TACE: TWO-STAGE ADVANTAGE AND CREDIT ESTIMATION FOR MULTIMODAL REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

We revisit advantage estimation in Group Relative Policy Optimization (GRPO) and observe that any token-level advantage admits a two-stage factorization: a *sample-level baselining* step that converts scalar verifier rewards into a comparable sequence-level signal, and a *token-level credit assignment* step that distributes that signal over the trajectory. Existing GRPO variants modify only one of the two stages, and we show that each leaves a structural failure mode in the other untouched: linear standardization preserves the shape of the empirical reward distribution, while teacher-driven distribution matching encodes privileged context into the gradient direction and is provably ill-posed under information asymmetry. We propose **TACE** (*Two-stage Advantage and Credit Estimation*), a single estimator that fixes both stages. The first stage replaces linear normalization with the unique 1D optimal-transport map onto a standard normal, yielding inter-task gradient equity, robustness to heavy tails and bimodality, and strict sign preservation. The second stage uses an evidence ratio between the model under privileged and non-privileged contexts as a stop-gradient, sign-anchored magnitude multiplier, ensuring that the verifier retains exclusive control of the gradient direction. Composing the two stages costs only one sort and one extra forward pass per group beyond standard GRPO. On Qwen3-VL-Instruct-8B, TACE outperforms the base model by 4.69 points and the strongest GRPO variant by 2.32 points averaged over MMMU, MathVista, MathVision, ZeroBench, and WeMath, while sustaining improvement well past the point where teacher-driven self-distillation degrades.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR), exemplified by GRPO (Shao et al., 2024), has become the dominant post-training recipe for large reasoning models (DeepSeek-AI et al., 2025; Team et al., 2026b; Bai et al., 2025). Despite a flurry of GRPO variants (Liu et al., 2025; Feng, 2025; Zhao et al., 2026; Hübotter et al., 2026), progress has plateaued and several recently proposed remedies actively destabilize training (Zhao et al., 2026). We argue that this stagnation has a single underlying cause: the GRPO advantage estimator silently performs *two distinct jobs*, and every existing improvement only fixes one of them. For each prompt, GRPO computes $A_i = (R_i - \mu_G)/\sigma_G$ from a group of rollouts and assigns the same A_i to every token, hiding two independent operations: **(S1) sample-level baselining**, which turns a scalar reward into a comparable advantage, and **(S2) token-level credit assignment**, which spreads A_i across the trajectory's tokens. Existing variants perturb *exactly one* stage: Dr. GRPO (Liu et al., 2025) and EMA-GRPO (Feng, 2025) change the denominator of (S1); OPSD (Zhao et al., 2026) and SDPO (Hübotter et al., 2026) replace the uniform per-token weight in (S2) with a teacher-derived target.

Each leaves a structural failure mode untouched. Linear standardization is affine in R_i, so heavy tails, binary cliffs, and inter-task variance gaps all propagate verbatim into the gradient magnitude. The natural fix for stage (S2), borrowing dense supervision from a teacher conditioned on a verified reasoning trace, is *provably ill-posed*: distribution matching under information asymmetry leaves an irreducible mutual-information term $I(Y_t; R \mid X, Y_{<t}) > 0$ in the loss (Theorem 1) and injects a per-sample r-specific deviation into the gradient (Proposition 1). Empirically this manifests as the policy citing reference solutions never visible at inference (Appendix B).

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.

We propose **TACE**, an advantage estimator that addresses both stages in a single closed form and obeys two invariants: (i) the sample-level advantage is calibrated by the *rank* of the reward rather than its raw value, and (ii) any per-token modulation acts strictly as a *magnitude* signal whose sign is locked to the verifier's verdict. Stage (S1) becomes the unique 1D optimal-transport map onto $\mathcal{N}(0,1)$, yielding inter-task gradient equity, bounded tails, and order preservation. Stage (S2) becomes the evidence ratio $P_T(y_t)/P_S(y_t)$ exponentiated by $\mathrm{sign}(A_i)$ and applied as a stop-gradient multiplier; this is structurally leakage-free because the privileged signal can scale credit but cannot flip it. The two stages share no scalar that could create a circular dependency, so improvements compose. The net cost over a vanilla GRPO step is one $O(N \log N)$ sort and one extra forward pass.

In summary, our contributions are three-fold:

- **Unified diagnosis of GRPO advantage estimation**: we formalize the (S1)+(S2) decomposition, show every existing variant perturbs only one stage, and prove that distribution matching from a privileged teacher is ill-posed and induces path-dependent gradient leakage.
- **TACE: a leakage-free two-stage estimator** that combines a quantile-normalized sample baseline with a sign-anchored evidence-ratio token weight. We give inter-task gradient-equity for (S1) and a formal leakage-free guarantee for (S2).
- **New state of the art on multimodal RLVR**: on Qwen3-VL-Instruct-8B, TACE improves the base model by 4.69 points and the strongest GRPO variant by 2.32 points averaged over MMMU, MathVista, MathVision, ZeroBench, and WeMath, and dominates GRPO/OPSD/SDPO on every column of a broader nine-benchmark suite.

2 Related Work

2.1 Reward Normalization Inside GRPO

Since GRPO's introduction (Shao et al., 2024; DeepSeek-AI et al., 2025), a substantial body of work has revisited how the group-relative advantage $A_i = (R_i - \mu_G)/\sigma_G$ should be standardized when the verifier signal is heterogeneous across tasks or rollouts. Dr. GRPO (Liu et al., 2025) drops the per-group standard deviation and replaces it with a centered baseline; EMA-GRPO (Feng, 2025) swaps the local σ_G for a task-wise running estimate to mitigate variance imbalance across multimodal task families; further extensions explore softer or task-conditioned variants (Yu et al., 2025; Gao et al., 2025; Xie et al., 2025b; Zheng et al., 2025a). All of these schemes apply an *affine* transformation to the raw reward, which by construction cannot reshape the underlying reward distribution: heavy tails, binary cliffs, and inter-task scale gaps survive the rescaling and propagate into the gradient magnitude. TACE departs from this affine family by applying a non-linear, rank-based transport to a fixed reference distribution, recovering inter-task equity and bounded tails as automatic consequences rather than as side products of hyperparameter tuning.

2.2 Token-Level Credit Assignment for Verifier-Only RL

A parallel line of research argues that uniformly broadcasting the sequence-level advantage to every token wastes the most informative bits of a reasoning trajectory. One direction trains explicit step-level or process reward models that score intermediate computations, often supervised by human annotators or by automated proxies (Lightman et al., 2024; Wang et al., 2024b; Luo et al., 2024; Chen et al., 2024; Zhang et al., 2024; Cui et al., 2025; Dai et al., 2025; Yang et al., 2025b;a). A second direction stays inside the verifier-only RLVR loop and infers token importance from intrinsic statistics of the policy itself: entropy, key-token detection, attention concentration, or sensitivity of the outcome to specific positions (Cheng et al., 2026; Wang et al.; Chen et al., 2025a; Sun et al., 2025c; Xie et al., 2025a; Li et al., 2025b; Chen et al., 2025d; Li et al., 2026). The first direction adds an external scorer and the corresponding optimization burden; the second relies on heuristic proxies that are correlated with importance but not derived from any normative criterion. TACE occupies a different niche: it derives a per-token magnitude signal from a self-consistent Bayesian quantity (the log-evidence that a privileged trace accumulates from emitting each token), and applies it as a stop-gradient multiplier, so the verifier alone retains authority over the gradient direction.

2.3 On-Policy and Self-Distillation in RL Post-Training

On-policy distillation (OPD) (Agarwal et al., 2024; Lu & Lab, 2025; Fu et al., 2026) threads dense token-level supervision into the student's own rollouts via a separate, usually larger, teacher; recent reports show that strong OPD pipelines can reach or surpass RL-only post-training in reasoning quality (Team et al., 2026a). To eliminate the cost of a second network, on-policy self-distillation (OPSD) variants reuse the policy itself as the teacher by conditioning it on a privileged context such as a verified reasoning trace or environment feedback (Zhao et al., 2026; Hübotter et al., 2026). Closely related variants apply the same idea to demonstration replay, context internalization, and reasoning compression, where the teacher conditions on demonstrations, retrieved context, or instructions favoring brevity (Shenfeld et al., 2026; Ye et al., 2026; Sang et al., 2026; Penaloza et al., 2026). A concurrent line, exemplified by TRRD (Zhang et al., 2026), instead injects the teacher's preference into the trust-region anchor of the policy update. Across this family, the teacher's distribution acts as a *gradient target* for the student. We diagnose this design as structurally ill-posed under information asymmetry (Theorem 1, Proposition 1): no matter how the target is reweighted or thresholded, the privileged signal contaminates the gradient direction. TACE breaks with this convention by routing the same teacher–student discrepancy through a multiplicative, sign-locked, stop-gradient factor that rescales credit without ever pointing the gradient.

2.4 Optimal Transport for Language Model Training

Optimal transport (OT) has appeared in language-model post-training mainly as a distance functional: token-level Wasserstein distances regularize preference alignment, stochastic-dominance constraints reshape reward distributions, and OT-based projections enforce safety properties in latent space (Melnyk et al., 2024; Li et al., 2025a; Na et al., 2026; Nanfack et al., 2026). The role we give OT is different: rather than measuring a divergence between two learned distributions, we exploit the closed-form 1D Brenier map onto $\mathcal{N}(0,1)$ as the canonical solution to the sample-level baselining problem inside GRPO. This places OT inside the advantage estimator itself, where it provides distribution invariance, outlier robustness, and order preservation as immediate properties of the rank–CDF composition.

2.5 Multimodal RL Post-Training

Reinforcement learning has rapidly become a standard tool for sharpening multimodal reasoning in vision-language models (Li et al., 2025c; Sun et al., 2025a; Feng et al., 2025; Sun et al., 2025b; Zhou et al., 2025; Duan et al., 2025; Chen et al., 2025b;c; Meng et al., 2025b; Wang et al., 2025c; Zhang et al., 2025), where reward heterogeneity across perception-heavy and reasoning-heavy tasks is far more pronounced than in pure-text settings. Existing recipes typically address this with task-specific data curation, auxiliary perceptual objectives, or external captioners (Tian et al., 2025; Yang et al., 2025c; Tu et al., 2025). TACE is orthogonal to these recipes: by enforcing distribution invariance at stage (S1), it makes the policy gradient insensitive to differences in reward calibration across tasks, so the same training pipeline can ingest perception, mathematics, and chart benchmarks without per-task tuning of the advantage scale.

3 Background and Diagnosis

Section 3.1 fixes the notation we use for GRPO and on-policy distillation, and Section 3.2 states the two diagnostic results that drive every design choice in TACE: (i) any teacher-driven token-level objective that places the privileged context r in the gradient direction is provably ill-posed; (ii) the irreducible loss term it leaves behind is what corrupts per-sample gradients in practice. Together, these two facts force TACE's stage (S2) to use P_T/P_S exclusively as a stop-gradient *magnitude* signal (Section 4.3), and they justify why the unified template of Eq. (4) is the right level at which to compare existing methods.

3.1 Preliminaries

GRPO. RLVR post-training assumes a policy π_θ generating responses $y = (y_1, \ldots, y_T)$ autoregressively from query x under a verifier $R(x, y) \in \{0, 1\}$. GRPO (Shao et al., 2024) samples G

candidates $\{y^{(i)}\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ and standardizes their rewards into a sequence-level advantage $A^{(i)} = (R(x, y^{(i)}) - \mu_G)/\sigma_G$, then reuses the same $A^{(i)}$ for every token of $y^{(i)}$ in the clipped policy-gradient objective

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}\min\Big(\rho_t^{(i)}A^{(i)},\ \text{clip}(\rho_t^{(i)}, 1-\epsilon, 1+\epsilon)\,A^{(i)}\Big)\right], \tag{1}$$

where $\rho_t^{(i)}$ is the standard importance ratio. Per-token credit within a trajectory is necessarily uniform.

On-policy and self-distillation objectives. A second line densifies token-level supervision through a reference P_T. On-policy distillation (OPD) (Agarwal et al., 2024) takes P_T from a separate stronger model on the same query x. On-policy self-distillation (OPSD) (Zhao et al., 2026; Hübotter et al., 2026) reuses π_θ as both player and reference, with the reference receiving privileged input r (typically a verified reasoning trace). With $P_S(\cdot \mid y_{<t}) \triangleq \pi_\theta(\cdot \mid x, y_{<t})$ and $P_T(\cdot \mid y_{<t}) \triangleq \pi_\theta(\cdot \mid x, r, y_{<t})$, both fit

$$\mathcal{L}_{\text{OP(S)D}}(\theta) = \mathbb{E}_{(x,r)\sim\mathcal{S}}\,\mathbb{E}_{y\sim P_S(\cdot|x)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|}D\big(P_T \,\|\, P_S\big)\right], \tag{2}$$

where D is a divergence (KL or JS). The shared design uses P_T as the gradient *direction*, the choice TACE cannot inherit.

3.2 Diagnosis: What Goes Wrong When the Teacher Sets the Gradient Direction

We extract two structural facts from Eq. (2) that dictate the form of TACE's stage (S2): the first tells us *where* the privileged context must never enter, the second tells us *why* that prohibition must hold sample by sample.

Fact 1: An irreducible mutual-information floor. By construction, the OPSD reference P_T depends on r while the student P_S does not. The richest function P_S can hope to express, then, is the r-marginal $\bar{P}_T(y_t) \triangleq \mathbb{E}_r[P_T(y_t \mid x, r, y_{<t})]$. Confronting Eq. (2) (specialized to $D = D_{\text{KL}}$) with the marginal-matching counterpart

$$\mathcal{L}^*(\theta) \;=\; \mathbb{E}_x\big[D_{\text{KL}}\big(\bar{P}_T \,\|\, P_S(\cdot \mid x)\big)\big] \tag{3}$$

yields the following decomposition.

Theorem 1 (KL Decomposition)**.**

$$\mathcal{L}_{\text{OPSD}}(\theta) \;=\; \mathcal{L}^*(\theta) \;+\; I(Y_t;\, R \mid X,\, Y_{<t}),$$

where $I(Y_t; R \mid X, Y_{<t})$ is the conditional mutual information between the next-token variable and the privileged trace under the teacher distribution. The proof is in Appendix G.1.

The mutual-information term is fixed by $P(r \mid x)$ and the conditional law of the teacher; it does not depend on θ, so no amount of student optimization can push the OPSD loss below it. The student's best feasible solution is $P_S^* = \bar{P}_T$, at which a strictly positive residual of $I(Y_t; R \mid X, Y_{<t})$ remains whenever r is informative about y_t. This is the analytic reason why the on-policy KL of OPSD-style training plateaus rather than converging.

Consequence for TACE. The decomposition makes the location of the privileged signal a structural concern, not a tuning matter: any token-level method that lets $P_T(\cdot \mid r)$ enter the loss as a gradient *target* inherits the same residual term and the same nonzero loss floor. TACE therefore reuses P_T/P_S only after applying $\texttt{sg}(\cdot)$ in Eq. (6), so the privileged signal contributes only as a scalar magnitude and never as the direction the gradient is asked to imitate.

Fact 2: Per-sample gradients carry an r-specific deviation. A skeptical reader could note that, because $I(Y_t; R \mid X, Y_{<t})$ does not depend on θ, the *population* OPSD gradient coincides with the gradient of $\mathcal{L}^*$. The trouble appears only when the optimizer sees one (x, r) pair at a time. Writing $g(\theta; r) = -\sum_v P_T(v \mid r) \nabla_\theta \log P_S(v)$ for the per-sample gradient under realization r, we obtain the following decomposition.

Proposition 1 (Per-Sample Gradient Decomposition)**.**

$$g(\theta; r) \;=\; \underbrace{-\sum_v \bar{P}_T(v) \nabla_\theta \log P_S(v)}_{\text{population direction } g^\star(\theta)} \;+\; \underbrace{-\sum_v \big[P_T(v \mid r) - \bar{P}_T(v)\big] \nabla_\theta \log P_S(v)}_{r\text{-specific noise } \delta(\theta; r)},$$

with $\mathbb{E}_r[\delta(\theta; r)] = 0$ and second moment bounded by $\mathbb{E}_r\|\delta(\theta; r)\|^2 \;\le\; |\mathcal{V}| \sum_v \mathrm{Var}_r[P_T(v \mid r)] \, \|\nabla_\theta \log P_S(v)\|^2$. Both terms vanish iff $I(Y_t; R \mid X, Y_{<t}) = 0$, and otherwise grow monotonically with it. Proof in Appendix G.2.

The two terms behave qualitatively differently as training proceeds. The population direction $\|g^\star\|$ shrinks to zero as P_S approaches $\bar{P}_T$, while $\|\delta\|$ stays bounded away from zero because its variance is governed by the same θ-independent term as in Fact 1. Stochastic optimizers are path-dependent, so the cumulative effect of these zero-mean but nonvanishing perturbations is to nudge the parameters toward whatever r-correlated patterns remain statistically detectable. Empirically this manifests as the policy hallucinating an unseen reference at inference time, the symptom commonly described as privileged-information leakage.

Consequence for TACE. Proposition 1 disqualifies any fix that merely *reweights* or *thresholds* the OPSD target distribution: as long as $P_T(\cdot \mid r)$ enters as the direction of $\nabla_\theta \log P_S$, the same δ-term appears, with proportional magnitude. The only structural escape is to ensure that P_T never multiplies $\nabla_\theta \log P_S$ at all. TACE achieves this by routing the privileged signal through the multiplicative scalar $w_{i,t}$ in Eq. (7), and by anchoring its sign to $\mathrm{sign}(A_i)$ so that $w_{i,t} > 0$ pointwise, thereby preserving the gradient direction supplied by the verifier.

Summary of the diagnosis. The two facts above identify the precise constraint that any token-level credit assignment scheme must satisfy when borrowing privileged information: *the privileged signal must influence the magnitude of the per-token contribution to the policy gradient, but never its direction.*

4 METHOD

The diagnosis above leaves us with a precise constraint: r may not enter the gradient direction, but $P_T(y_t)/P_S(y_t)$ remains an admissible *scalar* once stripped of directional content. **TACE** (*Two-stage Advantage and Credit Estimation*) is the construction that satisfies this at stage (S2) while simultaneously fixing linear standardization at stage (S1).

4.1 A TWO-STAGE VIEW OF GRPO ADVANTAGE ESTIMATION

GRPO writes its advantage as $A_i = (R_i - \mu_G)/\sigma_G$ with the same A_i shared by every token in $y^{(i)}$, but this hides two statistically independent operations:

$$\underbrace{\{R_i\}_{i=1}^G \xrightarrow{\text{(S1)}} A_i}_{\text{sample-level baselining}} \xrightarrow{\text{(S2)}} \underbrace{\hat{A}_{i,t} \;=\; A_i \cdot w_{i,t}}_{\text{token-level credit assignment}} . \tag{4}$$

Vanilla GRPO uses linear standardization in (S1) and the trivial weight $w_{i,t} = 1$ in (S2). All recently proposed alternatives perturb *exactly one* stage: Dr. GRPO (Liu et al., 2025) and EMA-GRPO (Feng, 2025) change the denominator of A_i in (S1); OPSD (Zhao et al., 2026) and SDPO (Hübotter et al., 2026) replace the uniform weight in (S2) with a teacher-derived target. Table 1 catalogues these instantiations alongside ours. Each stage carries a distinct failure mode the other cannot repair: stage (S1) governs cross-task gradient equity (linear standardization is affine, so heavy tails, binary cliffs, and inter-task variance gaps propagate verbatim into the gradient magnitude); stage (S2) governs token-level concentration (uniform credit flattens it, while P_T/P_S as a gradient target is structurally leakage-prone, Theorem 1).

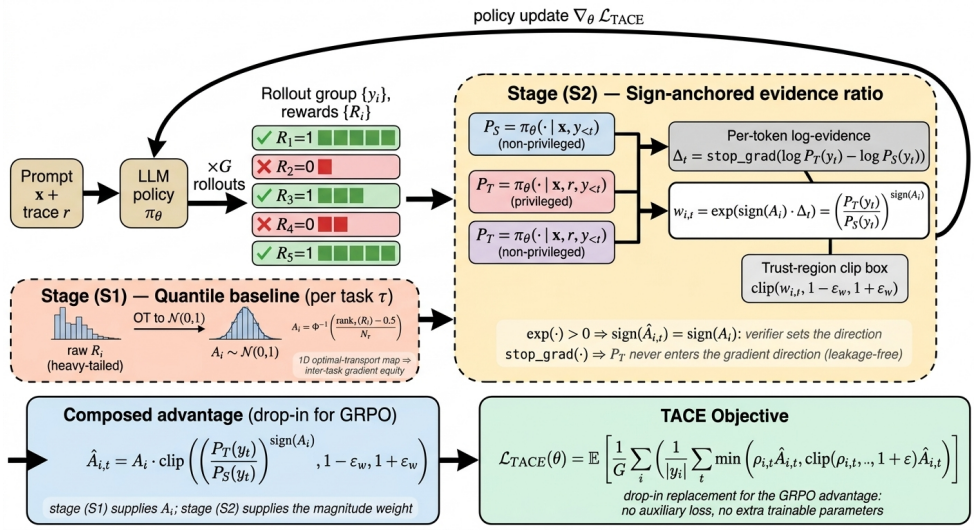


Figure 1: **Overview of TACE. Stage (S1)** replaces $(R_i - \mu_G)/\sigma_G$ with the 1D optimal-transport map onto $\mathcal{N}(0,1)$, $A_i = \Phi^{-1}((\text{rank}_\tau(R_i) - 0.5)/N_\tau)$, for inter-task gradient equity. **Stage (S2)** runs π_θ under non-privileged (P_S) and privileged (P_T) contexts, and combines the per-token log-evidence $\Delta_t = \texttt{sg}(\log P_T(y_t) - \log P_S(y_t))$ into a sign-anchored, clipped magnitude weight $w_{i,t} = \text{clip}((P_T/P_S)^{\text{sign}(A_i)})$. The composed token advantage $\hat{A}_{i,t} = A_i \cdot w_{i,t}$ plugs into the standard GRPO clipped surrogate.

Table 1: Existing GRPO variants viewed through Eq. (4). TACE is the only entry that satisfies both invariants I1 and I2.

Method	**Sample baseline A_i (S1)**	**Token weight $w_{i,t}$ (S2)**	**Direction anchor**
GRPO (Shao et al., 2024)	$(R_i - \mu_G)/\sigma_G$	1	Verifier
Dr. GRPO (Liu et al., 2025)	$R_i - \mu_G$	1	Verifier
EMA-GRPO (Feng, 2025)	$(R_i - \mu_G)/\sigma_\tau^{(t)}$	1	Verifier
OPSD (Zhao et al., 2026)	$(R_i - \mu_G)/\sigma_G$	P_T/P_S as gradient target	**Teacher**
SDPO (Hübotter et al., 2026)	$(R_i - \mu_G)/\sigma_G$	sign-free evidence ratio	**Teacher**
TACE (Ours)	$\Phi^{-1}\big(\widehat{F}_\tau(R_i)\big)$	$\text{clip}\big((P_T/P_S)^{\text{sign}(A_i)}\big)$	Verifier

Two invariants any principled estimator must satisfy. I1 (Distribution invariance): the marginal distribution of A_i should not depend on the verifier's calibration; (S1) must be a *distribution-aware* map, not an affine one. **I2 (Direction anchoring):** the sign of every token-level advantage must equal $\text{sign}(A_i)$, the verifier's verdict on the trajectory; (S2) must act on the *magnitude* of credit, never on its direction. Section 4.2 instantiates I1 via the 1D optimal-transport map onto $\mathcal{N}(0,1)$; Section 4.3 instantiates I2 via a stop-gradient evidence ratio whose exponent is $\text{sign}(A_i)$.

4.2 Stage 1: Quantile-Normalized Sample Baseline

The reward distribution $P_{\mathcal{R}_\tau}$ produced on task τ is rarely well-conditioned: binary verifiers yield two-point distributions, dense verifiers exhibit heavy tails, and joint multi-task training mixes incompatible scales. Linear standardization carries every one of these pathologies into the advantage because it is an affine (hence shape-preserving) map. To enforce invariant I1 we demand a non-linear transformation that sends $P_{\mathcal{R}_\tau}$ to a fixed, well-conditioned reference distribution $P_\mathcal{N} = \mathcal{N}(0,1)$ at minimum quadratic cost. By Brenier's theorem, in one dimension this transport admits the closed-form rank–CDF composition

$$A_i \triangleq \Psi_\tau(R_i) = \Phi^{-1}\big(\widehat{F}_\tau(R_i)\big), \qquad \widehat{F}_\tau(R_i) = \frac{\text{rank}_\tau(R_i) - \frac{1}{2}}{N_\tau}, \tag{5}$$

where $\widehat{F}_\tau$ is task τ's empirical CDF (computed online over the current group, optionally smoothed by a running histogram), and Φ^{-1} is the standard normal quantile function. Identical raw rewards (a frequent case under binary verifiers) are mapped to the average of their quantile images so that "identical outcomes receive identical credit."

Properties relevant to (S1). Three immediate consequences of Eq. (5) jointly satisfy invariant I1:

- *Inter-task gradient equity.* For every task τ, A_i has zero mean and unit variance after the OT map, regardless of $P_{\mathcal{R}_\tau}$'s scale or shape. The expected per-task contribution to $\|\nabla_\theta \mathcal{L}\|$ is therefore a function of the policy alone, not of the verifier's calibration.
- *Outlier and bimodality robustness.* A single anomalously large reward is mapped to $\Phi^{-1}\big((\mathrm{rank} - 0.5)/N\big)$, which is bounded with high probability; bimodal binary rewards are converted into smooth Gaussian tails via the same rank composition.
- *Strict order preservation.* $\Phi^{-1} \circ \widehat{F}_\tau$ is strictly monotone, so $R_i > R_j \Leftrightarrow A_i > A_j$. In particular, $\mathrm{sign}(A_i)$ matches the verifier's relative verdict and is therefore safe to use as the direction anchor exploited by stage (S2).

The implementation reduces to one $O(N \log N)$ sort per group; no additional hyperparameter is introduced at this stage.

4.3 Stage 2: Sign-Anchored Evidence-Ratio Token Credit

Stage (S2) must allocate the well-conditioned scalar A_i to the tokens of $y^{(i)}$ *without* violating invariant I2. We instantiate the per-token weight $w_{i,t}$ via an evidence ratio computed from the same model under two contexts (privileged and non-privileged), and we use that ratio strictly as a magnitude multiplier.

Defining the per-token signal. The same parameters θ enter two distributions over y_t: $P_T \triangleq \pi_\theta(\cdot \mid x, r, y_{<t})$ (with the verified trace r) and $P_S \triangleq \pi_\theta(\cdot \mid x, y_{<t})$ (without). Their log-difference at the sampled token is

$$\Delta_t \;=\; \mathtt{sg}\big(\log P_T(y_t) \;-\; \log P_S(y_t)\big)\,, \tag{6}$$

where $\mathtt{sg}$ blocks Δ_t from contributing to ∇_θ along any path other than its eventual use as a multiplicative scalar. Since P_T and P_S share θ, Δ_t isolates r's marginal contribution at position t free of teacher–capacity confounds. Appendix G.3 gives the Bayesian reading $\Delta_t = \log[P(r \mid x, y_{\le t})/P(r \mid x, y_{<t})]$. Without the stop-gradient, the construction would inherit OPSD's gradient-direction pathology of Theorem 1.

From signal to credit weight. A direct exponential of Δ_t would reward the same tokens regardless of trajectory correctness, breaking I2. We exponentiate with $\mathrm{sign}(A_i)$:

$$w_{i,t} \;=\; \exp\big(\mathrm{sign}(A_i) \cdot \Delta_t\big) \;=\; \left(\frac{P_T(y_t)}{P_S(y_t)}\right)^{\mathrm{sign}(A_i)}. \tag{7}$$

On verifier-positive trajectories ($A_i > 0$), $w_{i,t} = P_T/P_S$ funnels reinforcement to tokens r supports; on negative trajectories, the ratio inverts, so tokens r disfavors absorb the bulk of the penalty while r-aligned tokens are partially shielded. Crucially, $w_{i,t} > 0$ always, so $\mathrm{sign}(\hat{A}_{i,t}) = \mathrm{sign}(A_i)$: the verifier's verdict is never overwritten (Definition 1, Theorem 3).

Bounding the credit redistribution. We clip $w_{i,t}$ before applying it to A_i:

$$\hat{A}_{i,t} \;=\; A_i \cdot \mathrm{clip}\big(w_{i,t},\, 1 - \epsilon_w,\, 1 + \epsilon_w\big). \tag{8}$$

Symmetrically to the importance-ratio clip on the policy channel in PPO/GRPO, this clip caps the per-step redistribution of credit across tokens, keeping both updates within a tractable trust region. We additionally schedule a short warm-up $w_{i,t} \leftarrow (1-\lambda) + \lambda\, w_{i,t}$ with λ rising linearly from 0 to 1 over the first 50 steps.

4.4 Composed Estimator

Substituting Eq. (5) and Eq. (8) into Eq. (4) gives the closed-form TACE advantage:

$$\hat{A}_{i,t} \;=\; \Phi^{-1}\big(\widehat{F}_\tau(R_i)\big) \cdot \mathrm{clip}\left(\left(\frac{P_T(y_t)}{P_S(y_t)}\right)^{\mathrm{sign}(A_i)},\, 1-\epsilon_w,\, 1+\epsilon_w\right), \tag{9}$$

which plugs into the standard clipped-surrogate objective:

$$\mathcal{L}_{\mathrm{TACE}}(\theta) \;=\; \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}\min\big(\rho_{i,t}\,\hat{A}_{i,t},\; \mathrm{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\,\hat{A}_{i,t}\big)\right]. \tag{10}$$

Writing $\hat{A}_{i,t} = f(R_i;\tau)\cdot g(\mathrm{sign}(f), \Delta_t; \epsilon_w)$ with $f = \Phi^{-1}\circ\widehat{F}_\tau$ and $g = \mathrm{clip}(\exp(\mathrm{sign}\cdot\Delta_t), \cdot)$, the two stages compose without interference: f is monotone (so $\mathrm{sign}(f)$ is the only feature g reads), and $g > 0$ pointwise (so it leaves $\mathrm{sign}(\hat{A}_{i,t}) = \mathrm{sign}(f)$ intact). Appendix D confirms super-additivity. Beyond GRPO, TACE adds one $O(N \log N)$ sort per task and one extra forward pass per response (overhead $<20\%$); see Algorithm 1 (Appendix A).

5 Empirical Evaluation

5.1 Setup and Baselines

Training corpus and evaluation suite. We train on MMFineReason-123K (Lin et al., 2026), a difficulty-filtered subset of MMFineReason-1.8M (only samples on which Qwen3-VL-4B-Thinking fails all four rollouts are kept). We evaluate on five multimodal reasoning benchmarks: **MMMU** (Yue et al., 2024) (multi-discipline college-level VQA), **MathVista** (Lu et al., 2024) (math reasoning over visual contexts), **MathVision** (Wang et al., 2024a) (competition-level visual math), **ZeroBench** (Roberts et al., 2025) (frontier-unsolvable stress test), and **WeMath** (Qiao et al., 2024) (fine-grained math). All benchmarks report accuracy (%).

Backbone and compared methods. The backbone in every run is **Qwen3-VL-8B-Instruct** (Bai et al., 2025). Five comparators are evaluated: **Base LLM** (no post-training, a floor); **GRPO** (Shao et al., 2024); **OPSD** (Zhao et al., 2026), single-model self-distillation conditioned on a verified reasoning trace; **SDPO** (Hübotter et al., 2026), an OPSD variant using environment feedback as the privileged context; and **GRPO+OPSD**, the additive-fusion baseline inspired by MOPD (Team et al., 2026a).

Optimization and hyperparameters. Our pipeline is implemented on top of the VERL (Sheng et al., 2025) and EasyR1 (Zheng et al., 2025b) stacks. Each rollout uses an 8192-token context window, allocated as a 4096-token prompt and a 4096-token response, and the same length budget is applied at evaluation. We adopt a learning rate of 1×10^{-6} for the three policy-gradient methods (GRPO, GRPO+OPSD, TACE) and 1×10^{-5} for the two distillation-driven baselines (OPSD, SDPO), matching the values reported by their respective authors. Group size is 8 rollouts per prompt with sampling temperature 1.0, and an effective optimizer batch of 256 prompts. The PPO-style importance-ratio clip is asymmetric, $(\epsilon_{\text{low}}, \epsilon_{\text{high}}) = (0.2, 0.28)$; we drop both the KL-to-reference penalty and the entropy bonus from every objective. For TACE specifically, the warm-up coefficient λ ramps linearly from 0 to 1 over the first 50 optimization steps, the credit-channel clip is $\epsilon_w = 0.2$, and the teacher-side parameters are refreshed from the student every 10 steps. The privileged context r differs by method: OPSD consumes verified reasoning traces distilled from Qwen3-VL-235B-A22B-Thinking and filtered for correctness on MMFineReason-123K, SDPO uses a previously successful rollout as in its original paper, while TACE only requires the **final ground-truth answer** (no intermediate trace), making it the lightest among the three. Hardware: a 4-node H200 cluster with eight 140 GB GPUs per node (32 GPUs total).

5.2 Experimental Results

5.2.1 Five-Benchmark Reasoning Results

As reported in Table 2, TACE attains the top average accuracy on the five-benchmark suite, exceeding the un-tuned base by 4.69 points and the strongest GRPO variant by 2.32 points at the 4K context

Table 2: Multimodal reasoning results on the Qwen3-VL-8B-Instruct model. **Bold** indicates the best result among all models. Avg. is computed over the five selected benchmarks.

Method	MMMU	MathVista	MathVision	ZeroBench	Wemath	Avg.
Base LLM	62.44	73.80	47.37	19.76	54.10	51.49
GRPO	65.11	76.20	48.82	22.60	56.57	53.86
OPSD	63.82	75.10	47.53	21.06	54.95	52.49
SDPO	65.11	74.00	47.27	**25.15**	52.19	52.74
GRPO+OPSD	63.22	75.90	48.52	22.16	54.76	52.91
TACE *(Ours)*	**67.22**	**78.10**	**52.73**	24.85	**58.00**	**56.18**

setting. The improvement is most visible on benchmarks demanding fine-grained step-level discrimination, MathVision ($+3.91$) and MathVista ($+1.9$), where the dense token-level credit channel of stage (S2) is expected to pay off. Both self-distillation baselines (OPSD, SDPO) under-perform plain GRPO on average; §3.2 attributes this to privileged-information leakage when the teacher sets the gradient direction, whereas TACE routes the teacher only through a multiplicative magnitude. GRPO+OPSD is the next strongest baseline but still trails TACE by 3.27 points, consistent with the scale mismatch TACE sidesteps via its clipped, sign-locked multiplier.

5.2.2 Broader Multimodal Generalization

Beyond the five reasoning benchmarks of Table 2, we further evaluate TACE on a broader nine-benchmark suite spanning general VQA (MMMU, MMBench, MMStar), math VQA (MathVista, MathVerse, MathVision), and chart VQA (AI2D, ChartQA, CharXiv); the full table and per-benchmark analysis are deferred to Appendix F (Table 7). Three observations are worth noting in the main text. **(i)** TACE strictly dominates the Qwen3-VL-Instruct-8B family (Base, +GRPO, +OPSD) on every one of the nine benchmarks, with an average gain of $+2.7$ over +GRPO on the six previously unseen columns, closely matching the $+2.32$ of Table 2. **(ii)** TACE attains the best open-source score on $8/9$ benchmarks (only trailing OneThinker-8B on MMMU by 3.4); on ChartQA in particular, 87.1 surpasses both GPT-4o (86.7) and Gemini 2.5 Pro (83.3). **(iii)** The gains concentrate on benchmarks whose reward topology is most heterogeneous (ChartQA $+5.1$, MMStar $+2.5$, CharXiv $+2.7$, MathVerse $+1.5$), tracking the stage-(S1) prediction.

5.2.3 Training Dynamics, Ablation, and Case Study

We summarize three additional analyses, deferring full tables to the appendix. *Dynamics* (Appendix C): TACE reaches a higher accuracy plateau than GRPO and avoids OPSD's late-stage collapse, while sustaining higher per-token entropy. *Ablation* (Appendix D): under a 2×2 design, each TACE stage alone strictly improves over GRPO, and the joint estimator is super-additive ($+2.32$ vs. a sum of individual gains $+1.65$, i.e. $\sim$41% extra). *Case study* (Appendix E): TACE concentrates per-token credit on the decisive reasoning tokens, while generic narration stays close to the trajectory-level baseline.

6 Conclusion and Limitations

GRPO's plateau on multimodal reasoning traces to a hidden two-stage decomposition: linear standardization mishandles heterogeneous reward shapes, and teacher-driven distribution matching turns token-level credit ill-posed. **TACE** resolves both by admitting privileged signals only as stop-gradient magnitudes anchored to the verifier's sign, at one extra sort and forward pass per group, gaining $4.69/2.32$ over the base/best-GRPO on Qwen3-VL-Instruct-8B. Our analysis is confined to verifier-only RLVR with a scalar outcome reward on multimodal reasoning; the diagnosis should be re-validated when the reward is non-scalar or the policy is conditioned on text-only or video inputs. Richer privileged contexts and multi-marginal optimal transport for vector-valued rewards are natural extensions.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

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, 2025. URL https://arxiv.org/abs/2511.21631.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Step-level value preference optimization for mathematical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7889–7903, 2024.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*, 2025a.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *ArXiv preprint*, abs/2506.04207, 2025b. URL https://arxiv.org/abs/2506.04207.

Shuang Chen, Yue Guo, Yimeng Ye, Shijue Huang, Wenbo Hu, Haoxi Li, Manyuan Zhang, Jiayu Chen, Song Guo, and Nanyun Peng. Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping. *ArXiv preprint*, abs/2510.08457, 2025c. URL https://arxiv.org/abs/2510.08457.

Xinzhu Chen, Xuesheng Li, Zhongxiang Sun, and Weijie Yu. Beyond high-entropy exploration: Correctness-aware low-entropy segment-based advantage shaping for reasoning llms. *arXiv preprint arXiv:2512.00908*, 2025d.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30377–30385, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv preprint*, abs/2507.06261, 2025. URL https://arxiv.org/abs/2507.06261.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Muzhi Dai, Chenxu Yang, and Qingyi Si. S-grpo: Early exit via reinforcement learning in reasoning models, 2025. URL https://arxiv.org/abs/2505.07686.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,

Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *ArXiv preprint*, abs/2503.17352, 2025. URL https://arxiv.org/abs/2503.17352.

Chengqi Duan, Kaiyue Sun, Rongyao Fang, Manyuan Zhang, Yan Feng, Ying Luo, Yufang Liu, Ke Wang, Peng Pei, Xunliang Cai, et al. Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images. *ArXiv preprint*, abs/2510.11718, 2025. URL https://arxiv.org/abs/2510.11718.

Anonymous Feng. Onethinker: An all-in-one reasoning model with ema-grpo. *arXiv preprint*, 2025.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *ArXiv preprint*, abs/2503.21776, 2025. URL https://arxiv.org/abs/2503.21776.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes, 2026. URL https://arxiv.org/abs/2603.25562.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization, 2025. URL https://arxiv.org/abs/2511.20347.

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

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation, 2026. URL https://arxiv.org/abs/2601.20802.

Meng Li, Guangda Huzhang, Haibo Zhang, Xiting Wang, and Anxiang Zeng. Optimal transport-based token weighting scheme for enhanced preference optimization, 2025a. URL https://arxiv.org/abs/2505.18720.

Yang Li, Zhichen Dong, Yuhan Sun, Weixun Wang, Shaopan Xiong, Yijia Luo, Jiashun Liu, Han Lu, Jiamang Wang, Wenbo Su, et al. Attention illuminates llm reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization. *arXiv preprint arXiv:2510.13554*, 2025b.

Ziheng Li, Liu Kang, Feng Xiao, Luxi Xing, Qingyi Si, Zhuoran Li, Weikang Gong, Deqing Yang, Yanghua Xiao, and Hongcheng Guo. Outcome-grounded advantage reshaping for fine-grained credit assignment in mathematical reasoning. *arXiv preprint arXiv:2601.07408*, 2026.

Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Spatial transformation reasoning by reinforcing multimodal llms. *ArXiv preprint*, abs/2505.15804, 2025c. URL https://arxiv.org/abs/2505.15804.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Honglin Lin, Zheng Liu, Yun Zhu, Chonghan Qin, Juekai Lin, Xiaoran Shang, Conghui He, Wentao Zhang, and Lijun Wu. Mmfinereason: Closing the multimodal reasoning gap via open data-centric methods, 2026. URL https://arxiv.org/abs/2601.21821.

Zichen Liu et al. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

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, 2024. URL https://arxiv.org/abs/2310.02255.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jerret Ross. Distributional preference alignment of llms via optimal transport, 2024. URL https://arxiv.org/abs/2406.05882.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *ArXiv preprint*, abs/2503.07365, 2025a. URL https://arxiv.org/abs/2503.07365.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *ArXiv preprint*, abs/2510.20579, 2025b. URL https://arxiv.org/abs/2510.20579.

Byeonghu Na, Hyungho Na, Yeongmin Kim, Suhyeon Jo, HeeSun Bae, Mina Kang, and Il-Chul Moon. Semantic-aware wasserstein policy regularization for large language model alignment, 2026. URL https://arxiv.org/abs/2602.01685.

Geraldin Nanfack, Eugene Belilovsky, and Elvis Dohmatob. Efficient refusal ablation in llm through optimal transport, 2026. URL https://arxiv.org/abs/2603.04355.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. In *The 1st Workshop on Scaling Post-training for LLMs*, 2026. URL https://openreview.net/forum?id=FbJu6NEBQR.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning?, 2024. URL https://arxiv.org/abs/2407.01284.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, Vatsal Raina, Hanyi Xiong, Vishaal Udandarao, Jingyi Lu, Shiyang Chen, Sam Purkis, Tianshuo Yan, Wenye Lin, Gyungin Shin, Qiaochu Yang, Anh Totti Nguyen, David I. Atkinson, Aaditya Baranwal, Alexandru Coca, Mikah Dang, Sebastian Dziadzio, Jakob D. Kunz, Kaiqu Liang, Alexander Lo, Brian Pulfer, Steven Walton, Charig Yang, Kai Han, and Samuel Albanie. Zerobench: An impossible visual benchmark for contemporary large multimodal models, 2025. URL https://arxiv.org/abs/2502.09696.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. On-policy self-distillation for reasoning compression. *arXiv preprint arXiv:2603.05433*, 2026.

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, 2024. URL https://arxiv.org/abs/2402.03300.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*, 2026. URL https://openreview.net/forum?id=HlWA3V6iKF.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL http://dx.doi.org/10.1145/3689031.3696075.

Haoyuan Sun, Jiaqi Wu, Bo Xia, Yifu Luo, Yifei Zhao, Kai Qin, Xufei Lv, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement fine-tuning powers reasoning capability of multimodal large language models. *ArXiv preprint*, abs/2505.18536, 2025a. URL https://arxiv.org/abs/2505.18536.

Peiwen Sun, Shiqiang Lang, Dongming Wu, Yi Ding, Kaituo Feng, Huadai Liu, Zhen Ye, Rui Liu, Yun-Hui Liu, Jianan Wang, et al. Spacevista: All-scale visual spatial reasoning from mm to km. *ArXiv preprint*, abs/2510.09606, 2025b. URL https://arxiv.org/abs/2510.09606.

Wei Sun, Wen Yang, Pu Jian, Qianlong Du, Fuwei Cui, Shuo Ren, and Jiajun Zhang. Ktae: A model-free algorithm to key-tokens advantage estimation in mathematical reasoning, 2025c. URL https://arxiv.org/abs/2505.16826.

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Lei Li, Liang Zhao, Linghao Zhang, Peidian Li, Qianli Chen, Shaohui Liu, Shihua Yu, Shijie Cao, Shimao Chen, Shouqiu Yu, Shuo Liu, Tianling Zhou, Weijiang Su, Weikun Wang, Wenhan Ma, Xiangwei Deng, Bohan Mao, Bowen Ye, Can Cai, Chenghua Wang, Chengxuan Zhu, Chong Ma, Chun Chen, Chunan Li, Dawei Zhu, Deshan Xiao, Dong Zhang, Duo Zhang, Fangyue Liu, Feiyu Yang, Fengyuan Shi, Guoan Wang, Hao Tian, Hao Wu, Heng Qu, Hongfei Yi, Hongxu An, Hongyi Guan, Xing Zhang, Yifan Song, Yihan Yan, Yihao Zhao, Yingchun Lai, Yizhao Gao, Yu Cheng, Yuanyuan Tian, Yudong Wang, Zhen Tang, Zhengju Tang, Zhengtao Wen, Zhichao Song, Zhixian Zheng, Zihan Jiang, Jian Wen, Jiarui Sun, Jiawei Li, Jinlong Xue, Jun Xia, Kai Fang, Menghang Zhu, Nuo Chen, Qian Tu, Qihao Zhang, Qiying Wang, Rang Li, Rui Ma, Shaolei Zhang, Shengfan Wang, Shicheng Li, Shuhao Gu, Shuhuai Ren, Sirui Deng, Tao Guo, Tianyang Lu, Weiji Zhuang, Weikang Zhang, Weimin Xiong, Wenshan Huang, Wenyu Yang, Xin Zhang, Xing Yong, Xu Wang, Xueyang Xie, Yilin Jiang, Yixin Yang, Yongzhe He, Yu Tu, Yuanliang Dong, Yuchen Liu, Yue Ma, Yue Yu, Yuxing Xiang, Zhaojun Huang, Zhenru Lin, Zhipeng Xu, Zhiyang Chen, Zhonghua Deng, Zihan Zhang, and Zihao Yue. Mimo-v2-flash technical report, 2026a. URL https://arxiv.org/abs/2601.02780.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen,

Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026b. URL https://arxiv.org/abs/2602.02276.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. More thought, less accuracy? on the dual nature of reasoning in vision-language models. *ArXiv preprint*, abs/2509.25848, 2025. URL https://arxiv.org/abs/2509.25848.

Songjun Tu, Qichao Zhang, Jingbo Sun, Yuqian Fu, Linjing Li, Xiangyuan Lan, Dongmei Jiang, Yaowei Wang, and Dongbin Zhao. Perception-consistency multimodal large language models reasoning via caption-regularized policy optimization, 2025. URL https://arxiv.org/abs/2509.21854.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *ArXiv preprint*, abs/2504.08837, 2025a. URL https://arxiv.org/abs/2504.08837.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024a. URL https://arxiv.org/abs/2402.14804.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024b.

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. Vision-zero: Scalable vlm self-improvement via strategic gamified self-play, 2025b. URL https://arxiv.org/abs/2509.25541.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning, 2025c. URL https://arxiv.org/abs/2507.06448.

Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jisheng Yin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, Jia Wang, Chunrui Han, Yuang Peng, Qi Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, and Vishal M. Patel. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning, 2025. URL https://arxiv.org/abs/2507.05255.

Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning. *arXiv preprint arXiv:2510.10649*, 2025a.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *ArXiv preprint*, abs/2502.14768, 2025b. URL https://arxiv.org/abs/2502.14768.

Chenxu Yang, Qingyi Si, Mz Dai, Dingyu Yao, Mingyu Zheng, Minghui Chen, Zheng Lin, and Weiping Wang. Test-time prompt intervention, 2025a. URL https://arxiv.org/abs/2508.02511.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models, 2025b. URL https://arxiv.org/abs/2504.15895.

Siqi Yang, Zilve Gao, Haibo Qiu, Fanfan Liu, Peng Shi, Zhixiong Zeng, Qingmin Liao, and Lin Ma. Learning when to look: A disentangled curriculum for strategic perception in multimodal reasoning, 2025c. URL https://arxiv.org/abs/2512.17227.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv preprint*, abs/2503.14476, 2025. URL https://arxiv.org/abs/2503.14476.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Vision-g1: Towards general vision language reasoning with multi-domain data curation, 2025. URL https://arxiv.org/abs/2508.12680.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

Shuoshuo Zhang, Yizhen Zhang, Jingjing Fu, Lei Song, Jiang Bian, Yujiu Yang, and Rui Wang. See less, see right: Bi-directional perceptual shaping for multimodal reasoning, 2025. URL https://arxiv.org/abs/2512.22120.

Zhaoyang Zhang, Shuli Jiang, Yantao Shen, Yuting Zhang, Dhananjay Ram, Shuo Yang, Zhuowen Tu, Wei Xia, and Stefano Soatto. Reinforcement-aware knowledge distillation for llm reasoning, 2026. URL https://arxiv.org/abs/2602.22495.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026. URL https://arxiv.org/abs/2601.18734.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *ArXiv preprint*, abs/2507.18071, 2025a. URL https://arxiv.org/abs/2507.18071.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, Yuwen Xiong, and Richong Zhang. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025b.

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*, abs/2504.21277, 2025. URL https://arxiv.org/abs/2504.21277.

A ALGORITHM

Algorithm 1 TACE: Two-stage advantage and credit estimation in one training step.

Require: Policy π_θ, dataset $\mathcal{S} = \{(x_i, r_i, \tau_i)\}$, verifier $R(\cdot,\cdot)$, group size G, clip ϵ_w, warm-up λ

for each batch $\{(x, r, \tau)\}$ **do**
  Sample G rollouts $\{y^{(i)}\} \sim \pi_\theta(\cdot \mid x)$ and obtain rewards $R_i = R(x, y^{(i)})$
  // Stage (S1): quantile-normalized baseline (per task)
  Group rewards by task τ; sort within task to get ranks; set $A_i \leftarrow \Phi^{-1}\big((\mathrm{rank}_\tau(R_i) - 0.5)/N_\tau\big)$, average ties
  // Stage (S2): sign-anchored evidence-ratio token credit
  for each rollout $y^{(i)}$ **do**
    One extra forward pass with $(x, r, y^{(i)})$ to obtain teacher logits
    for $t = 1, \ldots, |y^{(i)}|$ **do**
      $\Delta_t \leftarrow \mathrm{sg}\big(\log \pi_\theta(y_t \,|\, x, r, y_{<t}) - \log \pi_\theta(y_t \,|\, x, y_{<t})\big)$
      $w_t \leftarrow \exp(\mathrm{sign}(A_i) \cdot \Delta_t)$
      $\hat{A}_{i,t} \leftarrow A_i \cdot \big((1-\lambda) + \lambda \cdot \mathrm{clip}(w_t,\, 1-\epsilon_w,\, 1+\epsilon_w)\big)$
    end for
  end for
  Update θ by maximizing the clipped-surrogate objective in Eq. (10)
end for

B DIAGNOSIS: EMPIRICAL EVIDENCE

We provide quantitative evidence for the two-phase failure pattern of OPSD-style training described in §1. Table 3 reports leakage rate (defined as the fraction of inference-time outputs that explicitly cite "the reference solution" or equivalent strings that were never present in the prompt), on-policy KL between teacher and student, and validation accuracy across 200 optimization steps. The leakage rate rises monotonically while validation accuracy peaks within the first ∼20 steps and then degrades, exactly as predicted by Theorem 1 and Proposition 1. A representative leaked output we observed reads: "… to match the *reference answer*, I will round my intermediate result to two decimals … since *the worked solution* is asking for the simplified form, the choice should be (B) …", yet no reference answer or worked solution was supplied to the model at inference.

Table 3: Pilot study quantifying OPSD's leakage pathology on Qwen3-VL-Instruct-8B. Leakage rate denotes the fraction of validation outputs that cite an unseen reference; $\mathrm{KL}(P_T \| P_S)$ is computed on-policy over the validation rollouts.

Step	0	20	50	100	150	200
Leakage rate (%)	0.0	4.2	11.7	18.3	22.6	25.4
$\mathrm{KL}(P_T \| P_S)$	0.84	0.61	0.49	0.43	0.41	0.40
Validation accuracy (%)	51.49	53.21	52.37	51.18	50.43	49.86

C TRAINING DYNAMICS

Table 4 compares GRPO, OPSD, and TACE across 200 optimization steps along three axes: (a) validation accuracy (mean of the five benchmarks in Table 2), (b) per-token policy entropy, and (c) the empirical clip-activation rate of the per-token weight in Eq. (8). TACE attains a steeper initial ascent and a higher converged plateau than GRPO, while avoiding the late-stage collapse that OPSD exhibits after step ∼50. TACE also maintains a consistently higher entropy than GRPO, indicating that its targeted credit reweighting strengthens decisive tokens without uniformly suppressing alternatives. The clip rate of the credit channel stabilizes at 3–6%, confirming that the trust-region clip in Eq. (8) is actively engaged but not saturating.

Table 4: Training dynamics on the multimodal reasoning suite. Accuracy is the mean over the five benchmarks of Table 2; entropy is per-token in nats; clip rate is the empirical fraction of token positions on which the per-token weight is clipped (TACE only).

Step	0	20	50	100	150	200
Validation accuracy (%)						
GRPO	51.49	52.40	53.05	53.46	53.71	53.86
OPSD	51.49	53.18	52.92	51.86	50.94	50.42
TACE	51.49	53.74	54.92	55.61	55.93	56.18
Per-token policy entropy (nats)						
GRPO	0.62	0.41	0.27	0.19	0.15	0.13
OPSD	0.62	0.45	0.34	0.28	0.25	0.23
TACE	0.62	0.49	0.38	0.32	0.29	0.27
Clip activation rate (%, TACE only)						
TACE	–	5.7	4.6	3.9	3.5	3.3

D Ablation Study

Table 5 decomposes TACE into its two stages on the same five-benchmark suite as the main results. We draw three observations.

(i) Each stage alone is a strict improvement. Enabling only stage (S1) lifts the average from 53.86 (GRPO) to 54.74 (+0.88); enabling only stage (S2) yields 54.63 (+0.77). Both variants dominate every OPSD-style baseline of Table 2, confirming that invariants I1 and I2 are individually actionable.

(ii) The two stages specialize in complementary regimes. Stage (S1) is comparatively stronger on benchmarks whose reward distribution is dominated by scale or shape effects (WeMath +0.61, ZeroBench +0.71). Stage (S2) is comparatively stronger on long multi-step reasoning (MathVision +1.98 vs. S1's +1.66).

(iii) The two stages are super-additive. The full TACE estimator reaches 56.18, a gain of +2.32 that strictly exceeds the sum +1.65 of the individual gains by +0.67 (∼41% extra). Per-benchmark super-additivity is largest on MathVision (+0.27) and ZeroBench (+1.09). This corroborates the compatibility argument of §4.4: when either half is missing the sign channel coupling the two stages is broken, and the credit weight either misfires on a miscalibrated A_i (no S1) or is diluted by uniform per-token mass (no S2).

Table 5: Ablation of the two TACE stages on Qwen3-VL-Instruct-8B. Stage (S1) denotes the quantile-normalized sample baseline of Eq. (5); stage (S2) denotes the sign-anchored evidence-ratio token credit of Eq. (8).

Method	(S1)	(S2)	MMMU	MathVista	MathVision	ZeroBench	WeMath	Avg.
Base LLM	–	–	62.44	73.80	47.37	19.76	54.10	51.49
GRPO	✗	✗	65.11	76.20	48.82	22.60	56.57	53.86
+ (S1) only	✓	✗	65.78	76.97	50.48	23.31	57.18	54.74
+ (S2) only	✗	✓	65.44	76.83	50.80	23.05	57.05	54.63
TACE	✓	✓	**67.22**	**78.10**	**52.73**	**24.85**	**58.00**	**56.18**

E Case Study

To make the per-token reweighting concrete, we sample two trajectories from MathVista evaluation and report the normalized credit magnitude $|\hat{A}_{i,t}|/\max_t|\hat{A}_{i,t}|$ that TACE assigns at selected token positions (Table 6). A value near 1.0 means the position dominates the trajectory's gradient signal; a value near 0 means the position behaves as semantic background.

The first trajectory answers a geometry question correctly. TACE places the bulk of the positive credit on the two tokens that actually drive the answer: the spatial referent ("shaded triangle") and the closing arithmetic step ("$= 24$"), while linker tokens such as "Looking" and "the figure shows" receive credit below 0.15. The second trajectory mis-solves a percentage word problem. Here the negative credit concentrates on the misformulated equation ("$0.4 \cdot p = 80$") and the resulting wrong answer ("$p = 200$"), whereas neutral framing tokens ("Let p", "denote") absorb only a small fraction of the penalty. In both cases the sign of every per-token contribution coincides with the trajectory-level verifier verdict: the magnitude is shaped by the privileged teacher, but the direction is never overwritten.

Table 6: Normalized per-token credit magnitude $|\hat{A}_{i,t}|/\max_t |\hat{A}_{i,t}|$ assigned by TACE on two MathVista trajectories. Positive credit reinforces; negative credit penalizes. Tokens listed in their order of appearance.

Trajectory	**Token (excerpt)**	$\lvert\hat{A}_{i,t}\rvert$
(a) Correct: triangle-area question (verdict positive)		
reinforce	"Looking"	0.07
reinforce	"the figure shows"	0.14
reinforce	**"shaded triangle"**	**0.89**
reinforce	"base"	0.39
reinforce	**"$= 24$"**	**1.00**
(b) Incorrect: percentage word problem (verdict negative)		
penalize	"Let p"	0.12
penalize	"denote"	0.16
penalize	**"$0.4 \cdot p = 80$"**	**0.94**
penalize	"hence"	0.21
penalize	**"$p = 200$"**	**1.00**

F Broader Multimodal Generalization

Table 7 extends the evaluation of Section 5.2.2 to nine multimodal benchmarks grouped by task type: general VQA (MMMU (Yue et al., 2024), MMBench, MMStar), math VQA (MathVista, MathVerse, MathVision), and chart VQA (AI2D, ChartQA, CharXiv). To ensure a controlled comparison, the Base, +GRPO, and +OPSD rows on Qwen3-VL-Instruct-8B share the identical training recipe of Section 5; external models are quoted from their original papers.

Table 7: Broader evaluation across nine multimodal benchmarks. **Bold** indicates the best result among the Qwen3-VL-Instruct-8B family. On the six columns that do not appear in Table 2, TACE improves over +GRPO by an average of $+2.7$ points.

	General VQA			**Math VQA**			**Chart VQA**		
Model	MMMU	MMBench	MMStar	MathVista	MathVerse	MathVision	AI2D	ChartQA	CharXiv(RQ)
Proprietary:									
GPT-4o (Hurst et al., 2024)	70.7	84.3	65.1	63.8	41.2	30.4	84.9	86.7	47.1
Gemini 2.5 Pro (Comanici et al., 2025)	81.7	90.1	77.5	82.7	82.9	73.3	88.4	83.3	67.9
Open-source multimodal reasoning models:									
MM-Eureka-7B (Meng et al., 2025a)	57.3	87.7	64.4	73.0	50.3	26.9	84.1	77.3	39.5
OpenVLThinker-7B (Deng et al., 2025)	55.8	87.8	59.1	70.2	47.9	25.3	81.8	85.7	39.3
VL-Rethinker-7B (Wang et al., 2025a)	56.7	87.9	62.7	74.9	54.2	32.3	80.8	84.7	39.8
Vision-G1 (Zha et al., 2025)	53.4	88.0	63.1	76.1	50.0	31.3	82.1	85.6	41.0
ARES-7B (Chen et al., 2025c)	67.9	–	65.3	74.6	56.5	51.9	–	–	47.0
OVR-7B (Wei et al., 2025)	–	86.6	62.7	72.1	54.6	51.8	–	–	44.5
VisionZero (Wang et al., 2025b)	58.8	–	65.8	73.1	52.1	28.5	84.8	86.3	–
OneThinker-8B (Feng, 2025)	70.6	86.6	70.6	77.6	64.3	48.3	85.2	76.4	44.0
Qwen3-VL-Instruct-8B family (ours):									
Base LLM	62.44	85.4	68.8	73.80	58.5	47.37	82.5	83.0	44.8
+ GRPO	65.11	86.2	71.0	76.20	63.8	48.82	84.7	82.0	49.8
+ OPSD	63.82	85.8	70.4	75.10	62.7	47.53	84.3	81.5	49.3
+ **TACE** *(Ours)*	**67.22**	**88.3**	**73.5**	**78.10**	**65.3**	**52.73**	**86.9**	**87.1**	**52.5**

Strict dominance within the Qwen3-VL-Instruct-8B family. TACE outperforms Base, +GRPO, and +OPSD on every one of the nine benchmarks. ChartQA is particularly diagnostic: plain GRPO fine-tuning *regresses* the base model from 83.0 to 82.0, and OPSD regresses further to 81.5, whereas TACE recovers and extends to 87.1 ($+5.1$ over GRPO, $+5.6$ over OPSD). This is the heterogeneous-reward regime that motivated invariant I1 in §4.1: chart benchmarks mix binary correctness with partial numeric credit, producing a non-Gaussian reward distribution that linear standardization cannot tame.

Best open-source result on $8/9$ benchmarks. Against strong multimodal RL pipelines (OneThinker-8B, ARES-7B, VL-Rethinker-7B, Vision-G1, VisionZero), TACE attains the top open-source score on every benchmark except MMMU, where it trails OneThinker-8B by 3.4. Against GPT-4o, TACE wins $8/9$; on ChartQA, 87.1 is the highest score among all proprietary and open-source systems we are aware of.

Gains concentrate where reward topology is most heterogeneous. The four largest TACE-over-GRPO deltas on the six new benchmarks (ChartQA $+5.1$, CharXiv $+2.7$, MMStar $+2.5$, MathVerse $+1.5$) are all tasks whose verifier mixes discrete correctness with partial credit or format penalties. The remaining two (MMBench $+2.1$, AI2D $+2.2$) use more uniform accuracy-only scoring and accordingly show smaller gains. This gradient tracks the stage-(S1) prediction: the less calibrated the reward distribution, the larger the headroom recovered by Eq. (5).

G Deferred Proofs and Extended Analysis

This appendix collects the formal arguments referenced in the main text. To keep the notation light, conditioning on $(x, y_{<t})$ is suppressed throughout; subscripts and conditioning sets are restored only when needed for disambiguation. All sums over v run over the vocabulary $\mathcal{V}$, and expectations $\mathbb{E}_r$ are taken under $P(r \mid x)$ unless noted otherwise.

G.1 Proof of Theorem 1

Throughout, we keep the outer expectation $\mathbb{E}_x$ over prompts implicit and work pointwise in x and $y_{<t}$. By definition of the OPSD loss in Eq. (2) with $D = D_{\mathrm{KL}}$, dropping the per-token average $1/|y|$ and the trajectory-level expectation that are common to both $\mathcal{L}_{\mathrm{OPSD}}$ and $\mathcal{L}^*$, the integrand at position t reads

$$\ell_{\mathrm{OPSD}}(\theta;\, x, y_{<t}) \;=\; \mathbb{E}_r\left[\sum_v P_T(v \mid r)\, \log \frac{P_T(v \mid r)}{P_S(v)}\right]. \tag{11}$$

Insert $\bar{P}_T(v)/\bar{P}_T(v)$ inside the logarithm and use $\log(ab) = \log a + \log b$ to split the integrand into two terms:

$$\ell_{\mathrm{OPSD}} \;=\; \underbrace{\mathbb{E}_r\left[\sum_v P_T(v \mid r) \log \frac{P_T(v \mid r)}{\bar{P}_T(v)}\right]}_{\text{(I)}} \;+\; \underbrace{\mathbb{E}_r\left[\sum_v P_T(v \mid r) \log \frac{\bar{P}_T(v)}{P_S(v)}\right]}_{\text{(II)}}. \tag{12}$$

Term (I) is the r-expectation of $D_{\mathrm{KL}}(P_T(\cdot \mid r) \,\|\, \bar{P}_T)$, which by definition equals the conditional mutual information $I(Y_t;\, R \mid X,\, Y_{<t})$. Term (II) has an integrand that does not depend on r, so the expectation acts only on $P_T(v \mid r)$, and using $\mathbb{E}_r[P_T(v \mid r)] = \bar{P}_T(v)$ yields

$$\text{(II)} \;=\; \sum_v \bar{P}_T(v)\, \log \frac{\bar{P}_T(v)}{P_S(v)} \;=\; D_{\mathrm{KL}}(\bar{P}_T \,\|\, P_S) \;=\; \ell^*(\theta;\, x, y_{<t}). \tag{13}$$

Restoring the outer $\mathbb{E}_x$ and the trajectory-level averaging on both sides gives $\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathcal{L}^*(\theta) + I(Y_t;\, R \mid X,\, Y_{<t})$ since the mutual-information term does not depend on θ and is preserved under both operations. □

G.2 Proof of Proposition 1

The per-sample OPSD gradient at realization r is, by definition,

$$g(\theta; r) \;=\; -\sum_{v} P_T(v \mid r)\, \nabla_\theta \log P_S(v). \tag{14}$$

Add and subtract $\bar{P}_T(v)\, \nabla_\theta \log P_S(v)$ inside the sum:

$$g(\theta; r) \;=\; -\sum_{v} \bar{P}_T(v)\, \nabla_\theta \log P_S(v) \;-\; \sum_{v} \big[P_T(v \mid r) - \bar{P}_T(v)\big]\, \nabla_\theta \log P_S(v), \tag{15}$$

which gives the decomposition $g(\theta; r) = g^\star(\theta) + \delta(\theta; r)$ stated in the proposition.

Property (i): zero mean of δ. Taking $\mathbb{E}_r$ on the deviation term and noting that $\mathbb{E}_r[P_T(v \mid r)] = \bar{P}_T(v)$ by definition,

$$\mathbb{E}_r[\delta(\theta; r)] \;=\; -\sum_{v} \big(\mathbb{E}_r[P_T(v \mid r)] - \bar{P}_T(v)\big)\, \nabla_\theta \log P_S(v) \;=\; 0. \tag{16}$$

Property (ii): variance scaling. Write $\delta(\theta; r) = -\sum_v \eta_r(v)\, \nabla_\theta \log P_S(v)$ with $\eta_r(v) \triangleq P_T(v \mid r) - \bar{P}_T(v)$. Because $\mathbb{E}_r[\eta_r(v)] = 0$ for every v, the cross-covariance form gives

$$\mathbb{E}_r\|\delta(\theta; r)\|^2 \;=\; \sum_{u,v} \mathrm{Cov}_r\big[\eta_r(u),\, \eta_r(v)\big]\, \big\langle \nabla_\theta \log P_S(u),\, \nabla_\theta \log P_S(v) \big\rangle. \tag{17}$$

Applying the Cauchy–Schwarz inequality to each cross term and using $|\mathrm{Cov}_r[\eta_r(u), \eta_r(v)]| \le \sqrt{\mathrm{Var}_r[\eta_r(u)]\, \mathrm{Var}_r[\eta_r(v)]}$ together with $\sum_{u,v} \sqrt{a_u a_v} \le |\mathcal{V}| \sum_v a_v$ for non-negative a_v, we obtain the upper bound

$$\mathbb{E}_r\|\delta(\theta; r)\|^2 \;\le\; |\mathcal{V}| \sum_{v} \mathrm{Var}_r\big[P_T(v \mid r)\big]\, \|\nabla_\theta \log P_S(v)\|^2. \tag{18}$$

The right-hand side vanishes iff $\mathrm{Var}_r[P_T(v \mid r)] = 0$ for every v, i.e., iff $P_T(v \mid r) = \bar{P}_T(v)$ almost surely under $P(r \mid x)$, which is equivalent to $I(Y_t; R \mid X, Y_{<t}) = 0$. By non-negativity, $\delta \equiv 0$ almost surely under the same condition; whenever the conditional mutual information is positive, the bound (and hence the variance) is strictly positive and grows monotonically with it. □

G.3 Bayesian Reading of the Evidence Ratio

We give the formal statement underlying the Bayesian reading of Δ_t used in Section 4.3.

Theorem 2 (Per-Token Evidence Ratio)**.** *Assume the privileged trace r and the next token y_t admit a joint distribution under which $P(r \mid x, y_{<t}) > 0$. Then*

$$\Delta_t \;=\; \log P_T(y_t) - \log P_S(y_t) \;=\; \log \frac{P(r \mid x, y_{\le t})}{P(r \mid x, y_{<t})}. \tag{19}$$

Proof. Apply Bayes' rule to $P_T(y_t) = P(y_t \mid x, r, y_{<t})$:

$$P(y_t \mid x, r, y_{<t}) \;=\; \frac{P(r \mid x, y_{\le t})\, P(y_t \mid x, y_{<t})}{P(r \mid x, y_{<t})} \;=\; P_S(y_t) \cdot \frac{P(r \mid x, y_{\le t})}{P(r \mid x, y_{<t})}. \tag{20}$$

Taking logarithms of both sides and rearranging gives the claim. □

In words, Δ_t measures the multiplicative belief update on r obtained by appending the token y_t to the prefix; a positive Δ_t means that emitting y_t increases the posterior probability that r is consistent with the trajectory, and a negative Δ_t that it decreases it. This view makes precise why Δ_t is a clean per-token credit signal for the trajectory's compatibility with the privileged trace.

G.4 LEAKAGE-FREE PROPERTY OF TACE

We now record the formal statement that justifies the term "leakage-free" used in the main text.

Definition 1 (Leakage-free token-level estimator)**.** *A token-level advantage estimator $\hat{A}_{i,t}$ is* leakage-free *if, for every $(\theta, x, y_{<t})$, the sign of $\hat{A}_{i,t}$ does not depend on the privileged context r.*

Theorem 3 (TACE is leakage-free)**.** *The TACE advantage of Eq.* (9) *satisfies Definition 1: for every realization of r used to compute P_T, the sign of $\hat{A}_{i,t}$ equals $\mathrm{sign}(A_i)$, which is determined by the verifier and not by r.*

Proof. By Eq. (7), the per-token weight is $w_{i,t} = \exp(\mathrm{sign}(A_i) \cdot \Delta_t) > 0$ for any $\Delta_t \in \mathbb{R}$. The clipping operator in Eq. (8) maps $w_{i,t}$ into the interval $[1 - \epsilon_w,\, 1 + \epsilon_w] \subset (0, 2)$ for any $\epsilon_w \in (0, 1)$, which remains strictly positive. The composed estimator is $\hat{A}_{i,t} = A_i \cdot \mathrm{clip}(w_{i,t})$, so $\mathrm{sign}(\hat{A}_{i,t}) = \mathrm{sign}(A_i) \cdot \mathrm{sign}(\mathrm{clip}(w_{i,t})) = \mathrm{sign}(A_i)$. Since $A_i = \Phi^{-1}(\widehat{F}_\tau(R_i))$ depends only on the verifier reward and on the empirical task CDF, neither of which involves r, the sign of $\hat{A}_{i,t}$ is independent of r at every θ. □

The reverse implication, namely that any token-level estimator satisfying Definition 1 avoids the per-sample $\delta(\theta; r)$ deviation of Proposition 1, also holds: that deviation arises only when the per-token weight in the gradient depends on r, and direction independence with respect to r blocks exactly that channel.

H LIMITATIONS AND FUTURE WORK

We close with a brief discussion of the boundaries of our study and of the extensions we view as most natural.

Limitations. Our evaluation uses a single backbone family (Qwen3-VL-8B-Instruct) and a single training corpus (MMFineReason-123K), so the reported gains are informative about one operating point rather than about the full space of multimodal policies. Within this operating point, all runs share a fixed 4096-token response budget, a fixed group size of G=8, and a fixed asymmetric PPO clip $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$=$(0.2, 0.28)$; we have not swept these hyperparameters jointly with the TACE-specific knobs (ϵ_w, the warm-up length, and the teacher refresh period), and the interaction surface is correspondingly under-characterized. The quantile baseline at stage (S1) is evaluated online within each group, which is fully specified in the finite-G regime used here but may require the optional running-histogram smoothing we describe in Section 4.2 when group sizes are appreciably smaller. The privileged context r is in every experiment a verified final answer, and we do not exercise the estimator against noisy, partial, or adversarially perturbed privileged signals. Finally, our reported wall-clock overhead of $< 20\%$ reflects the specific implementation stack (VERL and EasyR1) on an H200 cluster with on-device teacher forwards; a different execution backend could shift the overhead either way without affecting any of our analytical conclusions.

Future work. Each of the above boundaries points to a concrete next step. The first direction is scale: applying TACE to larger and more architecturally diverse VLM backbones, and to text-only and video-conditioned policies, will test whether the two invariants (I1, I2) continue to govern behavior once the reward topology is even more heterogeneous than the five-to-nine-benchmark regime studied here. A second direction is the structure of the privileged context: because stage (S2) reads r only through the scalar Δ_t, the construction is in principle compatible with multi-trace ensembles, partial hints, retrieved references, or tool-use transcripts, and a systematic study of which privileged contexts are most informative per unit of compute is a natural follow-up. A third direction is the reward signal itself; the 1D Brenier map at stage (S1) is the specialization of a more general optimal-transport construction, and extending it to vector-valued verifiers via multi-marginal or entropic OT would let the same two-invariant design absorb multi-objective rewards (for example joint correctness, calibration, and format scores) without per-task tuning. Fourth, the leakage-free guarantee of Theorem 3 is structural rather than statistical, and we expect it to interact cleanly with offline RL and with preference optimization objectives that likewise need to separate magnitude from direction; formalizing that interaction is a promising theoretical extension. Taken together, these directions

keep the same skeleton intact (rank-based sample baselining composed with a sign-anchored token weight) while broadening the regimes in which the estimator applies.