Diffusion Probe: Generated Image Result Prediction Using CNN Probes

Supplementary Material

5. More Results about Diffusion Probe

5.1. The performance of Diffusion Probe at 512×512 resoluton

While our main experiments are conducted at a resolution of 1024×1024, we additionally evaluate the Diffusion Probe at a lower resolution of 512×512 to assess its robustness across input scales. Even under this reduced-resolution setting, the probe exhibits stable performance and remains well aligned with the target quality metrics, indicating that its predictive capability does not depend on the high-resolution regime. The corresponding results are presented in Table 4.

5.2. More Results of Ablation Study

Diffusion Probe trained with another metric. In the main text, the diffusion probe is trained using image–reward annotations as supervision. To further examine the probe's flexibility, we additionally train it with several alternative image–quality metrics as labels. Across all cases, the probe is able to accurately approximate the corresponding quality indicators, demonstrating its robustness to different supervisory signals. The results are summarized in Table 5.

Ablation Study of Steps Window. We examine the effect of varying the diffusion steps window on the performance of the Diffusion Probe. Our primary goal is to demonstrate that the Diffusion Probe does not only exhibit strong predictive accuracy at the fifth diffusion step, but also performs well across a range of neighboring steps, indicating that the effective window extends beyond a single step. Specifically, we analyze the probe's ability to predict image quality metrics at steps in the vicinity of step 5, such as steps 3, 4, 6, and 7.

By systematically evaluating the model across these steps, we aim to show that the probe captures relevant image quality features consistently over a broader set of diffusion stages, rather than relying solely on step 5. This extended effective window suggests that the probe's attention features are stable and robust, maintaining high predictive accuracy across multiple diffusion steps.

As shown in Table 4, the Diffusion Probe maintains similar levels of predictive accuracy across the steps near step 5, with only marginal variations in performance. This result confirms that the probe remains effective over a range of neighboring steps, highlighting the flexibility and robustness of our method in capturing quality-related features at different stages of the diffusion process.

5.3. Results of Efficient Flow-GRPO

In this section, we present the empirical results of integrating our **Diffusion Probe** into the Flow-GRPO framework, demonstrating the practical benefits of our early-stage quality predictor in a resource-intensive application. Our evaluation focuses on two key aspects of efficiency that are directly impacted by the probe's filtering capability:

- **Training Efficiency Improvement:** We quantify the acceleration of the overall training process, measuring the reduction in computational time and resources required to achieve comparable or superior model performance.
- **Increase in Effective Sample Ratio:** We analyze how our Diffusion Probe enhances the sample efficiency during the training loop, specifically within the Flow-GRPO framework. The sampling process in Flow-GRPO aims to capture both positive and negative samples, encouraging diversity and a balanced exploration of the reward space. In this context, we measure the variance of the Pick Score across the sampled data at each training step. A higher variance in the Pick Score indicates a better distinction between positive and negative samples, which is crucial for training stability and performance. By using our probe to filter out low-quality samples early in the process, the variance of the Pick Score becomes more focused on high-quality, informative samples. As a result, the proportion of valid training data increases by 40%, reflecting a significant improvement in the sample efficiency and diversity, without sacrificing the effectiveness of positive-negative sample separation.

Figure 5 that our approach not only accelerates the training pipeline but also makes it more effective by optimizing the data used for policy updates.

5.4. More Qualitative Results about Diffusion Probe

In Figure 9, we provide additional detailed qualitative examples of our Diffusion Probe, showcasing its ability to evaluate images generated at different quality levels. We rank 10 images generated from different prompts and compare these rankings with the corresponding original image quality metrics. This comparison highlights the Diffusion Probe's effectiveness in assessing both the visual consistency between the text prompt and the generated image, as well as the overall aesthetic quality of the images.

Furthermore, we demonstrate how well the Diffusion Probe's rankings align with existing image quality metrics, such as those based on human perception. The results illustrate that the Diffusion Probe can effectively capture both text-image consistency and image appeal, making it a reli-

Table 3. **Computational Cost Analysis.** This table compares the computational cost of naive brute-force workflows against our Diffusion Probe-guided approach. 'Single Generation' serves as the baseline, representing the cost of one complete image generation (14.70s), while the subsequent row highlights the negligible overhead of a single probe prediction (+0.05s). We evaluate two scenarios: a **10-candidate Seed Selection** and a **4-candidate Prompt Optimization**, demonstrating significant savings in both latency and FLOPS.

Task	Workflow	Cost Breakdown	Total FLOPS (T)	Total Latency (s)
Reference	Single Generation	1 × Full Gen.	1877.56	14.70
	Single Gen. + 1 × Pred.	1 × Full Gen. + 1 × Pred.	1877.57	14.75
Seed Selection	Naive Oversampling	10 × Full Gen.	18775.60	147.00
	Diffusion Probe Guided (Ours)	**1 × Full Gen. + 10 × Pred.**	**5280.43**	**42.62**
Prompt Optimization	Naive Comparison	4 × Full Gen.	7,510.25	58.00
	Diffusion Probe Guided (Ours)	**1 × Full Gen. + 4 × Pred.**	**3026.42**	**28.29**

Table 4. Diffusion Probe's Predictive Accuracy for External Image Quality Metrics across Diffusion Steps near 5, evaluated on two resolutions (1024×1024 and 512×512). This table reports the predictive alignment between the Diffusion Probe's internal scores and standard external image quality metrics under different input resolutions. Higher values indicate better predictive accuracy.

Base Model	Resolution	Step	SRCC ↑	AUC-ROC ↑	KTC ↑	PCC ↑
FLUX	1024×1024	3	0.69	0.83	0.53	0.68
		4	0.75	0.88	0.60	0.74
		6	0.77	0.90	0.62	0.70
		7	0.74	0.87	0.58	0.72
	512×512	3	0.61	0.78	0.46	0.60
		4	0.66	0.81	0.50	0.64
		6	0.69	0.83	0.52	0.67
		7	0.67	0.82	0.51	0.65

Table 5. Diffusion Probe's Predictive Accuracy for External Image Quality Metrics across Diffusion Step 5 (FLUX Model, 1024×1024 Resolution) with Different Label Categories. This table reports the predictive alignment between the Diffusion Probe's internal scores and external image quality metrics, trained under two different label categories: Aesthetic Score and CLIP-SCORE. Higher values indicate better predictive accuracy.

Base Model	Resolution	Step	Label Category	SRCC ↑	AUC-ROC ↑	KTC ↑	PCC ↑
FLUX	1024×1024	5	**Aesthetic Score**	0.74	0.82	0.58	0.73
		5	**CLIP-SCORE**	0.77	0.84	0.62	0.76

able tool for quality assessment. The alignment with established metrics further validates the probe's predictive accuracy, emphasizing its potential as a robust quality evaluator for generated images.

5.5. Computation Cost Analysis

We show the computation cost in Table 3. A hallmark of Diffusion Probe is its exceptional computational efficiency. As detailed in Table 3, a single probe prediction requires only 0.0036 TFLOPS and 0.05s—orders of magnitude less than the 1877.56 TFLOPS and 14.70s demanded by a full generation.

This efficiency enables dramatic accelerations in real-world workflows. For a 10-candidate Seed Selection task, our probe-guided approach slashes latency from 147.00s to 42.62s, resulting in a **3.45**× speedup. Similarly, in a 4-candidate Prompt Optimization task, it reduces latency from 58.00s to 28.29s, yielding a **2.05**× speedup.

By strategically substituting expensive full-generation rollouts with near-instantaneous probe predictions for all but the final candidate, Diffusion Probe provides substantial computational and temporal savings, validating its role as a practical and powerful tool for optimizing large-scale generative workflows.

5.6. Robustness and Generalization Analysis

To verify the reliability of our proposed probe, we conduct a comprehensive robustness analysis across various dimensions, including model architectures, sampling trajectories, and semantic complexity. Although trained exclusively on **FLUX**, the following results demonstrate its exceptional zero-shot generalization.

Table 6. **Robustness across Architectures, Resolutions, and Steps.** Our probe is fixed (trained on FLUX). Results across IR and CLIP metrics.

Model	Res.	N	Ext. t	Met.	SRCC	AUC	PCC	KTC
SDXL	768	25	5	IR	0.70	0.81	0.51	0.69
SDXL	768	25	5	**CLP**	**0.63**	**0.74**	**0.42**	**0.61**
			5	IR	**0.79**	**0.91**	**0.64**	**0.76**
			5	**CLP**	**0.72**	**0.82**	**0.56**	**0.71**
FLUX	2048	25	21	IR	0.76	0.88	0.61	0.72
			22	IR	0.75	0.89	0.60	0.73
			24	IR	0.74	0.88	0.59	0.70

Table 7. **Generalization and Human Alignment.** Zero-shot performance across prompt complexities and sampling steps.

Prompt	Res.	N	Ext. t	SRCC	AUC	PCC	KTC	User
		10	2	0.73	0.82	0.71	0.58	79.2%
Simple	2048	25	5	0.72	0.87	0.71	0.56	83.5%
		50	10	0.74	0.88	0.72	0.57	81.3%
		10	2	0.69	0.79	0.67	0.53	76.0%
Complex	2048	25	5	0.68	0.81	0.66	0.52	79.1%
		50	10	0.70	0.83	0.68	0.54	77.5%

Universal Generalization across Architectures and Steps. As detailed in Tab. 6 and Tab. 7, the probe exhibits significant structural and temporal invariance. It maintains **consistent robustness** across diverse backbones and pixel densities, capturing universal quality signals rather than over-fitting to model-specific artifacts. Notably, the performance remains stable across different sampling budgets (**10, 25, or 50 steps**) at equivalent noise levels. This *step-invariance* proves that the probe relies on SNR-linked features rather than specific sampling trajectories. Furthermore, without fine-tuning, the probe generalizes to **CLIP Score (SRCC 0.72)**, effectively distilling universal features such as aesthetic quality and semantic alignment.

Architectural Insights and Prompt Complexity. Our design choices are grounded in structural necessity. Ablation studies show that cross-attention significantly outperforms self-attention (**SRCC 0.76 vs. 0.61**) in capturing semantic-structural alignment. The probe also sustains high performance (**AUC** > 0.78) even when processing long, multi-object prompts. When filtering by specific Parts-of-Speech (POS) tags (e.g., nouns and adjectives), accuracy drops by **only** $\sim$**10%**, suggesting a reliance on holistic context rather than isolated keywords. Additionally, a **three-head design** maintains high accuracy (**SRCC** $\approx$ **0.76**), enabling multi-dimensional assessment of both alignment and aesthetics.

Efficiency vs. Trajectory Robustness. The probe provides valid predictive power as early as **Step 3/25**, while remaining accurate (**SRCC** > 0.70) for artifacts emerging late in the trajectory (e.g., $t \in [21, 24]$). Identifying these "bad cases" early **drastically reduces inference costs** by avoiding redundant decoding. Regarding image diversity, the number of semantic clusters (N_{cls}) remains stable post-optimization (**5.3** $\rightarrow$ **5.2**), confirming that our method prunes low-quality samples while **preserving generative variety** without inducing mode collapse.

Correlation with Human Perception. Finally, our **user study** (Tab. 7) confirms a **74% agreement** with human preferences. This validates the probe as a **reliable perceptual proxy** for real-world generative applications, ensuring that the automated quality assessment aligns with actual human judgment.

5.7. More failure modes of generated images and corresponding cross attention maps

As shown in Figure 8 and Figure 7. In addition to the examples discussed in the main text, we further illustrate two representative failure modes of generated images: attribute mismatch and quantity mismatch. For attribute-mismatch cases, we find that the cross-attention maps corresponding to the specific attribute tokens become noticeably diffuse, suggesting that the model is unable to localize the intended attribute to the correct visual regions.

In contrast, quantity-mismatch cases exhibit different attention behaviors. When the model generates fewer objects than specified, the attention associated with the quantity or category tokens typically collapses onto a single region, indicating a failure to distribute attention across multiple instances. Conversely, when the model produces more objects than required, the attention maps often become fragmented into several weak hotspots. These patterns highlight how deviations in attention allocation correlate with different forms of generation errors.

6. Details about the experiments

6.1. Details about our metric

To quantitatively assess the performance of our **Diffusion Probe**, we follow a carefully controlled evaluation pipeline to compute the correlation and classification metrics. The procedure ensures both fairness and reproducibility.

1. **Ground-Truth Data Preparation:** For a given set of prompts, we first generate a large collection of final images using the base diffusion model. Each rendered image is then evaluated using a pre-trained aesthetic scoring model (e.g., the LAION aesthetic predictor) to obtain the ground-truth quality score S_{gt}. **To avoid distributional collapse in the test set—where scores might cluster heavily within a narrow interval—we deliberately adjust the selection of test samples so that the resulting distribution of S_{gt} spans a broad range of quality levels, including both low- and high-quality**

images. This ensures that evaluation metrics are not biased toward any particular score region.

2. **Probe Prediction:** During the generation process of the same set of images, our Diffusion Probe is activated at an early stage (typically within the first 10–20% of denoising steps). The probe analyzes intermediate attention features and outputs a predictive score S_{probe} for each instance.
3. **Metric Computation:** Given the paired scores (S_{probe}, S_{gt}), we compute the evaluation metrics as follows:
   - **SRCC and KTC:** We compute the Spearman Rank Correlation Coefficient (SRCC) and Kendall Tau Coefficient (KTC) between the vectors of all S_{probe} and S_{gt} values. These metrics measure the consistency of the probe's predicted ranking with the ground-truth ranking.
   - **AUC-ROC:** To assess classification performance, we binarize the ground-truth scores using the median of all S_{gt} values as the threshold. Images above the median are labeled as high-quality (class 1), and those below as low-quality (class 0). Using these binary labels and the probe's continuous predictions S_{probe}, we compute the Area Under the ROC Curve (AUC-ROC), which reflects the probe's ability to separate high- and low-quality samples.

This controlled process—especially the balanced construction of the ground-truth distribution—ensures that our evaluation reliably reflects the predictive capability of the Diffusion Probe across a wide spectrum of image qualities.

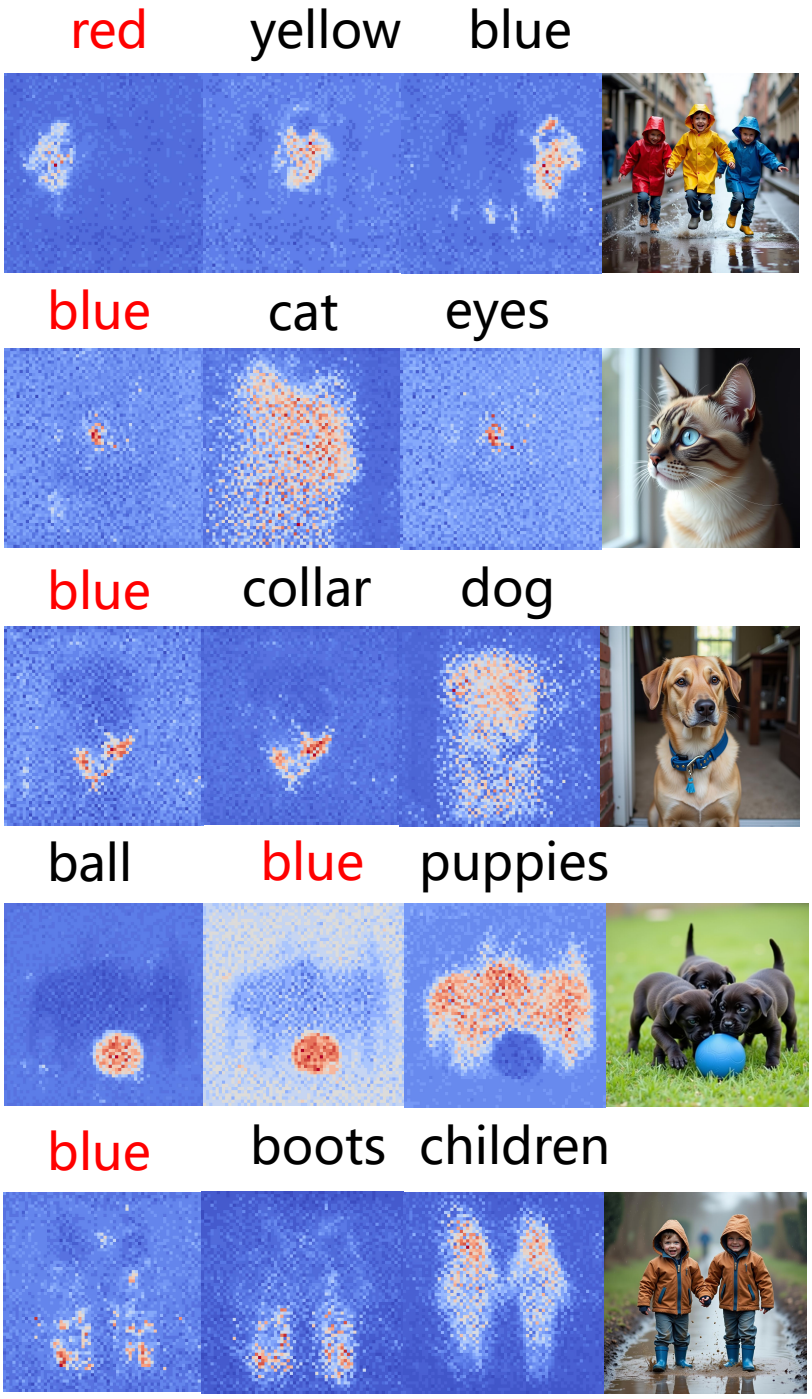


Figure 7. Normal case illustrating generated images and their corresponding cross-attention maps when the image generation is successful with no quality issues. In these cases, the cross-attention maps are well-focused, highlighting the specific regions of the image that correspond to the key features of the prompt. This indicates that when the generated image quality is high, the attention mechanism remains concentrated on the relevant visual attributes, reflecting the model's proper alignment with the textual description.

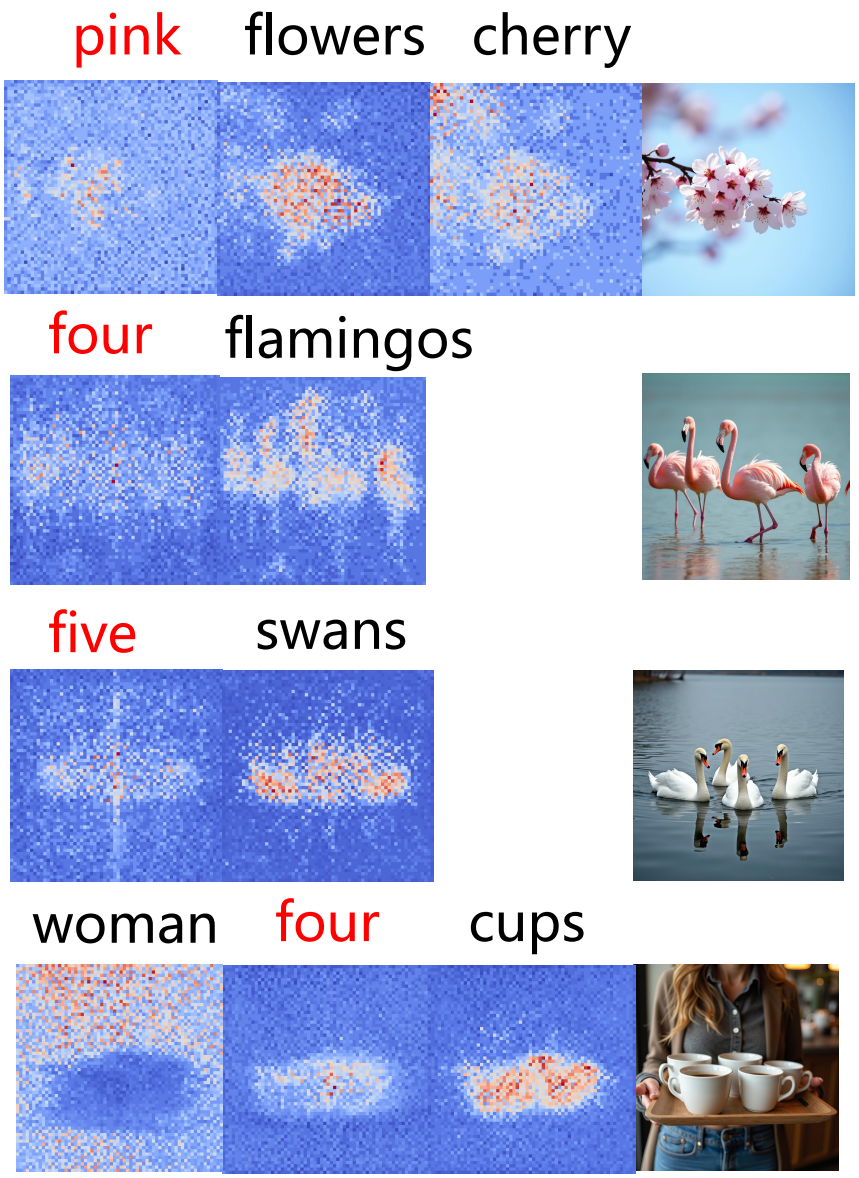


Figure 8. We list some cases of generation failures. When the generated image has attributes that do not match the prompt, the corresponding cross-attention map visualization exhibits a dissipation characteristic.

Figure 9. More qualitative results about Diffusion Probe

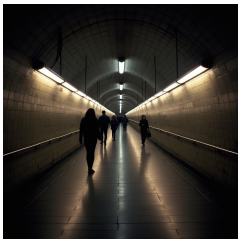
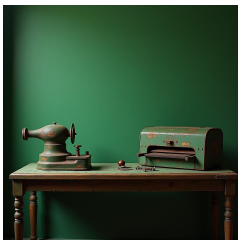
Prompt&Score	Good Cases	Prompt&Score	Bad Cases
An orange reddish rose in a vase filled with water on top of a table. Predicted Score: 1.45		A long subway with people in it is lit up. Predicted Score: 0.57	
A white vase with yellow tulips against a grey background. Predicted Score: 1.42		A GREEN TABLE WITH A OLD RUSTY BLENDER AND A PRINTER Predicted Score: 0.52	
A sandwich with salad on a plate and a cup of pop. Predicted Score: 1.39		Four giraffes look around a rock corner in their enclosure. Predicted Score: 0.55	
A woman in a blue dress with no shoes, seated with her legs crossed on a chair in the middle of a room. Predicted Score: 1.37		A couple fighting each other over a wii remote control. Predicted Score: 0.59	

Figure 10. More qualitative results about Diffusion Probe