A. Appendix

A.1. Construction of Pick-High Dataset

A.1.1. Refining Prompt Construction

We selected 360,000 relatively short prompts from the pickapic_v2 dataset's prompt collection as our base prompts. As shown in Figure 9, to refine these base prompts, we first employed a GPT-2 model (Prompt-Extend) trained on diffusion model prompts, which generates appropriate style and detail descriptions based on the prompt's main theme. To further enhance the prompts' coherence and granularity, we designed a chain-of-thought template and utilized the Claude-3.5-sonnet for chain-of-thought reasoning. Through this process, we filtered out descriptions that didn't match the base prompt's style and any inadvertently introduced anomalous content, while simultaneously enriching the prompts with thematically appropriate aesthetic and stylistic information.

A.1.2. Image Collection and Dataset construction

Given the limitations in generalization performance of vision language models[24, 33, 45], we additionally verified whether the refined prompts fully preserve the core concepts of the base prompts while introducing no conflicting information.. Our verification included: (1) Filtering out NSFW contents rejected by Claude-3.5-sonnet (∼2%); (2) Applying a binary verification CoT template to exclude non-compliant samples (∼7%); Additionally, expert evaluation of 1000 randomly sampled Pick-High items confirmed 97% prompt and 95% image compliance with requirements. Based on this, we input all refined prompts into the Stable Diffusion-3.5-large model to generate 360,000 high-quality images, forming our proposed Pick-High dataset. Since the base prompts in the Pick-High dataset originate from the filtered results of the pickapic_v2 dataset, the natural fusion of these two datasets creates a training dataset with ternary preference relationships and significant quality variations.

A.2. Experiment Details

Our experimental framework comprises three sequential phases: ICT model training, HP model training, and diffusion model optimization.

ICT Model Training In the first phase, we fine-tune all parameters of the CLIP-H model [4] using MSE loss to optimize ICT scores. Training is conducted on 8 NVIDIA A800 GPUs for a total of 40,000 iterations. We employ the AdamW optimizer [16] with a learning rate of 3e-5. The smoothing coefficient α and threshold parameter β in the negative sample smoothing function are set to 20 and 6, respectively, while the balancing factor λ in the loss function is 0.1. **HP Model Training** In the second phase, we keep the ICT model parameters fixed and only fine-tune the latter half of the CLIP-H model [4] parameters and its connected MLP layers. Training is similarly performed on 8 NVIDIA A800 GPUs for 50,000 iterations. The margin threshold m is set to 0.2, using the AdamW optimizer [16] with a learning rate of 3e-6.

Diffusion Model Optimization In the final phase, we optimize the diffusion model using the trained reward models. We select Stable Diffusion-3.5-turbo as the base model and conduct training in half-precision (FP16). The diffusion process is configured with 8 sampling steps and a Guidance Scale of 0.0. Following the DRaFT-K method [5], we only propagate gradients through the last 3 denoising steps to optimize the LoRA parameters in the transformer layers, while keeping all other base model parameters frozen. Training is performed on 5 nodes equipped with 8 NVIDIA A800 GPUs each, for a total of 3,000 iterations, with a training time of approximately 24 hours. We utilize the AdamW optimizer [16] with a learning rate of 5e-6.

A.3. Comprehensive Reward Scoring for Original and Refine Images Across Diffusion Models

We randomly selected 800 base prompts from DiffusionDB [36] and COCO Captions [6], and obtained refined prompts through optimization by large language models. These two sets of prompts were input into six diffusion models with diverse architectures (SD1.5, SDXL, SD3.5-Turbo, SD3.5-Large-Turbo, FLUX.1-schnell, and FLUX-1.dev) for image generation. We use the suffix "e" to denote images generated with base prompts and "r" for images generated with refined prompts.

As shown in Table 6, when evaluating images generated from refined prompts, scores decrease across both basic multimodal models (CLIP and BLIP) and all human preference models. This indicates that images generated from refined prompts contain richer information, resulting in reduced text-image similarity. Notably, our refined prompts, filtered through the chain-of-thought process of large language models, do not contain semantically irrelevant subjects or style words. Therefore, refine images do not introduce text-image misalignment, but rather enhance aesthetic qualities, details, and texture-related information.

Experimental results demonstrate that aesthetic metrics based solely on the image modality show improvement across all test cases, confirming that images generated from refined prompts indeed contain richer visual information. Since pure image modality evaluation is not affected by explicit or implicit text-image

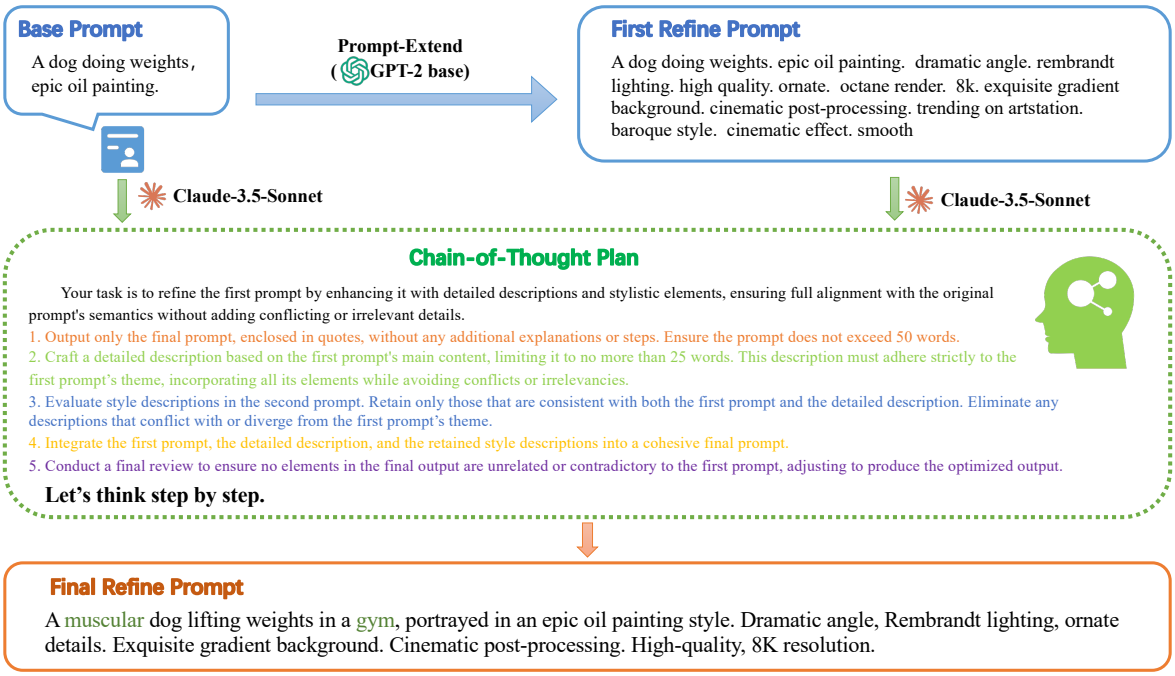


Figure 9. The Overview of Prompt Refinement Pipeline via Large Language Model Chain-of-Thought.

Model	CLIP↑	ITM↑	ImgRwd↑	HPS↑	Pick↑	Aes↑	ICT↑	HP↑	ICT-HP↑
$SD1.5_e$	0.336	0.697	0.188	0.267	20.689	5.884	0.667	0.672	0.457
$SD1.5_r$	0.309↓	0.575↓	−0.078↓	0.264↓	20.395↓	6.169↑	0.680↑	0.685↑	0.477↑
$SDXL_e$	0.367	0.807	0.719	0.272	21.853	6.446	0.898	0.772	0.695
$SDXL_r$	0.338↓	0.715↓	0.687↓	0.271↓	21.789↓	6.857↑	0.908↑	0.776↑	0.705↑
$SD3.5\text{-}T_e$	0.340	0.831	0.955	0.277	22.034	6.555	0.910	0.777	0.717
$SD3.5\text{-}T_r$	0.326↓	0.757↓	0.914↓	0.276↓	22.012↓	6.780↑	0.923↑	0.781↑	0.718↑
$SD3.5\text{-}L_e$	0.356	0.881	1.038	0.277	21.950	6.403	0.943	0.777	0.732
$SD3.5\text{-}L_r$	0.334↓	0.800↓	0.955↓	0.276↓	21.853↓	6.704↑	0.939↓	0.779↑	0.731↓
$FLUX\_S_e$	0.349	0.866	0.974	0.277	21.739	6.439	0.891	0.763	0.683
$FLUX\_S_r$	0.335↓	0.800↓	0.893↓	0.276↓	21.657↓	6.691↑	0.909↑	0.773↑	0.704↑
$FLUX_e$	0.333	0.818	0.982	0.279	22.021	6.642	0.906	0.775	0.703
$FLUX_r$	0.326↓	0.791↓	0.967↓	0.278↓	21.899↓	6.846↑	0.919↑	0.778↑	0.715↑

Table 6. **Quantitative Results of Image Generation Models**.

comparison objectives, its assessment results align with human preference trends.

Our trained ICT model, HP model, and the combined ICT-HP model show score improvements across almost all tested models, strongly demonstrating that the ICT training objective successfully addresses the inherent deficiency between instance-level text-image alignment and human preferences. The results indicate that as the generation model quality reaches high levels, the ICT metric maintains stable high values, suggesting our ICT scoring mechanism does not negatively evaluate high-quality images after reaching the text-image alignment inflection point. The HP model, as a reward model trained solely on the image modality, ensures higher scores for refine images in all test scenarios. In the current implementation, the ICT-HP score is calculated through a simple product of ICT and HP scores; future research could explore more optimal methods for integrating these two models.

Model	Mean ↑	Single↑	Two↑	Counting↑	Colors↑	Position↑	Color Attribution↑
SDXL	0.55	0.98	0.74	0.39	0.85	0.15	0.23
DALL-E 2	0.52	0.94	0.66	0.49	0.77	0.10	0.19
DALL-E 3	0.67	0.96	0.87	0.47	0.83	**0.43**	0.45
SD3	0.68	0.98	0.84	0.66	0.74	0.40	0.43
FLUX.1-Schnell	0.68	0.99	0.88	0.62	0.76	0.30	0.51
+ ICT-HP (Ours)	**0.69**	**0.99**	**0.88**	**0.62**	**0.81**	0.29	**0.81**
+ ICT (Ours)	**0.69**	**0.99**	**0.88**	0.60	**0.81**	0.29	0.55
FLUX.1-dev	0.66	0.97	0.82	0.71	0.78	0.22	0.45
+ ICT-HP (Ours)	**0.67**	**0.99**	**0.82**	**0.74**	**0.80**	0.20	**0.51**
+ ICT (Ours)	**0.67**	0.98	0.81	0.73	**0.80**	0.19	0.50

Table 7. **Quantitative GenEval Results of FLUX.1-schnell with RM-Optimized LoRA (HP/ICT-HP) and Transfer Performance on FLUX.1-dev.**

A.4. Optimizing FLUX.1-schnell with HP and ICT-HP Models: Implementation and Evaluation

We applied our proposed HP model and ICT-HP model to optimize the flux.1-schnell architecture. During training, we employed half-precision (FP16) to enhance computational efficiency. The diffusion process was configured with a 4-step sampling procedure and a Guidance Scale of 0.0. Following the DRaFT-K methodology [5], we selectively propagated gradients through only the final denoising step to optimize the LoRA parameters within the transformer layers, while maintaining the integrity of other parameters in the base model. The experimental setup consisted of 4 computing nodes, each equipped with 8 NVIDIA A800 GPUs, supporting a total of 3,000 training iterations completed in approximately 24 hours. For optimization, we utilized the AdamW optimizer [16] with a learning rate of 5e-6. Herein, we present both qualitative analysis and quantitative evaluation results on the GenEval benchmark.

Table 7 presents the quantitative results of our FLUX.1-schnell optimization using both the ICT-HP model and ICT model. Additionally, we directly transferred the LoRA weights trained on FLUX.1-schnell to FLUX.1-dev to obtain quantitative performance metrics within the multidimensional GenEval evaluation framework. Our ICT-HP and ICT models demonstrated notable advantages compared to the baseline models, with particularly significant improvements in color-related scores. These results indicate that our reward model effectively enhances the color fidelity of the FLUX model series.

A.5. Comprehensive Presentation of Diverse Qualitative Results

Simple Element Generation Qualitative Results. To evaluate our model's performance in text-image alignment, particularly under minimal prompt conditions, we provide additional qualitative results of simple element generation, thoroughly demonstrating our method's excellence in text-image consistency. As shown in Figure 11, we compared the original SD3.5-turbo with results optimized by CLIP model, ICT model, and ICT-HP model. Through detailed observation, it is evident that the original SD3.5-turbo exhibits significant limitations when executing minimal instructions; while the CLIP model-optimized version improves text-image alignment in some scenarios, it performs inconsistently across various situations. Among all variants, the ICT model optimization solution demonstrates the most precise and efficient effects, perfectly meeting all prompt requirements; meanwhile, the ICT-HP model optimization solution achieves near-optimal performance, with overall quality significantly superior to the CLIP model-optimized version. These experimental results strongly confirm that our innovative method not only significantly enhances the model's overall performance, but also successfully maintains and strengthens the base model's accurate understanding of clear, minimal concepts.

Qualitative Comparison of Optimization Results. In Figure 12, we provide more examples of various reward optimizations. Based on our observations, ImageReward struggles to further optimize the high-performance diffusion model SD3.5-Large-Turbo, therefore we have excluded the qualitative results of ImageReward in this instance.

Qualitative Comparison Between Optimization Results and Refine images. In Figure 13, we present comparative displays of additional original images, refine images, and results optimized through our model, confirming that our reward model can surpass the performance of prompt refinement.

Qualitative Comparison of Style Injection Re-

sults. In Figure 14, we showcase additional qualitative results demonstrating style transfer achieved by extracting stylistic elements using Image-Encoders from both the HP model and the ICT model.

B. Mathematical Framework and Theoretical Foundation of ICT-HP Reward System

B.1. Information Saturation Hypothesis and ICT Metric Formulation

We propose the Information Saturation Hypothesis as mathematical foundation of ICT:

Hypothesis 1 (***Information Saturation Hypothesis***). *For any image-text pair (v, t), there exists a mutual information critical value $I^*(v, t)$ such that v semantically aligns with t if and only if $I(v; t) \geq I^*(v, t)$.*

As detailed in Section 3, the CLIP score is formulated as

$$\text{CLIP}(v, t) \approx \frac{I(v; t)}{\sqrt{(I(v; t) + I(v|t)) \cdot I(t)}} \tag{14}$$

When $I(v; t) = I^*(v, t)$, according to the Information Saturation Hypothesis, the image fully contains the textual information, and

$$\text{CLIP}^*(v, t) \approx \frac{I^*(v, t)}{\sqrt{(I^*(v, t) + I(v|t)) \cdot I(t)}} \tag{15}$$

Since this function decreases monotonically with $I(v|t)$, its minimum value occurs at the critical threshold when $I(v|t) = I_{\max}(v|t)$. To ensure all semantically aligned image-text pairs receive an ICT score of 1, we set

$$\theta^* = \text{CLIP}^*_{\min} \approx \frac{I^*(v, t)}{\sqrt{(I^*(v, t) + I_{\max}(v|t)) \cdot I(t)}} \tag{16}$$

Physically, ICT reflects boundary saturation effect in human perception, where once an image adequately represents text, further details don't reduce alignment score.

B.2. Preference Modeling and HP Metric Derivation

Our data contains image triplets (I_1, I_2, I_3) from identical prompts with preference hierarchy $I_3 \succ I_2 \succ I_1$, justifying HP's image-only approach. Bradley-Terry models preferences as:

$$P(I_j \succ I_i) = \frac{1}{1 + \exp(-(s(I_j) - s(I_i)))} \tag{17}$$

The log-likelihood:

$$\mathcal{L} = \log P(I_2 \succ I_1) + \log P(I_3 \succ I_2) \tag{18}$$

has negative upper bound via convex function theory, yielding ranking loss of HP Metric:

$$\begin{aligned} L_{margin} = \sum [&\max(0, -\Delta(I_2, I_1) + m) \\ &+ \max(0, -\Delta(I_3, I_2) + m)] \end{aligned} \tag{19}$$

B.3. Multiplicative Integration Theory and System Properties

Based on our analysis of CLIP scoring limitations, we propose a multiplicative dual-metric evaluation system:

$$\text{Reward}(v, t) = \text{ICT}(v, t) \cdot \text{HP}(v), \tag{20}$$

where ICT and HP are defined as:

$$\text{ICT}(v, t) = \frac{I(v; t)}{I(t)}, \tag{21}$$

$$\text{HP}(v) = f(I(v|t)). \tag{22}$$

This multiplicative formulation offers significant theoretical advantages:

1. **Complementary Constraint**: ICT ensures faithful textual expression while HP evaluates aesthetic quality; their product requires simultaneous satisfaction of both conditions.
2. **Threshold Effect**: When either metric approaches zero, the overall reward approaches zero, preventing optimization of one aspect at the expense of the other.
3. **Non-linear Gain**: When both ICT and HP improve simultaneously, the reward function exhibits accelerated growth, incentivizing concurrent enhancement of text containment and image quality.

In information-theoretic terms, the multiplicative form can be expressed as:

$$\text{Reward}(v, t) \approx \frac{I(v; t)}{I(t)} \cdot f(I(v|t)), \tag{23}$$

Compared to CLIP scoring:

$$\text{CLIP}(v, t) \approx \frac{I(v; t)}{\sqrt{I(t) \cdot (I(v; t) + I(v|t))}}. \tag{24}$$

Our multiplicative formulation avoids the negative impact of $I(v|t)$ on the denominator in CLIP, instead positively utilizing $I(v|t)$ through the HP term, enabling accurate assessment of high-quality images.

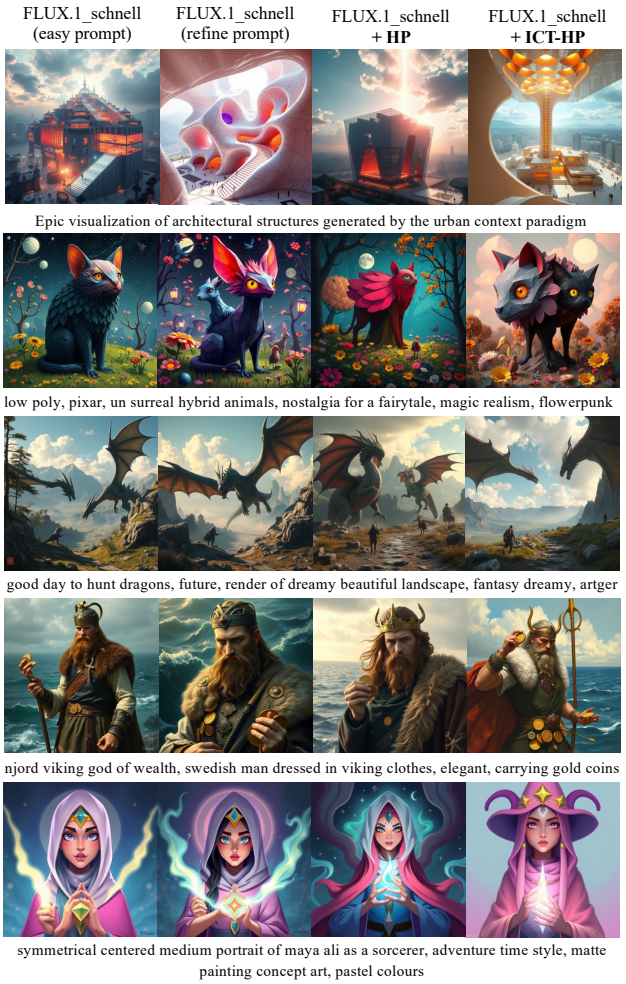


Figure 10. Qualitative Results of Optimizing FLUX.1-schnell.

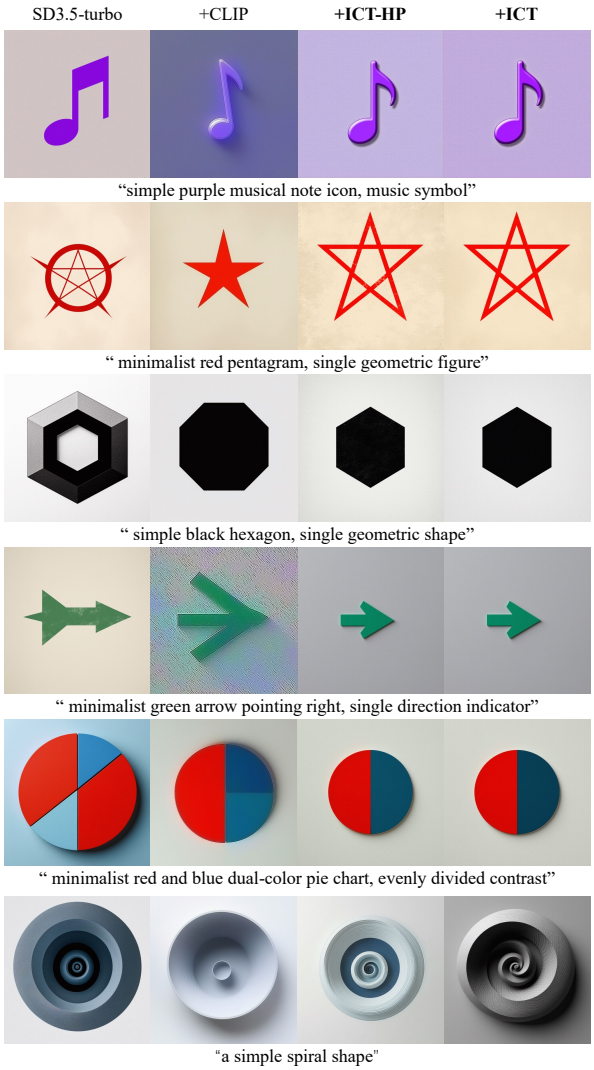


Figure 11. More simple elements generate results.

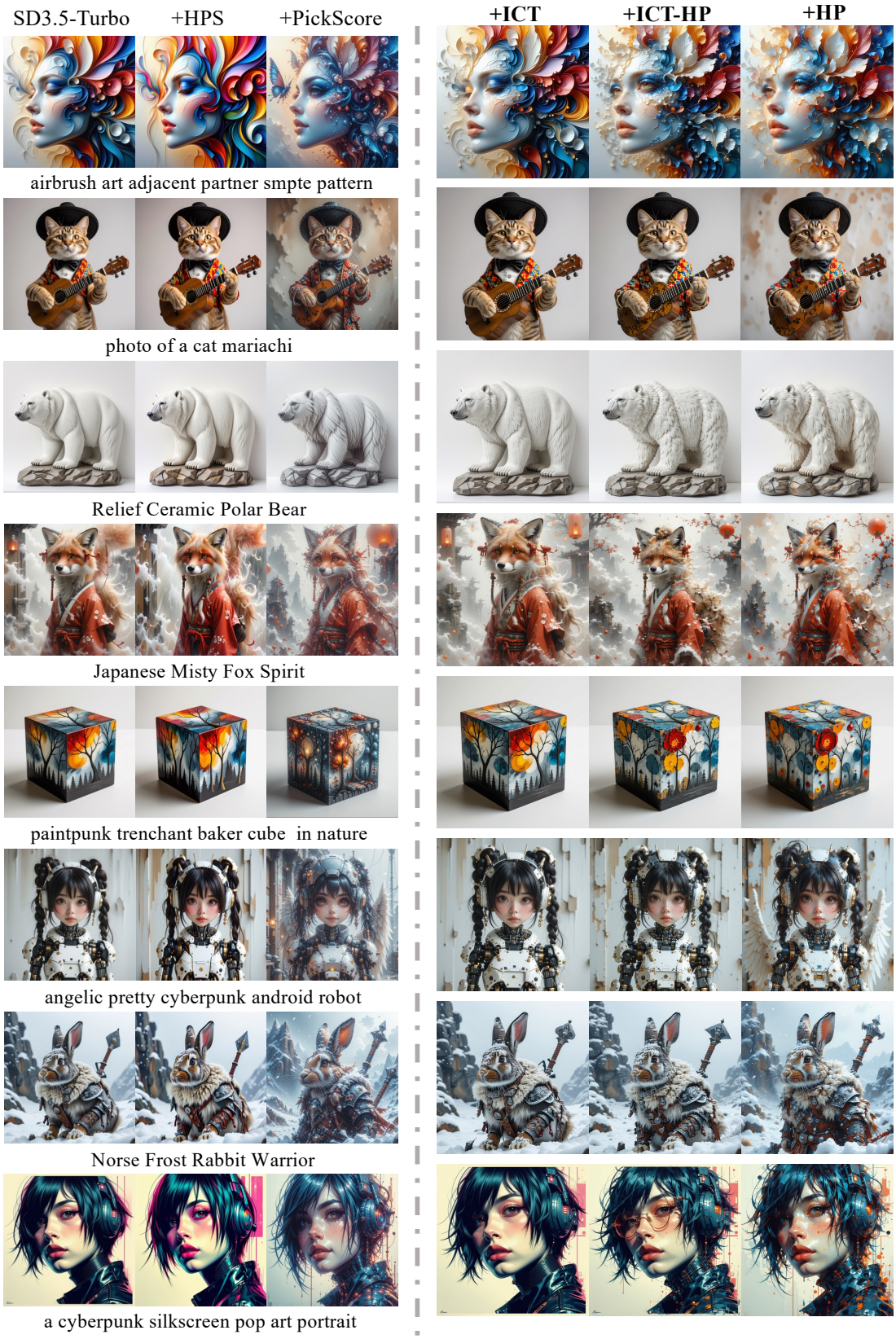


Figure 12. Qualitative Comparison of Optimization Results Across Reward Models Using Real User Prompts.

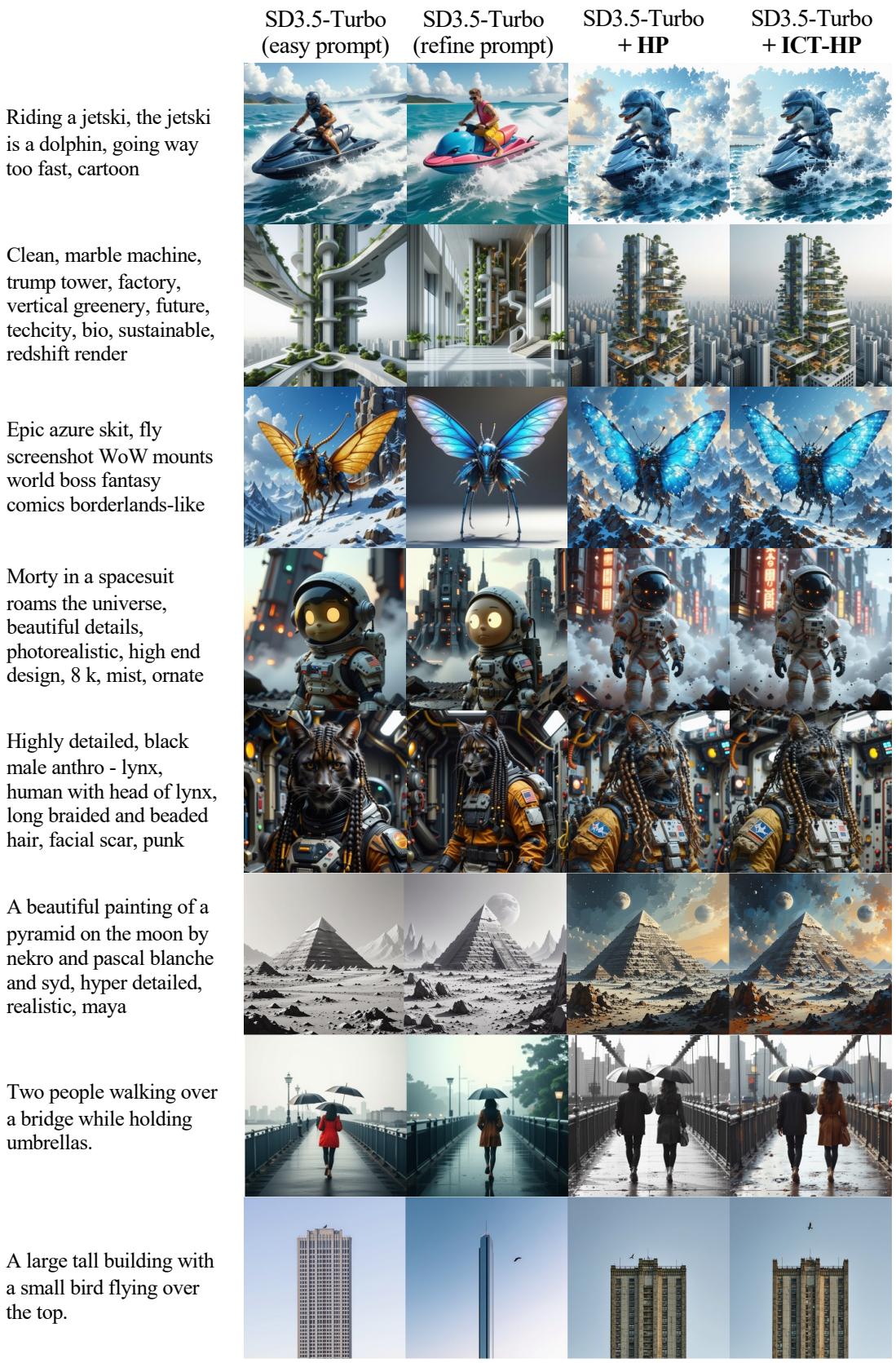
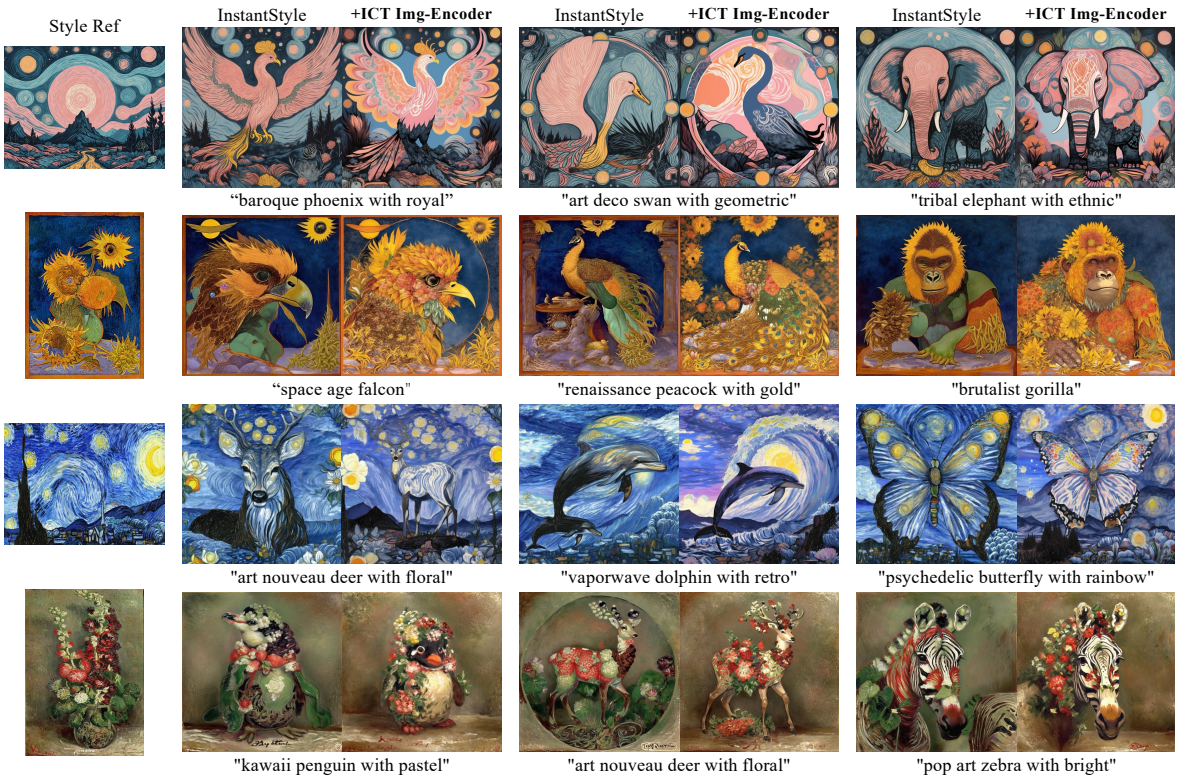


Figure 13. Qualitative Comparison: Origin Images, Refine Images, and Generations Optimized by Our Reward Models.



(a) Qualitative Comparison of ICT Image-Encoder Performance in Style Injection Tasks.

(b) Qualitative Comparison of HP Image-Encoder Performance in Style Injection Tasks.

Figure 14. Qualitative Results of Style Injection Tasks.