DON'T GENERATE JUST TO DISCARD: PROMPT-SIDE AESTHETIC FORECASTING FOR TEXT-TO-IMAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Achieving satisfactory text-to-image (T2I) results often requires costly trial and error: users repeatedly revise prompts, generate images, and evaluate them manually or with quality assessment models, wasting computation and waiting time on unsatisfactory outputs. This leaves a practical but underexplored question: ***can we assess a prompt's quality potential before generating any images?*** Existing prompt-side studies find meaningful correlations between prompt embeddings and image quality averaged across repeated generations, suggesting that expected generation quality can be partially forecast from text alone. Yet limitations remain at three levels. At the quantification level, they collapse stochastic outcomes into an average score, obscuring variation across images from the same prompt; at the modeling level, coarse prompt-to-score mappings do not capture contextual interactions or address the cost of obtaining reliable multi-sample supervision; and at the interpretability level, average attribution makes it difficult to identify which prompt components improve or impair image quality. Motivated by these limitations, we formulate Prompt-Side Aesthetics Forecasting (PSAF), a pre-generation task that ***estimates a given generator's expected aesthetic response from prompt text alone***, and instantiate it with AesPrompt. AesPrompt uses learnable Aesthetic Response Queries to distill distributional aesthetic evidence from multiple generated samples into prompt-side representations, together with a Chunk-Aware Prompt Adapter that captures contextual interactions in long prompts. We further contribute PSAF-540K, a dataset of 540K samples designed for the PSAF task. During dataset construction, adaptive Bayesian response sampling directs the limited generation budget toward prompts expected to provide the most informative supervision. With only 95.7M trainable parameters, approximately 0.3% of a 30B-scale generator, AesPrompt achieves a Spearman rank correlation (SRCC) of 0.951 and a pairwise ranking accuracy (PairAcc) of 0.944, without sampling images at inference time. In typical prompt exploration workflows, filtering by AesPrompt can reduce rendered candidates and GPU time by 60%–85% while retaining high-quality prompts for final generation. The code, datasets, and demo are all included in the supplementary materials.

1 INTRODUCTION

Text-to-image (T2I) generators have become widely used tools for visual content creation, particularly in aesthetically demanding applications such as concept art, advertising design, game asset production, fashion visualization, and cinematic storyboarding. In these settings, prompt exploration typically follows an iterative, render-first process of aesthetic refinement: users, prompt engineers, or automated systems repeatedly adjust descriptions of composition, lighting, style, and rendering quality, generate images, evaluate their aesthetic quality, and retain only the prompts that yield satisfactory outputs. Because the aesthetic quality of the resulting images can usually be assessed only after generation, computational resources and time have already been consumed even when the outputs are aesthetically unsatisfactory. For computationally intensive or closed-source generators, this workflow turns prompt design into a costly feedback loop involving per-query charges, queueing delays, GPU computation, and manual review.

To reduce prompt exploration costs, a promising direction is to forecast a prompt's likely generation quality and avoid rendering low-quality candidates. As shown in Fig. 1, output-side methods use

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.

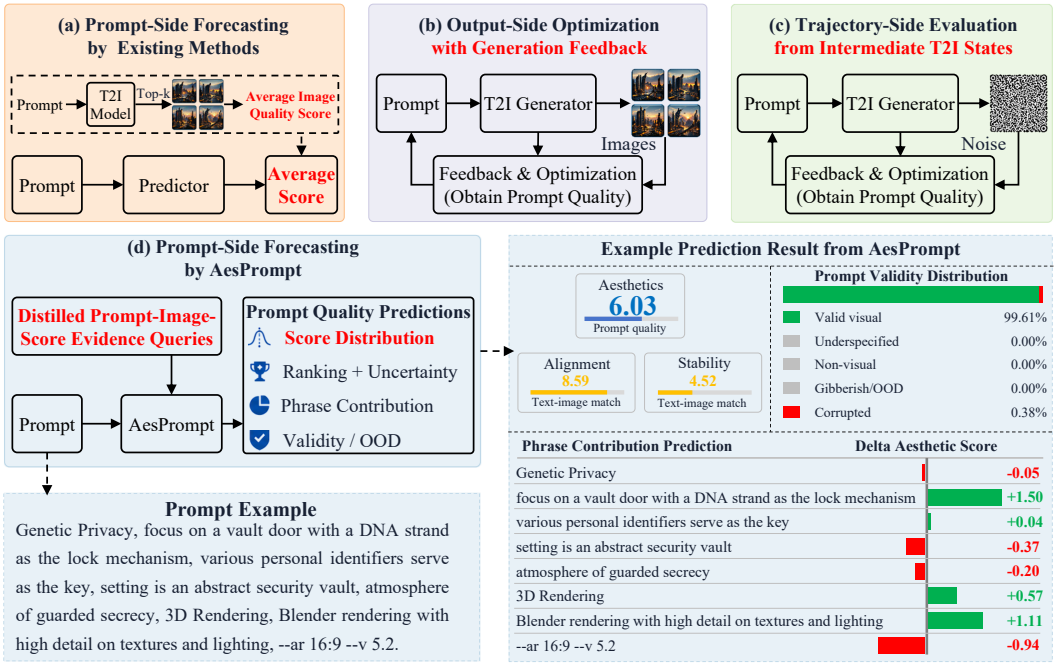

Figure 1: **Comparison of prompt quality forecasting paradigms.** (a) Existing prompt-side methods collapse stochastic outcomes into a global score. (b) Output-side methods require rendered images and feedback. (c) Trajectory-side methods require noise or intermediate generator states and assess individual generation paths. (d) AesPrompt distills prompt–image–score supervision and forecasts distributional aesthetic responses, phrase contributions, and validity/OOD risk from text alone without generation at inference.

rendered images or reward feedback to forecast prompt quality indirectly (Hao et al., 2023; Cao et al., 2023), but still incur generation costs. Trajectory-side methods use initial noise or intermediate generator states to predict the quality of individual seeds or diffusion paths (Kim et al., 2026; Kim & Mills, 2026; Cui et al., 2026), requiring partial generator execution and white-box access. Prompt-side methods instead forecast aggregate generation quality, human-judged performance, or generator suitability directly from text (Bizzozzero et al., 2024; Poesina et al., 2025; Lee & Lee, 2024), but reduce stochastic outcomes to global targets. This aggregation obscures quality variation across images generated from the same prompt and does not fully exploit prompt–image–quality evidence to model latent aesthetic responses and contextual phrase effects. Together, these limitations motivate Prompt-Side Aesthetics Forecasting (PSAF) in this paper and raise three key challenges.

1) Stochastic aesthetic response. PSAF seeks to infer a fixed generator's latent aesthetic response distribution from prompt text alone, without observing any realized images at inference time. This requires transferring distributional evidence learned from multiple training samples into representations of prompts while preserving differences in expected quality, variability, and the likelihood of outcomes with high aesthetic quality. Capturing stochastic response patterns specific to a generator is considerably more challenging than predicting a single aggregate score.

2) Context-sensitive aesthetic attribution of prompt phrases. The aesthetic quality of a generated image often depends strongly on a few prompt phrases describing lighting, composition, camera settings, style, or rendering quality. However, the contribution of a phrase cannot be assessed independently of the surrounding prompt: the same modifier may improve the generated result in one context but become redundant, conflicting, or harmful when combined with different subjects, styles, constraints, or negative clauses. Moreover, phrase-level effects may produce only small differences in aesthetic scores and can be obscured by noisy prompt-level supervision, making it difficult to identify which phrases contribute to the observed quality differences.

3) Budget-limited response sampling. Reliable prompt-level supervision would ideally be estimated from a large number of images generated and evaluated for each prompt. However, exhaustive generation is prohibitively expensive for computationally intensive or closed-source generators

because each query incurs rendering time, monetary cost, or queueing delay. Uniform sampling also wastes budget by repeatedly sampling prompts whose response estimates are already stable or whose additional samples are unlikely to change their relative rankings. The challenge is therefore to allocate a limited rendering budget to prompts for which additional samples would be most informative for estimating aesthetic responses and determining relative rankings.

To address these challenges, we introduce **AesPrompt**, a compact model for prompt aesthetic forecasting that uses only text and requires no image generation at inference time. First, to address the uncertainty caused by stochastic aesthetic responses, we introduce **Aesthetic Response Pretraining (ARP)**, which constructs a response memory that represents distributional information from prompt, image, and score evidence collected during training, and distills this memory into learnable Aesthetic Response Queries. Second, to capture contextual interactions among clauses in long prompts, we derive localized aesthetic contributions through occlusion that preserves context. Specifically, we measure the score change after masking one phrase while keeping the remaining prompt unchanged. The resulting contribution targets are then distilled into an efficient **Phrase Contribution Query Module**, enabling direct attribution at the phrase level without repeated occlusion. Third, to operate under a limited rendering budget, we formulate dataset curation as a resource allocation problem with access to the generator only as a black box, and use **adaptive Bayesian response sampling** to allocate additional generation queries where they are most informative for aesthetic response estimation and prompt ranking. Our key contributions are summarized as follows:

- **A new task for the T2I community.** We formulate the underexplored problem of assessing a prompt's aesthetic potential before image generation as PSAF. PSAF estimates the expected aesthetic response of a fixed T2I generator using only prompt text, moving aesthetic assessment from generated images to prompts before rendering.
- **A compact model for PSAF.** We propose AesPrompt, which distills costly aesthetic supervision from generated images into a compact predictor that uses only prompt text. Aesthetic Response Pretraining and learnable Aesthetic Response Queries model stochastic aesthetic responses, while the Phrase Contribution Query Module captures contextual phrase effects. To acquire reliable supervision within a limited rendering budget, adaptive Bayesian response sampling directs additional generation queries to the most informative prompts.
- **A large-scale PSAF dataset and benchmark.** We construct **PSAF-540K**, comprising 39,748 seed prompts, 120,000 long-prompt variants, 94,748 validity and OOD annotations, and 285,778 phrase-contribution records. Using this dataset, we evaluate 10 representative baseline methods and establish, to our knowledge, the first comprehensive benchmark for PSAF.
- **Strong performance and practical value.** With only 95.7M trainable parameters, approximately 0.3% of those in a 30B-scale generator, AesPrompt achieves strong rank agreement with PSAF score targets without image sampling, reducing the number of rendered candidates and GPU time by 60%–85% while retaining high-quality prompts for final generation. Together, PSAF, AesPrompt, and PSAF-540K provide a foundation for further research on efficient prompt assessment without image generation.

2 Related Work

To forecast prompt quality, existing approaches can be categorized by the generator-side evidence required for prompt quality forecasting: prompt text alone (prompt-side), completed generation outcomes (output-side), or partial generation trajectories (trajectory-side), as shown in Fig. 1.

Prompt-Side Forecasting. Prompt-side prompt quality forecasting aims to forecast T2I generation quality from prompt text alone before rendering. PPP (Bizzozzero et al., 2024) demonstrates the feasibility of this setting by predicting quality scores averaged across repeated generations from prompt representations. PQPP (Poesina et al., 2025) further introduces human judgments of generation relevance and realism together with pre-generation predictors, while other prompt-level quality predictors use textual features to select the most suitable generator for each prompt (Lee & Lee, 2024). These formulations mainly produce global scalar targets or generator-selection labels, **typically collapsing inherent stochastic variation across T2I generations into a single aggregate score.** Such aggregation discards the fact that the same prompt can induce diverse visual outcomes with different aesthetic qualities, resulting in the loss of distributional information that is essential for accurate

prompt-level quality estimation. Moreover, although these methods require only prompt text during inference, they do not exploit the rich cross-modal information contained in prompt-image-score pairs during supervision, where generated images reveal latent aesthetic responses and contextual phrase effects associated with each prompt. In contrast, AesPrompt distills multiple training observations linking prompts, images, and scores into a text-only predictor, preserving distributional cues and contextual phrase diagnosis while maintaining prompt-only inference.

Output-Side Optimization. Some prompt optimization methods learn to rewrite or enrich prompts from curated prompt pairs and image-based rewards, and iteratively render images to evaluate prompt candidates during search (Hao et al., 2023; Cao et al., 2023; Mo et al., 2024; Mañas et al., 2024; Yang et al., 2025). Other methods use generated-image feedback to rank candidate keywords or identify prompt words that are poorly represented in the resulting images (Zhang et al., 2024b; Liang et al., 2024). However, these output-side optimization methods aim to produce better prompts by diagnosing rendered outputs rather than directly optimizing the aesthetic quality of an existing prompt before image sampling. Whether such output-side feedback is obtained during training or inference, estimating prompt quality requires image generation and therefore incurs rendering costs. In contrast, AesPrompt eliminates these costs by optimizing directly from prompt text.

Trajectory-Side Evaluation. A complementary line of work predicts generation quality before the full diffusion process is complete. Noise-aware predictors evaluate human preference scores from a prompt together with sampled initial noise and use these estimates to select promising seeds (Kim et al., 2026; Kim & Mills, 2026). Early-stage predictors instead extract intermediate generator signals, such as cross-attention statistics from initial denoising steps, to forecast final image quality and terminate low-potential trajectories (Cui et al., 2026). These trajectory-side evaluation methods require sampled noise, intermediate generator states, or partial diffusion computation. In particular, methods that rely on intermediate states require access to internal activations and therefore effectively assume white-box access, whereas AesPrompt requires no white-box access to the generator.

3 Problem Formulation

To formally ground pre-generation prompt aesthetic evaluation, we cast the PSAF task into a text-centric probabilistic framework. Specifically, under a frozen T2I generator $\mathcal{G}$ with a fixed configuration, an input prompt p induces a latent aesthetic score distribution $\mathcal{P}_a(s \mid p)$ over possible generated images, where $s \in \mathbb{R}$ denotes the aesthetic score assigned to one sampled image. Rather than mapping to a deterministic scalar, the prompt induces this distribution due to the inherent stochasticity of diffusion sampling. Consequently, a robust prompt-level metric should encapsulate the holistic characteristics of $\mathcal{P}_a(s \mid p)$, including its central tendency, dispersion, and tail behaviors. To operationalize this paradigm under a finite computational budget, we approximate $\mathcal{P}_a(s \mid p)$ via a multi-dimensional, **distribution-aware response profile** $\mathbf{R}_{\mathrm{PSAF}}(p)$:

$$\mathbf{R}_{\mathrm{PSAF}}(p) = \Big(y_{\mathrm{PSAF}}(p),\, s_p^{\max},\, s_p^{\min},\, \sigma_p,\, \pi_p\Big), \quad \text{where } \; y_{\mathrm{PSAF}}(p) = \frac{1}{K_p}\sum_{k=1}^{K_p} s_{p,k}. \tag{1}$$

Here, the profile variables explicitly parameterize the empirical properties of $\mathcal{P}_a(s \mid p)$. The image set for prompt p is denoted as $\mathcal{X}_p = \{\mathbf{x}_{p,k}\}_{k=1}^{K_p}$, where K_p is the number of rendered samples and $s_{p,k} = \mathcal{A}(\mathbf{x}_{p,k}, p)$ is the aesthetic score assigned by the image-level evaluator $\mathcal{A}(\cdot)$. The mean response $y_{\mathrm{PSAF}}(p) \in \mathbb{R}$ estimates $\mathbb{E}[s]$ from these samples; $s_p^{\max}$ and $s_p^{\min}$ bound the aesthetic ceiling and floor; σ_p quantifies distributional dispersion (generation instability); and $\pi_p = \frac{1}{K_p}\sum_{k=1}^{K_p} \mathbb{I}[s_{p,k} > \tau_{\mathrm{HQ}}]$ denotes the empirical upper-tail probability above the high-quality threshold τ_{HQ}.

Accordingly, rather than relying on naive scalar regression, we reframe the core challenge as jointly modeling and fitting ***the latent statistical properties of*** $\mathcal{P}_a(s \mid p)$ ***purely from textual inputs, bypassing image rendering during inference.***

4 Adaptive Bayesian Response Sampling

To construct the PSAF dataset, fully characterizing the stochastic response of the generator $\mathcal{G}$ would require extensively sampling each prompt with a large number of generated images ($K_p \gg 1$), which

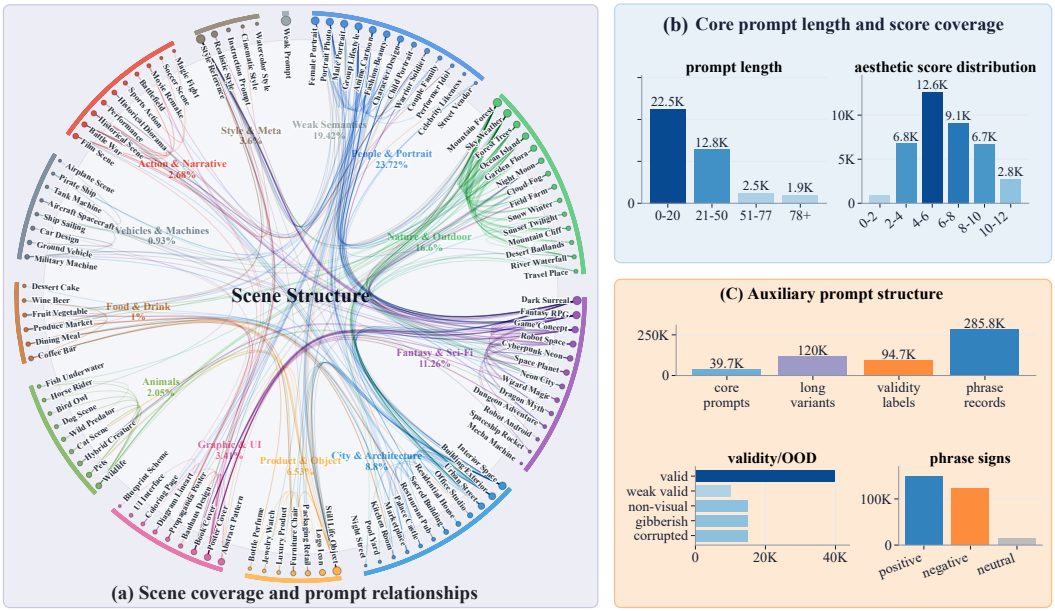

Figure 2: **Comprehensive statistical breakdown of PSAF-540K.** The curated prompt library encapsulates a highly diversified spectrum of semantic concepts, compositional prompt lengths, stratified PSAF scores, structural validity taxonomies, and localized phrase-level attribution labels.

is computationally prohibitive at the scale of PSAF-540K. We therefore formulate dataset curation as a resource-constrained optimization problem and introduce adaptive Bayesian response sampling to allocate rendering budgets more efficiently. Given an initial per-prompt rendering budget K_0, each rendered image provides a noisy observation of the unknown response distribution of $\mathcal{G}$ under prompt p. At iteration t, we fit a lightweight Bayesian surrogate over the accumulated history $\mathcal{D}_t$:

$$q_\phi(r_p \mid p, \mathcal{D}_t) = \mathcal{N}\big(\mu_\phi(p), \sigma_\phi^2(p)\big), \tag{2}$$

where r_p denotes the latent prompt-level response summarized by $y_{\mathrm{PSAF}}(p)$, $\mu_\phi(p)$ estimates its posterior mean, and $\sigma_\phi^2(p)$ captures posterior uncertainty and response variability induced by stochastic generation. We then use an acquisition function to select prompts for additional rendering, prioritizing prompts with high upper-confidence response estimates, large epistemic uncertainty, high aleatoric variance, or proximity to ranking decision boundaries. After the selected prompts are rendered, the new prompt-image-score observations are added to $\mathcal{D}_t$, and the surrogate is updated for the next allocation round. Crucially, this surrogate network is strictly restricted to the data collection phase and remains decoupled from downstream PSAF modeling. Finally, following the above adaptive sampling strategy, we construct PSAF-540K (Fig. 2), with detailed dataset construction and annotation procedures provided in Appendix A.1.

5 Architecture of AesPrompt

The architectural design of AesPrompt (Fig. 3) is grounded in three core considerations: ***1) Contextual Attribution.*** Capturing a fine-grained understanding of the non-linear aesthetic contributions yielded by both isolated phrases and global prompts (Sec. 5.1); ***2) Stochastic Mitigation.*** Leveraging dense cross-modal evidence to alleviate the inherent stochasticity of diffusion sampling and text-only estimation uncertainty (Sec. 5.2); and **3) Diagnostic Alignment.** Synthesizing multi-task predictive signals under ranking loss to optimize training robustness and PSAF performance (Sec. 5.3).

5.1 Prompt Representation Learner (PRL)

Prompt Chunking. To capture the aesthetic nuances between isolated phrases and holistic prompts, our Prompt Chunking module introduces a dynamic parsing mechanism. For a concise prompt, it is

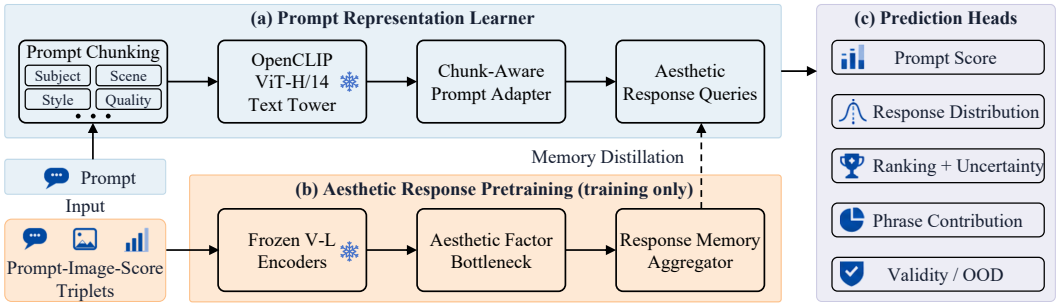

Figure 3: **Overview of AesPrompt.** AesPrompt contains (a) a Prompt Representation Learner that maps text prompts into response-aware query features, (b) training-only Aesthetic Response Pretraining that distills prompt–image–score evidence into a prompt-level response memory, and (c) Prediction Heads for score, distribution, ranking uncertainty, phrase contribution, and validity/OOD risk. Inference is prompt-only and uses no image generation or online image scoring.

encoded directly by the frozen text backbone to yield token-level features $\mathbf{H}_p = E_T(p)$. Conversely, ***a long prompt is partitioned into a prompt-specific sequence of semantic chunks*** $\mathcal{C}_p = \{c_j\}_{j=1}^{J_p}$, where J_p denotes the number of chunks for prompt p and each constituent chunk is augmented with learnable position and section-type embeddings to preserve its structural role and long-range contextual dependencies. Subsequently, these structured chunk segments are forwarded through the text backbone $E_T(\cdot)$ and aggregated into a unified backbone feature space, which we also denote as $\mathbf{H}_p$ for notation simplicity. To leverage rich pretrained generative priors while maintaining computational efficiency, this text backbone (OpenCLIP ViT-H/14) is kept frozen, focusing the downstream gradient optimization exclusively on the prompt-side adapters.

Chunk-Aware Prompt Adapter. Since existing CLIP encoders are optimized for cross-modal alignment rather than text-only aesthetic prediction, we introduce a trainable Chunk-Aware Prompt Adapter $F_\theta(\cdot)$ to refine and realign the semantic space $\mathbf{Z}_p = F_\theta(\mathbf{H}_p)$, where $\mathbf{Z}_p$ denotes the adapted prompt representation. Structurally, $F_\theta(\cdot)$ comprises a 6-layer transformer block tailored for token-level refinement, integrated with a chunk-level aggregation module that infuses spatial and functional priors. This design preserves the robust linguistic foundation of the frozen text tower while driving the trainable parameters to obtain aesthetic-predictive clauses.

Aesthetic Response Queries. To synthesize the refined, high-dimensional representation $\mathbf{Z}_p$ into a compact, unified prompt embedding $\mathbf{z}_p$, we employ a set of learnable aesthetic response queries via a cross-attention pooling mechanism. Rather than acting as passive pooling tokens, these queries are jointly optimized under score regression, distribution alignment, ranking, and memory distillation objectives. Consequently, they function as active semantic carriers that encapsulate latent quality ceilings, degradation risks, and variance cues distilled from the ARP stage (Sec. 5.2). The resulting comprehensive embedding $\mathbf{z}_p$ is subsequently distributed as the core conditioning vector to all downstream prediction heads detailed in Sec. 5.3.

5.2 Aesthetic Response Pretraining (ARP)

ARP serves as an auxiliary training phase formulated to ***distill cross-modal visual evidence into the prompt-side predictor***. Concretely, the Aesthetic Factor Bottleneck compresses high-dimensional prompt-image-score triplets into compact visual-preference factors, while the Response Memory Aggregator compiles these factors into a distillation memory that supervises the text-only model.

Aesthetic Factor Bottleneck. The bottleneck mechanism is motivated by the inherent stochasticity of T2I generation, where a singular scalar aesthetic score fails to capture why identical prompts can yield substantial visual variance across samples. For each generated image $\mathbf{x}_{p,k}$, we extract the prompt feature $\mathbf{t}_p = E_T(p)$ and image feature $\mathbf{v}_{p,k} = E_I(\mathbf{x}_{p,k})$, where $E_T(\cdot)$ and $E_I(\cdot)$ denote the frozen text and image encoders, respectively. AesPrompt then constructs an image-conditioned aes-

thetic factor $\mathbf{a}_{p,k}$ by fusing prompt semantics, visual characteristics, and cross-modal interactions:

$$\mathbf{a}_{p,k} = B_\psi\left(\left[\mathbf{t}_p,\, \mathbf{v}_{p,k},\, \mathbf{t}_p \odot \mathbf{v}_{p,k},\, |\mathbf{t}_p - \mathbf{v}_{p,k}|\right]\right), \tag{3}$$

where $[\cdot]$ denotes concatenation, $\odot$ represents element-wise multiplication, and $B_\psi(\cdot)$ is a parameterized bottleneck mapping. The bottleneck is optimized to predict image-level aesthetic scores while explicitly preserving within-prompt preferences. Specifically, for any pair of images generated from the same prompt satisfying $s_{p,i} > s_{p,j} + \Delta$, where Δ is the minimum image-level score gap for treating a preference pair as reliable, ***a margin ranking loss is enforced on*** $\mathbf{a}_{p,k}$ ***to capture the visual variations that dictate aesthetic quality***. Consequently, the bottleneck distills high-dimensional cross-modal evidence into compact, structurally aligned factors.

Response Memory Aggregator. To synthesize the stochastic behavior of a prompt into a unified signature, the Response Memory Aggregator compiles the sampled factors into a distribution-aware response memory $\mathbf{M}(p)$:

$$\mathbf{M}(p) = \left[\bar{\mathbf{a}}(p),\, \mathbf{a}_{\text{top}}(p),\, \mathbf{a}_{\text{low}}(p),\, \boldsymbol{\sigma}_a(p)\right], \tag{4}$$

where $\bar{\mathbf{a}}(p)$ represents the score-weighted mean factor, $\mathbf{a}_{\text{top}}(p)$ and $\mathbf{a}_{\text{low}}(p)$ characterize the upper (high-quality) and lower (degraded) response bounds, and $\boldsymbol{\sigma}_a(p)$ quantifies the generation variance. Compared to scalar targets, $\mathbf{M}(p)$ yields a richer optimization landscape by encoding the tail distributions and inherent uncertainty of the prompt-induced visual space. The prompt-side model is trained to reconstruct $\mathbf{M}(p)$ solely from text inputs via the joint objective:

$$\mathcal{L}_{\text{ARP}} = \mathcal{L}_{\text{score}} + \lambda_{\text{img}}\mathcal{L}_{\text{img-rank}} + \lambda_{\text{mem}}\mathcal{L}_{\text{mem}}. \tag{5}$$

Here, λ_{img} and λ_{mem} are nonnegative trade-off weights for image-level ranking and response-memory distillation. Crucially, upon completion of the ARP phase, the frozen visual encoders, bottleneck module, and memory aggregator are pruned from the inference graph. The distilled cross-modal knowledge is retained within the prompt-side representations and aesthetic response queries, ensuring a highly efficient, image-free deployment.

5.3 Prediction Heads and Loss Functions

The Prediction Heads are lightweight multilayer perceptrons (MLPs) operating on the shared prompt representation $\mathbf{z}_p$. They predict five diagnostic outputs: 1) the expected aesthetic score $\hat{y}_{\text{PSAF}}(p)$, 2) response statistics $(\hat{s}_p^{\max}, \hat{s}_p^{\min}, \hat{\sigma}_p, \hat{\pi}_p)$, 3) ranking uncertainty, 4) phrase contributions, and 5) validity/OOD logits. Sharing a response-aware representation encourages consistency among these outputs. To jointly learn aesthetic responses and their relative ordering, we integrate a pairwise ranking loss into a multi-task objective covering score regression, response statistics, memory distillation, phrase attribution, and validity/OOD prediction. The complete formulation is provided in Appendix A.4. At inference time, the ARP visual pipeline is removed.

6 Experiments

Baselines. To the best of our knowledge, no publicly available model is designed specifically for PSAF. We therefore construct prompt-only baselines from representative backbone families: prompt-side prompt quality forecasting methods (PPP, PQPP, CEMS), vision-language text encoders (LongCLIP, SigLIP-SO400M), generator-side text stacks (SDXL and FLUX), and large language or multimodal quality backbones (Qwen2.5, Q-Align, ROC4MLLM). To adapt them to PSAF, we introduce task-specific architectural modifications while preserving their core backbones. All methods use the same fixed PSAF-540K prompt split: 80%/10%/10% for training, validation, and testing.

Evaluation Metrics. Distributional performance is summarized by $\text{DistErr} = \frac{1}{5}\sum_{r\in\{y,s^{\max},s^{\min},\sigma,\pi\}} NormMAE(\hat{r}, r)$, which averages the normalized MAE over the expected score, maximum, minimum, standard deviation, and high-quality probability. For overall score prediction, PLCC measures linear agreement with the target scores, while SRCC and PairAcc evaluate relative prompt ordering. Top-10 Recall measures the proportion of ground-truth high-quality prompts retained among the top 10% of predicted candidates.

Table 1: Performance comparison of prompt-level aesthetic forecasting on the PSAF-540K test set. All methods operate strictly in a pre-generation, text-only manner. The best results are **bolded**.

Method	Params	Prompt-level forecasting				
		DistErr ↓	PLCC ↑	SRCC ↑	PairAcc ↑	Top-10 Recall ↑
PPP (Bizzozzero et al., 2024)	427M	0.068	0.753	0.763	0.810	0.380
PQPP (Poesina et al., 2025)	428M	0.068	0.813	0.823	0.844	0.451
CEMS (Lee & Lee, 2024)	150M	0.073	0.737	0.745	0.801	0.404
LongCLIP-L (Zhang et al., 2024a)	427M	0.069	0.846	0.853	0.865	0.519
SDXL (Podell et al., 2024)	817M	0.063	0.855	0.862	0.872	0.526
FLUX text stack	5.0B	0.061	0.880	0.885	0.887	0.554
SigLIP (Zhai et al., 2023)	878M	0.075	0.818	0.822	0.848	0.512
Qwen2.5-7B (Wang et al., 2024)	7.6B	0.059	0.879	0.883	0.884	0.549
Q-Align (Wu et al., 2023a)	8.20B	0.063	0.848	0.855	0.863	0.491
ROC4MLLM (Ma et al., 2026)	8.20B	0.059	0.852	0.857	0.865	0.460
AesPrompt	**95.7M**	**0.045**	**0.945**	**0.951**	**0.944**	**0.695**

6.1 Performance Evaluation

Performance Evaluations. As shown in Table 1, AesPrompt outperforms all generic text and multimodal backbones across all metrics. Three key observations emerge: **1) High Accuracy with Parameter Efficiency.** Despite its compact size of 95.7M parameters, AesPrompt achieves a 0.951 SRCC and 0.944 PairAcc, supporting the effectiveness of aesthetic response pretraining and ranking-oriented objectives. AesPrompt also achieves the lowest DistErr, indicating more accurate forecasting of distributional response statistics. Detailed results for individual response statistics and probabilistic calibration are provided in Appendix A.2. **2) Task Feasibility.** Competitive baselines such as FLUX and Qwen2.5 achieve meaningful ranking agreement (∼0.88 SRCC), demonstrating the feasibility of prompt-side aesthetic forecasting. **3) Pre-generation Filtering Utility.** Top-10 Recall directly measures practical utility under constrained rendering budgets. While the strongest baselines plateau around 0.55, AesPrompt achieves 0.695, indicating that it can discard roughly 90% of candidate renders while retaining a substantial proportion of high-quality prompts. A prompt-length analysis is provided in Appendix A.3, showing that very long prompts primarily challenge top-candidate retrieval rather than global ranking.

Cross-Generator Adaptation. Table 2 evaluates the transferability of AesPrompt across various T2I architectures. In zero-shot settings, performance decreases due to domain shifts caused by variations in style priors, safety filters, and text rendering behaviors. This baseline performance is also sensitive to the parameter scale and complexity of the target models. Despite these cross-domain variations, fine-tuning the model with approximately 1,000 samples improves both SRCC and PairAcc across all generator scales. This adaptation demonstrates the utility of AesPrompt as an efficient calibration tool that aligns text features with new generation engines without requiring online rendering.

Table 2: Cross-generator adaptation.

T2I generator	Scale	Zero-shot		1K-adapted	
		SRCC ↑	PairAcc ↑	SRCC ↑	PairAcc ↑
Sana-0.6B	0.59B	0.46	0.65	0.61	0.73
SD-1.5	0.9B	0.44	0.64	0.59	0.72
SD3.5-Medium	2.5B	0.38	0.61	0.55	0.70
SDXL	2.6B	0.41	0.63	0.57	0.71
FLUX.1-schnell	12B	0.35	0.59	0.52	0.68
Gemini Nano Banana	≈ 80B	0.32	0.58	0.50	0.67
ChatGPT Image 2	Unknown	0.29	0.59	0.43	0.59

Phrase Attribution Validation. For each phrase, we render both the original prompt and the phrase-removed prompt under the same fixed generator configuration, collect manual aesthetic scores for the resulting images, and use the aggregated score gap as the phrase contribution target. The phrase query head is evaluated by how well its predictions match these external occlusion-derived labels in contribution value and within-prompt ranking. It achieves an MAE of 0.372, a PLCC of 0.772, an SRCC of 0.554, and a within-prompt phrase ranking accuracy of 0.727. These results show that the head approximates the expensive re-rendering-based attribution protocol without repeated masking at inference time. Since phrase removal can also change global prompt semantics, we interpret these labels as context-dependent empirical marginal effects rather than strict isolated causal effects.

Ablation Study. As shown in Table 3, eliminating the ARP stage decreases the SRCC from 0.951 to 0.827. This reduction demonstrates that transferring cross-modal distribution boundaries during training is necessary for a text-only model to infer latent visual qualities from text alone. Additionally, removing the pairwise ranking loss lowers the Top-10 Recall to 0.548 and the PairAcc to 0.822. This change confirms that ranking-oriented supervision is required to calibrate the model for fine-grained prompt selection and high-quality candidate retrieval.

Table 3: Core component ablation.

Variant	PLCC ↑	SRCC ↑	PairAcc ↑	Top-10 Recall ↑
Full AesPrompt	**0.945**	**0.951**	**0.944**	**0.695**
w/o ARP	0.818	0.827	0.834	0.572
w/o ranking loss	0.801	0.814	0.822	0.548

Validity and OOD Risk Detection. To evaluate the validity head, we measure its ability to identify nonvisual, corrupted, and OOD prompts. The module achieves an AUROC of 0.740 for OOD detection and a top-1 accuracy of 0.719 when classifying validity across five categories. These results show that the module provides an informative risk signal for flagging nonvisual prompts or prompts affected by distribution shifts before generation.

Adaptive Sampling Ablation. We use 1,000 prompts with 32 renders each as a dense reference and replay data collection under restricted budgets. As shown in Figure 4, adaptive sampling achieves lower DistErr and higher Top-10 Recall than uniform, score-based, and variance-based allocation. With four renders per prompt, it approaches the label recovery of uniform sampling with eight renders while achieving higher top-candidate recall, suggesting comparable supervision quality with roughly half the rendering budget. We further provide a small-scale end-to-end retraining validation on a 1,000-prompt subset in Appendix A.5, showing that the labels collected by adaptive sampling are expected to provide stronger downstream supervision under the same rendering budget.

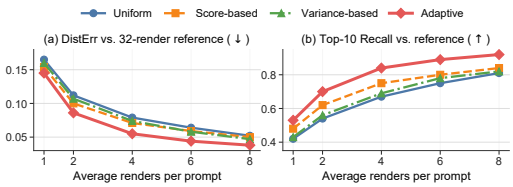

Figure 4: Adaptive sampling achieves lower DistErr and higher Top-10 Recall.

6.2 Complementarity with Prompt Optimizers

We evaluate AesPrompt as a pre-generation selector for Promptist (Hao et al., 2023) and BeautifulPrompt (Cao et al., 2023) on an exploratory subset of 200 prompts. Each optimizer generates eight variants, from which *Default* selects the first, *Random* samples uniformly, and *AesPrompt* selects the highest-scoring candidate for one Stable Diffusion 1.5 (Rombach et al., 2022) rendering.

Table 4: AesPrompt selection for Promptist and BeautifulPrompt on 200 prompts.

Optimizer	Selector	PickScore ↑	HPS ↑	CLIP Score ↑
Promptist	Default	20.448	0.259	0.258
Promptist	Random	20.281	0.256	0.251
Promptist	**AesPrompt**	**20.609**	**0.261**	**0.262**
BeautifulPrompt	Default	20.063	0.255	0.232
BeautifulPrompt	Random	19.770	0.251	0.222
BeautifulPrompt	**AesPrompt**	**20.323**	**0.259**	**0.243**

We evaluate the rendered images using PickScore (Kirstain et al., 2023), HPS (Wu et al., 2023b), and CLIP Score (Radford et al., 2021). Table 4 shows that AesPrompt improves all three metrics over default and random selection for both optimizers. Thus, AesPrompt complements prompt optimization by improving the selected render without increasing the rendering budget.

7 Conclusion

In this paper, we formulate PSAF to forecast a T2I generator's expected aesthetic response from prompt text and introduce AesPrompt, a compact model requiring no inference-time image generation. AesPrompt distills image-derived aesthetic evidence into learnable Aesthetic Response Queries and models contextual phrase effects through a Phrase Contribution Query Module. Adaptive Bayesian response sampling prioritizes informative prompts during dataset construction, while PSAF-540K supports aesthetic forecasting, robustness evaluation, and phrase-level attribution. Experiments show that AesPrompt accurately ranks prompts and reduces rendered candidates and GPU time while retaining high-quality prompts. The current formulation remains generator-specific and requires generated images for prompt-level supervision; future work could improve cross-generator transfer and reduce adaptation requirements.

AI USE STATEMENT

In this work, we used generative AI tools to improve the language and presentation of the manuscript, assist with method implementation, and support dataset organization. All AI-assisted outputs were reviewed and verified by the authors. We take full responsibility for the final manuscript, claims, experimental results, code, data, and other artifacts.

REFERENCES

Nicolas Bizzozzero, Ihab Bendidi, and Olivier Risser-Maroix. Prompt performance prediction for image generation. In *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 2208–2214. IEEE, 2024.

Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1–11, 2023.

Benlei Cui, Bukun Huang, Zhizeng Ye, Xuemei Dong, Tuo Chen, Hui Xue, Dingkang Yang, Longtao Huang, Jingqun Tang, and Haiwen Hong. Diffusion probe: Generated image result prediction using cnn probes. *arXiv preprint arXiv:2602.23783*, 2026.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.

Joong Ho Kim and Keith G Mills. Can we predict the human preference for text-to-image content prior to generation and is it even useful to do so? *arXiv preprint arXiv:2606.05478*, 2026.

Joong Ho Kim, Nicholas Thai, Souhardya Saha Dip, Dong Lao, and Keith G Mills. Naïve paine: Lightweight text-to-image generation improvement with prompt evaluation. *arXiv preprint arXiv:2603.12506*, 2026.

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.

Dongkeun Lee and Wonjun Lee. Learning prompt-level quality variance for cost-effective text-to-image generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3847–3851, 2024.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024.

Xingyuan Ma, Shuai He, Anlong Ming, Haobin Zhong, and Huadong Ma. Regression over classification: Assessing image aesthetics via multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7827–7835, 2026.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*, 2024.

Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. Dynamic prompt optimizing for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26627–26636, 2024.

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. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Eduard Poesina, Adriana Valentina Costache, Adrian-Gabriel Chifu, Josiane Mothe, and Radu Tudor Ionescu. Pqpp: A joint benchmark for text-to-image prompt and query performance prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28651–28661. IEEE, 2025.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023a.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023b.

Hongji Yang, Yucheng Zhou, Wencheng Han, and Jianbing Shen. Self-rewarding large vision-language models for optimizing prompts in text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7332–7349, 2025.

Zhichao Yang, Tianjiao Gu, Jianjie Wang, Feiyu Lin, Xiangfei Sheng, Pengfei Chen, and Leida Li. Longt2ibench: A benchmark for evaluating long text-to-image generation with graph-structured annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11820–11828, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pp. 310–325. Springer, 2024a.

Haiyang Zhang, Mengchao Wang, Shuai He, and Anlong Ming. Ak4prompts: aesthetics-driven automatically keywords-ranking for prompts in text-to-image models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 1661–1669, 2024b.