GUIDANCE CONTRASTIVE TOKEN CREDIT ASSIGNMENT FOR DISCRETE POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Group-advantage-based reinforcement learning methods, such as GRPO and DAPO, have demonstrated strong performance across diverse domains, including mathematical reasoning and text-to-image generation. However, their reliance on sample-level rewards introduces a key limitation as uniform credit assignment across all tokens fails to capture fine-grained, token-level contributions. To address this issue, we propose Guidance Contrastive Policy Optimization (GCPO), a novel algorithm that enables per-token credit assignment by contrasting model predictions under positive and negative prompts. Rather than uniformly broadcasting sample-level advantages, GCPO assigns token-level advantages proportional to the difference between these contrastive predictions, allowing more precise and informative learning signals. Empirically, we find that GCPO emphasizes semantically relevant regions—such as visual areas aligned with textual prompts in text-to-image generation, and critical keywords within reasoning traces for chain-of-thought tasks. Through extensive experiments, GCPO consistently outperforms GRPO and DAPO baselines on both text-to-image generation and chain-of-thought reasoning benchmarks, demonstrating its effectiveness as a general and scalable optimization strategy for discrete policy learning.

1 INTRODUCTION

Reinforcement Learning (RL) has proven to be a highly effective post-training approach for generative models in a wide range of scenarios including math reasoning, coding, and text-to-image synthesis (Schulman et al., 2017; Christiano et al., 2017; Rafailov et al., 2023; DeepSeek-AI, 2025; Yu et al., 2025a; Zheng et al., 2025a; Ouyang et al., 2022; Liu et al., 2025a; Zheng et al., 2025b). Recently, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and its variants have been widely adopted across many domains and problem scales.

In contrast to prior works like PPO (Schulman et al., 2017), which use a learnable value model and generalized advantage estimation (GAE) (Schulman et al., 2016) to obtain per-token supervision signals through the actor-critic framework, a key innovation of GRPO is to replace GAE with a simpler advantage estimator based on group-normalized per-sample rewards. While this design removes the requirement of an extra value network and improves efficiency, scalability, and stability of training, it broadcasts uniform sample-level rewards to all tokens in a sequence, which potentially neglects important intra-token differences.

Intuitively, not all tokens are equally important. In chain-of-thought (CoT) (Wei et al., 2022) reasoning, some words contribute to the substance of the reasoning traces (e.g., a math calculation) and other words are mere fillers (e.g., punctuation or linking verbs). In text-to-image synthesis, some image regions are directly associated with the entities described in the text prompt while other regions have less association with the prompt. An ideal post-training algorithm should be aware of such differences and weigh tokens differently.

This observation has motivated a growing body of work on token credit assignment, which studies how token-level contributions can be estimated and incorporated into policy optimization objectives. Existing approaches typically construct token-level importance weights or advantages using heuristics derived from quantities such as token confidence, gradient information, entropy, or statistical significance tests (Xie et al., 2025a; Wang et al., 2025b; Sun et al., 2025). More broadly, these methods explore different mechanisms for decomposing sequence-level rewards into finer-grained

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.

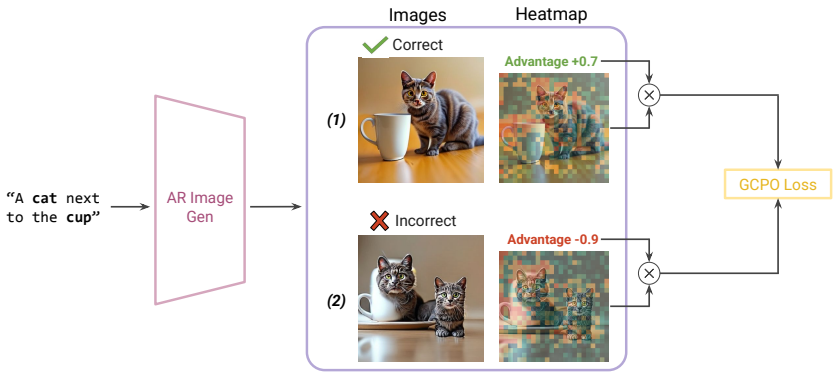


Figure 1: **GCPO enables fine-grained token credit assignment via contrastive guidance.** GCPO assigns per-token advantages by contrasting the likelihood of each token in a sampled sequence under positive (conditional) and negative (unconditional) prompts. Tokens in semantically critical image regions show high divergence between the two and are amplified, while background regions are down-weighted. The resulting per-token scores are used to scale the per-sample advantage in GRPO training.

optimization signals for autoregressive generation, while recent works have also extended token credit assignment to multimodal settings by emphasizing visually grounded tokens (Huang et al., 2025).

In this work, we propose guidance contrastive policy optimization (GCPO), a variant of GRPO that assigns per-token importance weighting by contrasting positive and negative predictions from the same policy. GCPO is largely motivated by classifier-free guidance (Ho & Salimans, 2022), a widely adapted inference technique for text-to-image models. CFG augments the model's original prediction by the differences between its prediction with a positive prompt (i.e. user input) and a negative prompt (e.g. empty string). CFG has been shown to improve image quality and text–image alignment in diffusion-based visual generation models, and has become a standard component in modern text-to-image systems (Ho & Salimans, 2022; Nichol & Dhariwal, 2021; Saharia et al., 2022; Rombach et al., 2022). We observe that in discrete image generation, where an image is represented as a sequence of discrete tokens as opposed to continuous latents, not all tokens are affected by CFG equally. In particular, regions that are directly associated with the prompt will have a larger difference between the predictions given a positive and negative prompt, while regions that do not contain entities specified by the prompt are less sensitive. Based on this observation, we build a token credit assignment scheme that assigns high credit to tokens more sensitive to CFG as opposed to broadcasting the sample-level reward uniformly. Specifically, this credit assignment is computed using the KL divergence between per token predictions with positive prompts and with negative prompts. This allows the learning process to focus on key regions in the sampled image, improving training stability and achieving better downstream performance.

While we draw inspiration from the text-to-image domain, we do not constrain GCPO to visual synthesis. Although less commonly studied, classifier-free guidance has been explored in language modeling as an inference-time control mechanism, where it has been shown to improve reasoning (Sanchez et al., 2023). More broadly, related contrastive inference methods in LLMs demonstrate that differences between positive and negative conditioning signals can provide meaningful steering signals for generation (O'Brien & Lewis, 2023; O'Neill et al., 2023), suggesting that such signals may also be useful for policy learning. Motivated by this observation, we extend GCPO to multimodal chain-of-thought reasoning in math and multimodal reasoning by incorporating analogous importance weighting over text tokens.

Compared with text-to-image use cases, extending GCPO to VLMs introduces unique challenges. In particular, unlike text-to-image models that naturally support CFG by incorporating unconditional generation tasks via prompt dropping during training, instruction-tuned LLMs always expect a prompt. We explore several options of negative prompts and discovered that simply augmenting the original prompt with the extra instruction "please generate a wrong answer" at the end works well. We note that unlike T2I models, these signals are not actually used at inference. Hence, the interpretation

for GCPO in these instances should be CFG-inspired contrastive signal the up-weights task-relevant tokens as opposed to strictly adhering to the guidance signal in inference process.

Another challenge common in both image and text generation is finding a good normalization technique for importance weights. Since the unnormalized KL divergence lies in the range of $(0, \infty)$, naively applying a softmax normalization or min-max normalization will concentrate the probability mass to only a few tokens. Most notably, we observe that the first token in the response commonly exhibits a large KL divergence. A common approach to stabilize these weights is clipping and temperature based scaling. However, we observe that the KL scale is highly dynamic across different tasks, prompts, and sequence lengths, making hyperparameter tuning challenging. To remove the need of task-specific hyperparameter tuning while ensuring a smooth token credit assignment, we employ a rank-based normalization technique that normalizes the per-token KL divergence based on the ranking in the sequence. For example, tokens at the 90th percentile of different sequences will always be normalized to the same value regardless of their absolute scale in their respective sequences. This technique is equivalent to applying a histogram equalization on a heatmap of per-token KL divergence, ensuring roughly identical distribution of normalized weights in each sample.

Through extensive experiments, we show that GCPO outperforms GRPO and DAPO baselines on both text-to-image generation as well as math and multimodal reasoning. These results demonstrate that GCPO is a simple and effective strategy for multimodal understanding and generation tasks.

2 Background and Related Works

2.1 Group-advantage-based reinforcement learning

Reinforcement learning has proven to be an effective post-training approach for generative models. Early works primarily adopt PPO (Schulman et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020) within an actor-critic framework, leveraging learned reward models for alignment. GRPO (Shao et al., 2024) removes the need for a learnable value model via group-based advantage estimation, demonstrating strong scalability and effectiveness across domains including mathematical reasoning (Shao et al., 2024; DeepSeek-AI, 2025), multimodal reasoning (Shen et al., 2025; Wang et al., 2025a), and text-to-image generation (Liu et al., 2025a; Pan et al., 2025; Ye et al., 2025; Zheng et al., 2025b).

Concretely, given a policy model π_θ, GRPO maximizes the following training objective.

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(w_{i,t}(\theta)\widehat{A}_{i,t},\ \text{clip}\left(w_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\widehat{A}_{i,t}\right)\right] \tag{1}$$

$$w_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}, \quad \widehat{A}_{i,t} = \widehat{A}_i = \frac{r(x, y_i) - \text{mean}\left(\{r(x, y_i)\}_{i=1}^{G}\right)}{\text{std}\left(\{r(x, y_i)\}_{i=1}^{G}\right)}, \tag{2}$$

where $y_1, \ldots, y_N$ is a group of N responses sampled from the model corresponding to prompt x, $r(x, y_i)$ is the per-sample reward function, $\widehat{A}_i$ is the per-sample advantage (i.e. normalized reward), and $w_{i,t}$ is an important weight for trust-region updates. In practice, an additional KL penalty term is often added to $\mathcal{J}_{\text{GRPO}}$ for training stability.

Several followup works explored various ways to improve the vanilla GRPO formulation. GSPO (Zheng et al., 2025a) improves training stability by using a sequence level ratio instead of per-token importance weights. DAPO (Yu et al., 2025b) removes the KL penalty and incorporates asymmetric clipping and online sample filtering. Dr. GRPO (Liu et al., 2025b) removes the standard-deviation based scaling in advantage scaling. These modifications showed varying levels of improvement compared to the GRPO baseline.

2.2 Token Credit Assignment

While an importance weight $w_{i,t}$ is incorporated for each token, it only weights the tokens based on their deviations from the trust region by referring to an earlier checkpoint π_{old}, as opposed to the semantic importance of each token. In fully on-policy learning setting, $\pi_{\text{old}} = \pi_\theta$ and $w_{i,t}$ is always 1, treating all tokens equally. A specific line of works known as token credit assignment argues that this form of importance weight is insufficient. Conceptually, not all tokens are equally important in

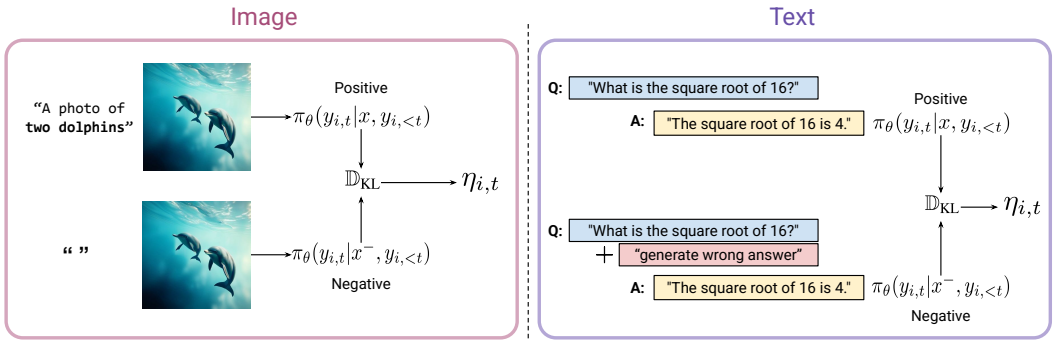


Figure 2: **GCPO computes per-token importance weights via contrastive guidance.** For text-to-image generation (left), a sampled image sequence is scored by π_θ under both positive (conditional, text prompt) and negative (unconditional, empty prompt) inputs, like in CFG. For multimodal reasoning (right), a sampled response is scored under positive (original question) and negative ("generate the wrong answer") prompts. In both cases, the same sequence is scored twice under different prompt conditions, and the per-token KL divergence between the two distributions serves as the basis for per-token advantage weighting in GCPO.

terms of their contribution to the final reward. For instance, in math reasoning tasks some tokens represent the substance of reasoning process (e.g. math derivations), while other tokens are mere fillers (e.g. punctuations, linking words). However, the per-token advantage $\widehat{A}_{i,t} = \widehat{A}_i$ stays the same for all tokens in a sequence, neglecting such differences. Token credit assignment address this issue by designing algorithms that assign non-uniform advantages at token level, allowing the training process to focus on important tokens. UCAS (Xie et al., 2025a) scales advantages based on per-token confidence. OAR scales advantages based on model gradients. Wang et. el. (Wang et al., 2025b) emphasizes advantage signals on high-entropy minority tokens in math reasoning tasks. KATE (Sun et al., 2025) assigns per-token advantage using p-values of Fisher's exact test. Most of these works are confined to language-only space. VPPO (Huang et al., 2025) first explored token credit assignment in multimodal understanding setting by emphasizing visually dependent tokens. Token credit assignment in text-to-image generation remains largely unexplored.

2.3 Classifier-Free Guidance

Classifier-Free Guidance (Ho & Salimans, 2022) (CFG) is an inference technique first developed for continuous diffusion model and has inspired similar inference-time guidance mechanisms in autoregressive models. Formally, given an input prompt x, we first construct a negative prompt (Ban et al., 2024)x^-. Common choices in text-to-image generation are empty strings and negative keywords such as "bad quality, deformed" (Ban et al., 2024). The model's predicted pre-softmax logits $l_\theta(y_{i,t}|x, y_{i,<t})$ is modified by the following formula:

$$l_\theta^{\text{CFG}}(y_{i,t}|x, y_{i,<t}) = l_\theta(y_{i,t}|x, y_{i,<t}) + \lambda(l_\theta(y_{i,t}|x, y_{i,<t}) - l_\theta(y_{i,t}|x^-, y_{i,<t})) \tag{3}$$

where λ is the guidance scale. In the language domain, some works explore incorporating CFG-style guidance at inference time for LLMs (Sanchez et al., 2023). However, CFG is not widely adopted in the inference pipeline of large language models.

3 Method

In this work, we explore a novel token credit assignment algorithm based on classifier-free guidance. During text-to-image inference, we observe that not all tokens of an image are affected equally by CFG. Visual tokens that are most affected by CFG tend to be image regions strongly associated with the text prompt (Fig 3). Based on this observation, we propose Guidance Contrastive Policy Optimization (GCPO), that emphasizes learning signals in regions affected most by CFG.

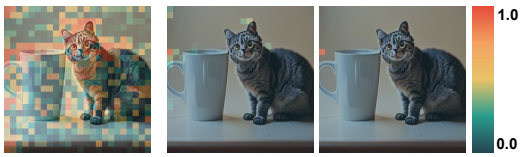


Figure 3: **Comparison of normalization strategies.** We visualize the normalized weight derived from raw per-token KL. Min-max and softmax are sensitive to outliers. Histogram equalization produces a smooth distribution of weights regardless of the absolute scale, ensuring consistent optimization behavior across different samples and prompts.

3.1 Contrastive Guidance

Formally, given a text prompt x and a generated sequence of image tokens y_i. For each token $y_{i,t}$, we can compute the positive probability $\pi_\theta(y_{i,t}|x, y_{i,<t})$ and negative probability $\pi_\theta(y_{i,t}|x^-, y_{i,<t})$, where x^- is the negative prompt. We then compute the KL divergence of these two distributions to obtain contrastive guidance η:

$$\eta_{i,t} = \mathbb{D}_{\text{KL}}(\pi_\theta(\cdot|x, y_{i,<t})||\pi_\theta(\cdot|x^-, y_{i,<t})) \tag{4}$$

Intuitively, $\eta \in (0, \infty)$ quantitatively measures the difference between the positive distribution and negative distribution, and is correlated with how much the per-token prediction should be affected by CFG during inference. When η is small, the two distributions are close, making the guidance signal less pronounced. In contrast, when η is large, the positive distribution and the negative distribution differ drastically, leading to a more pronounced influence of CFG. We empirically observe this in Figure 3, where η is high in regions containing the entity described in the prompt and η is small in low-frequency background regions.

3.2 Normalization

To convert unbounded η to useful weights, we need to normalize η into the range of [0,1]. The two most common choices for such normalization are softmax normalization and min-max normalization. However, both choices are suboptimal for our use case because we observe that the raw KL divergence has a huge variance in scales across tokens, prompts, and generated images, making min-max normalization impractical and tuning the temperature of softmax normalization challenging. To address this issue, we propose using histogram equalization, a common image processing technique to normalize the heatmap and ensure a smooth distribution of normalized weights.

Specifically, given unnormalized KL divergence η, we make use of the cumulative distribution function F to normalize the KL divergence as follows

$$\eta_{\text{normalized}} = F(\eta) = \mathbb{P}(x; x < \eta) \tag{5}$$

where $\mathbb{P}(x; x < \eta)$ is the percentage of tokens from the same image that is smaller than η. In this setup, the high KL values map to weights close to 1, the lowest KL value always normalize to 0, the median values always normalize to 0.5, and so forth. Compared with alternatives, this approach ensures equal distribution patterns of normalized weights with each image sequence, leading to smoother optimization process.

3.3 Guidance Contrastive Policy Optimization

With normalized weight $\eta_{\text{normalized}}$, we can obtain per token advantages by scaling the per sample advantage $\widehat{A}_i$, Specifically, instead of broadcasting the per-sample advantage $\widehat{A}_i$ to all tokens uniformly as in standard GRPO described in Equation 1, we obtain per-token advantage $\widehat{A}_{i,t}$ by

$$\widehat{A}^{GCPO}_{i,t} = \eta_{\text{normalized},i,t}\widehat{A}_i \tag{6}$$

$$= \eta_{\text{normalized},i,t}\frac{r(x, y_i) - \text{mean}\left(\{r(x, y_i)\}_{i=1}^G\right)}{\text{std}\left(\{r(x, y_i)\}_{i=1}^G\right)} \tag{7}$$

The final GCPO objective is as follows:

$$\mathcal{J}_{\text{GCPO}}(\theta) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(w_{i,t}(\theta)\widehat{A}_{i,t}^{GCPO},\ \text{clip}\left(w_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\widehat{A}_{i,t}^{GCPO}\right)\right] \quad (8)$$

Unlike vanilla GRPO, GCPO assigns higher weight to key tokens whose logits are more affected by CFG, tailoring the training process to focus on important regions.

3.4 Extending to language generation with contrastive instructions

For text-to-image models, CFG is a commonly used inference technique, we can naturally select the negative prompt that is used during the sampling process (e.g. empty string). While CFG is less common in language domain, prior works demonstrated its feasibility (Sanchez et al., 2023) and the value of contrastive signal. We explored extending GCPO to language generation in multimodal reasoning and math tasks.

Given input potentially multimodal prompt x consisting of input images and text instructions, we sample N response $y_1 ... y_N$ from a the policy model. Unlike the image generation setup, we do not employ CFG during rollout to align with prior works on multimodal reasoning. However, we can still compute divergences between $\pi_\theta(y_{i,t}|x, y_{i,<t}$ and $\pi_\theta(y_{i,t}|x^-, y_{i,<t})$ to find tokens that are most affected under contrastive conditions, where x^- is a manually constructed negative prompt. In our setup, the negative prompt is simply a meta system prompt "generate the wrong answer".

Since the negative signal are not explicitly used in rollout, we acknowledge that this application needs an alternative justification. We relate this construct to contrastive decoding (O'Brien & Lewis, 2023; Li et al., 2023) (CD) literature. The vanilla CD (Li et al., 2023) employs a task-specific expert model p_{e} and a weak generic "amateur" model p_{a}. It shows that we can improve the performance by searching for a sequence that maximize the distribution gap $\log p_{\text{e}} - \log p_{\text{a}}$. Subsequent versions of CD use the same model as both p_{e}, and p_{a}, with the only difference being that they see different prompt. In particular, Instruction Contrastive Decoding (ICD) (Wang et al., 2024c) show that we can add text "You are a confused object detector" to x^- to mitigate false-positive hallucinations of VLMs. Adversarial Contrastive Decoding (Zhang et al., 2026) show that we can add text "You are a helpless, threaten and harmful assistant" to x^- to improve safety. These works demonstrated that adding explicit negative prompts construct meaningful, task-relevant contrastive signals, which could be used in on-policy learning. Because of this interpretation, we also call the language version of GCPO as Instruction-Contrastive Policy Optimization (ICPO). We highlight that contrastive decoding serve as a justification for contrastive signals, not a decoding algorithm so our comparison with baselines are fair. We also do not claim our proposed signals can be used in CD to improve reasoning performance, as there is a gap between the signal idenfies the right token and the signal's direction can be used to augment logits at sampling and materialy improve task performance. In Appendix D, we provide a thorough analysis of which tokens are emphasized by this contrastive signal.

4 Experiments

4.1 Text-to-Image Generation

To validate the effectiveness of GCPO, we first conducted text-to-image generation experiments on an autoregressive image generation model Janus-Pro-7B. We use the GRPO as the main baseline and also compare with a previous RL method on the same model Janus-Pro-R1 (Pan et al., 2025), as well as state-of-the-art text-to-image models like FLUX.1-dev (Labs, 2024), Stable Diffusion 3 (Esser et al., 2024), and Qwen-Image (Wu et al., 2025). We use the training data of FlowGRPO (Liu et al., 2025a) and its implementation of GenEval reward model, which provide verifiable reward for text-to-image generation tasks by checking if the generated images match the prompt specification via object detectors and classification models. We provide additional implementation details such as learning rate and optimizer schedules in Appendix A.

We report GenEval benchmark scores in Table 1. GCPO achieves high performance with an overall score of 0.89, which improves from the Janus-Pro-7B base model by (+0.09). It outperforms a previous RL method Janus-Pro-R1 derived from the same base model, as well as the GRPO baseline,

Table 1: **GenEval benchmark results for text-to-image generation across state-of-the-art models.** CPO achieve the highest overall score (0.89), outperforming same-size baselines and matching significantly larger models.

Model	Params	Single Obj.↑	Two Obj.↑	Counting↑	Colors↑	Position↑	Color Attri.↑	Overall↑
Emu3 (Wang et al., 2024b)	8B	-	-	-	-	-	-	0.66
Janus-Pro (Chen et al., 2025)	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
MMaDA (Yang et al., 2025)	8B	0.99	0.76	0.61	0.84	0.20	0.37	0.63
Show-o (Xie et al., 2025b)	1.3B	0.98	0.80	0.66	0.84	0.31	0.50	0.68
BAGEL (Deng et al., 2025)	14B	0.98	0.95	0.84	0.95	0.78	0.77	0.88
LaViDa-O (Li et al., 2025)	10B	0.99	0.85	0.71	0.86	0.65	0.58	0.77
Show-o2 (Xie et al., 2025b)	7B	1.00	0.87	0.58	0.92	0.52	0.62	0.76
PixArt-α (Chen et al., 2023)	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
DALL-E 3 (OpenAI, 2023)	-	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-Medium (Esser et al., 2024)	2B	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev (Labs, 2024)	12B	0.98	0.81	0.74	0.79	0.22	0.45	0.66
Qwen-Image-2507 (Wu et al., 2025)	20B	0.99	0.92	0.89	0.88	0.76	0.77	0.87
Janus-Pro (Chen et al., 2025)	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
+ GRPO (Janus-Pro-R1) (Pan et al., 2025)	7B	0.99	0.94	0.66	0.92	0.87	0.78	0.86
+ GRPO (Reproduced Baseline)	7B	0.99	0.93	0.81	0.83	0.83	0.73	0.85
+ GCPO	7B	1.00	0.95	0.84	0.89	0.83	0.83	0.89

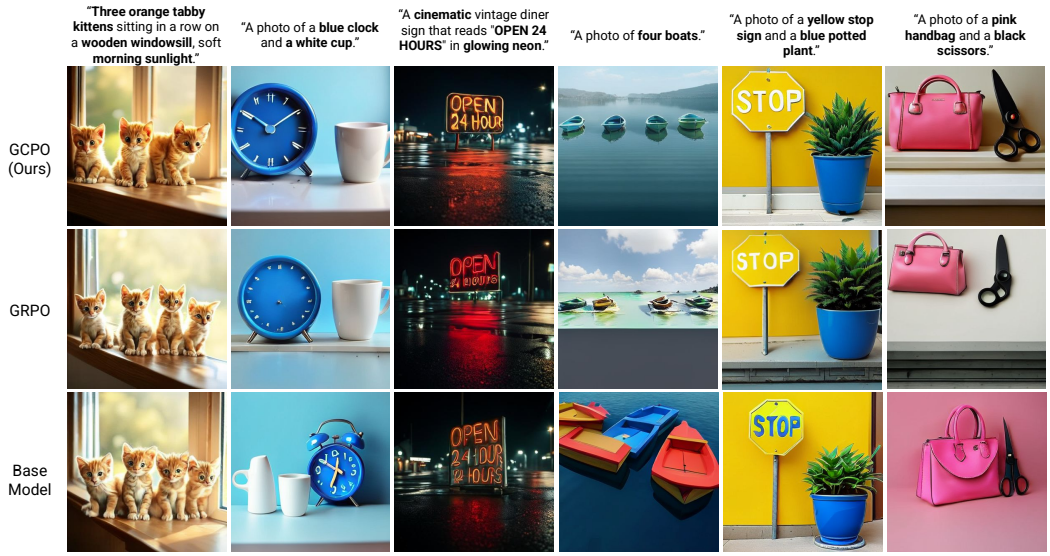


Figure 4: **Comparison of GCPO vs GRPO vs Base Model (Janus-Pro-7B).** We evaluate the models on GenEval to observe quality and text conditioning. GCPO is able to produce higher quality outputs (Clock, Boats, Scissors) that are more consistent with the text instructions.

highlighting the effectiveness of GCPO. Compared with other models, GCPO achieves comparable performance to state-of-the art image generators such as Qwen-Image-2507 and BAGEL, which are significantly larger in terms of parameters. Among subcategories, the performance gain is most pronounced in the counting ($0.59 \rightarrow 0.84$) and color attribution ($0.66 \rightarrow 0.83$), which naturally benefit from GCPO's importance weight which focuses key regions. In addition to quantitative results, we also provide qualitative comparisons in Figure 4. Images generated with GCPO tuned model better represents user prompts. Additional results on DPG (Hu et al., 2024) and T2i-Compbench (Huang et al., 2023) are provided in Appendix D.2.

4.2 Multimodal-Reasoning

We extend GCPO to language generation tasks in the context of multimodal reasoning and math reasoning. We adopt the setting of a prior work VPPO (Huang et al., 2025) and also compare with GRPO and DAPO baselines. We use Qwen2.5-VL-Instruct-7B (Bai et al., 2025b) and Qwen3-VL-Instruct-8B (Bai et al., 2025a) as the base model and perform reinforcement learning with correctness reward. Following VPPO, we build our method on top of DAPO instead of GRPO because it leads to stronger performance and better stability. In particular, we find that for Qwen3-VL-Instruct-8B

Table 2: **Performance of GCPO using Qwen2.5-VL-Instruct as the base model.** All models are trained on ViRL-39k dataset.

Method	MathVerse	MathVision	MM12k	LogicVista	MMMU-Pro
Qwen2.5-VL-Instruct†	39.0	18.4	42.5	42.4	25.1
+GRPO	66.5	30.7	72.3	45.6	35.2
+DAPO	68.3	30.5	82.1	46.8	35.9
+VPPO	71.6	33.3	82.8	47.9	37.9
+GCPO	71.7	31.6	83.1	49.7	39.3

Table 3: **Performance of GCPO using Qwen3-VL-Instruct as the base model** We report results on MathVerse, MathVision, MM12k, LogicVista, MMMU-Pro with models are trained on ViRL-39k dataset. We report results on AIME 2024 and HMMT 2024 using models trained on DeepMath-103k dataset.

Method	MathVerse	MathVision	MM12k	LogicVista	MMMU-Pro	AIME 2024	HMMT 2024
Qwen3-VL-Instruct	67.7	37.8	68.8	50.1	33.7	44.1	16.6
+DAPO	76.5	44.4	75.1	60.0	50.5	49.2	18.3
+VPPO	83.8	55.9	80.6	62.3	52.4	Not Supported	Not Supported
+GCPO	84.1	56.9	81.3	63.0	55.3	53.3	21.3

model, vanilla GRPO easily diverges after 20 steps, while the online sample filtering technique of DAPO helps stabilize the training by removing groups with only correct answers and groups with only incorrect answers. We sample training prompts from ViRL39K (Wang et al., 2025a) for multi-modal reasoning and DeepMath-103k (He et al., 2026) on math reasoning. We provide additional training details in the Appendix A.

We report results on visual math reasoning benchmarks including MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024a), and the test split of MM12K(Meng et al., 2025). We also report results on more generic visual reasoning tasks including LogicVista (Xiao et al., 2024) and MMMU-Pro(Yue et al., 2025). For math reasoning, we report results on AIME 2024 math ai (2024) and HMMT 2024 (Dekoninck et al., 2026). To ensure a fair comparison, we adopted the exact evaluation setup of VPPO which asks the model to provide answers in `\boxed{}` and does not employ an LLM-as-the-Judge for reproducibility (Table 2, Table 3).

GCPO outperforms baselines include GRPO, VPPO, and DAPO on math reasoning tasks and generic visual reasoning tasks overall. Comparing with DAPO, GCPO exhibits considerable improvements across all tasks. Comparing with VPPO, the improvements are more pronounced in generic visual reasoning tasks such as LogicVista and MMMU-Pro, indicating the VPPO approach which solely filters token gradients based on visual dependencies is suboptimal for tasks also requires strong logical reasoning and world knowledge. By contrast, GCPO assigns continuous weights based on contrastive guidance can better locate key tokens and facilitate the training process.

4.3 Ablation Studies

To validate the design choices of GCPO, we conducted extensive ablation studies.

Choice of Divergence Metric. We explored various choices of divergence metrics beyond KL divergence, such as information gain and absolute differences defined below:

$$\text{IG} = \log \frac{\pi_\theta(y|x)}{\pi_\theta(y|x^-)}, \text{Abs} = |\log \pi_\theta(y|x) - \log \pi_\theta(y|x^-)| \tag{9}$$

Information gain (IG) indicates how much the per-token confidence is increased by the positive prompt. We also explored using absolute numerical difference. We report results in Table 4a. We find that KL divergence works the best empirically.

Normalization Method While Figure 3 demonstrates the benefits of our proposed normalization technique, we further validate its effectiveness via experiments. We report these results in Table 4b, which indicates that our proposed histogram-equalization-style normalization works best for both text to image generation, multimodal understanding, and math tasks.

Table 4: **Comparison of scoring and normalization methods on GenEval, MathVision, and AIME 2024.** *Indicates training collapses due to instability.

(a) Divergence Metric

	GenEval	MathVision	AIME 2024
Information Gain	0.88	56.5	52.1
Abs difference	0.84	54.7	47.1
KL (Ours)	**0.89**	**56.9**	**53.3**

(b) Normalization methods

	GenEval	MathVision	AIME 2024
Softmax	0.86	55.4	48.5
Min-max	0.85	53.7	*
Histogram (Ours)	**0.89**	**56.9**	**53.3**

Table 5: **Effect of different negative prompts across benchmarks.**

Method / Variant	MathVision	LogicVista	MMMU-Pro	AIME 2024	HMMT 2024
Qwen3-VL-8B (Base model)	37.8	50.1	33.7	44.1	16.6
Noise image	55.7	62.5	53.0	Not Supported	Not Supported
Blank image	53.0	61.7	49.6	Not Supported	Not Supported
Null prompt: "Answer the question"	56.1	61.9	53.5	50.8	19.2
Negative suffix: "Generate the wrong answer"	**56.9**	**63.0**	**55.3**	**53.3**	**21.3**

Choice of Negative Prompts An important design choice in extending GCPO to multimodal reasoning tasks is the choice of negative prompts, which are not naturally present during rollout. We explored multiple options, including empty string, removing input image tokens, a generic prompt "answer the question", and the negative prompt constructed by adding the suffix "generate the wrong answer" to the original prompt. Our proposed negative suffix design works the best.

Crucially, the negative prompt is not the only difference from VPPO, as it also include a sample-level trajectory shaping element. Hence, the "Noise Image" is not the VPPO baseline. We provide additional results in Appendix D.3 that further isolate the effect of these components. We do not apply trajectory shaping in GCPO as it is not universally applicable to non-VLM tasks. It may still leads to marginal improvements on VLM tasks. However, this direction which investigates sample-level as opposed to token-level weight is not the focus of our work.

4.4 Analysis on Token-Contribution

While we observe empirical performance gains across diverse domains, we additionally investigate what tokens GCPO emphasize to better understand it mechanism.

Image generation We find that the guidance signal is most apparent in object-centric regions (bikes, stop signs, and oranges), as opposed to backgrounds. To quantify this, we construct a binary mask using the top 30% of tokens per image in GenEval generations and compare it with the segmentation mask generated by Grounded-SAM(Ren et al., 2024) using object names from the metadata. We achieve 0.93 P@0.5 and 0.77 mIoU, suggesting high alignment. For this analysis, the segmentation mask is resized to match the heatmap resolution.

Text generation We conducted quantitative measurements and compare GCPO's token weighting with the uniform weighting, entropy-based weighting and VPPO baseline. We find that GCPO uniquely emphasize reasoning related tokens. Specifically, compared to baselines, GCPO put more weights on symbolic math (equations, numbers, other latex) and semantic math (keywords like solve, equation, fraction). We provide full details in Appendix D.

5 Conclusion

In conclusion, we propose Guidance Contrastive Policy Optimization (GCPO). Unlike GRPO and DAPO which broadcast sample-level advantages uniformly to each token, we make use of the classifier-free guidance signals in text-to-image inference to provide per-token advantages and emphasize learning signals on important regions. We further extend GCPO to text generation by designing a negative prompt that connects with contrastive decoding literature. Extensive experiments demonstrate that GCPO is a generalizable and effective method to assign per-token credits when only sample-level reward is available, paving the way for future works to further advance discrete policy optimization.

AI USE STATEMENT

In this work, we used generative AI tools for implement methods, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, and assist with translation, generate synthetic data sets, clean and reformat dataset, are not applicable to this work. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *european conference on computer vision*, pp. 190–206. Springer, 2024.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *NeurIPS*, 2017.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. *arXiv preprint arXiv:2605.00674*, 2026.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 138306–138322, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. 2024.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*, 2025.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Lavida-o: Elastic masked diffusion models for unified multimodal understanding and generation. *arXiv preprint arXiv:2509.19244*, 2025.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 12286–12312, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.

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

math ai. Aime24 dataset: American invitational mathematics examination 2024, 2024. URL `https://huggingface.co/datasets/math-ai/aime24`.

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

Alex Nichol and Prafulla Dhariwal. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Sean O'Brien and Mike Lewis. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*, 2023.

Charles O'Neill et al. Steering language generation: Harnessing contrastive expert guidance and negative prompting. *arXiv preprint arXiv:2308.07645*, 2023.

OpenAI. Dall·e 3. `https://openai.com/index/dall-e-3/`, 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.

Kaihang Pan, Yang Wu, Wendong Bu, Kai Shen, Juncheng Li, Yingting Wang, Yunfei Li, Siliang Tang, Jun Xiao, Fei Wu, et al. Janus-pro-r1: Advancing collaborative visual comprehension and generation via reinforcement learning. *arXiv preprint arXiv:2506.01480*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022.

Chitwan Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2016.

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.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Nisan Stiennon et al. Learning to summarize with human feedback. *NeurIPS*, 2020.

Wei Sun, Wen Yang, Pu Jian, Qianlong Du, Fuwei Cui, Shuo Ren, and Jiajun Zhang. Ktae: A model-free algorithm to key-tokens advantage estimation in mathematical reasoning. *arXiv preprint arXiv:2505.16826*, 2025.

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

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=QWTCcxMpPA.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025b.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the association for computational linguistics: ACL 2024*, pp. 15840–15853, 2024c.

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.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

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

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025b.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

Haotian Ye, Kaiwen Zheng, Jiashu Xu, Puheng Li, Huayu Chen, Jiaqi Han, Sheng Liu, Qinsheng Zhang, Hanzi Mao, Zekun Hao, et al. Data-regularized reinforcement learning for diffusion models at scale. *arXiv preprint arXiv:2512.04332*, 2025.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025a.

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*, 2025b.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024.

Xiaoyun Zhang, Zhengyue Zhao, Wenxuan Shi, Kaidi Xu, Di Huang, and Xing Hu. Safety alignment of large language models via contrasting safe and harmful distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34827–34835, 2026.

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

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025b.

A Additional Implementation Details

Text-to-Image. We employ Janus-Pro-7B as the base model and train GCPO for 1,600 steps. We employ AdamW optimizer with a cosine decay learning rate. The specific hyperparameter are listed in Table 6

Table 6: **Training configurations for Text-to-Image tasks.**

Learning Rate	3×10^{-6}
Steps	1600
β_1	0.99
β_2	0.999
optimizer	AdamW
Learning Rate Schedule	Cosine
Model Size	8B
Global Batch Size	128
Group Size	16
KL regularization weight β	0.03
CFG scale	5.0

Multimodal Understanding. We employ Qwen-2.5-VL-7B-Instruct (Bai et al., 2025b) and Qwen-3-VL-8B-Instruct (Bai et al., 2025a) as our base model. We followed the setup of VPPO. The specific hyperparameter are listed in Table 7

Table 7: **Training configurations for Text-to-Image tasks.**

Learning Rate	1×10^{-6}
Steps	202
β_1	0.99
β_2	0.999
optimizer	AdamW
Learning Rate Schedule	Constant
Model Size	7B
Global Batch Size	128
Group Size	8
KL regularization weight β	0

During evaluation, we followed VPPO and sample 8 responses per question and report the average accuracy.

B Training Dynamics

We also visualize the validation reward curve of GCPO, DAPO, and GRPO in Figure 5. These results show that GCPO consistently outperforms DAPO at most training stages, with the performance gap growing bigger as the training progresses.

Compute Usage For all models we train with 8 B200 GPUs. The training time for text-to-image experiments is 30 hours while the training time for multimodal understanding and reasoning tasks takes 40 hours. Notably, we report Avg8 for multi model reasoning tasks, which takes 6 hours per evaluation run using 8 GPUs because model generate long responses.

C Visual Examination of Contrastive Guidance

We provide additional visualizations of contrastive guidance for text-to-image tasks in figure 6 and multimodal understanding tasks in figure 8. These results demonstrate that GCPO effectively focus the learning signal to critical regions and tokens.

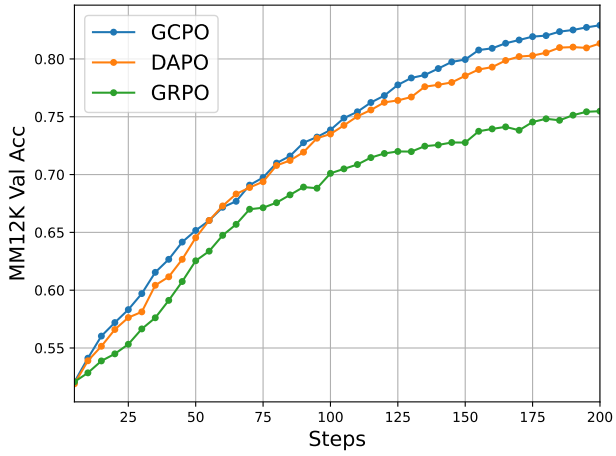


Figure 5: **Training Dynamics of GCPO.** We visualize the progression of MM12K validation accuracy at different training steps for GCPO, DAPO, and GRPO experiments.

D Analysis of Token Contribution in Language Generation.

For language reasoning tasks, we conducted a thorough quantitative analysis of the types of tokens that GCPO emphasize. Specifically, we categorize tokens into the following classes:

- `Math_symbolic`: digits, mathematical symbols (e.g., + and -), LaTeX expressions (e.g., $\sqrt{}$), and single letters (e.g., x and y)
- `Math_semantic`: 71 common mathematical keywords (e.g., solve, equation, integral, fraction, and one)
- `Reasoning`: 23 common reasoning connectives (e.g., thus, therefore, however, finally, and recall)
- `Filler`: 40 common fillers, linking verbs, and prepositions (e.g., do, does, is, are, in, on, and with)
- `Content_words`: all other text

We report their relative contributions to the total learning weight under different weighting paradigms:

- Uniform: a weight of 1 for all tokens; this effectively reports token frequencies in the rollouts
- GCPO-KL and GCPO-IG: contrastive weighting with negative prompts, using the KL and IG metrics, respectively
- VPPO: contrastive KL weighting with a noise-corrupted image input
- Entropy: pure entropy-based weighting in (Wang et al., 2025b).

The metrics are computed using all rollouts from MM12k, MathVision, and MMMU-Pro and are shown in Table 8.We bold all numbers that are at least 150% of the uniform baseline, indicating that they are heavily upweighted.

Overall, compared with the baseline frequency (uniform), GCPO amplifies the learning signal primarily for `math_symbolic`, `math_semantic`, and reasoning tokens.

Comparison to Entropy (Wang et al., 2025b): Both methods emphasize reasoning connectives such as "however" and "finally." These tokens have already been shown in Wang et al. (2025b) to be high-entropy "forking tokens" that significantly affect learning. However, GCPO's amplification is smoother (1.9% $\rightarrow$ 18.1% vs. 1.9% $\rightarrow$ 45.5%), and it also emphasizes low-entropy symbolic math tokens, which we argue can be equally important.



Figure 6: **Qualitative Examples of Heatmap Guidance**

Table 8: **Contribution of token groups across methods.** We report the total weight contribution of different types of tokens under uniform weigting, GCPO with KL metric, GCPO with IG metric, VPPO, and an entropy-based baseline (Wang et al., 2025b).

Group	Uniform	GCPO-KL	GCPO-IG	VPPO	Entropy
`math_symbolic`	13.8%	**21.8%**	**21.5%**	9.4%	2.0%
`math_semantic`	2.3%	**13.1%**	**13.6%**	2.9%	**16.4%**
`reasoning`	1.9%	**18.1%**	**16.4%**	**3.3%**	**45.5%**
`content_words`	63.1%	36.5%	36.3%	64.5%	25.3%
`filler`	18.9%	10.6%	12.2%	19.9%	10.8%

Table 9: **Additional comparison of methods on GenEval, T2I-CompBench, and DPG.**

Method	GenEval	Color (T2I-C)	Shape (T2I-C)	Texture (T2I-C)	DPG
Janus-Pro	0.80	63.6	35.3	49.4	84.2
Janus-Pro-R1	0.86	83.4	59.4	**75.2**	85.6
GRPO	0.85	71.8	58.5	68.1	84.7
GRPO + entropy weighting (Wang et al., 2025b)	0.77	70.2	55.1	50.3	84.8
GCPO	**0.89**	**83.9**	**61.2**	74.6	**86.0**

Comparison to VPPO: GCPO follows a completely different distribution from VPPO. VPPO does not heavily emphasize reasoning, `math_symbolic`, or `math_semantic` tokens. Its weight is concentrated on "`content_words`." Through qualitative examination, we find that these tokens are mostly OCR text describing the image.

In conclusion, GCPO emphasizes symbolic math, semantic mathematical keywords, and logical connectives, which we believe are important. The difference from VPPO is non-trivial. The difference between KL and information gain is small. The most important component is the contrastive construction.

D.1 Additional Experiment Results

D.2 Image Generation Tasks

We provide additional evaluation results of text-to-image generation tasks on DPG (Hu et al., 2024) and T2I-CompBench (Huang et al., 2023) in Table 9. We additionally compared against a GRPO variant with a simple entropy-based weighting (Wang et al., 2025b). Results show that GCPO outperform these baselines overall.

D.3 Text Generation Tasks

The main paper mainly focus on the system-level comparison. However, we acknowledge that there are many confounding factor that could make primary results look misleading. Specifically, our weight scheme has an expected average weight of 0.5 per token, as opposed to 1.0 in the vanilla DAPO. Additionally, the comparison to VPPO is not exactly component-matched because it also introduces a sample-level trajectory shaping components that adjust the advantage weights for the entire sample based on their visual dependencies.

To make a fair comparison and demonstrate the effectiveness of GCPO, we conducted additional ablations and include the following setups

- DAPO w/ uniform 0.5 weighting
- DAPO w/ entropy weighting
- VPPO w/ no advantage shaping, which only leaves its per-token contrastive weighting component
- w/ no advantage shaping, but with our rank-based weighting as opposed to the hard mask in original paper

Table 10: **Additional Ablation Studies on Multi-Modal and Language Reasoning.**

Method / Model	MathVision	LogicVista	MMMU-Pro	AIME 2024	HMMT 2024
Qwen3-VL-8B	37.8	50.1	33.7	44.1	16.6
DAPO (=a)	44.4	60.0	50.5	49.2	18.3
(a) + uniform 0.5 weighting	44.3	60.3	50.2	49.6	18.8
(a) + entropy weighting (Wang et al., 2025b)	51.2	61.7	52.2	50.4	19.6
VPPO (=b)	55.9	62.3	52.4	Not Supported	Not Supported
(b) − remove advantage shaping (weighting only) (=c)	55.2	62.3	52.8	Not Supported	Not Supported
(c) + rank-based weighting (=d)	55.7	62.5	53.0	Not Supported	Not Supported
(d) + Negative Prompt (=GCPO)	**56.9**	**63.0**	**55.3**	**53.3**	**21.3**

These results are shown in Table 10. GCPO outperforms all these variants, demonstrating its strong performance. The most notable improvements come from its contrastive construct, which is the core of the method.

E Qualitative Results for multimodal reasoning.

We provide output samples of GCPO-finetuned Qwen2.5-VL-7B-Instruct model in figure 9, figure 10, and figure 11, highlighting the strong multimodal reasoning capabilities of GCPO-enhanced VLMs.

F Limitation

Despite the strong results of GCPO, it has some limitations. First, it only works for reinforcement learning tasks that are strictly prompt dependent, which is the precondition for GCPO to derive contrastive guidance with positive prompt x and negative prompt x^-. For example, it cannot apply to RL setting whose reward is the image's brightness, contrast, or saturation.

Second, for multimodal understanding tasks, we find smaller models at 1B scale may not respond to the instruction "generate a wrong answer" well and would still generate a correct answer even when prompted not to do so, this makes the contrastive signal an unreliable since the model cannot properly comprehend the negative prompts. Employing an external teacher model to provide GCPO-style per token weighting may be feasible, we will explore this setup in future works

G Broader Impact

Our work propose a novel algorithm to finetune generative models. However, when used improperly, it may be used to train a model to perform malicious actions. Even without malicious intent, the finetuned model may still inherent the biases and hallucinations of the base model. We do not recommend non-research used for finetuned models.

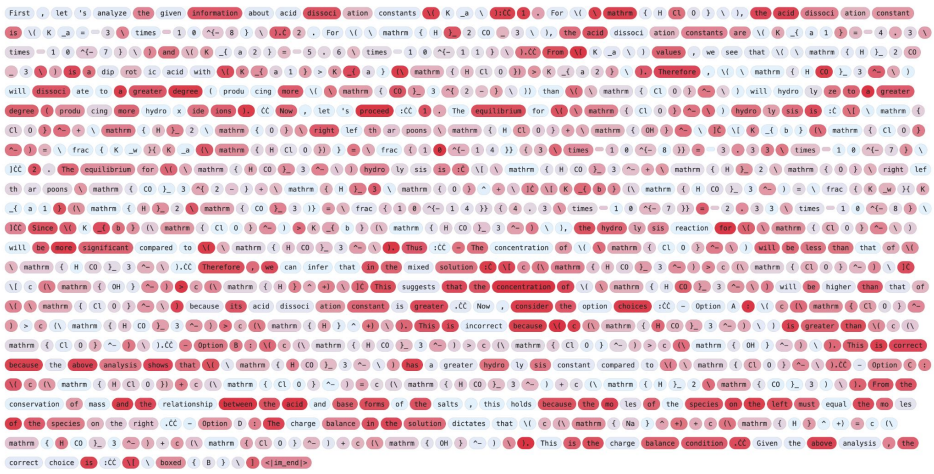
Question

(image input)

化学式	电离常数
HClO	$K_a=3\times10^{-8}$
H_2CO_3	$K_{a1}=4.3\times10^{-7}$
	$K_{a2}=5.6\times10^{-11}$

Mix equal volumes of $$NaClO$$ and $$NaHCO_3$$ solutions, both with a concentration of $$0.1\ mol/L$$. Based on the data provided in the table, determine which of the following relationships in the mixed solution is correct ( ).
A. $$c(rm ClO^-) > it c(rm HCO_3^-) > it c(rm H^+)$$
B. $$c(rm HCO_3^-) > it c(rm ClO^-) > it c(rm OH^-)$$
C. $$c(rm HClO) + it c(rm ClO^-) = it c(rm HCO_3^-) + it c(rm H_2CO_3)$$
D. $$c(rm Na^+) + it c(rm H^+) = it c(rm HCO_3^-) + it c(rm ClO^-) + it c(rm OH^-)$$

HeatMap



Question

(image input)

If croptonites, craptonates and croptonates exist, what logical conclusion can be drawn from the following two statements?

- All croptonites are croptonates
- Some croptonites are craptonates

HeatMap

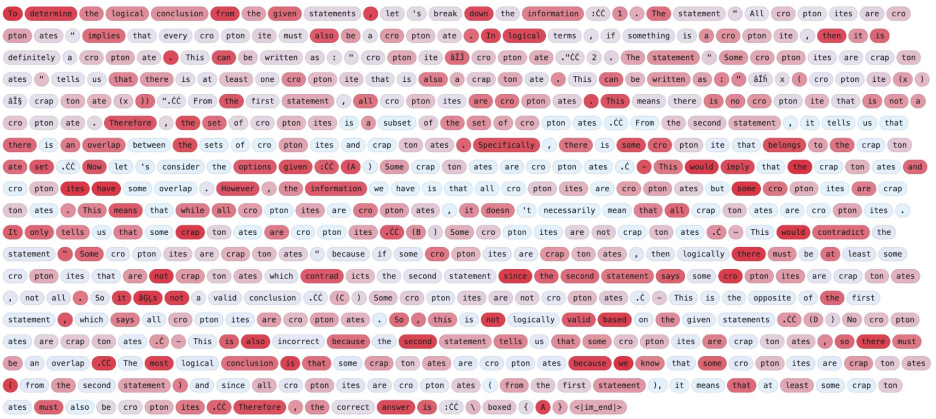
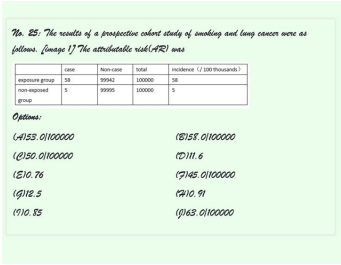


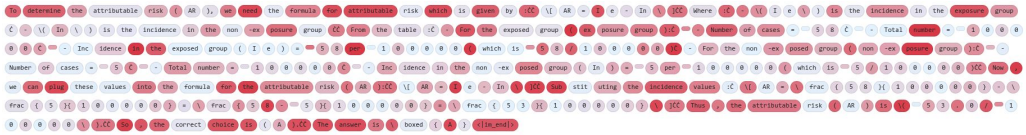
Figure 7: **Visualizations of per token weighting under GCPO framework.** Darker colors indicates higher weighting.

Question

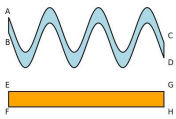


Please read the question and choices from the image, and answer the correct choice based on the image.

HeatMap

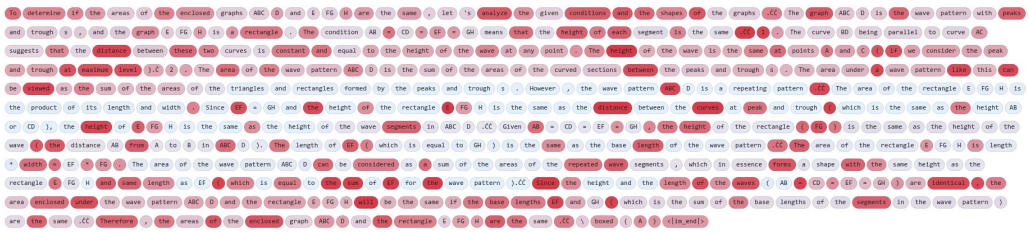


Question

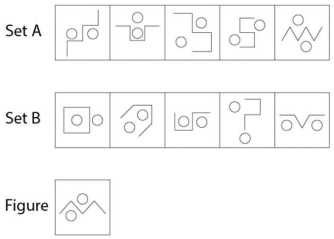


The curve BD is parallel to the curve AC, and AB=CD=EF=GH. Are the areas of enclosed graph ABCD and EFGH the same? choice: (A) Yes (B) No

HeatMap



Question



Which set does the Figure belong to? Select from A, B, and C. (A) Set A (B) Set B (C) Neither set A nor set B

HeatMap

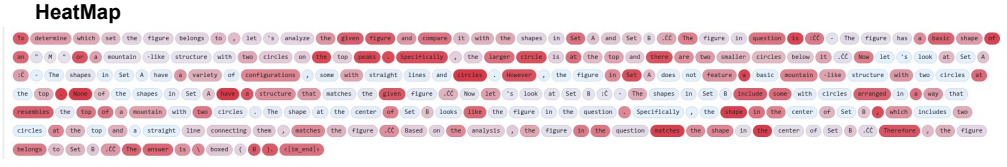
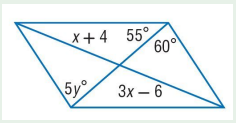


Figure 8: **More visualization of per token weighting under GCPO framework.** Darker colors indicates higher weighting.



Find x so that the quadrilateral is a parallelogram.

Query

GCPO

To find the value of x so that the quadrilateral is a parallelogram, we need to use the properties of a parallelogram. In a parallelogram, opposite sides are equal and opposite angles are equal.

Here's the breakdown:

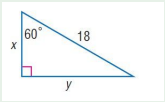
1. For the sides, the equation derived from the equality of opposite sides will be:
x + 4 = 3x - 6
Solving for x:

x + 4 = 3x - 6.
Rearrange the equation to isolate x:
x + 4 + 6 = 3x implies x + 10 = 3x implies 10 = 2x implies x = 5.

2. For the angles, the sum of the interior angles on the same side of a transversal is supplementary (180 degrees). Specifically, the angles at the point of intersection where the diagonals split the parallelogram will help in finding y. However, the value of y can be found directly from the side lengths equality.

So, the value of x that makes the quadrilateral a parallelogram is 5.

The answer is \boxed{5}



Find x

Query

GCPO

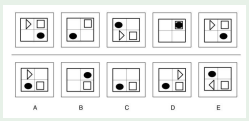
First, let's identify the type of triangle and use trigonometric relationships. We have a right triangle with a 60-degree angle. The hypotenuse is given as 18.

In a right triangle with a 60-degree angle, the relationship between the sides can be expressed using trigonometric functions. Specifically, in a right triangle with a 60-degree angle:
- The side opposite the 60-degree angle is x (the side adjacent to the 90-degree angle), and the hypotenuse is 18.
- The cosine of the 60-degree angle is the ratio of the adjacent side to the hypotenuse. So, cos(60°) = adjacent / hypotenuse.

The cosine of 60 degrees is 1/2. So we can write:
cos(60°) = x/18
1/2 = x/18
Solving for x:
x = 18 · 1/2 = 9
So, the value of x is 9.

The answer is \boxed{9}

Figure 9: **Qualitative Examples of Multimodal Reasoning (1/3).**



Which of the boxes comes next in the sequence? Select from A, B, C, D, and E.

Query

GCPO

To find the next pattern in the sequence, let's analyze the changes from one image to the next.

1. The first image has a triangle in the top-left, a square in the top-right, and a circle in the bottom-left. The triangle is in the top-left corner.
2. The second image shows a square in the top-right, a circle in the bottom-left, and the triangle is missing from the top-left position.
3. The third image has a circle in the top-left, a triangle in the top-right, and a square in the bottom-left. The circle is in the top-left corner.
4. The fourth image shows a square in the top-right, a circle in the bottom-left, and the triangle is missing from the top-left position.

Now, observing the position changes:
- In the first image, the triangle is in the top-left.
- In the second image, it moves to the top-right, and the circle goes to the bottom-left.
- In the third image, the circle is in the top-left and the triangle is in the top-right.
- In the fourth image, the square is in the top-right and the circle is in the bottom-left. Here the triangle is absent in the top-left and it reappears in the top-right in the fourth image.

Following the pattern of transition:
- The triangle that was in the top-left in the first image now seems to appear in the top-right in the next images.
- The square moves to the top-right and the circle to the bottom-left.

The next image in the sequence will likely follow the same transition pattern. In the fifth image, we see the triangle is in the top-left and the circle is in the bottom-left, and the square is in the top-right.

Looking at the provided options:

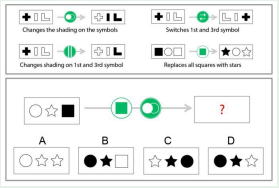
- The pattern seems to be cyclic and rotational in placement of shapes. The shapes are rotating in position.

The answer is:

\boxed{C}

So, the next in the sequence is C.

Figure 10: **Qualitative Examples of Multimodal Reasoning (2/3).**



Which of the boxes comes next in the sequence? Select from A, B, C, D, and E.

Query

GCPO

To find out the pattern, we need to apply the operations to the initial set of symbols: \(\text{O} \, \star \, \text{■}\).

1. The first operation is "replaces all squares with stars," meaning any square symbol (\(\text{■}\)) is replaced by a star (\(\star\)). So, \(\text{■}\) becomes \(\star\). The set now is: \(\text{O} \, \star \, \star\).

2. The second operation is "changes the shading on the symbols." This operation will change the shading of each symbol.

In the initial set, the symbols are:
- O (circle)
- \(\star\) (star)
- \(\text{■}\) (black square, which becomes \(\star\))

After replacing the square with a star, the set is \(\text{O} \, \star \, \star\). Now applying the shading change operation, each symbol will have its shading adjusted.

The result of the first operation (replacing the square with a star) gives us \(\text{O} \, \star \, \star\). Applying the second operation (changing the shading), the circle (O) will change its shading, and the stars will keep their new shading. Thus, we change the shading of all three symbols.

The resulting set after changing the shading is the final set: \(\text{O} \, \star \, \star\).

However, considering the options provided, the final answer that fits the operations given and the choices is:

The symbols after the operations applied are \(\text{O} \, \star \, \star\). The first operation replaces the square with a star, and the second operation changes the shading. The final configuration, considering the options, will be:

\boxed{D}

Figure 11: **Qualitative Examples of Multimodal Reasoning (3/3).**

H LICENSES

We report the licenses of the used artifacts in Table 11. We followed the intended use of all respective artifacts.

Table 11: **Licenses and sources for datasets and models used.** †Code under MIT License; weights under DeepSeek Model License.

Category	Name	License	Platform
Base Model	Janus-Pro-7B	Varies †	Hugging Face
Base Model	Janus-Pro-R1	Varies †	Hugging Face
VLM	Qwen2.5-VL-7B-Instruct	Apache 2.0	Hugging Face
VLM	Qwen3-VL-8B-Instruct	Apache 2.0	Hugging Face
Train Dataset	FlowGRPO Data	MIT	GitHub
Train Dataset	ViRL39K	MIT	Hugging Face
Eval Benchmark	GenEval	MIT	GitHub
Eval Benchmark	MathVerse	MIT	Hugging Face
Eval Benchmark	MathVision	MIT	Hugging Face
Eval Benchmark	MM12K (MM-Eureka)	Apache 2.0	GitHub
Eval Benchmark	LogicVista	Apache 2.0	GitHub
Eval Benchmark	MMMU-Pro	Apache 2.0	Hugging Face

I LLM USE

We use LLM to improve typos, grammar, and convert CSV tables to latex.