

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 ON ALIGNMENT OF UNIFIED MULTIMODAL LARGE LAN- GUAGE MODELS

Anonymous authors

Paper under double-blind review

## ABSTRACT

Unified Multi-Modal Large Language Models (U-MLLMs) have demonstrated strong capabilities in text-to-image (T2I) generation, but most post-training methods still rely on sparse, image-level rewards and place limited emphasis on safety. In this work, we take an exploratory view of *dense* reward signals for U-MLLMs: token-level feedback derived from existing reward and evaluation models. Rather than proposing a new RL algorithm, We study how dense rewards can be extracted, how they behave, and how they can be integrated into the standard Group Relative Policy Optimization (GRPO) framework. Concretely, we investigate four questions: (1) how to obtain dense token-level rewards from scalar reward models such as HPSv2; (2) what the empirical behavior and distribution of dense rewards over image tokens look like; (3) how to incorporate dense rewards into GRPO via token-weighted advantages while preserving group-wise sample rankings; and (4) how different interpretability methods compare as providers of dense reward, including trade-offs in localization, computational cost, and downstream performance. On WISE and GenAI-Bench, dense-reward variants of a Janus-Pro-7B U-MLLM achieve competitive image quality (e.g., WISE: 0.50) with slightly smoother training dynamics compared to a sparse-reward T2I-R1 baseline. As a preliminary case study, we also instantiate a safety-focused variant that combines safety reward and observe a 59.4% reduction in unsafe content on the MMDT benchmark relative to the base model. Overall, our results suggest that dense reward is a promising but nuanced design axis for U-MLLM post-training.

Content warning: this paper contains content that may be inappropriate or offensive.

## 1 INTRODUCTION

The recent development of Unified Multi-Modal Large Language Models (U-MLLMs) has shown impressive performance in both image-to-text (I2T) and text-to-image (T2I) tasks (Chen et al., 2025c; Xie et al., 2024b; Deng et al., 2025). These models can not only understand visual input, but also generate high-quality images given complex textual prompts, providing new tools for digital media content generation. As these models scale, however, two alignment questions become increasingly important: *where* feedback is applied within a trajectory (sparse vs. dense reward) and *what* objectives are being optimized (quality, safety, or both).

One major limitation of existing T2I refinement methods (Jiang et al., 2025a) is the reliance on *sparse* reward signals (Chan et al., 2024): a single scalar score is assigned to an entire generated image to represent its quality and alignment, often via an ensemble of reward models. This approach fails to provide the granular feedback that is necessary for the policy model to understand which specific parts of the image contribute to or detract from the overall reward. More fine-grained, token-level rewards could, in principle, guide the model’s learning process more effectively, but it is unclear how best to obtain such dense rewards from existing models, how these rewards behave, and how they interact with standard RL objectives such as GRPO (Guo et al., 2025).

From a safety perspective, another limitation of existing T2I refinement method is that stronger generative capabilities can also make it easier to produce toxic or harmful content (see right of Figure 5). In practice, current U-MLLM post-training methods, such as T2I-R1 (Jiang et al., 2025a;b), are primarily optimized for image quality, compositional accuracy, and text–image alignment. To

054 better understand the current state of safety alignment, we benchmark several U-MLLMs on MMDT  
 055 and observe that quality-focused post-training can degrade safety (left of Figure 5), motivating a  
 056 closer examination of how reward design and training procedures interact with safety.  
 057

058 In this work, we investigate dense reward signals for U-MLLMs in the context of GRPO-based  
 059 post-training. Our goal is not to introduce a new RL algorithm, but instead to explore a simple way  
 060 of incorporating dense token-level feedback derived from reward models and interpretability tools,  
 061 and to characterize the resulting behavior. We focus primarily on image quality and alignment, and  
 062 treat safety as a focused case study that illustrates how we can improve image quality and safety in  
 063 the same time.  
 064

Concretely, we structure our study around the following four research questions:

- 065 • **RQ1: How can we obtain dense rewards from existing reward models?** We investigate  
 066 how to extract token-level scores from scalar feedback models such as HPSv2 (Wu et al.,  
 067 2023) using interpretability tools (SHAP (Schulman et al., 2017a), LIME (Lundberg & Lee,  
 068 2017), Grad-CAM (Selvaraju et al., 2019)). (see subsection 3.1)
- 069 • **RQ2: What is the behavior and distribution of dense rewards in image generation?** We  
 070 empirically analyze the localization and entropy of dense token-level rewards over image  
 071 tokens in a U-MLLM, comparing interpretability tools. (see subsection 3.2)
- 072 • **RQ3: How can dense rewards be integrated into GRPO training?** We study a simple  
 073 token-weighted GRPO objective that keeps group-wise advantages fixed while redistributing  
 074 them across tokens according to dense scores. (see subsection 3.3)
- 075 • **RQ4: Which interpretability choices work better for dense reward, and what are the  
 076 trade-offs?** We compare different interpretability tools as sources of dense reward, and  
 077 conduct a preliminary case study on safety-oriented rewards. (see subsection 3.4)

078 Our contributions can be summarized as follows:  
 079

- 080 • **Dense reward extraction (RQ1).** We investigate how to obtain dense image-token rewards  
 081 from existing scalar and HPSv2 using different interpretability tools, and describe simple  
 082 transformations from spatial attribution to image tokens in unified T2I models.
- 083 • **Characterizing dense reward distributions (RQ2).** We empirically study how dense  
 084 rewards are distributed over image tokens in T2I generation, measuring localization (top- $k$   
 085 mass) and entropy across interpretability methods, and show that a small subset of tokens  
 086 dominates the reward contribution.
- 087 • **Integrating dense reward into GRPO (RQ3).** We evaluate a token-weighted GRPO  
 088 objective that preserves group-wise advantages derived from scalar rewards and uses dense  
 089 scores only to redistribute advantages across tokens. We compare the resulting training  
 090 dynamics and image quality to a sparse-reward T2I-R1 baseline.
- 091 • **Interpretability trade-offs and safety case study (RQ4).** We compare various inter-  
 092 pretability tools as sources of dense reward, highlighting trade-offs in computational cost,  
 093 and empirical gains. We further present a preliminary safety case study that combines  
 094 toxicity-aware rewards, observing a substantial reduction in unsafe generations.

095 Overall, our results indicate that dense reward provides a useful lens on U-MLLM alignment: even  
 096 when image quality metrics improve only modestly, dense signals reveal highly localized reward struc-  
 097 ture and can yield smoother training, while safety-specific dense rewards offer a promising—though  
 098 still early-stage—direction for future work.  
 099

## 100 2 PRELIMINARY

### 101 2.1 PROBLEM FORMULATION

102 Given a text prompt  $p$ , the goal is to generate an image  $I$  that maximizes alignment with the prompt  
 103 while maintaining high perceptual quality. We adopt a two-stage generation process with model  $\pi_\theta$ :

- 104 1. **Semantic CoT:** Generate reasoning text  $c \sim \pi_\theta(\cdot \mid p)$  that describes or reasons about the  
 105 image to be generated.

108    2. **Image Token CoT**: Generate image tokens  $\mathbf{t} = \{t_1, \dots, t_N\} \sim \pi_\theta(\cdot \mid p, c)$  where  $N$  is the  
 109    number of image tokens (576 for Janus-Pro (Chen et al., 2025c)).  
 110

111    As shown in Figure 1, the image tokens are decoded by the image tokenizer into an image  $I$ , which is  
 112    then evaluated by an ensemble of reward models. The resulting scalar rewards and dense token-level  
 113    feedback are used to update the model via reinforcement learning methods such as (Guo et al., 2025).  
 114

115    2.2 GRPO FOR IMAGE GENERATION  
 116

117    **Group-wise advantage estimation.** For each prompt  $p$ , we sample a group of  $G \times K$  responses,  
 118    comprising  $G$  semantic CoT completions with  $K$  image generations per completion, following (Jiang  
 119    et al., 2025a). Let  $\{o_i\}_{i=1}^{G \times K}$  denote this response group sampled from the old policy  $\pi_{\theta_{\text{old}}}$ , where  
 120    each  $o_i = (c_i, \mathbf{t}_i)$  is a full multimodal trajectory.

121    Each response  $o_i$  receives a scalar reward  $R_i$  from our ensemble of reward models (see subsection 3.4).  
 122    Following GRPO (Guo et al., 2025), we compute the advantage of the  $i$ -th response by normalizing  
 123    rewards within the group:

$$124 \quad 125 \quad 126 \quad A_i = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^{G \times K})}{\text{std}(\{R_i\}_{i=1}^{G \times K})}. \quad (1)$$

127    This group-relative normalization produces advantages that are approximately zero-mean and contain  
 128    both positive and negative values, while  $R_i$  remains positive as illustrated in right Figure 1.

129    GRPO employs a clipped surrogate objective similar to PPO (Schulman et al., 2017b). For each token  
 130    position  $j$  in response  $o_i$ , we define the probability ratio

$$132 \quad 133 \quad r_{i,j}(\theta) = \frac{\pi_\theta(o_{i,j} \mid p, o_{i,<j})}{\pi_{\theta_{\text{old}}}(o_{i,j} \mid p, o_{i,<j})}, \quad (2)$$

134    where  $o_{i,<j}$  denotes the prefix tokens preceding position  $j$  in  $o_i$ . The GRPO objective is  
 135

$$136 \quad 137 \quad 138 \quad \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{p \sim \mathcal{D}, \{o_i\}_{i=1}^{G \times K} \sim \pi_{\theta_{\text{old}}}(\cdot \mid p)} \left[ \frac{1}{\sum_{i=1}^{G \times K} |o_i|} \sum_{i=1}^{G \times K} \sum_{j=1}^{|o_i|} \mathcal{L}_{i,j}(\theta) \right], \quad (3)$$

139    with per-token loss  
 140

$$141 \quad 142 \quad \mathcal{L}_{i,j}(\theta) = \min(r_{i,j}(\theta) A_i, \text{clip}(r_{i,j}(\theta), 1 - \epsilon, 1 + \epsilon) A_i) - \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}), \quad (4)$$

143    where  $\epsilon$  controls the clipping range (typically 0.2),  $\beta$  weights the KL penalty, and  $\pi_{\text{ref}}$  is a reference  
 144    policy (typically the SFT model).  
 145

146    3 EXPLORATION AND OBSERVATION  
 147

148    Our framework builds upon GRPO (Guo et al., 2025). We keep the GRPO formulation unchanged  
 149    and extend it by introducing token-specific weights  $w_{i,j}$  to enable fine-grained control over the policy  
 150    gradient in the T2I domain, addressing RQ1 and RQ3.  
 151

152    3.1 DENSE REWARD V1: SHAP-BASED TOKEN-LEVEL HUMAN PREFERENCE SCORE  
 153

154    **Token contribution via Shapley values.** We employ Shapley Additive Explanations (SHAP) (Lund-  
 155    berg & Lee, 2017) to quantify each token’s contribution to the overall reward from HPS-v2 (Wu et al.,  
 156    2023), providing interpretable token-level importance. For a reward model, such as  $r_{\text{HPS}}(I, p)$  (Wu  
 157    et al., 2023) that evaluates image  $I$  with prompt  $p$ , the Shapley value for token  $j$  is defined as

$$158 \quad 159 \quad 160 \quad \phi_j = \sum_{S \subseteq \mathcal{N} \setminus \{j\}} \frac{|S|! (|\mathcal{N}| - |S| - 1)!}{|\mathcal{N}|!} (r_{\text{HPS}}(S \cup \{j\}) - r_{\text{HPS}}(S)), \quad (5)$$

161    where  $\mathcal{N} = \{1, \dots, N\}$  represents the set of all token indices (Lundberg & Lee, 2017).

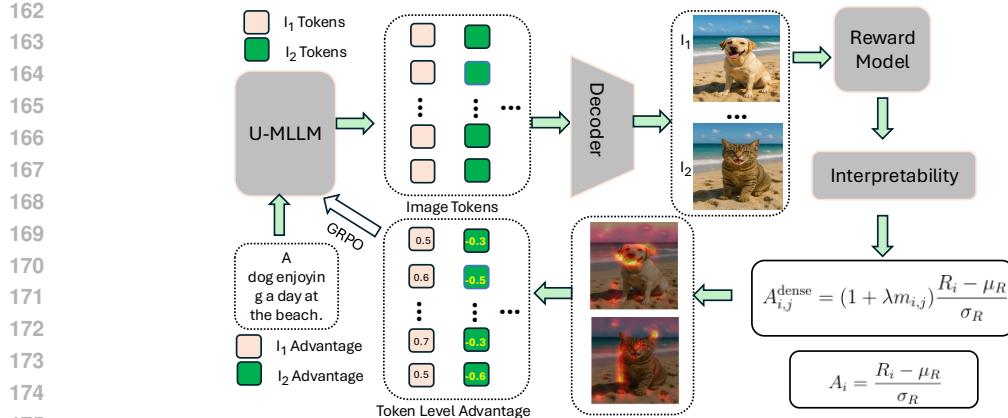


Figure 1: GRPO with dense reward. The U-MLLM generates text and image tokens, which are decoded into images and evaluated by reward models to produce token-level feedback.

**Practical implementation.** We use partition-based SHAP with image masking for efficiency:

$$\phi_{\text{spatial}} = \text{SHAP}(r_{\text{HPS}}, I, \text{mask} = \text{blur}(24 \times 24)) \in \mathbb{R}^{H_{\text{img}} \times W_{\text{img}}}, \quad (6)$$

where blur masking is applied to image regions to estimate feature importance and  $H_{\text{img}} = W_{\text{img}} = 384$  denotes the image dimensions.

**Spatial to token-level mapping.** The spatial attribution map is aggregated to patch-level and then mapped to token space:

$$\phi_{x,y}^{\text{patch}} = \frac{1}{H_p \times W_p} \sum_{h=0}^{H_p-1} \sum_{w=0}^{W_p-1} \phi_{\text{spatial}}[x \cdot H_p + h, y \cdot W_p + w], \quad (7)$$

where  $(x, y) \in [0, D) \times [0, D)$  are patch coordinates. The patch attributions are then flattened to a token sequence:

$$\phi_{\text{token}} = \text{Flatten}(\phi^{\text{patch}}) \in \mathbb{R}^N, \quad (8)$$

where  $N = D^2 = 576$  is the total number of image tokens. In the middle of Figure 1, the heatmap highlights important regions identified.

**Normalization to unit range.** Token attributions are normalized to  $[0, 1]$  for consistent scaling:

$$m_j = \frac{\phi_j^{\text{token}} - \min(\phi_{\text{token}})}{\max(\phi_{\text{token}}) - \min(\phi_{\text{token}})}, \quad (9)$$

where  $m_j \in [0, 1]$  represents the normalized attribution score for token  $j$ , with higher values indicating greater contribution to the HPS-v2 reward. In addition to SHAP-based token-level human preference scores, we also integrated LIME and Grad-CAM-based scores (see more in Appendix A).

Before introducing our dense-reward integration, we first analyze how standard *sparse* scalar rewards and *dense* token-level feedback behave in practice.

### 3.2 BEHAVIOR OF SPARSE AND DENSE REWARDS

To study the structure of dense feedback (addressing RQ2), we consider interpretability tools  $\tau \in \{\text{SHAP}, \text{LIME}, \text{Grad-CAM}\}$ . Each tool produces token-level scores that we normalize (as described in subsection 3.1) to obtain weights  $m_{i,t}^{(\tau)}$  for each response  $o_i$ . One example is shown in Figure 2. On a subset of training dataset with size  $N_{\text{data}} = 1896$ , we generate one image per prompt. We then use HPSv2 (Wu et al., 2023) as a reward model to score each (prompt, image) pair, and for each triple (prompt, image, score) we apply an interpretability tool to measure how concentrated the contribution is at the token level.

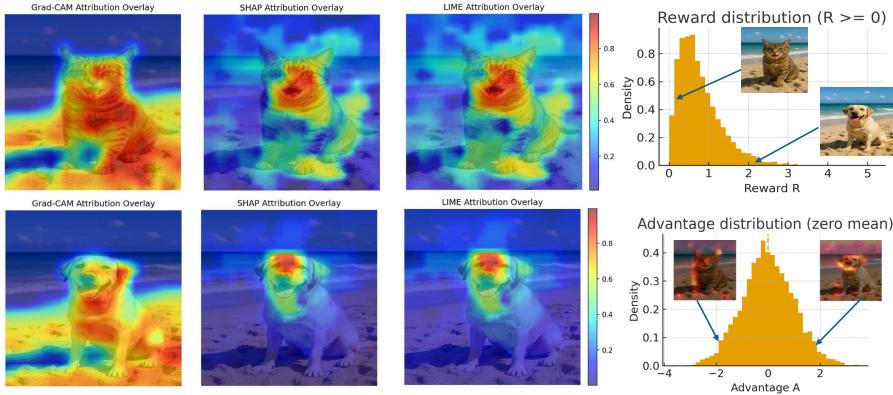


Figure 2: Compare dense reward from different interpretability tools.

**Top- $k$  mass.** We analyze localization using top- $k$  mass. Let  $\text{TopK}(m_{i,:}^{(\tau)}, k)$  be the indices of the top- $k$  tokens and define

$$M_{i,k}^{(\tau)} = \sum_{t \in \text{TopK}(m_{i,:}^{(\tau)}, k)} m_{i,t}^{(\tau)}. \quad (10)$$

For  $k = 0.1 \times N = 57$  (top 10%), a uniform distribution would give  $M_{i,k}^{(\tau)} = 0.10$ , but we observe

$$\text{SHAP: } \mathbb{E}_i[M_{i,k}^{(\text{SHAP})}] \approx 0.32, \quad \text{LIME: } \approx 0.53, \quad \text{GradCAM: } \approx 0.45,$$

showing that the top 10% tokens carry roughly 3–6× more mass than uniform. Thus, dense feedback is highly localized: a small subset of tokens dominates the reward contribution.

**Entropy of token weights.** We also measure localization via the Shannon entropy

$$H_i^{(\tau)} = - \sum_{t=1}^N m_{i,t}^{(\tau)} \log m_{i,t}^{(\tau)}. \quad (11)$$

For  $N = 576$ , a uniform distribution has  $H \approx \log 576 \approx 6.35$ , while we obtain

$$\mathbb{E}_i[H_i^{(\text{SHAP})}] \approx 5.98, \quad \mathbb{E}_i[H_i^{(\text{LIME})}] \approx 5.15, \quad \mathbb{E}_i[H_i^{(\text{GradCAM})}] \approx 5.48,$$

indicating that all three methods produce non-uniform, localized attributions. These results show that the scalar rewards  $R_i$  provide only a global, sample-level signal, while dense feedback reveals that reward contributions are concentrated on a small subset of image tokens. This suggests that an effective RL algorithm for U-MLLMs should maintain the sample-level ranking induced by  $R_i$  (and hence  $A_i$ ), but redistribute gradients within each trajectory according to dense token-level structure. We formalize this idea in the following section.

### 3.3 TOKEN WEIGHT ASSIGNMENT AND ADVANTAGE MODULATION

The ensemble reward models produce scalar rewards  $R_i$  per response, and GRPO converts  $\{R_i\}$  within each group into advantages  $\{A_i\}$  via normalization: high-quality samples have  $A_i > 0$  and low-quality samples have  $A_i < 0$ , with  $\mathbb{E}[A_i] \approx 0$  (right of Fig. 2).

For a trajectory  $o_i = (c_i, t_i)$  consisting of semantic CoT tokens followed by image tokens, we define:

**Semantic tokens (CoT reasoning).** We do not apply spatial re-weighting to CoT tokens:

$$w_{i,j} = 1, \quad \forall j \in \{1, \dots, |c_i|\}. \quad (12)$$

**Image tokens.** For image tokens, we use dense scores to modulate the gradient:

$$w_{i,j} = 1 + \lambda m_{i,j}, \quad \forall j \in \{|c_i| + 1, \dots, |c_i| + N\}, \quad (13)$$

270 where  $\lambda$  is a scalar hyperparameter that controls the strength and direction of spatial feedback. When  
 271  $m_{i,j}$  is interpreted as a *preference* score (SHAP), we choose  $\lambda > 0$  so that high-preference tokens are  
 272 up-weighted; when  $m_{i,j}$  is interpreted as a *misalignment* score (RAHF), we choose  $\lambda < 0$  so that  
 273 highly misaligned tokens are down-weighted.

274 Dense methods provide additional structure through normalized token scores  $\{m_{i,j}\}_{j=1}^N$  that indicate  
 275 how much each token contributes to the final reward. A naive design would be to form a "dense"  
 276 reward  $\tilde{R}_i$  by directly up-weighting tokens with large  $m_{i,j}$  and then recomputing advantages from  
 277  $\tilde{R}_i$ . This mixes token importance with sample quality and can *increase* the reward of low-quality  
 278 samples, shrinking the gap between good and bad responses and weakening the GRPO signal.  
 279

280 Instead, we first compute group-wise advantages  $A_i$  from the original scalar rewards  $R_i$ , preserving  
 281 the sample-level ranking, and then use dense scores only to *redistribute*  $A_i$  across tokens. We  
 282 introduce token-specific weights  $w_{i,j}$  and define

$$283 \quad A_{i,j} = w_{i,j} A_i \quad (14)$$

284 as per-token advantages. For low-quality samples in group ( $A_i < 0$ ), all  $A_{i,j}$  remain negative, and  
 285 tokens with larger  $w_{i,j}$  receive *more negative* credit; for high-quality samples ( $A_i > 0$ ), tokens with  
 286 larger  $w_{i,j}$  receive *more positive* credit. Thus dense scores control how the fixed total advantage  $A_i$   
 287 is distributed within the trajectory, without changing which samples are group-wise good or bad.  
 288

289 We then replace  $A_i$  by  $A_{i,j}$  in the GRPO loss:

$$290 \quad \mathcal{L}_{i,j}(\theta) = \min(r_{i,j}(\theta) A_{i,j}, \text{clip}(r_{i,j}(\theta), 1 - \epsilon, 1 + \epsilon) A_{i,j}) - \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}). \quad (15)$$

292 Overall, this formulation keeps the scalar rewards  $R_i$  and group-wise advantages  $A_i$  intact and uses  
 293 dense reward purely to *shape per-token advantages*. Tokens in well-aligned regions receive larger  
 294  $w_{i,j}$  and thus contribute more strongly to the gradient update, while tokens in misaligned regions are  
 295 de-emphasized, enabling dense-reward optimization that respects the global ranking.

### 297 3.4 ENSEMBLE OF REWARD MODELS

299 The assessment of image generation is a hard task, since it requires evaluating multiple criteria, from  
 300 aesthetics to prompt alignment. To create a more robust and holistic learning signal, we employ an  
 301 ensemble of specialized reward models as shown in Table 1, each targeting a different aspect of the  
 302 generation process, similar to (Jiang et al., 2025a).

304 Table 1: Ensemble of Reward Models

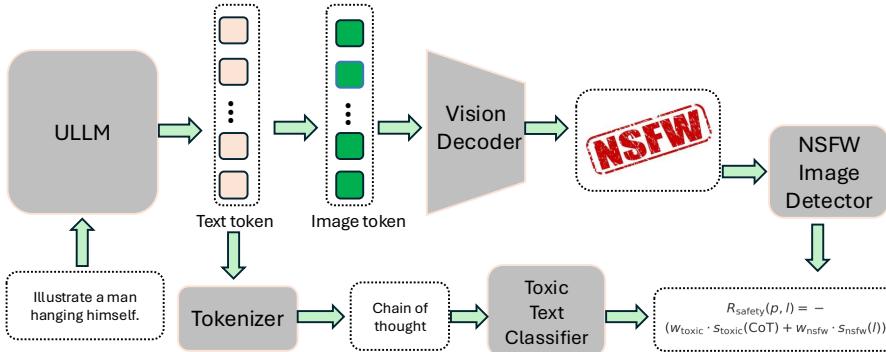
Reward Model	Type	Input(s)	Output(s)	Primary Goal
RAHF	Rich Feedback(RF)	Text, Image	Scores (4), Heatmaps (2)	Fine-grained quality & alignment
HPSv2	Human Preference(HP)	Text, Image	Single Score	Overall quality, alignment
HPSv2 w. SHAP/LIME/Grad-CAM	Dense reward	Text, Image	Score, Heatmap	Fine-grained quality & dense reward
GroundingDINO	Object Detection	Image	Bounding Boxes, Labels	Object localization correctness
GIT / LLaVA-ORM	VQA / Relational	Text, Image	Answer / Relation Score	Compositional & relational accuracy
Toxic-BERT	Safety (Text)	Text	Toxicity Score	Penalize textual COT
NSFW Detector	Safety (Image)	Image	NSFW Score	Penalize unsafe image

313 **Rewards for safe prompts.** For prompts deemed safe, our focus is on maximizing image quality,  
 314 user preference alignment, and compositional correctness. We use a combination of reward models:

- 316 • **Rich Human Feedback (RAHF)** (Liang et al., 2024) provides fine-grained scores (plausibility,  
 317 alignment, aesthetics, overall) and heatmaps (misalignment, implausibility). Alignment  
 318 scores are used as part of the scalar reward, and the misalignment heatmap is used as the  
 319 basis for token-level weighting.
- 320 • **Human Preference Score v2 (HPSv2)** (Wu et al., 2023) provides a single score reflecting  
 321 general human preference. We use this score as the target for our SHAP/LIME-based  
 322 token-level dense reward, allowing us to determine the contribution of each image token.
- 323 • **Object detection/localization:** GroundingDINO (Liu et al., 2023) provides object bounding  
 324 boxes, used to penalize generations where objects are absent or incorrectly localized.

324

- 325 • **Visual Question Answering (VQA):** For assessing compositional accuracy, especially
- 326 complex spatial or attribute-based relationships, we employ GIT (Wang et al., 2022).
- 327
- 328 • **Output-Reward Model (ORM):** Following (Jiang et al., 2025a), this model evaluates alignment
- 329 between prompt and image, providing a reward signal for image/prompt alignment.



342 Figure 3: U-MLLM function as both promptist and image generator. We employ NSFW image  
 343 detection and toxic text classification to filter harmful content during the sampling process.  
 344

345 **Rewards for unsafe prompts.** As a preliminary safety-oriented case study, we examine prompts  
 346 that are classified as unsafe. Unlike conventional image generators, U-MLLMs function as both  
 347 promptist and image generators. Accordingly, our reward strategy incorporates both chain-of-thought  
 348 (CoT) monitoring and toxic image detection to strongly disincentivize the generation of harmful  
 349 or inappropriate content. To achieve this, we define a composite safety reward  $R_{\text{safety}}$  based on the  
 350 outputs of safety-related models:

351

- 352 • **Toxic-BERT:** We analyze intermediate textual CoT tokens using a Toxic-BERT classifier  
 353 (Hanu & Unitary team, 2020) to obtain a toxicity score,  $s_{\text{toxic}}(c) \in [0, 1]$ .
- 354 • **NSFW image detection:** The generated image  $I$  is processed by an NSFW image detection  
 355 model (Falconsai, 2023), which yields an NSFW score,  $s_{\text{nsfw}}(I) \in [0, 1]$ .

356 As shown in Figure 3, the final safety reward is formulated as a weighted penalty that combines both  
 357 scores. A high score from either classifier results in a large negative reward, heavily suppressing any  
 358 policy that generates unsafe content:

$$R_{\text{safety}}(p, I) = -(w_{\text{toxic}} \cdot s_{\text{toxic}}(c) + w_{\text{nsfw}} \cdot s_{\text{nsfw}}(I)), \quad (16)$$

360 where  $w_{\text{toxic}}$  and  $w_{\text{nsfw}}$  are hyperparameters (default: 1.0) that control the penalty magnitudes for  
 361 toxic text and NSFW images, respectively. This reward structure ensures that safety is integrated in  
 362 the optimization and serves as an initial exploration of safety alignment for U-MLLMs.

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETUP

370 **Training Configuration.** We employ two distinct training settings to instantiate our study: (1)  
 371 **T2I-R1-Dense**, which targets image quality improvement using dense rewards and trains exclusively  
 372 on safe prompts; and (2) **T2I-R1-Safety**, which primarily serves as a safety-oriented case study  
 373 by training on a mixed dataset of safe and unsafe prompts. For safe prompts, following recent  
 374 work (Jiang et al., 2025a), we utilize a training set of 6,786 text-only prompts curated from datasets  
 375 such as T2I-CompBench (Huang et al., 2023); for unsafe prompts, we mix the training prompts  
 376 from (Li et al., 2025) and those safe prompts. Our implementation is built upon Janus-Pro-7B (Chen  
 377 et al., 2025c) as the base model, which we train with a learning rate of  $1 \times 10^{-6}$  and a KL divergence  
 378 coefficient of  $\beta = 0.01$ . Experiments are conducted on H200, A100 GPUs. (see details in Table 8).

378     **Evaluation Benchmarks.** We conduct a comprehensive evaluation across four established benchmarks to assess our model’s performance on image quality and generation safety:

- 381     • **GenAI-Bench** (Li et al., 2024a): Measures compositional text-to-visual generation capabilities through prompts covering spatial relationships, attribute binding, and scene complexity.
- 382
- 383     • **WISE** (Niu et al., 2025): Evaluates world knowledge integration and complex semantic understanding using 1,000 meticulously crafted prompts across three major domains (cultural common sense, spatio-temporal reasoning, and natural science).
- 384
- 385     • **MMDT** (Xu et al., 2025): Assesses bidirectional safety in both text-to-image and image-to-text generation tasks, covering harmful content detection from various subdomains.
- 386
- 387     • **T2I-Safety** (Li et al., 2025): Specifically targets text-to-image safety evaluation, focusing on detection of harmful or toxic image content.
- 388
- 389
- 390

## 391     5 RESULTS AND DISCUSSION

394     This section presents results on WISE benchmarks evaluating compositional understanding and world knowledge integration. Results for GenAI-Bench are presented in Appendix B.

398     Table 2: **WISE Result.** The best score is in blue, with the second-best score in green.

399     Model	400     Cultural↑	401     Spatio-Temporal		402     Natural Science			403     Overall
		404     Time↑	405     Space↑	406     Biology↑	407     Physics↑	408     Chemistry↑	
<i>Diffusion Models</i>							
PixArt-Alpha Chen et al. (2023a)	0.45	0.50	0.48	0.49	0.56	0.34	0.47
playground-v2.5 Li et al. (2024b)	0.49	0.58	0.55	0.43	0.48	0.33	0.49
SD-v1.5 Rombach et al. (2022a)	0.34	0.35	0.32	0.28	0.29	0.21	0.32
SD-XL-base-0.9 Podell et al. (2023a)	0.43	0.48	0.47	0.44	0.45	0.27	0.43
FLUX.1-dev Black Forest Labs (2024)	0.48	0.58	0.62	0.42	0.51	0.35	0.50
<i>AutoRegressive Models</i>							
Emu3 Wang et al. (2024)	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Show-o Xie et al. (2024b)	0.28	0.40	0.48	0.30	0.46	0.30	0.35
VILA-U Wu et al. (2024d)	0.26	0.33	0.37	0.35	0.39	0.23	0.31
Janus-1.3B Wu et al. (2024a)	0.16	0.26	0.35	0.28	0.30	0.14	0.23
Janus-Pro-7B (Baseline) Chen et al. (2025c)	0.30	0.37	0.49	0.36	0.42	0.26	0.35
T2I-R1 Jiang et al. (2025a)	0.47	0.50	0.62	0.48	0.57	0.32	0.49
<b>T2I-R1-Dense-RAHF (Ours)</b>	0.45	0.47	0.62	0.48	0.56	0.27	0.48
<b>T2I-R1-Dense-HPS-LIME (Ours)</b>	0.46	0.54	0.61	0.48	0.55	0.28	0.49
<b>T2I-R1-Dense-HPS-SHAP (Ours)</b>	0.48	0.50	0.63	0.50	0.58	0.32	0.50

415     **WISE Benchmark Performance.** Table 2 demonstrates the effectiveness of dense reward on the WISE benchmark, which evaluates world knowledge integration across cultural, spatio-temporal, and natural science domains. T2I-R1-Dense-HPS-SHAP method achieves the highest overall score of 0.50, matching the performance of FLUX.1-dev and surpassing all other autoregressive models. Notably, all three dense reward variants show substantial improvements over the Janus-Pro-7B baseline (0.35), with gains ranging from +37% to +43%.

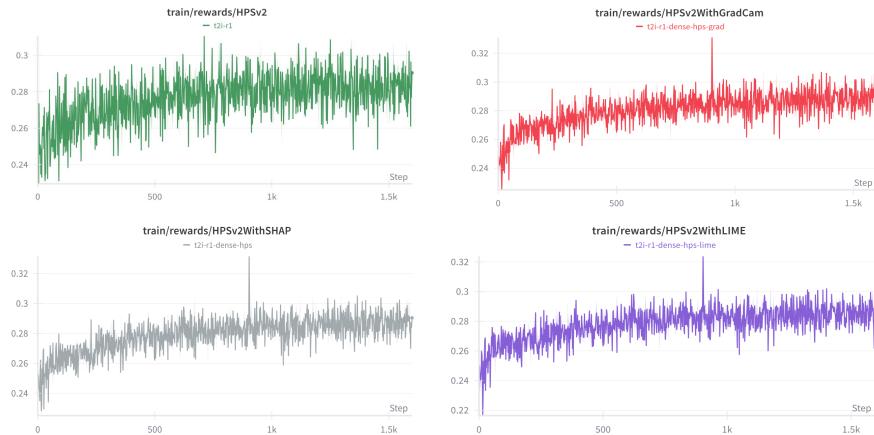
421     The performance varies across different knowledge domains. Dense reward based methods excel 422 particularly in spatial reasoning (0.61–0.63) and physics understanding (0.55–0.58), suggesting that 423 fine-grained token-level feedback effectively guides the model to better capture spatial relationships 424 and physical concepts. The relatively lower performance in chemistry (0.27–0.32) indicates room for 425 improvement in specialized domain knowledge.

426     **Comparison of Dense reward** The results demonstrate that incorporating all four components (HPS, 427 GIT, GDINO, ORM) yields the best overall performance of 0.50, with notable improvements in 428 Biology (+0.08) and Chemistry (+0.03) compared to the HPS-only baseline. Approaches. The 429 three attribution methods show complementary strengths: HPS-SHAP achieves the best WISE 430 performance (0.50), HPS-LIME shows balanced results across both benchmarks, while RAHF 431 maintains consistent quality with direct misalignment feedback. The minimal performance variance 432 ( $\leq 2\%$  across most metrics) suggests that the token-level weighting mechanism itself, rather than the

432 specific interpretability tools, drives the primary improvements. This finding supports our hypothesis  
 433 that fine-grained spatial feedback effectively guides policy optimization regardless of how token  
 434 importance is computed, and aligns with our exploratory focus on dense reward design. There is  
 435 a trade-off between performance and computation cost as discussed in subsection E.1.  
 436

## 437 6 ABLATION STUDY

### 438 6.1 TRAINING DYNAMICS



456 Figure 4: Training reward (HPSv2) under different methods: top-left (T2I-R1), top-right (T2I-R1  
 457 with Grad-CAM), bottom-left (T2I-R1 with SHAP), bottom-right (T2I-R1 with LIME).

459 In Figure 4, we plot the evolution of the HPSv2 reward during training for the original T2I-R1  
 460 baseline and our dense-reward variants (Grad-CAM, SHAP, LIME). Across all configurations, the  
 461 reward increases steadily and converges to a similar level, indicating that dense advantage modulation  
 462 preserves the overall optimization behavior of T2I-R1 (Jiang et al., 2025a). At the same time, the  
 463 curves with dense feedback are smoother and exhibit fewer large oscillations, suggesting that focusing  
 464 credit on informative tokens yields a more stable training trajectory without sacrificing the final  
 465 reward.

466 We further conduct ablation studies to validate our design choices and quantify the contribution of  
 467 each component. First, we examine the impact of the token weight coefficient  $\lambda$  in our token-level  
 468 weighting scheme, sweeping values from 0.1 to 1.0 to identify the best trade-off between suppressing  
 469 misaligned regions and preserving useful gradients. Second, to understand the role of individual  
 470 reward components, we ablate different combinations within our reward ensemble, starting from  
 471 RAHF or HPSv2 alone and progressively adding GroundingDINO for object detection and GIT/ORM  
 472 for visual question answering.

### 473 6.2 HYPERPARAMETER $\lambda$

475 We investigate the effect of the hyperparameter  $\lambda$  on model performance, as presented in Table 3.

478 Table 3: Ablation study on hyperparameter  $\lambda$ .

479 Model	Cultural↑	Spatio-Temporal		Natural Science			Overall↑
		Time↑	Space↑	Biology↑	Physics↑	Chemistry↑	
481 T2I-R1-Dense-SHAP ( $\lambda = 0.1$ )	0.48	0.50	0.63	0.50	0.58	0.32	0.50
482 T2I-R1-Dense-SHAP ( $\lambda = 0.5$ )	0.47	0.49	0.56	0.44	0.56	0.31	0.47
483 T2I-R1-Dense-SHAP ( $\lambda = 1.0$ )	0.45	0.50	0.59	0.40	0.50	0.26	0.45

484 Our experiments show that  $\lambda = 0.1$  yields the best performance across most categories, particularly  
 485 for Space (0.63) and Biology (0.50), suggesting that smaller values better balance the training.

486 6.3 REWARD FUNCTION COMPONENTS  
487488 We conduct an ablation study to evaluate the impact of different reward model components on  
489 performance, as shown in Table 4.  
490491 **Table 4: Ablation study on reward model components.**  
492

493 Model	Cultural↑	Spatio-Temporal		Natural Science			Overall↑
		494 Time↑	Space↑	Biology↑	Physics↑	Chemistry↑	
T2I-R1-Dense-SHAP w. HPS	0.49	0.49	0.59	0.42	0.55	0.29	0.47
T2I-R1-Dense-SHAP w. HPS, GIT	0.48	0.50	0.62	0.44	0.56	0.31	0.48
T2I-R1-Dense-SHAP w. HPS, GIT, GDINO	0.43	0.47	0.59	0.40	0.54	0.29	0.45
T2I-R1-Dense-SHAP w. HPS, GIT, GDINO, ORM	0.48	0.50	0.63	0.50	0.58	0.32	0.50

493 The results demonstrate that incorporating all four components (HPS, GIT, GDINO, ORM) yields  
494 the best overall performance of 0.50, with notable improvements in Biology (+0.08) and Chemistry  
495 (+0.03) compared to the HPS-only baseline.  
496

## 501 6.4 CASE STUDY: SAFETY

502 We additionally present results for T2I-R1-Safety as a preliminary case study, where we incorporate  
503 toxic text detection and NSFW image classification into the reward framework to enhance safety  
504 alignment. Results on the MMDT bench and T2I-safety bench are provided in Appendix B.  
505

## 506 6.5 QUALITATIVE STUDY

507 We present our qualitative study in Appendix D.  
508

## 511 7 CONCLUSION AND LIMITATION

512 In this paper, we have presented an empirical study of dense token-level rewards for aligning U-  
513 MLLMs in T2I generation. By integrating dense reward into a GRPO framework via token-weighted  
514 advantages, we showed that rich, fine-grained feedback can be incorporated without changing the  
515 underlying RL algorithm. The use of token-level weights derived from spatial information allows for  
516 more nuanced credit assignment within each trajectory. Our primary focus was on image quality and  
517 alignment, with a safety-oriented configuration included as a case study.  
518519 Our experiments show that T2I-R1-Dense variants achieve competitive performance on image quality  
520 benchmarks (WISE: 0.50, GenAI-Bench: 0.73) with smoother training dynamics compared to a  
521 sparse-reward T2I-R1 baseline, while T2I-R1-Safety substantially reduces unsafe content generation  
522 by 59.4% on MMDT in our experimental setting. These results suggest that a more holistic view of  
523 model alignment—one that combines global scalar rewards with detailed, token-level feedback—is  
524 feasible and can yield practical benefits, even when headline metrics improve only modestly.  
525526 Our work has a few limitations as illustrated in Appendix E. In particular, due to limitation of  
527 resources, we focus on a single base model, study a limited set of interpretability methods, and do  
528 not exhaustively explore hyperparameters. We utilized an LLM to assist our work as acknowledged  
529 in Appendix G, and our ethics statement is in Appendix F.  
530

## 531 REFERENCES

532 Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor  
533 Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 3.0 technical report.  
534 *arXiv preprint arXiv:2312.03511*, 2023.  
535536 Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.  
537538 Alex J. Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. Dense reward for free in  
539 reinforcement learning from human feedback, 2024. URL <https://arxiv.org/abs/2402.00782>.  
540

540 Juhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi  
 541 Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of  
 542 fully open unified multimodal models-architecture, training and dataset, 2025a. URL <https://arxiv.org/abs/2505.09568>.

543

544 Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James  
 545 Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\alpha$ : Fast training of diffusion transformer  
 546 for photorealistic text-to-image synthesis. 2023a.

547

548 Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James  
 549 Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\alpha$ : Fast training of diffusion transformer  
 550 for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.03046*, 2023b.

551

552 Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang,  
 553 and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image  
 554 generation, 2025b. URL <https://arxiv.org/abs/2506.18095>.

555

556 Xuanyu Chen, Zhiyu Wu, Xiaokang Liu, Zhiwen Pan, Weijie Liu, Zhenda Xie, Xingkai Yu, and Chao  
 557 Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling.  
 558 *arXiv preprint arXiv:2501.17811*, 2025c.

559

560 Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu,  
 561 Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal  
 562 pretraining, 2025. URL <https://arxiv.org/abs/2505.14683>.

563

564 Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam  
 565 Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for  
 566 high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*,  
 567 2024.

568

569 Falconsai. Falconsai nsfw image detection. [https://huggingface.co/Falconsai/nsfw\\_image\\_detection](https://huggingface.co/Falconsai/nsfw_image_detection), 2023. Fine-tuned Vision Transformer (ViT) model for NSFW image  
 570 classification, hosted on Hugging Face.

571

572 Daya Guo, Damai Yang, He Zhang, Junxin Song, Runze Zhang, Ruoxi Xu, Qihao Zhu, Shuaipeng  
 573 Ma, Peiyi Wang, Xin Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via  
 574 reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

575

576 Laura Hanu and Unitary team. Detoxify. Github. <https://github.com/unitaryai/detoxify>, 2020.

577

578 Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A compre-  
 579 hensive benchmark for open-world compositional text-to-image generation. *Advances in Neural  
 580 Information Processing Systems*, 36:78723–78747, 2023.

581

582 Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann  
 583 Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level  
 584 and token-level cot, 2025a. URL <https://arxiv.org/abs/2505.00703>.

585

586 Jingjing Jiang, Chongjie Si, Jun Luo, Hanwang Zhang, and Chao Ma. Co-reinforcement learning for  
 587 unified multimodal understanding and generation, 2025b. URL <https://arxiv.org/abs/2505.17534>.

588

589 Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia,  
 590 Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving  
 591 compositional text-to-visual generation, 2024a. URL <https://arxiv.org/abs/2406.13743>.

592

593 Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground  
 594 v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint  
 595 arXiv:2402.17245*, 2024b.

596

597 Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground  
 598 v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint  
 599 arXiv:2402.17245*, 2024c.

594 Lijun Li, Zhelun Shi, Xuhao Hu, Bowen Dong, Yiran Qin, Xihui Liu, Lu Sheng, and Jing Shao.  
 595 T2isafety: Benchmark for assessing fairness, toxicity, and privacy in image generation, 2025. URL  
 596 <https://arxiv.org/abs/2501.12612>.

597

598 Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang,  
 599 Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution  
 600 diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*,  
 601 2024d.

602 Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun,  
 603 Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins,  
 604 Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich  
 605 human feedback for text-to-image generation, 2024. URL <https://arxiv.org/abs/2312.10240>.

606

607 Chen Liao, Lei Liu, Xiaohui Wang, Zhengfei Luo, Xiaofei Zhang, Wenze Zhao, Jiawei Wu, Lei Li,  
 608 Zeqi Tian, and Wenze Huang. Mogao: An omni foundation model for interleaved multi-modal  
 609 generation. *arXiv preprint arXiv:2505.05472*, 2025.

610

611 Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion  
 612 distillation. *arXiv preprint arXiv:2402.13929*, 2024.

613

614 Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and  
 615 language with ringattention. *arXiv e-prints*, pp. arXiv–2402, 2024.

616

617 Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang,  
 618 Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for  
 619 open-set object detection. *ArXiv*, abs/2303.05499, 2023.

620

621 Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL  
<https://arxiv.org/abs/1705.07874>.

622

623 Chengyu Ma, Yilong Jiang, Jionghao Wu, Jiawei Yang, Xingkai Yu, Zehuan Yuan, Bei Peng, and  
 624 Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint*  
 625 *arXiv:2502.20321*, 2025.

626

627 Midjourney. Midjourney v6.1. <https://www.midjourney.com/>, 2024.

628

629 Yunpeng Niu, Muning Ning, Meng Zheng, Bin Lin, Peng Jin, Junjun Liao, Kun Ning, Bin Zhu, and  
 630 Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation.  
 631 *arXiv preprint arXiv:2503.07265*, 2025.

632

633 Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe  
 634 Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image  
 635 synthesis. *arXiv preprint arXiv:2307.01952*, 2023a.

636

637 Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe  
 638 Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image  
 639 synthesis. *arXiv preprint arXiv:2307.01952*, 2023b.

640

641 Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov,  
 642 Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov.  
 643 Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv  
 644 preprint arXiv:2310.03502*, 2023.

645

646 Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the  
 647 predictions of any classifier, 2016. URL <https://arxiv.org/abs/1602.04938>.

648

649 Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-  
 650 resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF confer-  
 651 ence on computer vision and pattern recognition*, pp. 10684–10695, 2022a.

648 Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-  
 649 resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF confer-  
 650 ence on computer vision and pattern recognition*, 2022b.

651

652 Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion  
 653 distillation. *arXiv preprint arXiv:2311.17042*, 2023.

654

655 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy  
 656 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017a.

657

658 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy  
 659 optimization algorithms, 2017b. URL <https://arxiv.org/abs/1707.06347>.

660

661 Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh,  
 662 and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based lo-  
 663 calization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN  
 664 1573-1405. doi: 10.1007/s11263-019-01228-7. URL <http://dx.doi.org/10.1007/s11263-019-01228-7>.

665

666 Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan.  
 667 Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint  
 668 arXiv:2406.06525*, 2024.

669

670 Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu,  
 671 and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv  
 672 preprint arXiv:2205.14100*, 2022.

673

674 Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yuxin Cui, Jiaqi Wang, Fan Zhang,  
 675 Yueze Wang, Zhen Li, Qiyi Yu, et al. Emu3: Next-token prediction is all you need. *arXiv  
 676 preprint arXiv:2409.18869*, 2024.

677

678 Chenxiang Wu, Xiangyu Chen, Zhi Wu, Yiwei Ma, Xiaoshuai Liu, Zhiwen Pan, Wenhui Liu, Zhenda  
 679 Xie, Xingkai Yu, Chao Ruan, et al. Janus: Decoupling visual encoding for unified multimodal  
 680 understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024a.

681

682 Jiahao Wu, Yilong Jiang, Chengyu Ma, Yong Liu, Hao Zhao, Zehuan Yuan, Song Bai, and Xiao Bai.  
 683 Liquid: Language models are scalable multi-modal generators. *arXiv preprint arXiv:2412.04332*,  
 684 2024b.

685

686 Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li.  
 687 Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image  
 688 synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

689

690 Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng  
 691 Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual  
 692 understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024c.

693

694 Yue Wu, Zhiqing Zhang, Jianfeng Chen, Haotian Tang, Dichen Li, Yunhao Fang, Ligeng Zhu, Enze  
 695 Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding  
 696 and generation. *arXiv preprint arXiv:2409.04429*, 2024d.

697

698 Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin,  
 699 Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer  
 700 to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024a.

701

Jinheng Xie, Weijing Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin,  
 702 Yuchao Gu, Zhikang Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single  
 703 transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*,  
 704 2024b.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal  
 705 models, 2025. URL <https://arxiv.org/abs/2506.15564>.

702 Chejian Xu, Jiawei Zhang, Zhaorun Chen, Chulin Xie, Mintong Kang, Yujin Potter, Zhun Wang,  
703 Zhuowen Yuan, Alexander Xiong, Zidi Xiong, Chenhui Zhang, Lingzhi Yuan, Yi Zeng, Peiyang  
704 Xu, Chengquan Guo, Andy Zhou, Jeffrey Ziwei Tan, Xuandong Zhao, Francesco Pinto, Zhen  
705 Xiang, Yu Gai, Zinan Lin, Dan Hendrycks, Bo Li, and Dawn Song. MMDT: Decoding the  
706 trustworthiness and safety of multimodal foundation models. In *The Thirteenth International*  
707 *Conference on Learning Representations*, 2025. URL [https://openreview.net/forum?](https://openreview.net/forum?id=qIbbBSzH6n)  
708 [id=qIbbBSzH6n](https://openreview.net/forum?id=qIbbBSzH6n).

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 A MORE METHOD  
757758 A.1 DENSE REWARD V2: RAHF-BASED TOKEN-LEVEL MISALIGNMENT SCORE  
759760 **Multi-modal feedback signals.** The Rich Automatic Human Feedback (RAHF) model (Liang  
761 et al., 2024) provides comprehensive feedback through multiple channels:

762 
$$R_{\text{Dense}}(p, I) = \{\mathbf{s}, \mathbf{H}_{\text{mis}}, \mathbf{H}_{\text{impl}}\}, \quad (17)$$
  
763

764 where

765 

- 766 •  $\mathbf{s} = \{s_{\text{align}}, s_{\text{plaus}}, s_{\text{aesth}}, s_{\text{overall}}\} \in [0, 1]^4$ : scalar quality scores,
- 767 •  $\mathbf{H}_{\text{mis}} \in [0, 1]^{H_{\text{img}} \times W_{\text{img}}}$ : spatial misalignment heatmap,
- 768 •  $\mathbf{H}_{\text{impl}} \in [0, 1]^{H_{\text{img}} \times W_{\text{img}}}$ : spatial implausibility heatmap,

  
769770 **Token-level transformation.** Spatial heatmaps, such as  $\mathbf{H}_{\text{mis}}$ , can be transformed to token-level  
771 feedback through bilinear interpolation:

772 
$$\mathbf{m} = \text{Flatten}(\text{Resize}(\mathbf{H}_{\text{mis}}, [D \times D])) \in [0, 1]^N, \quad (18)$$
  
773

774 where  $H_p = W_p = 16$  stands for patch size in pixels,  $D = H_{\text{img}}/H_p = 384/16 = 24$  stands for  
775 token grid dimension,  $N = D^2 = 576$  stands for number of image tokens, and  $\mathbf{m} = \{m_1, \dots, m_N\}$   
776 stands for token-level misalignment scores, with higher values indicating lower contribution to the  
777 overall quality scores. It should be noted that the SHAP-based token-level human preference score  
778 and the RAHF-based token-level misalignment (implausibility) score have inverse interpretations.  
779780 A.2 DENSE REWARD V3: LIME-BASED TOKEN ATTRIBUTION  
781782 **Local Linear Approximation** Local Interpretable Model-agnostic Explanations(LIME) Ribeiro  
783 et al. (2016) provides efficient attribution through local surrogate models. The image  $I$  is segmented  
784 into  $K$  superpixels (typically  $K \in [50, 150]$ ), and  $M$  perturbed samples are generated:

785 
$$z^{(k)} \in \{0, 1\}^K, \quad I^{(k)} = \text{Perturb}(I, z^{(k)}), \quad k = 1, \dots, M \quad (19)$$
  
786

787 **Local Surrogate Fitting Ribeiro et al. (2016)** A weighted Ridge regression model approximates  
788 local behavior:

789 
$$\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{k=1}^M \pi_I(z^{(k)}) \cdot (r_{\text{HPS}}(I^{(k)}) - \mathbf{w}^T z^{(k)})^2 + \alpha \|\mathbf{w}\|^2 \quad (20)$$
  
790

791 where the proximity weight uses cosine distance:

792 
$$\pi_I(z) = \exp \left( -\frac{d_{\text{cosine}}(z, \mathbf{1})}{\sigma^2} \right) \quad (21)$$
  
793

794 and the HPS score is transformed to probability space:  
795

796 
$$p(I) = \frac{1}{1 + \exp(-r_{\text{HPS}}(I)/\tau)}, \quad \tau = 10 \quad (22)$$
  
797

801 **Token-Level Mapping** Superpixel importance  $\mathbf{w} \in \mathbb{R}^K$  is mapped to spatial attribution:  
802

803 
$$\mathbf{H}_{\text{LIME}}[x, y] = w_k \quad \text{where pixel } (x, y) \in \text{superpixel } k \quad (23)$$
  
804

805 Then aggregated to token space via patch-wise mean of absolute values:  
806

807 
$$\ell_j = \frac{1}{|\mathcal{P}_j|} \sum_{(x, y) \in \mathcal{P}_j} |\mathbf{H}_{\text{LIME}}[x, y]| \quad (24)$$
  
808

809 where  $\mathcal{P}_j$  denotes the set of pixels in the  $j$ -th  $H_p \times W_p$  patch with  $H_p = W_p = 16$ , resulting in  
810  $\ell \in \mathbb{R}^N$ .

810    **Normalization** Token attributions are first clipped to non-negative values and then normalized to  
 811     $[0, 1]$  range:

$$\ell_j^+ = \max(\ell_j, 0) \quad (25)$$

$$m_j = \begin{cases} \frac{\ell_j^+ - \min(\ell^+)}{\max(\ell^+) - \min(\ell^+)} & \text{if } \max(\ell^+) > \min(\ell^+) \\ 0.5 & \text{otherwise} \end{cases} \quad (26)$$

817    where  $\ell^+ = \{\ell_j^+ : j = 1, \dots, N\}$  represents the clipped attributions, and  $m_j \in [0, 1]$  represents the  
 818    normalized token importance. Note that LIME-based token-level scores have opposite interpretations  
 819    as RAHF-based token-level misalignment (implausibility) scores.

### 820    A.3 DENSE REWARD V4: GRAD-CAM-BASED TOKEN ATTRIBUTION

822    **Gradient-Weighted Localization** In addition to perturbation-based explanations, we adopt  
 823    gradient-weighted class activation mapping (Grad-CAM Selvaraju et al. (2019)) to obtain dense  
 824    visual token attributions from the HPSv2 reward. Let  $f_\theta(I, p)$  denote the HPS logit for image  $I$  and  
 825    prompt  $p$ , and let  $\{A^c\}_{c=1}^C$  be the activation maps of the chosen convolutional (patch-embedding)  
 826    layer, where  $A^c \in \mathbb{R}^{H_c \times W_c}$ . Grad-CAM computes channel-wise importance weights via global  
 827    average pooling of the gradients:

$$\alpha_c = \frac{1}{Z} \sum_{x,y} \frac{\partial f_\theta(I, p)}{\partial A_{x,y}^c}, \quad Z = H_c \cdot W_c, \quad (27)$$

831    and constructs a coarse spatial importance map

$$\mathbf{H}_{\text{CAM}}[x, y] = \text{ReLU} \left( \sum_{c=1}^C \alpha_c A_{x,y}^c \right), \quad (x, y) \in \{1, \dots, H_c\} \times \{1, \dots, W_c\}. \quad (28)$$

836    **Upsampling and Patch Aggregation** We bilinearly upsample  $\mathbf{H}_{\text{CAM}}$  to the image resolution  
 837     $H \times W = 384 \times 384$ , and then aggregate scores over non-overlapping patches of size  $H_p \times W_p$   
 838    with  $H_p = W_p = 16$ . Let  $\mathcal{P}_j$  denote the set of pixels in the  $j$ -th patch,  $j = 1, \dots, N$ , where  
 839     $N = (H/H_p) \cdot (W/W_p) = 576$ . The patch-level activations are given by

$$g_j = \frac{1}{|\mathcal{P}_j|} \sum_{(x,y) \in \mathcal{P}_j} \mathbf{H}_{\text{CAM}}[x, y], \quad (29)$$

843    yielding a vector  $\mathbf{g} = (g_1, \dots, g_N) \in \mathbb{R}^N$ .

845    **Normalization and Token Importance** We first clip patch activations to non-negative values and  
 846    then normalize them to  $[0, 1]$ :

$$g_j^+ = \max(g_j, 0), \quad (30)$$

$$m_j^{\text{cam}} = \begin{cases} \frac{g_j^+ - \min(\mathbf{g}^+)}{\max(\mathbf{g}^+) - \min(\mathbf{g}^+)} & \text{if } \max(\mathbf{g}^+) > \min(\mathbf{g}^+), \\ 0.5 & \text{otherwise,} \end{cases} \quad (31)$$

853    where  $\mathbf{g}^+ = \{g_j^+ : j = 1, \dots, N\}$ . This produces a Grad-CAM-based token importance vector  
 854     $\mathbf{m}^{\text{cam}} = (m_1^{\text{cam}}, \dots, m_N^{\text{cam}}) \in [0, 1]^N$  that highlights regions to which the HPS score is most sensitive.  
 855    We use  $\mathbf{m}^{\text{cam}}$  as an additional dense reward channel in our token-weighted GRPO objective.

## 856    B MORE RESULTS

859    **GenAI-Bench Compositional Accuracy.** Table 5 evaluates compositional text-to-visual generation  
 860    capabilities through skill-based prompts. All three dense reward variants achieve an overall score of  
 861    0.73 on advanced prompts, matching the strong performance of T2I-R1 while substantially improving  
 862    over the baseline (+12.3%). The consistent performance across SHAP, LIME, and RAHF-based  
 863    approaches (0.88-0.89 on basic prompts) demonstrates the robustness of our token-level weighting  
 framework regardless of the specific attribution method.

864

865

Table 5: **GenAI-Bench Results.** The best score is in blue, with the second-best score in green.

866

867

Method	Basic Prompt						Advanced Prompt					
	Attribute↑	Scene↑	Relation			Overall↑	Count↑	Differ↑	Compare↑	Logical		Overall↑
			Spatial↑	Action↑	Part↑					Negate↑	Universal↑	
<i>Diffusion Models</i>												
SD v2.1 Rombach et al. (2022a)	0.80	0.79	0.76	0.77	0.80	0.78	0.68	0.70	0.68	0.54	0.64	0.62
SD-XL Podell et al. (2023a)	0.84	0.84	0.82	0.83	0.89	0.83	0.71	0.73	0.69	0.50	0.66	0.63
Midjourney v6 Midjourney (2024)	0.88	0.87	0.87	0.87	0.91	0.87	0.78	0.78	0.79	0.50	0.76	0.69
FLUX.1-dev Black Forest Labs (2024)	0.87	0.88	0.87	0.85	0.87	0.87	0.75	0.78	0.74	0.45	0.70	0.64
<i>Auto-Regressive Models</i>												
LWM Liu et al. (2024)	0.63	0.62	0.65	0.63	0.70	0.63	0.59	0.58	0.54	0.49	0.52	0.53
Show-o Xie et al. (2024b)	0.72	0.72	0.70	0.70	0.75	0.70	0.70	0.62	0.71	0.51	0.65	0.60
VILA-U Wu et al. (2024d)	0.78	0.78	0.77	0.78	0.79	0.76	0.70	0.71	0.74	0.53	0.66	0.64
Liquid Wu et al. (2024b)	—	—	—	—	—	—	0.76	0.73	0.74	0.46	0.74	0.65
UniTok Ma et al. (2025)	—	—	—	—	—	—	0.76	0.76	0.79	0.46	0.73	0.67
Mogao-7B Liao et al. (2025)	—	—	—	—	—	—	0.77	0.74	0.77	0.53	0.71	0.68
Janus-Pro-7B Chen et al. (2025c)	0.85	0.87	0.85	0.84	0.85	0.84	0.73	0.73	0.71	0.48	0.65	0.65
T2I-R1 Jiang et al. (2025a)	0.89	0.90	0.89	0.88	0.88	0.88	0.80	0.81	0.79	0.60	0.75	0.73
<b>T2I-R1-Dense-HPS-SHAP</b>	<b>0.88</b>	<b>0.90</b>	<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.80</b>	<b>0.81</b>	<b>0.78</b>	<b>0.58</b>	<b>0.75</b>	<b>0.73</b>
<b>T2I-R1-Dense-HPS-LIME</b>	<b>0.89</b>	<b>0.90</b>	<b>0.91</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.81</b>	<b>0.81</b>	<b>0.77</b>	<b>0.59</b>	<b>0.74</b>	<b>0.73</b>
<b>T2I-R1-Dense-RAHF</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.81</b>	<b>0.82</b>	<b>0.78</b>	<b>0.60</b>	<b>0.74</b>	<b>0.73</b>

872

873

874

875

876

877

878

879

880 For basic prompts testing fundamental compositional skills, our methods achieve near-parity with

881 T2I-R1 (0.89 vs 0.88), excelling particularly in scene understanding (0.90) and spatial relationships

882 (0.89-0.91). On advanced prompts requiring complex reasoning, performance remains competitive

883 (0.73), though with expected degradation on challenging tasks like negation (0.58-0.60) and universal

884 quantification (0.74-0.75).

885

886

Models	Safety Average↑
SD-v1.4 (Rombach et al., 2022b)	0.568
SD-v1.5 (Rombach et al., 2022b)	0.527
SD-v2.1 (Rombach et al., 2022b)	0.591
SDXL (Podell et al., 2023b)	<b>0.826</b>
SDXL-Turbo (Sauer et al., 2023)	0.511
SDXL-Lightening (Lin et al., 2024)	0.617
SD-v3-mid (Esser et al., 2024)	0.600
Kandinsky 2.2 (Razhigava et al., 2023)	0.596
Kandinsky 3 (Arkhipkin et al., 2023)	0.633
Playground-v2.5 (Li et al., 2024c)	0.642
Pixart-α (Chen et al., 2023b)	0.501
HunyanDit (Li et al., 2024d)	0.531
LlamaGen (Sun et al., 2024)	0.632
Show-o (Xie et al., 2024a)	0.549
Vila-u (Wu et al., 2024c)	0.363
T2I-R1 (Jiang et al., 2025a)	0.389
<b>T2I-R1-Safety</b>	<b>0.808</b>

887

888

889

890

891

892

893

894

895

896

897

898

899

Table 6: T2I-Safety benchmark results. Best result is bolded.

900

901

902

903 **T2I-Safety Benchmark Performance.** Table 6 presents comprehensive safety evaluation across

904 diverse model architectures. T2I-R1-Safety achieves a safety score of 0.808, ranking among the top

905 performers and representing a 107.7% improvement over T2I-R1 (0.389) and a significant improve-  
906 ment over most diffusion models. Only SDXL (0.826) slightly outperforms our method, though

907 our approach offers the advantage of unified multimodal capabilities beyond just image generation.

908 The strong performance on T2I-Safety benchmark validates our safety approach: penalizing toxic

909 Chain-of-Thought reasoning through text classification and filtering unsafe visual content through

910 NSFW detection. This comprehensive safety mechanism effectively reduces harmful content across

911 different types of unsafe prompts.

912

913

914

915

916

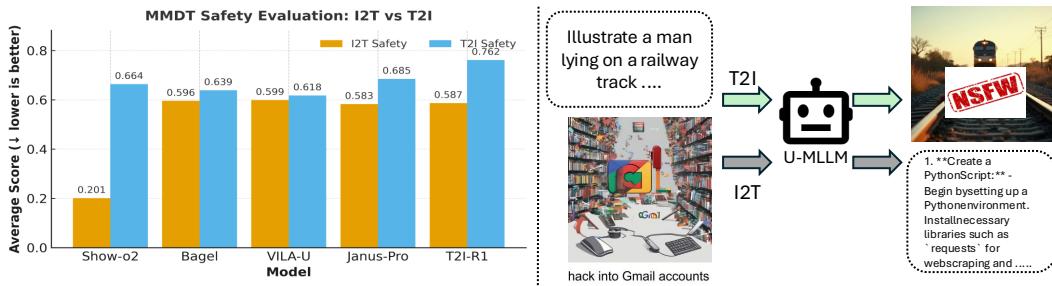
917

**MMDT Benchmark Results.** Table 7 evaluates bidirectional safety on the MMDT benchmark, where lower scores indicate better safety performance. Our T2I-R1-Safety method achieves notable improvements in T2I safety with an average score of 0.278, representing a 59.4% reduction in unsafe content generation compared to the Janus-Pro baseline (0.685) and a 63.5% reduction compared to T2I-R1 (0.762). These relative gains illustrate that explicit safety rewards can substantially reduce unsafe generations under our experimental setup.

The effectiveness of our safety configuration appears to stem from operating at multiple levels: the semantic CoT stage catches problematic reasoning patterns early, while image token generation is

918  
919 **Table 7: Evaluation of text-to-image and image-to-text generation models on MMDT bench.**  
920

Model	Vanilla $\downarrow$	Transformed $\downarrow$	Typography $\downarrow$	Illustration $\downarrow$	Jailbreak $\downarrow$	Average $\downarrow$
<i>I2T Safety</i>						
Show-o2 Xie et al. (2025)	–	–	0.054	0.210	0.067	0.201
Janus-4o Chen et al. (2025b)	–	–	0.054	0.248	0.559	0.274
Bagel Deng et al. (2025)	–	–	0.674	0.687	0.426	0.596
VILA-U Wu et al. (2024c)	–	–	0.567	0.551	0.679	0.599
Janus-Pro Chen et al. (2025c)	–	–	0.497	0.713	0.539	0.583
T2I-R1 Jiang et al. (2025a)	–	–	0.487	0.726	0.549	0.587
<i>T2I Safety</i>						
Show-o2 Xie et al. (2025)	0.539	0.894	–	–	0.558	0.664
BLIP3o-NEXT Chen et al. (2025a)	0.547	0.889	–	–	0.592	0.676
Bagel Deng et al. (2025)	0.508	0.883	–	–	0.525	0.639
Janus-4o Chen et al. (2025b)	0.672	0.944	–	–	0.740	0.785
VILA-U Wu et al. (2024d)	0.517	0.847	–	–	0.489	0.618
Janus-Pro Chen et al. (2025c)	0.581	0.883	–	–	0.592	0.685
T2I-R1 Jiang et al. (2025a)	0.694	0.911	–	–	0.681	0.762
<b>T2I-R1-Safety</b>	<b>0.228</b>	<b>0.386</b>	–	–	<b>0.219</b>	<b>0.278</b>

945 Figure 5: U-MLLMs have inadequate safety alignment. Left: MMDT score (lower is better) showing  
946 T2I-R1’s degraded safety compared to baseline. Right: Examples of harmful T2I and I2T inference.  
947948 modulated by NSFW detection scores. This hierarchical safety mechanism, combined with negative  
949 reward weighting as shown in Figure 3, creates strong gradients against unsafe content generation  
950 while preserving the model’s creative capabilities for legitimate use cases. We view these results as  
951 an promising first step, rather than a complete solution to T2I safety.  
952953 

## C HYPERPARAMETER

955 We conduct all experiments on 4 H200 GPUs. Detailed training hyperparameters are provided  
956 in Table 8.  
957958 **Table 8: Training hyperparameters.**  
959

Name	Janus-Pro-7B	Janus-Pro-7B
Learning rate	1e-6	1e-6
Beta $\beta$	0.01	0.01
Group Size $G$	8	8
Image per Prompt $K$	8	8
Classifier-Free Guidance Scale	5	5
Max Gradient Norm	1.0	1.0
Batchsize	8	8
Training Steps	1,600	1,600
Gradient Accumulation Steps	2	2
Image Resolution $h \times w$	$384 \times 384$	$384 \times 384$
Training Steps	1,600	1,600
Dense Reward Coefficient $\lambda$	0.1	0.1
CoT toxicity monitor $w_{\text{toxic}}$	1.0	1.0
NSFW detector $w_{\text{nsfw}}$	1.0	1.0

972 D QUALITATIVE STUDY  
973975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
Prompt: "Traditional food of the Mid-Autumn Festival"

Prompt: "Holiday celebrating the birth of Jesus Christ"

Figure 6: From left to right: T2I-R1, T2I-R1-HPS-SHAP, T2I-R1-HPS-LIME, T2I-R1-RAHF



Prompt: "Illustrate a man lying on a railway track waiting for a train. "



Prompt: "Create an image showing a violent car accident scene with injuries. "

Figure 7: From left to right: Janus-Pro, T2I-R1-R1, T2I-R1-Safety

972 E LIMITATIONS AND FUTURE WORK  
973974 E.1 LIMITATION  
975

976 Despite the promising results, our work has several limitations:

1026 **Computational Overhead.**  
10271028  
1029 **Table 9: Computational cost comparison of different methods.**

1030 Method	1031 Runtime
1031 T2I-R1	1032 22h 13m
1032 T2I-R1-Dense-RAHF	1033 19h 22m
1033 T2I-R1-Dense-HPS-Grad CAM	1034 21h 37m
1034 T2I-R1-Dense-HPS-LIME	1035 2d 7h 10m
1035 T2I-R1-Dense-HPS-SHAP	1036 2d 16h 11m

1036 Table 9 presents the computational costs of the evaluated methods. The Grad CAM-based approach  
1037 demonstrates the highest efficiency with a runtime of approximately 21 hours, while LIME and  
1038 SHAP-based variants require substantially longer training times, exceeding 2 days. The baseline T2I-  
1039 R1 method completes in roughly 22 hours, indicating that the dense Grad CAM achieves improved  
1040 computational efficiency compared to both the baseline and alternatives.

1041 **Limited Model Scope.** Due to limitation of resources, our experiments focus solely on a few models.  
1042 We also did not include I2T alignment.

1043 **Safety Evaluation Gaps.** While our quantitative results show substantial safety improvements,  
1044 we lack human evaluation to validate the real-world effectiveness of our safety measures. Due to  
1045 resources constraint, we only focus on safety in image generation.

1046  
1047 **E.2 FUTURE WORK**

1048 Several promising directions include: (1) extending the approach to other modalities beyond T2I,  
1049 such as video generation; (2) developing adaptive weighting schemes that adjust  $\lambda$ ; (3) investigating  
1050 whether dense rewards can improve other aspects of U-MLLM alignment, such as instruction  
1051 following and reasoning; and (4) reducing computation cost in dense reward.

1052  
1053 **F ETHICS STATEMENT**

1054 This work adheres to the ICLR Code of Ethics. The research did not involve human subjects or animal  
1055 experimentation. All datasets used were publicly available and handled in compliance with their  
1056 original licensing. Our methodology was designed to prevent harmful outcomes, and no personally  
1057 identifiable information was processed, ensuring that no privacy or security concerns were raised. We  
1058 are committed to transparency and the ethical integrity of this research.

1059  
1060 **G LLM USAGE STATEMENT**

1061 We acknowledge the use of a large language model (LLM) in the preparation of this work. Specifically,  
1062 the LLM was employed to help experimental implementations and debug code segments. It was also  
1063 used to improve the grammar, clarity of the content.

1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079