TFPO: TOKEN-LEVEL OBJECTIVE FUSION FOR STABLE PREFERENCE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference optimization is usually applied uniformly to entire responses, although explanatory text, final answers, formatting fields, and code need not benefit from the same training signal. We propose TFPO (Token-Fused Preference Optimization), which learns a lightweight gate that routes each response token between a DPO-style preference objective and a chosen-response likelihood anchor. Training uses no token-level supervision and includes ratio, smoothness, and entropy regularization. On the matched ten-benchmark suite, TFPO scores 88.00 versus 83.08 for SimPO, the strongest baseline without NLL anchoring, and 83.52 for SimPO+NLL (three-seed means). It also improves external alignment across Qwen, Llama, and Mistral backbones and multimodal performance. Under repeated sampling, TFPO raises answer agreement and majority-answer accuracy while preserving explanation diversity. On 1,000 responses from five extractable-answer tasks, annotated blind to gate values and model identity, its content gate recovers answer spans at 0.85 AUPRC versus 0.55 for the strongest strict position-only control. These results support token-level objective routing as a practical way to improve both preference alignment and answer stability.

1 INTRODUCTION

Preference optimization has become a standard post-training stage for large language models (LLMs) and multimodal large language models (MLLMs). Methods such as DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024), and ORPO (Hong et al., 2024) offer efficient alternatives to reinforcement learning from human feedback. Yet most preference objectives still treat a completion as a homogeneous sequence that receives one scalar preference decision. This creates a coarse credit-assignment problem for a functionally heterogeneous output.

LLM responses are not homogeneous token strings. Their spans range from explanatory prose and answer-bearing fields to formatting delimiters, schema keys, and executable code. These regions should not necessarily receive the same optimization signal. Reasoning tokens can benefit from preference shaping: they determine whether an explanation is helpful, concise, and well justified. In contrast, final answer fields, option letters, short numeric answers, JSON keys, or code fragments often need likelihood anchoring to remain parseable and stable. When all tokens inherit the same sequence-level preference pressure, these roles can interfere: the model may learn a more preferred style while destabilizing the small span that actually determines correctness.

We introduce **Token-Fused Preference Optimization (TFPO)**, a token-level objective routing framework for this functional heterogeneity. TFPO learns a small gate over response tokens and uses it to route tokens between two objectives: a preference objective for reasoning-sensitive regions and a chosen-response likelihood anchor for answer-bearing regions. The method does not require explicit labels for reasoning or conclusion spans. Instead, the gate learns a latent soft partition that minimizes the fused objective.

The central design question is *where* each objective should act within a response. TFPO treats this choice as token-level functional credit assignment, allocating preference shaping and answer anchoring across different spans. This routing design aims to preserve final-answer stability while improving alignment quality. Figure 2 gives a visual overview.

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.

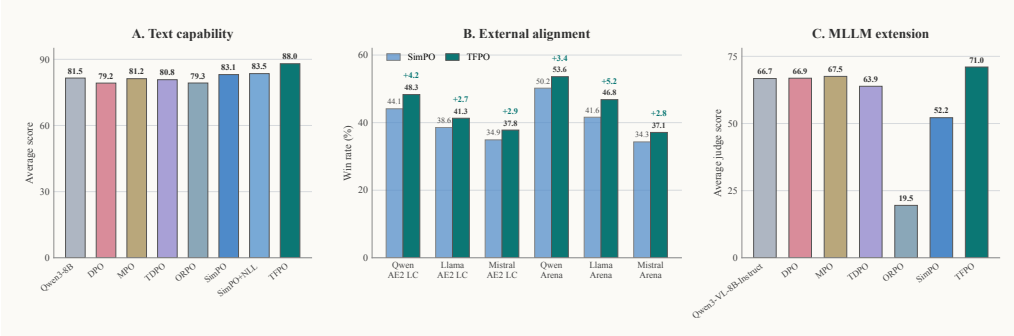


Figure 1: Summary of the main empirical picture. Left: TFPO achieves the best main ten-benchmark average among all compared methods, outperforming both SimPO and the NLL-anchored SimPO+NLL baseline on Qwen3-8B. Middle: TFPO consistently improves over SimPO on judge-based alignment benchmarks across Qwen, Llama, and Mistral backbones. Right: under multimodal preference tuning, TFPO yields the best average among the compared methods.

Our experiments evaluate TFPO across diverse settings and examine the behavior of its learned route. Across text capability, cross-family alignment, and multimodal preference tuning, TFPO improves strong preference-optimization baselines; targeted routing analyses and repeated sampling support content-sensitive allocation and more stable answers.

The broader diagnostic lesson is that aggregate capability alone does not reveal route-mass violations, excessive local switching, or premature saturation. The targeted diagnostics in Appendix D show that TFPO's regularizers control these complementary routing properties. Preference optimization should ask not just whether overall performance improves, but where optimization pressure is applied within the response.

Contributions. We make three contributions:

1. We formulate stable preference alignment as a token-level functional credit-assignment problem and instantiate it as TFPO, a routing method between preference optimization and likelihood anchoring.
2. We introduce a lightweight gate regularization recipe—ratio, smoothness, and entropy control—that controls route mass, local switching, and premature saturation while remaining easy to implement.
3. We provide broad empirical evidence across text, cross-family external alignment, and multimodal settings, together with strict position controls, gate-blind gold-span annotations, counterfactuals, and repeated-sampling stability tests that directly probe the routing claim.

2 Related Work

Preference alignment with learned rewards. RLHF pipelines train reward models from comparisons and optimize policies with PPO-style updates (Schulman et al., 2017; Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022); RLAIF and constitutional variants substitute AI feedback or rules (Bai et al., 2022). Recent work revisits the update rule through REINFORCE- and GRPO/RLVR-style reasoning pipelines (Ahmadian et al., 2024; Shao et al., 2024; DeepSeek-AI et al., 2025; Yu et al., 2025). TFPO instead uses offline preference pairs, while sharing the view that scalar rewards can hide unstable credit assignment.

Sequence-level preference optimization. DPO (Rafailov et al., 2023) learns from chosen and rejected responses relative to a frozen reference. SLiC-HF (Zhao et al., 2023), RRHF (Yuan et al., 2023), IPO (Azar et al., 2023), KTO (Ethayarajh et al., 2024), WPO (Zhou et al., 2024), SimPO (Meng et al., 2024), and ORPO (Hong et al., 2024) alter the surrogate, reference usage, weighting, or normalization, but still assign credit at response level.

Table 1: Positioning of TFPO among preference-optimization methods. The key distinction is not whether a method uses preference data, but whether it can assign different objectives to different token roles inside the same response.

Method	Granularity	Token-level mechanism	Reference?	NLL anchor?
DPO	Sequence-level	None	Yes	No
SimPO	Sequence-level	None	No	No
ORPO	Sequence-level	None	No	Yes, chosen NLL
MPO-style	Sequence-level	Mixed objectives on full sequence	Yes	Yes, full-response SFT/NLL
TDPO	Token-level	Tokenized preference objective	Yes	No
TIS-DPO	Token-level	Token importance weighting inside DPO	Yes	No
TFPO	Token-level	Functional objective routing via gate	Yes	Yes, token-routed chosen NLL

Mixed-objective preference training. Mixed-loss recipes combine preference optimization with auxiliary generation losses. MPO-style training in recent multimodal reasoning work (Wang et al., 2024b), for example, mixes preference and SFT-style objectives to reduce the instability of preference-only tuning. TFPO shares the motivation that preference optimization can disturb answer quality, but differs by applying objective mixing at token granularity rather than uniformly across the whole completion.

Token-level preference modeling. TDPO (Zeng et al., 2024) and TIS-DPO (Liu et al., 2025) both move beyond treating the response as an atomic unit. TDPO formulates token-level preference optimization, while TIS-DPO estimates token importance weights inside a preference loss. TFPO is complementary but structurally different: it does not only reweight tokens within a preference objective; it routes tokens *between* preference and likelihood objectives.

Stability under alignment. Preference tuning can improve style or helpfulness while hurting exact answer formatting, output schemas, or short-answer correctness. We use *stability* to mean reduced answer/format drift under preference training, especially for final answer fields that are easy to parse and verify. This framing is central to TFPO: the method is designed to preserve stable conclusions while leaving reasoning tokens available for preference shaping. One alternative is to identify final-answer spans with task-specific parsers or templates, but such supervision is brittle across open-ended, code, schema, and multimodal settings. TFPO instead learns a latent route from preference data without requiring token-level span labels.

Table 1 summarizes the gap that TFPO targets. Prior work either improves the sequence-level preference surrogate, mixes objectives at full-response granularity, or reweights tokens inside a preference loss. TFPO instead treats the output as functionally heterogeneous: different token roles can be assigned to different objectives inside the same response.

3 METHOD

3.1 SETUP

Let x be a prompt, y^+ a preferred response, and y^- a rejected response. We train a policy model π_θ initialized from an instruction-tuned checkpoint, with a frozen reference model π_{ref} when using the reference-based objective. Standard DPO forms a preference margin from the full log-probability of each response. This is elegant, but it also binds heterogeneous token roles to one scalar preference decision: a final numeric answer, a chain-of-thought explanation, a markdown delimiter, and a code token all contribute to the same sequence-level margin.

TFPO changes this credit-assignment step. It keeps the familiar preference-pair setup, but treats the response as a mixture of functional token roles and asks the model to learn a soft route over them. Tokens that look like reasoning, elaboration, or preference-sensitive style are routed toward the preference objective. Tokens that look like final answer fields, schema fields, fragile output formats, or code/output tokens are routed toward a chosen-response likelihood anchor. The partition is latent: no training example tells the model which tokens are "reasoning" or "conclusion" tokens. In this sense, TFPO is not simply a token weighting variant; it is an instance of token-level objective routing for functionally heterogeneous completions.

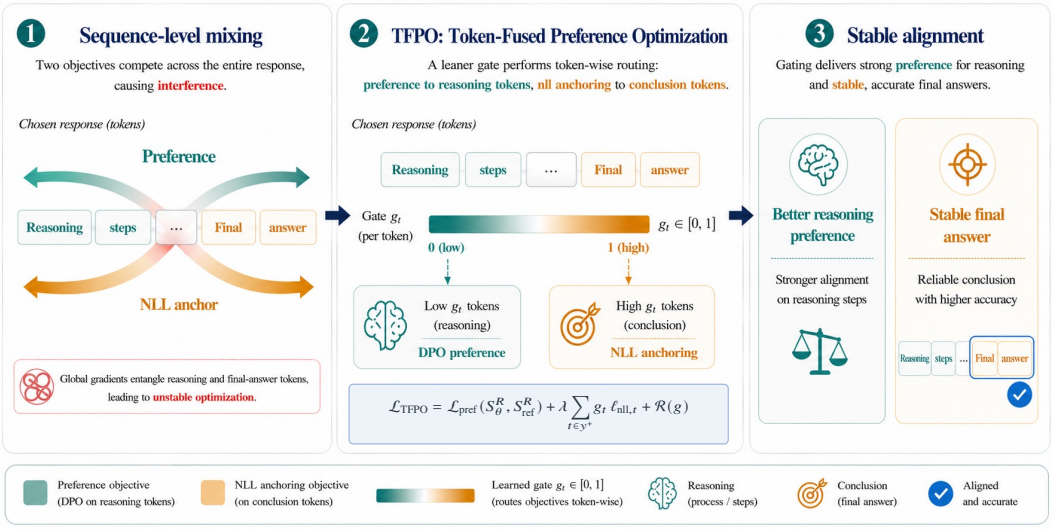


Figure 2: Overview of TFPO. A learned gate routes response tokens between preference shaping and likelihood anchoring, targeting reasoning-sensitive and answer-bearing regions, respectively. The displayed loss is schematic: routed DPO sequence scores are defined in Eqs. (2)–(4), and Eq. (9) gives the full objective.

3.2 Token Gate

For each response token, TFPO predicts a scalar gate

$$g_t = \sigma(\text{sg}(\mathbf{h}_t)\mathbf{W}_g + b_g) \in (0, 1), \tag{1}$$

where $\mathbf{h}_t$ is the decoder hidden state, $\text{sg}(\cdot)$ denotes stop-gradient, and $\mathbf{W}_g, b_g$ are a lightweight gate head. We use the convention that a large g_t routes token t toward the conclusion / likelihood objective, while a small g_t routes it toward the reasoning / preference objective.

Two design choices are worth spelling out. First, the default gate is deliberately small: a single linear head over hidden states. This makes it cheap and limits the risk that the gate becomes a second model. Second, we stop gradients through $\mathbf{h}_t$ on the gate path. This blocks gradients from gate predictions into the backbone while preserving gradients from the policy losses; Appendix A gives the exact computational graph.

3.3 Token-Level Objective Fusion

From sequence scores to token-routed scores. The SFT/NLL loss is naturally token-level: each token has a local negative log-likelihood. DPO, by contrast, first sums token log-probabilities into one sequence score and then applies a pairwise sigmoid. TFPO keeps this DPO structure, but replaces the unweighted sequence sum with a token-weighted sum. Define conclusion and reasoning weights as $w^C_t = g_t$ and $w^R_t = 1 - g_t$. For any response y, the reasoning-weighted log-probability is

$$S^R_\theta(y \mid x) = \sum_{t=1}^{T_y} (1 - g_t) \log \pi_\theta(y_t \mid x, y_{<t}). \tag{2}$$

Tokens with low gate values contribute strongly to this score; tokens with high gate values are mostly excluded from the preference comparison. This is the central move of TFPO: it does not change the preference-pair data format, but it changes which functional regions are responsible for the preference margin.

Preference objective. For the preference term, we replace the four full-response log-probabilities in DPO with the reasoning-weighted scores:

$$\Delta^{TFPO}_\theta = \left[S^R_\theta(y^+ \mid x) - S^R_\theta(y^- \mid x)\right] - \left[S^R_{\text{ref}}(y^+ \mid x) - S^R_{\text{ref}}(y^- \mid x)\right], \tag{3}$$

and the preference loss is

$$\mathcal{L}_{\text{pref}}^{TFPO} = -\log \sigma\big(\beta \Delta_\theta^{TFPO}\big)\,. \tag{4}$$

This preference term is therefore not a sum of independent token-level preference losses. The gate first forms reasoning-weighted sequence scores, and those scores then enter the same DPO-style pairwise sigmoid. The chosen and rejected responses each receive their own gate values from their own hidden states; there is no token-to-token alignment assumption between them.

We reuse the policy-side weights for the corresponding reference log-probabilities and treat those weights as constants for the frozen reference model. In implementation, the reused policy-side gates are detached when multiplying reference log-probabilities, so the reference score does not introduce additional gradients through the frozen reference term. This keeps policy and reference scores in the same routed token subspace. If the policy score were routed but the reference score remained unweighted, conclusion tokens could leak back into the preference margin through the reference term.

Likelihood objective. The likelihood term provides the complementary signal. It is computed only on the chosen response:

$$\mathcal{L}_{\text{nll}}^{TFPO} = \sum_{t=1}^{T_{y^+}} g_t \Big(-\log \pi_\theta(y_t^+ \mid x, y_{<t}^+) \Big). \tag{5}$$

This chosen-side NLL term anchors tokens that should remain stable, such as final answers, schema fields, and code-like output fragments. For chosen responses, the gate allocates token mass between the routed preference margin and the likelihood anchor. Rejected tokens participate only in the routed preference margin and are never imitated.

3.4 Regularization and full objective

The objective above admits trivial shortcuts. If every token receives the preference objective, TFPO reduces toward token-weighted DPO and loses answer anchoring. If every token receives the NLL objective, it becomes mostly imitation and loses preference learning. If the gate changes every token, the routing becomes hard to interpret and unstable. We therefore use three simple regularizers.

Ratio regularization. Let $\bar{g} = \frac{1}{T}\sum_t g_t$. The ratio constraint keeps the conclusion-side mass inside a broad range:

$$\mathcal{L}_{\text{len}} = \big[\max(0, r_{\min} - \bar{g})\big]^2 + \big[\max(0, \bar{g} - r_{\max})\big]^2. \tag{6}$$

The bounds are intentionally loose: they control route mass without prescribing where conclusion tokens should be routed.

Smoothness regularization. The smoothness penalty is a total-variation term:

$$\mathcal{L}_{\text{tv}} = \sum_{t=2}^{T} |g_t - g_{t-1}|. \tag{7}$$

This encourages contiguous spans rather than salt-and-pepper token assignments, which also makes the learned routing easier to inspect in heatmaps.

Entropy regularization. Finally, we maximize a small Bernoulli entropy term early in training:

$$\mathcal{L}_{\text{ent}} = \frac{1}{T}\sum_{t=1}^{T} \left[-g_t \log g_t - (1-g_t)\log(1-g_t)\right]. \tag{8}$$

The entropy term discourages premature hard decisions while the two objectives are calibrating.

Full objective. The full objective is

$$\mathcal{L}^{TFPO} = \lambda_{\text{pref}}\mathcal{L}_{\text{pref}}^{TFPO} + \lambda_{\text{nll}}\mathcal{L}_{\text{nll}}^{TFPO} + \lambda_{\text{len}}\mathcal{L}_{\text{len}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}} - \lambda_{\text{ent}}\mathcal{L}_{\text{ent}}. \tag{9}$$

Our default recipe uses $\beta = 0.1$, $\lambda_{\text{pref}} = 1$, $\lambda_{\text{nll}} = 0.2$, $[r_{\min}, r_{\max}] = [0.05, 0.9]$, with TV and entropy regularization enabled. These settings combine soft objective routing with constraints on route mass, smoothness, and early saturation. Appendix A gives the exact computational graph and implementation notes; Appendix D ablates each regularizer and all three together, linking $\mathcal{L}_{\text{len}}$ to route-budget compliance, TV to local smoothness, and entropy to avoiding early saturation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and training data. The main comparison uses the post-trained `Qwen/Qwen3-8B` checkpoint (Yang et al., 2025). All preference methods are trained on the same cleaned UltraFeedback-style preference split (Cui et al., 2023) with a shared backbone, data budget, and LoRA setup; method-specific objective settings and exact run details are in Appendix B and Tables 6–7. Appendix C reports robustness on a larger 2M preference mixture. Cross-family checks use Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) under the same external alignment protocol. The multimodal extension trains preference methods from Qwen3-VL-8B-Instruct (Bai et al., 2025).

Baselines. We compare TFPO with DPO, MPO-style mixed preference optimization, SimPO, TDPO, ORPO, and SimPO+NLL as an NLL-anchored baseline. The off-the-shelf model is included when it is the natural initialization baseline.

Capability benchmarks. We evaluate on ten benchmarks spanning reasoning, instruction following, QA/exam tasks, math, and code: BBH (Suzgun et al., 2022), GSM8K (Cobbe et al., 2021), HumanEval (Chen et al., 2021), IFEval (Zhou et al., 2023), MMLU (Hendrycks et al., 2021a), RACE (Lai et al., 2017), C3 (Sun et al., 2020), Gaokao (Zhang et al., 2023), MATH (Hendrycks et al., 2021b), and MBPP (Austin et al., 2021). The MATH column uses `EleutherAI/hendrycks_math`, not MATH-500; Appendix E reports a separate standard MATH-500 bridge. The main aggregate is the arithmetic mean across the ten benchmark scores, computed before display rounding. We use one frozen, method-independent prompt, decoding, extraction, and scoring protocol per benchmark; all fine-tuned values in the capability table are equal-weight means over 3 training seeds. BBH, GSM8K, and MATH use fixed 200-prompt held-out subsets with four aligned samples per prompt and report AvgAcc@4. HumanEval and MBPP use 20 samples per evaluated problem and the standard unbiased pass@1 estimator; the remaining tasks use one sample per evaluated item. Generative tasks use Qwen3's non-thinking sampling configuration and a 32,768-token output cap; Table 7 records the task protocols, and Appendix E reports the generation-completeness audit.

External alignment benchmarks. For AlpacaEval 2.0 (Dubois et al., 2024), we generate model outputs in the official format and evaluate with the official length-controlled annotator configuration `weighted_alpaca_eval_gpt4_turbo_new`. For Arena-Hard-v2.0-Preview (Li et al., 2024), we generate official-format answers and use the official GPT-4.1 judge with markdown and length controls.

Multimodal benchmarks. For Qwen3-VL-8B-Instruct, we evaluate ten multimodal benchmarks: MathVista (Lu et al., 2023), MathVision (Wang et al., 2024a), MMMU (Yue et al., 2024), AI2D (Kembhavi et al., 2016), MMVet (Yu et al., 2024), LLaVA-Bench (Liu et al., 2023a), MMBench (Liu et al., 2023b), MMStar (Chen et al., 2024), POPE (Li et al., 2023), and HallusionBench (Guan et al., 2024). We use `gpt-4o-mini` judge-based scoring and report the VLM Score for LLaVA-Bench rather than the relative score.

4.2 MAIN CAPABILITY RESULTS

Figure 1 summarizes the main results, and Table 2 gives the primary capability comparison. Under the shared evaluation protocol, TFPO achieves the highest ten-benchmark average. SimPO+NLL provides a conventional full-response NLL-anchored baseline.

The improvements span instruction following, math, general academic reasoning, reading/exam-style tasks, and code generation. TFPO improves both HumanEval and MBPP, consistent with the motivation for likelihood anchoring of answer-bearing outputs. Overall, it exceeds SimPO by 4.92 points and the NLL-anchored SimPO+NLL baseline by 4.48 points under the shared evaluation protocol.

Table 2: Matched capability evaluation on post-trained Qwen3-8B. Fine-tuned rows are 3-seed means under one frozen, method-independent protocol; BBH, GSM8K, and MATH use fixed 200-prompt held-out subsets. Best values are bolded and second-best values are underlined.

Method	Avg.	BBH	GSM8K	HumanEval	IFEval	MMLU	RACE	C3	Gaokao	MATH	MBPP
Qwen3-8B	81.50	79.2	92.1	78.4	<u>83.2</u>	<u>79.5</u>	84.1	88.5	70.2	83.7	76.1
DPO	79.22	78.4	88.2	78.0	80.3	75.2	84.5	87.1	66.2	81.5	72.8
MPO	81.22	79.4	89.2	84.1	80.9	75.0	86.0	88.0	70.4	83.0	76.2
SimPO	83.08	<u>80.5</u>	92.4	86.1	82.1	78.1	88.0	<u>91.2</u>	70.6	84.3	77.5
TDPO	80.76	79.8	90.6	82.0	81.3	76.0	85.3	89.0	68.3	82.1	73.2
ORPO	79.26	78.4	88.0	80.3	79.6	75.0	84.0	88.6	66.0	80.7	72.0
SimPO+NLL	<u>83.52</u>	<u>80.5</u>	<u>92.5</u>	<u>86.5</u>	82.3	79.0	<u>88.5</u>	91.0	<u>71.4</u>	<u>85.3</u>	<u>78.2</u>
TFPO	**88.00**	**86.0**	**96.1**	**91.3**	**88.1**	**84.6**	**91.0**	**94.3**	**76.0**	**89.6**	**83.0**

Table 3: Cross-family external alignment results. AlpacaEval 2.0 uses the official length-controlled weighted GPT-4-Turbo annotator configuration; Arena-Hard-v2.0-Preview uses the official GPT-4.1 judge with markdown and length controls. For AlpacaEval, LC and WR denote length-controlled and raw win rate; Arena-Hard WR is the official control-adjusted win rate.

Method	Qwen3-8B			Llama-3.1-8B			Mistral-7B-v0.3		
	AlpacaEval		Arena-Hard	AlpacaEval		Arena-Hard	AlpacaEval		Arena-Hard
	LC	WR	WR	LC	WR	WR	LC	WR	WR
SFT/Instruct	19.8	31.2	34.1	7.1	23.1	27.5	11.3	19.2	13.8
DPO	24.1	34.6	38.0	10.6	25.7	30.1	12.6	22.7	19.8
MPO	30.8	37.2	41.5	23.4	28.8	32.4	17.4	25.0	20.1
SimPO	44.1	49.7	50.2	38.6	44.9	41.6	34.9	39.1	34.3
TDPO	36.9	41.6	42.1	27.1	38.2	35.3	25.1	36.2	25.7
ORPO	17.4	19.7	36.1	9.4	8.7	28.6	10.1	12.5	14.9
TFPO	**48.3**	**56.1**	**53.6**	**41.3**	**48.6**	**46.8**	**37.8**	**41.0**	**37.1**

4.3 External alignment across model families

Table 3 tests whether the gains persist on open-ended judge-based alignment benchmarks and across model families. TFPO is strongest on every reported AlpacaEval 2.0 and Arena-Hard-v2.0-Preview column. This matters because AlpacaEval 2.0 length control reduces verbosity bias, while Arena-Hard-v2.0-Preview is a harder pairwise benchmark with official length and markdown controls.

The cross-family pattern is important for the method claim. The gains replicate when the same routing recipe is trained independently on Qwen, Llama, and Mistral backbones, suggesting that the result is not tied to one tokenizer, one instruction-tuning recipe, or one model family. Appendix E reports the 3-seed robustness check, while Appendix C separately reports external alignment on the larger preference mixture.

4.4 Core routing evidence

We next examine content-dependent routing. Table 4 compares the learned gate with strict position-only controls using three complementary evaluations independent of the capability-scoring pipeline. Among initially correct responses, RR-Preserve measures answer retention after removing low-score reasoning-route tokens; CR-Failure (conclusion-removal failure) measures answer failure after removing high-score conclusion-route tokens.

The strict position-only controls read only normalized position and response-length features. Across equal deletion budgets, the content gate improves over the stronger position MLP by 12.3 RR-Preserve and 21.5 CR-Failure points. Under method-independent gold removal on each variant's own responses, its joint score is 84.4 versus 69.9. On the separate gold-recovery evaluation, it improves by 0.30 AUPRC (95% response-bootstrap CI $[0.27, 0.33]$) and 0.22 F1 ($[0.19, 0.25]$). Appendix D additionally reports a *content+position* auxiliary-feature variant; because it reads hidden-state content, it is not used as position-only evidence.

Table 4: Core routing evidence from three complementary evaluations comparing the TFPO content gate with strict position-only controls. Panel A averages equal-token-budget interventions over 10–40% removal budgets. Panel B uses method-independent gold spans for removal. Panel C evaluates gate scores against blind answer-span annotations on five extractable-answer tasks.

A. Equal-removal-budget intervention (mean over 10–40% budgets)

Router	RR-Preserve↑	CR-Failure↑	ΔRR vs. MLP↑	ΔCR vs. MLP↑
Pure-position linear	68.9	41.2	–	–
Pure-position MLP	72.6	46.0	0.0	0.0
TFPO content gate	**84.9**	**67.5**	**+12.3**	**+21.5**

B. Method-independent gold-span intervention

Routing variant	Reasoning preserve↑	Answer failure↑	Joint geom.↑	ΔJoint vs. MLP↑
Pure-position linear route	73.1	59.1	65.7	–
Pure-position MLP route	76.5	63.8	69.9	0.0
Content-gated route	**88.6**	**80.4**	**84.4**	**+14.5**

C. Gold answer-span recovery (1,000 responses)

Router	AUPRC↑	Precision↑	Recall↑	F1↑
Pure-position linear	0.50	0.58	0.52	0.55
Pure-position MLP	0.55	0.63	0.57	0.60
TFPO content gate	**0.85**	**0.84**	**0.80**	**0.82**

4.5 Answer-position robustness and stability

We further examine how the learned gate responds to changes in answer position and whether the trained models produce stable answers under repeated sampling.

Gold spans and position controls. Two annotators, blind to gate values and model identity, label attempted answer-bearing spans independently on 300 responses before adjudication ($\kappa = 0.92$, span IoU $= 0.89$). The final evaluation contains 1,000 responses (200 from each of five extractable-answer tasks) after 18 prespecified, score-blind integrity exclusions. The 1,000-response gold labels are used only for evaluation; AUPRC is the primary threshold-free metric, and the F1 threshold is fixed once on disjoint development data. Table 4 shows that TFPO reaches 0.85 AUPRC versus 0.55 for the strongest strict position-only MLP. In a content-preserving counterfactual that moves the same answer to the beginning, middle, or end, TFPO remains at 0.83/0.84/0.85 AUPRC, whereas a pure-position linear router varies from 0.25 to 0.62. The near-constant TFPO scores across placements are consistent with routing by answer content; the position-only router varies strongly with location.

Figure 3 illustrates the same pattern qualitatively: the matched-budget position baseline places mass at a fixed head or tail span, whereas the learned gate better tracks local answer-bearing structure. Together with the quantitative position controls and independently annotated gold spans in Table 4, this supports content-sensitive rather than positional routing. The sensitivity row is descriptive; semantic evaluation uses the gold spans.

Final-answer stability and explanation diversity. For BBH, GSM8K, and MATH, we draw $K = 4$ matched non-thinking responses per prompt for each of 3 training seeds. AnswerAgreement@4 is the modal canonical-answer fraction; MajorityAcc@4 scores the unique plurality answer against gold; AvgAcc@4 is mean per-generation correctness; and ExplanationDiv@4 is pairwise diversity after removing final-answer fields. Appendix A gives the edge-case rules.

Relative to SimPO+NLL, TFPO improves answer agreement by 2.96 points, majority accuracy by 4.61 points, and average accuracy by 4.46 points, with higher explanation diversity. These gains hold on each of the three tasks. At matched 10–40% deletion budgets on initially correct responses, high-score conclusion-route removal causes substantially more answer failures than low-score reasoning-route removal (67.5% versus 15.1% on average). Together with the blind gold-span recovery in Table 4, this supports functional answer-bearing routing rather than stylistic segmentation. Appendix E gives per-task results, a MATH-500 robustness check, strict-position threshold/equal-budget sweeps, and method-independent gold interventions.

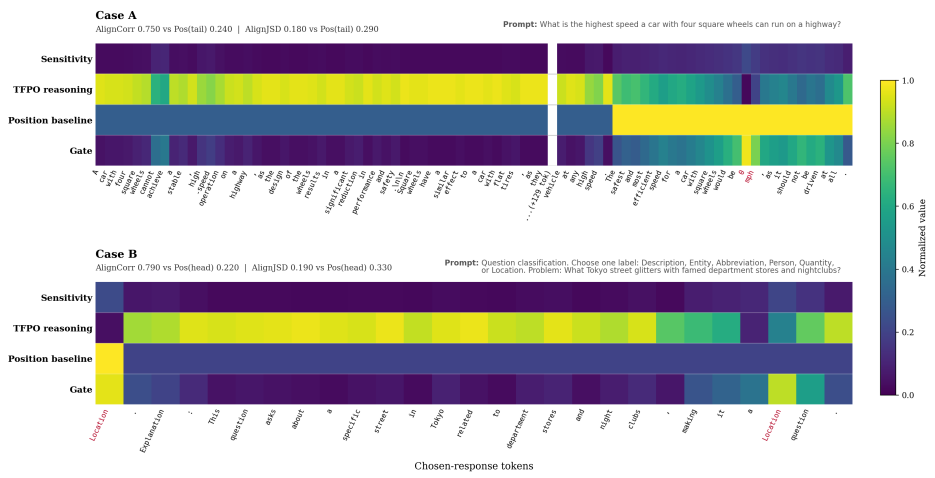


Figure 3: Held-out token-routing examples. *Sensitivity* denotes the normalized absolute policy–reference log-probability difference $|\Delta \log p|$. Other rows show TFPO reasoning weights, a matched-budget position baseline, and the gate. The gate highlights a final span in Case A and the answer-first label in Case B.

Table 5: Repeated-sampling stability on BBH, GSM8K, and MATH. Results are averaged first over tasks within each seed and then over 3 training seeds. TFPO achieves the highest answer agreement and accuracy while also improving explanation diversity. All methods use the same frozen answer canonicalizer, with missing/invalid generations retained in the denominator.

Method	Agreement@4↑	MajorityAcc@4↑	AvgAcc@4↑	Div.@4↑	Valid↑
SimPO	88.33	80.67	85.74	59.01	99.67
SimPO+NLL	89.01	81.72	86.10	59.72	99.71
TFPO	**91.97**	**86.33**	**90.56**	**62.33**	**99.79**

4.6 Multimodal Extension

Multimodal preference tuning amplifies the need for token-level routing because VLM benchmarks combine visual reasoning with fragile answer fields such as OCR/counting answers, option letters, structured VQA fields, and yes/no judgments. TFPO is therefore a natural fit: visual reasoning tokens can be preference-shaped while final answers and schema-like tokens remain likelihood anchored. Figure 1 summarizes the result: TFPO reaches the strongest multimodal average (71.03), improving over MPO by 3.49 points and over Qwen3-VL-8B-Instruct by 4.29 points. Appendix F gives the full table and qualitative examples of format drift, wrong option selection, and excessive verbosity.

5 Limitations and Conclusion

Limitations and broader impact. Gold-span semantics are evaluated only on five tasks with extractable answer fields and should not be read as a unique partition for arbitrary open-ended text. The learned route is latent and not necessarily identifiable, and our model-scale evidence is limited to 7–8B LoRA settings. Judge-based evaluations remain judge-dependent. TFPO can improve parseability and answer stability, but it does not remove risks such as hallucination, bias, or misuse.

Conclusion. TFPO reframes preference optimization as token-level credit assignment between preference shaping and likelihood anchoring. Capability and repeated-sampling results show improvements in performance and answer stability. Strict position-only controls, blind gold spans, and answer-position counterfactuals provide complementary evidence that the learned route tracks answer-bearing content. Cross-family and multimodal results further support the effectiveness of this design across the tested model families and modalities.

AI USE STATEMENT

Generative AI tools were used for language editing, translation, and structural suggestions, and for limited feedback during experiment planning, result interpretation, and selected code implementation and consistency checks. The authors made all methodological decisions, conducted the experiments, performed the primary analyses, and determined the scientific claims and conclusions. All AI-assisted materials were reviewed by the authors, and reported values were checked against the underlying experiment records. The authors take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

Sections 4–5 specify the evaluation scope and claims; Appendix B records frozen training, checkpoint-selection, decoding, and high-level scoring settings; and the remaining appendices provide per-task robustness and mechanism audits. The anonymized supplementary package includes explicit training entry points, evaluation utilities, provenance schemas, and scripts for stability and multimodal aggregation.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to Basics: Revisiting REINFORCE-Style Optimization for Learning from Human Feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program Synthesis with Large Language Models. *arXiv preprint arXiv:2108.07732*, 2021. URL https://arxiv.org/abs/2108.07732.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A General Theoretical Paradigm to Understand Learning from Human Preferences. *arXiv preprint arXiv:2310.12036*, 2023. URL https://arxiv.org/abs/2310.12036.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting Language Models with Scaled AI Feedback. *arXiv preprint arXiv:2310.01377*, 2023. URL https://arxiv.org/abs/2310.01377.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. *arXiv preprint arXiv:2404.04475*, 2024. URL https://arxiv.org/abs/2404.04475.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model Alignment as Prospect Theoretic Optimization. *arXiv preprint arXiv:2402.01306*, 2024. URL https://arxiv.org/abs/2402.01306.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint arXiv:2103.03874*, 2021b. URL https://arxiv.org/abs/2103.03874.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic Preference Optimization without Reference Model. *arXiv preprint arXiv:2403.07691*, 2024. URL https://arxiv.org/abs/2403.07691.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram is Worth a Dozen Images. In *Computer Vision – ECCV 2016*, pp. 235–251, 2016. doi: 10.1007/978-3-319-46493-0_15. URL https://link.springer.com/chapter/10.1007/978-3-319-46493-0_15.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, 2017. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082/.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline. *arXiv preprint arXiv:2406.11939*, 2024. URL https://arxiv.org/abs/2406.11939.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Xiaoming Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, Philip Yu, and Meng Cao. TIS-DPO: Token-level Importance Sampling for Direct Preference Optimization With Estimated Weights. In *International Conference on Learning Representations*, pp. 51339–51368, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7fb9f39075a5202472676a7531568212-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*, 2023b. URL https://arxiv.org/abs/2307.06281.

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. *arXiv preprint arXiv:2310.02255*, 2023. URL https://arxiv.org/abs/2310.02255.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple Preference Optimization with a Reference-Free Reward. *arXiv preprint arXiv:2405.14734*, 2024. URL https://arxiv.org/abs/2405.14734.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290*, 2023. URL https://arxiv.org/abs/2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155, 2020. doi: 10.1162/tacl_a_00305. URL https://aclanthology.org/2020.tacl-1.10/.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *arXiv preprint arXiv:2210.09261*, 2022. URL https://arxiv.org/abs/2210.09261.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. *arXiv preprint arXiv:2402.14804*, 2024a. URL https://arxiv.org/abs/2402.14804.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization. *arXiv preprint arXiv:2411.10442*, 2024b. URL https://arxiv.org/abs/2411.10442.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

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 arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 57730–57754, 2024. URL https://proceedings.mlr.press/v235/yu24o.html.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. RRHF: Rank Responses to Align Language Models with Human Feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/23e6f78bdec844a9f7b6c957de2aae91-Abstract-Conference.html.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level Direct Preference Optimization. *arXiv preprint arXiv:2404.11999*, 2024. URL https://arxiv.org/abs/2404.11999.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the Performance of Large Language Models on GAOKAO Benchmark. *arXiv preprint arXiv:2305.12474*, 2023. URL https://arxiv.org/abs/2305.12474v3.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence Likelihood Calibration with Human Feedback. *arXiv preprint arXiv:2305.10425*, 2023. URL https://arxiv.org/abs/2305.10425.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A Scalable lightWeight Infrastructure for Fine-Tuning. *arXiv preprint arXiv:2408.05517*, 2024. URL https://arxiv.org/abs/2408.05517.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911*, 2023. URL https://arxiv.org/abs/2311.07911.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. WPO: Enhancing RLHF with Weighted Preference Optimization. *arXiv preprint arXiv:2406.11827*, 2024. URL https://arxiv.org/abs/2406.11827.

APPENDIX

- Appendix A: Method details and derivations
- Appendix B: Experimental details and reproducibility
- Appendix C: Robustness to a larger preference mixture
- Appendix D: Ablations and training diagnostics
- Appendix E: Full mechanistic evaluation results
- Appendix F: MLLM evaluation details and qualitative examples

A METHOD DETAILS AND DERIVATIONS

Exact routed computation. The policy and reference scores use the same numerical routing weights but different gradient paths. For either chosen or rejected response y, the implementation computes

$$S_\theta^R(y \mid x) = \sum_t (1 - g_t) \log \pi_\theta(y_t \mid x, y_{<t}),$$

$$S_{\text{ref}}^R(y \mid x) = \sum_t \text{sg}(1 - g_t) \log \pi_{\text{ref}}(y_t \mid x, y_{<t}),$$

where sg denotes stop-gradient. These four sequence scores enter the single DPO-style sigmoid in Eqs. (3)–(4). The reference model is frozen and the detached reference weights do not update the policy gate. The chosen-only NLL branch is the weighted token sum in Eq. (5); rejected tokens are never imitated. Regularizers are computed separately on chosen and rejected gates and then averaged.

Descriptive policy–reference diagnostic. For the descriptive diagnostic in Figure 3, define $\delta_t = \log \pi_\theta(y_t \mid x, y_{<t}) - \log \pi_{\text{ref}}(y_t \mid x, y_{<t})$. AlignCorr is the response-averaged Spearman correlation between the normalized $|\delta_t|$ map and the normalized reasoning allocation $(1 - g_t)$; higher values indicate more similar token rankings. AlignJSD is the response-averaged Jensen–Shannon divergence between the same normalized maps; lower values indicate more similar token-weight distributions. We use these diagnostics only to describe policy–reference shift; semantic correspondence is evaluated separately against gold spans.

Gold answer-span evaluation. Gold labels mark the span in which a response attempts to express the task-schema answer, independent of whether that attempted answer is correct. Two annotators are blind to model identity and all gate scores. AUPRC is the primary threshold-free measure; the F1 threshold is selected once on disjoint development data. The evaluation is restricted to GSM8K, MATH, RACE, ARC-Challenge, and BoolQ, where an attempted answer-bearing span is operationally extractable. It does not assume a unique semantic partition for arbitrary open-ended responses.

Stability and interventions. For canonical answer equivalence $f(\cdot)$ and K generations, Answer-Agreement is

$$\text{Agreement@}K(x) = \max_{a \in \mathcal{A}_{\text{valid}}} \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}[f(y^{(k)}) = a].$$

Here $\mathcal{A}_{\text{valid}}$ contains only valid canonical answers. Missing/invalid generations remain in the denominator but cast no vote; an all-invalid prompt has agreement zero. MajorityAcc is one only when a unique plurality answer is gold-equivalent (ties are zero); AvgAcc averages the K correctness indicators. ExplanationDiv is the mean pairwise token-set Jaccard distance over frozen reasoning text after final-answer-field stripping. Valid-answer rate is the fraction of generations with extractor status `valid` and a canonical answer. Tables abbreviate AnswerAgreement, ExplanationDiv, and valid-answer rate as Agreement, Div., and Valid. On responses that are correct before intervention, RR-Preserve asks whether deleting low-score reasoning-route tokens preserves the answer, whereas CR-Failure asks whether deleting high-score conclusion-route tokens breaks it. We report threshold and equal-token-budget sweeps using each router's own scores, plus method-independent gold-span removal; no sequence-level baseline is evaluated with a shared TFPO segmenter.

Implementation notes. Before routing, the implementation applies the same deterministic short-response safeguard to every routed TFPO variant. For a response of length $T \geq 2$, it clamps the response mean $\bar{g}$ to $[1/T, 1 - 1/T]$, shifts every response-token gate by the resulting mean difference, and clips only for numerical safety. Thus each objective retains approximately one token of soft mass even when the ratio penalty is weak on a short response. The gate head is attached as lightweight trainable state before LoRA wrapping. TFPO disables padding-free RLHF and precomputed reference log-probabilities because the reference terms must be dynamically reweighted by the policy gate. SFT replay, when enabled, bypasses the gate and contributes only plain NLL.

B EXPERIMENTAL DETAILS AND REPRODUCIBILITY

Training split. The main experiments use 60k cleaned UltraFeedback-style preference pairs. This setting supports controlled comparisons, ablations, and repeated mechanism evaluations at a manageable training cost. The larger 2M preference mixture is reported separately as a robustness check.

Compute. All reported training runs use NVIDIA A100-80GB GPUs. The main setup runs for one epoch on 8 GPUs and takes roughly 3 hours (24 A100 GPU-hours per run); the larger 2M mixture runs for one epoch on 16 GPUs and takes roughly 50 hours (800 GPU-hours per run). In the matched 60k efficiency audit, TFPO uses 3.00 hours and 66.16 GiB peak memory per GPU versus DPO's 2.83 hours and 63.0 GiB—a 1.06× time and 1.05× memory overhead. Mechanism evaluation reuses the same checkpoint family and runs offline on cached benchmark subsets.

Table 6: Frozen run card for main training and checkpoint selection. Baselines use the shared launcher settings unless the method-specific objective requires otherwise.

Item	Setting
Optimizer / schedule	AdamW through the MS-SWIFT trainer (Zhao et al., 2024); learning rate 1×10^{-4}; cosine scheduler; warmup ratio 0.05.
Batching	Eight GPUs; LoRA micro-batch 2 per GPU with gradient accumulation 2, giving effective global batch size 32.
Length / precision	Maximum sequence length 4096; bfloat16 training; gradient checkpointing enabled.
LoRA	Rank 8, alpha 32, dropout 0.05, target modules `all-linear`.
Training preprocessing	`qwen3_nothinking` template with `add_non_thinking_prefix=false`; the same tokenizer and preprocessing are used across methods.
Training horizon	One epoch; evaluation and checkpointing every 10% of the epoch; step 0 is evaluated for diagnostics but is not a selection candidate; save limit 100.
Search budget	Broad methods receive the same nine-configuration search budget; NLL-anchored baselines use one fixed configuration over 3 seeds. TDPO retains its prescribed extra SFT stage, so search counts rather than total training stages are matched.
Validation / checkpoint selection	A fixed set of 2,000 preference pairs, disjoint from training and reported tests, selects among the ten checkpoints at $10\%, 20\%, \ldots, 100\%$. Within each method and seed, the selector is $0.5\,\mathrm{rank}(\text{pairwise accuracy}) + 0.5\,\mathrm{rank}(-\text{chosen-token-mean NLL})$, using percentile ranks over those ten candidates; exact ties select the earlier checkpoint. Reported test benchmarks never enter selection.
TFPO objective	$\beta = 0.1$, $\lambda_{\mathrm{pref}} = 1.0$, $\lambda_{\mathrm{nll}} = 0.2$, $\lambda_{\mathrm{len}} = 1.0$, $\lambda_{\mathrm{tv}} = 0.02$, entropy init $\lambda_{\mathrm{ent}} = 0.01$.
Gate regularization	Linear gate with stop-gradient; $r_{\min} = 0.05$, $r_{\max} = 0.9$; entropy weight decays linearly from 0.01 to 0 over the first 35% of training. The short-response safeguard keeps approximately one token of soft mass in each branch.

Table 7: Frozen run card for objective controls and evaluation protocols.

Item	Setting
NLL-anchored baselines	DPO+NLL uses $\beta = 0.1$ and `rpo_alpha=0.2`; SimPO+NLL uses $\beta = 2.0$, $\gamma = 1.0$, and `cpo_alpha=0.2`. Both add chosen-response token-mean NLL with coefficient 0.2 and contain no gate or TFPO routing regularizers ($\mathcal{L}_{\text{len}}$, $\mathcal{L}_{\text{tv}}$, or entropy).
Reference handling	Reference-based methods use a frozen reference model; SimPO, SimPO+NLL, and ORPO are reference-free. In TFPO, policy-side reasoning weights are reused for reference log-probabilities and detached in the reference term.
Capability evaluation	Qwen3 non-thinking mode; temperature 0.7, top-$p = 0.8$, top-$k = 20$, min-$p = 0$, presence penalty 1.5, max new tokens $32{,}768$. BBH/GSM8K/MATH use fixed 200-prompt held-out subsets and report AvgAcc@4 ($n = 4$); HumanEval/MBPP use the standard unbiased pass@1 estimator ($n = 20$); remaining tasks use $n = 1$. Prompt/scoring protocols are 3-shot CoT for BBH, 4-shot CoT for GSM8K, 5-shot for MMLU, 3-shot for MBPP, official HumanEval/IFEval checkers, fixed-choice extraction for RACE/C3/Gaokao, and boxed-answer normalization for MATH.
External alignment decoding	AlpacaEval 2.0: temperature 0, top-$p = 1$, max new tokens 1024; Arena-Hard-v2.0-Preview: temperature 0, top-$p = 1$, max new tokens 4096.
Mechanism decoding	Stability uses 200 held-out prompts per task and $K = 4$ samples under the non-thinking capability protocol above on BBH/GSM8K/MATH for each of 3 training seeds. Interventions report a 0.3–0.7 threshold sweep and equal-token budgets of 10–40%.
MATH-500 bridge	Standard 500-example test split with $n = 8$ aligned samples. Non-thinking uses the capability decoding above; thinking uses temperature 0.6, top-$p = 0.95$, top-$k = 20$, min-$p = 0$, and presence penalty 0. Both use max new tokens $32{,}768$; the length audit additionally evaluates caps 512 and $4{,}096$.

External benchmark workflow. We generate AlpacaEval 2.0 and Arena-Hard-v2.0-Preview model outputs in their official answer formats and evaluate them with the corresponding public judging pipelines. AlpacaEval uses `weighted_alpaca_eval_gpt4_turbo_new`; Arena-Hard-v2.0-Preview uses the official GPT-4.1 judge with markdown and length controls. We report AlpacaEval length-controlled win rate (LC) and raw win rate (WR), and the official control-adjusted Arena-Hard win rate (WR).

Benchmark subsets. For BBH, GSM8K, and MATH, we randomly sample 200 held-out prompts per benchmark using a fixed seed and reuse the same subsets across methods and training seeds.

Loss normalization. TFPO uses the weighted token-sum NLL in Eq. (5), averaged over response pairs in a minibatch. DPO+NLL and SimPO+NLL use full-response token-mean NLL with the coefficients specified in their respective training recipes.

Supplementary material. The supplementary package includes TFPO training and evaluation code, external-benchmark generation utilities, canonical-generation and bootstrap schemas, an offline stability reanalysis entry point, and the gold-span annotation protocol and worker disclosure. Asset sources and licenses are listed in `reproducibility/asset_license_inventory.md`.

C Robustness to a larger preference mixture

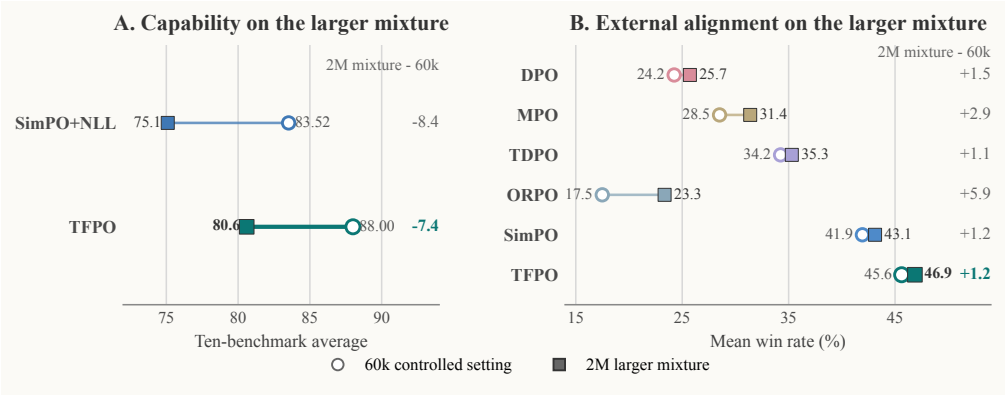


Figure 4: Robustness to a larger 2M preference mixture. Left: matched ten-benchmark averages for SimPO+NLL and TFPO in the controlled 60k setting and under the larger mixture. Right: external-alignment composites for the same settings, computed from AlpacaEval 2.0 LC/WR and Arena-Hard-v2.0-Preview WR across Qwen, Llama, and Mistral. Connecting lines compare the controlled 60k setting with the larger mixture, whose composition also differs.

Table 8: Qwen3-8B capability comparison between the controlled 60k setting and the larger preference mixture. Both columns use the same non-thinking evaluation.

Method	60k avg.↑	Larger-mixture avg.↑	Parser-invalid↓
SimPO+NLL	83.52	75.1	$\leq 1.4\%$
TFPO	**88.00**	**80.6**	$\leq 1.4\%$
TFPO gap	**+4.48**	**+5.5**	–

Figure 4 and Table 8 show that the main conclusion remains robust under the larger preference mixture: TFPO stays 5.5 points above the NLL-anchored baseline, with parser-invalid outputs at or below 1.4%.

External alignment under the larger mixture. Table 9 reports judge-based evaluation under the same larger preference mixture. The pattern is consistent with the controlled 60k setting: TFPO remains strongest across all three model families and all reported external-alignment metrics.

Table 9: External alignment under the larger 2M preference mixture. AlpacaEval 2.0 uses LC/raw WR; Arena-Hard-v2.0-Preview reports its official control-adjusted WR.

Method	Qwen3-8B			Llama-3.1-8B			Mistral-7B-v0.3		
	AlpacaEval		Arena-Hard	AlpacaEval		Arena-Hard	AlpacaEval		Arena-Hard
	LC	WR	WR	LC	WR	WR	LC	WR	WR
SFT/Instruct	19.8	31.2	34.1	7.1	23.1	27.5	11.3	19.2	13.8
DPO	26.5	35.8	40.3	12.4	28.4	32.4	14.6	23.2	17.9
MPO	35.1	41.7	46.5	25.1	29.8	35.1	19.1	26.5	23.6
SimPO	46.3	52.4	52.6	39.1	43.2	44.1	34.1	40.1	36.2
TDPO	36.7	42.6	46.0	31.1	37.6	36.3	27.2	34.6	25.6
ORPO	25.8	26.1	36.2	19.8	21.2	30.6	15.2	16.8	18.4
TFPO	**49.1**	**57.2**	**56.7**	**43.7**	**49.1**	**49.9**	**35.2**	**43.0**	**37.9**

D Ablations and training diagnostics

Secondary routing diagnostics. Table 10 complements the strict position-only comparison in Table 4. These rows directly evaluate where a route places its mass against the same blind gold spans. Constant, random, and shuffled assignments all preserve some aspect of the gate budget

while breaking token–content correspondence. The *content+position* variant reads both hidden-state content and normalized position, so it is an auxiliary-feature diagnostic rather than a strict position-only control.

Table 10: Secondary routing diagnostics on the 1,000-response blind gold-span evaluation. The strict pure-position linear/MLP rows are the controls used in the main claim; content+position is not a position-only baseline. Bold marks the best learned route.

Route / score	Construction	Gold AUPRC↑
Constant-g	Same route mass at every response position	0.18
Matched-ratio random	Same per-response budget, randomly assigned tokens	0.18
Shuffled learned route	Same learned gate-value multiset, permuted token locations	0.24
Matched-tail heuristic	Fixed tail allocation with matched budget	0.42
Pure-position linear	Normalized position only	0.50
Pure-position MLP	Position/length features only; strongest strict-position control	0.55
Content+position	Hidden-state content plus normalized position	0.83
TFPO content gate	Hidden-state content; no explicit position feature	**0.85**

What each term controls. Table 11 isolates the failure mode targeted by each regularizer. Budget violation is the fraction of responses whose mean gate falls outside $[r_{\min}, r_{\max}]$; transitions/100 counts thresholded route switches at the fixed 0.5 threshold per 100 response tokens; intra-span TV averages $|g_t - g_{t-1}|$ over adjacent tokens sharing the same gold label. Early saturation is the fraction of tokens with $g < 0.05$ or $g > 0.95$ at 10% training progress. Removing $\mathcal{L}_{\text{len}}$, with the shared one-token-per-branch safeguard retained, increases budget violations from 1.6% to 36.1%; removing TV increases assignment transitions from 3.1 to 15.8 per 100 tokens and intra-span TV from 0.022 to 0.126; removing entropy raises early saturation from 6.4% to 52.1%. Figure 5 visualizes these three intended effects. Across 3 training seeds, the default configuration produced no non-finite values or branch-budget collapse.

Table 11: Regularizer-specific diagnostics. Each ablation is evaluated on the failure mode its removed term is designed to control. Values are means over 3 training seeds. The one-token-per-branch safeguard is retained in every row.

Variant	Gold AUPRC↑	Budget violation↓	Transitions/100↓	Intra-span TV↓	Early saturation↓
Default	**0.85**	**1.6%**	**3.1**	**0.022**	**6.4%**
w/o $\mathcal{L}_{\text{len}}$	0.56	36.1%	6.3	0.049	26.9%
w/o TV	0.68	3.5%	15.8	0.126	11.4%
w/o entropy	0.59	17.1%	8.9	0.071	52.1%
w/o all regularizers	0.42	49.2%	18.2	0.151	66.7%

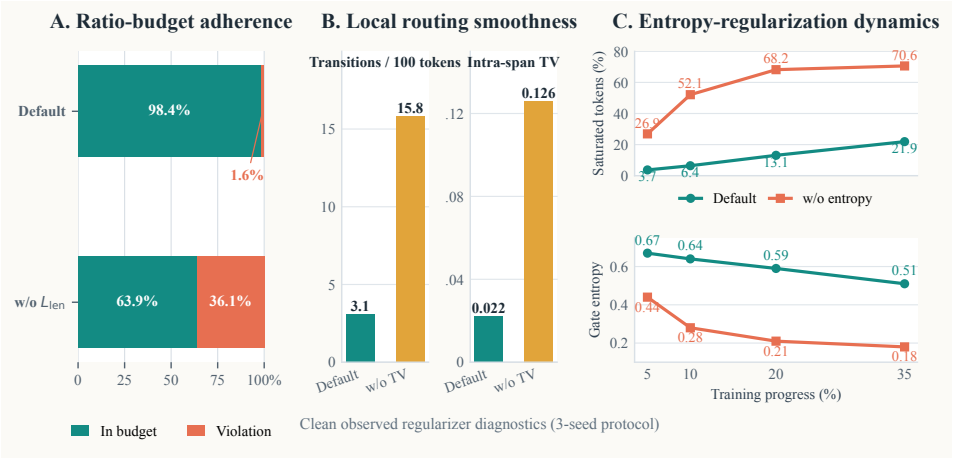


Figure 5: Regularizer-specific diagnostics. A: 98.4% of default responses satisfy the gate budget versus 63.9% without $\mathcal{L}_{\text{len}}$. B: removing TV increases transitions/100 tokens and intra-span TV from $3.1/0.022$ to $15.8/0.126$. C: without entropy regularization, 52.1% of tokens are saturated at 10% training progress, versus 6.4% by default.

E Full Mechanistic Evaluation Results

Final-answer stability. Table 12 expands the compact scorecard in Table 5. The same matched generation records drive every metric, with aggregation performed prompt → task → training seed. Relative to SimPO and SimPO+NLL, TFPO improves answer agreement, majority-answer accuracy, average accuracy, and explanation diversity on BBH, GSM8K, and MATH, showing that the stability gain is consistent across tasks without reducing explanation diversity.

Table 12: Per-task repeated-sampling stability on 200 held-out prompts per task, averaged over 3 training seeds. Each prompt has $K = 4$ aligned non-thinking samples; missing/invalid answers remain in the denominator. Overall is the task macro-average; all values are percentages.

Split	Method	Agreement@4↑	MajorityAcc@4↑	AvgAcc@4↑	Div.@4↑	Valid↑
Overall	SimPO	88.33	80.67	85.74	59.01	99.67
	SimPO+NLL	89.01	81.72	86.10	59.72	99.71
	TFPO	**91.97**	**86.33**	**90.56**	**62.33**	**99.79**
BBH	SimPO	84.04	80.00	80.50	59.03	99.71
	SimPO+NLL	84.42	80.50	80.50	60.00	99.71
	TFPO	**88.00**	**85.00**	**86.00**	**61.97**	**99.79**
GSM8K	SimPO	94.96	90.00	92.42	61.03	99.71
	SimPO+NLL	95.58	90.50	92.50	61.00	99.71
	TFPO	**97.71**	**94.00**	**96.08**	**64.03**	**99.79**
MATH	SimPO	86.00	72.00	84.29	56.97	99.58
	SimPO+NLL	87.04	74.17	85.29	58.17	99.71
	TFPO	**90.21**	**80.00**	**89.58**	**61.00**	**99.79**

Stability comparisons, MATH-500, and replication across model families. Table 13 summarizes complementary checks. Panel A reports differences between the three-seed task-macro means; these are point estimates, not seed-level significance tests. Panel B repeats the stability test at $K = 8$ on all 500 test prompts of standard MATH-500. Panel C reports 3-seed external-alignment composites across Qwen, Llama, and Mistral, showing consistent gains across all three backbones.

Table 13: Robustness of the stability and cross-family results. External composite is the mean of AlpacaEval 2.0 LC/raw WR and Arena-Hard-v2.0-Preview control-adjusted WR. Panels A–B use equal-weight means over 3 training seeds for fine-tuned methods; the Qwen3-8B reference in Panel B uses one checkpoint. Values after $\pm$ in Panel C are sample standard deviations over 3 training seeds. Scores are percentages; gaps are percentage points computed before rounding.

A. Three-task $K = 4$ mean differences

Comparison	Agreement gap	MajAcc gap	AvgAcc gap	Div. gap
TFPO − SimPO	+3.64	+5.67	+4.82	+3.32
TFPO − SimPO+NLL	+2.96	+4.61	+4.46	+2.61

B. Standard MATH-500, non-thinking $K = 8$

Method	AvgAcc@8↑	Agreement@8↑	MajorityAcc@8↑	Div.@8↑	Valid↑
Qwen3-8B	87.25	90.45	95.80	64.25	99.70
SimPO+NLL	86.00	89.75	95.13	63.40	99.70
TFPO	**90.40**	**92.52**	**97.20**	**65.30**	**99.80**

C. 3-seed external-alignment composite

Backbone	SimPO	TFPO	Gap
Qwen3-8B	48.0 ± 0.4	52.7 ± 0.4	+4.7
Llama-3.1-8B-Instruct	41.7 ± 0.3	45.6 ± 0.4	+3.9
Mistral-7B-Instruct-v0.3	36.1 ± 0.3	38.6 ± 0.3	+2.5

Functional token interventions. RR-Preserve asks whether the answer survives after low-score reasoning-route tokens are removed; CR-Failure asks whether deleting high-score conclusion-route tokens breaks it. Table 14 evaluates this signature with router-specific and gate-blind gold-span interventions. Fixed-threshold and equal-budget sweeps use each router's own scores, while the final panel applies gate-blind gold answer spans to each method or routed variant's own responses. TFPO beats the stronger position MLP on both metrics at all five thresholds and all four deletion budgets.

Table 14: Threshold, equal-token-budget, and method-independent gold interventions. Panels A–B use each learned router's own scores; Panel C applies gate-blind gold answer spans to each method or routed variant's own responses. Joint geom. is the geometric mean of the two percentage rates.

A. Fixed-threshold sweep

Threshold	Pure-position linear		Pure-position MLP		TFPO	
	RR↑	CR↑	RR↑	CR↑	RR↑	CR↑
0.3	73.7	33.3	77.2	38.5	**91.1**	**62.7**
0.4	69.9	29.6	73.4	34.7	**86.2**	**55.9**
0.5	66.5	27.4	69.8	31.5	**81.98**	**47.94**
0.6	60.2	23.9	63.7	27.6	**78.3**	**42.1**
0.7	54.1	18.7	57.2	22.1	**72.8**	**36.5**

B. Equal-removal-budget sweep

Removed tokens	Pure-position linear		Pure-position MLP		TFPO	
	RR↑	CR↑	RR↑	CR↑	RR↑	CR↑
10%	78.6	24.7	82.2	29.5	**94.2**	**47.9**
20%	72.9	35.5	76.5	40.2	**89.3**	**62.8**
30%	65.4	46.1	69.2	51.0	**82.1**	**75.2**
40%	58.5	58.3	62.4	63.2	**74.0**	**84.1**
Mean	68.9	41.2	72.6	46.0	**84.9**	**67.5**

C. Method-independent gold-span intervention

Method / router	Reasoning preserve↑	Answer failure↑	Joint geom.↑
DPO+NLL	69.2	54.7	61.5
SimPO+NLL	71.9	57.1	64.1
TFPO (pure-position linear route)	73.1	59.1	65.7
TFPO (pure-position MLP route)	76.5	63.8	69.9
TFPO (content-gated route)	**88.6**	**80.4**	**84.4**

Table 16 reports the population and annotation audit together with natural-response recovery and answer-position counterfactual results.

MATH-500 protocol and length audit. Table 15 documents the bridge between our main MATH column and the standard MATH-500 reference used in Qwen3 reporting. The main table uses `EleutherAI/hendrycks_math`; the bridge uses all 500 examples of the standard MATH-500 test split and mode-specific Qwen3 sampling. With prompts and sampling indices fixed, varying only the output cap confirms that 512 tokens is restrictive for every method. At 32,768 tokens, truncation is negligible and TFPO remains above SimPO+NLL in both modes. Across the full ten-task capability suite, overall truncation, missing-answer, and parser-invalid rates are at most 0.1%, 0.4%, and 0.5%, respectively, and no individual per-task rate among these categories exceeds 0.8%.

Table 15: MATH-500 model/mode and output-length audit. AvgAcc@8 averages correctness over eight aligned samples; empirical pass@8 is the fraction of problems with at least one correct sample. Trunc. denotes `finish_reason=length` or reaching the cap without EOS. Missing counts all generations for which the frozen extractor finds no answer, including truncated outputs; it may overlap with Trunc. Fine-tuned values are equal-weight means over 3 training seeds.

A. Mode-specific MATH-500 bridge at max new tokens 32,768

Method	AvgAcc@8↑	Emp. pass@8↑	Trunc.↓	Missing↓	Median/p95 tokens
Non-thinking					
Qwen3-8B	87.25	99.8	0.1%	0.3%	281/953
SimPO+NLL	86.0	99.6	0.1%	0.3%	272/850
TFPO	**90.4**	**99.9**	0.1%	**0.2%**	**255/725**
Thinking					
Qwen3-8B	97.2	**100.0**	0.2%	0.2%	502/4,203
SimPO+NLL	97.0	99.9	0.2%	0.2%	481/4,502
TFPO	**97.5**	**100.0**	**0.1%**	**0.1%**	**433/3,703**

B. Length sensitivity with aligned prompts and sampling indices
Output caps: 512 / 4096 / 32768 tokens (values in this order)

Method	AvgAcc@8↑	Trunc.↓	Missing↓
Non-thinking			
Qwen3-8B	82.0/87.2/87.25	18.2/0.4/0.1%	3.0/0.4/0.3%
SimPO+NLL	82.5/85.9/86.0	14.3/0.3/0.1%	2.5/0.4/0.3%
TFPO	**86.0/89.9/90.4**	10.1/0.8/0.1%	**1.5/0.2/0.2%**
Thinking			
Qwen3-8B	76.0/92.0/97.2	48.5/5.4/0.2%	8.0/1.1/0.2%
SimPO+NLL	77.0/90.8/97.0	45.7/6.6/0.2%	7.0/1.2/0.2%
TFPO	**81.0/93.8/97.5**	**38.3/4.2/0.1%**	**4.0/0.5/0.1%**

Table 16: Gold-span population, annotation, recovery, and position audit. AUPRC pools token labels/scores over responses; each paired bootstrap replicate resamples responses and recomputes the pooled metric, retaining structurally valid all-`other` responses.

A. Prespecified, model- and score-blind population audit

Record status	**Count**
Candidate responses	1,018
Exact duplicate removed	9
Irrecoverable token/gate alignment mismatch removed	5
Corrupted/incomplete serialization removed	4
Removed using correctness, extraction, identity, score, or downstream metric	0
Final population (five tasks $\times$ 200)	**1,000**

B. Annotation audit before adjudication

Item	**Value**
Double-annotated responses	300
Cohen's κ	0.92
Span IoU	0.89
Median boundary disagreement	1 token
Answer-bearing token prevalence	0.18

C. Natural-response recovery and paired response-bootstrap differences

Router	**AUPRC↑**	**Precision↑**	**Recall↑**	**F1↑**
Pure-position MLP	0.55	0.63	0.57	0.60
TFPO content gate	**0.85**	**0.84**	**0.80**	**0.82**
TFPO − MLP (95% CI)	$+0.30\ [0.27, 0.33]$	$+0.21\ [0.18, 0.24]$	$+0.23\ [0.20, 0.26]$	$+0.22\ [0.19, 0.25]$

D. Content-preserving answer-position counterfactual (AUPRC)

Router / score	**Answer first**	**Answer middle**	**Answer last**	**Max gap↓**
Matched-tail heuristic	0.18	0.38	0.68	0.50
Pure-position linear	0.25	0.42	0.62	0.37
TFPO content gate	**0.83**	**0.84**	**0.85**	**0.02**

Gold-span annotation protocol and worker information. The annotation rules and worker disclosure are provided in the supplementary file `reproducibility/gold_span_annotation_protocol.md`. The labels were produced by two employees of an external annotation company in mainland China. The rate was RMB 10 per annotation item, and both annotators agreed that their annotations could be used for research. Both annotators were blind to model identity and all gate scores. The 300 double-annotated responses were labeled independently by both annotators before disagreements were adjudicated; before adjudication, Cohen's κ was 0.92, span IoU was 0.89, and the median boundary disagreement was one token.

The candidate order and structural exclusion rules were fixed before inspecting gate values, correctness, extraction outcomes, or downstream metrics. Integrity checks used only candidate IDs, serialization status, and token/offset/gate-array consistency, and were applied once to the shared candidate manifest before any router comparison. Incorrect responses with an attempted answer-bearing span were retained; structurally valid responses without an attempted answer-bearing span were retained with all response tokens labeled `other`. In the counterfactual, reasoning and answer content are held fixed while the answer is moved to the beginning, middle, or end.

F MLLM EVALUATION DETAILS AND QUALITATIVE EXAMPLES

Table 17 gives the per-benchmark breakdown behind the aggregate MLLM result in Section 4.6. The multimodal runs use the processed MM-RLHF and OpenBMB RLAIF-V preference mixture for one epoch with global batch size 16 and maximum sequence length 4096; the ViT and aligner remain frozen. The reported multimodal TFPO run uses $\beta = 0.05$ and $\lambda_{\text{nll}} = 0.2$. We evaluate the `MathVista_MINI`, `MathVision`, `MMMU_DEV_VAL`, `AI2D_TEST`, `MMVet`, `LLaVABench`, `MMBench`, `MMStar`, `POPE`, and `HallusionBench` splits. We report only the judge-based `gpt-4o-mini` MLLM scores in the main text, because strict exact matching underestimates many free-form multimodal answers.

Table 17: Per-benchmark multimodal results for preference methods trained from Qwen3-VL-8B-Instruct. Scores are judge-based with `gpt-4o-mini`.

A. Overall and multimodal reasoning

Method	Avg.	MathVista	MathVision	MMMU
Qwen3-VL-8B-Instruct	66.74	69.20	15.00	52.67
DPO	66.86	**71.30**	17.07	48.67
MPO	67.54	67.60	20.49	52.11
SimPO	52.16	55.40	12.96	38.44
TDPO	63.86	59.88	26.34	48.12
ORPO	19.52	23.60	5.07	0.56
TFPO	**71.03**	70.90	**30.93**	**55.67**

B. General VQA

Method	AI2D	MMVet	LLaVA-Bench	MMBench	MMStar
Qwen3-VL-8B-Instruct	80.57	49.68	**88.80**	84.70	64.87
DPO	81.99	50.64	88.30	84.02	65.87
MPO	85.07	53.62	85.70	85.57	65.73
SimPO	61.14	37.25	69.20	70.57	45.33
TDPO	80.12	50.89	81.20	82.50	59.28
ORPO	3.34	**73.39**	10.00	0.11	0.00
TFPO	**85.90**	59.22	86.30	**87.28**	**69.20**

C. Hallucination

Method	POPE	HallusionBench
Qwen3-VL-8B-Instruct	87.98	73.92
DPO	87.90	72.87
MPO	88.00	71.50
SimPO	84.20	47.11
TDPO	81.26	69.00
ORPO	26.60	52.58
TFPO	**90.35**	**74.50**

Figure 6 presents three complementary cases: Case 1 illustrates answer-format discipline, while Cases 2 and 3 show correct TFPO outputs against a wrong option or excessive verbosity from representative baselines. Together, they illustrate answer formatting, visual grounding, and response concision.

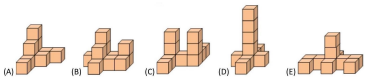
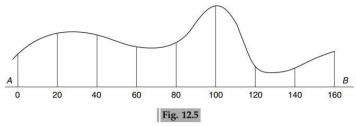
Visual input	Response comparison
Case 1: Visual counting. MathVision, dice solids  	**Failure mode: format drift.** **Question.** Max has 10 dice. Which one of the five solids can he build? The prompt asks for a final option letter. **TFPO raw excerpt.** "To determine which solid Max can build with 10 dice, we need to count the number of dice in each option: (A) The solid has 10 dice. (B) The solid has 11 dice. (C) The solid has 10 dice. (D) The solid has 11 dice. (E) The solid has 10 dice. Max has 10 dice, so he can build solids (A), (C), or (E). However, the question asks for the correct option letter, and the answer is: A." **ORPO raw excerpt.** "After counting the cubes, the fifth solid uses exactly 10 dice. Therefore, Max can build the fifth figure." **Interpretation.** This case concerns final-answer formatting, not visual-counting accuracy: TFPO supplies an explicit option letter, whereas ORPO gives an ordinal description.
Case 2: Map reasoning. MMBench, Silk Road map  	**Failure mode: wrong answer.** **Question.** The map shows Silk Road routes around 1300 CE. Which statement is true: connected East Asia/Middle East/Europe, connected East Asia and the Americas by sea, or only land routes? **TFPO raw excerpt.** "A. The Silk Road connected parts of East Asia, the Middle East, and Europe. This answer is correct based on the map shown. The map clearly illustrates the extensive network of overland routes (shown in dark lines) that connected East Asia with the Middle East and Europe. It also shows some sea routes (in pink) that connected different parts of the network." **TDPO raw excerpt.** "The map shows trade routes extending across oceans and connecting distant continents. Since the routes appear to cross the Pacific Ocean, the Silk Road connected East Asia and the Americas by sea. Answer: B." **Interpretation.** Here the failure is not formatting but content. TDPO produces a plausible-sounding explanation that conflicts with the map, while TFPO keeps the visual conclusion and option letter aligned.
Case 3: Numerical VQA. MMMU, surveying problem  Fig. 12.5 *Point* | *A* | | | | | | | | *B* *Distance* (m) | 0 | 20 | 40 | 60 | 80 | 100 | 120 | 140 | 160 *Offset length* (m) | 23 | 40 | 42 | 30 | 32 | 60 | 10 | 14 | 22	**Failure mode: excessive verbosity in a strong baseline.** **Question.** Use the offsets in the figures to compute the area between line AB and a meandering stream by both the trapezoidal rule and Simpson's rule. **TFPO raw excerpt.** "The offsets are: 23, 40, 42, 30, 32, 60, 10, 14, 22. The interval is h = 20 m. Using the trapezoidal rule: Area = h/2 × [first offset + last offset + 2 × sum of intermediate offsets] = 5010 m^2 = 0.5010 hectares. Using Simpson's rule: Area = h/3 × [y0 + y8 + 4(y1 + y3 + y5 + y7) + 2(y2 + y4 + y6)] = 5260 m^2 = 0.5260 hectares. Therefore, the correct answer is: C." **SimPO raw excerpt.** "This problem asks us to calculate the area between a straight baseline AB and a meandering stream. Since the offsets are taken at equal intervals, numerical integration methods can be used. The distance interval is 20 m, and the offsets are 23, 40, 42, 30, 32, 60, 10, 14, and 22 m. For the trapezoidal rule, the first and last offsets are taken once, while the intermediate offsets are doubled. This gives an area of 5010 square meters, which is equal to 0.5010 hectares. For Simpson's rule, because there are eight equal intervals, the rule can be applied. The odd-positioned offsets are multiplied by 4, and the even-positioned intermediate offsets are multiplied by 2. Carrying out the calculation gives 5260 square meters, which is 0.5260 hectares. Since option C contains these two values, the answer is C." **Interpretation.** SimPO is a strong baseline and reaches the right answer, but its raw response is much longer than necessary for an option-selection task. This case illustrates concise presentation of a correct final answer.

Figure 6: Qualitative multimodal examples. We show the visual input and raw response excerpts for TFPO and representative baselines. The three cases illustrate answer formatting, visual grounding, and response concision in the multimodal evaluation of Section 4.6.