

# BeautyGRPO: Aesthetic Alignment for Face Retouching via Dynamic Path Guidance and Fine-Grained Preference Modeling

## Supplementary Material

### A. Mathematical Derivations for Dynamic Path Guidance

#### A.1. From Flow-Matching ODE to SDE (Flow-GRPO)

**Flow-Matching ODE.** In flow-matching generative models, a deterministic ordinary differential equation (ODE) is used to transport a simple noise distribution  $p_1(x)$  at time  $t = 1$  to the data distribution  $p_0(x)$  at  $t = 0$  along a continuous trajectory. The model learns a time-dependent vector field  $v_\theta(x, t)$  such that integrating

$$dx_t = v_\theta(x_t, t) dt \quad (11)$$

evolves  $x_t$  from pure noise at  $t = 1$  to a realistic sample  $x_0$  at  $t = 0$ . Strictly following this learned ODE trajectory typically yields high-fidelity images. However, a deterministic path is incompatible with exploration in a reinforcement learning (RL) setting: we need stochasticity to explore alternative outcomes and obtain diverse reward signals.

**ODE-to-SDE conversion (FlowGRPO).** FlowGRPO [21] converts the deterministic ODE into a stochastic differential equation (SDE) while preserving the marginal distributions of the original ODE at every time  $t$ . Specifically, it injects Gaussian noise at each sampling step in exploration process:

$$dx_t = \left[ v_\theta(x_t, t) + \frac{\sigma_t^2}{2t} (x_t + (1-t)v_\theta(x_t, t)) \right] dt + \sigma_t d\omega_t, \quad (12)$$

where  $d\omega_t$  denotes the increment of a standard Wiener process, which can be approximated as  $d\omega_t \approx \sqrt{dt} \epsilon_t$ , with  $\{\epsilon_t\}$  i.i.d.  $\mathcal{N}(0, I)$  across time steps, and  $\sigma_t = \eta \sqrt{\frac{t}{1-t}}$  controls the noise magnitude at time  $t$ .

Here the SDE in Eq. (12) is formulated in reverse time (from  $t = 1$  to  $t = 0$ ). For a small timestep  $\Delta t$ , the Euler-Maruyama discretization of Eq. (12) yields the conditional mean

$$\mu_t = x_t + \left[ v_\theta(x_t, t) + \frac{\sigma_t^2}{2t} (x_t + (1-t)v_\theta(x_t, t)) \right] \Delta t, \quad (13)$$

with a step-wise standard deviation

$$\sigma_{\text{step}} = \sigma_t \sqrt{\Delta t}. \quad (14)$$

The one-step update can therefore be written as

$$x_{t-\Delta t} = \mu_t + \sigma_{\text{step}} z_t, \quad z_t \sim \mathcal{N}(0, I), \quad (15)$$

where  $z_t$  is a standard normal random vector. Eqs. (13)–(15) show that each update is centered at the model prediction  $\mu_t$  and then perturbed by isotropic Gaussian noise, so iterating this stochastic update produces a distribution over possible final states  $x_0$  rather than a single deterministic outcome, which in turn enables the policy to explore multiple image variants under the same condition and obtain diverse reward feedback in the RL setting.

#### A.2. Fidelity–Exploration Conflict

While the SDE in FlowGRPO provides the stochasticity needed for online RL, it conflicts with the high-fidelity requirements of face retouching. From the discrete update in Eq. 15, each noise sample  $z_t$  perturbs the state away from the model-predicted mean  $\mu_t$ , and the cumulative term

$$\sum_t \sigma_{\text{step}}(t) z_t$$

over  $T$  sampling steps (we discretize the exploration process into  $T$  steps) gradually drives the trajectory  $\{x_t\}$  off the high-quality image manifold. Here  $\sigma_{\text{step}}(t) = \sigma_t \sqrt{\Delta t}$  denotes the step-wise standard deviation at time  $t$ . In generic text-to-image generation this drift can be tolerable or even helpful for diversity, but in face retouching it mainly appears as noise artifacts. Reducing  $\sigma_t$  mitigates drift but also suppresses exploration, limiting the ability of RL to find higher-reward solutions. We therefore introduce a mechanism that keeps the trajectory close to a high-fidelity manifold while still enabling controlled stochastic exploration.

#### A.3. Dynamic Path Guidance (DPG)

To resolve the fidelity–exploration conflict, we introduce Dynamic Path Guidance (DPG), which performs anchored exploration by reshaping the noise in the FlowGRPO SDE so that the sampling trajectory remains close to a high-fidelity reference path while still preserving stochasticity.

**Anchor-guided target.** We assume access to a high-quality anchor sample  $x_0^{\text{anchor}}$  that lies on the desired high-fidelity manifold (e.g., a top-rated retouched face from our preference dataset). At sampling time  $t$  with current state  $x_t$ , we construct a straight-line path connecting  $x_t$  to  $x_0^{\text{anchor}}$  over the remaining sampling-time horizon. The corresponding target at the next time  $t - \Delta t$  on this anchor-guided path is defined as

$$x_{t-\Delta t}^* = \left(1 - \frac{\Delta t}{t}\right) x_t + \frac{\Delta t}{t} x_0^{\text{anchor}}, \quad (16)$$

which is a linear interpolation that moves a fraction  $\Delta t/t$  of the distance from  $x_t$  toward  $x_0^{\text{anchor}}$ . Geometrically,  $x_{t-\Delta t}^*$  lies on the segment joining  $x_t$  and  $x_0^{\text{anchor}}$  and specifies an anchor-consistent reference direction for the next update. In practice,  $t$  is discretized into  $T$  sampling timesteps in the exploration process and we only apply the anchor-guided update for  $t > 0$ , so the division by  $t$  is well-defined and the final step at  $t = 0$  is treated as the terminal state.

**Anchor correction noise.** Recall that the FlowGRPO discretization in Eq. (15) can be written as

$$x_{t-\Delta t} = \mu_t + \sigma_{\text{step}} z_t, \quad z_t \sim \mathcal{N}(0, I), \quad (17)$$

where  $\mu_t$  is given by Eq. (13) and  $\sigma_{\text{step}} = \sigma_t \sqrt{\Delta t}$ . Conditioned on  $x_t$ , this induces the Gaussian transition

$$x_{t-\Delta t} | x_t \sim \mathcal{N}(\mu_t, \sigma_{\text{step}}^2 I). \quad (18)$$

The noise that maps the next state exactly to the anchor-guided target  $x_{t-\Delta t}^*$  in Eq. (16) is obtained by enforcing

$$x_{t-\Delta t}^* = \mu_t + \sigma_{\text{step}} z_t, \quad (19)$$

which yields the anchor correction vector

$$z_t^{\text{anchor}} = \frac{x_{t-\Delta t}^* - \mu_t}{\sigma_{\text{step}}}. \quad (20)$$

The vector  $z_t^{\text{anchor}}$  in Eq. (20) precisely maps the next state onto the anchor-guided target  $x_{t-\Delta t}^*$ . However, if this correction is applied deterministically at every step, i.e.,  $z_t = z_t^{\text{anchor}}$  for all  $t$ , then  $x_{t-\Delta t} = x_{t-\Delta t}^*$  holds throughout the sampling process, causing the trajectory to collapse onto the deterministic anchor path in Eq. (16) and leaving no space for stochastic exploration.

**Mixed noise and time-dependent guidance.** To retain stochasticity, DPG replaces the original standard Gaussian  $z_t$  with a time-dependent mixture of the anchor correction and an independent standard normal sample. Let  $z_t^{\text{std}} \sim \mathcal{N}(0, I)$  be independent of  $x_t$  and define

$$z_t^{\text{mix}} = \lambda_t z_t^{\text{anchor}} + (1 - \lambda_t) z_t^{\text{std}}, \quad 0 \leq \lambda_t \leq 1, \quad (21)$$

where  $\lambda_t$  is a time-dependent guidance weight. In practice, we discretize the reverse-time sampling process into  $T$  steps indexed by integers  $t \in \{0, \dots, T-1\}$ , where  $T$  denotes the total number of steps in the exploration process. We adopt the same linear schedule as in the main text,

$$\lambda_t = \frac{t}{\max(1, T-1)}. \quad (22)$$

Sampling proceeds in reverse time from  $t = T-1$  (early, high-noise steps) down to  $t = 0$  (late refinement), so the effective guidance weight  $\lambda_t$  is large at early sampling timesteps and gradually decreases to 0 toward the final refinement steps. For notational simplicity, we continue to write

$\lambda_t$  in the derivations below to denote this time-dependent weight. We then replace  $z_t$  in the discretization above by  $z_t^{\text{mix}}$  and obtain the DPG transition

$$x_{t-\Delta t} = \mu_t + \sigma_{\text{step}} z_t^{\text{mix}}. \quad (23)$$

Conditioned on  $x_t$ , the anchor correction  $z_t^{\text{anchor}}$  is a deterministic function of  $(x_t, x_0^{\text{anchor}})$  and the only stochastic component is  $z_t^{\text{std}}$ . Taking conditional expectation with respect to  $z_t^{\text{std}}$  and using  $\mathbb{E}[z_t^{\text{std}}] = 0$ , we obtain

$$\mathbb{E}[x_{t-\Delta t} | x_t] = \mu_t + \sigma_{\text{step}} \mathbb{E}[z_t^{\text{mix}} | x_t] \quad (24)$$

$$= \mu_t + \sigma_{\text{step}} \lambda_t z_t^{\text{anchor}} \quad (25)$$

$$= \mu_t + \lambda_t (x_{t-\Delta t}^* - \mu_t) \quad (26)$$

$$= (1 - \lambda_t) \mu_t + \lambda_t x_{t-\Delta t}^*. \quad (27)$$

Thus, in expectation, DPG interpolates the transition mean between the original FlowGRPO mean  $\mu_t$  and the anchor-guided target  $x_{t-\Delta t}^*$ , with interpolation weight  $\lambda_t$ .

Similarly, still conditioning on  $x_t$  and treating  $z_t^{\text{anchor}}$  as deterministic, the covariance of  $z_t^{\text{mix}}$  is

$$\text{Cov}[z_t^{\text{mix}} | x_t] = (1 - \lambda_t)^2 I, \quad (28)$$

since  $z_t^{\text{anchor}}$  has zero conditional variance and  $\text{Cov}[z_t^{\text{std}}] = I$ . This implies

$$x_{t-\Delta t} | x_t \sim \mathcal{N}\left((1 - \lambda_t) \mu_t + \lambda_t x_{t-\Delta t}^*, ((1 - \lambda_t) \sigma_{\text{step}})^2 I\right). \quad (29)$$

When  $\lambda_t = 0$ , Eq. (29) reduces exactly to the original FlowGRPO kernel with mean  $\mu_t$  and variance  $\sigma_{\text{step}}^2 I$ . When  $\lambda_t = 1$ , the variance collapses to zero and the update becomes a deterministic step to  $x_{t-\Delta t}^*$ . For intermediate values  $0 < \lambda_t < 1$ , the trajectory is biased toward the anchor path while retaining a reduced but nonzero variance.

**Effect on fidelity and exploration.** Eqs. (27) and (29) show that DPG simultaneously (i) pulls the mean transition toward an anchor-consistent direction and (ii) shrinks the step variance by a factor of  $(1 - \lambda_t)^2$ . By choosing  $\lambda_t$  large when  $t$  is large and the state is still highly noisy, the sampling trajectory in the exploration process is tightly constrained around the high-fidelity anchor manifold during the early, structure-forming stages. As  $t \rightarrow 0$  and  $\lambda_t \rightarrow 0$ , the variance gradually recovers to its original level, allowing finer stochastic exploration on top of an already well-structured face. In this way, DPG reconciles the fidelity requirements of portrait retouching with the exploration needs of online reinforcement learning, purely through a sampling-time modification of the FlowGRPO transition kernel without changing the underlying flow-matching policy  $v_\theta$ .

## B. Anchored Exploration with Standard GRPO Optimization

It is important to emphasize that the anchor image is employed exclusively during the sampling-based exploration

phase that generates trajectories for GRPO, and it is never used as a supervision target in the optimization objective. During the exploration phase, for each input portrait in FRPref-10K, we select a high-quality retouched image as the anchor  $x_0^{\text{anchor}}$  solely to construct the DPG-guided sampling path. Specifically, at each sampling timestep  $t$ , DPG computes the anchor-guided target  $x_{t-\Delta t}^*$  and the corresponding correction noise  $z_t^{\text{anchor}}$  (Eqs. (16)–(20)), forms the mixed noise  $z_t^{\text{mix}}$  in Eq. (21), and updates  $x_{t-\Delta t}$  via Eq. (23). This anchored exploration only changes how trajectories  $\{x_t\}$  are sampled during RL rollouts and does not add any reconstruction or regression term toward the anchor in the loss.

In the optimization phase, we apply the standard GRPO objective to the policy model  $\pi_\theta$ . The reward model scores the final retouched outputs  $x_0$ , and these scores are converted into advantages to optimize the policy via the clipped GRPO loss. This optimization process is entirely independent of the anchor image. In particular, the anchor image  $x_0^{\text{anchor}}$  is never used as a pixel-level or feature-level target, and there is no supervised alignment between the policy output and the anchor. At inference time, DPG is discarded entirely, and we utilize standard ODE sampling. This ensures that the inference process requires no reference images and incurs no additional computational cost.

### C. Relation to Reference-Guided Methods, CFG, and Path Regularization

Although DPG uses an anchor image during the exploration phase, its role and implementation are fundamentally different from typical reference-guided editing, classifier-free guidance (CFG), or path regularization.

**Distinction from Reference-Guided Editing.** Conventional reference-based methods explicitly inject the reference image into the model architecture as an additional conditioning signal (e.g., via concatenation, cross-attention, or style codes), often training the network to strictly imitate the reference appearance through supervised losses. In contrast, our anchor image is never passed through the retouching network as an input condition and never appears in the optimization objective. Instead, it serves purely as a geometric anchor to define an anchor-based ODE path in the latent space and to compute the correction noise  $z_t^{\text{anchor}}$  within the sampler. The policy model  $\pi_\theta$  remains conditioned solely on the original input portrait and textual prompt, learning aesthetic alignment from reward signals rather than through direct supervision against the anchor.

**Distinction from Classifier-Free Guidance (CFG).** CFG modifies the sampling dynamics by adding a guidance term proportional to the difference between conditional and unconditional model outputs. This effectively steers the generation toward regions with higher likelihood under the condition. In contrast, DPG operates without an external classifier or dual-forward passes. Instead of altering the

Table 5. Distortion vs. perceptual quality metrics on FFHQ.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	MUSIQ $\uparrow$	MANIQA $\uparrow$	NRQM $\uparrow$	
SUPERVISED	RetouchFormer	42.088	<b>0.997</b>	0.024	4.465	1.036	8.175
	Xu (2025) [52]	45.480	0.993	0.008	–	–	–
	RetouchGPT	<b>46.210</b>	–	<b>0.002</b>	–	–	–
GENERATIVE	ICEdit	32.463	0.882	0.152	4.192	0.978	7.424
	SeedDream	25.598	0.739	0.170	4.503	0.993	7.538
	NanoBanana	28.841	0.824	0.106	4.681	1.009	8.065
	Flux.K w/LoRA	36.001	0.956	0.063	4.459	1.035	8.009
	<b>Ours</b>	33.697	0.927	0.102	<b>4.906</b>	<b>1.079</b>	<b>8.401</b>

vector field, DPG reshapes the stochastic transition kernel by linearly mixing a deterministic anchor correction with standard Gaussian noise via a time-dependent weight  $\lambda_t$ . This simultaneously biases the transition mean toward the anchor-guided direction (Eq. (27)) and shrinks the variance (Eq. (29)), keeping the trajectory on a high-fidelity manifold during early sampling while still allowing exploration. All of this is achieved without requiring auxiliary guidance networks or modifying the drift prediction.

**Distinction from Path Regularization.** Path regularization methods typically enforce that the learned vector field or sampling trajectory stay close to a predefined reference path by adding explicit regularization terms to the training loss, penalizing deviations over the entire trajectory. Our DPG does not add any path penalty or consistency term to the loss function. The anchor-based ODE path is used only at sampling time to construct  $x_{t-\Delta t}^*$  and  $z_t^{\text{anchor}}$ , and the influence of the anchor is further modulated by the schedule  $\lambda_t$  that decays along the sampling timesteps. Thus, DPG acts as a sampling-time constraint on exploration rather than a training-time path regularizer. It regularizes the rollout distribution from which GRPO collects experience, while the GRPO objective itself remains unchanged.

**Summary.** DPG distinguishes itself from these methods in that: (i) the anchor image is neither a supervised target nor an explicit conditioning input; (ii) no auxiliary guidance network or classifier is employed; and (iii) no regularization terms are added to the training loss. Instead, DPG is a lightweight exploration-time mechanism that reshapes the SDE sampling kernel using anchor-induced geometry, enabling high-fidelity yet exploratory RL updates while keeping inference-time sampling identical to standard ODE-based flow matching.

## D. Evaluation Metrics

### D.1. Why Not Use Full-Reference Metrics

Previous face retouching works [19, 43, 45, 53] typically adopt classical full-reference (FR) metrics, such as PSNR, SSIM, and LPIPS, to assess pixel-level fidelity between the retouched output and a ground-truth reference. However, we maintain that FR metrics are fundamentally ill-suited for

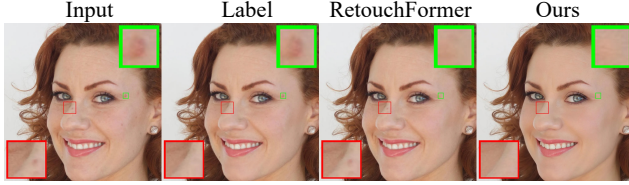


Figure 8. Visual examples of imperfect ground-truth references in the FFHQ dataset, which are actively corrected by our method.

evaluating aesthetic enhancement tasks like face retouching, supported by both theoretical and empirical evidence:

- **Theoretical Conflict (The Perception-Distortion Tradeoff).** As theoretically justified by Blau and Michaeli [5], there is an inherent mathematical conflict between maximizing perceptual quality and minimizing pixel-level distortion (e.g., PSNR). Specifically, models strictly minimizing distortion tend to average out high-frequency details, often resulting in overly smoothed and perceptually flat skin textures. In contrast, synthesizing realistic facial details, such as natural pores and lighting variations, inevitably introduces pixel-wise deviations from the reference. This tradeoff is clearly evident in Table 5: compared to our supervised baseline (Flux.K w/LoRA), our RL-optimized BeautyGRPO trades pixel fidelity (PSNR dropping from 36.00 to 33.70 dB) for significant aesthetic gains (MUSIQ improving from 4.46 to 4.91). Optimizing strictly for FR metrics fundamentally restricts a model’s ability to achieve peak perceptual quality.
- **Imperfect Ground Truth and Defect Replication.** FR metrics implicitly assume that target reference images are flawless. In reality, existing datasets like FFHQ often contain ground-truth references with residual blemishes or suboptimal aesthetic qualities (see Fig. 8). Models optimizing for high PSNR are therefore forced to replicate these defects to strictly match the reference. In contrast, BeautyGRPO actively corrects these flaws, moving beyond the dataset’s inherent limitations to improve naturalness and visual appeal. Consequently, FR metrics unfairly penalize our model for deviating from these flawed references.

Consequently, we focus on no-reference (NR) metrics, which directly evaluate naturalness, realism, and aesthetic quality, providing a much more accurate and reliable reflection of true human preference in face retouching.

## D.2. No-Reference Metrics

For these reasons, we instead focus on no-reference (NR) perceptual and aesthetic metrics that better approximate human judgments in our evaluation.

- **NIQE [28] (↓).** NIQE models the statistical regularities of high-quality natural images and measures how strongly a test image deviates from them. For face retouching, it mainly reflects global perceptual naturalness, including

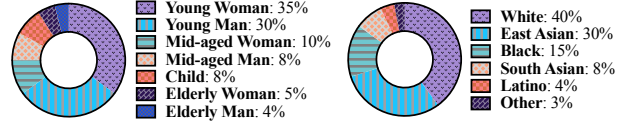


Figure 9. Age and racial distribution of the FRPref-10K dataset.

whether skin texture appears artificial, overly smooth, or oily, and whether it contains visible AI artifacts. A **lower** score indicates better perceptual naturalness.

- **NIMA [37] (↑).** NIMA predicts the distribution of human opinion scores and uses the expected rating as the aesthetic quality measure. The AVA-trained model emphasizes overall visual appeal, including composition, color harmony, and aesthetic preference. A **higher** score indicates better aesthetic quality.
- **MUSIQ [16] (↑).** MUSIQ provides a holistic assessment of perceptual quality by jointly considering texture fidelity, clarity, and overall aesthetic impression. The AVA-trained model is sensitive to global balance as well as fine-grained visual quality, making it suitable for evaluating portrait enhancements. A **higher** score indicates better perceptual quality.
- **NRQM [26] (↑).** NRQM predicts a single perceptual quality score learned from subjective studies. In face retouching, it reflects perceived sharpness, texture realism, and local structure fidelity, especially in facial regions. A **higher** score indicates better perceptual quality.
- **MANIQA [56] (↑).** MANIQA is designed to detect complex distortions and generative artifacts. It is particularly sensitive to texture integrity, fine facial detail, and subtle artifacts introduced during retouching. A **higher** score indicates better perceptual and textural quality.
- **TOPIQ [6] (↑).** TOPIQ follows a top-down procedure that emphasizes distortions in semantically important regions. Using the GFIQA-trained face model, it focuses on face-region clarity, texture naturalness, and identity preservation. A **higher** score indicates better face-specific perceptual quality.

## E. Composition and Statistics of FRPref-10K

To ensure a diverse and representative training corpus for aesthetic alignment, we carefully curated the FRPref-10K dataset. The dataset comprises a mixture of sources: 70% of the images are sampled from the official training split of the FFHQ dataset, while the remaining 30% are sourced from a proprietary high-resolution portrait collection. This hybrid approach specifically helps mitigate racial and demographic biases [27] present in FFHQ datasets. To provide a comprehensive understanding of the dataset’s diversity, we detail its demographic distribution in Fig. 9.

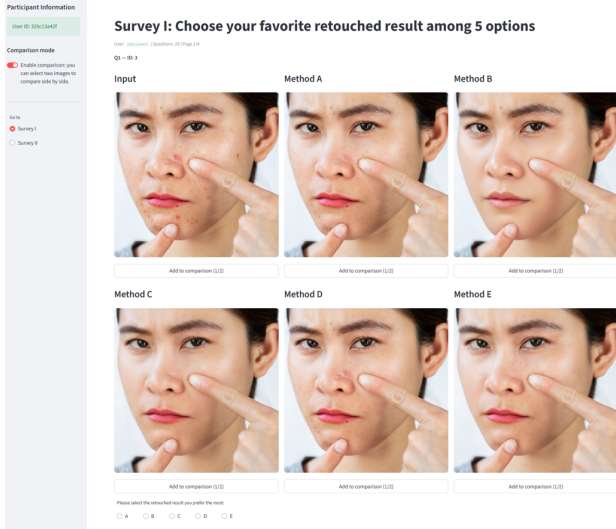


Figure 10. User interface of Questionnaire I.

## F. User Study Details

### F.1. Study Protocol

We design two complementary questionnaires to evaluate users’ perceptual preferences and to assess how well different reward models align with human judgments.

**Questionnaire I: Overall Preference.** The first questionnaire consists of a fixed set of 25 questions. Each question presents one original input portrait and five retouched results produced by different methods, including our approach and representative baselines. For each question, participants are asked to select one retouched image that they personally prefer the most. We do not impose any additional instructions or constraints, and users are encouraged to follow their first impression and choose according to their own subjective aesthetic preferences. The 25 questions cover a diverse range of ages, genders, and ethnicities, and the candidate images in each question include both clearly different cases and more subtle differences in texture and details. To reduce position bias, the order of the five candidate images within each question is randomized. The user interface of Questionnaire I is illustrated in Figure 10.

**Questionnaire II: Dimension-wise Ratings for Reward Evaluation.** The second questionnaire is designed to measure the consistency between different reward models and human judgments on fine-grained perceptual dimensions. It is constructed from an additional pool of 20 input–output pairs. For each participant, 5 questions are randomly sampled from this pool. Each question shows one original portrait and one retouched result, and users rate the result along five dimensions: skin smoothing, blemish removal, skin texture quality, clarity, and identity preservation. For each dimen-

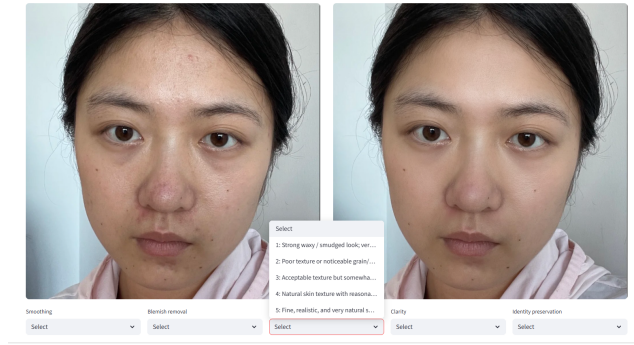


Figure 11. User interface of Questionnaire II.

sion, participants assign an integer score in the range 0–5, where a larger value indicates better perceived quality. We provide a detailed textual description and concrete examples for every score level of each dimension (from 0 to 5), so that users have a clear and consistent guideline when rating. These human scores are later compared with the outputs of different reward models to quantify their alignment with human perception. The user interface of Questionnaire II is illustrated in Figure 11.

**Data Collection and Aggregation.** In total, we obtain 100 valid questionnaires from anonymous participants. For Questionnaire I, we compute the win rate of each method as the proportion of answers in which it is selected as the preferred result over all questions and participants. For Questionnaire II, we aggregate the human scores for each dimension by averaging over all participants and questions, and then compare these averages with the corresponding scores predicted by different reward models. This allows us to evaluate dimension-wise consistency between reward models and human judgments.

### F.2. Demographic Analysis of User Preferences

Figure 12 presents the win rates for Questionnaire I, stratified by user age group and gender. Our method achieves the highest win rate in every demographic group, consistently capturing at least 50% of user preferences and reaching about 75% for women aged 18–24 and 45+ years. Competing methods obtain much smaller shares, with none exceeding 25% in any group or showing consistent dominance for any demographic. These results indicate that the preference advantage of our method is robust across ages and genders, rather than being driven by a single specific subgroup.

## G. Training Details

### G.1. Reward Model Training Details

We adopt Qwen2.5-VL-7B-Instruct [2] as our base reward model. For both SFT stages (Stage 1 and Stage 2), training

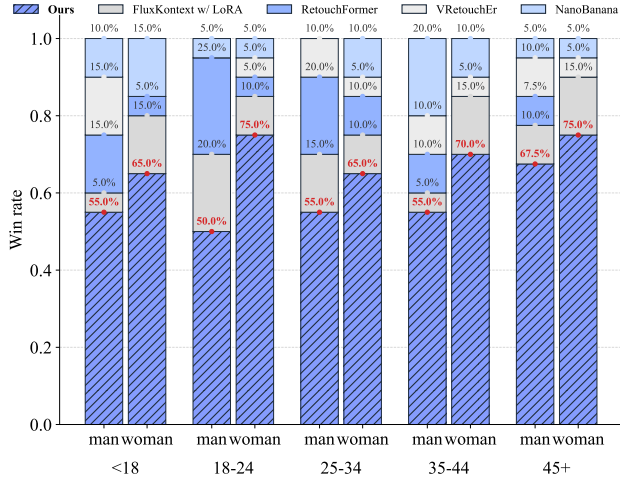


Figure 12. Comparison of user preference win rates for various face retouching methods across different ages and gender.

is performed with a per-device batch size of 1, 8 gradient accumulation steps, a learning rate of  $5 \times 10^{-5}$ , and a warm-up ratio of 0.1. For the GRPO stage, we use a per-device batch size of 2, 4 gradient accumulation steps, a learning rate of  $1 \times 10^{-6}$ , and a warm-up ratio of 0.05, with a KL penalty coefficient  $\beta = 0.04$ . The number of generated responses  $N$  is set to 8.

## G.2. Retouching Model Training Configuration

We adopt Flux.1 Kontext [18] as our base retouching model and first perform LoRA adaptation on our face retouching corpus. The model is trained at a resolution of  $1024 \times 1024$  with LoRA rank 32 and  $\alpha = 32$ , using AdamW with a learning rate of  $1 \times 10^{-5}$ , cosine learning rate decay, 500 warm-up steps, weight decay 0.01, a per-device batch size of 4, and 2 gradient accumulation steps. For BeautyGRPO, we initialize from this LoRA-adapted Flux.1 Kontext and run online RL at  $1024 \times 1024$  resolution with a per-device batch size of 2 and 8 images per prompt, using  $T = 10$  sampling steps and a noise level  $\eta = 0.7$ .

## H. CoT Reasoning Examples of Reward Model

We provide qualitative examples of our reward model’s Chain-of-Thought (CoT) reasoning process in Figure 13.

## I. More Visual Results

Figures 14–16 provide additional qualitative comparisons across different face retouching methods, further highlighting the perceptual advantages of our BeautyGRPO.

## J. Discussion and Future Work

### J.1. Anchor Selection and Quality

In our main experiments, we use the top-1 high-preference sample as the anchor  $x_0^{\text{anchor}}$  for each input, which keeps the implementation simple and ensures that the anchor lies on a high-fidelity manifold. Conceptually, DPG only requires the anchor to be a reasonably good exemplar that constrains exploration to a high-quality neighborhood while still allowing the policy to move beyond it when the reward signal suggests a better solution. Systematic studies of alternative anchor selection strategies (e.g., sampling from a top- $k$  set, diversity-aware anchors, or combining multiple anchors) are promising directions for future work.

### J.2. Scenarios Without Explicit Anchors

In this work, we focus on a practical face-retouching regime where each input portrait is accompanied by at least one high-quality retouched exemplar (e.g., curated or highly rated images in FRPref-10K), which naturally provides anchors for DPG. In scenarios where only raw inputs are available and no such exemplars exist, one could instead construct pseudo-anchors using a strong reward model, self-distilled EMA policies, or other generative priors, and then apply the same anchored-exploration principle. Exploring such anchor-free or weakly anchored settings is an interesting extension of DPG that we leave to future work.

### J.3. DPG at Inference Time

We apply DPG only during the sampling-based exploration phase of RL training. After optimization, standard ODE sampling uses the improved vector field to produce high-fidelity results without anchors or extra computation at inference time. In principle, one could design inference-time variants of DPG by supplying reference images or synthesizing anchors on the fly, but this would introduce additional assumptions and complexity. We therefore adopt pure ODE sampling for deployment and leave inference-time extensions of DPG as future work.



Input





Image1





Image2



You are a professional portrait beauty enhancement evaluation expert with expertise in semantic image understanding, facial structure analysis, and aesthetic judgment.

Your task is: Based on the user-provided three images: one original image (Input) and two beautified result images (Image1 and Image2), systematically analyze the editing operations applied to the two beautified images relative to the Input. For each dimension, evaluate both the editing intensity and the editing quality.

You must zoom in on local regions to examine pixel-level details, and pay special attention to the following five dimensions:

1. Skin smoothing effect: Whether the Input's facial skin has been moderately smoothed, reasonably reducing pores and fine lines in the forehead, cheeks, nose wings, corners of the mouth, and eye area, while keeping it within a natural range and avoiding plastic-like texture or complete loss of pores.
2. Blemish removal effect: Whether facial blemishes in the Input, such as spots, acne, pimples, or uneven skin tone, have been effectively removed.
3. Skin tone and texture effect: Evaluate the skin tone, facial glow, and texture to determine whether the skin looks more aesthetically pleasing.
4. Preservation of personal features: Whether important personal features are fully preserved (e.g., male facial hair, female makeup, elderly wrinkles, moles, and freckles), without deletion or weakening.
5. Overall clarity: Refers to the global image quality, i.e., whether the contours of facial features (eyes, lips, nose bridge) are sharp and clear, and whether there is any blurriness, noise, or AI artifacts.

For each dimension, compare Image1 and Image2 relative to the Input, and assign an integer score from 1 to 5 (higher scores indicate better results). Strictly use the following output structure:

1. All reasoning and analysis must be written inside:  
`<think> ... </think>`

2. All scoring and the final judgment must be written inside:  
`<answer>...</answer>`

[Scoring Section]

1. Skin smoothing effect: Image1 = X, Image2 = Y
2. Blemish removal effect: Image1 = X, Image2 = Y
3. Skin tone and texture effect: Image1 = X, Image2 = Y
4. Preservation of personal features: Image1 = X, Image2 = Y
5. Overall clarity: Image1 = X, Image2 = Y

Image1 = X.XX, Image2 = Y.YY

[Preference Judgment]

ImageX is better

`</answer>`

`<think>`

1. Skin smoothing effect:
  - Image1 applies a light to moderate smoothing effect. Fine lines on the forehead and around the eyes are slightly reduced, but skin texture is still visible, maintaining a natural look.
  - Image2 applies a much stronger smoothing effect. The forehead lines are nearly erased, and the cheeks and under-eye areas appear overly airbrushed, resulting in a plastic-like texture and loss of natural skin detail.
2. Blemish removal effect:
  - Image1 removes minor blemishes and evens out the skin tone slightly, but some unevenness remains, especially around the cheeks and jawline.
  - Image2 removes all visible blemishes and discoloration, achieving a very clean and uniform skin appearance.
3. Skin tone and texture effect:
  - Image1 retains a natural skin tone with slight enhancement in brightness and warmth, giving a healthy and realistic appearance.
  - Image2 has a flatter, more artificial tone. The skin appears waxy and lacks natural glow or texture, making it look less lifelike.
4. Preservation of personal features:
  - Image1 preserves all personal features such as facial structure, hair texture, and natural lines.
  - Image2 slightly alters the facial structure, especially around the eyes and cheeks, due to over-smoothing. Some natural contours are lost.
5. Overall clarity:
  - Image1 maintains sharpness in the eyes, lips, and hair. The image is crisp and detailed.
  - Image2 shows slight blurring in the facial features, especially around the eyes and mouth, likely due to aggressive smoothing.

`</think>`

`<answer>`

[Scoring Section]

1. Skin smoothing effect: Image1 = 4, Image2 = 2 \n2. Blemish removal effect: Image1 = 4, Image2 = 5 \n3. Skin tone and texture effect: Image1 = 5, Image2 = 1 \n4. Preservation of personal features: Image1 = 5, Image2 = 3 \n5. Overall clarity: Image1 = 5, Image2 = 3

[Average Score] Image1 = 4.6, Image2 = 2.8

[Preference Judgment] Image1 is better.

`</answer>`

Figure 13. Qualitative examples of face retouching CoT reward reasoning. Given an input image and two retouched candidates, our reward model performs chain-of-thought quality assessment across five dimensions: skin smoothing, blemish removal, skin texture quality, clarity, and identity preservation.



Figure 14. Visual comparison of face retouching results across different methods on the FFHQR dataset, where Flux.K denotes FluxKontext. Compared with existing methods that suffer from incomplete blemish removal, over-smoothing, or unnatural visual appearance, our BeautyGRPO cleanly removes blemishes while preserving identity-consistent natural skin texture, gloss, and fine details such as wrinkles.

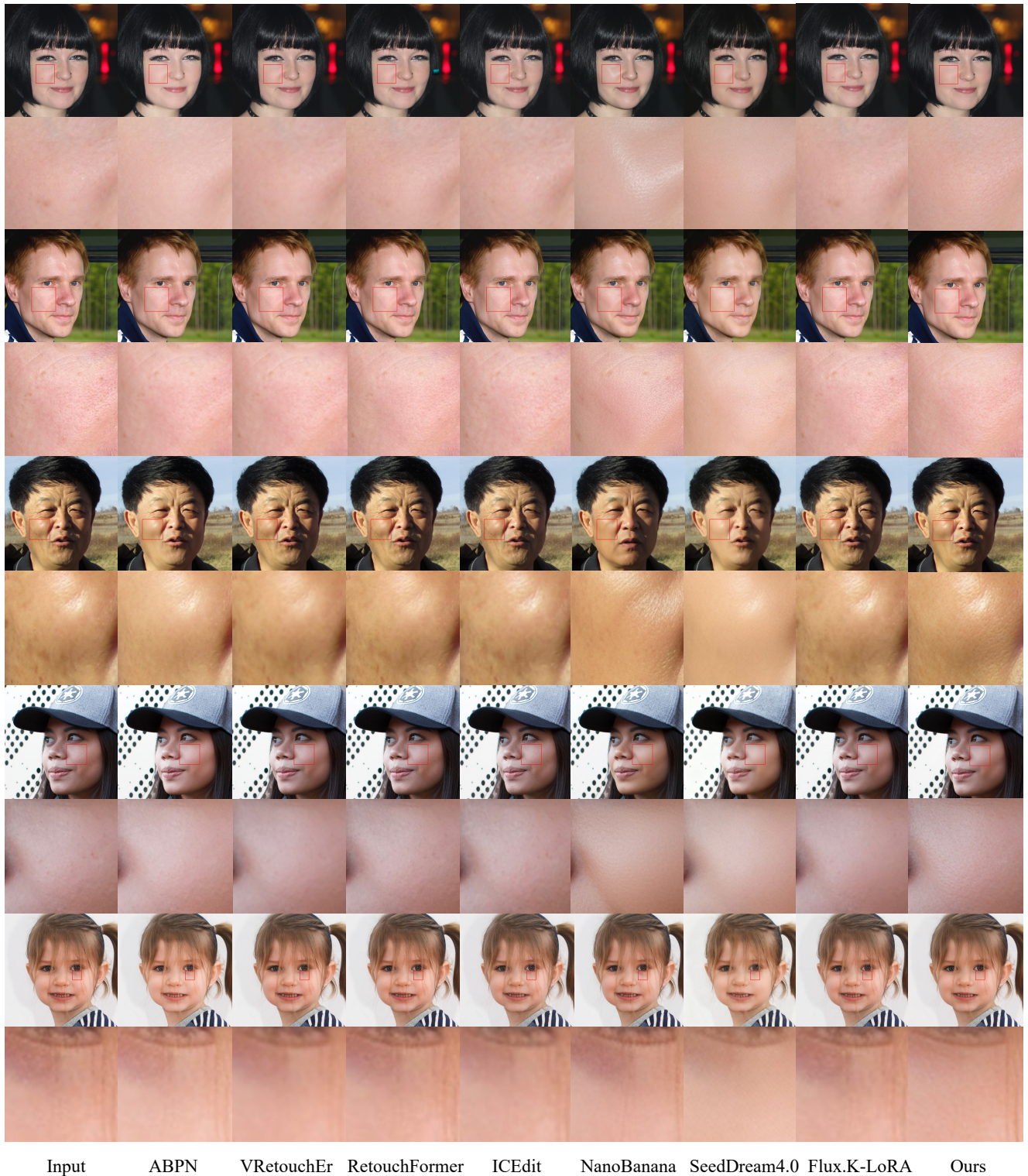


Figure 15. Visual comparison of face retouching results across different methods on the FFHQR dataset, where Flux.K denotes FluxKontext. Compared with existing methods that suffer from incomplete blemish removal, over-smoothing, or unnatural visual appearance, our BeautyGRPO cleanly removes blemishes while preserving identity-consistent natural skin texture, gloss, and fine details such as wrinkles.

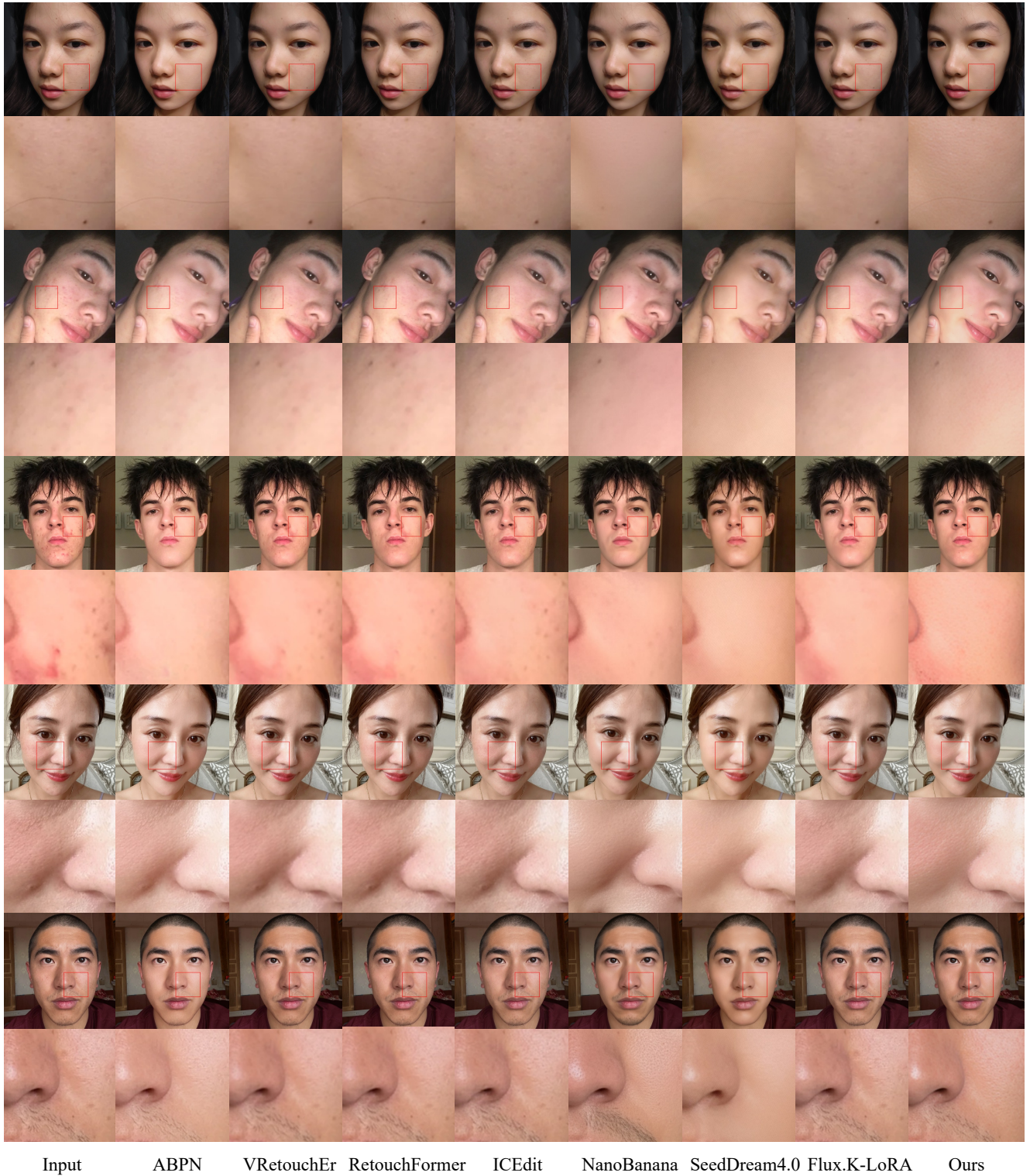


Figure 16. Visual comparison of face retouching results across different methods on in-the-wild datasets, where Flux.K denotes FluxKontext. Compared with existing methods that suffer from incomplete blemish removal, over-smoothing, or unnatural visual appearance, our BeautyGRPO cleanly removes blemishes while preserving identity-consistent natural skin texture, gloss, and fine details such as moles.