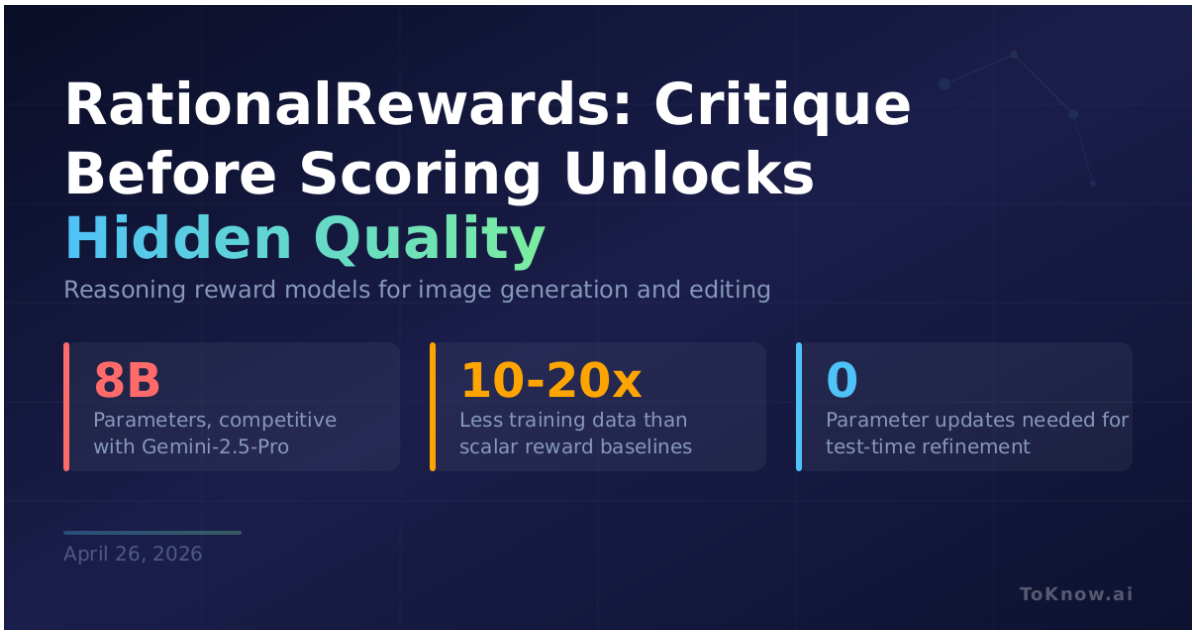
RationalRewards: Reward Models That Critique Before Scoring Unlock Hidden Image Generator Quality

Kabui, Charles

2026-04-26

Read at ToKnow.ai



TIGER-AI-Lab released RationalRewards, an 8B reward model built on Qwen3-VL-Instruct-8B that generates structured critiques before scoring image outputs. Instead of collapsing

human judgment into one number, it evaluates text faithfulness, visual quality, and text rendering separately, backing each score with specific visual evidence. Training uses PARROT (Preference-Anchored Rationalization), which recovers reasoning supervision from cheap preference data in three phases: a teacher model proposes critique candidates anchored to known labels, inconsistent explanations are filtered out, and a student learns to critique without seeing labels. This needs 10-20x less training data than comparable scalar baselines. At training time, these structured rationales serve as reinforcement learning rewards. At test time, the same critiques drive a Generate-Critique-Refine loop that rewrites prompts to fix identified flaws, with no retraining required. RationalRewards 8B achieves state-of-the-art preference prediction among open-source reward models, competitive with Gemini-2.5-Pro across both text-to-image and image-editing tasks.

Test-time prompt refinement, requiring zero parameter updates, matches or exceeds full RL fine-tuning on several benchmarks. Existing image generators have significant latent capability that bad prompts fail to surface. A designer can run a critique-and-refine loop at inference instead of spending hundreds of GPU-hours on retraining. The critique structure also resists reward hacking: because every score must be justified by concrete visual evidence, gaming the reward through superficial tricks becomes much harder. Both models and all training data are publicly released.

Quality gains in image generation may be shifting from model capacity to prompt design and reward structure. Meta's process-driven generation showed that inspect-and-fix loops improve compositional accuracy during generation. RationalRewards demonstrates the same principle at evaluation time: structured reasoning unlocks gains that bigger models or longer training alone cannot.

Sources:

- RationalRewards Paper (arXiv:2604.11626)
- RationalRewards Project Page
- RationalRewards GitHub Repository
- RationalRewards-8B-T2I Model (HuggingFace)
