

---

# Co-Reinforcement Learning for Unified Multimodal Understanding and Generation

## Supplementary Material

---

Anonymous Author(s)

Affiliation

Address

email

## 1 A Appendix

### 2 A.1 Training Data

3 **Training Data for Unified Reinforcement Learning.** To support synergistic multimodal modeling  
4 during unified RL, we curate a dataset (*i.e.*, `x2x_rft_22k`) that simultaneously involves text-to-image  
5 generation and multimodal understanding tasks. As illustrated in Figure 1, each sample includes *a*  
6 *real image*, *a prompt* for generation, and *a problem* for understanding. The real images are sourced  
7 from the COCO 2017 train split [5], while the problems and their corresponding solutions are adapted  
8 from A-OKVQA [8] and GPT-VQA [10]. In addition, prompts are selected from the original COCO  
9 captions based on their entity coverage with the problem solutions.

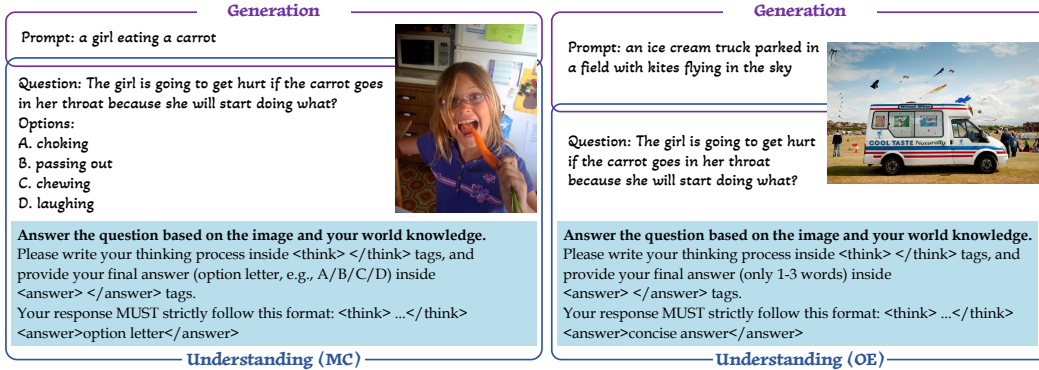


Figure 1: Representative training samples used in unified reinforcement learning.

10 **Training Data for Refined Reinforcement Learning.** In this stage, we collect three specialized  
11 datasets for task-specific RL. For text-to-image generation, we continue constructing a dataset (*i.e.*,  
12 `x2x_rft_16k`) with prompts derived from COCO captions. Moreover, we curate `mcot_r1_mcq` and  
13 `mcot_r1_vqa` for multiple-choice and open-ended multimodal understanding, respectively. These  
14 two datasets encompass a diverse range of multimodal tasks, including mathematical reasoning,  
15 science-problem solving, and visual commonsense reasoning, across multiple source datasets. Specif-  
16 ically, `mcot_r1_mcq` consists of A-OKVQA [8], M<sup>3</sup>CoT [1], SQA-IMG (train) [6], ArxivQA [4],  
17 TabMWP (MC) [7], and MAVIS-Instruct (MC) [9], while `mcot_r1_vqa` includes GeomVerse [3],  
18 R-CoT [2], TabMWP (OE) [7], and MAVIS-Instruct (OE) [9].

## A.2 Supplementary Experimental Setups

Table 1 provides detailed hyperparameter settings for ULM-R1’s RL training.

Table 1: **Training hyperparameter setting.**

Configuration	Unified RL	Refined RL (T2I)	Refined RL (MM2T-MC)	Refined RL (MM2T-OE)
Number of sampled outputs ( $G$ )	8	16	16	16
Regularization coefficient of $\mathbb{D}_{\text{KL}}$ ( $\beta$ )	0	0.02	0.02	0.02
Max prompt length	1024	256	1024	1024
Max completion length	512	/	512	512
Batch size	16	16	32	32
Peak learning rate	4e-6	1e-6	1e-6	1e-6
Epoch	1	1	1	1

## References

- [1] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M<sup>3</sup>CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv:2405.16473*, 2024. 1
- [2] Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv:2410.17885*, 2024. 1
- [3] Mehran Kazemi, Hamidreza Alvani, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv:2312.12241*, 2023. 1
- [4] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv:2403.00231*, 2024. 1
- [5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1
- [6] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pages 2507–2521, 2022. 1
- [7] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *ICLR*, 2023. 1
- [8] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162, 2022. 1
- [9] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv:2407.08739*, 2024. 1
- [10] Zhiyuan Zhao, Linke Ouyang, Bin Wang, Siyuan Huang, Pan Zhang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Mllm-dataengine: An iterative refinement approach for mllm. *arXiv:2308.13566*, 2023. 1