RealGen: Photorealistic Text-to-Image Generation via Detector-Guided Rewards

Junyan Ye[1,2,3,*], Leqi Zhu[1,4,*], Yuncheng Guo[1], Dongzhi Jiang[5], Zilong Huang[3], Yifan Zhang[2], Zhiyuan Yan[6], Haohuan Fu[2], Conghui He[1], and Weijia Li[2,1†]

[1] Shanghai AI Lab, China; [2] Shenzhen International Graduate School, Tsinghua University, China; [3] Sun Yat-Sen University, China; [4] Nanjing University, China; [5] CUHK MMLab, China; [6] Peking University, China
[*] Equal contribution. [†] Corresponding author.

Fig. 1: Images generated by our proposed RealGen. RealGen achieves superior photorealism and enhanced details, outperforming both powerful T2I models, such as Qwen-Image, and specialized photorealistic models, like FLUX-Krea.

Abstract. Rapid advancements in image generation have led models like GPT-Image and Qwen-Image to excel in text-to-image consistency and world knowledge. However, they still struggle with photorealism, often producing "fake" images laden with obvious AI artifacts, such as overly smooth skin or unnatural oily sheens. To recapture the original goal of "indistinguishable-from-reality" generation, we propose RealGen, a photorealistic text-to-image framework. RealGen integrates an LLM component for prompt optimization and a diffusion model for realistic image generation. Inspired by adversarial generation, RealGen introduces a "Detector Reward" mechanism, which quantifies artifacts and assesses realism using both semantic and feature-level synthetic image detectors. We leverage this reward with the GRPO algorithm to optimize the entire generation pipeline, significantly enhancing image realism and detail. Furthermore, we propose RealBench, an automated evaluation benchmark employing Detector-Scoring and Arena-Scoring. It enables human-free photorealism assessment, yielding results that are more accurate and aligned with real user experience. Experiments show that RealGen produces photorealistic images with enhanced realism, detail, and aesthetics.

Keywords: Image Generation, Reinforcement learning

1 Introduction

Image generation has undergone a significant evolution from GANs [11] to Diffusion models [7, 30], leading to a generation of powerful models such as GPT-Image-1 [26], Nano-Banana [8], and Qwen-Image [41]. These models demonstrate exceptional capabilities in areas like precise attribute control and complex text rendering. However, recent research has focused on enhancing prompt fidelity and leveraging world knowledge for complex content generation. Despite these advancements, even existing powerful generative models like FLUX.1 Pro [18] and Qwen-Image [41] still tend to produce images that lack photorealism, such as "overly smooth skin" and "oily facial sheens", as shown in Fig. 2. This arguably deviates from the original goal of image generation: ***"to produce visuals that are indistinguishable from reality,"*** which is precisely the focus of our work.

To pursue higher photorealism, existing research has attempted to combine generative models with reinforcement learning (RL) and human preference scores [31, 35]. For instance, methods like DanceGRPO [45] and FlowGRPO [23] utilize reward models such as PickScore [17] and HPSv2 [42] to align generation with human preferences. However, trained on human preference data may introduce prior biases, such as an excessive preference for colors like red or purple. More critically, human preference scores do not directly equate to photorealism. A candid snapshot might be highly realistic but receive a low score due to a lack of aesthetic appeal, inadvertently encouraging the model to favor artistic or anime styles. Meanwhile, approaches like FLUX-Krea [20] employ large-scale, human-curated, high-quality image samples combined with TPO to enhance realism. However, this method is not only cost-prohibitive but also highly dependent on the subjective preferences of annotators. Therefore, the critical bottleneck in leveraging RL for photorealism lies in *how to establish an objective, scalable, and human-free metric for image realism.*

Inspired by adversarial generation, a promising approach is to quantify photorealism using synthetic image detectors. The core rationale is that ***a more realistic image should possess fewer AI artifacts and thus be more difficult for a detector to classify as "Fake."*** In recent years, as generative models have rapidly advanced, corresponding image detectors have evolved in parallel, achieving robust detection performance. Furthermore, some MLLM-based methods have extended this capability to semantic-level, interpretable detection, such as analyzing "AI-feel" skin textures. Therefore, we posit that two distinct classes of detectors can be employed to measure image realism: one for analyzing visible, semantic-level artifacts, and another for deep feature-level artifacts. As shown in Figure 2, during the RL process, we define these detectors as the reward model to guide the generator to "escape" detection, which effectively reduces artifacts and enhances image realism.

In addition, the quality of image synthesis is highly correlated with the complexity of the text prompt [2]. Expert users, through sophisticated "prompt engineering," can produce photorealistic, cinematic-quality images. Conversely, simple prompts from casual users often yield low-quality results that lack realism. This discrepancy arises because T2I models often rely on complex prompt

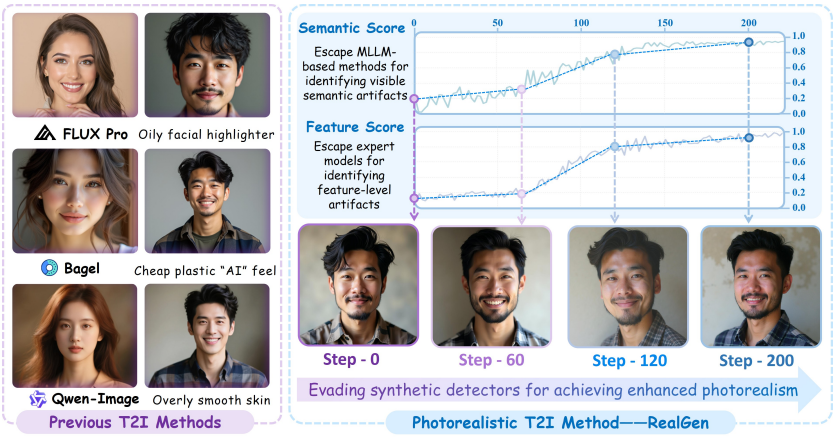

Fig. 2: From Synthetic Artifacts to Photorealism. Contrasting the common "fake-feel" AI artifacts in previous T2I methods , our proposed RealGen achieves enhanced photorealism by progressively evading semantic and feature-level detectors.

structures; simple prompts provide low information entropy, causing the model to default to its high-frequency priors, which results in images lacking detail and specificity [57]. Therefore, automatically rewriting or enriching the user's input directive presents another critical pathway to enhancing the quality and realism of the subsequently generated images [36].

In summary, we propose ***RealGen***, a framework for photorealistic text-to-image generation. The framework includes a Large Language Model (LLM) component for optimizing user prompts and a diffusion model dedicated to generating realistic images. We utilize detector models as rewards, optimizing both the LLM and the generative model via Generalized Reinforcement Policy Optimization (GRPO) algorithm [23, 45] to effectively reduce artifacts in the generated images. Specifically: (1) With the generative model held fixed, we train the LLM to optimize the input prompts, using the detector model to score the final synthesized image; this encourages the LLM to generate richer and more effective prompts. (2) While maintaining text input consistency, we optimize the diffusion model itself, enabling it to "escape" the detectors and generate more realistic and detailed images. As illustrated in Fig. 1, RealGen produces images with more realistic textures and richer visual details.

Furthermore, we also propose ***RealBench***, a new benchmark for evaluating the photorealism of generated images. RealBench contains a diverse set of text prompts, supplemented by high-quality, real-world image references. Addressing the lack of effective evaluation methods beyond manual human assessment, RealBench employs two automated evaluation protocols: Detector-Scoring and Arena-Scoring. First, Detector-Scoring utilizes multiple synthetic image detectors (including those held out from the reward) to score the images; images with fewer AI artifacts receive higher scores. Second, for Arena-Scoring, we adapt

the LMArena methodology, using several different MLLMs to simulate user preferences in pairwise comparisons, selecting the result that appears more realistic. Outputs from each model are put into pairwise "battles" against other models or real-world images for at least 3000 random pairings to determine a final win rate. The inclusion of real-world data not only enhances scoring stability but also validates the effectiveness of this adversarial scoring approach. Our main contributions are as follows:

- We propose ***RealGen***, a text-to-image generator capable of producing highly convincing photorealistic images. It leverages a Detector Reward-guided GRPO post-training to escape detector identification, thereby reducing artifacts and enhancing image realism and detail.
- We introduce ***RealBench***, a new benchmark for evaluating photorealism that achieves human-free automated scoring through Detector-Scoring and Arena-Scoring.
- RealGen significantly outperforms both general image models (like GPT-Image-1, Qwen-Image) and specialized realistic models (like FLUX-Krea) in realism, details, and aesthetics on the T2I task.

2 Related Work

2.1 Image Generation Models

In recent years, diffusion models have achieved remarkable success in generative modeling. Representative models, such as the Stable Diffusion (SD) series [7,28,30], FLUX [18, 19], Emu [5, 32, 39], and DALL·E [29], have demonstrated robust text-to-image generation capabilities, laying the foundation for controllable and scalable generation. To enhance controllability, models like ControlNet [56] and T2I-Adapter [25] introduce external conditioning modules, while StyleShot [9] and InstructPix2Pix [1] focus on fine-grained editing and instruction-driven generation. These advancements highlight the increasing flexibility of diffusion models under various conditioning settings [49, 50]. As research gradually shifts toward multimodal generative models, these models achieve understanding and generation through unified architectures [3, 27, 43]. Although powerful models like GPT-Image-1 [26], Bagel [6] and, and Show-o [43, 44] can generate precise objects and complex text, challenges remain in generating realistic images, especially with the strong oily appearance of human faces. RealGen focuses on enhancing realistic image generation, optimizing image realism and detail.

2.2 Reinforcement Learning and Human Preferences in Image Generation

Recently, with the development of reinforcement learning, more methods have begun to explore its application in the Diffusion image generation domain [45, 46]. For example, DiffusionDPO [35], DiffusionNFT [58] and FlowGRPO [23] use human preference data as a reward function to generate images that align

with human aesthetics. However, reward models based on human preferences can introduce prior biases, such as HPSv2 [42] favoring red-toned images and PickScore [17] preferring purple ones. While human-captured images are generally realistic, they do not always score highly in terms of preference. SRPO [31] has made some progress in enhancing image realism but still lacks in aesthetic quality. FLUX-Krea [20], through large-scale manual collection of high-quality image samples and TPO reinforcement learning, optimizes the realism of generated images. However, this method is limited by the costs of data labeling and the personal preferences of annotators. In contrast, RealGen does not rely on human preference data. Instead, it uses advanced AIGC detection models to guide the generative model away from AI artifacts at both semantic and feature levels, further improving image realism.

2.3 Synthesis detection for Image generation

As the realism of image generation continuously increases, corresponding detection techniques have rapidly evolved. From early detectors like CNNSpot [38] advanced methods leveraging powerful pre-trained vision models (e.g., OmniAID [13], Effort [47]), these detectors have demonstrated success rates exceeding 80% [52, 59]. Recently, detector based on MLLMs have emerged, such as Fakevlm [40], Forensic-chat [22] and LEGION [16], which not only achieve good detection performance but also enhance the explainability of artifact detection. Since synthetic image detection and generation are a "cat-and-mouse" game, some studies have explored using detector feedback to optimize generation quality. However, these methods are currently more limited to post-processing techniques, such as Inpainting, to optimize already generated images [54, 55]. In contrast, our work combines reinforcement learning and synthetic detectors to directly optimize the generative model itself.

3 RealGen

3.1 Model architecture

As illustrated in Figure 3, the architecture of RealGen comprises two core components: a LLM for understanding and refining user intent, and a diffusion model for realistic image synthesis. First, the LLM receives the initial user instruction and performs "thought and planning": it expands the short prompt into a longer, more diverse text description by adding rich details. Subsequently, the diffusion model utilizes this refined text as its input condition, executing the denoising and decoding process to generate the final image.

For implementation, we employ Qwen-3 4B [48] as the base LLM. For image generation, we utilize the advanced pre-trained diffusion model, FLUX.1-dev dev [18], integrated with fine-tuned LoRA layers. Both the LLM and the diffusion model first undergo specialized Supervised Fine-Tuning (SFT) as a cold-start phase, followed by Reinforcement Learning optimization via GRPO. The subsequent sections will detail our designed detection reward function and the RL training process.

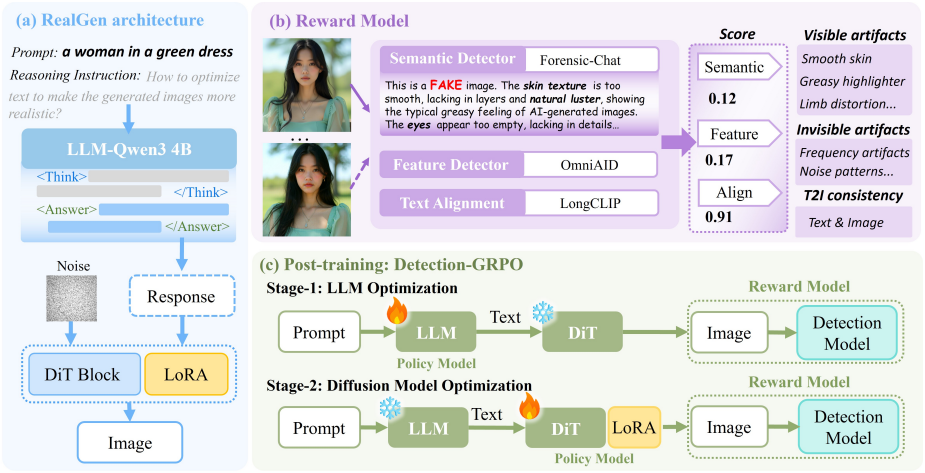

Fig. 3: Overview of the RealGen Method. (a) The architecture of RealGen, consisting of an LLM component and a Diffusion component. (b) Our detector-based reward model, which evaluates images based on visible artifacts, feature-level artifacts, and text-image alignment. (c) The two-stage post-training process guided by this reward model, which respectively optimizes the LLM and Diffusion components.

3.2 Detection Reward

To steer the model optimization towards high-fidelity realism, the design of the reward function is critical, as it must accurately quantify authenticity. We adopt a "detection-as-reward" paradigm inspired by adversarial generation, designing a multi-objective reward function. As shown in Fig. 3(b), this function combines detectors at two distinct levels—semantic and feature—to penalize both perceptible artifacts and imperceptible synthesis traces, respectively.

Semantic Detector: We employ Forensic-Chat [22], a generalizable and interpretable detector optimized from Qwen2.5-VL-7B [37]. It assesses authenticity by analyzing image content (e.g., smooth greasy skin, artifacts in faces/hands, unnatural background blur). We define the semantic reward R_{semantic} as the normalized probability of its output "real" token probability:

$$R_{\text{semantic}} = \text{softmax}\left(\left[L_{(\text{"fake"},\text{"Fake"})}, L_{(\text{"real"},\text{"Real"})}\right]\right)_1 \tag{1}$$

Feature Detector: We utilize the advanced expert detector OmniAID [13], which achieves stable and accurate detection by being pre-trained on large-scale real and synthetic datasets. Feature-level artifacts are primarily associated with frequency artifacts and abnormal noise patterns. We define the feature-level reward (R_{feature}) as one minus its output "fake" probability, i.e., $1 - P_{\text{OmniAID}}(\text{fake})$.

Text Alignment: Finally, we include the Long-CLIP [53] score as an auxiliary reward (R_{align}) to maintain text-image alignment. This prevents the model from sacrificing fidelity to the input prompt in its pursuit of realism.

For the GRPO process, we combine the three reward functions defined above: $\{R_{\text{semantic}}, R_{\text{feature}}, R_{\text{align}}\}$. For a batch of N samples $\{I_i\}_{i=1}^{N}$, we first evaluate each sample I_i to obtain its raw scores $\{r_i^{\text{sem}}, r_i^{\text{feat}}, r_i^{\text{align}}\}$. Since each reward function operates on a different scale and distribution, we first normalize the scores for each reward dimension within the batch. Then, we sum these normalized scores to fuse them into a single advantage function $A(I_i)$, as Equation (2):

$$A(I_i) = \sum_{k \in \{\text{sem, feat, align}\}} \frac{r_i^k - \text{Mean}(\{r_j^k\}_{j=1}^{N})}{\text{Std}(\{r_j^k\}_{j=1}^{N})} \tag{2}$$

3.3 Post-training: Detection-GRPO

LLM Optimization. In the first stage, we optimize the LLM for user intent refinement, while the parameters of the image generation model remain frozen. In this setup, the LLM π_θ acts as the policy network. We first conduct SFT as a cold-start to teach it the "think-plan-then-generate" pattern required by the system prompt. During the RL phase, for a given text input x, the policy LLM samples and rewrites it, generating N optimized prompts $\{y_i\}_{i=1}^{N}$. Subsequently, the frozen diffusion model generates an image I conditioned on y. This image I is then evaluated by our multi-objective reward function to obtain the reward $R(I, y)$. In this context, each trajectory $\{o_i\}_{i=1}^{N}$ corresponds to the text sequence $y_i = (y_{1,i}, \ldots, y_{T,i})$ sampled by the LLM. We update the LLM's parameters by maximizing the GRPO objective function defined in Equation (3).

$$\mathcal{J}(\theta) = \mathbf{E}_{\substack{c\sim\mathcal{C}, \\ \{o_i\}_{i=1}^{N}\sim\pi_{\theta_{\text{old}}}}} \left[\frac{1}{NT} \sum_{i,t} \left(\mathcal{J}_{GRPO} - \beta \mathbf{D}_{\text{KL}}(\pi_\theta \| \pi_{\theta_{\text{ref}}}) \right) \right], \tag{3}$$
$$\text{where} \quad \mathcal{J}_{GRPO} = \min\left(\rho_{t,i} A_i, \text{clip}(\rho_{t,i}, 1-\varepsilon, 1+\varepsilon) A_i\right)$$

where A_i is the fused advantage function obtained from Equation (2) and $\rho_{t,i}$ denotes the importance sampling ratio between policies, which is defined in Equation (4).

$$\rho_{t,i} = \frac{\pi_\theta(y_{t,i}|x, y_{<t,i})}{\pi_{\theta_{\text{old}}}(y_{t,i}|x, y_{<t,i})} \tag{4}$$

The optimized LLM learns to explore and generate superior prompts, for instance: (1) adding rich scenic details, (2) naturally incorporating "flaws and imperfections," and (3) appending specific auxiliary words (e.g., "shot on iPhone") that aid the diffusion model in producing realistic images.

Diffusion Model Optimization. In the second stage, we optimize the diffusion model for image generation, while keeping the parameters of the LLM frozen. We employed a diffusion model RL framework Flow-GRPO [23], however, due to the stochastic nature of RL, images generated through limited denoising steps or full-trajectory exploration tend to be noisy and blurry [31]. This unstable process subsequently misleads the judgment of detection models, which are typically

trained on datasets lacking exposure to such noisy or ambiguous artifacts. To address this training-inference inconsistency, for a given input x, the policy model p_θ executes a **complete denoising inference** $(x_T, x_{T-1}, \ldots, x_0)$ for evaluation. Simultaneously, several consecutive steps Δt from this process is designated for random exploration and obtain N trajectories $(x_{T',i}, x_{T'-1,i}, \ldots, x_{T'-\Delta t,i})$, which effectively balance RL explorativeness with training efficiency. We optimize the policy model through GRPO process the same as stage-1, as defined in Equation (5), and the probability ratio $\rho_{t,i}$ is given by:

$$\rho_{t,i} = \frac{p_\theta(x_{t-1,i}|x_{t,i}, c)}{p_{\theta_{\text{old}}}(x_{t-1,i}|x_{t,i}, c)} \tag{5}$$

SDE sampling provides stochasticity to the reverse process. During GRPO training, a portion of short texts are processed by the text-rewrite LLM, while the rest bypass this component. This strategy enhances the model's generalization robustness across prompts of varying lengths. Finally, detector reward optimization enables diffusion models to significantly reduce semantic and feature-level artifacts, consequently generating more realistic images.

4 RealBench

4.1 Dataset overview

We introduce RealBench, a benchmark platform designed to comprehensively evaluate the photorealism of T2I synthesized images. As illustrated in Fig. 4, RealBench features a meticulously curated dataset of 1000 high-quality, real-world images and their corresponding captions, sourced from the internet and free photography websites [1]. This dataset spans seven distinct categories. Recognizing that "Portrait" is one of the most common and challenging categories in user T2I prompts, we have significantly increased its proportion while ensuring diversity across other categories. Unlike existing benchmarks focusing on instruction following (e.g., GenEval [10], DPG-Bench [14]) or human preference (e.g., HPD V2 [42]), RealBench focuses exclusively on assessing the photorealism of generated results. RealBench comprises two key evaluation protocols: Protocol-1, detector-based realism quantification (Detector-Scoring); and Protocol-2, arena-style preference evaluation (Arena-Scoring).

4.2 Detector-Based Realism Quantification

The core motivation for this protocol is that more photorealistic images should better evade advanced synthetic detectors, thus lowering their probability of being classified as "fake." Therefore, we utilize the probability of an image being deemed "real" as its photorealism score. In our assessment, we deploy detectors at both the

[1] https://pixabay.com/ (accessed June 28, 2026); https://www.pexels.com/ (accessed June 28, 2026).

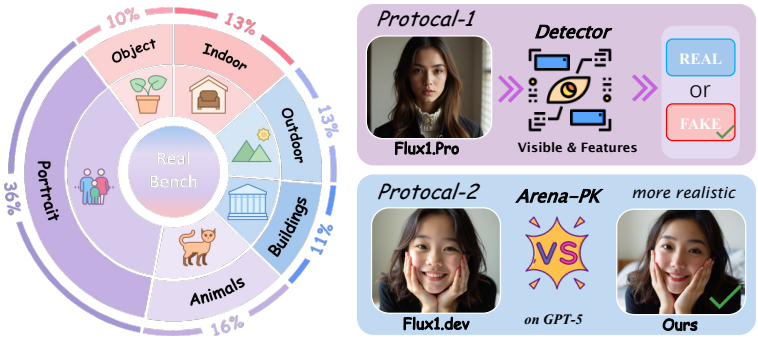

Fig. 4: Overview of the RealBench. The left shows the categorical data composition. The right details its evaluation protocol.

visible semantic layer and the invisible feature layer. First, we include Forensic-Chat [22] and OmniAID [13], which are consistent with our reward function. At the semantic level, we additionally employ SIDA [15], an out-of-domain MLLM-based synthetic image detector, utilizing rigorous prompt engineering to guide its attention toward common artifact regions. At the feature level, we utilize Effort [47], another out-of-domain expert synthetic image detector renowned for its robust generalization capabilities in assessing image authenticity. The inclusion of these two reward-independent detectors helps ensure the comprehensiveness and robustness of the evaluation.

4.3 Arena-Style Preference Evaluation

Given the inherent difficulty of obtaining reliable absolute scores for single images from existing models, we introduce Arena-Scoring, a pairwise comparison protocol inspired by LM-Arena [2]. In this protocol, we employ GPT-5 as the "judge model" to simulate user preferences (validated by >80% agreement with Gemini 3.0 to mitigate model-specific biases). During evaluation, the judge model is presented with two semantically similar images corresponding to the identical text prompt(e.g., Model A vs. Model B, or Model A vs. Real) and is compelled to make a forced-choice decision, selecting the one it perceives as more realistic. Finally, the images generated by each model undergo at least 3000 random pairwise "battles" against outputs from other models and real-world data. We then calculate the final win rate. The inclusion of real images in the comparison pool not only enhances scoring stability but also allows us to validate the plausibility and reliability of the MLLM judge's preferences.

4.4 Human Evaluation and Validation of RealBench Protocol

To validate our proposed Detector-Scoring and Arena-Scoring protocols, we invited 15 non-expert evaluators for human assessment. We selected four repre-

[2] https://lmarena.ai/leaderboard/text-to-image (accessed June 28, 2026).

Table 1: Evaluation of image generation ability on RealBench. **Bold** indicates the best result, and underlined denotes the second best. * indicates that we used an LLM component for prompt optimization, while † denotes the use of synthetic image detectors unseen during training.

Model	Detector-Scoring				Arena-Scoring		Other metrics			
	Forensic-chat	OmniAID	Effort†	SIDA†	VS Real	VS Others	PickScore	CLIP	HPSv2.1	HPSv3
Closed-Sourced T2I Model										
FLUX-Pro [18]	57.45	21.55	**20.94**	24.71	18.20	-	23.68	86.85	30.79	12.78
Nano-Banana [8]	46.43	31.02	11.74	24.51	**42.17**	-	23.67	84.02	30.90	**12.95**
SeedDream 3.0 [33]	63.47	23.73	8.91	**28.53**	36.40	-	**23.92**	**88.61**	**31.48**	12.02
GPT-Image-1 [26]	**75.63**	**33.59**	5.70	11.30	33.71	-	23.89	88.57	31.25	12.48
Open-Sourced T2I Model										
SDXL [28]	43.37	24.44	8.44	25.84	9.22	35.20	23.02	84.65	28.44	9.87
SD-3.5-Large [7]	48.63	20.02	17.45	33.33	23.82	55.58	23.46	88.23	30.68	12.06
FLUX.1-Dev [18]	40.91	21.32	14.85	26.03	12.61	43.60	23.59	86.33	31.21	13.58
FLUX.1-Kontext [19]	37.20	20.68	20.55	10.76	3.65	20.40	22.80	84.20	30.08	11.61
Echo-4o [51]	38.86	16.93	15.71	20.07	6.34	34.35	23.53	**89.25**	31.69	12.27
Bagel [6]	39.47	17.77	14.47	26.89	5.47	22.95	23.39	87.92	31.19	12.43
Qwen-Image [41]	57.47	36.82	17.10	30.93	18.25	47.35	21.97	82.83	25.50	8.15
SRPO [31]	64.14	40.09	24.73	50.62	40.70	64.30	23.55	86.16	29.66	12.43
FLUX-Krea [20]	57.10	32.44	18.42	37.46	37.60	66.40	**23.77**	87.60	30.75	12.50
Ours	70.59	37.85	31.71	67.43	43.41	74.80	23.58	86.80	**31.87**	**13.61**
Ours*	**80.84**	**47.20**	**38.35**	**73.10**	**50.15**	**84.85**	21.75	87.69	28.24	11.11

sentative methods: RealGen (Ours), FLUX-Krea, Qwen-Image, and SD-3.5-Large. For each method, we randomly sampled 100 generated images and paired them with their corresponding real images. The human evaluators were tasked with two distinct objectives: identifying whether a single given image is real, and selecting the more photorealistic image from a real-fake pair. To ensure fairness, the data assigned to the same evaluator across these two tasks did not overlap. Results demonstrate that Detector-Scoring (OmniAID) and Arena-Scoring (GPT-5) align with human visual preferences at 75% and 87%, respectively. This mitigates the risk of metric overfitting, preventing models from merely fooling detectors while generating images that still appear visually synthetic to human observers. Furthermore, despite marginal differences in absolute scores, the relative model rankings between the automated protocols and human judgments are perfectly consistent. This strongly indicates that RealBench serves as a robust and accurate benchmark for measuring the relative capabilities of generation models.

5 Experiments

5.1 Experimental Setup

Implement Details. Our RealGen model employs Qwen3-4B-Instruct [48] as the LLM component and FLUX.1.dev [18] as the base image generator. The training data for SFT and RL is primarily sourced from the real image subset of HPD v3 [24]. All experiments are conducted on 8 H200 GPUs. For the first stage

of GRPO, we set the batch size to 32; for the second stage, the batch size is 12. The entire RL training process iterates for approximately 230 steps. To ensure a fair evaluation, we assessed model effectiveness on two datasets independent of the HPD v3 training data: our newly proposed RealBench and the "Photo" subset of HPD v2 [42].

Comparison Method. We conduct extensive comparisons between RealGen and current T2I models. These baselines include general generative models, such as the closed-source GPT-Image-1 [26] and Nano Banana [8], as well as open-source models like Flux.1 dev [18] and Qwen-Image [41]. For a fair comparison, we also benchmark against methods specifically employing RL optimization for human preference or realism, such as Flux-Krea [20] and SRPO [31]. To ensure a fair comparison, all baselines utilize their official settings.

Evaluation Protocol. We employ a comprehensive evaluation protocol composed of three main categories of metrics: (1) Detector-Scoring, (2) Arena-Scoring, and (3) Other Metrics. The detailed configurations for the first two categories are described in Section 4 (RealBench). The third category, Other Metrics, includes preference alignment scores such as Pick-Score [17], HPSv2.1 [42], and HPSv3 [24], as well as a LongCLIP [53] image-text alignment metric. It is worth noting that HPSv3, unlike its predecessors, incorporates real-world images into its training data, allowing it to more comprehensively reflect human preferences for both photorealism and overall quality.

5.2 Main Results

Table 1 presents the evaluation results of different methods on **RealBench.** Our proposed RealGen outperforms existing leading T2I models across multiple key photorealism metrics, regardless of whether the LLM prompt rewrite component is used. For instance, images generated by general-purpose open-source T2I models, such as FLUX.1-kontext and Bagel, are easily identified by synthetic detectors, exposing obvious AI artifacts. Methods specialized for photorealism, like Flux-Krea and SRPO, achieve sub-optimal performance, but still a significant gap remains compared to our method. Crucially, strong performance on held-out evaluators (SIDA and Effort), coupled with the high human alignment demonstrated in Sec. 4.4, which were excluded from training, validates that RealGen learned robust realism rather than merely reward hacking or generating adversarial artifacts to fool detectors.

In the arena-style pairwise comparisons, we first analyze the "battle" results against Real images. RealGen demonstrates a significant advantage in this comparison, achieving a win rate approaching 50%, which suggests its outputs are capable of being confused with reality. In contrast, 8 of the 13 competing models achieved win rates below 30% against real images, clearly exposing their lack of photorealism. This stark disparity also validates the effectiveness of our Arena-Scoring as a reliable arbiter for realism. Furthermore, Figure 5 displays the win-rate matrix from the model-vs-model battles. The matrix confirms that RealGen achieves the highest overall win rate, indicating it is consistently selected as the more realistic option when compared directly against its peers.

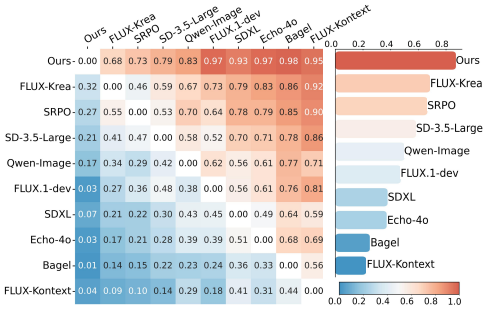

Fig. 5: Pairwise Win-Rate Matrix of Arena-Style Realism Battles.

Table 2: Evaluation on the "Photo" subset of HPDv2 [42].

Model	Det-Score		Aes-Score	
	Foren.	Omni.	HPS2.1	HPS3
FLUX-Pro	66.57	40.13	28.57	11.64
Nano-Banana	37.44	33.09	29.37	12.23
GPT-Image-1	63.75	27.56	29.73	12.19
SDXL	44.53	32.22	25.92	7.27
Qwen-Image	48.85	32.17	27.60	9.44
SRPO	62.65	48.23	27.88	11.11
FLUX-Krea	58.00	44.96	28.76	11.39
Ours	**71.34**	**56.93**	**30.18**	**13.10**

Table 3: HPDv3 Benchmark of popular image generation models. The upper section presents the generation methods originally included in the HPDv3 benchmark. The lower section details emerging methods independent of the HPSv3 training distribution, where we emphasize relative performance over absolute scores.

Models	Characters	Architecture	Animals	Natural.	Plants	Food	Overall
Stable Diffusion XL [28]	8.67	8.57	8.18	7.76	8.32	8.43	8.32
Stable Diffusion v2.0 [30]	-0.34	-0.24	-0.54	-0.32	-0.01	-0.38	-0.31
Stable Diffusion 3 [7]	6.70	5.25	4.09	5.24	5.84	6.01	5.52
Hunyuan [21]	7.96	8.71	7.24	7.86	8.28	8.31	8.06
PixArt-Σ [4]	10.08	9.83	8.86	8.87	9.52	9.73	9.48
Gemini 2.0 Flash [12]	9.98	10.11	9.42	9.01	9.55	10.16	9.71
Flux-Dev [18]	11.70	**10.93**	10.38	10.01	10.21	10.38	10.60
Kolors [34]	**11.79**	10.82	**10.60**	9.89	**10.50**	**10.63**	**10.71**
SD-3.5-Large [7]	9.79	9.41	8.91	8.63	9.07	9.79	9.27
Qwen-Image [41]	7.66	7.65	6.99	7.55	7.18	7.87	7.48
Nano-Banana [8]	9.84	10.02	9.17	9.40	9.02	9.98	9.57
FLUX-Krea [20]	9.53	9.71	9.53	9.11	9.09	9.69	9.44
Ours	**10.83**	**10.75**	**9.98**	**10.05**	**9.95**	**10.65**	**10.37**

To further validate the generalization ability of RealGen, we conducted an additional evaluation on the "photo" subset of **HPDv2**, with results presented in Table 2. The results clearly demonstrate that images synthesized by RealGen not only achieve exceptionally high detector-based realism scores but also rank among the top in human aesthetic preference scores. This provides evidence that our method can significantly enhance photorealism while simultaneously maintaining high aesthetic quality.

We further extend our evaluation to the **HPDv3** [24] benchmark to specifically assess the aesthetic quality of the photorealistic images generated by RealGen. Unlike previous metrics (e.g., HPSv2 or PickScore), HPDv3 employs the HPSv3 score and incorporates real image data into its training pipeline, resulting in a higher correlation with human perception. In this experiment, we exclude the "art" style category from the benchmark, with detailed results presented in Table 3. However, it is crucial to note the potential bias inherent in such

Fig. 6: Qualitative comparison of different methods on RealBench.

preference-based metrics. Specifically, models like Flux.1 Dev and Kolors have been exposed to the HPSv3 training data, which often leads to artificially inflated scores that do not necessarily reflect superior qualitative performance. Therefore, our comparative analysis primarily focuses on the emerging methods listed in the lower half of the table, which are independent of the HPSv3 training distribution. The results demonstrate that RealGen achieves superior performance by striking an optimal balance between high aesthetic quality and photorealism.

Fig. 6 illustrates the qualitative visual comparisons. Base models like FLUX-dev and Bagel tend to produce images with excessive oiliness and unnatural highlights. Qwen-Image often generates overly smooth skin, exhibiting a typical AI artifact. Beyond a distinct "AI plastic" feel, GPT-Image-1 also shows an unnatural warm color cast skewed towards yellow-green. In contrast, the results from RealGen are superior in both texture and detail, appearing visually closer to actual photographs.

5.3 Ablation experiments

As shown in Fig. 7, we conducted a progressive ablation study to explore the impact of the LLM optimization component and the Diffusion optimization com-

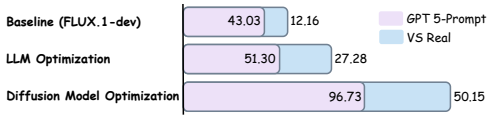

Fig. 7: Ablation on different key components.

Table 4: Ablation on Rewards.

Method	Effort	SIDA	VS Real	CLIP
Base (FLUX.1)	14.85	26.03	12.61	86.33
+ PickScore	12.75	29.30	22.96	86.17
+ HPSv2.1	11.46	28.82	19.12	86.24
+ Detector-Score	**31.71**	**67.43**	**43.41**	**86.80**

Fig. 8: Effectiveness of different components for photorealistic generation.

Fig. 9: Visualization of the effects of different reward functions.

ponent. For quantitative analysis, we employed two reward-independent metrics: the GPT-prompt Score and vs. Real images Arena-Scoring. The quantitative results indicate that applying only the Phase 1 LLM prompt optimization already leads to more realistic and detailed images by generating richer and more diverse prompts. Building on this, the subsequent inclusion of the Diffusion model optimization further improves image quality. The visualization in Fig. 8 shows the effect of component optimization: the LLM adds realistic descriptions to enrich detail, while optimizing the Diffusion model further enhances the image realism of the person, and glass.

Furthermore, we explored the distinct impacts of different reward functions on model optimization, comparing human preference rewards against our Detector-score. As shown in Table 4, when evaluated on multiple metrics held-out from the optimization objective, our Detector-Score demonstrates a clear advantage, proving it guides the model towards superior photorealism. The qualitative results in Figure 9 intuitively explain this gap: reward functions like PickScore and HPSv2.1 tend to bias the model towards cartoonish or artistic styles and still produce "oily" textures on portraits. In contrast, our proposed Detector-Score consistently leads to more realistic and photorealistic results. This comparison focuses on the aggregate Detector-Score rather than isolating each individual reward component.

6 Conclusion

To address the photorealism gap in current T2I models, we introduced RealGen. Our core contribution is the "Detector Reward" mechanism, inspired by adversar-

ial generation, which utilizes both semantic and feature-level detectors to quantify image realism. By leveraging this reward signal with the GRPO algorithm, we successfully optimized the entire generation pipeline, enhancing image realism and detail; this effectively realizes a **"Detection for Generation" paradigm in the RL era.** Furthermore, we proposed RealBench, an automated benchmark employing Detector-Scoring and Arena-Scoring, which enables human-free photorealism evaluation.

Acknowledgments. This work was partially supported by Guangdong S&T Program (2024B0101040005), National Natural Science Foundation of China (Grant No. T2125006, 62571560) and Shanghai Artificial Intelligence Laboratory.

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
2. Cao, T., Wang, C., Liu, B., Wu, Z., Zhu, J., Huang, J.: Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. arXiv preprint arXiv:2311.06752 (2023)
3. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
4. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
5. Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., et al.: Emu3. 5: Native multimodal models are world learners. arXiv preprint arXiv:2510.26583 (2025)
6. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning (2024)
8. Fortin, A., Vernade, G., Kampf, K., Reshi, A.: Introducing gemini 2.5 flash image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/ (2025), accessed: 2026-06-28
9. Gao, J., Liu, Y., Sun, Y., Tang, Y., Zeng, Y., Chen, K., Zhao, C.: Styleshot: A snapshot on any style. arXiv preprint arXiv:2407.01414 (2024)
10. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
11. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

12. Google: Gemini 2.0 flash. https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation (2025), accessed: 2026-06-28
13. Guo, Y., Ye, J., Zhang, C., Kang, H., Fu, H., He, C., Li, W.: Omniaid: Decoupling semantic and artifacts for universal ai-generated image detection in the wild (2025). https://doi.org/10.48550/arXiv.2511.08423, https://arxiv.org/abs/2511.08423, accessed: 2026-06-28
14. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
15. Huang, Z., Hu, J., Li, X., He, Y., Zhao, X., Peng, B., Wu, B., Huang, X., Cheng, G.: Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28831–28841 (2025)
16. Kang, H., Wen, S., Wen, Z., Ye, J., Li, W., Feng, P., Zhou, B., Wang, B., Lin, D., Zhang, L., et al.: Legion: Learning to ground and explain for synthetic image detection. arXiv preprint arXiv:2503.15264 (2025)
17. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in neural information processing systems **36**, 36652–36663 (2023)
18. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024), accessed: 2026-06-28
19. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742, accessed: 2026-06-28
20. Lee, S., Millon, E., Ebbecke, T., Beddow, W., Zhuo, L., García-Ferrero, I., Esparraguera, L., Petrescu, M., Saß, G., Menezes, G., Perez, V.: Flux.1 krea [dev]. https://github.com/krea-ai/flux-krea (2025), accessed: 2026-06-28
21. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024)
22. Lin, K., Yan, Z., Chen, R., Ye, J., Zhang, K.Y., Zhou, Y., Jin, P., Li, B., Yao, T., Ding, S.: Seeing before reasoning: A unified framework for generalizable and explainable fake image detection. arXiv preprint arXiv:2509.25502 (2025)
23. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)
24. Ma, Y., Wu, X., Sun, K., Li, H.: Hpsv3: Towards wide-spectrum human preference score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15086–15095 (2025)
25. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models (2023), https://arxiv.org/abs/2302.08453, accessed: 2026-06-28
26. OpenAI: Introducing our latest image generation model in the api. https://openai.com/index/image-generation-api/ (2025), accessed: 2026-06-28
27. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al.: Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256 (2025)
28. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

29. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
31. Shen, X., Li, Z., Yang, Z., Zhang, S., Zhang, Y., Li, D., Wang, C., Lu, Q., Tang, Y.: Directly aligning the full diffusion trajectory with fine-grained human preference. arXiv preprint arXiv:2509.06942 (2025)
32. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8871–8879 (2024)
33. Team, B.S.: Seedream 3.0 technical report (2025), https://arxiv.org/abs/2504.11346, accessed: 2026-06-28
34. Team, K.: Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. arXiv preprint (2024)
35. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
36. Wang, L., Xing, X., Cheng, Y., Zhao, Z., Li, D., Hang, T., Tao, J., Wang, Q., Li, R., Chen, C., et al.: Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. arXiv preprint arXiv:2509.04545 (2025)
37. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
38. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020)
39. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
40. Wen, S., Ye, J., Feng, P., Kang, H., Wen, Z., Chen, Y., Wu, J., Wu, W., He, C., Li, W.: Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. arXiv preprint arXiv:2503.14905 (2025)
41. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
42. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis (2023), https://arxiv.org/abs/2306.09341, accessed: 2026-06-28
43. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
44. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564 (2025)
45. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)

46. Yan, Z., Lin, K., Li, Z., Ye, J., Han, H., Wang, H., Wang, Z., Lin, B., Li, H., Xiao, X., et al.: Unified multimodal models as auto-encoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41903–41912 (2026)
47. Yan, Z., Wang, J., Jin, P., Zhang, K.Y., Liu, C., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L.: Orthogonal subspace decomposition for generalizable ai-generated image detection. arXiv preprint arXiv:2411.15633 (2024)
48. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
49. Ye, J., He, J., Li, W., Lv, Z., Yu, J., Yang, H., He, C.: Skydiffusion: Street-to-satellite image synthesis with diffusion models and bev paradigm. arXiv e-prints pp. arXiv–2408 (2024)
50. Ye, J., He, J., Zhang, X., Lin, Y., Lin, H., He, C., Li, W.: Satellite image synthesis from street view with fine-grained spatial textual guidance: A novel framework. IEEE Geoscience and Remote Sensing Magazine (2025)
51. Ye, J., Jiang, D., Wang, Z., Zhu, L., Hu, Z., Huang, Z., He, J., Yan, Z., Yu, J., Li, H., et al.: Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. arXiv preprint arXiv:2508.09987 (2025)
52. Ye, J., Zhou, B., Huang, Z., Zhang, J., Bai, T., Kang, H., He, J., Lin, H., Wang, Z., Wu, T., et al.: Loki: A comprehensive synthetic data detection benchmark using large multimodal models. arXiv preprint arXiv:2410.09732 (2024)
53. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of clip. In: European conference on computer vision. pp. 310–325. Springer (2024)
54. Zhang, L., Xu, Z., Barnes, C., Zhou, Y., Liu, Q., Zhang, H., Amirghodsi, S., Lin, Z., Shechtman, E., Shi, J.: Perceptual artifacts localization for image synthesis tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7579–7590 (2023)
55. Zhang, L., Zhou, Y., Barnes, C., Amirghodsi, S., Lin, Z., Shechtman, E., Shi, J.: Perceptual artifacts localization for inpainting. In: European Conference on Computer Vision. pp. 146–164. Springer (2022)
56. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3836–3847 (October 2023)
57. Zhang, X., Courville, A., Drozdzal, M., Romero-Soriano, A.: The intricate dance of prompt complexity, quality, diversity, and consistency in t2i models. arXiv preprint arXiv:2510.19557 (2025)
58. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: Diffusionnft: Online diffusion reinforcement with forward process. arXiv preprint arXiv:2509.16117 (2025)
59. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image. Advances in Neural Information Processing Systems **36**, 77771–77782 (2023)