (+1.0)	**69.6** (+1.2)	7.2 (-1.3)	66.3 (+1.6)	42.4 (-0.6)	**42.4** (-2.4)	67.2 (+0.4)	49.9 (0.0)
		D-Test	53.5 (+0.4)	21.4 (-47.0)	4.4 (-4.1)	**66.1** (+1.4)	40.4 (-2.6)	37.3 (-7.5)	46.1 (-20.7)	38.5 (-11.4)
	GRPO	Std.	54.9	**70.1**	7.5	**73.5**	48.1	42.1	69.1	52.2
		T-Stress	54.1 (-0.8)	65.7 (-4.4)	5.9 (-1.6)	71.2 (-2.3)	41.4 (-6.7)	39.7 (-2.4)	66.7 (-2.4)	49.2 (-2.9)
		D-Test	51.3 (-3.6)	**21.6** (-48.5)	4.4 (**-3.1**)	61.4 (-12.1)	42.3 (-5.8)	34.5 (-7.6)	46.3 (-22.8)	37.4 (-14.8)
	GRPO-MT	Std.	54.9	70.0	7.8	71.0	45.8	42.2	68.9	51.5
		T-Stress	54.0 (-0.9)	69.3 (-0.7)	7.6 (-0.2)	69.3 (-1.7)	43.2 (-2.6)	40.9 (**-1.3**)	68.0 (-0.9)	50.3 (-1.2)
		D-Test	**54.2** (**-0.7**)	21.3 (-48.7)	4.3 (-3.5)	61.9 (-9.1)	42.1 (-3.7)	34.5 (-7.7)	47.7 (-21.2)	38.0 (-13.5)
	PIRL	Std.	**55.3**	68.6	**9.6**	72.6	**49.3**	**44.8**	**69.5**	**52.8**
		T-Stress	**55.7** (**+0.4**)	68.9 (**+0.3**)	**10.7** (**+1.1**)	**73.4** (**+0.8**)	**48.5** (**-0.8**)	42.1 (-2.7)	**69.7** (**+0.2**)	**52.7** (**-0.1**)
		D-Test	**54.2** (-1.1)	21.4 (**-47.2**)	**5.5** (-4.1)	65.3 (**-7.3**)	**46.2** (**-3.1**)	**37.6** (**-7.2**)	**49.0** (**-20.5**)	**39.9** (**-12.9**)
Qwen3 VL-8B	Base	Std.	60.3	68.9	25.3	68.1	49.1	46.0	70.3	55.4
		T-Stress	59.9 (-0.4)	69.6 (+0.7)	25.7 (+0.4)	68.7 (+0.6)	49.2 (+0.1)	**46.2** (+0.2)	71.0 (+0.7)	55.8 (+0.3)
		D-Test	**60.2** (-0.1)	21.8 (-47.1)	20.7 (-4.6)	69.0 (+0.9)	46.7 (-2.4)	38.6 (-7.4)	49.8 (-20.5)	43.8 (-11.6)
	GRPO	Std.	60.5	69.1	27.1	77.2	49.9	46.2	72.7	57.5
		T-Stress	60.0 (**-0.5**)	67.5 (-1.6)	26.0 (-1.1)	76.1 (-1.1)	48.1 (-1.8)	44.8 (**-1.4**)	71.4 (-1.3)	56.3 (-1.3)
		D-Test	56.5 (-4.0)	20.8 (-48.3)	**23.7** (-3.4)	64.9 (-12.3)	44.1 (-5.8)	38.9 (-7.3)	50.7 (-22.0)	42.8 (-14.7)
	GRPO-MT	Std.	58.7	**72.8**	21.5	77.4	49.8	46.2	72.6	57.0
		T-Stress	57.8 (-0.9)	**73.1** (+0.3)	21.5 (**0.0**)	76.1 (-1.3)	47.0 (-2.8)	44.8 (**-1.4**)	72.2 (-0.4)	56.1 (-0.9)
		D-Test	58.4 (**-0.3**)	23.8 (-49.0)	18.2 (**-3.3**)	68.2 (-9.2)	46.2 (-3.6)	38.6 (-7.6)	51.8 (-20.8)	43.6 (-13.4)
	PIRL	Std.	**60.9**	72.0	**27.4**	**78.8**	**50.0**	**46.3**	**73.1**	**58.4**
		T-Stress	**60.1** (-0.8)	72.4 (**+0.4**)	**26.3** (-1.1)	**79.8** (**+1.0**)	**49.4** (**-0.6**)	44.0 (-2.3)	**73.6** (**+0.5**)	**57.9** (**-0.4**)
		D-Test	60.1 (-0.8)	**24.6** (**-47.4**)	22.8 (-4.6)	**71.2** (**-7.6**)	**49.0** (**-1.0**)	**39.5** (**-6.8**)	**52.8** (**-20.3**)	**45.7** (**-12.6**)
	s.d. of ΔT / ΔD		$\pm$0.4 / 0.5	$\pm$0.4 / 0.8	$\pm$0.7 / 0.7	$\pm$0.5 / 0.6	$\pm$0.6 / 0.6	$\pm$0.6 / 0.6	$\pm$0.5 / 0.5	

Table 2: **Ablation** on Qwen2.5-VL-7B of multi-template training (MT), dynamic trinary reward (DTR, with DAN), and the embedding adversary with consistency loss (ADV). Row 1 is GRPO; row 2 is single-template PIRL; row 4 is full PIRL (the configuration of Table 1). Mean$\pm$s.d. over 3 seeds (ΔT within $\pm$1.5 pp inconclusive). Exam VQA Avg = mean of MMK12/MathVista/Olym-Phys. Bold: best per column.

Components			Exam VQA Avg.					Olym-Phys				
MT	**ADV**	**DTR**	**Std.**	**T-Stress**	**D-Test**	**ΔT**	**ΔD**	**Std.**	**T-Stress**	**D-Test**	**ΔT**	**ΔD**
×	×	×	$44.2_{\pm 0.4}$	$41.9_{\pm 0.5}$	$25.8_{\pm 0.6}$	$-2.3_{\pm 0.5}$	$-18.4_{\pm 0.7}$	$7.5_{\pm 0.5}$	$5.9_{\pm 0.5}$	$4.4_{\pm 0.5}$	$-1.6_{\pm 0.5}$	$\mathbf{-3.1}_{\pm 0.5}$
×	✓	✓	$44.3_{\pm 0.4}$	$43.4_{\pm 0.5}$	$26.0_{\pm 0.7}$	$-0.9_{\pm 0.5}$	$-18.3_{\pm 0.7}$	$\mathbf{10.3}_{\pm 0.5}$	$10.4_{\pm 0.5}$	$4.4_{\pm 0.5}$	$+0.1_{\pm 0.5}$	$-5.9_{\pm 0.5}$
✓	×	✓	$44.4_{\pm 0.5}$	$44.4_{\pm 0.5}$	$26.8_{\pm 0.7}$	$0.0_{\pm 0.5}$	$-17.6_{\pm 0.7}$	$10.1_{\pm 0.5}$	$9.8_{\pm 0.5}$	$4.2_{\pm 0.5}$	$-0.3_{\pm 0.5}$	$-5.9_{\pm 0.5}$
✓	✓	✓	$\mathbf{44.5}_{\pm 0.4}$	$\mathbf{45.1}_{\pm 0.5}$	$\mathbf{27.0}_{\pm 0.7}$	$\mathbf{+0.6}_{\pm 0.5}$	$\mathbf{-17.5}_{\pm 0.7}$	$9.6_{\pm 0.5}$	$\mathbf{10.7}_{\pm 0.5}$	$\mathbf{5.5}_{\pm 0.5}$	$\mathbf{+1.1}_{\pm 0.5}$	$-4.1_{\pm 0.5}$

where $|\boldsymbol{y}|$ is the response length and $\mathrm{sg}(\cdot)$ stops gradients through the clean anchor. For the same-delimiter pair $(\tau_0(\boldsymbol{x}), \tau_\phi(\boldsymbol{x}))$, the per-token sum, in expectation under $\pi_\theta(\cdot \mid \tau_\phi(\boldsymbol{x}))$, upper-bounds the answer-level KL (Proposition B.14).

3.3 Overall Objective

Let $\mathcal{L}_{\mathrm{G}}^{\mathrm{DAN}}(\pi_\theta, q; R_t, R_s)$ denote the clipped GRPO loss on rollouts of prompts drawn from q, with DAN advantages (Eq. (7)); the optional reference-KL penalty (Appendix F) has coefficient zero in

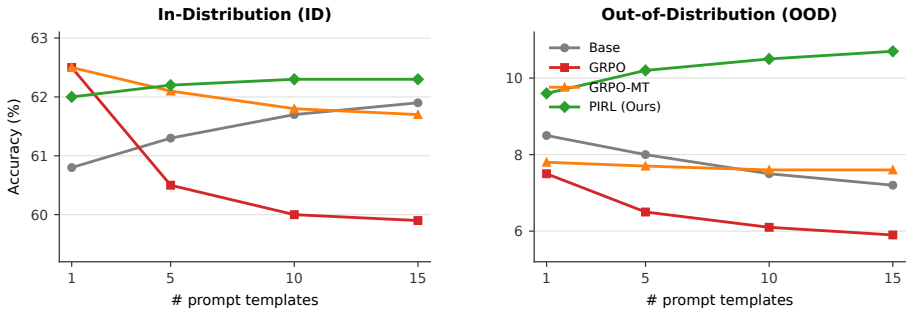


Figure 3: **Effect of prompt-template count** on Exam VQA accuracy (Qwen2.5-VL-7B; ID = mean of MMK12 and MathVista, OOD = Olym-Phys). $N{=}1$ and $N{=}15$ are the Standard and T-Stress entries of Table 1. PIRL has the flattest ID curve and is highest on OOD at every N.

our experiments. The adversarial objective is

$$\max_{\phi} \; \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\, \mathbb{E}_{\tau\sim q_\phi(\cdot|\boldsymbol{x})} \mathbb{E}_{\boldsymbol{y}\sim\pi_\theta(\cdot|\tau(\boldsymbol{x}))} \big[-R_{\mathrm{DTR}}(\boldsymbol{y}, \tau(\boldsymbol{x}))\big], \tag{9}$$

with q_ϕ supported on the ϵ_{emb} instruction-embedding ball. Let $\widehat{\phi}_{K_{\mathrm{adv}}}(\theta)$ be the parameter returned by K_{adv} projected score-function ascent steps on Eq. (9) (Appendix C), and $\bar{\phi}_{\boldsymbol{x}} = \mathrm{sg}(\widehat{\phi}_{K_{\mathrm{adv}}}(\theta))$ the frozen perturbation for prompt $\boldsymbol{x}$. The policy objective is

$$\min_{\theta} \; \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Big[\mathcal{L}_{\mathrm{G}}^{\mathrm{DAN}}\big(\pi_\theta, q_{\bar{\phi}_{\boldsymbol{x}}}(\cdot \mid \boldsymbol{x}); R_t, R_s\big) + \lambda\, \mathcal{L}_{\mathrm{cons}}(\theta; \boldsymbol{x}, \bar{\phi}_{\boldsymbol{x}})\Big]. \tag{10}$$

The GRPO term uses the G rollouts sampled under $\tau_{\bar{\phi}_{\boldsymbol{x}}}(\boldsymbol{x})$, the clean prompt $\tau_0(\boldsymbol{x})$ enters only as the anchor of $\mathcal{L}_{\mathrm{cons}}$, $\lambda > 0$ weights the invariance penalty, and no gradient flows through the inner ascent. At $\boldsymbol{\delta} = 0$, $\mathcal{L}_{\mathrm{cons}} = 0$ and PIRL reduces to GRPO-MT with DTR and DAN. Algorithm 2 gives the full loop.

4 Experiments

4.1 Setup

We evaluate on three blocks: **Exam VQA** (train/ID eval MMK12 (Meng et al., 2025), ID eval MathVista (Lu et al., 2024), OOD eval Olym-Phys (He et al., 2024)), **Medical VQA** (train Lingshu (LASA Team et al., 2025); ID eval VQA-RAD (Lau et al., 2018) and PathVQA (He et al., 2020); OOD eval GMAI-MMBench (Chen et al., 2024a)), and **Legal VQA** (document VQA; train DUDE (Van Landeghem et al., 2023); OOD eval DocVQA (Mathew et al., 2021)). Base models: Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B (Bai et al., 2025a). Baselines: GRPO (Shao et al., 2024) and GRPO-MT (GRPO on the same five-template pool as PIRL). Protocols (Appendix E): **Standard** (each benchmark's default prompt; the bare question on MMK12), **T-Stress** (one of 15 semantics-preserving templates drawn per question), and **D-Test** (operator-based input mutations sampled per question, following Yang et al. (2025); Appendix D). ΔT and ΔD denote the Standard→T-Stress and Standard→D-Test gaps (negative = degradation), respectively. All reported accuracies are means over 3 run seeds.

4.2 Template Stress Test

Table 1 reports Standard and T-Stress accuracy, and Figure 4 (Appendix I) summarizes the mean gap per method. The base model is stable on average (mean ΔT 0.0 and $+0.3$), although single benchmarks move by up to 2.4 points (GMAI on Qwen2.5-VL-7B). On **general Exam VQA**, standard GRPO gains Standard accuracy on the ID sets but loses sharply under template shift: on Qwen2.5-VL-7B MathVista drops by 4.4 and Olym-Phys by 1.6, whereas PIRL is flat-to-positive

(ΔT $= +0.3$ on MathVista, $+1.1$ on Olym-Phys) and matches or beats GRPO-MT trained on the same five-template pool (MathVista $+0.3$ vs. -0.7). Without (C2), the training objective does not constrain the policy on templates outside its training support (Proposition B.6).

High-stakes VQA. The Medical and Legal blocks of Table 1 target deployment-oriented domains, where an answer that flips under a reworded instruction is of little use. On *medical VQA*, GRPO and GRPO-MT degrade under template stress (Qwen2.5-VL-7B VQA-RAD -2.3/-1.7, PathVQA -6.7/-2.6) while PIRL stays near flat (VQA-RAD $+0.8$, PathVQA -0.8); the exception is GMAI, where PIRL has the largest T-Stress drop among trained methods on both backbones (-2.7/-2.3 vs. GRPO-MT -1.3/-1.4). On *Legal VQA* (we use the legal-related instances in DocVQA), PIRL retains accuracy on both base models (ΔT $= +0.2$ on Qwen2.5-VL-7B, $+0.5$ on Qwen3-VL-8B) whereas GRPO degrades (-2.4 and -1.3) and GRPO-MT lies in between (-0.9 and -0.4).

Answer invariance. Answering every question under all 15 templates (Appendix L; Qwen2.5-VL-7B), PIRL flips the answer on $7.4/6.1/8.2\%$ of MathVista/VQA-RAD/PathVQA questions, against $12.1/11.4/14.8\%$ for GRPO-MT and $22.8/18.9/25.4\%$ for GRPO, and has the highest worst-template accuracy on all three; on GMAI its flip rate is close to GRPO-MT's (13.5 vs. 12.6%).

4.3 Effect of Template Count

Figure 3 sweeps the number N of evaluation templates on Exam VQA: as N grows from 1 (Standard) to 15 (T-Stress), GRPO's ID accuracy falls from 62.5 to 59.9, while PIRL has the flattest ID curve and the highest OOD accuracy at every N.

4.4 Dynamic Evaluation

Under operator-based input mutations (Table 1; Appendix D), on **general (Exam) VQA** every method collapses on MathVista (~ -47 to -49). Excluding MathVista, PIRL has the smallest mean ΔD among trained methods (-7.2 vs. GRPO-MT -7.7 and GRPO -9.2 on Qwen2.5-VL-7B; -6.9 vs. -7.5 and -9.1 on Qwen3-VL-8B).

High-stakes VQA. On *medical VQA*, PIRL is consistently the least-degraded trained method (Qwen2.5-VL-7B: VQA-RAD ΔD $= -7.3$ vs. GRPO-MT -9.1 and GRPO -12.1; PathVQA -3.1 vs. -3.7/-5.8; GMAI -7.2 vs. -7.7/-7.6). On *Legal VQA* (DocVQA), D-Test lowers every method by $\sim$20 points, and PIRL degrades least (ΔD $= -20.5$ and -20.3 vs. GRPO-MT -21.2 and -20.8 and GRPO -22.8 and -22.0).

4.5 Ablation

Table 2 removes multi-template training (MT) or the embedding-space adversary (ADV) from full PIRL on Exam VQA Avg. and Olym-Phys; DTR is on in rows 2–4, and the reward-side removals (binary verifier, DTR without DAN) are in Appendix K. Full PIRL has the strongest T-Stress result (ΔT $= +0.6$ vs. GRPO -2.3). The aggregate D-Test, dominated by MathVista, does not separate the configurations (ΔD from -18.4 to -17.5); on T-Stress, full PIRL ($+0.6$) and MT+DTR (0.0) both clearly beat GRPO (-2.3). On Olym-Phys, PIRL improves ΔT ($+1.1$ vs. -1.6) and D-Test accuracy (5.5 vs. 4.4) over GRPO; its ΔD is larger (-4.1 vs. -3.1) because its standard accuracy is higher. Matching compute does not close the gap: GRPO-MT with PIRL's 32 sampled sequences per prompt (G=32) or with its 8 backward passes (twice the iterations) gains 0.4–0.5% Standard accuracy but stays at ΔT $= -0.6$, against PIRL's $+0.6$. Within the ϵ_{emb}-ball the adversary lowers the mean R_{DTR} of MMK12 training rollouts from $+0.64$ to -0.18, while nearest-token decodings of the perturbed instructions keep the instruction in 94 of 100 inspected cases (Appendix L).

4.6 Discussion

Against the matched GRPO-MT baseline and the $\sim$1.5% inconclusive threshold, PIRL's ΔT advantage clears the threshold on VQA-RAD ($+2.5$/$+2.3$ on Qwen2.5-VL-7B/Qwen3-VL-8B) and PathVQA ($+1.8$/$+2.2$). The answer-level separation is clearer: on MathVista, VQA-RAD and PathVQA PIRL's flip rate is 4.7–6.6% lower and its worst-template accuracy 1.2–5.3% higher than

GRPO-MT's. Against GRPO, the advantage clears the threshold on five of seven benchmarks on Qwen2.5-VL-7B ($+2.6$ to $+5.9$) and on MathVista, VQA-RAD, and DocVQA on Qwen3-VL-8B ($+1.8$ to $+2.1$). On Pass@8 (Table 9) PIRL matches the base model and exceeds GRPO by $+0.6$ to $+2.5$, consistent with RLVR reweighting probability mass rather than enlarging the reasoning frontier (Yue et al., 2025; Liu et al., 2025b).

5 Related Work

RL for reasoning and its failure modes. RLVR is widely used to improve reasoning in LLMs and MLLMs (DeepSeek-AI, 2025; Meng et al., 2025; Zhang et al., 2025; Zhou et al., 2025). Outcome-based optimization also exposes Goodhart-style failures, where policies exploit reward or evaluator artifacts (Teney et al., 2020; Yue et al., 2025), including length bias and reward hacking (Hu et al., 2025; Wang et al., 2025a; Liu et al., 2025b). We add a mechanism specific to verifiable rewards: format–content entanglement and restricted prompt coverage.

Prompt sensitivity. Meaning-preserving changes to prompt formatting cause large accuracy swings in pretrained LLMs (Sclar et al., 2024), which motivates multi-prompt evaluation (Mizrahi et al., 2024). Adversarial prompt benchmarks (Zhu et al., 2023) and option-order sensitivity (Zheng et al., 2024) show the same brittleness, and scaling reasoning degrades instruction following in large reasoning models (Fu et al., 2025). These works *measure* sensitivity at inference time; we show that RLVR can *create* it in MLLMs whose base models lack it, and enforce prompt-equivalence invariance inside RLVR.

Embedding-space adversarial training and consistency. Adversarial perturbations of token embeddings with smoothness or consistency regularization are established tools for robust NLU fine-tuning (Miyato et al., 2017; Zhu et al., 2020; Jiang et al., 2020). Latent adversarial training perturbs hidden activations (Casper et al., 2024; Sheshadri et al., 2024). Closest to us, Fu & Barez (2025) perturb latent prompt representations with an inner/outer loop to improve paraphrase robustness in supervised fine-tuning. In PIRL, the adversary instead ascends the *negative verifiable RL reward* through a score-function estimator and is paired with a reward decomposition that addresses the verifier-side cause of the gap. Other RL-based robustness work searches for adversarial prompts or red-teams models (Chen et al., 2024b; Lin et al., 2025; Pan et al., 2025), targets perception-side perturbations (Li et al., 2025), or augments images in RL rollouts (Liu et al., 2025a); PIRL instead enforces invariance across equivalent prompts. R1-style recipes add a format reward to an accuracy reward (DeepSeek-AI, 2025; Meng et al., 2025); DTR's trinary term equals this sum up to a constant, and DTR adds a within-format-bucket semantic score with separate normalization (Section 3.1). Mao et al. (2025) split RLHF rewards into prompt-free and prompt-related parts; evolving reward models and prompts (Kim et al., 2025; Ye et al., 2025) and adversarial preference learning (Wang et al., 2025b) are complementary to PIRL.

6 Conclusion

RLVR can create a prompt-robustness gap absent from the base model: accuracy gained on the standard prompt is lost under semantically equivalent rewordings. We trace it to a binary verifier that entangles format with semantics and to restricted prompt coverage that cannot distinguish prompt-invariant from template-specific policies. PIRL meets the resulting design conditions, (C1) and (C2), with a dynamic trinary reward and an embedding-space adversary with consistency regularization. Across exam, medical, and document VQA on two backbones, PIRL returns the mean template-stress gap to the base model's level while improving standard accuracy, and on three of four benchmarks it flips fewer answers across templates than multi-template GRPO.

Limitations. (i) *Scale and model family.* We study 7–8B Qwen-family backbones, and the gap is smaller on the stronger backbone. (ii) *Theory.* The variance-floor and ordering results hold exactly for every finite rollout group; the identifiability and transfer results are population-level.

AI USE STATEMENT

Generative AI tools assisted with editing the text. The authors verified all AI-assisted content and take full responsibility for the paper, including its text, claims, and artifacts.

ETHICS STATEMENT

This work uses publicly available benchmarks under their research licenses (Appendix G.2). It involves no new human-subject data. Medical and legal VQA are studied as stress tests of robustness. Robustness is not correctness, and PIRL-trained models are not intended for autonomous clinical or legal decision-making. The embedding-space adversary and the evaluation operators are dual-use in principle; we discuss the risks and mitigations in Appendix H.

REPRODUCIBILITY STATEMENT

Full statements and proofs of all theoretical results are in Appendices A–B. The complete training loop is given as Algorithm 2, with adversary details in Appendix C. Datasets and splits, licenses, compute, all hyperparameters, and software versions are listed in Appendix G. The evaluation protocol is specified in Appendix E, and the dynamic-evaluation operators in Appendix D. Anonymized code and evaluation scripts are provided in the supplementary material.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 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, Mingkun 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. *arXiv preprint arXiv:2502.13923*, 2025b.

Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. Defending against unforeseen failure modes with latent adversarial training. *arXiv preprint arXiv:2403.05030*, 2024.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J. Seibel, Yu Qiao, and Junjun He. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024a.

Xuan Chen, Yuzhou Nie, Wenbo Guo, and Xiangyu Zhang. When llm meets drl: Advancing jailbreaking efficiency via drl-guided search. *Advances in Neural Information Processing Systems*, 37:26814–26845, 2024b.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tingchen Fu and Fazl Barez. Same question, different words: A latent adversarial framework for prompt robustness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31305–31319, 2025.

Tingchen Fu, Jiawei Gu, Yafu Li, Xiaoye Qu, and Yu Cheng. Scaling reasoning, losing control: Evaluating instruction following in large reasoning models. *arXiv preprint arXiv:2505.14810*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2177–2190, 2020.

Zae Myung Kim, Chanwoo Park, Vipul Raheja, Suin Kim, and Dongyeop Kang. Toward evaluative thinking: Meta policy optimization with evolving reward models. *arXiv preprint arXiv:2504.20157*, 2025.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251, 2018. doi: 10.1038/sdata.2018.251.

Yuting Li, Lai Wei, Kaipeng Zheng, Jingyuan Huang, Linghe Kong, Lichao Sun, and Weiran Huang. Vision matters: Simple visual perturbations can boost multimodal math reasoning. *arXiv preprint arXiv:2506.09736*, 2025.

Jiayi Lin, Liangcai Su, Junzhe Li, and Chenxiong Qian. R1-fuzz: Specializing language models for textual fuzzing via reinforcement learning. *arXiv preprint arXiv:2509.20384*, 2025.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. *arXiv preprint arXiv:2601.05242*, 2026.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. NoisyRollout: Reinforcing visual reasoning with data augmentation. In *Advances in Neural Information Processing Systems*, 2025a. arXiv:2504.13055.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024.

Liyuan Mao, Haoran Xu, Amy Zhang, Weinan Zhang, and Chenjia Bai. Information-theoretic reward decomposition for generalizable rlhf. *arXiv preprint arXiv:2504.06020*, 2025.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2199–2208, 2021. doi: 10.1109/WACV48630.2021.00225.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*, 2017.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Jiazhen Pan, Bailiang Jian, Paul Hager, Yundi Zhang, Che Liu, Friedrike Jungmann, Hongwei Bran Li, Chenyu You, Junde Wu, Jiayuan Zhu, Fenglin Liu, Yuyuan Liu, Niklas Bubeck, Christian Wachinger, Chen Chen, Zhenyu Gong, Cheng Ouyang, Georgios Kaissis, Benedikt Wiestler, and Daniel Rueckert. Beyond benchmarks: Dynamic, automatic and systematic red-teaming agents for trustworthy medical language models. *arXiv preprint arXiv:2508.00923*, 2025.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent adversarial training improves robustness to persistent harmful behaviors in LLMs. *arXiv preprint arXiv:2407.15549*, 2024.

Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of Goodhart's law. In *Advances in Neural Information Processing Systems*, volume 33, pp. 407–417, 2020.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew B. Blaschko, Sien Moens, and Tomasz Stanisławek. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19528–19540, 2023.

Xiangqi Wang, Yue Huang, Yujun Zhou, Xiaonan Luo, Kehan Guo, and Xiangliang Zhang. Causally-enhanced reinforcement policy optimization. *arXiv preprint arXiv:2509.23095*, 2025a.

Yuanfu Wang, Pengyu Wang, Chenyang Xi, Bo Tang, Junyi Zhu, Wenqiang Wei, Chen Chen, Chao Yang, Jingfeng Zhang, Chaochao Lu, Yijun Niu, Keming Mao, Zhiyu Li, Feiyu Xiong, Jie Hu, and Mingchuan Yang. Adversarial preference learning for robust llm alignment. *arXiv preprint arXiv:2505.24369*, 2025b.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1112–1122, 2018.

Yue Yang, Shuibai Zhang, Wenqi Shao, Kaipeng Zhang, Yi Bin, Yu Wang, and Ping Luo. Dynamic multimodal evaluation with flexible complexity by vision-language bootstrapping. In *International Conference on Learning Representations*, 2025.

Ziyu Ye, Rishabh Agarwal, Tianqi Liu, Rishabh Joshi, Sarmishta Velury, Quoc V Le, Qijun Tan, and Yuan Liu. Reward-guided prompt evolving in reinforcement learning for llms. In *Forty-second International Conference on Machine Learning*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2025.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, Fangfu Liu, Xiang Xu, Jiaze Ma, Xuekai Zhu, Ermo Hua, Yihao Liu, Zonglin Li, Huayu Chen, Xiaoye Qu, Yafu Li, Weize Chen, Zhenzhao Yuan, Junqi Gao, Dong Li, Zhiyuan Ma, Ganqu Cui, Zhiyuan Liu, Biqing Qi, Ning Ding, and Bowen Zhou. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024.

Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 2025.

Pengfei Zhou, Xiaopeng Peng, Fanrui Zhang, Zhaopan Xu, Jiaxin Ai, Yansheng Qiu, Wangbo Zhao, Jiajun Song, Chuanhao Li, Weidong Tang, et al. MDK12-Bench: A multi-discipline benchmark for evaluating reasoning in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28982–28990, 2026.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. FreeLB: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2020.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.

A Formal Definitions of (C1) and (C2)

We give the formal versions of the two design conditions of Section 2.4. Total variation is $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$ ($= \frac{1}{2}\|P - Q\|_1$ on countable spaces). Transformation distributions are probability measures on $(\mathcal{T}, \Sigma_{\mathcal{T}})$; finite-pool distributions are zero-extended to this space.

Joint buckets. Fix a prompt $\tau(\boldsymbol{x})$ and its extractor $\mathcal{A}_\tau$ (Definition A.2). For a response $\boldsymbol{y}$ put $V_f := \mathbf{1}\{\mathcal{A}_\tau(\boldsymbol{y}) \neq \emptyset\}$, $\widetilde{V}_c := \mathbf{1}\{\mathcal{A}_\tau(\boldsymbol{y}) = a^*\}$ and $B := (V_f, \widetilde{V}_c)$, the *joint bucket*. Since $\widetilde{V}_c = 1$ forces $V_f = 1$, B takes three values: $\mathrm{C} = (1,1)$ (correct), $\mathrm{W} = (1,0)$ (formatted, wrong) and $\mathrm{F} = (0,0)$ (format failure), and Eq. (5) reads $R_t = V_f + \widetilde{V}_c - 1$. For a format-exact extractor ($\mathcal{A}_\tau(\boldsymbol{y}) \neq \emptyset$ iff $\boldsymbol{y}$ matches $\mathcal{F}_\tau$), V_f is the format indicator of Eq. (2), and the binary verifier that reads the answer through the same extractor is $V = \widetilde{V}_c$.

Definition A.1 ((C1) Reward Decomposition)**.** The reward is supplied as a pair (R_t, R_s): $R_t \in \{-1, 0, +1\}$ is the format-aware trinary of Eq. (5), a function of B, and $R_s \in [0, 1]$ is a bounded measurable semantic score of the response, computed without the requested delimiter, with averaged within-bucket dispersion $\mathbb{E}[\mathrm{Var}(R_s \mid B)] \geq m$ for every rollout law in the class considered (the rollout law of a prompt, or the empirical law of one rollout group), with a constant $m > 0$ independent of p_f. The two components enter the advantage through separate group normalization (Eq. (7)); the scalar $R_{\mathrm{DTR}} = R_t + \alpha_{\mathrm{rew}} R_s$ of Eq. (6) is the reward that the adversary descends.

Only the bucket average is required: an Exact answer on C has $R_s = 1$, and the dispersion is carried by the failure buckets W and F, on which the binary verifier is constant. By the law of total variance and Popoviciu's inequality, $m \leq \mathrm{Var}(R_s) \leq \frac{1}{4}$; if R_s is constant on C, then $\mathbb{E}[\mathrm{Var}(R_s \mid B)] \leq \frac{1}{4} \Pr[B \neq \mathrm{C}]$, so every rollout law of the class has failure mass $\Pr[B \neq \mathrm{C}] \geq 4m$.

Definition A.2 ((C2) Invariance Control (answer-level))**.** Let $\mathcal{C}(\boldsymbol{x}) \subseteq \mathcal{T}$ be a finite semantics-preserving pool for the task instance $\boldsymbol{x}$ ($\tau(\boldsymbol{x}) \in [\![\boldsymbol{x}]\!]$ for $\tau \in \mathcal{C}(\boldsymbol{x})$), and equip $\mathcal{Y}_{\mathrm{ans}} \cup \{\emptyset\}$, with $\emptyset \notin \mathcal{Y}_{\mathrm{ans}}$, with the discrete sigma-algebra. For each τ, $\mathcal{A}_\tau(\boldsymbol{y}) := \mathcal{A}(\boldsymbol{y}; d_\tau)$ is the measurable deterministic extractor at the delimiter requested by τ (the extractor of Eq. (5)), $\mathcal{F}_\tau$ is the requested format, and $P^\theta_\tau := \pi_\theta^{\mathcal{A}_\tau}(\cdot \mid \tau(\boldsymbol{x})) := (\mathcal{A}_\tau)_\# \pi_\theta(\cdot \mid \tau(\boldsymbol{x}))$ is the answer marginal ($\sum_{\boldsymbol{y}:\mathcal{A}_\tau(\boldsymbol{y})=a} \pi_\theta(\boldsymbol{y} \mid \tau(\boldsymbol{x}))$ at a for countable response spaces). Extractors are delimiter-separated: on a response in $\mathcal{F}_\tau$ that encodes a, $\mathcal{A}_\tau$ returns a, and on a response in a format family disjoint from $\mathcal{F}_\tau$ it returns $\emptyset$. Let D be an f-divergence with a Pinsker-type bound $\mathrm{TV}(P, Q) \leq \kappa_D \sqrt{D(P\|Q)}$ ($\kappa_D = 1/\sqrt{2}$ for the natural-log KL; $\kappa_D = 1$ for $D = \mathrm{TV}$, since $\mathrm{TV} \leq \sqrt{\mathrm{TV}}$ on $[0, 1]$). For a non-degenerate coverage distribution $\rho_{\boldsymbol{x}} \in \Delta(\mathcal{C}(\boldsymbol{x}))$ (at least two atoms, which rules out the vacuous $\rho_{\boldsymbol{x}} = \delta_\tau$), (C2) with tolerance ε requires

$$\mathbb{E}_{\tau_1, \tau_2 \sim \rho_{\boldsymbol{x}}} \left[D\big(P^\theta_{\tau_1} \,\big\|\, P^\theta_{\tau_2}\big) \right] \;\leq\; \varepsilon.$$

Lemma A.3 (Data-processing inequality for the answer marginal)**.** *For any f-divergence D, any fixed deterministic extraction map $\mathcal{A}$ and any two prompts, with $P_i = \pi_\theta(\cdot \mid \tau_i(\boldsymbol{x}))$ the full sequence laws, $D(\mathcal{A}_\# P_1 \,\|\, \mathcal{A}_\# P_2) \leq D(P_1 \,\|\, P_2)$.*

Proof. $\boldsymbol{y} \mapsto \mathcal{A}(\boldsymbol{y})$ is a deterministic channel; apply the data processing inequality for f-divergences. □

The per-token estimator of Eq. (8). Let $P^\phi := \pi_\theta(\cdot \mid \tau_\phi(\boldsymbol{x}))$ and $P^0 := \pi_\theta(\cdot \mid \tau_0(\boldsymbol{x}))$, with the next-token conditionals p_t^ϕ, p_t^0 of Eq. (8), and $\ell_\theta(\boldsymbol{y}) := |\boldsymbol{y}|^{-1} \sum_{t=1}^{|\boldsymbol{y}|} D_{\mathrm{KL}}(p_t^\phi \,\|\, p_t^0)$. Both laws stop at EOS or at the cap $L_{\max}$ and are read by the same extractor $\mathcal{A}_{\tau_0}$, so D_{KL} between their answer marginals is at most $D_{\mathrm{KL}}(P^\phi \| P^0) = \mathbb{E}_{P^\phi}[|\boldsymbol{y}|\, \ell_\theta(\boldsymbol{y})] \leq L_{\max} \mathbb{E}_{P^\phi} \ell_\theta$ (Proposition B.14(v)). The released loss evaluates the full-vocabulary terms $D_{\mathrm{KL}}(p_t^\phi \,\|\, \mathrm{sg}(p_t^0))$ (the stop-gradient changes the update direction, not the value), averaged over the positions of each rollout and then over rollouts. The rollouts are drawn by $\pi_{\theta_{\mathrm{old}}}$ at temperature 0.9 and top-p 0.99, so the loss averages ℓ_θ under this sampler; it is the Monte-Carlo estimate of the bounded quantity $\mathbb{E}_{P^\phi} \ell_\theta$ at temperature 1, top-p 1 and $\theta = \theta_{\mathrm{old}}$, the first of the C_{adv} updates after a refresh.

B Full Proofs

B.1 Within-format share of the binary verifier

The binary verifier carries content signal only inside the formatted bucket, and its within-format share is small exactly when the zero-reward rollouts are mostly format failures. Here $V = V_f V_c$ as in Eq. (2), where V_c records content correctness for every response.

Lemma B.1 (Within-format share of the binary verifier)**.** *Let $p_f := \Pr[V_f{=}1] \in (0, 1)$ and $p := \Pr[V_c{=}1 \mid V_f{=}1] \in (0, 1]$, so that $\mathrm{Var}(V) > 0$; no independence of V_f and V_c is assumed. The law of total variance gives*

$$\mathrm{Var}(V) = \underbrace{p_f\, p(1{-}p)}_{\textit{within-format}} + \underbrace{p^2 p_f(1{-}p_f)}_{\textit{between-format}} = p_f\, p\,(1{-}p_f\, p), \tag{11}$$

and the within-format reward-variance share *is*

$$\Phi_{\mathrm{wf}} := \frac{\mathbb{E}[\mathrm{Var}(V \mid V_f)]}{\mathrm{Var}(V)} = \frac{1-p}{1-p_f\,p} = \frac{\pi_W}{p_f}, \qquad \pi_W := \Pr[V_f{=}1 \mid V{=}0], \tag{12}$$

strictly decreasing in p. For $p < 1$ it is strictly increasing in p_f. Under $V_f \perp V_c$, $p = p_c := \Pr[V_c{=}1]$ and $\Phi_{\mathrm{wf}} = (1-p_c)/(1-p_f p_c)$. At a fixed marginal p_c, positive correlation of V_f and V_c gives a smaller share than this formula and negative correlation a larger one. The share is also that of the GRPO advantage, since $\mathrm{Norm}_{\mathcal{G}}(V)$ is a positive affine function of V.

Proof. $V = V_c$ on $\{V_f = 1\}$ and $V = 0$ on $\{V_f = 0\}$, so $\mathbb{E}[\mathrm{Var}(V \mid V_f)] = p_f\,p(1-p)$ and $\mathrm{Var}(\mathbb{E}[V \mid V_f]) = p_f p^2 - (p_f p)^2 = p^2 p_f(1-p_f)$; their sum factors as in Eq. (11), and the ratio is $(1-p)/(1-p_f p)$. Also $\pi_W = \Pr[V_f{=}1, V_c{=}0]/\Pr[V{=}0] = p_f(1-p)/(1-p_f p)$. The derivatives $\partial_{p_f}\Phi_{\mathrm{wf}} = p(1-p)/(1-p_f p)^2$ (> 0 for $p < 1$) and $\partial_p \Phi_{\mathrm{wf}} = -(1-p_f)/(1-p_f p)^2 < 0$ give the monotonicity. Finally $\mathrm{Cov}(V_f, V_c) = p_f(p - p_c)$, so at fixed p_c positive correlation means $p > p_c$, and variance shares are invariant under positive affine maps. □

Proposition B.2 (Where the binary verifier discards content signal)**.** *Fix a prompt and a rollout law with $\mu := \mathbb{E}V \in (0,1)$ and $p_f \in (0,1)$, and let $\eta := \Pr[V_c{=}0]$ be the content error and $\bar{p}_f := 1 - p_f$ the format-miss rate.*

(i) Content-correct format failures. *These rollouts, which the binary reward scores like every other failure, have mass $\Pr[V_f{=}0, V_c{=}1] = 1 - \mu - \eta$, and*

$$\Phi_{\mathrm{wf}} \;\le\; \frac{\eta}{p_f(1-\mu)} \;\le\; \min\Big\{\frac{\eta}{\mu(1-\mu)},\, \frac{\eta}{\bar{p}_f(1-\bar{p}_f)}\Big\}; \qquad 1 - \mu \ge k\eta \implies \Phi_{\mathrm{wf}} \le \frac{1}{k p_f}.$$

(ii) Pool templates and adversarial prompts. *Under the same hypotheses, at a pool template τ, $\Phi_{\mathrm{wf}}(\tau) \le \eta(\tau)/(\bar{p}_f(\tau)(1-\bar{p}_f(\tau)))$; at an adversarial prompt with binary mean μ' and content error η', $\Phi_{\mathrm{wf}} \le \eta'/(\mu'(1-\mu'))$ with a mass $1 - \mu' - \eta'$ of content-correct format failures. A policy that knows the answer and misses the requested delimiter has a small binary share there.*

(iii) Persistence. *PIRL draws τ_0 per instance from the uniform pool and restarts the adversary at $\boldsymbol{\delta} = \mathbf{0}$ against the current policy at every refresh, so the prompts of (ii) stay in training until the policy follows every pool delimiter and is robust within the ϵ_{emb}-ball.*

Proof. Let $a := \Pr[V_f{=}1, V_c{=}0] \le \eta$. By Eq. (12), $\Phi_{\mathrm{wf}} = \pi_W/p_f = a/(p_f(1-\mu))$, and $\Pr[V_f{=}0, V_c{=}1] = \Pr[V_c{=}1] - \Pr[V{=}1] = 1 - \eta - \mu$. The minimum follows from $\mu = p_f p \le p_f$ and $1 - \mu \ge \Pr[V_f{=}0] = \bar{p}_f$, the last bound from $1 - \mu \ge k\eta$. (ii) is (i) at the respective prompts, and (iii) restates the sampler (Algorithm 2). □

On these prompts R_t alone places all format failures lowest without raising the within-format share (Corollary B.12), while DAN ranks format failures by R_s (Proposition B.13) and keeps the floor of Theorem B.10 for every $\bar{p}_f$.

B.2 Quantitative non-identifiability (Proposition B.6)

Off $\mathrm{supp}\, P_{\mathrm{train}}$ the RLVR objective cannot distinguish following the instruction from ignoring it; this blind spot is what (C2) constrains. We use $\mathcal{T}$ and $\mathcal{R}(\pi, \rho; \boldsymbol{x})$ of Section 2.3, and write $\mathcal{F}_\tau := \mathcal{F}(\mathcal{I}_\tau)$ for $\tau(\boldsymbol{x}) = (\boldsymbol{v}, \boldsymbol{q}, \mathcal{I}_\tau)$.

Assumption B.3 (Restricted prompt coverage)**.** There are sets $\emptyset \ne S_{\mathrm{train}} \subseteq \mathcal{T}_{\mathrm{train}} \subsetneq \mathcal{T}$ in $\Sigma_{\mathcal{T}}$ with $P_{\mathrm{train}}(S_{\mathrm{train}}) = 1$.

Assumption B.4 (Disjoint unseen format)**.** There is $\tau_B \in \mathcal{T} \setminus \mathcal{T}_{\mathrm{train}}$ whose format constraint $\mathcal{F}_B := \mathcal{F}_{\tau_B}$ is disjoint from $\mathcal{F}_{\mathrm{train}} := \bigcup_{\tau \in \mathcal{T}_{\mathrm{train}}} \mathcal{F}_\tau$.

Assumption B.5 (Policy-class extension realizability)**.** The policy class contains two measurable deterministic branches π_1, π_2 that coincide on $\mathcal{T} \setminus \{\tau_B\}$ and emit the correct answer a^* in the requested format on S_{train}. At τ_B, π_1 emits a^* in the format $\mathcal{F}_A := \mathcal{F}_{\tau_A}$ of some $\tau_A \in S_{\mathrm{train}}$, while π_2 emits a^* in $\mathcal{F}_B$.

Proposition B.6 (Quantitative non-identifiability). *Let Assumptions B.3–B.5 hold, and let the verifier be τ-dependent with reward $R \in [R_{\min}, R_{\max}]$, such that a correct response in $\mathcal{F}_\tau$ receives $R_{\max}$ and any response in a format constraint disjoint from $\mathcal{F}_\tau$ receives $R_{\min}$; put $\Delta_R := R_{\max} - R_{\min}$. Then the branches $\pi_1 \neq \pi_2$ both attain the maximal training reward $R_{\max}$, so they are P_{train}-optimal with*

$$\mathcal{R}(\pi_1, P_{train}; \boldsymbol{x}) = \mathcal{R}(\pi_2, P_{train}; \boldsymbol{x}), \tag{13}$$

and every $P_{test} \in \Delta(\mathcal{T})$ with atomic mass $m_B := P_{test}(\{\tau_B\}) > 0$ separates them:

$$\mathcal{R}(\pi_2, P_{test}; \boldsymbol{x}) - \mathcal{R}(\pi_1, P_{test}; \boldsymbol{x}) = m_B \Delta_R. \tag{14}$$

Hence the P_{train}-optimum set has P_{test}-diameter at least $m_B \Delta_R$.

Proof. Since $\tau_B \notin \mathcal{T}_{\text{train}} \supseteq S_{\text{train}}$, the branches coincide on S_{train}, where they emit a correct response in the requested format and receive $R_{\max}$; as $P_{\text{train}}(S_{\text{train}}) = 1$ and $R \leq R_{\max}$, both are optimal and Eq. (13) holds. At τ_B, π_1 answers in $\mathcal{F}_A \subseteq \mathcal{F}_{\text{train}}$, which is disjoint from $\mathcal{F}_B$, and receives $R_{\min}$, while π_2 receives $R_{\max}$. The pointwise difference is therefore $\Delta_R \mathbf{1}\{\tau = \tau_B\}$; integrating against P_{test} gives Eq. (14). □

Corollary B.7 (Sets of unseen formats). *Under the reward conditions of Proposition B.6, let $S_B \subseteq \mathcal{T} \setminus \mathcal{T}_{train}$ be measurable with $\mathcal{F}_\tau \cap \mathcal{F}_{train} = \emptyset$ for every $\tau \in S_B$, and let the policy class contain two measurable branches that coincide off S_B, emit a^* in the requested format on S_{train}, and at every $\tau \in S_B$ emit a^* in a training format (π_1) or in $\mathcal{F}_\tau$ (π_2). Then both are P_{train}-optimal and, for every $P_{test} \in \Delta(\mathcal{T})$, $\mathcal{R}(\pi_2, P_{test}; \boldsymbol{x}) - \mathcal{R}(\pi_1, P_{test}; \boldsymbol{x}) = P_{test}(S_B)\, \Delta_R$.*

Proof. $S_B \cap S_{\text{train}} = \emptyset$, so both branches receive $R_{\max}$ on S_{train}, and the pointwise difference is Δ_R on S_B (as at τ_B above) and 0 elsewhere. □

B.3 Measure-zero training support and low-complexity shortcuts

No finite template pool removes the blind spot, and a complexity tie-breaker turns non-identifiability into the selection of a shortcut.

Proposition B.8 (Measure-zero training support). *Let $\mathcal{W} \subseteq \mathbb{R}^n$ be measurable, let the deployment distribution satisfy $P_W \ll \text{Leb}_n$, and let the training distribution satisfy $P_{\text{tr}}(\Omega) = 1$ for a measurable $\Omega \subseteq \mathcal{W}$ that is finite or contained in a measurable d-dimensional embedded submanifold with $d < n$. Let $\mathcal{Y}_{\text{ans}}$ be discrete with at least two labels and $A^\star : \mathcal{W} \to \mathcal{Y}_{\text{ans}}$ the measurable deployment semantic rule. For every bounded measurable reward R, two policies whose conditional output laws agree on Ω have the same objective $J(\pi) = \mathbb{E}_{x \sim P_{\text{tr}}} \mathbb{E}_{y \sim \pi(\cdot|x)}[R(x, y)]$. Consequently there are infinitely many measurable shortcut rules $a : \mathcal{W} \to \mathcal{Y}_{\text{ans}}$ with $a = A^\star$ on Ω and $P_W(\{a \neq A^\star\}) > 0$, each with $\pi_a(\cdot \mid x) = \delta_{a(x)}$ training-equivalent to $\pi_{A^\star}$. If training-optimal policies are selected by $\arg\min_{\pi \in \Pi^\star} C(\pi)$ with $\Pi^\star = \{\pi : J(\pi) = J^\star\}$ and $\pi_{A^\star} \in \Pi^\star$, every such shortcut lies in $\Pi^\star$, and any one with $C(\pi_a) < C(\pi_{A^\star})$ is preferred to the semantic policy.*

Proof. Finite sets and lower-dimensional embedded submanifolds are Leb_n-null, so $P_W(\Omega) = 0$, while J integrates only over Ω. Fix distinct labels b_0, b_1. As P_W is non-atomic and $P_W(\mathcal{W} \setminus \Omega) = 1$, there are infinitely many measurable $S \subseteq \mathcal{W} \setminus \Omega$ with $P_W(S) > 0$. Let $a_S := A^\star$ off S and, on S, $a_S := b_1$ where $A^\star = b_0$ and $a_S := b_0$ elsewhere. Then a_S is measurable, $a_S = A^\star$ on Ω, $a_S \neq A^\star$ exactly on S (so distinct S give distinct rules), and $J(\pi_{a_S}) = J(\pi_{A^\star})$, which places π_{a_S} in $\Pi^\star$ whenever $\pi_{A^\star}$ is; the tie-breaker then prefers every such shortcut of lower complexity C. □

B.4 Dense rewards without invariant semantic coverage

A prompt-invariant score such as R_s leaves the blind spot of Proposition B.6 intact: when it scores the two correct branches equally, $\Delta_{\text{inv}} = 0$ below and the gap is $m_B \Delta_{\text{dep}}$. R_s is therefore a within-bucket ranking signal (C1), and invariance comes from (C2).

Proposition B.9 (Dense rewards without invariant semantic coverage). *Let the dense reward decompose, after absorbing component weights, as $R(\boldsymbol{y}, \tau(\boldsymbol{x})) = R_{\text{dep}}(\boldsymbol{y}, \tau(\boldsymbol{x})) + R_{\text{inv}}(\boldsymbol{y}, \boldsymbol{x})$, with measurable components integrable under the policies considered, where R_{dep} collects the prompt-dependent verifier and format terms and R_{inv} is prompt-invariant; write $\mathcal{R}_j(\pi, \tau; \boldsymbol{x})$ for the expected*

value of R_j under $\boldsymbol{y} \sim \pi(\cdot \mid \tau(\boldsymbol{x}))$. Let $\bar{\pi}$ be a measurable maximizer over the admissible class Π of $J_{\mathrm{dense}}(\pi) := \mathbb{E}_{\tau \sim P_{train}} \mathbb{E}_{\boldsymbol{y} \sim \pi(\cdot \mid \tau(\boldsymbol{x}))}[R(\boldsymbol{y}, \tau(\boldsymbol{x}))]$, and let Π contain two extensions π_1, π_2 of $\bar{\pi}$ that equal $\bar{\pi}$ P_{train}-a.s., coincide away from τ_B, and use at τ_B the two branches of Proposition B.6. With $\Delta_{\mathrm{dep}} := \mathcal{R}_{\mathrm{dep}}(\pi_2, \tau_B; \boldsymbol{x}) - \mathcal{R}_{\mathrm{dep}}(\pi_1, \tau_B; \boldsymbol{x})$ and $\Delta_{\mathrm{inv}} := \mathcal{R}_{\mathrm{inv}}(\pi_1, \tau_B; \boldsymbol{x}) - \mathcal{R}_{\mathrm{inv}}(\pi_2, \tau_B; \boldsymbol{x})$, both extensions are P_{train}-optimal, the training equality (13) holds for R and for each component, and for every P_{test} with $m_B = P_{test}(\{\tau_B\}) > 0$,

$$\mathcal{R}(\pi_2, P_{test}; \boldsymbol{x}) - \mathcal{R}(\pi_1, P_{test}; \boldsymbol{x}) = m_B(\Delta_{\mathrm{dep}} - \Delta_{\mathrm{inv}}),$$

which is positive whenever $\Delta_{\mathrm{dep}} > \Delta_{\mathrm{inv}}$.

Proof. J_{dense} and its components integrate only against P_{train}, and both extensions equal $\bar{\pi}$ P_{train}-a.s., so they share the optimal value of $\bar{\pi}$ and every training-side component. They coincide on every $\tau \neq \tau_B$, so the test difference is m_B times the difference at τ_B, which is $\Delta_{\mathrm{dep}} - \Delta_{\mathrm{inv}}$ by definition. □

B.5 Within-bucket share of the DAN advantage signal

Under (C1) the DAN signal keeps a within-format share bounded below independently of p_f and α_{rew}, while single normalization of R_{DTR} does not.

Setting. Fix a prompt and a probability law of the rollout: the population law, or the empirical law of one rollout group, whose moments are exactly those used by $\mathrm{Norm}_{\mathcal{G}}$. Let B be the joint bucket of Appendix A, and $\mu_t, \sigma_t^2, \mu_s, \sigma_s^2$ the means and variances of R_t and R_s. For $\sigma_t, \sigma_s > 0$ put $\tilde{U} := (R_t - \mu_t)/\sigma_t$, $\tilde{W} := (R_s - \mu_s)/\sigma_s$, $\rho_{ts} := \mathrm{Corr}(R_t, R_s)$ and $s := \mathbb{E}[\mathrm{Var}(R_s \mid B)]/\mathrm{Var}(R_s) \in [0, 1]$, the share of semantic-score variance *inside* the buckets. For a signal X with $\mathrm{Var}(X) > 0$ let $\Phi_{\mathrm{wf}}(X) := \mathbb{E}[\mathrm{Var}(X \mid V_f)]/\mathrm{Var}(X)$ and $\Phi_{\mathrm{wb}}(X) := \mathbb{E}[\mathrm{Var}(X \mid B)]/\mathrm{Var}(X)$. For $\lambda \geq 0$ let $Z_\lambda := \tilde{U} + \lambda \tilde{W}$ and $F_s(\lambda) := \lambda^2 s/(1 + \lambda^2 + 2\lambda\sqrt{1-s})$.

Theorem B.10 (Within-bucket share of the DAN advantage signal)**.** *Assume $\sigma_s > 0$ and $\epsilon_{\mathrm{norm}} = 0$; in (i)–(iii) assume also $\sigma_t > 0$.*

(i) Exact share. *For every $\lambda \geq 0$ with $\mathrm{Var}(Z_\lambda) > 0$ (which holds whenever $s > 0$), $\Phi_{\mathrm{wb}}(Z_\lambda) = \lambda^2 s/(1 + \lambda^2 + 2\lambda\rho_{ts})$ and $\Phi_{\mathrm{wf}}(Z_\lambda) \geq \Phi_{\mathrm{wb}}(Z_\lambda)$.*

(ii) Sharp floor. *$|\rho_{ts}| \leq \sqrt{1-s}$, hence $\Phi_{\mathrm{wf}}(Z_\lambda) \geq \Phi_{\mathrm{wb}}(Z_\lambda) \geq F_s(\lambda)$. $F_s(\lambda)$ is nondecreasing in λ and in s, strictly increasing in λ when $s > 0$. For every $s \in [0, 1]$ and $\lambda > 0$ some rollout law has within-bucket share s and $\Phi_{\mathrm{wf}}(Z_\lambda) = F_s(\lambda)$.*

(iii) The three estimators. *The DAN signal of Eq. (7) before batch normalization is $Z = Z_{\alpha_{\mathrm{adv}}}$; it does not involve α_{rew}. Single normalization of the DTR reward is $\mathrm{Norm}_{\mathcal{G}}(R_t + \alpha_{\mathrm{rew}} R_s) = Z_{\lambda_{\mathrm{rew}}}/\mathrm{sd}(Z_{\lambda_{\mathrm{rew}}})$ with $\lambda_{\mathrm{rew}} := \alpha_{\mathrm{rew}}\sigma_s/\sigma_t$. The binary GRPO signal $\mathrm{Norm}_{\mathcal{G}}(V)$, $V = \widetilde{V}_c$, is a function of B, so its within-bucket share is 0.*

(iv) Floor free of p_f and α_{rew}. *Under (C1), $\sigma_s^2 \geq m$, $s \geq m/\sigma_s^2 \geq 4m$ and, for every $p_f = \Pr[V_f{=}1] \in [0, 1]$,*

$$\Phi_{\mathrm{wf}}(Z) \;\geq\; F_{4m}(\alpha_{\mathrm{adv}}) \;\geq\; \frac{4\alpha_{\mathrm{adv}}^2\, m}{(1 + \alpha_{\mathrm{adv}})^2}, \qquad \Phi_{\mathrm{wf}}(Z)\big|_{\alpha_{\mathrm{adv}}=1} \;\geq\; \tfrac{1}{2}\big(1 - \sqrt{1 - 4m}\big) \;\geq\; m.$$

If $\sigma_t = 0$, the convention $\mathrm{Norm}_{\mathcal{G}}(R_t) = 0$ applies and $\Phi_{\mathrm{wf}}(Z) = 1$. By contrast, the single-normalized DTR signal has $\Phi_{\mathrm{wb}} \leq \lambda_{\mathrm{rew}}^2 s/(1 - \lambda_{\mathrm{rew}})^2$ whenever $\lambda_{\mathrm{rew}} < 1$.

Proof. (i). $\tilde{U}$ is B-measurable, so $\mathrm{Var}(Z_\lambda \mid B) = \lambda^2 \mathrm{Var}(\tilde{W} \mid B)$, whose mean is $\lambda^2 s$, and $\mathrm{Var}(Z_\lambda) = 1 + \lambda^2 + 2\lambda\rho_{ts}$. As V_f is a function of B, the conditional law of total variance gives $\mathrm{Var}(Z_\lambda \mid V_f) \geq \mathbb{E}[\mathrm{Var}(Z_\lambda \mid B) \mid V_f]$; take expectations. If $s > 0$, then $|\rho_{ts}| < 1$ by (ii) and $\mathrm{Var}(Z_\lambda) \geq (1 - \lambda)^2 + 2\lambda(1 - |\rho_{ts}|) > 0$ for $\lambda > 0$, while $\mathrm{Var}(Z_0) = 1$.

(ii). As $\tilde{U}$ is B-measurable with $\|\tilde{U}\|_2 = 1$ and $\mathbb{E}[\tilde{W} \mid B]$ has mean 0, Cauchy–Schwarz gives $|\rho_{ts}| = |\mathbb{E}[\tilde{U}\, \mathbb{E}[\tilde{W} \mid B]]| \leq (\mathrm{Var}\, \mathbb{E}[\tilde{W} \mid B])^{1/2} = (1 - \mathbb{E}[\mathrm{Var}(\tilde{W} \mid B)])^{1/2} = \sqrt{1-s}$, which bounds the denominator in (i) by $1 + \lambda^2 + 2\lambda\sqrt{1-s}$. With $r := \sqrt{1-s}$ and $D := 1 + \lambda^2 + 2\lambda r$,

$\partial_\lambda F_s = 2\lambda s(1+\lambda r)/D^2$; in s, the numerator of F_s increases and D does not. Attainment: let only the buckets F and C occur, with probabilities $b \in (0,1)$ and $1-b$, so $\tilde{U} = -\sqrt{(1-b)/b}$ on F and $\sqrt{b/(1-b)}$ on C. For a fair sign ξ independent of B let $\tilde{W} := \sqrt{1-s}\,\tilde{U} + \sqrt{s}\,\xi$ (mean 0, variance 1) and $R_s := \frac{1}{2} + c\tilde{W}$ with $0 < c \le \frac{1}{2}/\max|\tilde{W}|$, so $R_s \in [0,1]$ has standardization $\tilde{W}$, $\mathrm{Var}(\tilde{W} \mid B) = s$ in both buckets and $\rho_{ts} = \sqrt{1-s}$. Here B and V_f generate the same partition, so $\Phi_{\mathrm{wf}} = \Phi_{\mathrm{wb}} = F_s(\lambda)$.

(iii). With $\epsilon_{\mathrm{norm}} = 0$, $\mathrm{Norm}_{\mathcal{G}}(R_t) = \tilde{U}$ and $\mathrm{Norm}_{\mathcal{G}}(R_s) = \tilde{W}$ in Eq. (7); $R_t + \alpha_{\mathrm{rew}} R_s - \mathbb{E}[R_t + \alpha_{\mathrm{rew}} R_s] = \sigma_t Z_{\lambda_{\mathrm{rew}}}$ has standard deviation $\sigma_t\,\mathrm{sd}(Z_{\lambda_{\mathrm{rew}}})$; and $V = \widetilde{V}_c$ is a coordinate of B.

(iv). The law of total variance gives $\sigma_s^2 \ge \mathbb{E}[\mathrm{Var}(R_s \mid B)] \ge m$ and Popoviciu's inequality $\sigma_s^2 \le \frac{1}{4}$, so $s \ge m/\sigma_s^2 \ge 4m$. For $\sigma_t > 0$, (ii) and monotonicity in s give $\Phi_{\mathrm{wf}}(Z) \ge F_{4m}(\alpha_{\mathrm{adv}})$, and $\sqrt{1-4m} \le 1$ gives the second bound. $F_s(1) = s/(2+2\sqrt{1-s}) = \frac{1}{2}(1-\sqrt{1-s}) \ge s/4$ because $\sqrt{1-s} \le 1 - s/2$; take $s = 4m$. For $\sigma_t = 0$, R_t, hence B and V_f, is constant and $Z = \alpha_{\mathrm{adv}}\tilde{W}$, so $\mathbb{E}[\mathrm{Var}(Z \mid V_f)] = \mathrm{Var}(Z) = \alpha_{\mathrm{adv}}^2 > 0$. No step uses p_f or α_{rew}. The contrast follows from $1 + \lambda^2 + 2\lambda\rho_{ts} \ge (1-\lambda)^2$ in (i). □

Remark B.11 (Finite groups, ϵ_{norm}, batch normalization, degenerate groups). The theorem holds for every realized rollout group, with s replaced by the realized share $s_{\mathcal{G}}$. If R_t is not constant in a group of size G, then $\sigma_t \ge \sqrt{G-1}/G$ (the $k(G-k) \ge G-1$ pairs formed by a minimal and a non-minimal value of the integer-valued R_t differ by at least 1, and $\sum_{i<j}(R_t^{(i)} - R_t^{(j)})^2 = G^2\sigma_t^2$) and $\sigma_s \le \frac{1}{2}$, so $\lambda_{\mathrm{rew}} \le \alpha_{\mathrm{rew}} G/(2\sqrt{G-1}) < 0.0207$ at $G = 16$, $\alpha_{\mathrm{rew}} = 0.01$: single normalization keeps a within-bucket share of at most $4.45 \times 10^{-4} s_{\mathcal{G}}$, DAN at $\alpha_{\mathrm{adv}} = 1$ at least $F_{s_{\mathcal{G}}}(1) \ge s_{\mathcal{G}}/4$. With $\epsilon_{\mathrm{norm}} > 0$, $c_t := \sigma_t/(\sigma_t + \epsilon_{\mathrm{norm}})$ and $c_s := \sigma_s/(\sigma_s + \epsilon_{\mathrm{norm}})$, the implemented signal is $Z = c_t\, Z_{\alpha_{\mathrm{adv}} c_s/c_t}$; shares are scale invariant and F_s is nondecreasing, so $\Phi_{\mathrm{wf}}(Z) \ge F_s(\alpha_{\mathrm{adv}} c_s)$. Both components of Z are centred within each group, so $\mathrm{Norm}_{\mathcal{B}}(Z) = Z/(\sigma_{\mathcal{B}} + \epsilon_{\mathrm{norm}})$ is one positive multiple of Z for the whole batch and preserves every share and every order; the clip at ± 5 is monotone. If R_t is constant in a group and R_s is not, $\mathrm{Norm}_{\mathcal{G}}(R_t) = 0$ exactly and $Z = \alpha_{\mathrm{adv}}\mathrm{Norm}_{\mathcal{G}}(R_s)$: DAN ranks by R_s a group on which binary GRPO has no signal, with $\Phi_{\mathrm{wf}} = 1$ and, by Samuelson's inequality, $|Z_i| \le \alpha_{\mathrm{adv}}\sqrt{G-1}$ however small σ_s is. If R_s is constant, DAN reduces to $\mathrm{Norm}_{\mathcal{G}}(R_t)$; if both are constant, $Z \equiv 0$.

Corollary B.12 (Additive format rewards do not repair entanglement). *Let $V = \widetilde{V}_c$, $p := \Pr[\widetilde{V}_c{=}1 \mid V_f{=}1] \in (0,1)$ and $p_f \in (0,1)$. Then*

$$\Phi_{\mathrm{wf}}(R_t) = \Phi_{\mathrm{wf}}(V)\,\frac{\mathrm{Var}(V)}{\mathrm{Var}(R_t)}, \qquad \mathrm{Var}(R_t) = \mathrm{Var}(V) + p_f(1-p_f)(1+2p),$$

so $\Phi_{\mathrm{wf}}(R_t) < \Phi_{\mathrm{wf}}(V)$: the trinary reward separates the format bit more sharply but adds no content signal inside a format bucket, and its within-bucket share is 0. Within-bucket content signal comes only from R_s under separate normalization (Theorem B.10).

Proof. Given $V_f = 1$, $R_t = V$; given $V_f = 0$, $R_t = -1$ and $V = 0$. Hence $\mathbb{E}[\mathrm{Var}(R_t \mid V_f)] = \mathbb{E}[\mathrm{Var}(V \mid V_f)]$. From $R_t = V + V_f - 1$, $\mathrm{Var}(R_t) = \mathrm{Var}(V) + \mathrm{Var}(V_f) + 2\mathrm{Cov}(V, V_f)$ with $\mathrm{Var}(V_f) = p_f(1-p_f)$ and $\mathrm{Cov}(V, V_f) = (1-p_f)\mathbb{E}V = p_f p(1-p_f)$. R_t is a function of B. □

Proposition B.13 (DAN ranks by a group-calibrated DTR reward). *Consider a rollout group with $\sigma_t, \sigma_s > 0$ and $\epsilon_{\mathrm{norm}} = 0$, and put $\lambda_{\mathcal{G}} := \alpha_{\mathrm{adv}}\sigma_t/\sigma_s$. Then*

$$Z_i = \frac{1}{\sigma_t}\big(R_t^{(i)} + \lambda_{\mathcal{G}} R_s^{(i)}\big) - c_{\mathcal{G}}, \qquad Z = \mathrm{sd}(Z)\,\mathrm{Norm}_{\mathcal{G}}\big(R_t + \lambda_{\mathcal{G}} R_s\big),$$

where $c_{\mathcal{G}}$ and $\mathrm{sd}(Z) = \sqrt{1 + \alpha_{\mathrm{adv}}^2 + 2\alpha_{\mathrm{adv}}\rho_{ts}}$ are constant in the group: DAN is group normalization of the DTR reward with the dynamic weight *$\lambda_{\mathcal{G}}$ in place of α_{rew}, times a group-level scale. The advantages $\hat{A}_i$ are nondecreasing in Z_i across the batch, and strictly increasing off the clipping set. Label a rollout C (correct and formatted), W (formatted, wrong), F^+ (format failure with $R_s = 1$, i.e. the correct content in another form) or F^- (format failure with $R_s < 1$). Within the group:*

(a) Content survives a format failure. *Rollouts with the same trinary outcome are ranked by R_s; in particular $F^+ \succ F^-$.*

(b) Format compliance is strictly rewarded. *At equal R_s, C $\succ$ W $\succ$ F. No rollout ranks above a C rollout with $R_s = 1$.*

(c) Exchange rate. *A rollout one trinary level lower ranks above one a level higher if and only if its R_s exceeds the other's by more than $1/\lambda_{\mathcal{G}} = \sigma_s/(\alpha_{\text{adv}}\sigma_t)$ (by more than $2/\lambda_{\mathcal{G}}$ for two levels). Thus F$^+$ $\succ$ W if and only if $R_s(\text{W}) < 1 - 1/\lambda_{\mathcal{G}}$.*

(d) Role of α_{adv}. *If $\lambda_{\mathcal{G}} < 1$, i.e. $\alpha_{\text{adv}} < \sigma_s/\sigma_t$, the group is ranked as by R_{DTR} ($\alpha_{\text{rew}} = 0.01 < 1$): C $\succ$ W $\succ$ F whatever the realized R_s values. This holds in every group of size G if $\alpha_{\text{adv}} < 1/\sqrt{2G}$ (0.177 at $G = 16$). A smaller α_{adv} extends the reward-level order to more groups, at the price of a lower floor $F_s(\alpha_{\text{adv}})$ in Theorem B.10.*

If $\sigma_t = 0$ the group is ranked by R_s alone; if $\sigma_s = 0$, by R_t alone. With $\epsilon_{\text{norm}} > 0$, (a)–(d) hold with $\lambda_{\mathcal{G}} = \alpha_{\text{adv}}(\sigma_t + \epsilon_{\text{norm}})/(\sigma_s + \epsilon_{\text{norm}})$ and the uniform threshold $1/((1 + \epsilon_{\text{norm}})\sqrt{2G})$.

Proof. $Z_i = (R_t^{(i)} - \mu_t)/\sigma_t + \alpha_{\text{adv}}(R_s^{(i)} - \mu_s)/\sigma_s$ gives the first identity with $c_{\mathcal{G}} := \mu_t/\sigma_t + \alpha_{\text{adv}}\mu_s/\sigma_s$, and standardizing $R_t + \lambda_{\mathcal{G}} R_s$ divides the same centred quantity by $\sigma_t \operatorname{sd}(Z)$. $\text{Norm}_{\mathcal{B}}$ multiplies by one positive scalar (Remark B.11) and the clip is nondecreasing. Write $\lambda = \lambda_{\mathcal{G}}$. (a) and (b) are the order of $R_t + \lambda R_s$ with one coordinate held fixed; a C rollout with $R_s = 1$ scores $1 + \lambda$, at least the score of any rollout ($\leq \lambda$ for W, $\leq -1 + \lambda$ for F, $\leq 1 + \lambda$ for C). (c): if $R_t^{(i)} = R_t^{(j)} - k$, then $R_t^{(i)} + \lambda R_s^{(i)} > R_t^{(j)} + \lambda R_s^{(j)}$ if and only if $R_s^{(i)} - R_s^{(j)} > k/\lambda$. (d): for $\lambda < 1$ and $R_s \in [0, 1]$ the score gaps C$-$W and W$-$F are at least $1 - \lambda > 0$ and C$-$F at least $2 - \lambda$. For the uniform statement, $(x_{\max} - \bar{x})^2 + (x_{\min} - \bar{x})^2 \leq G\sigma^2$ and $(a - b)^2 \leq 2(a^2 + b^2)$ give $\max_i R_s^{(i)} - \min_i R_s^{(i)} \leq \sigma_s\sqrt{2G}$, and $\sigma_t \leq 1$ (Popoviciu on $[-1, 1]$), so for $R_t^{(i)} < R_t^{(j)}$, $Z_j - Z_i \geq 1/\sigma_t - \alpha_{\text{adv}}\sqrt{2G} \geq 1 - \alpha_{\text{adv}}\sqrt{2G} > 0$. With $\epsilon_{\text{norm}} > 0$ the same computation runs with $\sigma_t + \epsilon_{\text{norm}} \leq 1 + \epsilon_{\text{norm}}$ and $\sigma_s + \epsilon_{\text{norm}} \geq \sigma_s$ in the denominators. □

B.6 How PIRL enforces (C2) (Proposition B.14)

The pool reward makes instruction-ignoring strictly suboptimal and yields (C2) in TV on the pool, and the consistency term controls answer drift between a pool prompt and its same-delimiter embedding neighbors. Fix $\boldsymbol{x}$ and the pool $\mathcal{C} = \{\tau_1, \ldots, \tau_K\}$ with $P_{\text{train}} \in \Delta(\mathcal{C})$ of full support (uniform with $K = 5$ in our implementation). For a response $\boldsymbol{y}$ let $\text{Acc}(\boldsymbol{y}) := \{\tau \in \mathcal{C} : \mathcal{A}_\tau(\boldsymbol{y}) \neq \emptyset\}$ and define the *pool acceptance number* $\chi(\mathcal{C}) := \sup_{\boldsymbol{y}} P_{\text{train}}(\text{Acc}(\boldsymbol{y}))$. A policy is *instruction-ignoring on $\mathcal{C}$* if $\pi(\cdot \mid \tau(\boldsymbol{x}))$ is one law $\mu_{\boldsymbol{x}}$ for every $\tau \in \mathcal{C}$; the *format-fixed branch* $\pi_1^{f_0}$ is the instruction-ignoring policy that always emits one response $\boldsymbol{y}_0$ stating a^* in a format f_0, with $\mathcal{A}_\tau(\boldsymbol{y}_0) = a^*$ for $\tau \in \mathcal{C}_{f_0} := \text{Acc}(\boldsymbol{y}_0)$. Write $P_\tau^\pi := (\mathcal{A}_\tau)_\# \pi(\cdot \mid \tau(\boldsymbol{x}))$, $\text{err}_\tau(\pi) := 1 - P_\tau^\pi(\{a^*\})$ and let the *pool suboptimality* be $\zeta(\pi) := 1 - \mathbb{E}_{\tau \sim P_{\text{train}}} \mathbb{E}_{\boldsymbol{y} \sim \pi(\cdot \mid \tau(\boldsymbol{x}))} R_t(\boldsymbol{y}, \tau) \in [0, 2]$.

Proposition B.14 (How PIRL enforces (C2)). *(i)* The pool reward identifies delimiter-following on the pool. $\mathbb{E}_{\tau \sim P_{\text{train}}} \text{err}_\tau(\pi) \leq \zeta(\pi)$ *and* $\mathbb{E}_{\tau \sim P_{\text{train}}} P_\tau^\pi(\{\emptyset\}) \leq \zeta(\pi)/2$; *$\zeta(\pi) = 0$ if and only if $\mathcal{A}_\tau(\boldsymbol{y}) = a^*$ almost surely for every $\tau \in \mathcal{C}$. Hence, if the policy class contains a policy with $\zeta = 0$, then for every answer-level reward $R(\boldsymbol{y}, \tau) = r(\mathcal{A}_\tau(\boldsymbol{y}))$ with $r(a^*) = R_{\max}$ the R_t-optimal policies form a set of P_{test}-diameter 0 for every P_{test} supported on $\mathcal{C}$: the non-identifiability of Proposition B.6 is removed at every separated format τ_B added to the pool. The same holds for the binary V, and for R_{DTR} when the class contains a policy with $\zeta = 0$ and $R_s = 1$.*

(ii) The price of ignoring the instruction. *Every instruction-ignoring π has $\zeta(\pi) \geq 2(1 - \chi(\mathcal{C}))$, and $\zeta(\pi_1^{f_0}) = 2\big(1 - P_{\text{train}}(\mathcal{C}_{f_0})\big)$. If the separated format τ_B of Proposition B.6 is added to the pool, its branches have $\zeta(\pi_1) = 2P_{\text{train}}(\{\tau_B\}) > 0 = \zeta(\pi_2)$. For R_{DTR} the suboptimality $1 + \alpha_{\text{rew}} - \mathbb{E}R_{\text{DTR}}$ lies in $[\zeta, \zeta + \alpha_{\text{rew}}]$, so these lower bounds hold and the two branches differ by at least $2P_{\text{train}}(\{\tau_B\}) - \alpha_{\text{rew}}$.*

(iii) Near-optimality on the pool implies (C2). $\mathbb{E}_{\tau_1, \tau_2 \sim P_{\text{train}}} \text{TV}(P_{\tau_1}^\pi, P_{\tau_2}^\pi) \leq \mathbb{E} \max\{\text{err}_{\tau_1}, \text{err}_{\tau_2}\} \leq 2\zeta(\pi)$: *$\pi$ satisfies Definition A.2 with $D = \text{TV}$ ($\kappa_D = 1$), $\rho_{\boldsymbol{x}} = P_{\text{train}}$ and $\varepsilon = 2\zeta(\pi)$.*

(iv) Consistency is blind to instruction-ignoring policies. *If the next-token conditionals of π_θ do not depend on the instruction block, then $\mathcal{L}_{\text{cons}}(\theta; \boldsymbol{x}, \phi) = 0$ for every perturbation ϕ and every*

prefix law, in particular for $\pi_1^{f_0}$ realized with such conditionals. $\mathcal{L}_{\text{cons}}$ never compares two different pool templates: both of its arguments carry the same τ_0.

(v) Consistency controls same-delimiter answer drift. *With P^ϕ, P^0 and ℓ_θ as in Appendix A and $\mathcal{A} := \mathcal{A}_{\tau_0}$ ($d_{\tau_\phi} = d_{\tau_0}$),*

$$D_{\mathrm{KL}}\big(\mathcal{A}_\# P^\phi \,\|\, \mathcal{A}_\# P^0\big) \le D_{\mathrm{KL}}(P^\phi \,\|\, P^0) = \mathbb{E}_{\boldsymbol{y}\sim P^\phi}\big[|\boldsymbol{y}|\,\ell_\theta(\boldsymbol{y})\big] \le L_{\max}\,\mathbb{E}_{\boldsymbol{y}\sim P^\phi}\ell_\theta(\boldsymbol{y}).$$

Hence $\mathrm{TV}(\mathcal{A}_\# P^\phi, \mathcal{A}_\# P^0) \le \mathrm{TV}(P^\phi, P^0) \le \sqrt{L_{\max}\mathbb{E}\ell_\theta/2}$, and every reward $R(\boldsymbol{y})$ with range Δ satisfies $|\mathbb{E}_{P^\phi}R - \mathbb{E}_{P^0}R| \le \Delta\sqrt{L_{\max}\mathbb{E}\ell_\theta/2}$, which is $\sqrt{2L_{\max}\mathbb{E}\ell_\theta}$ for R_t, whether or not the answers are correct.

(vi) Sharpness. *(a) A policy whose next-token conditionals do not depend on the instruction block and whose every response $\boldsymbol{y}$ has $\mathcal{A}_\tau(\boldsymbol{y}) = a^*$ for every $\tau \in \mathcal{C}$ (possible only if $\chi(\mathcal{C}) = 1$) has $\zeta = 0$ and $\mathcal{L}_{\text{cons}} = 0$ and satisfies (C2) with $\varepsilon = 0$: it is answer-invariant on the pool, which is what (C2) asks for. (b) At a template whose delimiter family is outside the pool, Proposition B.6 applies verbatim with $\mathcal{T}_{train} = \mathcal{C}$.*

(vii) Finite-ε reward agreement. *For a common answer-level reward $r_{\boldsymbol{x}}$ with range Δ_R and (C2) with a Pinsker-type D, the $\rho_{\boldsymbol{x}}$-averaged pairwise expected-reward difference is at most $\Delta_R\kappa_D\sqrt{\varepsilon}$; with the TV form of (iii) it is at most $2\Delta_R\zeta$.*

Proof. (i). $R_t = +1$ exactly on $\{\mathcal{A}_\tau = a^*\}$, $R_t \le 0$ off it and $R_t = -1$ on $\{\mathcal{A}_\tau = \emptyset\}$, so pointwise $1 - R_t \ge \mathbf{1}\{\mathcal{A}_\tau(\boldsymbol{y}) \ne a^*\}$ and $1 - R_t \ge 2\cdot\mathbf{1}\{\mathcal{A}_\tau(\boldsymbol{y}) = \emptyset\}$; take expectations. $\zeta = 0$ forces $R_t = 1$ almost surely for P_{train}-almost every τ, i.e. for every $\tau \in \mathcal{C}$ by full support. If some policy has $\zeta = 0$, the R_t-optimal policies are exactly those with $\zeta = 0$; they all have $P_\tau^\pi = \delta_{a^*}$ and hence reward $R_{\max}$ at every $\tau \in \mathcal{C}$. The same argument applies to $V = \mathbf{1}\{\mathcal{A}_\tau = a^*\}$, and the optima of R_{DTR}, whose value $1 + \alpha_{\text{rew}}$ is then attained, have $R_t = 1$ and $R_s = 1$ almost surely, a subset of the R_t-optima.

(ii). Pointwise $R_t(\boldsymbol{y}, \tau) \le 2\cdot\mathbf{1}\{\tau \in \mathrm{Acc}(\boldsymbol{y})\} - 1$. For an instruction-ignoring π the response law does not depend on τ, so $\mathbb{E}_\tau\mathbb{E}_{\boldsymbol{y}\sim\mu_{\boldsymbol{x}}}R_t \le \mathbb{E}_{\boldsymbol{y}\sim\mu_{\boldsymbol{x}}}[2P_{\text{train}}(\mathrm{Acc}(\boldsymbol{y})) - 1] \le 2\chi(\mathcal{C}) - 1$. For $\pi_1^{f_0}$, $R_t = +1$ on $\mathcal{C}_{f_0}$ and -1 off it. With τ_B in the pool, π_1 has $R_t = +1$ on $\mathcal{C}\setminus\{\tau_B\}$ and $R_t = -1$ at τ_B, where its format is disjoint from $\mathcal{F}_B$ and the delimiter-separated extractor of Definition A.2 returns $\emptyset$; π_2 has $R_t = +1$ everywhere. For R_{DTR}, $1 + \alpha_{\text{rew}} - \mathbb{E}R_{\mathrm{DTR}} = \zeta + \alpha_{\text{rew}}(1 - \mathbb{E}R_s)$ with $R_s \in [0, 1]$.

(iii). If $P_i(\{a^*\}) \ge 1 - e_i$ ($i = 1, 2$), then for every event S, $P_1(S) - P_2(S) \le 1 - P_2(\{a^*\}) \le e_2$ if $a^* \in S$ and $P_1(S) - P_2(S) \le P_1(S) \le e_1$ otherwise, and symmetrically, so $\mathrm{TV}(P_1, P_2) \le \max\{e_1, e_2\} \le e_1 + e_2$. Average over independent $\tau_1, \tau_2 \sim P_{\text{train}}$ and apply (i); P_{train} is non-degenerate for $K \ge 2$.

(iv). If the conditionals ignore the instruction block, $p_t^\phi = p_t^0$ at every prefix. Both arguments of Eq. (8) are built on $\tau_0(\boldsymbol{x})$.

(v). The first inequality is Lemma A.3, since one deterministic $\mathcal{A}$ reads both laws. Both sequence laws are generated by the same stopping rule (EOS or $L_{\max}$) with full-support softmax conditionals, so $\log(P^\phi(\boldsymbol{y})/P^0(\boldsymbol{y})) = \sum_{t\le|\boldsymbol{y}|}\log(p_t^\phi(y_t)/p_t^0(y_t))$. The event $\{|\boldsymbol{y}| \ge t\}$ is determined by $\boldsymbol{y}_{<t}$, and given $\boldsymbol{y}_{<t}$ the token y_t has law p_t^ϕ under P^ϕ, so the tower property over $t \le L_{\max}$ gives $D_{\mathrm{KL}}(P^\phi\|P^0) = \mathbb{E}_{P^\phi}\sum_{t\le|\boldsymbol{y}|}D_{\mathrm{KL}}(p_t^\phi\|p_t^0) = \mathbb{E}_{P^\phi}[|\boldsymbol{y}|\,\ell_\theta(\boldsymbol{y})]$. Then use $|\boldsymbol{y}| \le L_{\max}$, the data-processing inequality for TV, Pinsker's inequality, and $|\mathbb{E}_P R - \mathbb{E}_Q R| \le \Delta\,\mathrm{TV}(P, Q)$; R_t has range 2.

(vi). (a): $R_t = 1$ at every pool template, so $\zeta = 0$ and $P_\tau^\pi = \delta_{a^*}$ for every $\tau \in \mathcal{C}$, whence every divergence in (C2) vanishes; (iv) gives $\mathcal{L}_{\text{cons}} = 0$. (b) is Proposition B.6.

(vii). For $P_i := P_{\tau_i}^\pi$ and $m_i := \mathbb{E}_{a\sim P_i}r_{\boldsymbol{x}}(a)$, $\mathbb{E}|m_1 - m_2| \le \Delta_R\,\mathbb{E}\,\mathrm{TV}(P_1, P_2) \le \Delta_R\kappa_D\,\mathbb{E}\sqrt{D(P_1\|P_2)} \le \Delta_R\kappa_D\sqrt{\mathbb{E}D(P_1\|P_2)} \le \Delta_R\kappa_D\sqrt{\varepsilon}$ by the Pinsker-type bound, Jensen's inequality and (C2); with $D = \mathrm{TV}$, the first step and (iii) give $2\Delta_R\zeta$. □

Remark B.15 (The released pool, and the division of labor). The pool has four delimited templates, requesting `\boxed{}`, `\answer{}`, `<answer> </answer>` and "`#### Answer:`", and one template without a delimiter whose extractor reads the last number or expression of the response;

each delimited extractor searches only for its own delimiter. A response that uses one of the four delimiters is therefore accepted by at most two pool templates, its own and the no-delimiter one, so every instruction-ignoring policy whose responses use at most one delimiter has $\zeta \geq 2(1 - \frac{2}{5}) = \frac{6}{5}$ (it loses at least 1.2 of the maximal 2 reward units), and a response without any of the four delimiters is accepted only by the no-delimiter template ($\zeta \geq \frac{8}{5}$). A response accepted by all five templates states the answer under all four delimiters; with a^* under every delimiter and as its last number or expression, it is answer-invariant on the pool and satisfies (C2) exactly (Proposition B.14(vi)(a)). The two ingredients of PIRL thus act on different template pairs: the pool reward enforces answer agreement across delimiters inside the pool, by (i)–(iii), and $\mathcal{L}_{\text{cons}}$ enforces it between each pool prompt and its same-delimiter embedding neighbors, by (v). Templates between and beyond the pool points are handled by Proposition B.16 and Lemma B.17.

B.7 Transfer to unseen templates

The reward shift from the pool to unseen templates of one grader class is bounded by an optimal-transport cost that is informative for disjoint templates. Fix $(\boldsymbol{v}, \boldsymbol{q}, a^*)$. A finite set $\mathcal{K}$ of templates is a *grader class* if the evaluated reward is one function of the response on all of it: $R(\boldsymbol{y}, \tau(\boldsymbol{x})) = R_{\mathcal{K}}(\boldsymbol{y}) \in [R_{\min}, R_{\max}]$ for $\tau \in \mathcal{K}$. Under strict delimiter extraction a grader class is a delimiter family; under a delimiter-agnostic grader (a content judge, or a parser that tries every delimiter) all templates form one class. Put $r_\pi(\tau) := \mathbb{E}_{\boldsymbol{y}\sim\pi(\cdot|\tau(\boldsymbol{x}))} R_{\mathcal{K}}(\boldsymbol{y})$ and $\Delta_R := R_{\max} - R_{\min}$. Let $\boldsymbol{E}_\tau := \boldsymbol{e}_{\mathcal{I}}(\tau(\boldsymbol{x})) \in \mathbb{R}^{n_\tau \times h}$ be the instruction-token embedding block, and let $d_e(\tau, \tau') := \|\boldsymbol{E}_\tau - \boldsymbol{E}_{\tau'}\|_F$ if $\tau(\boldsymbol{x})$ and $\tau'(\boldsymbol{x})$ coincide outside instruction blocks of equal length $n_\tau = n_{\tau'}$, and $d_e(\tau, \tau') := +\infty$ otherwise.

Proposition B.16 (Two-scale transfer bound)**.** *Let $P, Q \in \Delta(\mathcal{K})$ and let d be any (possibly $+\infty$-valued) pseudometric on $\mathcal{K}$, such as d_e. Let $L_\pi \geq 0$ satisfy $|r_\pi(\tau) - r_\pi(\tau')| \leq L_\pi\, d(\tau, \tau')$ for all $\tau \in \operatorname{supp} Q$ and $\tau' \in \operatorname{supp} P$ (with $0 \cdot \infty := \infty$). With the truncated cost $c_L(\tau, \tau') := \min\{L_\pi\, d(\tau, \tau'), \Delta_R\}$,*

$$\big|\mathcal{R}(\pi, P; \boldsymbol{x}) - \mathcal{R}(\pi, Q; \boldsymbol{x})\big| \;\leq\; W_{c_L}(Q, P) \;\leq\; \min\big\{\Delta_R\, \mathrm{TV}(P, Q),\; L_\pi\, W_1^d(Q, P)\big\}. \tag{15}$$

$W_{c_L}(Q, P)$ equals $\sup\{\mathbb{E}_P f - \mathbb{E}_Q f :\ |f(\tau) - f(\tau')| \leq c_L(\tau, \tau')\ \forall \tau, \tau' \in \operatorname{supp} P \cup \operatorname{supp} Q\}$, and every such f satisfies the hypotheses on r_π after a shift, so no better bound follows from (L_π, Δ_R, d) alone. When P and Q have disjoint supports, $\mathrm{TV} = 1$ and the bound is $\min\{\Delta_R, L_\pi W_1^d\}$, informative whenever $L_\pi W_1^d < \Delta_R$.

Proof. A coupling γ of (Q, P) is concentrated on $\operatorname{supp} Q \times \operatorname{supp} P$, where $|r_\pi(\tau') - r_\pi(\tau)| \leq c_L(\tau, \tau')$, so $|\mathbb{E}_P r_\pi - \mathbb{E}_Q r_\pi| \leq \mathbb{E}_\gamma c_L$; take the infimum over γ. The maximal coupling has $\gamma(\tau \neq \tau') = \mathrm{TV}(P, Q)$ and $c_L \leq \Delta_R$, with $c_L = 0$ on the diagonal; a W_1^d-optimal coupling costs at most $L_\pi W_1^d$. As $\min\{a + b, \Delta\} \leq \min\{a, \Delta\} + \min\{b, \Delta\}$, c_L is a pseudometric, and on the finite space $\operatorname{supp} P \cup \operatorname{supp} Q$ Kantorovich–Rubinstein (finite LP) duality identifies W_{c_L} with the stated supremum; such an f has oscillation at most Δ_R and satisfies the cross-pair Lipschitz condition. □

The TV bound is attained with non-degenerate coverage: for $Q = (1 - w)\delta_{\tau_1} + w\delta_{\tau_2}$ and $P = (1 - w - t)\delta_{\tau_1} + w\delta_{\tau_2} + t\delta_{\tau_3}$ ($w \in (0, 1)$, $0 < t \leq 1 - w$), a policy with identical answer marginals at τ_1, τ_2 (so 0-C2-feasible on $\operatorname{supp} Q$), reward $R_{\max}$ there and $R_{\min}$ at τ_3 has $\mathrm{TV}(P, Q) = t$ and reward shift exactly $t\Delta_R$.

B.8 What the adversary measures

The adversary ascends the one-sided drop of the reward over the ϵ_{emb}-ball, and a bound on this drop along paths from the pool takes the role of L_π in Proposition B.16. Fix a template τ of the grader class. For $\boldsymbol{E} \in \mathbb{R}^{n_\tau \times h}$ let $g_\pi(\boldsymbol{E}) := \mathbb{E}_{\boldsymbol{y}\sim\pi(\cdot|\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{E})} R_{\mathcal{K}}(\boldsymbol{y})$ be the reward when the instruction block of $\tau(\boldsymbol{x})$ carries the input embeddings $\boldsymbol{E}$, so that $r_\pi(\tau') = g_\pi(\boldsymbol{E}_{\tau'})$ whenever $d_e(\tau, \tau') < \infty$, and let $h_\pi(\boldsymbol{E}; \epsilon) := g_\pi(\boldsymbol{E}) - \inf_{\|\boldsymbol{\delta}\|_F \leq \epsilon} g_\pi(\boldsymbol{E} + \boldsymbol{\delta}) \geq 0$ be the one-sided ϵ-drop.

Lemma B.17 (Adversary-calibrated modulus)**.** *(a)* The adversary's objective is the drop. *Let $R_{\mathcal{K}} = R_{\mathrm{DTR}}$ read with the extractor of τ_0; it does not depend on $\boldsymbol{\delta}$, since the extractor stays at d_{τ_0} and R_s reads only $\boldsymbol{y}$ and a^*. At $\tau_0(\boldsymbol{x})$, the inner problem of Eq. (16) with a point-mass sampler has value $h_\pi(\boldsymbol{E}_{\tau_0}; \epsilon_{\text{emb}}) - g_\pi(\boldsymbol{E}_{\tau_0})$. For $G_{\text{adv}} \geq 2$ i.i.d. rollouts from the law $\pi(\cdot \mid \boldsymbol{v}, \boldsymbol{q}, \boldsymbol{E}_{\tau_0} + \boldsymbol{\delta})$*

whose score is differentiated (the release scores the temperature-0.9 softmax), the leave-one-out estimator of Eq. (17) is unbiased for $-\nabla_{\boldsymbol{\delta}} g_\pi(\boldsymbol{E}_{\tau_0} + \boldsymbol{\delta})$, with no pathwise term, and the realized drop $g_\pi(\boldsymbol{E}_{\tau_0}) - g_\pi(\boldsymbol{E}_{\tau_0} + \bar{\boldsymbol{\delta}})$ of the returned perturbation is a lower bound on $h_\pi(\boldsymbol{E}_{\tau_0}; \epsilon_{\mathrm{emb}})$.

(b) The drop is the local Lipschitz constant. *If g_π is C^1 with β_s-Lipschitz gradient on the ball $\mathbb{B}(\boldsymbol{E}, \epsilon)$, then*

$$\epsilon \|\nabla g_\pi(\boldsymbol{E})\|_F - \tfrac{\beta_s}{2}\epsilon^2 \;\le\; h_\pi(\boldsymbol{E}; \epsilon) \;\le\; \epsilon \sup_{\mathbb{B}(\boldsymbol{E}, \epsilon)} \|\nabla g_\pi\|_F.$$

For the untruncated softmax response law at any temperature, g_π is C^∞ in $\boldsymbol{E}$, so β_s exists on every ball.

(c) One radius. *If $d_e(\tau, \tau') \le \epsilon$, then $r_\pi(\tau') \ge r_\pi(\tau) - h_\pi(\boldsymbol{E}_\tau; \epsilon)$.*

(d) Chaining. *Let $d_e(\tau, \tau') < \infty$, $N := \lceil d_e(\tau, \tau')/\epsilon \rceil$ and $\boldsymbol{E}_k := \boldsymbol{E}_\tau + \frac{k}{N}(\boldsymbol{E}_{\tau'} - \boldsymbol{E}_\tau)$. If $h_\pi(\boldsymbol{E}_k; \epsilon) \le \omega$ for $k = 0, \dots, N-1$, then $r_\pi(\tau') \ge r_\pi(\tau) - N\omega$. If $h_\pi(\cdot\,; \epsilon) \le \omega$ on the whole segment, then $|r_\pi(\tau) - r_\pi(\tau')| \le \omega \lceil d_e/\epsilon \rceil \le (\omega/\epsilon)\, d_e + \omega$.*

(e) Templates of a different length. *Let $\xi \in [0, 1]$ and let $\tilde{\boldsymbol{E}} \in \mathbb{R}^{n_\tau \times h}$ be a ξ-proxy of τ', i.e. $\mathrm{TV}\big(\pi(\cdot \mid \boldsymbol{v}, \boldsymbol{q}, \tilde{\boldsymbol{E}}), \pi(\cdot \mid \tau'(\boldsymbol{x}))\big) \le \xi$, at distance $\varrho := \|\tilde{\boldsymbol{E}} - \boldsymbol{E}_\tau\|_F$. If the drop is at most ω at the $\lceil \varrho/\epsilon \rceil$ chaining points from $\boldsymbol{E}_\tau$ to $\tilde{\boldsymbol{E}}$, then $r_\pi(\tau') \ge r_\pi(\tau) - \omega\lceil \varrho/\epsilon \rceil - \Delta_R\, \xi$.*

Proof. (a). With a point mass at $\boldsymbol{\delta}$, Eq. (16) is $\max_{\|\boldsymbol{\delta}\|_F \le \epsilon} -g_\pi(\boldsymbol{E} + \boldsymbol{\delta}) = h_\pi - g_\pi(\boldsymbol{E})$. As $R_{\mathcal{K}}$ does not depend on $\boldsymbol{\delta}$ and the responses of length at most $L_{\max}$ form a finite set, $\nabla_{\boldsymbol{\delta}} g_\pi = \mathbb{E}[R\, \nabla_{\boldsymbol{\delta}} \log \pi(\boldsymbol{y} \mid \boldsymbol{E} + \boldsymbol{\delta})]$. $\bar{R}_{-i}$ is independent of $\boldsymbol{y}_i$ and $\mathbb{E}\nabla \log \pi(\boldsymbol{y}_i) = \nabla \sum_{\boldsymbol{y}} \pi(\boldsymbol{y}) = 0$, so $\mathbb{E}[(\bar{R}_{-i} - R_i)\nabla \log \pi(\boldsymbol{y}_i)] = -\nabla_{\boldsymbol{\delta}} g_\pi$ for every i. The returned $\bar{\boldsymbol{\delta}}$ lies in the ball.

(b). If $\nabla g_\pi(\boldsymbol{E}) \ne 0$, take $\boldsymbol{\delta} = -\epsilon \nabla g_\pi / \|\nabla g_\pi\|_F$ in the descent lemma $g(\boldsymbol{E} + \boldsymbol{\delta}) \le g(\boldsymbol{E}) + \langle \nabla g, \boldsymbol{\delta} \rangle + \frac{\beta_s}{2}\|\boldsymbol{\delta}\|_F^2$; otherwise the lower bound is nonpositive. The upper bound is the mean-value theorem on segments inside the ball. g_π is a finite sum over responses of products of softmax outputs of a network that is smooth in its input embeddings (linear embedding input, attention, SiLU MLPs, RMSNorm with a positive stabilizer, rotary encodings).

(c). $\boldsymbol{E}_{\tau'} = \boldsymbol{E}_\tau + \boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_F \le \epsilon$ and the prompts coincide outside the block, so $r_\pi(\tau') = g_\pi(\boldsymbol{E}_\tau + \boldsymbol{\delta}) \ge g_\pi(\boldsymbol{E}_\tau) - h_\pi(\boldsymbol{E}_\tau; \epsilon)$.

(d). $\|\boldsymbol{E}_{k+1} - \boldsymbol{E}_k\|_F = d_e/N \le \epsilon$, so $g_\pi(\boldsymbol{E}_{k+1}) \ge g_\pi(\boldsymbol{E}_k) - \omega$ as in (c); sum over k. The two-sided statement chains from either end, and $\lceil a \rceil \le a + 1$.

(e). Chain from $\boldsymbol{E}_\tau$ to $\tilde{\boldsymbol{E}}$ as in (d) and use $|g_\pi(\tilde{\boldsymbol{E}}) - r_\pi(\tau')| \le \Delta_R \xi$, since one $R_{\mathcal{K}}$ scores both response laws. □

Corollary B.18 (Transfer from the pool, calibrated by the adversary)**.** *Let Q be supported on pool (or Standard) templates and P on unseen templates of the same grader class, and assume the drop bounds of Lemma B.17(d) and (e). For every coupling γ of (Q, P),*

$$\mathcal{R}(\pi, P; \boldsymbol{x}) \;\ge\; \mathcal{R}(\pi, Q; \boldsymbol{x}) - \mathbb{E}_{(\tau, \tau') \sim \gamma} \min\big\{\Delta_R,\; c^+(\tau, \tau')\big\},$$

where $c^+(\tau, \tau') = \omega\lceil d_e(\tau, \tau')/\epsilon \rceil$ if $d_e < \infty$ and $c^+ = \omega\lceil \varrho_\xi/\epsilon \rceil + \Delta_R \xi$ otherwise, with ϱ_ξ the distance of a ξ-proxy of the unseen template. If $h_\pi(\cdot\,; \epsilon) \le \omega$ on all segments between $\operatorname{supp} Q$ and $\operatorname{supp} P$ and all these pairs have $d_e < \infty$, then

$$\big|\mathcal{R}(\pi, P; \boldsymbol{x}) - \mathcal{R}(\pi, Q; \boldsymbol{x})\big| \le \min\big\{\Delta_R\, \mathrm{TV}(P, Q),\; (\omega/\epsilon)\, W_1^{d_e}(Q, P) + \omega\big\},$$

so the adversary's drop per radius, ω/ϵ, takes the role of L_π in Proposition B.16. With Q the Standard template, P the law of T-Stress templates, R the evaluated accuracy and π the greedy evaluation decoder, $\mathcal{R}(\pi, P; \boldsymbol{x}) - \mathcal{R}(\pi, Q; \boldsymbol{x})$ averaged over instances is the T-Stress gap.

Proof. Pointwise, Lemma B.17(d)–(e) give $r_\pi(\tau') \ge r_\pi(\tau) - c^+(\tau, \tau')$, and $r_\pi(\tau') \ge r_\pi(\tau) - \Delta_R$ always; integrate against γ. For the two-sided form, the cost $\min\{\Delta_R, \omega\lceil d_e/\epsilon \rceil\}$ bounds $|r_\pi(\tau) - r_\pi(\tau')|$, vanishes on the diagonal, is at most Δ_R and at most $(\omega/\epsilon)d_e + \omega$; use the maximal coupling and a $W_1^{d_e}$-optimal coupling as in Proposition B.16. □

Remark B.19 (How training moves ω). At every refresh the adversary restarts from $\boldsymbol{\delta} = \mathbf{0}$ and ascends the drop at the pool points at radius ϵ_{emb} (Lemma B.17(a)); the GRPO loss is taken at the returned perturbation, which raises g_π where the drop is realized, and for $R_{\mathcal{K}} = R_t$ the consistency term bounds the realized drop of the untempered law by $\sqrt{2L_{\max}\mathbb{E}_{P^\phi}\ell_\theta}$ (Proposition B.14(v)). The drop bound ω along chaining paths turns this local signal into Corollary B.18; its scale is the number $\lceil d_e/\epsilon_{\text{emb}} \rceil$ (or $\lceil \varrho_\xi/\epsilon_{\text{emb}} \rceil$) of adversary radii between a test template and the pool, the first computable offline from the input-embedding matrix.

Lemma B.20 (When the adversary's estimator vanishes)**.** *At a fixed perturbation with i.i.d. rollouts $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{G_{\text{adv}}}$, the estimator of Eq. (17) is exactly 0 on the event $\{R_1 = \cdots = R_{G_{\text{adv}}}\}$. Under the binary reward this event has probability $\mu^{G_{\text{adv}}} + (1-\mu)^{G_{\text{adv}}}$ (0.66 at $\mu = 0.9$, $G_{\text{adv}} = 4$). Under R_{DTR} with $\alpha_{\text{rew}} < 1$, the event of equal R_{DTR} is contained in the event of equal V, and the difference is the event that V is constant but the group mixes W and F, or has constant B and non-constant R_s: the DTR adversary is silent no more often than the binary one, and strictly less often whenever that difference has positive probability.*

Proof. If all R_i are equal, every coefficient $\bar{R}_{-i} - R_i$ vanishes. For $\alpha_{\text{rew}} < 1$ the three trinary levels map to the disjoint intervals $[-1, -1+\alpha_{\text{rew}}]$, $[0, \alpha_{\text{rew}}]$ and $[1, 1+\alpha_{\text{rew}}]$, so equal R_{DTR} forces equal R_t, hence equal B and equal $V = \widetilde{V}_c$; conversely, equal B and equal R_s give equal R_{DTR}. □

C Embedding-Space Adversary Details

The adversary of Section 3.2 perturbs only the instruction-token input embeddings of the prompt $\boldsymbol{x}$ (after the template draw τ_0): $\tilde{\boldsymbol{e}}_{\mathcal{I}}(\boldsymbol{x}) = \boldsymbol{e}_{\mathcal{I}}(\boldsymbol{x}) + \boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_F \le \epsilon_{\text{emb}}$, where $\|\cdot\|_F$ is the Euclidean norm of the flattened $n_\tau \times h$ block (Lemma B.17). The perturbed prompt $\tau_\phi(\boldsymbol{x})$ carries $\tilde{\boldsymbol{e}}_{\mathcal{I}}(\boldsymbol{x})$; its image, question, answer target and extractor delimiter $d_{\tau_\phi} = d_{\tau_0}$ are those of the clean prompt. Every prompt has its own perturbation ($\phi = \boldsymbol{\delta}$, q_ϕ the point mass at it), so Eq. (9) decouples into the per-prompt problem

$$\max_{\|\boldsymbol{\delta}\|_F \le \epsilon_{\text{emb}}} \; -\mathbb{E}_{\boldsymbol{y}\sim\pi_\theta(\cdot|\tau_\phi(\boldsymbol{x}))} R_{\text{DTR}}(\boldsymbol{y}, \tau_\phi(\boldsymbol{x})), \tag{16}$$

which Algorithm 1 solves by K_{adv} projected ascent steps from $\boldsymbol{\delta} = \mathbf{0}$, restarted at every adversary refresh; Algorithm 2 is the full training loop.

Estimator. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{G_{\text{adv}}}$, $G_{\text{adv}} > 1$, be rollouts under $\tau_\phi(\boldsymbol{x})$ of the policy, which is frozen at $\theta = \theta_{\text{old}}$ during a refresh. Since R_{DTR} reads only $\boldsymbol{y}$, a^* and the delimiter of τ_0, none of which depends on $\boldsymbol{\delta}$, the gradient of Eq. (16) is estimated by the leave-one-out score-function estimator

$$\hat{\boldsymbol{g}} = \frac{1}{G_{\text{adv}}} \sum_{i=1}^{G_{\text{adv}}} (\bar{R}_{-i} - R_i)\, \nabla_{\boldsymbol{\delta}} \log \pi_\theta(\boldsymbol{y}_i \mid \tau_\phi(\boldsymbol{x})), \tag{17}$$

with $R_i = R_{\text{DTR}}(\boldsymbol{y}_i, \tau_\phi(\boldsymbol{x}))$ and $\bar{R}_{-i} = \frac{1}{G_{\text{adv}}-1}\sum_{j\neq i} R_j$. It is unbiased for $-\nabla_{\boldsymbol{\delta}}$ of the expected reward, has no pathwise term (Lemma B.17(a)), and is exactly zero on a group with constant reward (Lemma B.20).

Hyperparameters. At every refresh, $\epsilon_{\text{emb}} = 0.1 \cdot \text{median}_{\boldsymbol{x}} \|\boldsymbol{e}_{\mathcal{I}}(\boldsymbol{x})\|_F$ over the instruction-embedding blocks of the refresh batch. The adversary uses $K_{\text{adv}} = 4$ ascent steps, $G_{\text{adv}} = 4$ rollouts per step (the policy group is $G = 16$), and step size $\eta_{\text{adv}} = 5 \times 10^{-5}$ on the raw estimate $\hat{\boldsymbol{g}}$. Each refresh (all K_{adv} steps for every prompt of the refresh batch) is followed by $C_{\text{adv}} = 4$ policy updates.

Why embedding-space. Perturbing the instruction embedding is a differentiable, model-agnostic relaxation of the worst-case rewording in $[\![\boldsymbol{x}]\!]$ that needs no candidate enumeration. The perturbation is continuous, and its nearest-token projection is a near-paraphrase of the instruction (Appendix L).

Compute cost. Per refresh and prompt, the adversary samples $K_{\text{adv}}G_{\text{adv}} = 16$ rollouts in addition to the $G = 16$ policy rollouts, so PIRL samples 32 sequences per prompt where GRPO samples 16; each ascent step runs one forward–backward pass over its G_{adv} prompt–response sequences to obtain $\nabla_{\boldsymbol{\delta}} \log \pi_\theta$. No adversary rollout enters the policy update, and the K_{adv} forward–backward passes of a refresh are shared by its $C_{\text{adv}} = 4$ policy updates.

Algorithm 1 Embedding-space adversary: K_{adv} projected-ascent steps (per prompt $\boldsymbol{x}$)

Input: prompt $\boldsymbol{x}$ (carrying the drawn template τ_0), radius ϵ_{emb}, ascent steps K_{adv}, adversary group size G_{adv}, step size η_{adv}, rollout policy $\pi_{\theta_{\text{old}}}$.
Initialize $\boldsymbol{\delta} \leftarrow \mathbf{0}$.
for $j = 1, \ldots, K_{\text{adv}}$ **do**
  Sample $\{\boldsymbol{y}_i\}_{i=1}^{G_{\text{adv}}} \sim \pi_{\theta_{\text{old}}}(\cdot \mid \tau_\phi(\boldsymbol{x}))$ at the current $\boldsymbol{\delta}$; score $R_i = R_{\text{DTR}}(\boldsymbol{y}_i, \tau_\phi(\boldsymbol{x}))$ at the delimiter of τ_0.
  Estimate $\hat{\boldsymbol{g}}$ by Eq. (17).
  Ascend and project: $\boldsymbol{\delta} \leftarrow \Pi_{\|\cdot\|_F \le \epsilon_{\text{emb}}}(\boldsymbol{\delta} + \eta_{\text{adv}}\, \hat{\boldsymbol{g}})$.
end for
Return $\boldsymbol{\delta}$; the perturbed prompt $\tau_\phi(\boldsymbol{x})$ carries the instruction embedding $\boldsymbol{e}_{\mathcal{I}}(\boldsymbol{x}) + \boldsymbol{\delta}$.

Algorithm 2 PIRL training loop

Input: policy π_θ, dataset $\mathcal{D}$, template pool P_{train}, adversary settings $(K_{\text{adv}}, G_{\text{adv}}, \eta_{\text{adv}})$, policy group size G, refresh cadence C_{adv}, consistency weight λ.
for each adversary refresh **do**
  $\theta_{\text{old}} \leftarrow \theta$; sample a refresh batch $\mathcal{B} = \{(\boldsymbol{x}, a^*)\} \sim \mathcal{D}$.
  For each $\boldsymbol{x} \in \mathcal{B}$, draw $\tau_0 \sim P_{\text{train}}$ (the single template in the no-MT ablation) and write $\boldsymbol{x}$ for $\tau_0(\boldsymbol{x})$.
  $\epsilon_{\text{emb}} \leftarrow 0.1 \cdot \text{median}_{\boldsymbol{x} \in \mathcal{B}} \|\boldsymbol{e}_{\mathcal{I}}(\boldsymbol{x})\|_F$.
  for each $\boldsymbol{x} \in \mathcal{B}$ **do**
    $\boldsymbol{\delta}_{\boldsymbol{x}} \leftarrow$ Algorithm 1 on $\boldsymbol{x}$; freeze $\tau_{\bar{\phi}_{\boldsymbol{x}}}(\boldsymbol{x})$, the prompt with instruction embedding $\boldsymbol{e}_{\mathcal{I}}(\boldsymbol{x}) + \boldsymbol{\delta}_{\boldsymbol{x}}$.
    Sample $\{\tilde{\boldsymbol{y}}_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid \tau_{\bar{\phi}_{\boldsymbol{x}}}(\boldsymbol{x}))$ (the rollouts $\boldsymbol{y}$ of Eq. (8)); compute R_t (Eq. (5)) at the delimiter of τ_0 and the delimiter-free R_s, and the group scores Z_i (first line of Eq. (7)).
  end for
  $\hat{A}_i \leftarrow \text{clip}(\text{Norm}_{\mathcal{B}}(Z_i), -5, 5)$ over all rollouts of the refresh batch.
  for $c = 1, \ldots, C_{\text{adv}}$ **do**
    Update θ by one step on Eq. (10) over the c-th of C_{adv} shuffled minibatches of the stored rollouts, with the perturbed prompts fixed and $\mathcal{L}_{\text{cons}}$ teacher-forced on $\tilde{\boldsymbol{y}}$ against the stop-gradient clean prompt $\tau_0(\boldsymbol{x})$ (Eq. (8)).
  end for
end for

D Dynamic Evaluation: Operators and Label Handling

D-Test evaluates every model on label-tracked input mutations of the evaluation instances, in the dynamic-evaluation framing of Yang et al. (2025); the mutations are stored as an overlay that the benchmark loader reads in place of the original instances. Table 3 lists the operators as implemented, with the label each one assigns.

Configured operators and routing. The configurations use the five text operators of Table 3; the Exam benchmarks additionally use all six image operators, and the Medical D-Test is text-only. Each mutated instance receives exactly one text operator, drawn uniformly from the operators of its path, and on the Exam benchmarks one image operator per image. Each mutated instance therefore combines a content mutation with an instruction change. The harness also contains a QA↔MC type-transition operator, which is disabled in all our configurations.

E Evaluation Protocol

Every model is evaluated with the same prompts, scorers and decoding settings (Tables 4–6, Boxes 1–2); D-Test is specified in Appendix D.

Standard and T-Stress prompts. On VQA-RAD, PathVQA and GMAI-MMBench the instruction is drawn per item, uniformly from the first N entries of the list for the item's question type: the

Table 3: D-Test operators as implemented, and how each sets the label of the mutated instance. MC: multiple-choice instances.

Operator	Input	Implementation	Label
Rephrase	text	GPT-4.1-nano rewrites the question, instructed to make it more complex and longer while preserving its meaning	unchanged
Narrative distraction	text	GPT-4.1-nano generates an irrelevant background sentence that is appended to the question	unchanged
Option replacement	text (MC)	options randomly permuted	key letter follows the correct option
Answer negation	text (MC)	the text of the correct option is replaced by "None of the options are correct answers"	key letter kept: the key is the "None of the options" option
Choice expansion	text (MC)	options padded to six with fillers ("None of the above", "All of the above", "Insufficient information", "Cannot be determined from the context", "Both A and B", "Both C and D")	key letter kept
Resize	image	rescale to shorter side $\geq$ 448 px, black padding to a square	unchanged
Crop	image	center crop to 90%	unchanged
Rotation	image	15°, canvas expanded, black fill	unchanged
Salt-and-pepper	image	1% salt and 1% pepper pixels	unchanged
Color inversion	image	pixel-wise inversion	unchanged
Style transfer	image	OpenCV stylization filter	unchanged

Table 4: Benchmarks as run by the evaluation harness: split and number of items, question types, Standard prompt, and scoring. "Extracted answer": the content of `\boxed{}` or of `<answer>...</answer>` when present, the full response otherwise. MC/OE entry 0: entry 0 of the lists in Table 5. Judge prompts are given in Boxes 1–2 below.

Benchmark	Split (#items)	Question types	Standard prompt	Scoring (correct if rule *or* judge accepts)
MMK12	test; 2,000 (500 each math, physics, chemistry, biology)	open-ended	the question alone (no instruction)	extracted answer (`<answer>` first, then last `\boxed{}`); rule: letter, exact, numeric or normalized-string match; judge (Box 2) on every item
MathVista	testmini; 1,000	540 MC, 460 free-form	the MathVista query (hint, question, and choices for MC items), then `Please provide your answer enclosed in \boxed{}.`	MC: option letter or unique option text inferred from the response, else GPT-4.1-nano extraction with in-context examples; correct iff it is the key. Free-form: integer/float parse of the response, else GPT-4.1-nano extraction; exact numeric equality
Olym-Phys	OlympiadBench `OE_MM_physics_en_COMP`; 456	open-ended	competition preamble ending `So the final answer is \boxed{answer}.`, answer-type hint, context, question	extracted answer (`<answer>` first, then last `\boxed{}`); judge only (Box 2, with the question and the reference final answer)
VQA-RAD	test; 451	251 yes/no, 200 open	yes/no: fixed yes/no instruction; open: OE entry 0	extracted answer, lowercased. Yes/no: exactly one of the substrings `yes`/`no` occurs and it is the reference; open: exact match after normalization. Judge (Box 1) on every rule-failed item
PathVQA	test (deduplicated Hugging Face release); 6,719	3,362 yes/no, 3,357 open	as VQA-RAD	rules as VQA-RAD; judge (Box 1) on rule-failed open items; yes/no items are rule-only
GMAI-MMBench	VAL; 4,550	MC (A–E)	`Question:` question, `Options:` lettered options, then MC entry 0	rule only: option letter from the first or last paragraph, answer patterns, option letter or option text, else the option closest in string similarity

multiple-choice list for GMAI-MMBench and the open-ended list for the open questions of VQA-RAD and PathVQA (Table 5). The draw uses Python's random generator, seeded with seed + rank before each benchmark is loaded; items are sharded round-robin over ranks. Standard is N=1: every item receives entry 0. T-Stress is N=15: entries 0–14. Yes/no items receive the fixed instruction `Please answer with "yes" or "no" (no extra output).` Every instruction is appended to the question after a blank line, and no system prompt is used. The multiple-choice list varies the phrasing and the requested answer form (Table 5, right column).

Training templates. The Exam pool P_{train} of GRPO-MT and PIRL has the five instructions of Table 6, each appended to the question after a blank line and drawn uniformly; the trinary verifier

Table 5: The first 15 English instruction entries of the two T-Stress lists, verbatim (each is preceded by a blank line). Standard uses entry 0; T-Stress draws uniformly from entries 0–14. Entry 10 of the multiple-choice list is, as implemented, two sentences separated by a blank line.

	Multiple choice	**Requested form**
0	Put your final single letter answer in `\boxed{}`.	`\boxed{}`
1	Answer with the option's letter from the given choices directly in `\answer{}`.	`\answer{}`
2	Give your answer as the letter of the correct choice after `#### Answer:`	`#### Answer:`
3	To respond, transcribe ONLY one uppercase letter matching your selected option, and write the letter after `#### Answer:`	`#### Answer:`
4	Reply ONLY with the capital letter representing your selection. No explanation necessary.	letter only
5	Type the letter (A, B, C, etc.) of your answer without extra words or punctuation.	letter only
6	For your reply, render nothing except the isolated letter of the correct option.	letter only
7	Place the letter corresponding to your answer at the start of your reply. Refrain from further commentary.	letter first
8	Teleport yourself into the mind of the examiner, and submit as your answer a single letter for your chosen option, on a line by itself.	letter on its own line
9	Imagine that brevity is a virtue: express your choice solely as the letter label, no rationale, no period.	letter only
10	Output your choice by writing its letter between double asterisks, like `**B**`. ⟨blank line⟩ Please select the correct option by providing its letter.	`**X**`
11	Choose the right answer and respond with the corresponding letter.	letter
12	Indicate your answer by stating the letter of the correct option.	letter
13	Provide the letter that corresponds to the correct answer.	letter
14	Select the appropriate option and write its letter as your answer.	letter

	Open-ended (no delimiter requested)		
0	Please answer the question concisely.	8	Give a brief answer to the question.
1	Provide a brief and clear answer to the question.	9	Provide a short and clear response.
2	Answer the question in a concise manner.	10	Answer concisely and clearly.
3	Please give a concise response to the question.	11	Please provide a brief answer to the question.
4	Respond to the question briefly and clearly.	12	Reply succinctly and do not provide unnecessary details.
5	Provide a clear and concise answer.	13	Focus on clarity: your answer should be as brief as possible while remaining complete.
6	Answer the question briefly.	14	Craft a response that gets straight to the point — no extra explanation needed.
7	Please respond with a concise answer.		

Table 6: The five Exam training templates (verbatim) with the delimiter they request and the answer $a_\tau(\boldsymbol{y})$ extracted by the trinary verifier.

Instruction	**Delimiter**	**Extracted answer $a_\tau(\boldsymbol{y})$**
Put your final answer in `\boxed{}`.	`\boxed{}`	content of the last non-empty `\boxed{}` (brace-aware)
Provide only the answer in `\answer{}`.	`\answer{}`	content of the last non-empty `\answer{}` (brace-aware)
Give your answer in `<answer> </answer>`.	`<answer>...</answer>`	content of the last non-empty `<answer>` element
Write the answer after `#### Answer:`	`#### Answer:`	text after the last marker on its line, else the next non-empty line
Your final answer should be a number or mathematical expression.	none	last number or numeric expression (fraction, root, π multiple, coordinate pair) in the response

(Eq. (5)) extracts $a_\tau(\boldsymbol{y})$ at the delimiter requested by the drawn template and compares it with a^* after normalization or numeric comparison. On MMK12 the Standard prompt is the bare question, so it matches no training template; Standard is itself a template shift for every trained model, including single-template GRPO.

LLM judge. The judge is GPT-4.1-nano (2025-04-14), queried at temperature 0. VQA-RAD and PathVQA use the prompt of Box 1 (correct iff the returned `<judge>` tag is 0), and MMK12 and Olym-Phys the Yes/No prompt of Box 2. On MathVista the same model only extracts an answer when rule inference fails; GMAI-MMBench uses no judge.

Box 1: judge prompt for VQA-RAD and PathVQA (verbatim).

```
Your task is to determine whether the user's answer is correct based on the provided
questions and standard answers (for example, if the user expresses a similar meaning to
the standard answer, or another interpretation of the standard answer, it is considered
correct.)

The question is:  {question}

The standard answer:  {answer}

The user's answer:  {response}

Please strictly follow the following format for output(0 represents correct, 1 represents
incorrect):
<think>{your concise think step}</think>
<judge>{0/1}</judge>

for example:
<think>The standard answer is right, and the user's answer is right frontal lobe, they
express the same meaning, so it is correct.</think>
<judge>0</judge>
```

Box 2: judge prompt for MMK12 (verbatim; Olym-Phys uses a variant that calls the model output the "solution").

```
You are given a question, the correct answer and a model's answer.  Please determine if the
model's answer matches the correct answer.
Focus only on the mathematical or semantic correctness of the content.  Ignore any
differences in formatting, such as LaTeX syntax, symbols, styles, or additional wrappers
(e.g., \boxed, $...$, or similar).  Compare only the core mathematical or textual meaning of
the model's answer and the correct answer.
The process or reasoning leading to the Solution is irrelevant, Only the correctness of the
model's answer matters.
Return only "Yes" if the model's answer is correct or "No" if it is incorrect.
Only return "Yes" or "No" with no additional text or formatting.

Question:
{question}
--------------------------------
Correct Answer:
{answer}
--------------------------------
Model's Answer:
{solution}
--------------------------------
```

Decoding. Responses are generated with vLLM by greedy decoding (temperature 0, top-p 0.01, repetition penalty 1.0) with at most 1,024 new tokens and one response per question (pass@1); only Pass@8 (Appendix I) samples eight responses per question.

F Supplementary Method Details

F.1 GRPO Advantage and the Clipped Policy Loss

For a prompt $\boldsymbol{x}$, GRPO samples a group $\boldsymbol{Y} = \{\boldsymbol{y}^{(i)}\}_{i=1}^{G}$ from the rollout policy $\pi_{\theta_{\text{old}}}$ and assigns rollout i the group-relative advantage

$$A^{(i)} = \frac{r^{(i)} - \mu_r}{\sigma_r + \epsilon_{\text{norm}}}, \qquad \mu_r = \frac{1}{G}\sum_{j=1}^{G} r^{(j)}, \qquad \sigma_r^2 = \frac{1}{G}\sum_{j=1}^{G}\big(r^{(j)} - \mu_r\big)^2, \tag{18}$$

with $r^{(i)} = V(\boldsymbol{y}^{(i)}, \boldsymbol{x})$ and a small constant $\epsilon_{\text{norm}} > 0$; in the notation of Eq. (7), $A^{(i)} = \text{Norm}_{\mathcal{G}}(r^{(i)})$. The loss $\mathcal{L}_{\text{G}}^{\text{DAN}}(\pi_\theta, q; R_t, R_s)$ of Eq. (10) is the token-level clipped surrogate (Schulman et al., 2017) on groups sampled under $\tau(\boldsymbol{x})$, $\tau \sim q$, with the DAN advantages $\hat{A}_i$ of Eq. (7):

$$\begin{aligned}\mathcal{L}_{\text{G}}^{\text{DAN}} = -\,\mathbb{E}_{\tau\sim q}\,\mathbb{E}_{\boldsymbol{Y}\sim\pi_{\theta_{\text{old}}}^{\otimes G}(\cdot\mid\tau(\boldsymbol{x}))}\Big[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\boldsymbol{y}^{(i)}|}\sum_{t=1}^{|\boldsymbol{y}^{(i)}|}\min\big(w_{i,t}\hat{A}_i,\; c_{i,t}\hat{A}_i\big)\Big] \\ + \beta\,\mathbb{E}_{\tau\sim q}\,D_{\text{KL}}\big(\pi_\theta(\cdot\mid\tau(\boldsymbol{x}))\,\big\|\,\pi_{\text{ref}}(\cdot\mid\tau(\boldsymbol{x}))\big),\end{aligned}$$

where $w_{i,t} = \pi_\theta(y_t^{(i)} \mid \boldsymbol{y}_{<t}^{(i)}, \tau(\boldsymbol{x}))/\pi_{\theta_{\text{old}}}(y_t^{(i)} \mid \boldsymbol{y}_{<t}^{(i)}, \tau(\boldsymbol{x}))$ and $c_{i,t} = \text{clip}(w_{i,t}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})$. We use $(\epsilon_{\text{low}}, \epsilon_{\text{high}}) = (0.2, 0.28)$ and $\beta = 0$ (Table 8). Here q is the adversarial sampler, not the (C2) coverage $\rho_{\boldsymbol{x}}$, and $\hat{A}_i$ is normalized over the rollout batch of the current refresh. With $\tau(\boldsymbol{x}) = \boldsymbol{x}$ and $A^{(i)}$ in place of $\hat{A}_i$, the same expression is the GRPO loss (Shao et al., 2024).

F.2 Semantic Score of DTR

R_s is computed in the following order. *Candidates:* the contents of `\boxed{}` and `\answer{}`, the contents of `<answer></answer>`, the text after `#### Answer:`, and the text after an "(final) answer is" phrase (English or Chinese) up to the end of that sentence; if none is present, the last number in the response. A response with no candidate and no text is *Empty* and receives $R_s = 0$; every other response falls into exactly one of six graded categories, decided by the first tier that applies. *Exact* ($R_s = 1$): every explicit candidate matches a^* after normalization or numeric comparison. *Redundant* ($R_s = 0.5$): some candidate matches a^* but another explicit candidate gives a different answer. *Partial*: no candidate matches, and the best candidate earns graded credit, either the fraction of the ten relative-error tolerances $\{0.09, 0.08, \ldots, 0.01, 0\}$ that a numeric candidate meets or 0.5 for a normalized substring match. *Entail / Neutral / Contradict*: otherwise, a frozen NLI model (DeBERTa-large-MNLI (He et al., 2021)) scores the premise "The answer is $\langle c \rangle$.", with c the last explicit candidate (or the last three sentences of the response when there is none), against the hypothesis "The answer is a^*."; the category is the argmax label and R_s is the entailment probability. Numeric candidates are therefore graded by the tolerance ladder before the NLI tier: a numeric candidate within 9% relative error (e.g. 12.5 against 12, credit 0.5) never reaches the NLI model, and the NLI tier scores only candidates that earn no rule credit, namely non-numeric candidates without a substring match and numeric candidates beyond 9% relative error.

G Datasets, Licenses, and Reproducibility

G.1 Datasets and Statistics

Table 7 lists the datasets with their roles and sizes; the Legal VQA block is document VQA (DUDE for training, DocVQA for evaluation). Exam VQA training samples 200 MMK12 prompts per online-RL iteration, with $G = 16$ rollouts per prompt, over 10 iterations (Table 8).

G.2 Licenses and Intended Use

Both base models, Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct, are released under Apache-2.0; dataset licenses are listed in Table 7. All artifacts are used for non-commercial academic research consistent with their intended use, and benchmarks accessed under research-only terms (medical

Table 7: Datasets used, with role, domain, training-pool and evaluation-set sizes (#Train / #Eval; – = not used in that role), and license. Evaluation sizes are the item counts run by our evaluation harness; †: standard public split size. Licenses are as distributed by the original sources.

Dataset	Role	Domain	#Train / #Eval	License (as distributed)
MMK12 (Meng et al., 2025)	Train / ID eval	Exam (math/science)	15,616 / 2,000	Research use
MathVista (Lu et al., 2024)	ID eval	Exam (math)	– / 1,000	CC BY-SA 4.0
Olym-Phys (He et al., 2024)	OOD eval	Exam (physics)	– / 456	Apache-2.0
Lingshu (LASA Team et al., 2025)	Train	Medical	44,050 / –	Research use
VQA-RAD (Lau et al., 2018)	ID eval	Medical (radiology)	– / 451	CC0 1.0
PathVQA (He et al., 2020)	ID eval	Medical (pathology)	– / 6,719	MIT
GMAI-MMBench (Chen et al., 2024a)	OOD eval	Medical	– / 4,550	CC BY-NC-SA 4.0
DUDE (Van Landeghem et al., 2023)	Train	Document (Legal VQA block)	5,000 / –	CC BY 4.0
DocVQA (Mathew et al., 2021)	OOD eval	Document (Legal VQA block)	– / 5,349†	Research use

Table 8: Training, rollout, reward and adversary hyperparameters (GRPO with DTR/DAN and the embedding-space adversary).

Hyperparameter	Value	Hyperparameter	Value
Policy optimizer	GRPO	Rollout group size G	16
Learning rate	2×10^{-6}	MM-projector LR	1×10^{-5}
Optimizer	AdamW	$(\beta_1, \beta_2, \epsilon)$	$(0.9, 0.999, 10^{-8})$
Weight decay	0.01	Max grad norm	1.0
Clip range (low/high)	0.2/0.28	Reference-KL coef. β	0
Online-RL iterations	10	Prompts per iteration	200
Policy steps per iter.	≤ 50	Per-device batch	2
Grad. accumulation	2	Max sequence length	8192
Max new tokens (rollout)	4096	Precision	bf16
Rollout temperature	0.9	Rollout top-p	0.99
α_{rew} (in R_{DTR})	0.01	λ (consistency)	1.0
α_{adv} (DAN mixing)	1	NLI model for R_s	DeBERTa-large-MNLI
Adv. step size η_{adv}	5×10^{-5}	Adv. group size G_{adv}	4
Ascent steps K_{adv}	4	Updates per refresh C_{adv}	4
Radius ϵ_{emb}	$0.1\times$ median $\|e_{\mathcal{I}}\|_F$	ZeRO stage	3
Tuned modules	LLM + MLP	Frozen modules	vision enc.
Random seed	42/43/44	Eval decoding	greedy (T=0)

and document VQA) are not redistributed. The artifacts we create (trained policy checkpoints, dynamic-evaluation operators) are derivatives of research-accessed data and are released for research use only.

G.3 Models, Compute, and Infrastructure

The two policies, Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct (7–8B parameters each, including the vision encoder), are trained with DeepSpeed ZeRO-3 in bf16 mixed precision, with rollouts served by vLLM. The adversary's rollout and gradient cost is given in Appendix C (Compute cost).

G.4 Training Hyperparameters

Table 8 lists all training settings. The adversary's reward $R_{\text{DTR}} = R_t + \alpha_{\text{rew}} R_s$ ($\alpha_{\text{rew}} = 0.01$) is also the scalar reward of the w/o-DAN ablation; the policy advantage mixes the separately normalized R_t and R_s with $\alpha_{\text{adv}} = 1$ (Eq. (7); cf. Liu et al., 2026). The consistency weight $\lambda = 1.0$ multiplies the KL of Eq. (8) and is distinct from the reference-KL coefficient $\beta = 0$ (Appendix F); the decoupled clip range $0.2/0.28$ provides the trust region. Adversary settings are given in Appendix C; optimizer and clip settings follow GRPO.

G.5 Software and Packages

Training uses a custom Qwen-VL RL framework (PyTorch 2.7, Transformers 4.5x, DeepSpeed 0.16, FlashAttention 2.7.4) with vLLM for distributed rollout generation. Evaluation uses our improved MedEvalKit harness (vLLM, Transformers 4.52, `nltk` 3.9.1, `rouge` 1.0.1, `mathruler` 0.1.0, `qwen-vl-utils`), with `mathruler` providing the `\boxed{}` extraction; its prompts, scorers, GPT-4.1-nano judge and decoding settings are given in Appendix E.

Table 9: Pass@8 accuracy under standard evaluation (%), Qwen2.5-VL-7B backbone: a question counts as solved if any of 8 generated responses is correct. The Qwen2.5-VL-7B column is the base model. Cited in Section 4.6.

Benchmark	OOD	Qwen2.5-VL-7B	GRPO	PIRL (Ours)
VQA-RAD	×	76.99	75.66	**77.43**
PathVQA	×	58.81	58.22	**58.86**
GMAI-MMBench	✓	44.85	42.39	**44.90**

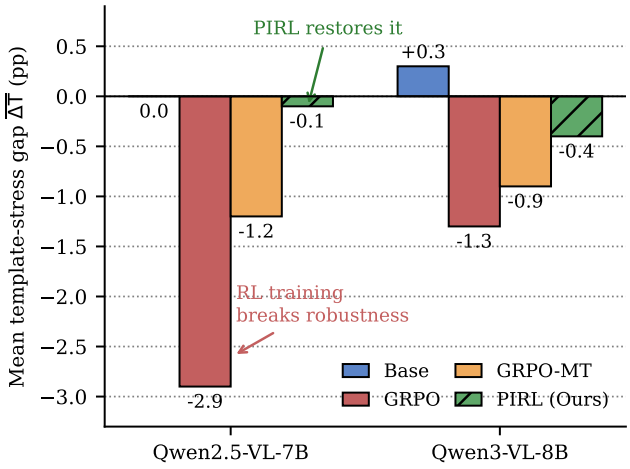


Figure 4: **Mean template-stress gap** $\overline{\Delta T}$ (over the 7 benchmarks; closer to 0 is more robust), recomputed from Table 1. Single-template GRPO sharply degrades robustness (Qwen2.5-VL-7B: $0.0 \rightarrow -2.9$ pp); GRPO-MT, trained on the same five-template pool as PIRL, partly mitigates it (-1.2); PIRL restores it to base level (-0.1). The effect is milder on the more robust Qwen3-VL-8B (GRPO -1.3, GRPO-MT -0.9, PIRL -0.4).

H Potential Risks

The embedding-space adversary, which searches for low-reward rewordings of the instruction, and the label-tracked dynamic-evaluation operators are dual-use: they could be repurposed to craft prompt perturbations that degrade a deployed model, or to probe for inputs on which it is brittle.

We mitigate this risk in three ways. First, the technique is defensive: its objective is to *reduce* the prompt-induced performance gap, and the released recipe strengthens robustness rather than providing a turnkey attack. Second, the search stays inside the semantic neighborhood of the instruction: the perturbation is continuous and bounded, its nearest-token projection is a near-paraphrase of the instruction (Appendix L), and it does not alter the question, the image or the label. Third, the dynamic-evaluation operators are label-tracked input mutations for measuring robustness, not for generating misleading or harmful content. We therefore release code and operators for research use only, documented as evaluation and robustness-training tools, and we recommend pairing any deployment with standard input validation and robustness monitoring.

Finally, robustness is not correctness: a prompt-stable model can still be confidently wrong, so in the medical and document domains studied here PIRL-trained policies are intended to support, not replace, expert judgment.

I PASS@8 RESULTS AND MEAN TEMPLATE-STRESS GAP

GRPO under D-Test. GRPO raises mean Standard accuracy over the base model (Table 1: $49.9 \rightarrow 52.2$ on Qwen2.5-VL-7B, $55.4 \rightarrow 57.5$ on Qwen3-VL-8B), but under D-Test it loses 8.5% on average over VQA-RAD, PathVQA and GMAI-MMBench on both backbones, about three times the drop of the base model (2.9 and 3.0%).

J WORKED EXAMPLE: ONE ROLLOUT GROUP THROUGH PIRL

We trace one *illustrative* rollout group (numbers chosen for exposition, not measured): a MathVista-style question with $a^* = 12$, a template τ_0 that requests `\boxed{}`, an instruction embedding already perturbed within the ϵ_{emb}-ball (Algorithm 2) while the extractor still reads `\boxed{}`, and $G = 4$ rollouts ($G = 16$ in training).

	Response (final segment)	$a_\tau(\boldsymbol{y})$	V	$\hat{A}^{\text{GRPO}}$	R_t	R_s	R_{DTR}	Z (DAN, α_{adv}=1)
$\boldsymbol{y}_1$	"…so the area is `\boxed{12}`" (Exact)	12	1	$+1.73$	$+1$	1.00	$+1.010$	$+2.32$
$\boldsymbol{y}_2$	"…`\boxed{12.5}`" (rounding slip; Partial)	12.5	0	-0.58	0	0.50	$+0.005$	-0.32
$\boldsymbol{y}_3$	"…`\boxed{3}`" (wrong reasoning; Contradict)	3	0	-0.58	0	0.02	$+0.000$	-1.50
$\boldsymbol{y}_4$	"…the answer is 12." (no delimiter; Exact)	$\emptyset$	0	-0.58	-1	1.00	-0.990	-0.51

Binary GRPO. $V = V_f V_c$ gives the three failures the same advantage (-0.58), so the update cannot separate the near-miss $\boldsymbol{y}_2$, the wrong-reasoning $\boldsymbol{y}_3$ and the correct-but-unformatted $\boldsymbol{y}_4$. **DTR.** R_t separates the format failure $\boldsymbol{y}_4$ from the formatted wrong answers, and R_s ranks *inside* a bucket ($\boldsymbol{y}_2$ above $\boldsymbol{y}_3$), the within-bucket variation of Theorem B.10. With $\alpha_{\text{rew}} = 0.01 < 1$, R_{DTR}, the reward the adversary descends, orders the group lexicographically: $\boldsymbol{y}_1 > \boldsymbol{y}_2 > \boldsymbol{y}_3 > \boldsymbol{y}_4$. **DAN.** The policy update ranks the group by $R_t + \lambda_{\mathcal{G}} R_s$ with $\lambda_{\mathcal{G}} = \alpha_{\text{adv}} \sigma_t / \sigma_s = 1.74$ (Proposition B.13): the correct-content $\boldsymbol{y}_4$ overtakes the wrong $\boldsymbol{y}_3$ because its semantic lead 0.98 exceeds $1/\lambda_{\mathcal{G}} = 0.58$, while $\boldsymbol{y}_2$ stays above $\boldsymbol{y}_4$ (lead 0.50). Separate z-scoring keeps the semantic term at unit scale; under a single normalization of R_{DTR} the $\boldsymbol{y}_2$–$\boldsymbol{y}_3$ difference of 0.005 would be negligible. Z is then re-normalized over the refresh batch and clipped (Eq. (7)). **Consistency.** The teacher-forced token distributions of the group under the perturbed prompt are pulled toward those under the stop-gradient clean prompt $\tau_0(\boldsymbol{x})$ (Eq. (8)), so a perturbation-induced drift toward $\boldsymbol{y}_4$-style unformatted answers is penalized directly, independently of the reward.

K EXTENDED COMPONENT ABLATION

Table 10 extends Table 2 with the cells that separate the reward side (DTR, i.e. R_t with R_s, and DAN) from the invariance side (ADV: embedding adversary with consistency loss) and from template augmentation (MT). "w/o DAN" normalizes $R_t + \alpha_{\text{rew}} R_s$ in a single group (Section 3.1); "× (binary)" keeps MT and ADV but replaces DTR by the binary verifier of Eq. (2), whose negative the adversary then ascends. The setup is that of Table 2 (Qwen2.5-VL-7B, Exam VQA, optimizer, $G = 16$ and hyperparameters of Appendix G.4); configurations with ADV add the $K_{\text{adv}} G_{\text{adv}} = 16$ adversary rollouts per prompt and refresh of Appendix C. Entries are mean±s.d. over three seeds.

Reward side (DTR and DAN). With a single template and no adversary, DTR moves the Exam ΔT from -2.3 to -1.9 and DAN further to -1.5; on OOD Olym-Phys the same two steps raise Standard accuracy from 7.5 to 8.2 to 8.9 and ΔT from -1.6 to -1.1 to -0.4. With MT and ADV present, the Exam ΔT is -0.5 with the binary verifier, $+0.2$ with DTR without DAN, and $+0.6$ with DTR and DAN (full PIRL); D-Test follows the same order ($26.3 < 26.6 < 27.0$). Among the components PIRL adds to GRPO-MT, replacing DTR by the binary reward is the most damaging removal on T-Stress ($45.1 \rightarrow 43.8$; 1.1% on ΔT).

Table 10: **Extended ablation** on Qwen2.5-VL-7B (Exam VQA Avg = mean of MMK12/MathVista/Olym-Phys). The upper block uses a single training template, and the lower block the five-template pool. †The MT-only row is GRPO-MT, averaged from Table 1, whose per-benchmark seed s.d. is given there. Gaps within ∼1.5% are individually inconclusive. Bold: best per column.

Components			Exam VQA Avg.					Olym-Phys (OOD)				
MT	ADV	DTR	Std.	T-Stress	D-Test	ΔT	ΔD	Std.	T-Stress	D-Test	ΔT	ΔD
×	×	× (GRPO)	44.2±0.4	41.9±0.5	25.8±0.6	−2.3±0.5	−18.4±0.7	7.5±0.5	5.9±0.5	4.4±0.5	−1.6±0.5	**−3.1±0.5**
×	×	w/o DAN	44.3±0.4	42.4±0.5	25.9±0.6	−1.9±0.5	−18.4±0.6	8.2±0.5	7.1±0.5	4.3±0.5	−1.1±0.5	−3.9±0.5
×	×	w/ DAN	44.3±0.4	42.8±0.5	26.0±0.6	−1.5±0.5	−18.3±0.6	8.9±0.5	8.5±0.5	4.4±0.5	−0.4±0.5	−4.5±0.5
×	✓	w/ DAN	44.3±0.4	43.4±0.5	26.0±0.7	−0.9±0.5	−18.3±0.7	**10.3±0.5**	10.4±0.5	4.4±0.5	+0.1±0.5	−5.9±0.5
✓	×	× (GRPO-MT)†	44.2	43.6	26.6	−0.6	−17.6	7.8	7.6	4.3	−0.2	−3.5
✓	×	w/ DAN	44.4±0.5	44.4±0.5	26.8±0.7	0.0±0.5	−17.6±0.7	10.1±0.5	9.8±0.5	4.2±0.5	−0.3±0.5	−5.9±0.5
✓	✓	× (binary)	44.3±0.4	43.8±0.5	26.3±0.6	−0.5±0.5	−18.0±0.7	8.8±0.5	8.9±0.5	4.8±0.5	+0.1±0.5	−4.0±0.5
✓	✓	w/o DAN	44.4±0.4	44.6±0.5	26.6±0.6	+0.2±0.5	−17.8±0.7	9.2±0.5	9.8±0.5	4.9±0.5	+0.6±0.5	−4.3±0.5
✓	✓	w/ DAN (**PIRL**)	**44.5±0.4**	**45.1±0.5**	**27.0±0.7**	**+0.6±0.5**	**−17.5±0.7**	9.6±0.5	**10.7±0.5**	**5.5±0.5**	**+1.1±0.5**	−4.1±0.5

Invariance side (ADV). Adding ADV to single-template DTR improves ΔT from -1.5 to -0.9 and Olym-Phys Standard accuracy from 8.9 to 10.3; adding it to MT+DTR improves ΔT from 0.0 to $+0.6$, T-Stress from 44.4 to 45.1, and D-Test from 26.8 to 27.0.

Template augmentation (MT). MT alone (GRPO-MT) reduces ΔT to -0.6; on top of it, DTR+DAN and ADV together add 1.2% of ΔT and 1.5% of T-Stress accuracy. Conversely, ADV+DTR *without* MT reaches $\Delta T = -0.9$, within noise of GRPO-MT, from a single template.

Summary. Removing any one component from full PIRL lowers T-Stress accuracy: ADV by 0.7, DAN by 0.5, DTR by 1.3, and MT by 1.7%. Each component contributes in the same direction, and the reward-side components account for a share comparable to the adversary's. Of the single-component steps, only the MT removal exceeds the ∼1.5% noise level on its own; the evidence for each other component is the consistent trend across configurations and both metrics, and the combined effect (ΔT from -2.3 to $+0.6$, T-Stress from 41.9 to 45.1) is well above noise. On the aggregate D-Test, all configurations lie within 1.5% (25.8 to 27.0).

L Answer-Level, Adversary, and Compute-Matched Diagnostics

Answer flips and worst-template accuracy. For every question of MathVista, VQA-RAD, PathVQA and GMAI-MMBench we generate one greedy response under each of the 15 T-Stress templates and extract its answer. The flip rate is the share of questions whose extracted answer is not the same under all 15 templates; the worst-template accuracy is the lowest of the 15 per-template accuracies. Both measure the answer-level invariance that (C2) asks for (Definition A.2) and that a T-Stress mean averages out. The yes/no items of VQA-RAD and PathVQA keep one instruction under every template (Appendix E), so only their open items can flip.

RL creates the gap at the answer level: GRPO raises the base model's flip rate by a factor of 1.5–1.7 on every benchmark, and its worst template costs it 7.1, 7.8 and 7.4% below its T-Stress accuracy on MathVista, VQA-RAD and PathVQA. On these three benchmarks PIRL has the lowest flip rate and the highest worst-template accuracy of all four models. Against GRPO-MT, trained on the same template pool, its flip rate is 4.7–6.6% lower and its worst-template accuracy 1.2–5.3% higher; on MathVista, where its T-Stress mean is 0.4% below GRPO-MT's, it flips 4.7% fewer answers. On GMAI-MMBench, PIRL's flip rate (13.5%) lies between those of GRPO-MT (12.6%) and GRPO (20.4%), and the base model is best in all three rows.

Adversary. On MMK12 training prompts, with $\epsilon_{\text{emb}} = 0.1 \cdot \text{median}\|\boldsymbol{e}_{\mathcal{I}}\|_F$ (Appendix C), the mean R_{DTR} of the policy's rollouts is $+0.64$ at $\boldsymbol{\delta} = \mathbf{0}$ and -0.18 after the $K_{\text{adv}} = 4$ ascent steps. The mean of R_{DTR} is $\Pr[\text{C}] - \Pr[\text{F}] + \alpha_{\text{rew}}\mathbb{E}R_s$ with $\alpha_{\text{rew}}\mathbb{E}R_s \in [0, 0.01]$, so correct answers lead format failures by 63–64 points of rollout mass before the attack, and format failures lead correct answers by 18–19 points after it. The realized drop, 0.82, is a lower bound on the mean ϵ_{emb}-drop h_π of Lemma B.17(a) at these prompts. The perturbations stay in the semantic neighborhood of the

Table 11: Answer-level invariance on Qwen2.5-VL-7B over the 15 T-Stress templates. T-Stress: accuracy with one template drawn per question (Table 1); Flip: share of questions whose answer changes across the 15 templates (↓); Worst: lowest per-template accuracy (↑). All entries in %; bold: best per row.

Benchmark	Metric	Base	GRPO	GRPO-MT	PIRL
MathVista	T-Stress	**69.6**	65.7	69.3	68.9
	Flip ↓	14.2	22.8	12.1	**7.4**
	Worst ↑	64.1	58.6	66.0	**67.2**
VQA-RAD	T-Stress	66.3	71.2	69.3	**73.4**
	Flip ↓	12.5	18.9	11.4	**6.1**
	Worst ↑	61.8	63.4	65.2	**70.5**
PathVQA	T-Stress	42.4	41.4	43.2	**48.5**
	Flip ↓	15.0	25.4	14.8	**8.2**
	Worst ↑	38.2	34.0	41.6	**46.1**
GMAI-MMBench	T-Stress	**42.4**	39.7	40.9	42.1
	Flip ↓	**11.8**	20.4	12.6	13.5
	Worst ↑	**38.6**	32.8	37.5	38.2

instruction: decoding a perturbed instruction by projecting each perturbed token embedding onto its nearest vocabulary embedding gives an instruction whose DeBERTa embedding has cosine similarity 0.912 with the original on average, and 94 of 100 manually inspected decodings still state the original instruction clearly, typically with repeated punctuation or near-synonym tokens.

Table 12: Compute-matched baselines on Qwen2.5-VL-7B (Exam VQA Avg. as in Table 2). Sequences and backward passes are counted per prompt; PIRL's include the adversary's $K_{\text{adv}}G_{\text{adv}} = 16$ rollouts and $K_{\text{adv}} = 4$ ascent passes.

Configuration	Sequences	Backward passes	Std.	T-Stress	ΔT
GRPO-MT ($G = 16$)	16	4	44.2	43.6	−0.6
GRPO-MT ($G = 32$)	32	4	44.6	44.0	−0.6
GRPO-MT (2× iterations)	16	8	**44.7**	44.1	−0.6
PIRL ($G = 16$)	32	8	44.5	**45.1**	**+0.6**

Compute-matched baselines. Table 12 gives GRPO-MT the sampling budget of PIRL (32 sequences per prompt: the 16 policy and 16 adversary rollouts of Appendix C) or its number of backward passes (twice the iterations). Both raise GRPO-MT's Standard accuracy by 0.4–0.5% and leave its ΔT at −0.6; PIRL reaches +0.6, and each baseline matches its budget in one of the two counts.