Boosting MLLM Reasoning with Text-Debiased Hint-GRPO
–Supplementary Material–

S1 Theoretical Analysis

This work uncovers a phenomenon (termed "text-bias") wherein the MLLM trained with GRPO tends to base their reasoning primarily on text condition while neglecting image condition. To tackle this problem, this work proposes text-bias calibration, which can directly emphasize the image condition **in test-time**. Inspired by CFG (classifier-free guidance) [2] in image generation, text-bias calibration is conducted by first generating token logits from the MLLM under two conditions: with and without image condition. Next, it calibrates the final token logits based on the differences between these two predictions.

Specifically, let q_{img} and q_{text} denote the image condition and text condition, then $\hat{\pi}_\theta(o_t|q_{\text{img}}) = \pi_\theta(o_t|q_{\text{img}}, q_{\text{text}}, o_{<t})$ and $\hat{\pi}_\theta(o_t) = \pi_\theta(o_t|q_{\text{text}}, o_{<t})$ represent the token logit predicted with and without image condition (note that $\hat{\pi}_\theta$ abbreviates the original π_θ). Finally, the calibrated token logit $\hat{\pi}_\theta^{\text{calibrated}}(o_t|q_{\text{img}})$ is calculated as below following CFG (γ is a hyper-parameter controlling the intensity of image condition):

$$\hat{\pi}_\theta^{\text{calibrated}}(o_t|q_{\text{img}}) = \hat{\pi}_\theta(o_t|q_{\text{img}}) + \gamma \cdot (\hat{\pi}_\theta(o_t|q_{\text{img}}) - \hat{\pi}_\theta(o_t)),$$

Theoretical analysis: To address the text-bias problem where image conditions are ignored, we follow Classifier Guidance [1] used in conditional image generation models to enhance the intensity of image condition on the final output. Classifier Guidance controls the intensity of conditional generation by adjusting the weight of the condition predictor with a γ coefficient. In the MLLM case, conditional control is calculated as below:

$$\hat{\pi}_\theta^{\text{calibrated}}(o_t|q_{\text{img}}) = \hat{\pi}_\theta(o_t) + \gamma \cdot (\hat{\pi}_\theta(q_{\text{img}}|o_t) - \hat{\pi}_\theta(q_{\text{img}})).$$

However, unlike image generation with classifier models for image-to-condition mapping, in the MLLM field there are no models for output-to-image mapping, *i.e.*, $\hat{\pi}_\theta(q_{\text{img}}|o_t)$. Therefore, we follow CFG (classifier-free guidance) that utilizes Bayes's theorem to tackle this problem. To this end, we first convert the token prediction logit $\hat{\pi}_\theta$ to token prediction probability $\hat{p}$, through a softmax operation.

$$\hat{p}(o_t|q_{\text{img}}) = \frac{\exp \hat{\pi}_\theta(o_t|q_{\text{img}})}{\sum_{o_t'} \exp \hat{\pi}_\theta(o_t'|q_{\text{img}})} = \frac{\exp \hat{\pi}_\theta(o_t|q_{\text{img}})}{\mathcal{Z}_{o_t|q_{\text{img}}}},$$

where $\mathcal{Z}_{o_t|q_{\text{img}}}$ denotes the normalization term. Next, after applying a log operation on it, we have:

$$\log \hat{p}(o_t|q_{\text{img}}) = \hat{\pi}_\theta(o_t|q_{\text{img}}) - \log \mathcal{Z}_{o_t|q_{\text{img}}}.$$

Other token prediction probability is calculated likewise, *e.g.*, $\log \hat{p}(o_t) = \hat{\pi}_\theta(o_t) - \log \mathcal{Z}_{o_t}$. Next, according to Bayes's theorem, we have:

$$\begin{aligned}
\hat{p}(o_t|q_{\text{img}})\hat{p}(q_{\text{img}}) &= \hat{p}(q_{\text{img}}|o_t)\hat{p}(o_t). \\
\log \hat{p}(o_t|q_{\text{img}}) + \log \hat{p}(q_{\text{img}}) &= \log \hat{p}(q_{\text{img}}|o_t) + \log \hat{p}(o_t). \\
\log \hat{p}(q_{\text{img}}|o_t) - \log \hat{p}(q_{\text{img}}) &= \log \hat{p}(o_t|q_{\text{img}}) - \log \hat{p}(o_t).
\end{aligned}$$

In a real implementation, we can approximate the normalization terms as being equal. Finally, $\hat{\pi}_\theta^{\text{calibrated}}(o_t|q_{\text{img}})$ is calculated as below:

$$
\begin{aligned}
\hat{\pi}_\theta^{\text{calibrated}}(o_t|q_{\text{img}}) &= \hat{\pi}_\theta(o_t) + \gamma \cdot (\hat{\pi}_\theta(q_{\text{img}}|o_t) - \hat{\pi}_\theta(q_{\text{img}})) \\
&= \hat{\pi}_\theta(o_t) + \gamma \cdot (\log \hat{p}(q_{\text{img}}|o_t) - \log \hat{p}(q_{\text{img}})) \\
&\quad + \gamma \cdot \left(\log \mathcal{Z}_{q_{\text{img}}|o_t} - \log \mathcal{Z}_{q_{\text{img}}}\right) \\
&\approx \hat{\pi}_\theta(o_t) + \gamma \cdot (\log \hat{p}(o_t|q_{\text{img}}) - \log \hat{p}(o_t)) \\
&\quad + \gamma \cdot \left(\log \mathcal{Z}_{o_t|q_{\text{img}}} - \log \mathcal{Z}_{o_t}\right) \\
&= \hat{\pi}_\theta(o_t) + \gamma \cdot (\hat{\pi}_\theta(o_t|q_{\text{img}}) - \hat{\pi}_\theta(o_t)).
\end{aligned}
$$

In practice, we find that this form of $\hat{\pi}_\theta^{\text{calibrated}}(o_t|q_{\text{img}})$ is difficult to control with a single parameter γ, as it deviates significantly from the original $\hat{\pi}_\theta(o_t|q_{\text{img}})$. Therefore, we replace $\hat{\pi}_\theta(o_t)$ by $\hat{\pi}_\theta(o_t|q_{\text{img}})$ to mitigate this problem, and the final form of $\hat{\pi}_\theta^{\text{calibrated}}(o_t|q_{\text{img}})$ is calculated as below:

$$
\hat{\pi}_\theta^{\text{calibrated}}(o_t|q_{\text{img}}) = \hat{\pi}_\theta(o_t|q_{\text{img}}) + \gamma \cdot (\hat{\pi}_\theta(o_t|q_{\text{img}}) - \hat{\pi}_\theta(o_t)).
$$

S2 More Experiments

S2.1 Training Time

Method	Qwen2-VL-7B	Qwen2.5-VL-3B
GRPO	10.4	8.9
Hint-GRPO	12.8 (+23.08%)	10.5 (+17.98%)

Table 1. Training time (Hours) comparison between the original GRPO and Hint-GRPO with adaptive hint strategy.

This section compares the training time between the original GRPO and Hint-GRPO with adaptive hint strategy in Table 1 ($M = 2$), indicating that when M is small, Hint-GRPO adds only slightly additional training time over GRPO.

References

[1] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS 2021*, pages 8780–8794, 2021. 1

[2] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1