

Too Vivid to Be Real? Benchmarking and Calibrating Generative Color Fidelity

Supplementary Material

A. Dataset Details

To support objective evaluation of color fidelity in realistic-style text-to-image (T2I) generation, we construct the large-scale **Color Fidelity Dataset (CFD)**, which provides high-quality real-world photographs, their corresponding textual captions, and progressively distorted synthetic counterparts generated under different guidance scales. This section provides detailed statistics and construction protocols for the dataset, including category distribution, model composition, and benchmark splits.

As summarized in Tab. 7, the **Color Fidelity Dataset (CFD)** comprises 189,490 high-quality real-world photographs evenly distributed across 12 semantic categories, ensuring comprehensive coverage of diverse color appearances encountered in everyday photography. These real images are carefully curated from open-source photographic collections with rigorous filtering to remove synthetic, over-processed, or low-resolution samples. Each image serves as a perceptual anchor for evaluating the color fidelity of its synthetic counterparts.

Tab. 6 further details the construction of the training and testing splits. Each real image is associated with six progressively distorted synthetic images generated under different classifier-free guidance (CFG) scales ($s \in 7.5, 10, 15, 20, 25, 30$), enabling ordered supervision of color realism levels. The **CFD-Train** subset contains 160,000 real photographs and over 1.1M synthetic images produced by seven representative text-to-image models, covering a balanced spectrum of diffusion-based architectures and training paradigms. The **CFD-Test** subset extends the model diversity to eleven generation systems, incorporating both commercial and open-source models to facilitate fair benchmarking across different generation pipelines. We also provide several visualization examples of CFD in Fig. 11 to illustrate the dataset composition and the progressive nature of the generated distortions.

In Sec. 6.3, CFD-SynPairs is formed by randomly sampling pairs of synthetic images within the same group that correspond to *adjacent* guidance scales (e.g., $s = 10$ vs. $s = 15$), which enables fine-grained assessment of color fidelity ranking consistency. In contrast, CFD-Real&Syn pairs each real image with its synthetic counterpart generated under the *lowest* guidance scale ($s = 7.5$), thus directly evaluating absolute color fidelity between real-world photographs and their minimally guided generations. Both subsets contain 5,000 image pairs and are used for model discrimination accuracy analysis.

Additionally, in Sec. 6.2, the **CFS Benchmark** for large-

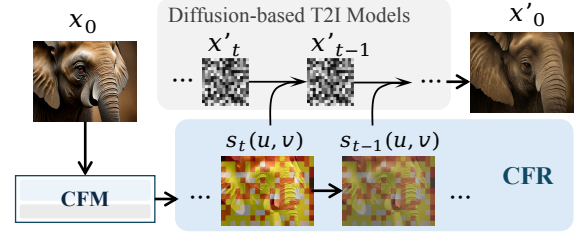


Figure 8. Framework of CFR.

scale evaluation of color realism across generation models is constructed by sampling 1,000 real image-caption pairs from each of the 12 semantic categories within CFD-Test, resulting in a total of 12,000 samples. This benchmark provides a balanced and category-diverse testbed for standardized comparison of fidelity scores and generation quality across different T2I systems.

B. Benchmark Comparison and Visualization

In Fig. 12 and Fig. 13, we present visualization examples of realistic-style generations produced by different T2I models, together with the corresponding scores assigned by our metric. Fig. 9 further reports the ranking results of various models as evaluated by CFM and by existing metrics. It can be observed that our CFM provides substantially more accurate assessments of color fidelity, consistently reflecting the degree to which synthesized images preserve natural color appearance. In contrast, aesthetic-based or semantics-oriented metrics exhibit rankings that deviate significantly from the color realism of the generated images, as they primarily focus on visual appeal or text-image alignment rather than faithfully capturing color authenticity. These discrepancies highlight the necessity of a dedicated color-fidelity metric and demonstrate the reliability of our method.

C. Details of CFR

To provide a clearer explanation of the Color Fidelity Refinement (CFR) mechanism introduced in Sec. 5, we illustrate the full refinement pipeline in Fig. 8. CFR leverages the cross-modal attention extracted by the Color Fidelity Metric (CFM) to identify spatial regions that exhibit deviations from natural photographic color characteristics, and uses this information to modulate the guidance behavior of diffusion models.

Specifically, CFM computes text-to-image attention from the multimodal embeddings produced by the Qwen2-

Table 6. Summary of CFD-Train and CFD-Test.

Subset	Real Images	Total Images	Models Used
CFD-Train	160,000	1,120,000	SDXL [20], SD3 [7], SD3.5, PixArt-Sigma [3], Kolors [28], CogView4 [4], Hunyuan-DiT [17]
CFD-Test	29,490	206,430	SDXL [20], SD3 [7], SD3.5, PixArt-Sigma [3], Kolors [28], CogView4 [4], Hunyuan-DiT [17], Flux-dev [15], Qwen-Image [31], Playground-v2.5 [16], SRPO [25]

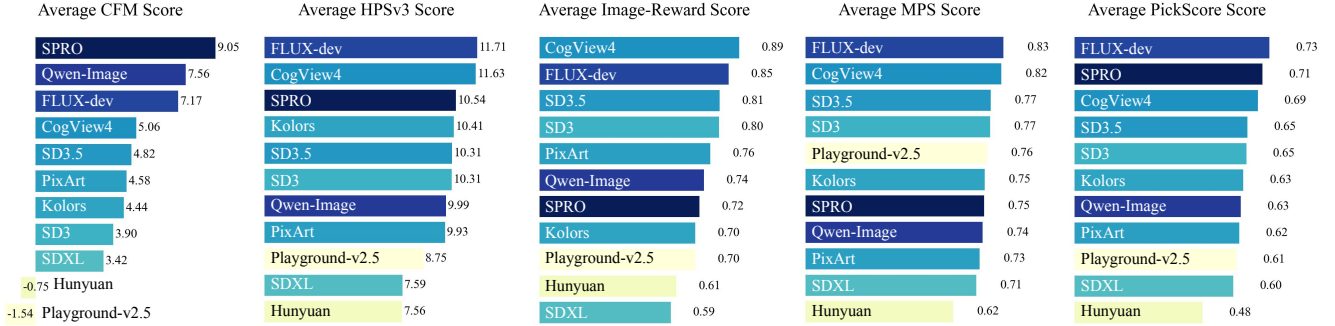


Figure 9. Cross-benchmark compairison.

Table 7. Category distribution of the Color Fidelity Dataset (CFD).

Category	Real Images	Percentage (%)
Human	21,422	11.3
Animals	17,105	9.0
Plants	16,158	8.5
Food	17,474	9.2
Vehicles	13,631	7.2
Sports	9,963	5.3
Architecture	16,526	8.7
Natural Scene	15,947	8.4
Street Scene	17,211	9.1
Indoor Scene	17,895	9.4
Night Scene	12,369	6.5
Others	13,789	7.3
Total	189,490	100.0

VL encoder. This attention distribution reflects the degree of *color realism discrepancy* at each spatial location. After normalization and upsampling, the attention map serves as a spatial mask that adjusts the classifier-free guidance strength during the sampling process: areas with higher discrepancy receive stronger attenuation, while regions with naturally aligned color statistics preserve a guidance level closer to the original value. In addition, a temporal decay factor gradually reduces the modulation amplitude along the denoising trajectory, ensuring that the refinement does not disrupt semantic structure in later sampling steps.

At each denoising step, the diffusion model recomputes the noise prediction using the spatially and temporally varying guidance field, enabling fine-grained correction of color

distortions without modifying model parameters. As a fully training-free method, CFR is compatible with any diffusion-based T2I model employing classifier-free guidance.

Furthermore, we visualize additional examples of CFR in Fig. 10. As shown, CFR substantially improves the perceptual realism of generated images by mitigating oversaturation and restoring natural color appearance. Correspondingly, the CFM scores also display consistent increases, demonstrating the effectiveness of CFR in enhancing generative color fidelity.

D. Details of Implementation

When training the CFM, we preserve the aspect ratio of each input image and resize it such that the longer edge is fixed to 448 pixels. We set the temperature hyperparameter of the softmax loss to $\tau = 0.1$. For the CFR pipeline, we use a temperature parameter of $\kappa = 10$ when computing the text-to-image attention matrix.

Limitations

In this work, we exploit the varying distortion effects produced by different T2I models under multiple classifier-free guidance (CFG) scales to diversify color fidelity variations as much as possible. However, color distortions simulated solely through CFG adjustments remain limited in scope and may not fully represent the wide range of color deviations in real-world generative outputs. In future work, we plan to explore more comprehensive and controllable distortion mechanisms tailored for T2I models to achieve a more holistic evaluation of color fidelity.

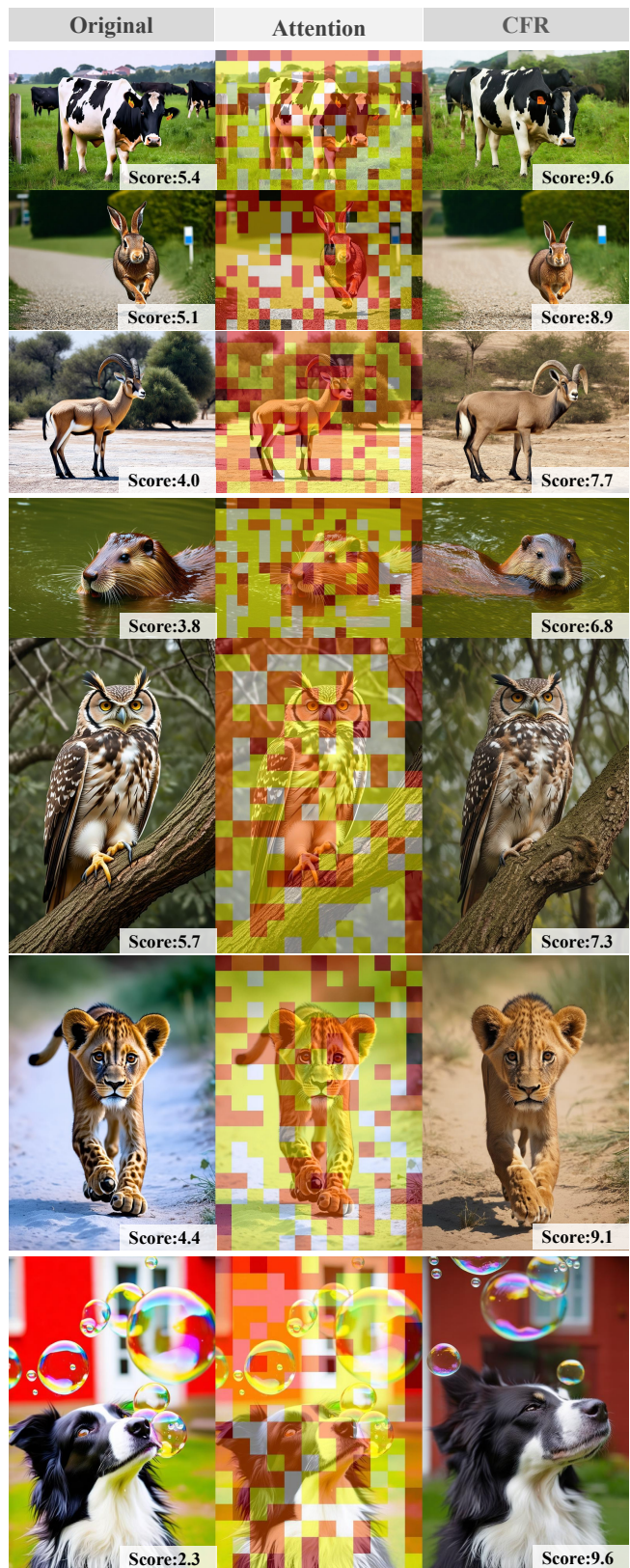


Figure 10. Visualization examples of CFR. The CFM score for each image is shown in the lower-right corner.

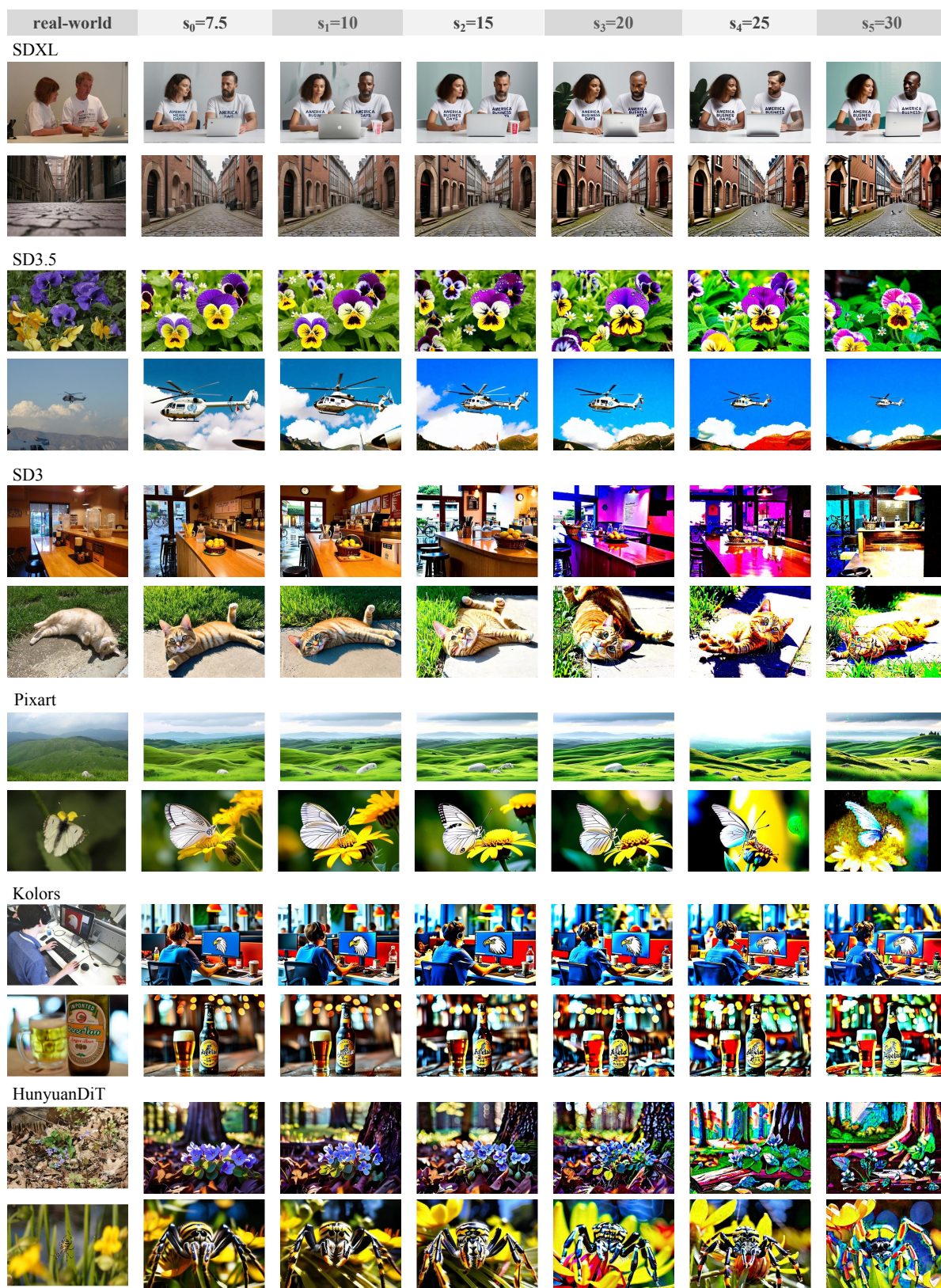


Figure 11. Visualization examples of CFD.

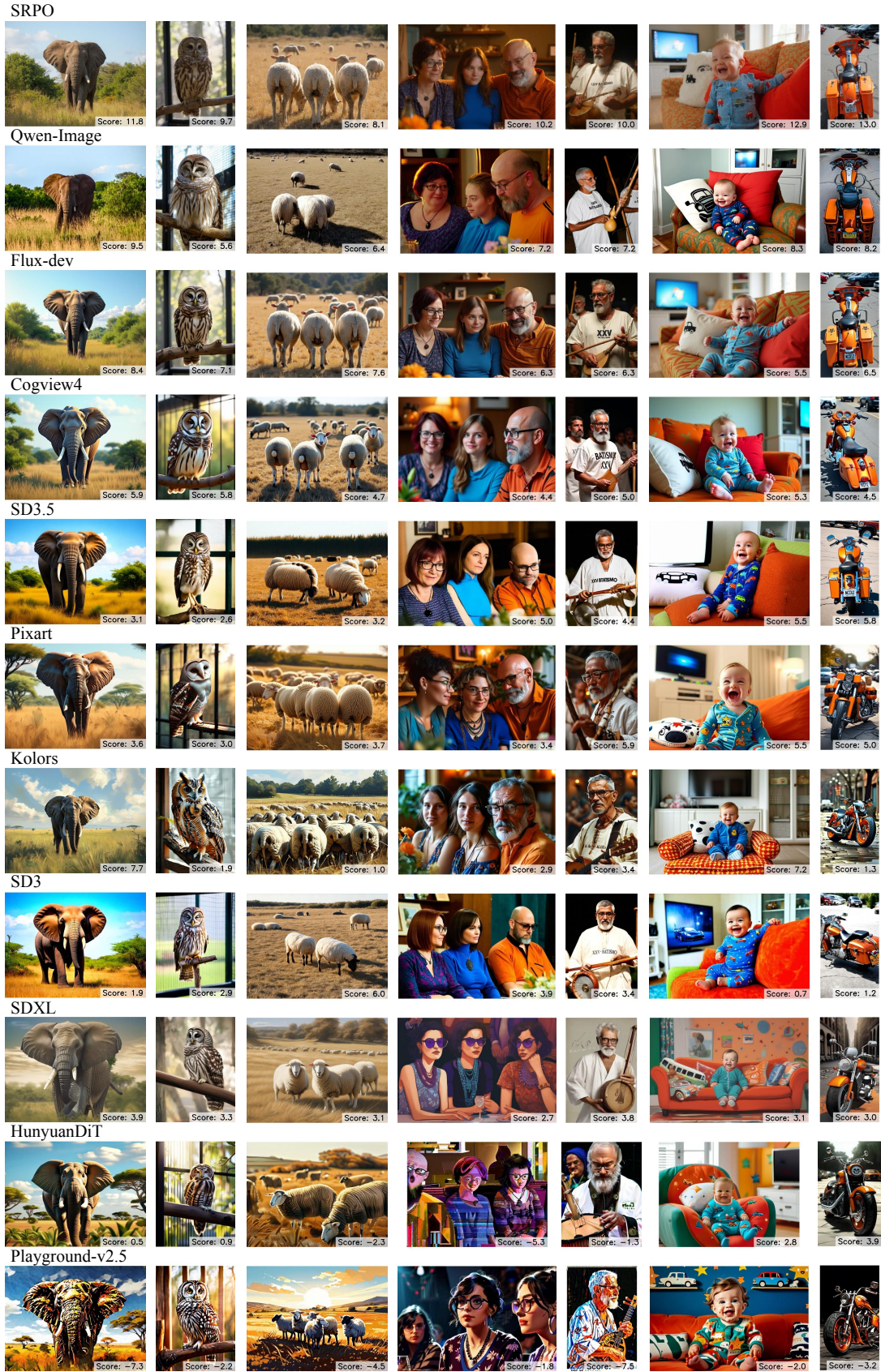
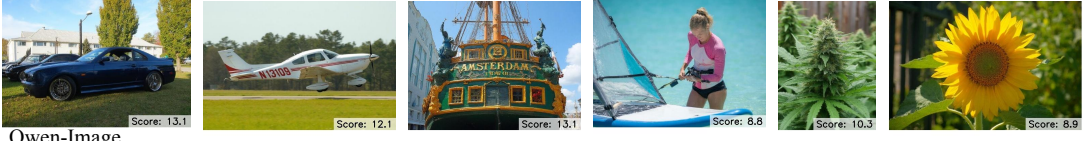
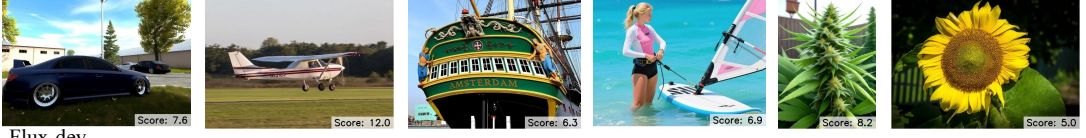


Figure 12. Visualization examples of CFM scores on realistic-style generations produced by different T2I models. The CFM score for each image is shown in the lower-right corner.

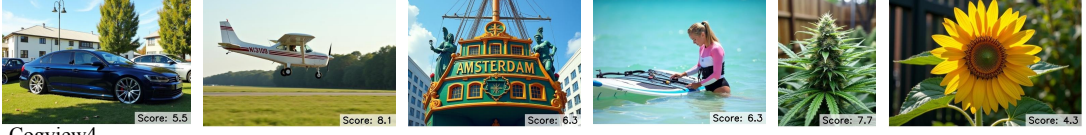
SRPO



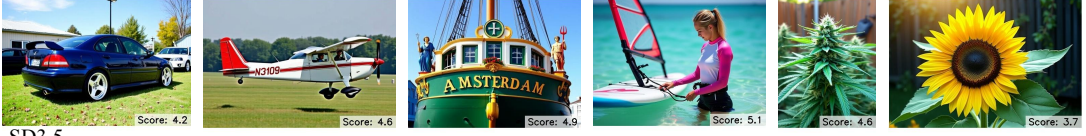
Qwen-Image



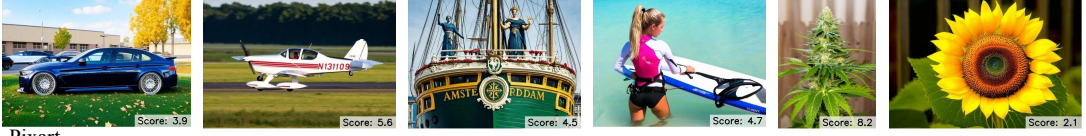
Flux-dev



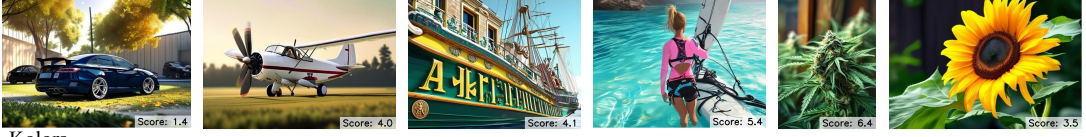
Cogview4



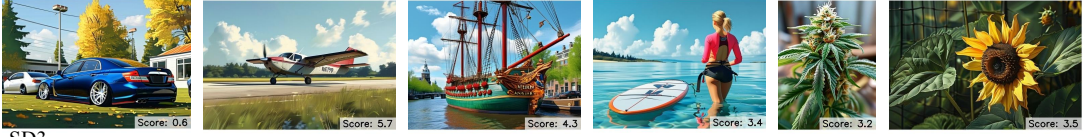
SD3.5



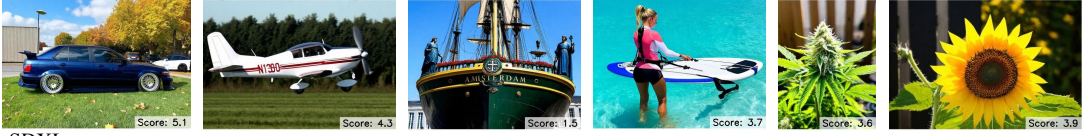
Pixart



Kolors



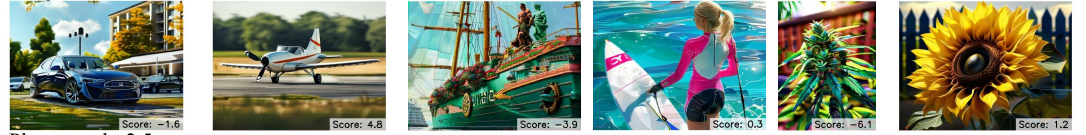
SD3



SDXL



HunyuanDiT



Playground-v2.5



Figure 13. Visualization examples of CFM scores on realistic-style generations produced by different T2I models. The CFM score for each image is shown in the lower-right corner.