CoGA-GRPO: Correctness-Gated Reward Shaping for Video Reasoning

Anonymous authors
Paper under double-blind review

Abstract

Video reasoning requires balancing answer quality against generation cost. Uniform length incentives may reward unsuccessful responses, while fixed policy regularization fails to account for variations in observed success across rollout groups. We introduce CoGA-GRPO, a correctness-gated extension of group relative policy optimization. It uses a shared rollout-derived allocation score to adapt KL regularization and applies bounded, group-relative concision shaping only to eligible correct responses. Starting from the same reproduced structured-reasoning checkpoint based on Qwen2.5-Omni-7B, we train Standard GRPO and CoGA-GRPO under matched settings and evaluate both using the same four-frame protocol. Relative to Standard GRPO, the CoGA-GRPO accuracy point estimate increases from 66.60% to 67.53%, while the mean generation length decreases from 452.98 to 449.14 tokens. Component ablations and mechanism stress tests further characterize the observed accuracy–length trade-offs associated with the proposed design. CoGA-GRPO also achieves 59.82% accuracy on Daily-Omni. Code and evaluation artifacts are available in an anonymous repository at `https://anonymous.4open.science/r/review-artifact-2027-81B5`.

1 Introduction

Video question answering often requires identifying relevant evidence from temporally distributed visual observations and interpreting dialogue, actions, and social context. A reasoning-capable model can articulate intermediate hypotheses, but the amount of text it produces is not a direct measure of evidence quality. A long response may repeatedly revisit an incorrect interpretation, whereas a short response may omit a necessary temporal relation. The training objective is therefore not simply to maximize or minimize reasoning length, but to shape generation without compromising answer quality.

Reinforcement learning provides a natural framework for addressing this trade-off. GRPO compares responses sampled from the same prompt and avoids training a separate value model (Shao et al., 2024). Reasoning-oriented systems further extend this paradigm to language and video tasks (Guo et al., 2025; Feng et al., 2025). However, an output-length reward is not inherently aligned with correctness. If difficult but incorrect responses receive positive rewards for increased length, optimization may reinforce unsuccessful reasoning trajectories. Conversely, a uniform concision incentive may penalize responses that require longer generation to arrive at a correct answer. A fixed KL coefficient imposes another uniform constraint: groups with substantially different observed success rates are subject to the same degree of resistance to policy updates.

Our starting point is an operational notion of difficulty. The fraction of correct responses within a rollout group measures how frequently the current policy succeeds on a given prompt. This policy-relative signal is available during training because reference answers are known, and it conditions both the regularization coefficient and eligibility for the group-relative length reward.

CoGA-GRPO combines these mechanisms through a shared allocation design. A bounded group score decreases with rollout correctness, and higher scores weaken the KL penalty for groups on which the sampled policy performs poorly. Length shaping is applied only to easy-enough, correct, and score-confident responses, comparing their generated lengths against the mean length of the corresponding rollout group for the same prompt. Hard-but-correct responses remain protected, while

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.

incorrect responses receive no explicit length bonus. The gate operates on the reward contribution rather than directly constraining the policy update. Figure 1 illustrates this contrast.

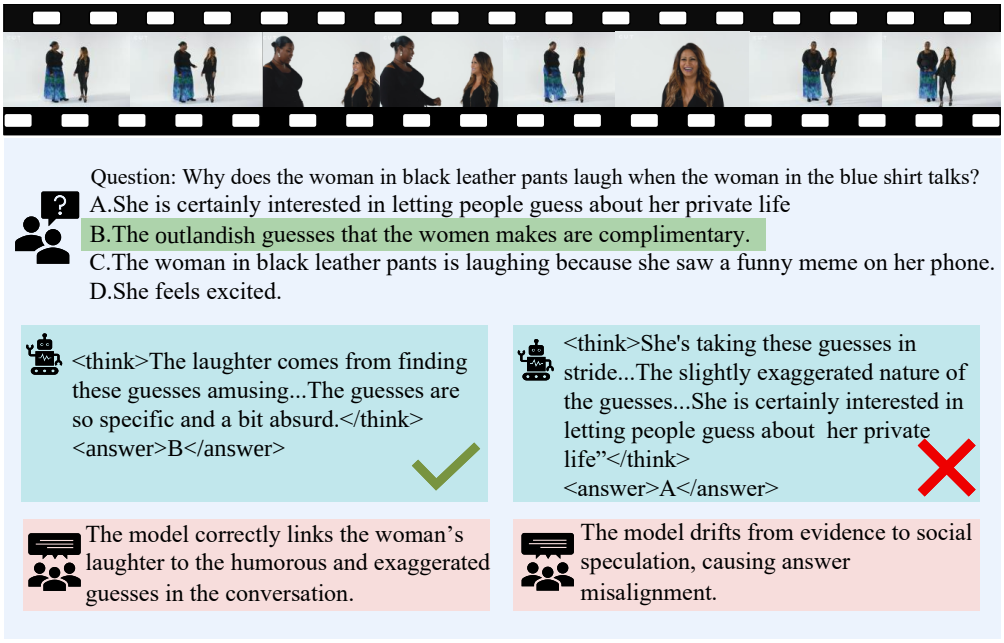


Figure 1: Motivating contrast between evidence-grounded and unsupported reasoning in video question answering.

Difficulty-aware regularization and fast–slow length rewards have been explored in vision-language reasoning (Xiao & Gan, 2025). Building on this line of work, we investigate correctness-gated, group-relative shaping with a shared allocation rule for video question answering, supported by component ablations and mechanism stress tests. Observable input load is evaluated separately because the number of frames or characters reflects processing load rather than reasoning difficulty.

The primary comparison evaluates Standard GRPO and the proposed CoGA-GRPO, both further trained from the same reproduced checkpoint using matched training and evaluation settings. Relative to Standard GRPO, CoGA-GRPO achieves a 0.93% higher accuracy point estimate while generating 0.85% fewer tokens.

The main contributions of this work are as follows:

• We propose CoGA-GRPO, which uses a shared rollout-correctness score to coordinate group-adaptive KL regularization and bounded, correctness-gated length shaping.

• We systematically characterize the allocation mechanism through component ablations, direction-reversal stress tests, and uncertainty analyses at both the question and video-cluster levels.

• Under a fixed evaluation protocol, CoGA-GRPO attains a 67.53% accuracy point estimate and generates 0.85% fewer tokens on average than Standard GRPO; it also attains 59.82% accuracy on the complete Daily-Omni evaluation.

2 Related Work

Multimodal Reasoning. Multimodal learning encompasses representation, alignment, and fusion (Baltrušaitis et al., 2019; Poria et al., 2017), with representative approaches including tensor fusion, modality-specific representations, and cross-modal attention (Zadeh et al., 2017; Hazarika et al., 2020; Tsai et al., 2019).Contrastive pretraining and vision-language interfaces bridge perception and language (Radford et al., 2021; Li et al., 2023; Liu et al., 2023), while video and omni-modal models further extend this capability (Maaz et al., 2024; Cheng et al., 2024; Xu et al., 2025). Video rein-

forcement learning and contextual reasoning further motivate our initialization (Feng et al., 2025; Yang et al., 2025).

Group-Relative Policy Optimization. Preference-based reinforcement learning aligns language-model outputs with human feedback (Ouyang et al., 2022), with related efforts extending alignment to multimodal settings (Sun et al., 2024). Proximal policy optimization (PPO) employs a clipped surrogate objective to constrain large policy updates (Schulman et al., 2017).GRPO derives relative advantages from groups of sampled responses (Shao et al., 2024), while reasoning systems apply outcome-based reinforcement learning (Guo et al., 2025).

Adaptive Computation and Length Control. Chain-of-thought prompting and self-consistency explore explicit reasoning and response aggregation (Wei et al., 2022; Wang et al., 2022), while multimodal chain-of-thought extends this paradigm to multimodal settings (Zhang et al., 2023). Adaptive computation and dynamic networks motivate input-dependent computation budgets (Graves, 2016; Han et al., 2022), distinct from general model-efficiency techniques (Menghani, 2023). Test-time scaling and difficulty-adaptive slow thinking explore adaptive reasoning allocation (Snell et al., 2024; Shen et al., 2025). FAST-GRPO combines difficulty-sensitive regularization with length-based incentives (Xiao & Gan, 2025).

3 METHOD

Figure 2 illustrates the training architecture, while the equations and algorithm define the evaluated controller.

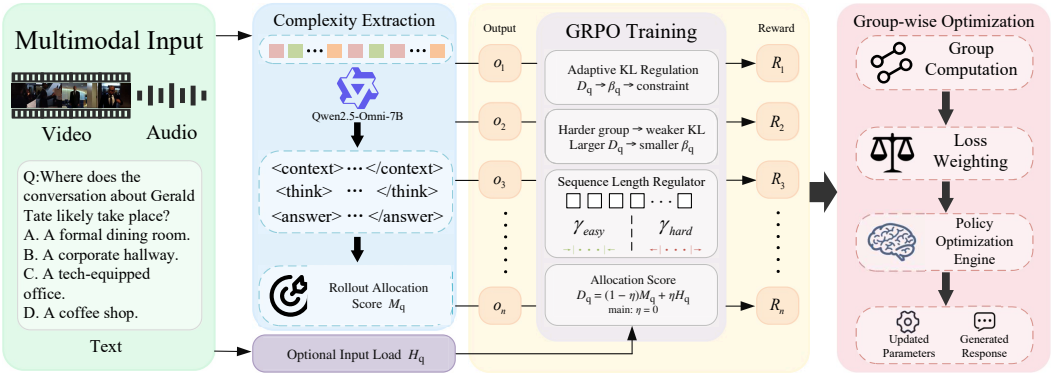


Figure 2: Training architecture. The rollout-correctness score M_q controls the main CoGA-GRPO configuration. The input-load branch H_q is used only in augmented variants, and audio is disabled during the reported training runs. The equations and Algorithm 1 define the evaluated controller.

3.1 ROLLOUT GROUPS AND TASK REWARDS

For a training prompt q with a known reference answer, we sample G responses o_i from the rollout policy π_{old}. Let $a_i \in [0, 1]$ denote the task-specific answer score, $f_i \in [0, 1]$ the structural-format score, and L_i the number of generated tokens selected by the completion mask. The answer scorer is task-dependent: multiple-choice answers use exact matching, whereas set-valued emotion outputs may receive partial credit. The repository also supports additional scoring backends. The allocation mechanism operates on the resulting bounded score.

We define the correctness gate $c_i = \mathbf{1}[a_i \geq \tau_a]$, with $\tau_a = 0.5$. Before length shaping, the scalar reward is

$$r_i^{\text{base}} = w_a a_i + c_i w_f f_i. \tag{1}$$

The reported launcher uses $w_a = w_f = 1$. The format term rewards the ordered `<context>`, `<think>`, and `<answer>` blocks and is disabled below the answer threshold. This gate controls

the format bonus rather than the sign of the relative advantage: partial answer scores and group centering can still determine that sign.

3.2 Policy-Relative Allocation and Adaptive Regularization

Let $p_q = G^{-1} \sum_i c_i$ be the observed correct fraction. Define rollout hardness $S_q = 1 - p_q$ and normalized population disagreement U_q:

$$U_q = 4 \operatorname{Var}_i(c_i) = 4p_q(1 - p_q), \qquad M_q = \lambda S_q + (1 - \lambda)U_q. \tag{2}$$

For binary indicators, disagreement is determined by p_q. The main configuration uses $\lambda = 0.8$, giving the simpler expression

$$M_q = 0.8(1 - p_q^2), \qquad M_q \in [0, 0.8]. \tag{3}$$

Thus, the main controller is a monotonic function of observed success. All-incorrect groups receive $M_q = 0.8$, whereas all-correct groups receive $M_q = 0$. With six rollouts, the score takes seven possible values before incorporating an input signal, so sampling noise and reward errors directly affect the allocation.

For the optional input-load augmentation, let $H_q \in [0, 1]$ combine bounded image, audio, video, and text processing descriptors. The augmented allocation score is

$$D_q = (1 - \eta)M_q + \eta H_q. \tag{4}$$

The main method uses $\eta = 0$, yielding $D_q = M_q$, whereas the augmented variants use $\eta = 0.25$. Their modality weights are 0.25, 0.05, 0.55, and 0.15 for image, audio, video, and text, respectively.

Appendix A describes these descriptors. We refer to H_q as input load rather than semantic complexity: a longer clip may be easy, whereas a brief interaction may be ambiguous. Historical success moving averages are logged as diagnostics and do not constitute an additional controller in the reported main method.

The shared allocation score also controls group-dependent regularization. Specifically, the main direction assigns smaller KL coefficients to groups with larger allocation scores:

$$\beta_q = \beta_{\min} + (\beta_{\max} - \beta_{\min})(1 - D_q). \tag{5}$$

We set $\beta_{\min} = 10^{-4}$ and $\beta_{\max} = 0.04$. For the main method, Equation 3 yields an attainable range of $0.00808, 0.04$. A smaller coefficient imposes weaker regularization relative to the reference policy, but does not guarantee a larger policy update or successful exploration. The reverse-direction stress test replaces $1 - D_q$ with D_q.

3.3 Correctness-Gated, Group-Relative Length Shaping

At optimization step t, the implementation first appends all allocation scores from the current rollout batch to a FIFO window W_t and retains only the most recent 32 values. It then computes a batch-level threshold $T_t = Q_{0.85}(W_t)$ using PyTorch's default linear quantile interpolation. A group is considered hard if $D_q > T_t$. Groups with $D_q = T_t$ remain eligible as easy. The fraction of hard groups is therefore not necessarily 15%, particularly when the scores are discrete or tied. Let $\bar{L}_q = \max(1, G^{-1} \sum_i L_i)$ and $g_i = \mathbf{1}[|a_i - \tau_a| > m]$, where $m = 0.05$. The length term is

$$\ell_i = w_{\text{len}} \tanh\left(\alpha \, g_i \, c_i \, \mathbf{1}[D_q \le T_t] \left(1 - \frac{L_i}{\bar{L}_q} \right) \right), \quad w_{\text{len}} = 0.25, \quad \alpha = 2. \tag{6}$$

Eligible correct responses shorter than the group mean receive positive length shaping, whereas longer responses receive negative shaping. Hard-but-correct and incorrect responses receive no length term. The bounded shaping term cannot increase without limit as response length decreases. The score-margin gate filters ambiguous partial-answer scores rather than estimating calibrated confidence.

The final reward is $r_i = r_i^{\text{base}} + \ell_i$, and the reported configuration uses the mean-centered advantage

$$A_i = r_i - \frac{1}{G} \sum_j r_j. \tag{7}$$

We do not normalize by the group reward standard deviation. A fully incorrect binary group has zero outcome advantage when all base rewards are zero and receives no length term. Reducing its KL coefficient alone therefore cannot create a positive task-learning signal.

Algorithm 1 CoGA-GRPO for one rollout batch

Require: Prompt batch $\mathcal{B}$; score window W; $G = 6$

for each prompt $q \in \mathcal{B}$ **do**
  Sample G responses and compute a_i, f_i, L_i, c_i
  Compute r_i^{base} using Equation 1
  Compute D_q using Equations 2–4
end for
Append $\{D_q : q \in \mathcal{B}\}$ to W and retain the latest 32 values
Set $T_t = Q_{0.85}(W)$
for each prompt $q \in \mathcal{B}$ **do**
  Compute β_q using Equation 5
  Compute ℓ_i using Equation 6
  Set $r_i = r_i^{\text{base}} + \ell_i$ and compute A_i using Equation 7
end for
Update θ using the batch loss in Equation 8

3.4 Implementation-Aligned Training Objective

For token t in response i, define the probability ratio $\rho_{it} = \pi_\theta(o_{it} \mid q, o_{i,<t}) / \pi_{\text{old}}(o_{it} \mid q, o_{i,<t})$. Let $z_{it} = \log \pi_{\text{ref}}(o_{it} \mid \cdot) - \log \pi_\theta(o_{it} \mid \cdot)$ and $K_{it} = e^{z_{it}} - z_{it} - 1$ denote the nonnegative sampled KL-penalty estimator. Using the completion mask b_{it}, the implementation minimizes

$$\mathcal{L}(\theta) = \frac{\sum_{i,t} b_{it} \left[-\min(\rho_{it} A_i, \text{clip}(\rho_{it}, 1-\epsilon, 1+\epsilon) A_i) + \beta_q K_{it} \right]}{\sum_{i,t} b_{it}}, \qquad \epsilon = 0.2. \tag{8}$$

The sum extends over all rollout groups in the optimization batch, using the corresponding β_q for each group. This yields a generated-token-weighted loss, in which longer responses contribute more token-level terms. It is therefore not an equally weighted average of per-response mean losses. We perform one update iteration per rollout batch. The reference policy is the frozen initialization policy, implemented by disabling the trainable adapters of the base model. Algorithm 1 summarizes the rollout-batch procedure.

4 Experiments and Results

4.1 Implementation Details

We use the Thinker component of Qwen2.5-Omni-7B (Xu et al., 2025). The public implementation follows a two-stage pipeline, consisting of multimodal supervised fine-tuning followed by controlled GRPO. The reproduced structured-reasoning checkpoint based on Qwen2.5-Omni-7B (Yang et al., 2025) serves as the initialization for the reported comparison. According to the saved data configuration, initialization uses the final 50% of the rewritten EMER and Social-IQ 2.0 annotations and the final 30% of the rewritten Video-R1 annotations, while further training uses the first 50% of EMER and Social-IQ 2.0.

All seven further-training variants use one epoch, 1,444 optimization steps, a learning rate of 10^{-6}, six responses per prompt, a batch size of six, and a 1,024-token completion cap. Low-rank adaptation (LoRA) uses rank 8 and a scaling factor of 16 for the query, key, value, output, gate, up, and down projections. The vision modules are frozen. The runs use bfloat16 precision, gradient checkpointing, SDPA attention, and 48-GB GPUs. During training, we sample at most six video frames at 1 FPS, with pixel bounds of 75,264 and 100,352. Additional settings are provided in Appendix A.

The reproduced structured-reasoning checkpoint serves as the common initialization for all further-training variants. Standard GRPO and CoGA-GRPO use the same training budget, base answer and format rewards, mean-centered group advantages, and frozen reference policy. Standard GRPO uses a fixed KL coefficient of $\beta = 0.04$ and no length-shaping term, whereas CoGA-GRPO uses the adaptive coefficient in Equation 5 together with the correctness-gated length term in Equation 6. The main CoGA-GRPO experiment uses the unaugmented configuration with $\eta = 0$.

Benchmarks and Decoding. IntentBench comprises 633 videos and 2,689 questions spanning intention, emotion, deception, and related human-centered reasoning tasks (Yang et al., 2025). It serves as the primary controlled benchmark. All compared local models use the same four-frame, 1-FPS protocol, pixel bounds of 50,176 and 75,264, greedy decoding, a 1,024-new-token cap, and a shared `<context>` assistant prefill. All 2,689 questions are included in the formal comparison. Binary correctness is the primary outcome. The archived rule-based answer score, which may include partial credit, is reported separately as the mean evaluation reward. Daily-Omni comprises 684 clips and 1,197 multiple-choice questions targeting audio-visual alignment and reasoning (Zhou et al., 2025). We evaluate all 1,197 questions using the official benchmark data. The evaluation uses the same four-frame visual sampling protocol as the controlled IntentBench runs.

Retrospective Statistical Analysis. For IntentBench, we align paired records by unique question ID and verify that their labels and video identities match. We compute paired differences in binary correctness, answer reward, and generated token count. A paired question-level bootstrap resamples the 2,689 questions, while a cluster-level bootstrap resamples the 633 videos and jointly includes all questions associated with each sampled video. Both use 30,000 bootstrap draws, and percentile-based 95% confidence intervals. The cluster-level bootstrap preserves the question-weighted estimand through ratio-of-sums aggregation.

4.2 Quantitative Results

The primary comparison evaluates Standard GRPO against CoGA-GRPO under matched training and evaluation protocols. Standard GRPO answers 1,791 questions correctly (66.60%), whereas CoGA-GRPO answers 1,816 correctly (67.53%). The mean answer reward increases from 0.6782 to 0.6880. The mean generation length decreases by 3.84 tokens, from 452.98 to 449.14, corresponding to a relative reduction of 0.85%.

As shown in Table 1, relative to Standard GRPO, the paired accuracy change comprises 107 errors corrected by CoGA-GRPO and 82 Standard-GRPO-correct answers that become incorrect. 1,709 questions are answered correctly by both models, while 791 are answered incorrectly by both.

Table 1: Paired correctness transitions on all 2,689 IntentBench questions.

	CoGA-GRPO correct	CoGA-GRPO incorrect
Standard GRPO correct	1,709	82
Standard GRPO incorrect	107	791

Category-Wise Behavior and Additional Benchmarks. Table 2 follows the benchmark comparison format and includes the main method. The Standard GRPO and CoGA-GRPO rows are evaluated locally under the same training and evaluation protocol, whereas all other comparison rows report published results. Relative to Standard GRPO in the controlled local comparison, CoGA-GRPO gains 10 correct answers in How, six in What, five in Other, two in Why, and one each in When, Who/Which, and Emotion, while losing one correct answer in Deception. The small When and Who/Which categories contain only 14 and 25 questions, respectively, so a single changed answer shifts their accuracy by 7.14% and 4.00%.

Table 2: Category-wise accuracy (%) on the official IntentBench benchmark.

Methods	LMM	Why	How	What	When	Who/Which	Other	Emotion	Deception	Avg
Proprietary MLLMs										
GPT-4o (Hurst et al., 2024)	–	61.46	55.69	60.00	35.71	76.00	63.31	60.99	59.00	59.98
Gemini-2.5-Pro (think) (Comanici et al., 2025)	–	68.57	67.41	65.12	57.14	64.00	70.03	68.23	60.00	67.15
Open-Source Video-Audio MLLMs										
MiniCPM-o (Yang et al., 2025)	8B	57.87	53.48	57.14	57.14	68.00	61.14	23.85	49.50	54.51
VITA-1.5 (Fu et al., 2026)	7B	53.15	49.36	51.66	**71.42**	64.00	61.14	53.20	59.50	54.17
Qwen2.5-Omni (Xu et al., 2025)	7B	62.60	63.44	63.53	57.14	76.00	69.03	**59.74**	**63.50**	64.20
Standard GRPO (matched baseline)	7B	66.19	64.56	66.88	57.14	80.00	73.37	54.89	**63.50**	66.60
CoGA-GRPO (Ours)	7B	**66.48**	**66.14**	**68.12**	64.29	**84.00**	**74.36**	55.64	63.00	**67.53**

Table 3: Category-wise accuracy (%) on the official Daily-Omni benchmark.

Methods	AV Event Alignment	Comparative Understanding	Context Understanding	Event Sequence	Inference	Reasoning	30s Subset	60s Subset	Avg
Proprietary MLLMs									
Gemini 2.0 Flash (Zhou et al., 2025)	62.18	73.28	63.73	63.72	76.62	75.43	62.29	56.57	67.84
Gemini 2.0 Flash Lite (Zhou et al., 2025)	55.04	64.89	58.03	54.25	74.03	72.00	60.57	53.01	61.32
Open-Source Video-Audio MLLMs									
Unified-IO-2 (Lu et al., 2024)	25.63	31.30	26.42	25.82	35.06	29.71	26.74	30.00	28.24
VideoLLaMA2 (Cheng et al., 2024)	35.71	35.88	35.75	31.70	40.91	34.29	38.02	31.82	35.17
Qwen2.5-Omni (Xu et al., 2025)	44.12	51.15	38.86	40.52	57.79	61.71	42.35	38.36	47.45
MiniCPM-o (Zhou et al., 2025)	40.33	61.06	49.22	48.36	68.83	61.14	54.09	52.00	53.13
HumanOmniV2 (Yang et al., 2025)	46.63	67.93	51.81	51.63	72.72	74.28	**63.06**	53.09	58.47
CoGA-GRPO (Ours)	**46.64**	**70.99**	**55.44**	**51.96**	**73.38**	**76.00**	61.67	**57.64**	**59.82**

Daily-Omni results are reported in Table 3. CoGA-GRPO achieves 59.82% overall accuracy, including 70.99% on Comparative Understanding and 76.00% on Reasoning. The accuracies on the 30-second and 60-second subsets are 61.67% and 57.64%, respectively. These subsets overlap with the task categories rather than forming additional categories to be averaged. The overall score follows the benchmark aggregation protocol rather than the arithmetic mean of the displayed columns. The main result exceeds the displayed values reported by published open-source systems.

4.3 Qualitative Results

Figure 3 illustrates how the model integrates temporally distributed cues when answering an affective video question. After one speaker jokingly refers to the venue as a "McMansion," the other speaker smiles and nods, responds with appreciative language, and remains visibly relaxed. The audience also cheers and applauds. Taken together, these cues support the conclusion that both speakers are amused and contradict alternatives suggesting confusion, anger, or exasperation.

The generated response makes this evidence chain explicit by identifying the relevant facial reaction, linking the appreciative reply to the preceding remark, and using the audience response as corroborating context before selecting option A. The example therefore illustrates the intended qualitative behavior of aggregating evidence across multiple moments rather than relying on a single isolated frame.

4.4 Ablation Studies

Ablation Configurations. Table 4 reports the component ablations and mechanism stress tests. Dynamic KL only applies Equation 5 without length shaping, whereas Safe length only applies Equation 6 with a fixed $\beta = 0.04$. CoGA-GRPO combines both components with $\eta = 0$. The input-load augmentation enables both controllers with $\eta = 0.25$. The unsafe and reverse-direction stress tests modify this augmented configuration by, respectively, adding a longer-incorrect term for hard groups and reversing the KL mapping.

Table 4: Component ablations and mechanism stress tests on IntentBench under the matched four-frame evaluation protocol. All further-trained variants share the same training setup. Accuracy differences are reported relative to Standard GRPO in percentage points, while length differences are reported in generated tokens.

Method	Correct / 2,689	Accuracy (%)	Δ acc. (%)	Reward	Tokens
Standard GRPO	1,791	66.60	–	0.6782	452.98
Dynamic KL only	1,811	67.35	+0.74	0.6853	454.40
Safe length only	1,804	67.09	+0.48	0.6830	452.30
Input-load augmentation	1,810	67.31	+0.71	0.6846	**447.82**
Unsafe incorrect-length incentive	1,774	65.97	-0.63	0.6722	454.00
Reverse KL direction	1,815	67.50	+0.90	0.6868	453.41
CoGA-GRPO (Ours)	**1,816**	**67.53**	**+0.93**	**0.6880**	449.14

Effects of Individual Components. Dynamic KL alone adds 20 correct answers but raises mean generation by 1.42 tokens. Safe length alone adds 13 correct answers and reduces mean generation



Question:How do the two men feel when the man in red white and blue call the place a McMansion?
A.The two men are amused by the man's comment.
B.The two men exchange a confused look.
C.The man in the glasses is infuriated by the quip uttered by the man in red white and blue.
D.The two men are exasperated by the man's comment.

<context>When the man in red, white, and blue refers to the place as a 'McMansion,' the man in glasses responds with a smile and a nod. The audience cheers and applauds in response. The man in red, white, and blue continues speaking with a relaxed expression. In the conversation, the man in red, white, and blue mentions cleaning up the place and turning it into a 'McMansion,' and adds that they spent a lot of money on it. The man in glasses responds with 'thanks so much' and 'it's nice,' and the conversation continues in a relaxed and pleasant tone.</context>

<think>Okay, I'm looking at this interaction. The man in red, white, and blue just said something about the place being a 'McMansion.' Now, I'm focusing on the man in glasses. **He's smiling and nodding.** Let me listen to what he says right after. Ah, **he says, 'thanks so much' and 'it's nice.' His tone sounds genuinely appreciative and pleased.** The audience, by the way, is cheering and applauding, which makes me think his reaction is positive too. So, if I consider other possibilities, like him being confused or angry, **his smile and the cheerful audience don't fit.** He's not exasperated either; his voice is calm, not agitated. **So, yeah, he seems genuinely amused and appreciative.**</think>
<answer>A</answer>

Figure 3: Qualitative example of affective video reasoning. The response associates the speaker's smile and nod, appreciative reply, and the audience's positive reaction with amusement, leading to the correct selection of option A. Bold text highlights the evidence emphasized in the generated reasoning.

by 0.69 tokens. Combining them gives 25 additional correct answers and a 3.84-token reduction. Input-load augmentation yields the shortest mean response length, at 447.82 tokens, but answers six fewer questions correctly than the main configuration. Its accuracy is 67.31%, with a mean reward of 0.6846.

Table 5: Video-cluster sensitivity analysis for the component configurations, reporting mean paired changes and 95% percentile intervals from 30,000 bootstrap draws that jointly resample all questions within each of the 633 videos.

Comparison with Standard GRPO	Δ accuracy (%)	Δ reward (×100)	Δ tokens
Dynamic KL only (E1)	+0.74 [-0.25, +1.76]	+0.70 [-0.26, +1.68]	+1.42 [-2.29, +5.22]
Safe length only (E2)	+0.48 [-0.49, +1.47]	+0.48 [-0.45, +1.42]	-0.69 [-4.36, +3.03]
CoGA-GRPO (E3)	+0.93 [-0.08, +1.98]	+0.98 [-0.0004, +1.99]	-3.84 [-7.59, -0.13]
Input-load augmentation (E4)	+0.71 [-0.30, +1.74]	+0.63 [-0.34, +1.63]	-5.16 [-8.93, -1.38]

Mechanism Stress Tests. The unsafe stress test achieves 65.97% accuracy with a mean generation length of 454.00 tokens. Relative to input-load augmentation, it answers 36 fewer questions correctly and generates 6.18 more tokens on average. This unsafe reward configuration exhibits a worse observed accuracy–length trade-off than the input-load augmentation configuration. The

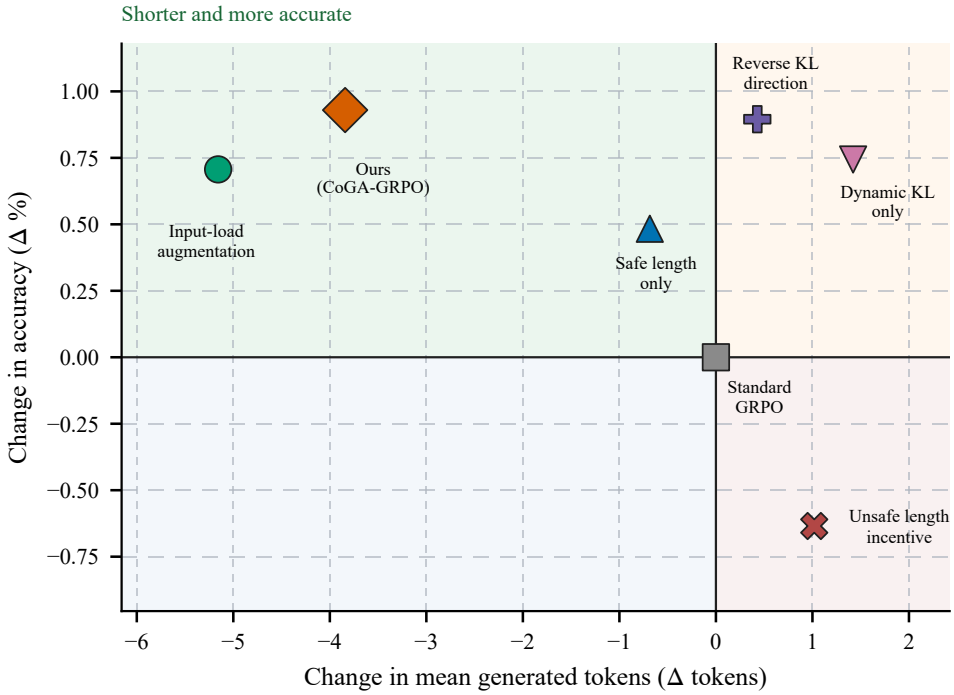


Figure 4: Accuracy–length operating points on IntentBench relative to the matched Standard GRPO baseline. The upper-left quadrant indicates higher accuracy with fewer generated tokens.

reverse-KL stress test achieves a 67.50% accuracy point estimate with a mean generation length of 453.41 tokens. Relative to the input-load augmentation configuration from which this stress test is derived, reversing the KL mapping increases the accuracy estimate by 0.19% while increasing the mean generation length by 5.59 tokens. The direction reversal therefore shifts the observed accuracy–length operating point.

Figure 4 places all component configurations in the joint accuracy–length plane. CoGA-GRPO lies in the upper-left quadrant, with a 0.93-percentage-point higher accuracy point estimate and a 3.84-token lower mean generation length than Standard GRPO. Input-load augmentation moves farther left, achieving the largest token reduction but a smaller accuracy gain. Dynamic KL only and the reverse-KL stress test lie to the right of the Standard GRPO baseline, indicating that their accuracy gains are accompanied by longer outputs. In contrast, the unsafe length incentive is the only configuration that produces longer outputs while also reducing accuracy. For CoGA-GRPO, the question-level bootstrap intervals are $[-0.07\%, 1.93\%]$ for accuracy and $-7.72, -0.04$ tokens for length, while the video-cluster bootstrap intervals are $[-0.08\%, 1.98\%]$ and $-7.59, -0.13$ tokens, respectively.

5 Conclusion

We introduced CoGA-GRPO, a group-relative optimization method that uses rollout correctness to coordinate adaptive KL regularization with correctness-gated length shaping. The shared allocation score weakens regularization for groups with lower observed correctness, while the length term encourages concise generation only for eligible correct responses. On IntentBench, CoGA-GRPO achieves 67.53% accuracy with a mean generation length of 449.14 tokens, compared with 66.60% accuracy and 452.98 tokens for Standard GRPO under matched training and evaluation settings. Component ablations and mechanism stress tests characterize the observed accuracy–length trade-offs associated with dynamic regularization, safe length shaping, and the direction of KL allocation. The model also achieves 59.82% accuracy on the complete Daily-Omni evaluation. These results provide an evaluated design for exploring accuracy–length operating points in video reasoning.

Ethics Statement

This study uses existing research benchmarks and introduces no new human-subject data. Inferences about intentions, emotions, or deception may be culturally biased and should not be treated as reliable assessments of an individual's internal state. The models have not been validated for employment, policing, clinical decision-making, or other consequential applications. Redistribution of media and annotations must comply with the original providers' licenses and privacy restrictions.

Reproducibility Statement

The main text specifies the controller, training settings, and controlled evaluation protocol. The supplementary materials and anonymous repository provide the implementation code, configuration files, and evaluation scripts. Reproducing the reported experiments additionally requires the initialization checkpoint, licensed media, rewritten annotations, and the corresponding computational resources.

Statement on the Use of Large Language Models

Large language models were used for manuscript language editing, code editing, and assistance with statistical-analysis scripts. They were not used to formulate research questions or hypotheses, design the methodology or experiments, develop theoretical models or mathematical claims, generate or clean experimental data, perform qualitative analysis, or interpret the experimental results. The authors reviewed and tested all AI-assisted code, independently verified the numerical results, citations, and technical claims, and take full responsibility for the manuscript and all associated artifacts.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

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 arXiv:2507.06261*, 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 arXiv:2503.21776*, 2025.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Advances in Neural Information Processing Systems*, 38:75300–75320, 2026.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

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

Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7436–7456, 2022.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 1122–1131, 2020.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Gaurav Menghani. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, 55(12):1–37, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

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

Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. Dast: Difficulty-adaptive slow-thinking for large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2322–2331, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wenyi Xiao and Leilei Gan. Fast-slow thinking grpo for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 1103–1114, 2017.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

A Additional Implementation Notes

Settings already specified in the method and experimental setup are not repeated here. Training rollouts use temperature 0.9, top-p 1.0, and top-k 50. These sampling settings are shared across further-training variants.

Input-Load Descriptors. The augmented controller combines normalized descriptors of image pixels, video duration/frame processing, audio duration when provided, and text characters. Available media metadata are clipped to the configured reference ranges. Missing modalities contribute zero. The weights are specified in Section 3.2. These descriptors quantify observable processing load rather than annotated reasoning difficulty. The main method does not use this augmentation, and video audio remains disabled during the reported GRPO further-training runs.

Checkpoint and Reward Semantics. Evaluation reward refers to the stored answer score rather than binary correctness or the full training reward. Paired analyses use 2,689 unique question IDs with identical reference labels.

B Additional Paired Summary

Figure 5 examines how generated length changes across the four paired correctness transitions between Standard GRPO and CoGA-GRPO. Among the 2,689 questions, both methods answer 1,709 correctly and 791 incorrectly, while CoGA-GRPO corrects 107 Standard-GRPO errors and loses 82 previously correct answers. The figure reports the proportions of shorter, unchanged, and longer CoGA-GRPO responses within each transition group, providing a descriptive view of how length changes are distributed across retained, corrected, lost, and persistently incorrect predictions.

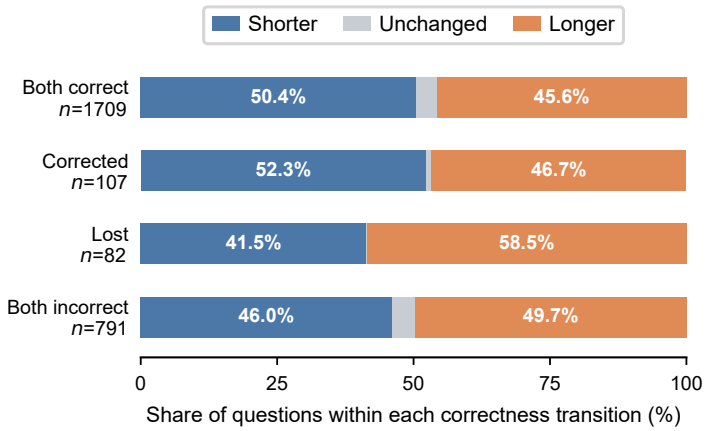


Figure 5: Direction of per-question generated-token changes from Standard GRPO to CoGA-GRPO on IntentBench, grouped by paired correctness transition. Each horizontal bar partitions the corresponding group into shorter, unchanged, and longer responses. The labels indicate group sizes, and the four groups together account for all 2,689 questions.

CoGA-GRPO has a median generation length of 477 tokens, with 90th and 95th percentiles of 624 and 665 tokens, respectively. None of the archived outputs in this main comparison reaches the 1,024-token cap. The source-data file separately preserves correctness, answer score, and generation length for each method and paired question, enabling alternative summaries.

For cluster resampling, each draw samples 633 video IDs with replacement. All associated questions are retained, including those from repeatedly sampled clusters. The estimate for each draw is computed as the sum of paired question-level differences divided by the total number of questions in that draw. This preserves the question-weighted benchmark estimand while allowing questions from the same video to covary. The reported intervals are defined by the 2.5th and 97.5th percentiles.

C Evaluation and Claim Boundaries

Table 6 summarizes the interpretations supported by the current evidence and the claims that the experiments do not establish.

Table 6: What the current evidence can and cannot establish.

Evidence	Supported interpretation	Not established
Paired primary outputs	Observed accuracy–token changes on the same questions	selection-adjusted superiority
Component variants	Descriptive differences under shared further-training settings	Allocation benefit independent of extra training
Mechanism stress tests	An unsafe reward configuration has a worse observed trade-off	Causal harm of length on individual answers
Official transfer scores	Performance of the main model on additional benchmarks	Matched gain over published models
Generated-token counts	Decoder-output demand under fixed decoding	End-to-end speed, energy, or reasoning faithfulness

D Original Statistical Figures

Figure 6 complements the video-cluster sensitivity analysis in Table 5 by reporting question-level paired bootstrap estimates for all 2,689 IntentBench questions. Relative to Standard GRPO, it summarizes the point-estimate changes and 95% percentile intervals for binary accuracy, mean answer reward, and generated-token count across the four component configurations. The accuracy intervals include zero for all configurations.

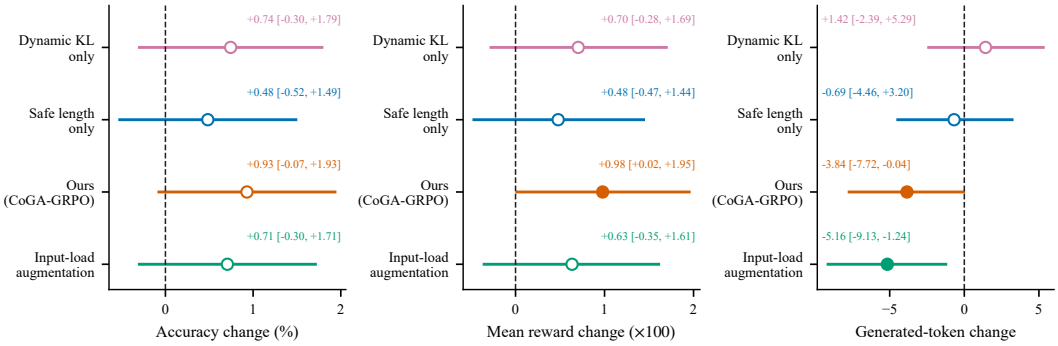


Figure 6: Original paired-effect figure for all 2,689 IntentBench questions. Points indicate changes in accuracy, answer reward, and generated token count. Intervals are obtained using paired question-level resampling.