PORTRAITGEN: ANCHORING GRPO WITH REAL-IMAGE TRAJECTORIES FOR PHOTOREALISTIC PORTRAIT GENERATION

Anonymous authors
Paper under double-blind review

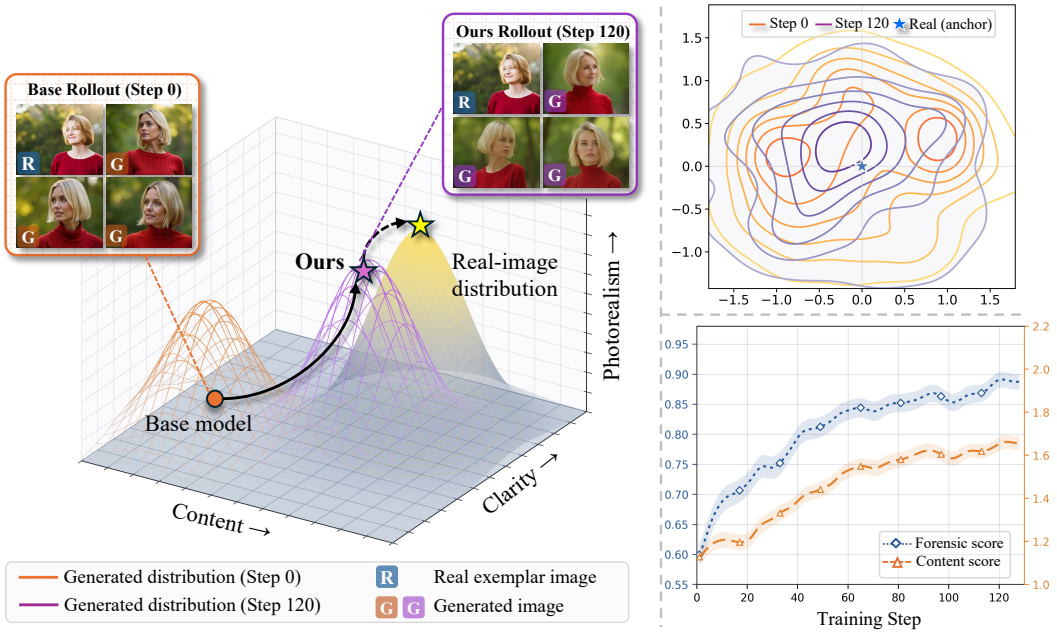

Figure 1: PortraitGen grounds GRPO in real-image trajectories, progressively transporting the rollout distribution toward that of photorealistic portraits. **(Left)** Successive training stages reveal a transition from base-model samples to increasingly photorealistic generations. **(Top right)** The discrepancy between generated and real-image distributions contracts throughout optimization. **(Bottom right)** Consistent gains in content fidelity and independently measured forensic realism provide quantitative evidence of this progressive alignment.

ABSTRACT

Reinforcement Learning like Group Relative Policy Optimization (GRPO) has significantly advanced text-to-image post-training. However, existing methods often favor superficial aesthetics, such as over-saturated colors, leaving critical flaws like AI artifacts and biological implausibilities unresolved. We attribute these limitations to two primary factors: (1) GRPO relies solely on self-sampled groups without real-image trajectory references, constraining optimization to the model's existing generative distribution; (2) existing reward signals are insufficiently sensitive to fine-grained portrait artifacts, such as anatomical inconsistencies and unnatural skin appearance. To address this, we propose PortraitGen, a novel framework tailored for photorealistic portrait generation. First, we introduce a real exemplar into each GRPO group and employ BELM inversion to construct its latent trajectory, enabling the exemplar to participate in both relative comparison and policy optimization. Second, to explicitly steer the model toward photorealism, we introduce a complementary dual-reward mechanism: OmniReward for general quality and AI-Portrait for human-centric fidelity. Furthermore, we curate PortraitBench, a comprehensive portrait-centric benchmark. Extensive experiments demonstrate that PortraitGen significantly outperforms existing baselines, effectively suppressing AI artifacts and improving portrait realism.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Recent advancements in text-to-image (T2I) generation have underscored the pivotal role of reinforcement learning (RL) Seedream et al. (2025); Black et al. (2023); Fan et al. (2023); Wallace et al. (2024b); Clark et al. (2023); Xu et al. (2026); Wallace et al. (2024a). Built upon pre-trained models, these frameworks can further align generated images with human preferences defined by reward models Kirstain et al. (2023); Xu et al. (2023); Ma et al. (2025). Specifically, Group Relative Policy Optimization (GRPO) Shao et al. (2024) has been extended to visual generation by methods such as Flow-GRPO Liu et al. (2025) and DanceGRPO Xue et al. (2025), enabling reward-guided optimization of flow matching models such as FLUX Labs (2025; 2024).

However, while these post-training techniques yield visually appealing images, they often induce superficial stylistic shifts, such as over-saturated colors, instead of rectifying underlying structural flaws Xu et al. (2023); Li et al. (2024); Hu et al. (2026). This limitation is particularly apparent in photorealistic portraits, where anatomical inconsistencies, over-smoothed or oily skin, and other synthetic signatures strongly affect perceived realism. Existing portrait generation works He et al. (2024); Li et al. (2023b) mainly focus on identity customization and do not directly address these post-training objectives. Although reinforcement learning has been explored for general T2I generation Jiang et al. (2025); Liang et al. (2025), comparatively little attention has been paid to jointly improving portrait quality, extremity integrity, and natural appearance.

Recent efforts have begun to pivot toward portrait realism. Ye et al. (2025) identifies that models like FLUX Labs (2024) and Qwen-Image Wu et al. (2025) often tend to produce over-smoothed, unnatural oily skin. However, by relying on deepfake detection tools, Ye et al. (2025) merely classifies images as "real" or "AI-generated" based on feature-level artifacts, lacking a specialized focus on human-centric nuances. Conversely, while Liao et al. (2025) evaluates higher-level aesthetic dimensions (e.g., facial beauty, physique, and environment), it remains aligned with traditional aesthetic standards and fails to quantify the AI artifacts or skin oiliness prevalent in flow matching generators. Consequently, there is an need for specialized reward models to explicitly quantify these AI artifacts. Beyond reward design, a more fundamental limitation of GRPO-based methods Liu et al. (2025); Xue et al. (2025); Wang et al. (2025b) is the high similarity of sampled data within a group. Because samples are drawn from the same generative distribution, the model never observes authentic photographs during optimization, making genuine photorealism an unreachable goal.

To this end, we propose PortraitGen, a post-training framework dedicated to suppressing AI artifacts and achieving superior photorealistic portrait generation. Unlike conventional GRPO groups composed entirely of model-generated samples, PortraitGen introduces a real image as an exemplar alongside $G - 1$ generated images. We employ BELM inversion to reconstruct the exemplar and obtain its latent trajectory, enabling it to participate in both group-relative comparison and policy optimization. However, relying solely on injected exemplars is insufficient, as the model lacks explicit feedback to guide the optimization. To actively propel the model toward genuine photorealism, we introduce two complementary reward models. On one hand, OmniReward provides point-wise evaluations across four fundamental dimensions: Content, Clarity, Lighting and Color, and Composition. On the other hand, the AI-Portrait reward explicitly penalizes synthetic signatures, i.e., the AI artifacts and skin oiliness, employing pairwise win-rate scoring within groups to amplify subtle perceptual differences. Additionally, we introduce PortraitBench, a diverse benchmark for photorealistic portrait generation. Extensive experiments demonstrate that PortraitGen improves portrait realism and suppresses common synthetic artifacts. As visualized in Fig. 1, the generated distribution progressively approaches the real-image distribution throughout training, further substantiating the improvement in photorealism. The primary contributions of this work are as follows:

- We propose Real-Image-Anchored GRPO, which augments policy-generated groups with deterministic BELM trajectories derived from real images, effectively breaking the inherent generative boundaries of standard GRPO methods.
- We design two complementary reward models, OmniReward for general generation and AI-Portrait reward for human-centric fidelity, serving as both optimization signals and independent evaluation metrics.
- We curate PortraitBench, a diverse benchmark spanning various demographics, to comprehensively evaluate portrait realism across rich scenarios.

2 Related Work

2.1 Reward Models for Image Generation

Reward models play a central role in aligning image generators with human preferences Wang et al. (2025c); Hartwig et al. (2025); Yu et al. (2025). Conventional metrics such as FID Heusel et al. (2017) and CLIP score Radford et al. (2021) offer coarse measures of visual quality and semantic alignment, but often fail to capture complex human preferences. Preference-aware methods, including PickScore Kirstain et al. (2023) and HPSv2 Wu et al. (2023), address this limitation by learning directly from human feedback. More recently, Multimodal Large Language Models (MLLMs) Bai et al. (2025); Grattafiori et al. (2024); Li et al. (2023a) have emerged as versatile reward-modeling backbones, supporting both score regression Xu et al. (2023); Wang et al. (2025c); Ma et al. (2025) and direct pairwise comparison Ku et al. (2024); Xiong et al. (2025). Despite these advancements, current models lack the sensitivity to assess portrait images or detect the specific synthetic artifacts associated with AI generation. We therefore formulate two complementary reward models to evaluate human-centric fidelity and detect AI artifacts.

2.2 Fine-tuning T2I Models with Rewards

Recent studies align T2I models with desired objectives through RL Jiang et al. (2025); Liang et al. (2025); Kaufmann et al. (2025), often by directly maximizing reward signals Li et al. (2025); Xu et al. (2023); Li et al. (2024). Flow-GRPO Liu et al. (2025) and DanceGRPO Xue et al. (2025) extend GRPO to flow matching models by converting deterministic ODE trajectories into stochastic SDEs for policy learning. Pref-GRPO Wang et al. (2025b) instead employs pairwise preference fitting to reduce reward hacking, but is not tailored to portrait-specific artifacts. RealGen Ye et al. (2025) employs synthetic-image detectors to suppress artifacts at both semantic and feature levels. While effective in alleviating skin oiliness, it remains limited in correcting structural distortions. Despite this progress, existing GRPO schemes sample exclusively from the model distribution and never observe photographic-quality images during optimization. We address this limitation with a new GRPO strategy designed to suppress persistent AI artifacts.

3 Reward Models

We introduce two complementary reward models for photorealistic portrait generation. OmniReward is a point-wise evaluator of general visual quality, whereas AI-Portrait Reward performs pairwise comparisons focused on portrait-specific synthetic artifacts.

3.1 OmniReward

Data Construction. Professional annotators score real-world images from 0 to 2 along four dimensions: Content, Clarity, Lighting and Color, and Composition. For samples assigned a score of 0, they additionally identify the cause of failure, such as structural artifacts or inappropriate content. After balancing the score distribution, we obtain 93,465 annotated samples. To provide interpretable supervision without costly free-form annotation, we use Qwen3-VL-235B-A22B-Instruct to generate visual rationales conditioned on the image, score, and human-provided failure tags. Each target response thus contains a concise rationale followed by the four scores. Representative training examples and further construction details are provided in Appendix B.

Reward Model Training. We utilize Qwen3-VL-8B-Instruct Bai et al. (2025) as the backbone, which is trained following the standard Supervised Fine-Tuning (SFT) protocol. Given an image x, instruction s, and target response y, the training objective is

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(x,s,y)\sim\mathcal{D}}\left[\sum_{t=1}^{T}\log P_\theta\left(y_t \mid x, s, y_{<t}\right)\right], \tag{1}$$

where θ denotes the learnable parameters and $y = \{y_1, y_2, \ldots, y_T\}$ contains the visual rationale and dimension-specific scores. Once trained, OmniReward provides multidimensional rewards for post-training and can also serve as an evaluation metric.

Table 1: Prediction accuracy (%) of our reward models and vanilla Qwen3-VL models on the test sets. "Avg." denotes average accuracy.

Model	OmniReward					AI-Portrait Reward		
	Content	Clarity	Color	Composition	Avg.	AI-feel	Oiliness	Avg.
Qwen3-VL-8B-Instruct	13.50	1.30	8.20	3.90	6.73	51.25	50.80	51.03
Qwen3-VL-235B-A22B-Instruct	33.93	4.00	2.30	10.21	12.61	53.71	54.22	53.97
OmniReward	**86.20**	**95.20**	**95.80**	**85.70**	**90.73**	-	-	-
AI-Portrait Reward	-	-	-	-	-	**96.2**	**96.5**	**96.35**

3.2 AI-Portrait Reward

Data Construction and Training. We sample approximately 10,000 real portraits and generate corresponding synthetic counterparts using FLUX.1-dev Labs (2024), Z-Image Cai et al. (2025), and Qwen-Image Wu et al. (2025), yielding 30,000 real-synthetic pairs. Based on the same Qwen3-VL-8B-Instruct backbone, AI-Portrait Reward receives a real-synthetic pair and identifies which image exhibits stronger AI artifacts and skin oiliness. Unlike OmniReward, it predicts only pairwise labels without rationales, avoiding potentially hallucinated defect descriptions for high-quality synthetic images. The data format can be found in Appendix B.

Evaluation. We evaluate both reward models on held-out test sets. OmniReward is assessed on 1,000 images with ground-truth annotations across its four scoring dimensions, while AI-Portrait Reward is evaluated on a disjoint set of 1,000 real-synthetic pairs whose presentation order is randomized to avoid positional bias. As shown in Tab. 1, task-specific supervision yields substantially more reliable multidimensional assessment and portrait-artifact discrimination than directly prompting the underlying vision-language models, including the substantially larger Qwen3-VL-235B-A22B-Instruct. This confirms that reliable reward estimation for fine-grained, precision-sensitive tasks benefits substantially from task-specific fine-tuning rather than model scaling alone.

4 Real-Image-Anchored GRPO

4.1 Preliminary

Group Relative Policy Optimization. The iterative image-denoising trajectory can be formulated as a Markov Decision Process (MDP), in which the generative model acts as the policy. Conditioned on a prompt c, GRPO samples a group of G rollouts $\{o_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$, and a reward model assigns each rollout a scalar reward r_i. Instead of training a value network, GRPO estimates the advantage using within-group reward statistics:

$$\hat{A}_t^i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^{G})}{\text{std}(\{r_i\}_{i=1}^{G})}. \tag{2}$$

The policy is then optimized with a clipped objective and a KL regularizer:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{c\sim\mathcal{P},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}\left[f(\theta,\hat{A})\right], \tag{3}$$

where $\mathcal{P}$ represents the set of prompts, and the inner function is defined as:

$$f(\theta,\hat{A}) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left[\min\left(\rho_t^i\hat{A}_t^i, \text{clip}(\rho_t^i, 1-\epsilon, 1+\epsilon)\hat{A}_t^i\right) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right]. \tag{4}$$

Here, $\rho_t^i = \frac{\pi_\theta(o_{t-1}^i|o_t^i,c)}{\pi_{\theta_{\text{old}}}(o_{t-1}^i|o_t^i,c)}$ is the importance ratio, ϵ controls clipping, and β weights the KL penalty.

Image Inversion for Flow Matching. Image inversion seeks an initial noise map and its associated trajectory from which a given image can be faithfully reconstructed. Conventional DDIM inversion Song et al. (2020), however, the numerical updates for inversion and reverse sampling are not exact inverses, which can introduce noticeable discrepancies between the original and reconstructed images. We instead adopt BELM Wang et al. (2024), which unifies both directions under the same

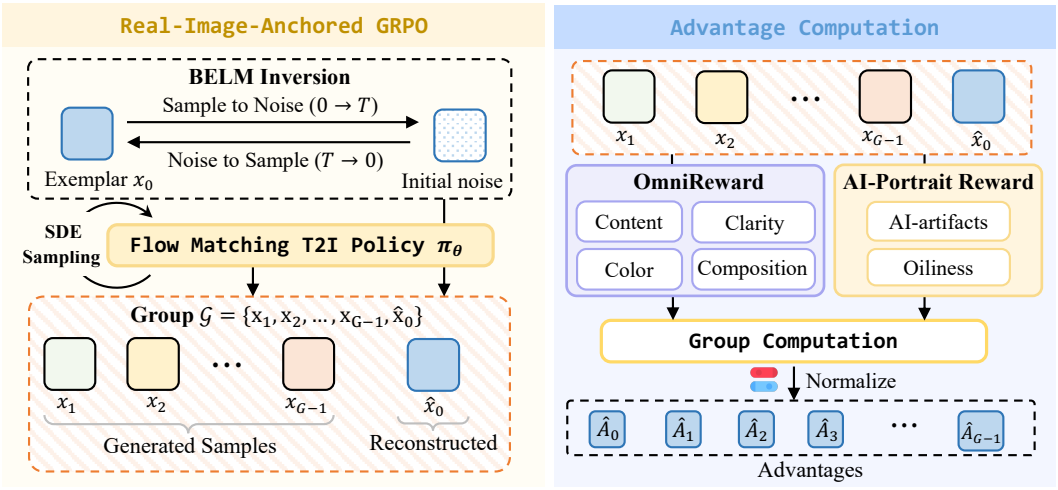

Figure 2: Overview of Real-Image-Anchored GRPO. The T2I model generates $G-1$ images, which form an augmented group with the exemplar reconstructed through BELM inversion.

linear multistep formulation. This bidirectional consistency enables near-lossless reconstruction while preserving the intermediate latents and transition probabilities. Given the flow-matching ODE $dx_t = v_\theta(x_t, t)dt$, the sampling and inversion updates are:

$$\mathbf{x}_{i-1} = \frac{h_i^2}{h_{i+1}^2}\mathbf{x}_{i+1} + \frac{h_{i+1}^2 - h_i^2}{h_{i+1}^2}\mathbf{x}_i - \frac{h_i(h_i + h_{i+1})}{h_{i+1}}v_\theta(\mathbf{x}_i, i) \tag{5}$$

$$\mathbf{x}_{i+1} = \frac{h_{i+1}^2}{h_i^2}\mathbf{x}_{i-1} + \frac{h_i^2 - h_{i+1}^2}{h_i^2}\mathbf{x}_i + \frac{h_{i+1}(h_i + h_{i+1})}{h_i}v_\theta(\mathbf{x}_i, i) \tag{6}$$

where the step size h_i is defined as $h_i = t_i - t_{i-1}$.

4.2 Overall Post-training Framework

Existing GRPO formulations introduce stochastic flow-matching rollouts for group-relative optimization Liu et al. (2025); Xue et al. (2025), but remain confined to the current policy distribution. We instead introduce a real-image trajectory as an external anchor for advantage estimation and policy optimization, progressively narrowing the gap between generated and real-image distributions (Fig. 1). The overall pipeline is shown in Fig. 2.

Exemplar-augmented Trajectory Groups. Given a text condition c and its paired real image x^{ref}, we construct a group that couples policy-sampled trajectories with an exemplar-induced trajectory under the same conditioning. Let $\tau_i = (z_0^i, \ldots, z_T^i)$ denote a denoising trajectory along noise levels $\sigma_0 > \cdots > \sigma_T = 0$. The augmented group is defined as

$$\mathcal{G}(c, x^{\text{ref}}) = \{\tau_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid c)\}_{i=1}^{G-1} \cup \{\tau_G = \mathcal{B}_{\theta_{\text{old}}}(E(x^{\text{ref}}) \mid c)\}. \tag{7}$$

Here $E(x^{\text{ref}})$ is a VAE-encoded latent, and $\mathcal{B}_{\theta_{\text{old}}}(\cdot \mid c)$ maps it to a reconstruction trajectory through BELM inversion and deterministic denoising. Unlike the other $G-1$ trajectories, τ_G follows the deterministic BELM reconstruction without additional SDE noise. Rewards are evaluated on the resulting decoded images o_i, including the reconstructed exemplar, rather than assigning the original photograph an artificial maximum reward.

Policy-density Evaluation of Reference Transitions. BELM determines the reference states, but its deterministic updates do not provide the stochastic transition density required for policy optimization. We score these states using the same Gaussian transition kernel as the generated trajectories in our GRPO-based FLUX implementation. Specifically, its velocity prediction defines the clean-latent estimate $\widehat{z}_{\text{clean},\theta} = z_t - \sigma_t v_\theta(z_t, \sigma_t, c)$ and the score estimate $s_\theta(z_t, \sigma_t, c) = -[z_t - (1-\sigma_t)\widehat{z}_{\text{clean},\theta}]/\sigma_t^2$. With step size $\Delta_t = \sigma_t - \sigma_{t+1} > 0$ and noise scale $\eta > 0$, the implemented SDE update has mean

$$m_{\theta,t} = z_t - \Delta_t v_\theta(z_t, \sigma_t, c) + \tfrac{1}{2}\eta^2 \Delta_t s_\theta(z_t, \sigma_t, c). \tag{8}$$

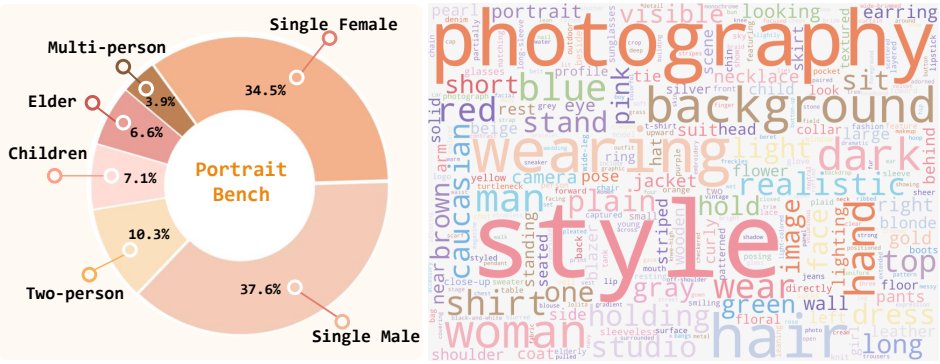

Figure 3: Distribution of our PortraitBench. Left: Categorical distribution of portrait scenarios. The benchmark encompasses diverse demographic groups, with percentages indicating their relative proportions. Right: Word cloud visualization of the benchmark.

The resulting transition kernel is

$$p_\theta(z_{t+1} \mid z_t, c) = \mathcal{N}(z_{t+1}; m_{\theta,t}, \nu_t^2 I), \tag{9}$$

where $\nu_t^2 = \eta^2 \Delta_t$. The distinction between the two branches lies in how the next state is obtained, not in how it is scored. Generated trajectories sample from this Gaussian, whereas the reference branch evaluates it at the fixed next state supplied by BELM. For a latent with d scalar coordinates, we use the averaged log density $\ell_{\theta,t}^i = d^{-1} \log p_\theta(z_{t+1}^i \mid z_t^i, c)$, which expands to

$$\ell_{\theta,t}^i = -\frac{\|z_{t+1}^i - m_{\theta,t}^i\|_2^2}{2d\nu_t^2} - \frac{1}{2}\log(2\pi\nu_t^2). \tag{10}$$

Substituting (z_t^G, z_{t+1}^G) scores the reconstructed exemplar under the current policy while keeping the reference states fixed. Gradients flow through $m_{\theta,t}^G$, not through BELM inversion. This gives the exemplar a direct trajectory-level training signal without perturbing its reconstruction with additional SDE noise.

Advantage Estimation in the Exemplar-augmented Group. Let $r_i^{(k)}$ denote the score of image o_i on reward signal $k \in \mathcal{K}$. OmniReward provides pointwise assessments of content, clarity, lighting and color, and composition. AI-Portrait Reward provides win rates over available within-group comparisons for AI artifacts and skin oiliness, with higher scores indicating fewer artifacts.

We adopt a GDPO-style Liu et al. (2026) advantage formulation, normalizing each reward dimension within the group before weighting and combining the resulting advantages. For each dimension, we compute the group mean $\mu_k = G^{-1}\sum_{j=1}^G r_j^{(k)}$ and standard deviation $s_k = \sqrt{\frac{1}{G}\sum_{j=1}^G (r_j^{(k)} - \mu_k)^2}$. Both statistics include the $G-1$ generated images and the reconstructed exemplar. The aggregated advantage is then obtained by weighting the standardized scores:

$$A_i = \sum_{k \in \mathcal{K}} w_k \frac{r_i^{(k)} - \mu_k}{s_k + \delta}, \tag{11}$$

where w_k balances the dimension-specific advantages and $\delta > 0$ ensures numerical stability. This formulation reduces sensitivity to reward scale. The exemplar participates in the same group statistics, and its advantage guides optimization of its BELM reference transitions.

5 Experiments

5.1 Experimental Settings

Dataset. We curate approximately 10,000 high-quality, photorealistic portraits from social media for training. For evaluation, PortraitBench comprises 1,000 samples spanning diverse genders, ages, and subject counts, as summarized in Fig. 3.

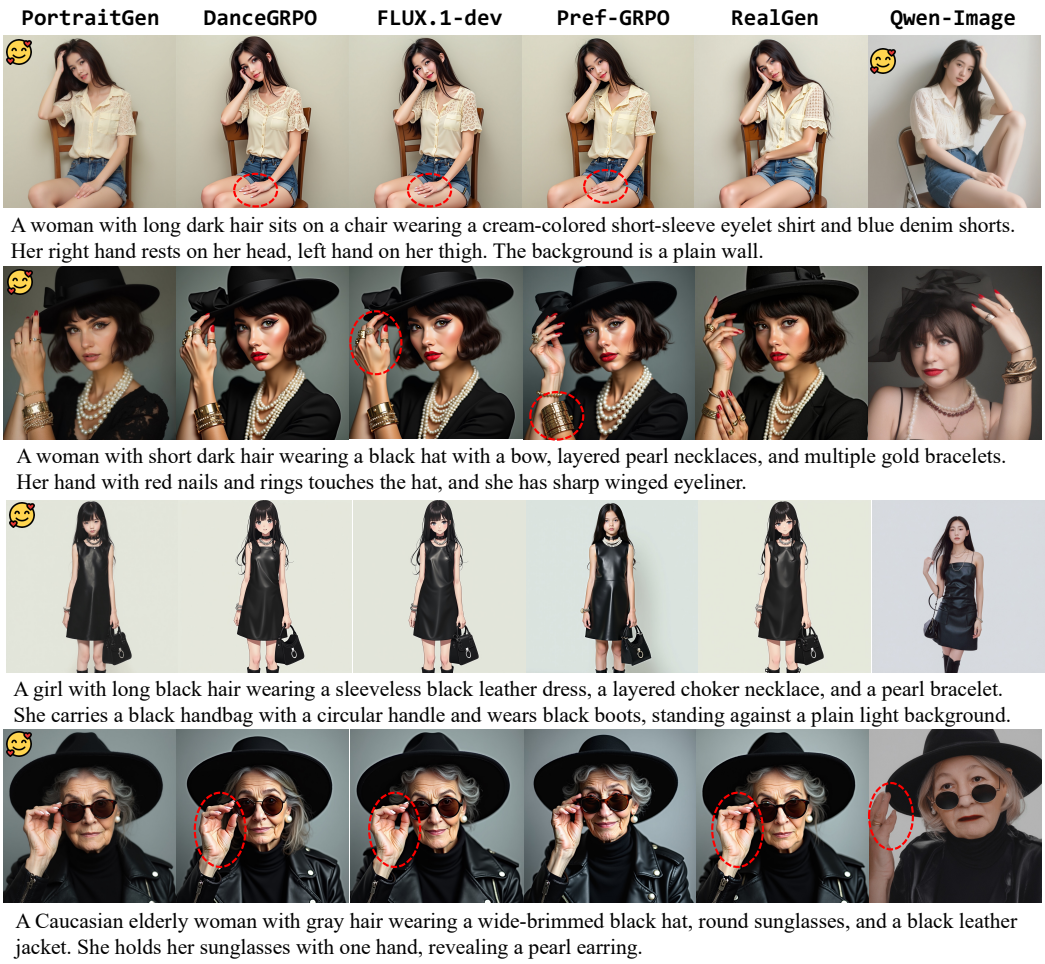

Figure 4: Qualitative comparisons between PortraitGen and other methods. The 🥰 icon denotes the superior generated image for each specific text prompt. Red dashed circles highlight obvious structural distortions in the generated limbs.

Implementation Details. We adopt FLUX.1-dev as our baseline, strictly freezing the backbone to exclusively optimize the LoRA parameters. To ensure the numerical accuracy required for BELM inversion, we perform optimization in FP32 with a learning rate of 5×10^{-5}, a mini-batch size of 1, and 12 gradient accumulation steps. Training is conducted on 16 NVIDIA H20 GPUs.

Metrics. Evaluation is structured around two complementary aspects: perceptual quality and synthetic appearance. The former is characterized by OmniReward across Content, Clarity, Lighting & Color, and Composition, supplemented by UnifiedReward Wang et al. (2025c) and PickScore Kirstain et al. (2023) for general quality and human preference. The latter is assessed by AI-Portrait Reward, together with Forensic-Chat and OmniAID from RealGen Ye et al. (2025), which target semantic- and feature-level artifacts, respectively.

5.2 Qualitative Results

To intuitively illustrate the comparative advantages of our method, we present a comprehensive qualitative evaluation. As shown in Fig. 4, baseline models including FLUX.1 dev, DanceGRPO, and RealGen frequently default to oil painting or anime visual styles. Despite explicit instructions enforcing a photorealistic style within all prompts, these architectures fundamentally fail to generate realistic appearances. Conversely, Qwen-Image delivers exceptional aesthetic quality and avoids skin oiliness when generating Asian subjects. However, its outputs appear heavily processed by a

Table 2: Quantitative comparison of PortraitGen against leading baselines on our benchmark. The best and second-best results are highlighted in **bold** and underlined, respectively. Higher values indicate better performance across all metrics.

Method	OmniReward				UnifiedReward			RealGen		PickScore
	Content	Clarity	Color	Composition	Alignment	Coherence	Style	Forensic.	OmniAID	
Base T2I Models										
SDXL Podell et al. (2023)	0.48	0.99	**0.97**	1.20	3.08	3.68	3.26	23.54	20.87	22.16
SD3.5-medium Esser et al. (2024)	0.94	**1.00**	0.94	1.19	3.43	3.66	3.19	39.43	19.36	22.17
Bagel Deng et al. (2025)	0.62	0.94	**0.97**	1.22	3.50	3.71	3.25	23.44	15.26	22.66
Qwen-Image Wu et al. (2025)	**0.97**	0.95	0.94	1.19	3.38	3.62	3.14	54.22	16.49	21.69
FLUX.1-dev Labs (2024)	0.63	0.96	0.95	1.25	3.48	3.79	3.29	28.42	20.23	22.66
Post-training Methods										
DanceGRPO Xue et al. (2025)	0.65	0.95	0.93	1.25	3.48	3.80	**3.30**	32.57	23.46	22.64
Pref-GRPO Wang et al. (2025b)	0.61	0.97	0.94	1.28	**3.51**	3.79	3.27	28.47	22.94	22.71
RealGen Ye et al. (2025)	0.73	0.87	0.94	1.24	3.47	3.81	3.26	43.90	23.16	22.53
PortraitGen	**0.97**	0.96	0.95	**1.30**	**3.51**	**3.83**	3.27	**56.62**	**24.58**	**22.77**

Table 3: Ablation study on the proposed reward models and real-image anchoring. Comp., Align., and Coher. denote Composition, Alignment, and Coherence, respectively.

Method	Reward Models	Real-Image Anchor	OmniReward				UnifiedReward			RealGen	
			Content	Clarity	Color	Comp.	Align.	Coher.	Style	Forensic.	OmniAID
Base T2I Models											
FLUX.1-dev Labs (2024)	-	-	0.63	**0.96**	**0.95**	1.25	3.48	3.79	3.29	28.42	20.23
DanceGRPO Xue et al. (2025)	-	-	0.65	0.95	0.93	1.25	3.48	3.80	**3.30**	32.57	23.46
Ours Methods											
PortraitGen	✓	×	0.51	**0.96**	**0.95**	1.24	3.49	3.79	3.28	21.70	22.81
	×	✓	0.61	0.92	0.94	1.24	3.49	3.78	3.29	25.56	20.17
	✓	✓	**0.97**	**0.96**	**0.95**	**1.30**	**3.51**	**3.83**	3.27	**56.62**	**24.58**

distinctive stylistic filter. Furthermore, the model is noticeably less effective at generating subjects of diverse ethnicities, which predictably leads to visually chaotic content. Additionally, as evidenced by the third row of Fig. 4, even when explicitly prompted to generate a young girl, Qwen-Image exhibits a persistent bias towards synthesizing adult females. Apart from our PortraitGen and Qwen-Image, all other evaluated methods exhibit severe skin oiliness. Ultimately, our proposed approach substantially suppresses common AI artifacts, particularly oily skin and structural limb distortions. Further comparative results are detailed in Appendix E.

5.3 Quantitative Results

The baselines are divided into base T2I models and FLUX.1-dev-based post-training methods, with the latter retrained on our dataset for a fair comparison. As shown in Tab. 2, PortraitGen achieves the best or second-best performance on most metrics, with notable gains in OmniReward Content and Composition as well as UnifiedReward Coherence. These results indicate improved structural plausibility and reduced synthetic artifacts. This advantage is further corroborated by the Forensic-Chat and OmniAID metrics adopted from RealGen, on which the highest scores are attained at both the semantic and feature levels.

Beyond these final-checkpoint results, Fig. 5 examines the optimization dynamics by evaluating intermediate checkpoints on the evaluation set. The Content, Composition, and Forensic-Chat scores generally improve and stabilize during training, while Clarity and Lighting & Color remain stable. To further assess portrait-specific realism, Fig. 6 reports pairwise comparisons under AI-Portrait Reward, where PortraitGen consistently outperforms the competing methods.

5.4 Ablation Study

Effect of Reward Models and Real-Image Anchoring. Tab. 3 isolates the contributions of our reward models and real-image anchoring. Neither component alone yields consistent gains: the former remains constrained by homogeneous model-generated samples, whereas the latter lacks sufficiently discriminative feedback on portrait artifacts. Their integration produces a clear complementary effect, attaining the strongest overall performance. In particular, the Forensic-Chat score increases from 32.57 to 56.62 over DanceGRPO, while the highest OmniAID score is also achieved, indicat-

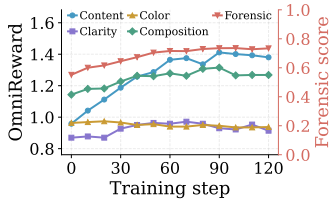

Figure 5: OmniReward and Forensic-Chat scores on the evaluation set.

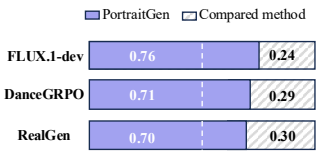

Figure 6: AI-Portrait Reward win rates of PortraitGen against competing methods.

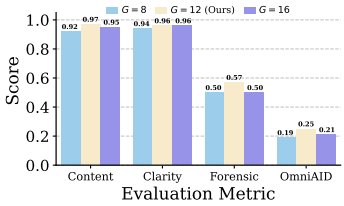

Figure 7: Effect of the GRPO group size on PortraitGen's performance.

Table 4: Generalization of the PortraitGen training paradigm to SD3.5-Medium on PortraitBench.

Method	OmniReward				RealGen	
	Cont.	Cla.	Color	Comp.	Forensic.	OmniAID
SD3.5-M	0.94	0.97	0.94	1.19	39.53	18.12
SD3.5-M + PortraitGen	0.97	0.98	0.94	1.22	46.15	18.75

Table 5: User study. Win rates (%) of PortraitGen against competing models.

	vs. FLUX	vs. DanceGRPO	vs. Qwen-Image
Semantic Align.	80.7	76.5	67.4
Aesthetics	71.4	66.8	60.9
Realism	87.5	78.2	63.1

ing substantial reductions in semantic- and feature-level artifacts. These results demonstrate that real-image exemplars provide an improved optimization target, while the proposed rewards supply the necessary guidance toward it.

Effect of Group Size. As shown in Fig. 7, $G = 12$ achieves the strongest overall performance, with particularly clear gains on the artifact-sensitive Forensic-Chat and OmniAID metrics. This behavior may be attributed to the less reliable relative comparisons induced by smaller groups, whereas excessively large groups dilute the contribution of the single real-image exemplar and thereby weaken its anchoring effect. We therefore use $G = 12$ to balance robust group-relative estimation with effective exemplar guidance.

Generalization across Generative Backbones. To assess its backbone-agnostic applicability, the proposed training paradigm is further instantiated on SD3.5-Medium. As shown in Tab. 4, this transfer yields consistent gains across nearly all OmniReward dimensions and both RealGen detectors, indicating improvements in perceptual quality and artifact suppression. The resulting gains substantiate that the effectiveness of our paradigm extends beyond the FLUX.1-dev backbone.

5.5 User Study

We recruit 52 participants for a user study comprising 100 pairwise comparisons between PortraitGen and FLUX.1-dev, DanceGRPO, or Qwen-Image. Given two outputs conditioned on the same prompt, participants indicate their preference independently in terms of semantic alignment, aesthetics, and realism. We then report the win rate of PortraitGen under each criterion. As shown in Tab. 5, PortraitGen is consistently preferred across all comparisons, with its most pronounced advantage in realism. This result aligns with our quantitative evaluation, indicating that the improvements in portrait realism and artifact suppression are also perceptible to human observers.

6 Conclusion

In this work, we introduce PortraitGen, a post-training framework that improves portrait realism by incorporating real-image trajectories into GRPO. Through BELM inversion, each exemplar is reconstructed as a reference trajectory that participates in both group-relative comparison and policy optimization, while OmniReward and AI-Portrait Reward provide complementary supervision over general visual quality and portrait-specific artifacts. We further introduce PortraitBench for evaluation across diverse portrait scenarios. Extensive experiments demonstrate that PortraitGen produces more natural portraits with fewer structural and synthetic artifacts than competing methods, and transfers effectively beyond the FLUX.1-dev backbone. These findings highlight real-image trajectories as a promising source of guidance for generative post-training. Future work will investigate domain-adaptive exemplar selection and general-purpose reward models for extending this paradigm beyond portraits, as well as temporally coherent exemplar trajectories for video generation.

AI USE STATEMENT

We use LLMs to assist with literature discovery, manuscript polishing, and selected implementation tasks. For retrieval and discovery, these tools help identify potentially relevant prior work and facilitate literature searches. All suggested references are independently verified against their original sources. They are also used to support routine and labor-intensive programming tasks, including drafting and refining evaluation scripts and repetitive data-processing utilities. All AI-assisted code is inspected, tested, and validated by the authors before use. Additionally, the tools are employed to proofread the manuscript and correct grammatical errors. All resulting text is reviewed and revised by the authors to ensure technical accuracy and consistency with the intended claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our training corpus contains approximately 10,000 portraits sourced from publicly accessible posts on Instagram, images hosted by Unsplash, and RedNote, for non-commercial academic research. We manually exclude suspected AI-generated, heavily retouched, sensitive, or otherwise inappropriate images. Nevertheless, online portrait data may retain platform-specific, geographic and demographic biases. We therefore do not claim that the corpus is fully representative, and all data use remains subject to the applicable licenses and platform terms.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility, we will publicly release our code and data required to replicate the reported experiments upon acceptance. The main paper establishes the methodological formulation and evaluation protocol, while the appendix provides the model configurations, hyperparameters, sampling settings, data-processing procedures, and reward-model construction details. We are committed to making all essential resources available under their applicable licenses, thereby enabling the research community to rigorously reproduce and extend this work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Poonam Poonam, Michael Glöckler, Alex Bäuerle, and Timo Ropinski. A survey on quality metrics for text-to-image generation. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):9464–9483, 2025. doi: 10.1109/TVCG.2025.3585077.

Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single- and multi-human image personalization, 2024. URL https://arxiv.org/abs/2408.05939.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Sujie Hu, Chubin Chen, Jiashu Zhu, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Embedding-perturbed exploration preference optimization for flow models. *arXiv preprint arXiv:2605.15803*, 2026.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback, 2025. URL https://arxiv.org/abs/2312.14925.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Xiaomin Li, Yixuan Liu, Takashi Isobe, Xu Jia, Qinpeng Cui, Dong Zhou, Dong Li, You He, Huchuan Lu, Zhongdao Wang, and Emad Barsoum. Reneg: Learning negative embedding with reward guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23636–23645, June 2025.

Yanyu Li, Xian Liu, Anil Kag, Ju Hu, Yerlan Idelbayev, Dhritiman Sagar, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Textcraftor: Your text encoder can be image quality controller. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7985–7995, June 2024.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding, 2023b. URL https://arxiv.org/abs/2312.04461.

Qian Liang, Yujia Wu, Kuncheng Li, Jiwei Wei, Shiyuan He, Jinyu Guo, and Ning Xie. Mm-r1: Unleashing the power of unified multimodal large language models for personalized image generation. *arXiv preprint arXiv:2508.11433*, 2025.

Zhichao Liao, Xiaokun Liu, Wenyu Qin, Qingyu Li, Qiulin Wang, Pengfei Wan, Di Zhang, Long Zeng, and Pingfa Feng. Humanaesexpert: Advancing a multi-modality foundation model for human image aesthetic assessment. *arXiv preprint arXiv:2503.23907*, 2025.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024a.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024b.

Fangyikang Wang, Hubery Yin, Yue-Jiang Dong, Huminhao Zhu, Hanbin Zhao, Hui Qian, Chen Li, et al. Belm: Bidirectional explicit linear multi-step sampler for exact inversion in diffusion models. *Advances in Neural Information Processing Systems*, 37:46118–46159, 2024.

Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025a.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025b.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025c.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

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

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models, 2025. URL https://arxiv.org/abs/2410.02712.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Yixian Xu, Kaiyuan Gao, Yuxiang Chen, Yilei Chen, Zecheng Tang, Zihao Liu, Zikai Zhou, Deqing Li, Hao Meng, Kuan Cao, et al. Qwen-image-2.0-rl technical report. *arXiv preprint arXiv:2606.27608*, 2026.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Junyan Ye, Leiqi Zhu, Yuncheng Guo, Dongzhi Jiang, Zilong Huang, Yifan Zhang, Zhiyuan Yan, Haohuan Fu, Conghui He, and Weijia Li. Realgen: Photorealistic text-to-image generation via detector-guided rewards. *arXiv preprint arXiv:2512.00473*, 2025.

Rui Yu, Shenghua Wan, Yucen Wang, Chen-Xiao Gao, Le Gan, Zongzhang Zhang, and De-Chuan Zhan. Reward models in deep reinforcement learning: A survey, 2025. URL https://arxiv.org/abs/2506.15421.

A More Implementation Details

A.1 Dataset Details

Image Captioning. The training dataset incorporates captions synthesized using the Qwen3-VL-235B-A22B-Instruct model Bai et al. (2025), categorized into three distinct granularities to capture varying levels of semantic density:

- Keywords: Comprising 4–8 non-repetitive, high-entropy descriptive tags.
- Short-form Captions: Concise and coherent narratives ranging from 50 to 80 words.
- Mid-form Captions: Detailed and comprehensive descriptions, typically spanning 100 to 150 words.

During training, we implement a strategic sampling approach tailored to the architectural constraints of our backbone. Given that the CLIP-L/14 text encoder Radford et al. (2021) in FLUX.1-dev Labs (2024) has an effective maximum sequence length of 77 tokens, we adopt a sampling ratio of $0.1 : 0.8 : 0.1$ for Keywords, Short-form, and Mid-form captions, respectively. This distribution prioritizes short-form captions to ensure high information density while remaining compatible with the encoder's token limit, thereby minimizing potential semantic truncation.

A.2 Evaluation Details

To ensure a comprehensive assessment, we employ OmniReward and AI-Portrait Reward alongside UnifiedReward Wang et al. (2025c), PickScore Kirstain et al. (2023), and the Forensic-Chat and OmniAID metrics introduced by RealGen Ye et al. (2025). Forensic-Chat and OmniAID assess synthetic artifacts at the semantic and feature levels, respectively. For UnifiedReward, we use the UnifiedReward-2.0-qwen3vl-32b variant with its default configuration. All test samples are drawn from PortraitBench. Notably, since AI-Portrait Reward is designed for pairwise image scoring, it is inherently unsuitable for reporting absolute scores for individual models in Tab. 2. Consequently, we present these results in Fig. 6 using a win-rate format. To ensure reproducibility and maintain a fair comparison across all experiments, we utilize a consistent random seed for all inference processes.

A.3 Experimental Settings

To ensure a fair and consistent comparison, all quantitative results in this section are reproduced using our proprietary training dataset. The key hyperparameters and configurations for each method is shown in Tab. 6:

PortraitGen. During training, samples retain the aspect ratio of their corresponding exemplars, with the longer edge capped at 1024 pixels for computational efficiency. Each group contains 12 samples generated using 16 sampling steps. At inference time, we use 28 sampling steps, a resolution of 1024, and a classifier-free guidance scale of 3.5.

DanceGRPO. Training is conducted on 16 H20 GPUs. The model is fine-tuned using LoRA Hu et al. (2022) with a rank of 128 and an alpha of 256. We set the learning rate to 3×10^{-4} and employ a batch size of 1 with 4 gradient accumulation steps. For each group of 12 samples, images are generated at a resolution of 1024×1024 using 16 sampling steps. HPSv2.1 serves as the reward model for this configuration.

Pref-GRPO. Consistent with DanceGRPO Xue et al. (2025), Pref-GRPO Wang et al. (2025b) is trained on 16 H20 GPUs by fine-tuning LoRA parameters while freezing the remaining network. We use a batch size of 1, a learning rate of 1×10^{-5}, and 4 gradient accumulation steps. The group size is set to 8, and the sampling steps for samples within each group remain at the default value of 25. The reward mechanism combines a UnifiedReward-Think model Wang et al. (2025a) and CLIP Radford et al. (2021) with a weighted ratio of $0.4 : 0.6$, respectively.

RealGen. RealGen Ye et al. (2025) is trained on 8 H20 GPUs with a learning rate of 3×10^{-4}. We fine-tune the model using LoRA with a rank of 64 and an alpha of 128. The training protocol incorporates a group size of 16, 12 gradient accumulation steps, and 28 sampling steps. To steer the optimization process, we integrate three specialized reward models:

Table 6: Hyperparameter settings and reward model selections for DanceGRPO, Pref-GRPO, and RealGen. "Grad. Accum. Steps" denotes gradient accumulation steps.

Parameters	**DanceGRPO**	**Pref-GRPO**	**RealGen**
GPUs (H20)	16	16	8
Learning Rate	3×10^{-4}	1×10^{-5}	3×10^{-4}
Batch Size	1	1	1
Grad. Accum. Steps	4	4	12
Group Size	12	8	16
Sampling Steps	16	25	28
LoRA Rank	128	128	64
LoRA Alpha	256	256	128
Resolution	1024×1024	1024×1024	1024×1024
Reward Models	HPSv2.1	UnifiedReward, CLIP	Forensic-Chat, OmniAID, Long-CLIP
Reward Weights	1.0	$0.4 : 0.6$	$1 : 1 : 1$

- **Forensic-Chat**: Evaluates image authenticity by analyzing visual cues such as skin texture, facial/hand artifacts, and unnatural background blur.
- **OmniAID**: Detects feature-level artifacts, primarily focusing on frequency anomalies and abnormal noise patterns.
- **Long-CLIP**: Maintains text-image alignment to prevent the model from sacrificing prompt fidelity while pursuing enhanced realism.

B Reward-Model Training Data

B.1 Supervision Formats

Figure 8 presents the supervision formats used by our two reward models. OmniReward is trained to produce a visual rationale followed by dimension-specific scores, whereas AI-Portrait Reward directly predicts pairwise preferences for AI artifacts and skin oiliness.

B.2 OmniReward Data Annotation Strategy

We collect a large-scale dataset comprising approximately one million real-world images. To ensure high-quality labels, we recruit professional designers as expert annotators. The custom annotation interface is illustrated in Fig. 9. Specifically, each image is evaluated across four primary dimensions: Content, Clarity, Lighting & Color, and Composition. Each dimension is graded on a 3-point scale: 0, 1, and 2. For a score of 0, annotators are required to select one or more specific attribute tags to identify the exact artifacts present in the image. Given that Content is the primary focus of our dataset, reason tags are also mandatory for a score of 1 in this category, whereas for other dimensions, reasons are only required for a score of 0. The detailed scoring criteria are as follows:

Content: Evaluates the intrinsic elements within the image, independent of photographic technique.

- *Score 0 (Unacceptable):* Contains severe defects such as heavy pixelation, sensitive/harmful content, or significant AI-generated artifacts. The latter are flagged only when they violate physical logic (e.g., distorted limbs, floating objects, or nonsensical text).
- *Score 1 (Tolerable):* Includes minor flaws such as small watermarks, slight stitching artifacts, or suspected AI generation that, while present, do not severely compromise the image's overall utility.
- *Score 2 (Optimal):* High-quality content with a clearly defined subject and a complete absence of the aforementioned defects.

Clarity: Measures the precision of details and visual sharpness, reflecting the impact of blur or noise.

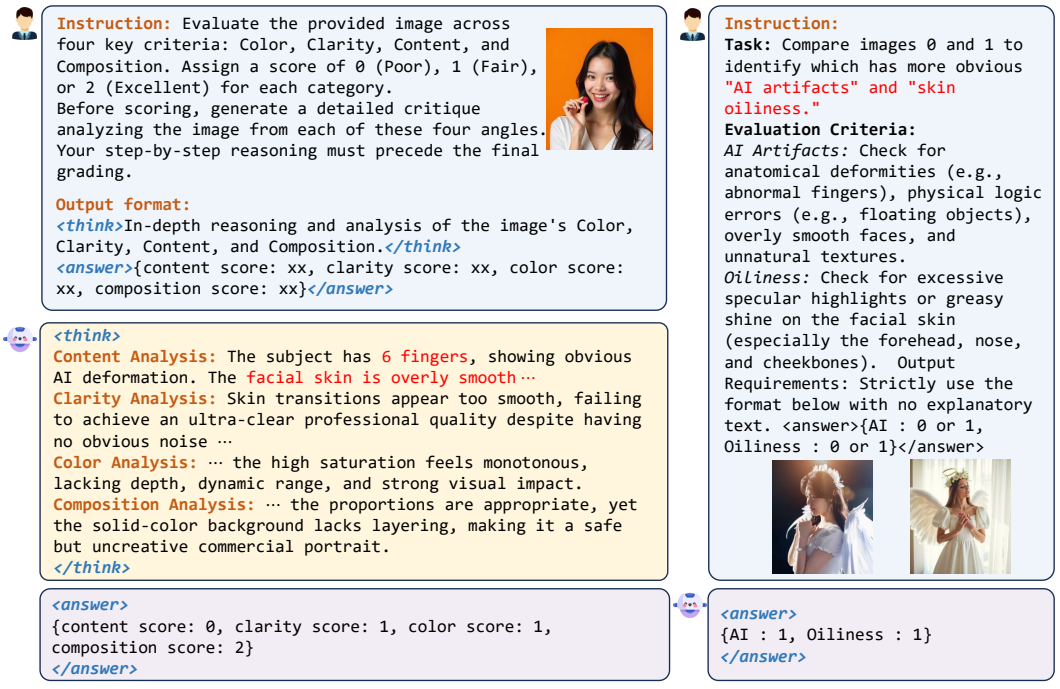

Figure 8: Training examples for OmniReward and AI-Portrait Reward. The OmniReward rationale is truncated for brevity.

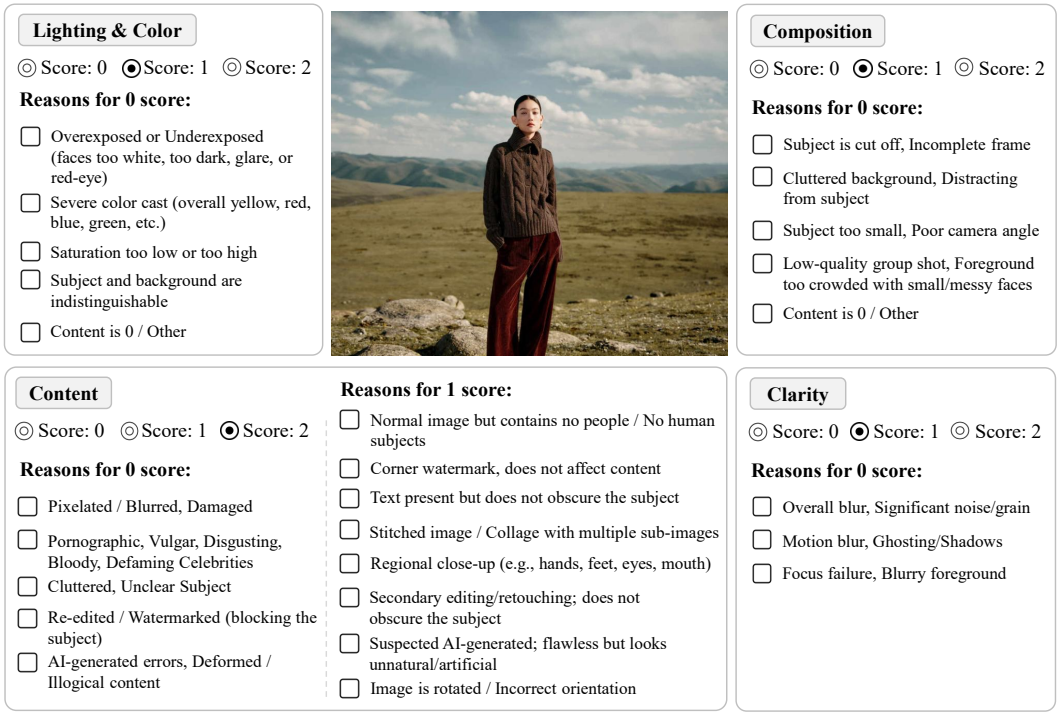

Figure 9: Data annotation interface for OmniReward. Annotators evaluate images across four dimensions. When a score of 0 is assigned, specific reason tags must be selected to justify the low-quality rating based on the visual content.

- *Score 0:* Characterized by severe motion blur, heavy grain, or out-of-focus subjects (particularly faces).
- *Score 1:* Exhibits acceptable clarity where the subject remains recognizable.
- *Score 2:* Demonstrates professional-grade sharpness where fine textures (e.g., hair, fabric) and edges are clearly discernible.

Lighting & Color: Assesses the light distribution and color harmony.

- *Score 0:* Significant over/underexposure (e.g., "blown-out" faces), severe color casts, or poor contrast between the subject and background.
- *Score 1:* Natural lighting and color that avoid Score 0 failures but lack professional polish.
- *Score 2:* Balanced exposure and harmonious tones characteristic of professional photography.

Composition: Refers to the spatial arrangement of subjects and background interference.

- *Score 0:* Poor framing where the subject is improperly cropped, too small, or obscured by a cluttered and distracting background.
- *Score 1:* Standard composition with no major structural flaws.
- *Score 2:* Excellent framing that follows aesthetic principles to effectively emphasize the primary subject.

B.3 Data Filtering and Balancing

The raw annotations exhibit a pronounced concentration around acceptable and high-quality scores. To prevent the reward model from learning this empirical bias, we construct a balanced training subset across the three score levels for each dimension, with particular emphasis on Content and Clarity because they are most directly associated with structural artifacts and perceptual degradation. This procedure yields 93,465 training samples.

B.4 Construction of Thinking Process

Beyond predicting quantitative scores, OmniReward is designed to provide interpretable visual evidence for its assessments. We therefore augment the standard image-score pairs with explicit reasoning traces. However, due to annotation costs, human feedback is limited to coarse-grained, multiple-choice failure tags for low-scoring samples and provides no descriptive rationale for higher-scoring images. To obtain comprehensive supervision, we employ Qwen3-VL-235B-A22B-Instruct to generate detailed rationales conditioned on the image, human-assigned scores, and available failure tags. The human tags serve as semantic anchors, while direct visual grounding enables the model to localize specific defects and articulate the strengths of higher-quality samples. The resulting rationale is serialized before the four final scores and used as the target response during OmniReward training.

B.5 Cross-Generator Generalization of AI-Portrait Reward

The synthetic negatives used to train AI-Portrait Reward are generated by FLUX.1-dev, Z-Image, and Qwen-Image. Since our primary generator is initialized from FLUX.1-dev, an important question is whether the reward model merely recognizes generator-specific artifacts rather than transferable indicators of synthetic appearance. We therefore conduct an additional evaluation on portraits generated by SD3.5-Medium and SDXL, neither of which is included in the reward-model training data. Following the original pairwise evaluation protocol, AI-Portrait Reward distinguishes each synthetic portrait from its real counterpart with respect to AI-feel and skin oiliness.

As shown in Tab. 7, AI-Portrait Reward achieves average accuracies of 96.80% and 97.35% on SD3.5-Medium and SDXL, respectively, comparable to its 96.35% accuracy on the original test distribution reported in Tab. 1. These results indicate that the model is not limited to recognizing FLUX.1-dev-specific signatures. Training with negatives from multiple generators encourages it to



Figure 10: Visualization of image reconstruction using BELM. The reported MSE values indicate a negligible pixel-level discrepancy between the original and reconstructed frames. This high-fidelity performance confirms that BELM enables near-lossless inversion, facilitating precise latent retrieval for the subsequent GRPO process.

Table 7: Cross-generator accuracy (%) of AI-Portrait Reward on synthetic portraits from held-out generators. Neither SD3.5-Medium nor SDXL is used to construct its training data.

Synthetic source	AI-feel Acc.	Oiliness Acc.	Avg.
SD3.5-Medium	96.90	96.70	96.80
SDXL	97.30	97.40	97.35

capture shared portrait artifacts, such as unnatural skin appearance and structural inconsistencies, which remain detectable in previously unseen generator distributions.

C Reconstruction of BELM

We incorporate exemplar images into GRPO groups via image inversion using BELM Wang et al. (2024). Due to its nearly lossless nature, BELM enables the extraction of precise latents and transition probabilities for exemplar images at each timestep. Unlike standard sampling, this process involves a noise-denoise cycle to retrieve the initial latent corresponding to the source image. As illustrated in Fig. 10, our approach yields a remarkably low Mean Squared Error (MSE). Notably, this high-fidelity reconstruction is achieved in just 16 sampling steps, aligning with the step count used for other samples within the group while ensuring near-perfect reconstruction.

D Analysis and Discussion

Choice of Exemplar Ratio. For $G = 12$, we use $K = 1$ real exemplar and 11 model-generated trajectories. This choice follows naturally from two considerations. At the data level, all trajectories in a GRPO group must share the same text condition, whereas real portraits are typically unique.

Constructing $K \geq 2$ groups would therefore require multiple real images with nearly identical semantics and attributes, which is difficult to satisfy reliably at scale. At the reward level, ranking multiple authentic images by the severity of AI artifacts is conceptually ill-defined and may introduce ambiguous supervision. Thus, $K = 1$ provides an unambiguous real-image anchor while preserving sufficient self-sampled trajectories.

Generality beyond Portrait Generation. Portrait generation serves as a rigorous instantiation of our paradigm, not its scope. Exemplar-driven GRPO is reward-agnostic: the exemplar acts as a target-distribution anchor that injects external evidence unavailable from self-sampled groups, thereby extending optimization beyond the model's current rollout distribution. Since the anchor represents the desired optimization target, the framework can be transferred by replacing both the exemplars and rewards with domain-specific counterparts, such as physical realism for general image generation or temporal consistency for video generation. PortraitGen therefore establishes a proof of concept for principled external anchoring, rather than a portrait-specific post-training recipe.

E More Visualization Results

Figs. 11-13 present more qualitative comparisons between our proposed PortraitGen and several representative baselines, highlighting a substantial enhancement over the FLUX.1-dev base model. While the base model frequently suffers from unnatural specular highlights and a characteristic oily facial texture, our approach consistently achieves superior photorealistic fidelity. Although Qwen-Image Wu et al. (2025) demonstrates commendable aesthetic performance on Asian subjects, it often exhibits structural degradation and anatomical inconsistencies when synthesizing other ethnicities. Furthermore, other post-training methods such as DanceGRPO, Pref-GRPO, and RealGen often produce images with an oily skin texture. As shown in the second case of Fig. 12, these baselines sometimes struggle to maintain the requested style, they may generate anime-like results even when a photography style is explicitly specified in the prompt.



A blonde Caucasian woman in yellow floral pajamas lies on a sofa with an Irish Wolfhound yawning on her lap. Her hand rests on the dog's back. A bookshelf and green curtain are visible in the background. Photography style.



A bride in an off-shoulder white wedding dress with ruffled sleeves, wearing a veil and a floral headpiece. She holds red roses, with a candelabra and lit candles in the background. Photography style.



A man in a studio portrait, wearing a black leather jacket and holding a NARS lipstick. He gazes off-camera against a solid terracotta background. Photography style.

Figure 11: Qualitative comparison between PortraitGen and five competitive baselines. Red dashed circles highlight artifacts such as limb deformities. Overall, PortraitGen significantly mitigates common AI artifacts, particularly reducing over-smoothed (oily) skin textures and improving structural integrity of limbs.



An elderly woman with short gray hair and tortoiseshell glasses sits on a wooden chair, wearing a black turtleneck and a single pearl earring. Her left arm rests on the chair's back, and she wears a white watch. Photography style.



A Caucasian girl with a brown bob haircut and a yellow hair clip on the right side, wearing a white collar with a black bow. She is looking slightly to the left against a dark background. Realistic photography style.



A Caucasian girl with a brown bob haircut and a yellow hair clip on the right side, wearing a white collar with a black bow. She is looking slightly to the left against a dark background. Realistic photography style.

Figure 12: Qualitative comparison between PortraitGen and five competitive baselines. Red dashed circles highlight artifacts such as limb deformities. Overall, PortraitGen significantly mitigates common AI artifacts, particularly reducing over-smoothed (oily) skin textures and improving structural integrity of limbs.



A man sits on the floor holding black cables, wearing a red and black leather jacket with white stripes, ripped light-blue jeans, black gloves, and sunglasses on his head. The background includes black monitor stands and tangled wires. Photography style.



A woman with dark wavy hair and bangs applies lip liner with a brush while holding a makeup compact. She wears a white long-sleeve blouse. The background shows a window with green foliage outside. Realistic style.



A Caucasian man with curly hair wearing a double-breasted suit, white shirt, and striped tie. He holds a coat over his left arm and wears rings on his right hand. The background is a paneled door. Photography style.

Figure 13: Qualitative comparison between PortraitGen and five competitive baselines. Red dashed circles highlight artifacts such as limb deformities. Overall, PortraitGen significantly mitigates common AI artifacts, particularly reducing over-smoothed (oily) skin textures and improving structural integrity of limbs.