Influence of KL Divergence

KL divergence is frequently employed in RL to mitigate policy degradation during reward maximization (Ouyang et al. 2022; Xiong et al. 2023). However, recent investigations in both LLMs and MLLMs reveal that the reasoning patterns of DeepSeek-R1 can be effectively replicated without the use of KL divergence (Meng et al. 2025; Hu et al. 2025). Notably, MM-Eureka (Meng et al. 2025) demonstrates that the inclusion of KL divergence not only reduces response length but also leads to diminished accuracy rewards in multimodal tasks. Therefore, it is imperative to examine the impact of KL divergence in CR^3 on compositional reasoning tasks.

In our detailed evaluation, we compared the performance of models with and without KL divergence, specifically using a KL coefficient of $\beta = 0.04$ on the Qwen2.5-VL-3B-Instruct model. As illustrated in Figure 4, the CR^3 model without KL divergence significantly outperforms its counterpart with KL divergence ($\beta = 0.04$), particularly on compositional reasoning benchmarks. This finding suggests that KL divergence is not a critical component of the objective in rule-based RL aimed at optimizing the compositional reasoning capabilities of MLLMs. Furthermore, this observation also indicates that KL divergence may sometimes hinder the generalization ability of RL.

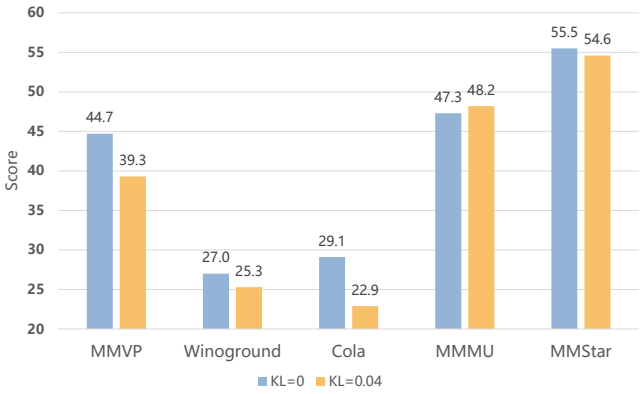

Figure 4: Performance comparison of CR^3 variants (with VS. without KL divergence) implemented on Qwen2.5-VL-3B-Instruct. The reported metrics of Winoground and Cola are group score.