2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

Reinforcement Learning-Driven Multimodal Medical Foundation Model for Unified Understanding, Reasoning and Generation

Wenting Chen, Lei Xing
Department of Radiation Oncology, Stanford University

ABSTRACT

Purpose: Build one multimodal medical foundation model for understanding, reasoning, and generation.

Approach: Train a discrete-diffusion MLLM with cycle-consistency and multimodal inpainting rewards.

Key results: On IU X-Ray, BLEU-1 +62.5%, ROUGE-L +50.0%, FID -4.0%, and NIQE -13.4%.

Significance: RL-based cross-modal alignment improves performance and interpretability.

INTRODUCTION

Clinical need: Existing medical foundation models recognize well but generate inconsistently and offer limited interpretability.

Hypothesis: A shared multimodal model can unify report generation, image synthesis, and grounded reasoning.

Contribution: RL rewards enforce coarse and fine image-text alignment instead of relying on fixed loss weights.

PROPOSED METHOD

Backbone: A discrete-diffusion multimodal LLM denoises image and text tokens in parallel.

Rewards: Cycle consistency and multimodal inpainting enforce coarse-to-fine alignment.

Use cases: Report generation, image synthesis, and grounded multimodal reasoning.

RESULTS

- Report generation improves strongly over baseline: BLEU-1 +62.5%, METEOR +10.2%, and ROUGE-L +50.0%.
- Image synthesis quality improves with lower FID (-4.0%) and NIQE (-13.4%).
- Grounded reasoning stays tied to visual evidence, enabling clinician-reviewable outputs.

Table 1 Performance on report generation task

Methods	BLEU-1↑	Meteor↑	Rouge-L↑
Baseline	0.2559	0.1468	0.2439
Ours	0.4161	0.1621	0.3656

Table 2 Performance on image synthesis task

Methods	FID↓	NIQE↓
Baseline	1.6597	3.8206
Ours	1.5938	3.3096

DISCUSSION

- RL links generated text and images to reconstructable multimodal evidence.
- Coarse-to-fine rewards encourage both semantic consistency and token-level alignment.
- Grounded reasoning improves interpretability, but broader external validation is still needed.

CONCLUSIONS

This work indicates the potential for a single RL-optimized medical foundation model to deliver high-performing and interpretable multimodal AI. Performance gains were observed across both recognition (report generation) and generative (text-to-image) tasks through RL-based cross-modal alignment, while visual grounding enabled transparent, spatially verifiable reasoning.

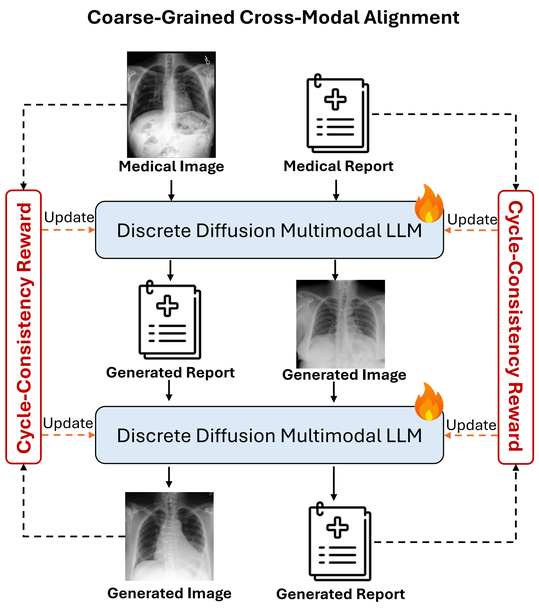

Figure 1 A cycle-consistency reward enforcing bidirectional coarse-grained cross-modal alignment.

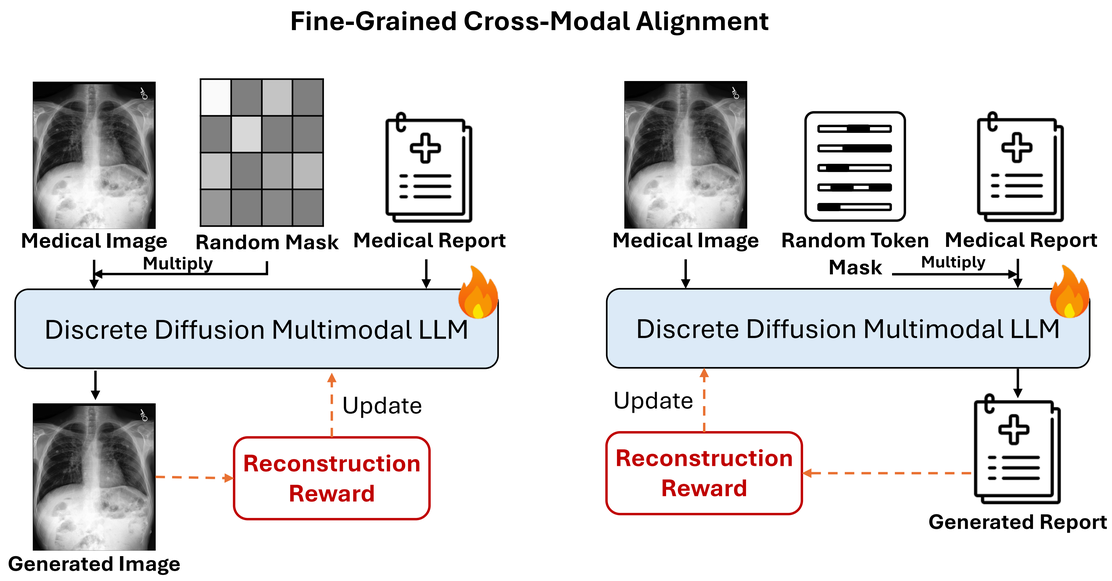

Figure 2 A multimodal inpainting reward enforcing fine-grained alignment.

REFERENCES

[1] Khan, Wasif, et al. "A comprehensive survey of foundation models in medicine." IEEE Reviews in Biomedical Engineering (2025).
[2] Kim, Kyungho, et al. "Image generation is may all you need for vqa." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.
[3] Zhang, Xintong, et al. "Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl." arXiv e-prints (2025): arXiv-2505.
[4] Teh, Yee, et al. "Distral: Robust multitask reinforcement learning." Advances in neural information processing systems 30 (2017).
[5] Yang, Ling, et al. "Mmada: Multimodal large diffusion language models." Advances in Neural Information Processing Systems 38 (2026): 138867-138907.

CONTACT

Wenting Chen
Department of Radiation Oncology
Stanford University