Prompt Stability Matters: A Benchmark for Quantifying Prompt Informativeness and Stability in Text-to-Image Models

Anonymous Authors[1]

Abstract

Recent Text-to-image (T2I) generative models have enabled users to produce strikingly realistic images from natural language prompts. However, we observe that prompt informativeness varies significantly across user proficiency levels, e.g., ambiguous or under-specified prompts often lead to unstable outputs that deviate from user intent, where limited efforts are made in the community to qualitatively investigate this phenomenon. Thus, we introduce **Authentic Prompt Benchmark (AP Bench)**, a large-scale benchmark of 17,580 authentic prompt-image pairs sourced from real-world web repositories, spanning from novice (e.g., short and informal words) to expert (e.g., highly detailed, professionally composed specifications) users. Unlike existing metrics that focus on prompt–image alignment, we position AP Bench as the first dedicated benchmark for investigating prompt-to-prompt transmission and evaluating user-oriented prompt stability in T2I generation. Building on the insights from our AP Bench, we further propose **NoxEye**, a novel end-to-end prompt optimization framework for enhancing T2I generation. Across AP Bench and other established benchmarks, NoxEye delivers improvements of up to 56.66% in mutual information, 18.71% in prompt entropy, and 19.98% in prompt energy. Importantly, we demonstrate that NoxEye can **genuinely improve authentic prompts written by real users**, serving as a plug-and-play framework that consistently boosts performances of existing state-of-the-art T2I generative models, as also verified by human evaluations. Our benchmark and model can be accessed at https://authpromptbench.github.io/.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Figure 1. A Taxonomy of Prompt-induced Instability. We show failure modes caused by various user prompt patterns (top), where a baseline prompt optimization method also fails (bottom).

1. Introduction

With the advent of generative models (Rombach et al., 2022; Ho et al., 2020; Ramesh et al., 2021; 2022; Saharia et al., 2022; Jiang et al., 2024; Lee et al., 2022; Han et al., 2025; Xie et al., 2025; Chen et al., 2025), text-to-image generation has become increasingly popular, enabling users to generate images based on a wide variety of textual prompts. The development of large language models (LLMs) has further enhanced this process by allowing for prompt tuning, leading to improved visual fidelity in the generated images (Hao et al., 2023; Wu et al., 2024; Yang et al., 2024).

Our rationale is motivated by the observation that most prior studies on prompt refinement (Hao et al., 2023; Cao et al., 2023; Rosenman et al., 2023; Wang et al., 2025) primarily focus on improving text–image alignment or aesthetic quality, while largely overlooking a fundamental question: **whether prompts written by real users can faithfully convey their underlying intent**. In practice, user prompting proficiency is highly heterogeneous, and such variability in prompt informativeness can substantially affect the stability of text-to-image generation. For example, expert users often craft long and detailed prompts with explicit attributes, whereas novice users tend to provide shorter, underspecified,

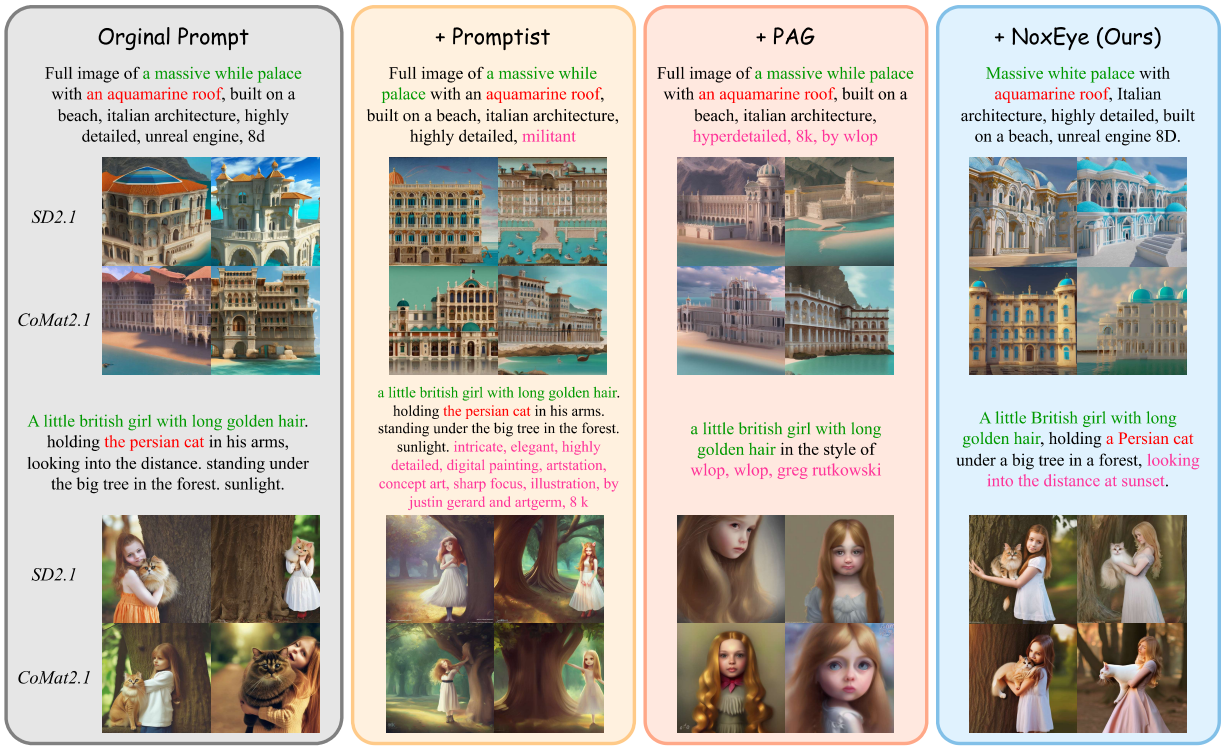


Figure 2. We observe that state-of-the-art generative models, such as Stable Diffusion 2.1 (SD2.1) (Rombach et al., 2022) and CoMat 2.1 (CoMat2.1) (Jiang et al., 2024), remain prone to generation instability under authentic user inputs. Our prompt-optimizing method, NoxEye, mitigates this limitation by systematically enhancing prompt–image alignment with user intent and outperforms the state-of-the-art prompt-refining method, e.g., Promptist (Hao et al., 2023) and PAG (Yun et al., 2025). Instead of stylistic injection caused by mode collapse (Yun et al., 2025), our approach focuses on semantic clarification and avoiding hallucinations.

and ambiguous descriptions.

Although recent approaches leverage LLMs to enhance or rewrite these authentic prompts for better expressiveness (Hao et al., 2023; Yun et al., 2025; Mo et al., 2024), we argue that such alignment-oriented optimization is often rather naive and may introduce **new failure modes**. In particular, we identify a taxonomy of prompt-induced instability commonly observed in real-world settings, as shown in Figure 1: (i) informational sparsity, where overly brief prompts lead LLM-based expansion to hallucinate unintended attributes and distort user intent; (ii) semantic imprecision and incompleteness, where non-expert or ambiguous descriptions omit critical specifications, causing optimized prompts to only partially reflect the desired semantics, as demonstrated in some pioneering research (Du et al., 2023; Chefer et al., 2023b); and (iii) lexical perturbation and noise sensitivity, where typos or minor word-level disturbances can mislead the text encoder and produce inconsistent or even opposite generations (Du et al., 2023). Together, these issues highlight that prompt optimization is not merely an alignment problem, but a robust intent-preservation challenge under realistic user conditions. Consequently, we argue that advancing the stability of T2I systems requires both the design of interpretable prompts and the development of robust evaluation metrics.

Thus, we cast the T2I prompt stabilizing problem as a prompt-to-prompt distribution matching problem, where the goal is to train a model such that the conditional distribution of the optimized prompt given user authentic prompts better resembles user intents. To this end, we propose **NoxEye**, a plug-and-play modular prompt optimization framework broadly compatible with diverse T2I models. The training of Noxeye proceeds in two stages: (1) Supervised fine-tuning, wherein a limited but high-quality dataset is leveraged to enable the LLM to learn alignment with the target task distribution (Zhou et al., 2023); and (2) GRPO-style on-policy preference optimization, which incorporates expert evaluations to guide the model toward reduced hallucination tendencies and to mitigate the squeezing effect (Shao et al., 2024; Ren & Sutherland, 2025).

To facilitate the study for the community, we additionally release a novel benchmark, named Authentic Prompt Benchmark, for assessing the stability of T2I generation through the lens of information propagation. Unlike prior efforts, our benchmark not only measures the alignment between gen-

erated images and user intent but also explicitly quantifies the informational content embedded in prompts. To support this, we curate a dataset of 17,580 real-world prompts crawled from authentic web cases (from which 2,048 were selected for evaluating), stratified into *novice* and *expert* subsets according to user proficiency in T2I prompt design.

Extensive experiments demonstrate that NoxEye effectively mitigates the adverse effects of ambiguous prompts, substantially improving both fidelity and stability of generated images. When compared to the open source state-of-the-art prompt optimization strategies, our framework yields 56.66%, 18.71% and 19.98% gains in mutual information, prompt entropy, and prompt energy metrics. Meanwhile, NoxEye can advance generative ability on existing T2I models. Our framwork excels both in easy and diffucult tasks such as "Two Object" (+5.23% over Flux.1 (Labs, 2024)), "Counting" (+5.10%), "Colors" (+6.07%) and "Position" (+19.39%).

2. Related Work

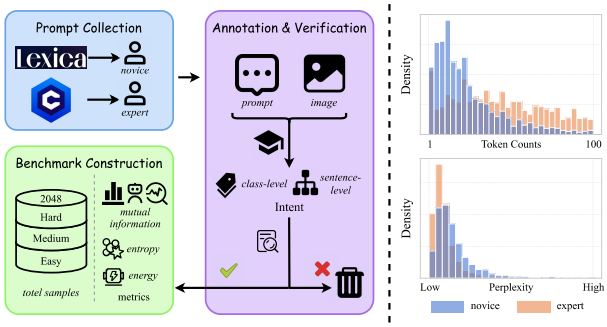
2.1. Text-to-Image Generation and Benchmarks

Text-to-image generation has progressed from early GANs (Goodfellow et al., 2014) and VAEs (Kingma & Welling, 2013) to diffusion-based models, with Stable Diffusion (Rombach et al., 2022) and CoMat (Jiang et al., 2024) exemplifying the current paradigm. These models typically employ frozen text encoders (e.g., CLIP (Radford et al., 2021)) to map prompts to embeddings that guide iterative denoising.

Benchmarking efforts have evolved alongside model capabilities. HEIM (Liang et al., 2022) evaluates twelve dimensions, including alignment, quality, reasoning, and fairness. T2I-CompBench (Huang et al., 2025) focuses on compositional generation with novel metrics and reward-driven fine-tuning (GORS), while GenEval (Ghosh et al., 2023) introduces object-centric evaluation for fine-grained analysis. Despite these advances, models still struggle to capture user intent accurately, motivating the proposed Authentic Prompt Benchmark for mapping ambiguous prompts to concrete object representations.

2.2. Prompt Optimization for Text-to-Image Generation

Some researchers have noticed that prompt design plays an essential role in making the model better understand our intentions and producing higher-quality results(Hao et al., 2023). Simultaneously, prompt optimization leverages LLMs (Schlegel et al., 2025; Xiang et al., 2025) to improve generated image quality. Promptist (Hao et al., 2023) fine-tunes GPT-2 (Radford et al., 2019) to reformulate user prompts via supervised fine-tuning (SFT) and direct preference optimization (DPO) using CLIP similarity and aesthetics scores. PromptCoT incorporates the Chain-of-



(a) A roadmap to Auth-Prompt Bench

(b) An visual comparision between novice and expert

Figure 3. Instability of real-world prompts and importance of AP Bench. We curate a dataset of 17,580 real-world prompts collected from authentic web sources, stratified into novice and expert subsets, and show that novice prompts exhibit more tokens and higher perplexity compared to expert prompts. As shown in (a), the prompt from the novice user leads to unstable T2I outcomes, and our method can mitigate the issue by aligning the prompt to a model-friendly distribution.

Thought (CoT) mechanism to learn high-quality prompt expressions (Yao et al., 2024). Self-Rewarding LVLMs (Yang et al., 2025) extend this two-stage paradigm with a self-reward mechanism, while PAG (Yun et al., 2025) uses GFlowNets to generate diverse adaptive prompts.

These approaches enhance prompt quality but largely focus on aesthetic and relevance objectives, often neglecting whether generated images faithfully reflect the user's underlying intent, the illusion problem of large language models (Kalai et al., 2025) and time cost (Venkatesh et al., 2025).

In addition, prior studies have explored prompt optimization for image generation through prompt embeddings and dynamically controlled prompts. For example, LLM4GEN (Liu et al., 2025) enhances CLIP embeddings by leveraging representations from large language models, thereby improving the generative performance of Stable Diffusion (Stability AI, 2023) models, while PAE (Mo et al., 2024) employs dynamically controlled prompts to guide the denoising process for more refined image synthesis. Although these approaches demonstrate effectiveness in specific settings, their generalizability is limited, and they cannot be readily extended to flow-matching frameworks or autoregressive generative models.

3. A Roadmap to AP Bench

3.1. Underlying Rationale

Our rationale stems from a key observation: due to user proficiency level, the informativeness of prompts exerts a substantial impact on the quality of text-to-image (T2I) generation, while the community has paid limited attention to systematically addressing this issue. Motivated by this gap, our objective is to design a comprehensive and principled methodology and benchmark for evaluating the quality of

prompts in T2I generation.

Inspired by recent preliminary work(Wang et al., 2025), we demonstrate how we quantify the informativeness of a text-to-image prompt P in conveying user intent Y to a generative model ϕ producing image I, using three carefully designed measures from an information-theoretic framework.

3.1.1. Mutual Information for User Intent Alignment

We model generation as a Markov chain $Y \to P \to I$, and define prompt stability via mutual information:

$$I(Y; I) = H(Y) - H(Y \mid I), \tag{1}$$

where larger $I(Y; I)$ indicates better preservation of user intent. Based on Equation 1, we use three metrics to characterize the stability of prompt from the perspective of mutual information to avoid the bias of a single evaluation metric.

CLIP Classification Accuracy (Mean). Following prior work (Du et al., 2023; Feng et al., 2022; Chefer et al., 2023a), user intent is approximated via **entity–template expansion**, $Y(e) = \{t_k(e)\}_{k=1}^{K}$, allowing stability assessment over a distribution of plausible prompts. For each prompt P, generate n images $\{I_1, \ldots, I_n\}$ and classify them into a set of $\mathcal{C}$, predefined categories using CLIP. Denote the predicted category of image I_j as $\hat{Y}_j$. The classification accuracy is defined as

$$\text{Acc}_{\text{mean}}(P) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\left[\hat{Y}_j = Y_c\right],$$

where $\mathbb{I}[\cdot]$ is the indicator function and Y_c is the real user intent category. Higher mean accuracy corresponds to lower conditional entropy $H(Y \mid I, P)$ and higher $I(Y; I \mid P)$.

CLIP Classification Accuracy (Standard Deviation). To capture variability across generated images, we compute the standard deviation of the classification results:

$$\text{Acc}_{\text{std}}(P) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \left\{ \mathbb{I}\left[\hat{Y}_j = Y_c\right] - \text{Acc}_{\text{mean}}(P) \right\}^2}.$$

This metric approximates the uncertainty or conditional entropy in the generated distribution. A high standard deviation indicates that the same prompt can produce semantically divergent outputs, reflecting ambiguity in the encoding P of the user intent Y.

MLLM Alignment Score. For a more fine-grained semantic alignment, we employ multimodal LLMs (MLLM) to score each generated image I_j against the short textual description of Y, producing a soft score $s_j \in [0, 1]$. The average score over n images is

$$S(P) = \frac{1}{n} \sum_{j=1}^{n} s_j.$$

This score provides a soft, continuous approximation of the mutual information between the full intent description and the generated images.

More details on how to derive these metrics from mutual information can be found in the Appendix A.1.

3.1.2. Prompt Entropy for T2I Reliability Assessment

By the data-processing inequality:

$$I(Y; I) \le I(Y; P),$$

indicates that the maximum achievable stability is constrained by the prompt information content and reveals the importance of prompt optimization.

To quantify the informativeness of user inputs (Farquhar et al., 2024; Cheng et al., 2025; Duan et al., 2023), we introduce the notion of **prompt entropy**. Intuitively, novice users often provide under-specified or ambiguous prompts that lack sufficient detail, making them harder to interpret and yielding unstable generations. In contrast, expert prompts tend to be more specific and constrained, thereby concentrating information and reducing uncertainty.

Thus, we introduce the T2I Prompt entropy $H(P)$ reflects the inherent information content of P:

$$H(P) \approx -\frac{1}{T} \sum_{t=1}^{T} \log p_\theta(w_t \mid w_{<t}),$$

where p_θ is a pretrained LM. Lower entropy prompts are more predictable, concentrate information, and typically yield more stable generations. See Appendix A.2 for derivation and theoretical connection to $I(Y; I)$.

3.1.3. Prompt Energy for T2I Stability Assessment

Existing stability metrics for text-to-image generation, such as mutual information or prompt entropy, capture either end-to-end information transfer or prompt *aleatoric uncertainty*, but fail to capture the model's *epistemic uncertainty*—uncertainty stemming from the model's lack of knowledge (Ma et al., 2025). To address this, we introduce **prompt energy** as a complementary measure: prompts with low energy correspond to concepts well-represented in the model, yielding stable generation, while high-energy prompts indicate unfamiliar or uncertain concepts. Formally,

Table 1. Example structure of AP Bench. Prompts are sourced from raw, real-world web cases, ensuring authenticity and diversity of user prompting.

Intent		**User Type**	**Challenge**	**Prompt**
class-level	*sentence-level*			
garbage truck	a futuristic blue garbage truck	Novice	Easy	New clean cyberpunk rubbish truck, blue colour
–	a studio photograph of a ripe tamarillo	Novice	Medium	Ripe tamarillo fruit in vibrant red, in photo studio
slug	a portrait of Sucy Manbavaran in a meadow with slugs	Expert	Hard	slugs, slug swarm, female character concept, fabulous artwork, best quality, high resolution, split-complementary color scheme, red sweater, black pants and sneakers, serene meadow, Sucy Manbavaran from *Little Witch Academia*

for a prompt sequence $x = (x_1, \ldots, x_T)$, the normalized sequence energy is

$$E(x) = -\frac{1}{T}\sum_{t=1}^{T} z_t(x_t),$$

where $z_t(x_t)$ denotes the model-assigned logit for token x_t. Lower $E(x)$ indicates higher confidence, whereas higher $E(x)$ signals uncertainty.

Combining prompt-level entropy and energy with end-to-end metrics such as $I(Y; I)$ provides a more comprehensive characterization of generation stability, directly linking user-provided information to image fidelity. Implementation details, derivations from model logits, and the connection to classical Boltzmann energy are provided in Appendix A.3.

3.2. Benchmark Construction

Building on our information-theoretic formulation, we design a benchmark to empirically evaluate prompt informativeness and generation stability. Inspired by ImageNet (Russakovsky et al., 2015), we curate 1,000 carefully selected intent categories, each paired with a set of text-to-image prompts and their corresponding outputs.

To capture variability in user expertise, prompts are carefully stratified into two groups: *novice* and *expert*. Novice prompts, sourced from `Lexica` (`https://lexica.art/`), reflect typical users who provide shorter, less informative descriptions. Expert prompts, collected from `Civitai` (`https://civitai.com/`), often specify detailed attributes, yielding richer, higher-information prompts. The two types of prompts are **manually annotation and re-verified** to ensure that: (1) novice and expert prompts strictly adhere to their intended styles, (2) in addition to class-level intents, sentence-level intents have also been added, and (3) the corresponding images are filtered to guarantee ethical compliance, safety, and the absence of sensitive content. Each intent category contains up to 10 instances, with each instance comprising (i) the user prompt, (ii) generated image URL, (iii) generation parameters, and (iv) auxiliary metadata.

We select 2,048 samples as our benchmark. Beyond category-level intent annotations, we further incorporate sentence-level intent annotations provided by human experts. Based on the difficulty of generation stability, the benchmark is stratified into three levels: (1) **Easy**, where the user intent is explicit and the category belongs to one of the 1,000 predefined categories; (2) **Medium**, where the user intent remains explicit but the category falls outside the predefined set; and (3) **Hard**, where the category is included in the predefined set, yet the user intent is ambiguous.

This benchmark enables systematic evaluation of how prompt informativeness—quantified via mutual information, prompt entropy, and prompt energy—affects generation stability. Mutual information measures the end-to-end alignment between user intent and generated images, reflecting whether the prompt provides sufficient information for semantically correct outputs. Prompt entropy, estimated using language model cross-entropy, captures the descriptive richness of the prompt. Prompt energy evaluates the model's internal "trust" in the input by measuring compatibility with its learned representation space. Together, these three metrics form a triangulated evaluation framework: mutual information provides an empirical upper bound of stability, prompt entropy assesses intrinsic informativeness, and prompt energy gauges the model's internal calibration. To avoid bias, Mutual information is evaluated via CLIP (Radford et al., 2021) and Qwen3-VL (Team, 2025), while prompt entropy and energy are measured with Llama-3,1-8B (Meta AI, 2024), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) and DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025) offering fine-grained insights into the impact of prompt characteristics on generation stability across user expertise levels.

4. NoxEye: An End-to-end Prompt Optimization Framework

We aim to improve the stability of text-to-image generation by aligning user-provided prompts with the preference

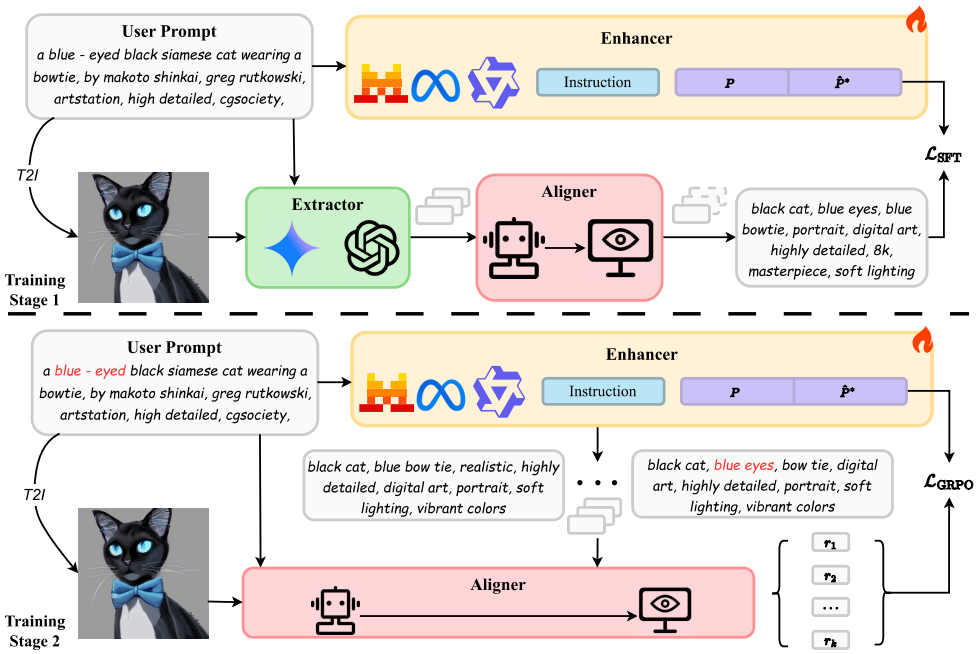


Figure 4. Overview of NoxEye training. It involves two stages, Supervised fine-tuning and GRPO-style on-policy preference optimization.

subspace of the target generative model. As shown in Section 3.1.2, the end-to-end stability is upper-bounded by the information encoded in the input prompt P. Therefore, to increase $I(Y;I)$, our rationale is to map user prompts closer to the **model's preferred prompt subspace**, a subset of the prompt space in which the generative model more reliably translates textual cues into visual concepts. Under a fixed generative model, prompts closer to this subspace effectively act as higher-quality carriers of intent-relevant information, resulting in improved generation stability.

Preference information extractor g_{ext}. To operationalize the notion of a model-preferred prompt subspace, we construct a structured prompt distribution that preserves intent-relevant information from the user prompt while conforming to the generative model's internal preference manifold. To this end, we introduce a **preference information extractor** that produces a high-fidelity textual proxy $\hat{P}^*$ conditioned on the user prompt P and the generated image I:

$$\hat{P}^* = g_{\text{ext}}(P, I).$$

In practice, g_{ext} is instantiated as a multimodal large language model (MLLM) that analyzes the visual output I and reconstructs a semantically precise textual description. These extracted descriptions approximate prompts lying closer to the model's preferred prompt subspace, and are used solely to estimate a proxy distribution of model-aligned prompts rather than as direct supervision for optimization.

Information enhancer π. The **information enhancer** learns a conditional transformation that re-expresses a user prompt P into a model-aligned representation lying closer to the model's preferred prompt subspace, while preserving the underlying user intent. Parameterized by θ, the enhancer defines a stochastic policy $\pi_\theta(\cdot \mid P)$ that generates a refined prompt: $P^* \sim \pi_\theta(\cdot \mid P)$.

Information aligner r_ϕ. To discourage hallucinations and preserve user intent in the refined prompt P^*, we introduce an **information aligner** that provides a preference-based reward signal $r_\phi(P^*, P)$ to guide prompt refinement. For a given user prompt P, we sample a group of candidate refined prompts $\{P_i\}_{i=1}^K \sim \pi(\cdot \mid P)$ and assign relative rewards $\{r_i\}$ using expert or model-based criteria that reflect generation stability and intent preservation.

Learning objective function. We first perform **supervised fine-tuning (SFT)** on a curated set of expert-refined prompt pairs $(P, \hat{P}^*)$ to provide a stable initialization for the prompt policy. Given these canonical refinements, SFT minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(P,\hat{P}^*)}\left[\log \pi_\theta(\hat{P}^* \mid P)\right].$$

To further optimize the prompts, mitigate model hallucinations, and prevent the introduction of extraneous information, we adopt **Group Relative Policy Optimization (GRPO)** as an intra-policy preference optimization stage with information aligner rewards $\{r_i\}$. GRPO optimizes the prompt policy by contrasting relative preferences within each group, encouraging transformations that consistently yield higher-quality and less hallucinations. Define group-relative advantages:

$$A_i = r_i - \frac{1}{K}\sum_{j=1}^{K} r_j.$$

Let π_{ref} be a frozen reference policy and

$$\rho_\theta(P_i \mid P) = \frac{\pi_\theta(P_i \mid P)}{\pi_{\text{ref}}(P_i \mid P)}.$$

The GRPO objective is

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_P \left[\frac{1}{K} \sum_{i=1}^{K} \min\left(\rho_i A_i, \tilde{\rho}_i A_i\right) \right],$$

where $\tilde{\rho}_i = \text{clip}\left(\rho_\theta(P_i \mid P), 1-\epsilon, 1+\epsilon\right)$.

5. Experiments

5.1. Settings

Data collection. We sample 5,000 prompt–image pairs from the *DiffusionDB* dataset and construct multiple candidate refined prompts using Gemini2.5 (Team et al., 2025) and GPT-4o(OpenAI et al., 2024). Each candidate is subsequently evaluated by human experts along three criteria: (i) semantic faithfulness to the original user prompt and alignment with the intended user intent, (ii) absence of hallucinated or unsupported content, and (iii) improvement in generation stability, measured by the consistency of images produced from the refined prompt.For supervised fine-tuning, the model is trained on 10,000 pairs of user prompts and refined prompts. In the GRPO phase, we adopt an on-policy preference optimization scheme, where human expert scores are used to construct preference signals over model-generated refined prompts, enabling GRPO-style updates guided by expert judgments.

Implementation Details. For the LLM backbone, we employ *Ministral3-8B* (Liu et al., 2026), *Llama-3.1-8B* (Meta AI, 2024) and *Qwen3-8B* (Team, 2025). Training details are shown in Appendix B.1. Evaluation is performed on multiple benchmark datasets by **NoxEye** (Figure 4) with *Llama-3.1-8B*, including *AP Bench* and *GenEval*(Huang et al., 2023), covering a wide range of prompt styles and complexities. Metrics include prompt and image stability, relevance, and diversity. All experiments are conducted on NVIDIA A100 40GB GPUs and the same seed 995 to ensure reproducibility and fair comparison.

Comparative Methods. We compare our approach against *Promptist* (Hao et al., 2023), *PAG* (Yun et al., 2025) and *PAE* (Mo et al., 2024). The text-to-image model use *SD 2.1* (Rombach et al., 2022), *CoMat 2.1* (Jiang et al., 2024), *Stable Diffusion 3 (SD3)* (Esser et al., 2024), *Flux.1* (Labs, 2024), *PixArt-Σ* (Saharia et al., 2022), *Infinity (Inf)* (Han et al., 2025) and *Show-O2* (Xie et al., 2025). For fairness, all models use their publicly released version, whereas *CoMat*, without a public version, is trained under the same experimental settings.

Table 2. Evaluation results about alignment scores ↑ on AP Bench. Bold values indicate **best**.

Method	SD2.1	CoMat2.1	SD3	Flux.1	PixArt-Σ	Inf	Show-o2
Promptist (Hao et al., 2023)	0.5813	0.5876	0.6412	0.6569	0.6620	0.6335	0.6062
PAG (Yun et al., 2025)	0.4174	0.4245	0.4580	0.4754	0.4817	0.4550	0.4369
PAE (Mo et al., 2024)	0.6246	–	–	–	–	–	–
NoxEye (Ours)	**0.6623**	**0.6690**	**0.7175**	**0.7352**	**0.7464**	**0.7152**	**0.6841**

Table 3. Evaluation results about energy (↓) and entropy (↓) on different models. Bold values indicate **best**.

Method	Energy ↓			Entropy ↓		
	Llama3	Mistral	DeepSeek	Llama3	Mistral	DeepSeek
Promptist (Hao et al., 2023)	-9.9156	-8.4298	-9.5479	5.0132	2.8596	3.8147
PAG (Yun et al., 2025)	-9.4217	-8.2512	-9.1560	5.3178	3.1739	3.9617
PAE (Mo et al., 2024)	-10.9813	-9.5547	-10.2767	4.6203	2.7240	3.6035
NoxEye (Ours)	**-11.9709**	**-9.8996**	**-10.6457**	**4.3549**	**2.6737**	**3.4519**

5.2. Results

State-of-the-art Comparison on AP Bench. As shown in Table 2 and Appendix B.2, our method consistently surpasses prompt-optimization baselines across all evaluation metrics for all generative models. For overall benchmark, *Ours+PixArt-Σ* improves alignment score from 0.4817 (*PAG+PixArt-Σ*) to 0.7464, while *Promptist+SD 2.1* achieves only 0.6620. The reductions in prompt entropy (from 3.9617 to 3.4519 by *PAG*) and prompt energy (from -9.1560 to -10.6457) by *DeepSeek-R1* indicate improved stability and higher fidelity of the generated images. For challenges of all difficulty levels, Our method achieves the best in mutual information, prompt energy and entropy.

Evaluating Advanced Generative Ability on GenEval. Beyond our proposed benchmark, we further evaluate our method on the widely adopted GenEval benchmark. As shown in Table 5 and Appendix B.3, our approach achieves consistent and competitive improvements across all prompt optimization methods, with NoxEye attaining the best overall scores on SD2.1, Flux.1, and PixArt-Σ. Especially, compared with prior prompt optimization methods, our approach attains state-of-the-art results on Single-object (+0.32%), Two-object (+5.23%), and Counting tasks (+3.67%). These results demonstrate that our method is not limited to our self-constructed benchmark, but also generalizes well to established, community-recognized evaluation protocols, validating its robustness and practical effectiveness.

Boosting Performances on Existing T2I Models. To assess the contributions of the *SFT* and *GRPO optimization* to generation quality and stability, we conduct a series of ablation experiments. as can be seen in Table 4, our model with SFT and GRPO optimization receive a better accuracy mean , accuracy standard deviation and alignment score in vast majority generative models for all levels.

Table 4. Impact of SFT and GRPO optimization. Bold values indicate **best**, and underlined values show second-best. NoxEye* refers to the model only training by SFT.

Method	SD2.1	CoMat2.1	SD3	Flux.1	PixArt-Σ	Inf	Show-o2
Llama3	0.6575	**0.6707**	0.7164	0.7308	0.7431	0.7115	0.6778
NoxEye*(Ours)	0.6613	0.6688	**0.7179**	0.7336	**0.7468**	**0.7159**	0.6828
NoxEye (Ours)	**0.6623**	0.6690	0.7175	**0.7352**	0.7464	0.7152	**0.6841**

Table 5. Evaluation results about overall scores ↑ on GenEval (Ghosh et al., 2023). Bold values indicate **best**.

Method	SD2.1	SD3	Flux.1	PixArt-Σ
Original prompt	0.48938	**0.71561**	0.64442	0.55127
Promptist (Hao et al., 2023)	0.48394	0.70670	0.63505	0.52257
PAG (Yun et al., 2025)	0.44156	0.65732	0.61077	0.52417
PAE (Mo et al., 2024)	0.38579	–	–	–
NoxEye (Ours)	**0.49240**	0.69266	**0.66890**	**0.56135**

Human Evaluation on Prompting Informativeness. We conducted a user study with 20 volunteers to compare our method with existing approaches from a human-centered perspective. Our approach was preferred most often, achieving scores of 0.44 (images) and 0.412 (prompts), compared to 0.417/0.143 and 0.366/0.222 for the baselines, respectively, demonstrating our method's superior alignment with human preferences in Figure 5.

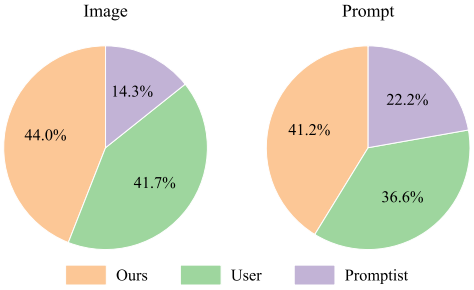


Figure 5. Human evaluation results. The result of NoxEye are preferred by human compared with the result of User Prompt and Promptist (Hao et al., 2023).

Inference Time Comparison. We further compare the inference efficiency of different prompt optimization methods. While Promptist/PAG and our method incur only marginal additional latency compared to the base SD 2.1 model, PromptCoT (Yao et al., 2024) significantly increases inference time due to its multi-step reasoning process. Notably, our approach achieves a favorable balance between efficiency and performance, adding only a small inference-time overhead. Quantitative results are summarized in Table 6.

5.3. Qualitative Analysis.

Figure 6 illustrates additional visual examples, where we can observe **how Noxeye enhances prompt generation stability**. Specifically, unlike other prompt optimization methods that largely retain the original prompt structure (e.g., Amateur...) and ignore the user intent (e.g., black cat),

Table 6. Inference time comparison across different methods.

Method	SD 2.1	+Promptist/PAG	+PromptCoT	+Ours
Inf. Time (s)	0.0327	0.0358	4.6114	0.0748

Noxeye re-constructs the prompt in a more explicit manner (e.g., black cat driving a 2023 Ford F-150), clarifying the core intent while systematically enhancing relevant visual specifications. Thus, our method effectively refines the main content and provides detailed descriptions of artistic style, lighting, and other visual attributes.

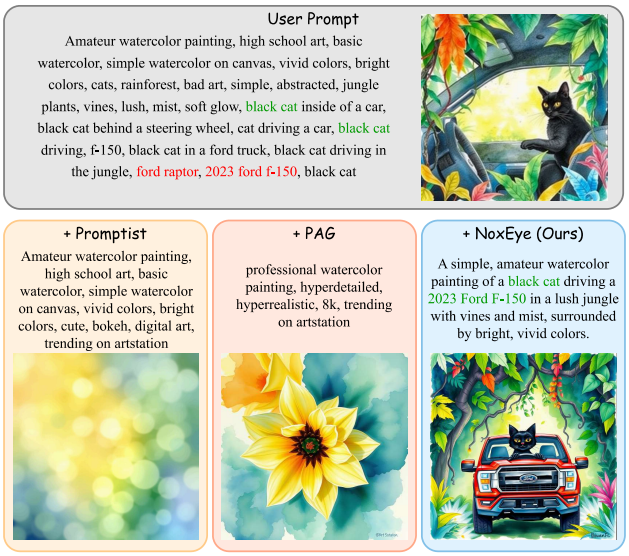


Figure 6. The generated images with the optimized prompts using our method. Each image generates by Flux.1 (Labs, 2024). More results are in the Appendix C.

6. Conclusion

In this work, we study the problem of user prompting proficiency and its quantification and taxonomy issues in the T2I generation. To this end, we construct **AP Bench**, a benchmark consisting of novice and expert prompts to comprehensively evaluating the informativeness of the user prompts. Furthermore, we design **NoxEye**, a plug-and-play modular prompt optimization framework broadly compatible with diverse T2I diffusion models. Extensive experiments across AP Bench and additional benchmarks demonstrate that our approach substantially improves both image quality and stability, without introducing significant inference overhead. Overall, our contributions provide not only a practical method for prompt optimization, but also a novel perspective on modeling information flow in text-to-image generation. We believe this work lays a foundation for future research on principled evaluation and optimization of user–model interactions in generative systems.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Cao, T., Wang, C., Liu, B., Wu, Z., Zhu, J., and Huang, J. Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. *arXiv preprint arXiv:2311.06752*, 2023.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023a.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023b.

Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., and Ruan, C. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. URL https://arxiv.org/abs/2501.17811.

Cheng, D., Huang, S., Zhu, X., Dai, B., Zhao, W. X., Zhang, Z., and Wei, F. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Du, C., Li, Y., Qiu, Z., and Xu, C. Stable diffusion is unstable. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:58648–58669, 2023.

Duan, J., Cheng, H., Wang, S., Zavalny, A., Wang, C., Xu, R., Kailkhura, B., and Xu, K. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. *arXiv preprint arXiv:2307.01379*, 2023.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., and Rombach, R. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

Ghosh, D., Hajishirzi, H., and Schmidt, L. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:52132–52152, 2023.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems (NeurIPS)*, 27, 2014.

Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., and Liu, X. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 15733–15744, 2025.

Hao, Y., Chi, Z., Dong, L., and Wei, F. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:66923–66939, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Huang, K., Sun, K., Xie, E., Li, Z., and Liu, X. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:78723–78747, 2023.

Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., and Liu, X. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation . *IEEE Transactions on Pattern Analysis Machine Intelligence (TPAMI)*, (01):1–17, January 2025. ISSN 1939-3539. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3531907.

Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Jiang, D., Song, G., Wu, X., Zhang, R., Shen, D., Zong, Z., Liu, Y., and Li, H. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:76177–76209, 2024.

Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Labs, B. F. Flux. https://github.com/black-forest-labs/flux, 2024.

Lee, D., Kim, C., Kim, S., Cho, M., and Han, W.-S. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11523–11532, 2022.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Liu, A. H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Liu, M., Ma, Y., Yang, Z., Dan, J., Yu, Y., Zhao, Z., Hu, Z., Liu, B., and Fan, C. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 5523–5531, 2025.

Ma, H., Pan, J., Liu, J., Chen, Y., Zhou, J. T., Wang, G., Hu, Q., Wu, H., Zhang, C., and Wang, H. Semantic energy: Detecting llm hallucination beyond entropy. *arXiv preprint arXiv:2508.14496*, 2025.

Meta AI. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/, July 2024. Accessed: 2024-07-23.

Mo, W., Zhang, T., Bai, Y., Su, B., Wen, J.-R., and Yang, Q. 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.

OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., et al. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International conference on machine learning (ICML)*, pp. 8821–8831. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Ren, Y. and Sutherland, D. J. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=tPNHOoZFl9.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Rosenman, S., Lal, V., and Howard, P. Neuroprompts: An adaptive framework to optimize prompts for text-to-image generation. *arXiv preprint arXiv:2311.12229*, 2023.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision (IJCV)*, 115(3):211–252, 2015.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems (NeurIPS)*, 35: 36479–36494, 2022.

Schlegel, K., Sommer, N. R., and Mortillaro, M. Large language models are proficient in solving and creating emotional intelligence tests. *Communications Psychology*, 3(1):80, 2025.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Stability AI. Stable diffusion public release. https://stability.ai/blog/stable-diffusion-public-release, 2023. Accessed: 2023-05-17.

Team, G. R., Abeyruwan, S., Ainslie, J., Alayrac, J.-B., et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Team, Q. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30, 2017.

Venkatesh, K., Dalva, Y., Lourentzou, I., and Yanardag, P. Ravel: Rare concept generation and editing via graph-driven relational guidance, 2025. URL https://arxiv.org/abs/2412.09614.

Wang, C., Franzese, G., Finamore, A., Gallo, M., and Michiardi, P. Information theoretic text-to-image alignment. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Wu, T.-H., Lian, L., Gonzalez, J. E., Li, B., and Darrell, T. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6327–6336, 2024.

Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

Xie, J., Yang, Z., and Shou, M. Z. Show-o2: Improved native unified multimodal models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=7VMg7Jb7AL.

Yang, H., Zhou, Y., Han, W., and Shen, J. Self-rewarding large vision-language models for optimizing prompts in text-to-image generation. *arXiv preprint arXiv:2505.16763*, 2025.

Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., and Cui, B. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *International Conference on Machine Learning (ICML)*, 2024.

Yao, J., Liu, Y., Dong, Z., Guo, M., Hu, H., Keutzer, K., Du, L., Zhou, D., and Zhang, S. Promptcot: Align prompt distribution via adapted chain-of-thought. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 7027–7037, 2024.

Yun, T., Zhang, D., Park, J., and Pan, L. Learning to sample effective and diverse prompts for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 23625–23635, 2025.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:55006–55021, 2023.

A. Additional Details on Stability Modeling

A.1. Building metrics from Mutual Information

From Mutual Information to CLIP Classification Accuracy. Let $Y \in \mathcal{Y}$ denote the discrete user intent category and $I \in \mathcal{I}$ the generated image. The mutual information between user intent and generated images is defined as

$$I(Y; I) = H(Y) - H(Y \mid I),$$

where $H(\cdot)$ denotes Shannon entropy. Since the prior distribution over intent categories is fixed across prompts and models, $H(Y)$ can be treated as a constant. Therefore, maximizing $I(Y; I)$ is equivalent to minimizing the conditional entropy $H(Y \mid I)$, which quantifies the remaining stability about the user intent after observing the generated image. However, the true posterior distribution $p(Y \mid I)$ is intractable. To obtain a computable surrogate, we introduce a classifier $f : \mathcal{I} \to \mathcal{Y}$, instantiated by CLIP (Radford et al., 2021), and define the predicted label $\hat{Y} = f(I)$. Let the classification error probability be

$$P_e = \Pr(\hat{Y} \neq Y).$$

By Fano's inequality, the conditional entropy is upper-bounded as

$$H(Y \mid I) \leq h(P_e) + P_e \log(|\mathcal{Y}| - 1), \tag{2}$$

where $h(\cdot)$ denotes the binary entropy function. Since the right-hand side is a monotonic function of P_e, reducing the classification error directly reduces an upper bound on $H(Y \mid I)$, thereby increasing a lower bound on the mutual information $I(Y; I)$.

In practice, we estimate P_e using the empirical classification accuracy of CLIP over generated images:

$$\mathrm{Acc} = \Pr(\hat{Y} = Y) = 1 - P_e.$$

Consequently, CLIP classification accuracy serves as a monotonic surrogate for mutual information between user intent and generated images. Higher accuracy implies lower conditional entropy $H(Y \mid I)$ and thus stronger intent-image alignment from an information-theoretic perspective.

Accuracy Variance and Mutual Information Stability. While the mean classification accuracy reflects the expected reduction of uncertainty in Y given I, it does not capture the variability of intent alignment across multiple samples generated from the same prompt. To characterize this aspect, we analyze the standard deviation of classification accuracy from an information-theoretic perspective.

Define a Bernoulli random variable

$$Z(I) = \mathbf{1}[\hat{Y}(I) = Y],$$

where $\hat{Y}(I)$ is the CLIP-predicted category. The empirical mean and variance of Z correspond to the classification accuracy and its standard deviation, respectively:

$$\mu = \mathbb{E}[Z], \quad \mathrm{Var}(Z) = \mu(1 - \mu).$$

From Equation 2, $H(Y \mid I)$ is upper-bounded by a monotonic function of the image-specific classification error probability $P_e(I)$. Since $Z(I) = 1 - P_e(I)$, the variance of Z directly reflects the variance of $P_e(I)$ and thus the variability of the entropy $H(Y \mid I)$ across generated images.

Therefore, the standard deviation of classification accuracy estimates the dispersion of entropy $H(Y \mid I)$. A low variance indicates that $H(Y \mid I)$ is concentrated, implying stable and consistent intent transmission, whereas a high variance suggests a multi-modal generation distribution in which different samples convey disparate or conflicting semantic interpretations of the user intent.

MLLM Alignment Score as a Variational Approximation of Mutual Information. While classification-based metrics rely on hard decoding of user intent categories, they provide only coarse-grained estimates of the conditional entropy $H(Y \mid I, P)$. To capture fine-grained semantic alignment between generated images and user intent descriptions, we employ a multi-modal large language model (MLLM) to produce a soft alignment score.

Let Y denote the corresponding intent description expressed in natural language. The entropy of interest is defined as

$$H(Y \mid I) = \mathbb{E}_{I,Y}\big[-\log p(Y \mid I)\big],$$

where $p(Y \mid I)$ is the true but intractable posterior distribution. We approximate this distribution using an MLLM, which induces a variational distribution $q_\theta(Y \mid I)$ over intent descriptions given an image.

The expected cross-entropy between the true distribution and its variational approximation admits the standard decomposition:

$$\mathbb{E}[-\log q_\theta(Y \mid I)] = H(Y \mid I) + \mathbb{E}\big[\mathrm{KL}(p \,\|\, q_\theta)\big],$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. Since the KL term is non-negative, the expected negative log-likelihood under q_θ provides an upper bound on the true entropy:

$$H(Y \mid I) \;\le\; \mathbb{E}[-\log q_\theta(Y \mid I)].$$

In practice, the MLLM produces a scalar alignment score $s(I, P) \in [0, 1]$, which can be interpreted as a calibrated estimate of $q_\theta(Y \mid I, P)$. Consequently, the expected negative log-alignment score,

$$\mathbb{E}[-\log s(I)],$$

serves as a computable surrogate for $H(Y \mid I, P)$. Since the conditional entropy appears as the only non-constant term in the prompt-conditioned mutual information,

$$I(Y; I \mid P = p) = H(Y \mid P = p) - H(Y \mid I, P = p),$$

maximizing the MLLM alignment score directly corresponds to increasing a variational lower bound on the mutual information between user intent and generated images.

Therefore, unlike classification accuracy which provides a discrete, upper-bounded estimate via Fano's inequality, the MLLM alignment score offers a continuous and semantically expressive approximation of mutual information, capturing fine-grained intent alignment beyond categorical correctness.

A.2. Prompt Entropy and Information-Theoretic Derivation

Consider the Markov chain $Y \to P \to I$, where Y is user intent, P is the prompt, and I is the generated image. By the data-processing inequality:

$$I(Y; I) \le I(Y; P),$$

indicating that the maximum achievable stability is constrained by the prompt information content.

Operationally, for a prompt $P = (w_1, \ldots, w_T)$, we approximate entropy using a pretrained LM:

$$H(P) \approx -\frac{1}{T} \sum_{t=1}^{T} \log p_\theta(w_t \mid w_1, \ldots, w_{t-1}).$$

Interpretation:

- *Low cross-entropy:* Predictable, concentrated prompt effectively conveys user intent, enhancing stability.
- *High cross-entropy:* Uncertain or dispersed prompt, less informative, reducing stability.

A.3. Derivation of Prompt Energy

In classical statistical mechanics, a system state $x_t^{(i)}$ follows a Boltzmann distribution:

$$p(x_t^{(i)}) = \frac{\exp(-E_t^{(i)}/k\tau)}{Z_t}.$$

An autoregressive LM with parameters θ defines the probability of token x_t as

$$p_\theta(x_t \mid x_{<t}) = \frac{\exp(z_t(x_t))}{\sum_{v\in\mathcal{V}} \exp(z_t(v))},$$

where $z_t(v)$ is the logit of token v.

Identifying logits with negative energies up to a normalization constant C_t:

$$z_t(v) = -\frac{1}{k\tau} E_t(v) + C_t.$$

Setting $k\tau = 1$ and $C_t = 0$ yields token-level energy

$$e_t := E_t(x_t) = -z_t(x_t),$$

and sequence-level prompt energy

$$E(x) = -\frac{1}{T}\sum_{t=1}^{T} z_t(x_t),$$

which measures the model's confidence in generating x. Lower $E(x)$ indicates familiar, well-represented concepts, whereas higher $E(x)$ indicates uncertain or out-of-distribution concepts.

Usage. Prompt energy complements entropy and end-to-end mutual information metrics, enabling a more complete characterization of text-to-image generation stability.

B. More Experiment Results

B.1. Experimental Setup Implementation Details

Training Hyperparameters Settings. We trained our model with the following hyperparameters: a learning rate of 1×10^{-5}, a batch size of 2, gradient accumulation steps of 16, and a total of 3 training epochs. The checkpoint with the lowest training loss was selected as the final model.

During LoRA fine-tuning, all parameters of the base model were frozen, and only the LoRA parameters were updated, specifically for query and value (Vaswani et al., 2017; Hu et al., 2022). The LoRA hyperparameters were set as follows: rank $r = 8$, $\alpha = 16$, a dropout rate of 0.1, and no bias. Training was performed using bf16 mixed precision.

We employed the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 1×10^{-2}.

NoxEye Prompt Template. To ensure consistency in model evaluation, we adopt the NoxEye Prompt Template, which specifies a unified structure system instruction, presenting tasks, inputs and outputs. The template is organized into four components:

- *System Instruction*: defines the global behavior and constraints that guide the model throughout the interaction.
- *Instruction*: defines the task description or objective to be performed.
- *Input*: provides the contextual information or query required to complete the task.
- *Response*: represents the expected model-generated output.

NoxEye Prompt Template

```
{system instruction}
### Instruction:
{instruction}

### Input:
{input}

### Response:
{output}
```

Figure 7. The prompt template of NoxEye.

Evaluation Settings. For evaluation on **AP Bench** and **GenEval**, the following hyperparameters were used:For SD2.1 and CoMat2.1, the number of sampling steps was set to 50, the CFG scale to 7.5, the image size to default size. For other models, the hyperparameters are default parameters.

Preference Information Extractor Setting. We present the prompt to transform the LVLM into the information extractor. The prompt for the preference information extractor in our model is illustrated in the Figure 8. As shown in Figure 9, the information of authentic prompts is significantly higher than that of prompts extracted by the preference information extractor, indicating that the latter produces prompts that are more concise and stable.

B.2. Full Results of AP Bench

Tables 7, 8 and 9 report the full evaluation results on the Easy, Medium, and Hard levels of AP Bench under both novice and expert user settings. Overall, **NoxEye consistently achieves the best or near-best performance across all difficulty levels and backbones**, particularly in terms of alignment score. On the Easy and Hard levels, NoxEye also attains the highest accuracy mean while maintaining competitive or lower accuracy variance, indicating improved robustness. On the Medium level, NoxEye shows clear advantages over prior methods for expert users and remains competitive for novice users. These results demonstrate that NoxEye generalizes well across user expertise, task difficulty, and diffusion backbones, outperforming existing prompt optimization approaches in both effectiveness and stability.

B.3. Full Results of GenEval

Table 12 reports the full evaluation results on the GenEval benchmark across multiple state-of-the-art text-to-image models. Overall, **NoxEye** demonstrates consistently strong and often superior performance across diverse model backbones and evaluation dimensions. On Stable Diffusion 2.1, NoxEye achieves the best results on *Single Object*, *Two Object*, *Counting*, and *Position*, leading to the highest overall score, which indicates improved compositional understanding and spatial grounding under complex prompts. For the more capable Stable Diffusion 3, while the original prompts yield the highest overall score, NoxEye remains highly competitive and attains near-best performance on fine-grained tasks such as *Counting* and *Two Object*, suggesting good generalization rather than over-specialization to a single model. Notably, on Flux.1 and PixArt-Σ, NoxEye consistently outperforms prior prompt optimization methods in terms of the overall metric, with clear gains on challenging attributes including multi-object composition, counting, and color reasoning. These results collectively indicate that NoxEye provides robust and model-agnostic improvements in text–image alignment, making it particularly effective for compositional and attribute-sensitive generation tasks, which aligns with the goals of controllable and reliable text-to-image synthesis.

Preference Distribution Extractor Prompt

You are a professional AI image analyst specializing in analyzing Stable Diffusion generated images. Please analyze this image and generate a prompt that could have been used to create this image.

Requirements:
1. The generated prompt should be concise, accurate, and suitable for CLIP model understanding
2. Use English with comma-separated keyword format
3. Include the following elements (if applicable):
   - Subject description (people, objects, scenes)
   - Art style (e.g., realistic, anime, oil painting, digital art, etc.)
   - Quality descriptors (e.g., highly detailed, 8k, masterpiece, etc.)
   - Composition description (e.g., portrait, full body, close-up, etc.)
   - Lighting effects (e.g., soft lighting, dramatic lighting, etc.)
   - Color characteristics (e.g., vibrant colors, monochrome, etc.)

4. Avoid overly complex descriptions, keep the prompt practical
5. Sort by importance, with the most important keywords first

Please output the prompt directly without additional explanations.

Figure 8. Preference information extractor prompt template.

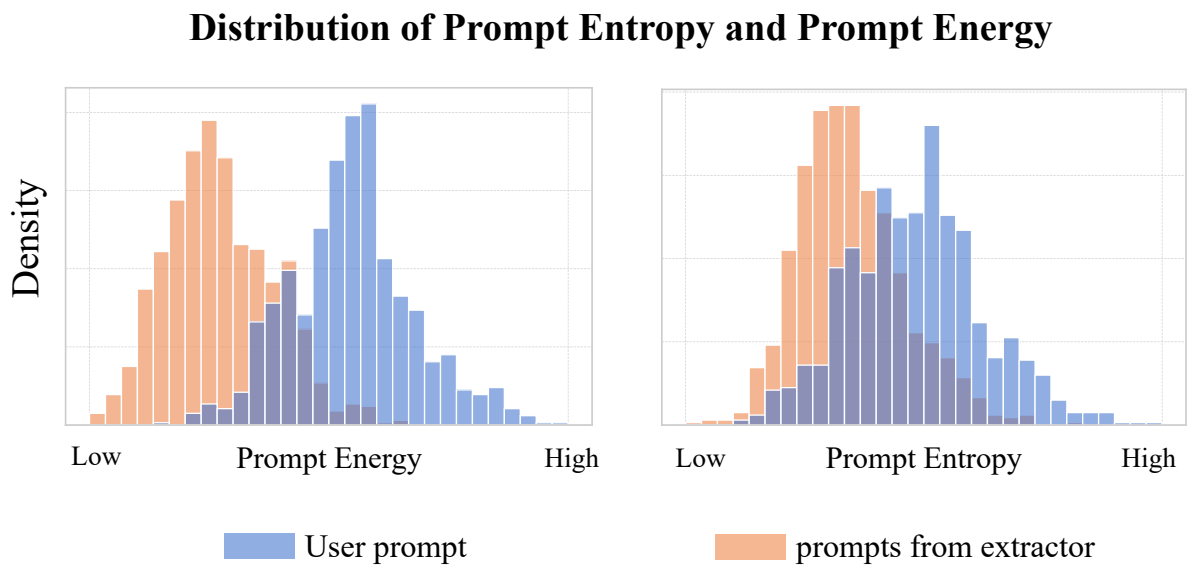


Figure 9. A visual comparison between authentic prompts and prompts from extractor.

Table 7. Full Evaluation results on Easy level AP Bench. Bold values indicate **best**.

User type		Novice			Expert		
Method	Model	Acc Mean ↑	Acc Std ↓	Align. ↑	Acc Mean ↑	Acc Std ↓	Align. ↑
Promptist (Hao et al., 2023)	SD2.1	0.3999	0.3742	0.6685	0.3450	0.3867	0.5237
	CoMat2.1	0.4188	0.3754	0.6740	0.3651	0.3922	0.5317
	SD3	0.4583	0.3719	0.7435	0.3847	0.3928	0.5759
	Flux.1	0.4296	0.3694	0.7537	0.3778	0.3878	0.6003
	PixArt-Σ	0.4402	0.3708	0.7628	0.3260	0.3740	0.5954
	Inf	0.4522	0.3758	0.7316	0.3483	0.3846	0.5655
	Show-O2	0.3995	0.3684	0.7079	0.3011	0.3669	0.5372
PAG (Yun et al., 2025)	SD2.1	0.3495	0.3774	0.5198	0.2754	0.3791	0.3462
	CoMat2.1	0.3625	0.3791	0.5305	0.2819	0.3801	0.3546
	SD3	0.3834	0.3834	0.5713	0.3120	0.3959	0.3875
	Flux.1	0.3629	0.3737	0.5970	0.2890	0.3793	0.3916
	PixArt-Σ	0.3594	0.3727	0.5996	0.2607	0.3653	0.4023
	Inf	0.3740	0.3795	0.5738	0.2710	0.3816	0.3722
	Show-O2	0.3352	0.3664	0.5507	0.2397	0.3565	0.3579
PAE (Mo et al., 2024)	SD2.1	0.3659	0.3648	0.6596	0.3211	0.3686	0.6167
NoxEye (Ours)	SD2.1	0.4268	0.3744	0.7047	0.3728	0.3832	0.6361
	CoMat2.1	0.4439	0.3732	0.7113	0.3852	0.3865	0.6429
	SD3	**0.4760**	0.3665	0.7670	**0.4215**	0.3851	0.6878
	Flux/1	0.4529	0.3670	0.7803	0.3958	0.3755	0.7018
	PixArt-Σ	0.4651	0.3723	**0.7890**	0.3583	0.3681	**0.7206**
	Inf	0.4627	0.3713	0.7659	0.3833	0.3794	0.6861
	Show-O2	0.4108	**0.3658**	0.7281	0.3169	**0.3540**	0.6562

Table 8. Full Evaluation results on Medium level AP Bench. Bold values indicate **best**.

User type		Novice	Expert
Method	Model	Align. ↑	Align. ↑
Promptist (Hao et al., 2023)	SD2.1	0.6416	0.5012
	CoMat2.1	0.6302	0.5139
	SD3	0.7035	0.5304
	Flux.1	0.7071	0.5387
	PixArt-Σ	**0.7222**	0.5541
	Inf	0.7013	0.5221
	Show-O2	0.6282	0.5053
PAG (Yun et al., 2025)	SD2.1	0.5222	0.3248
	CoMat2.1	0.5069	0.3293
	SD3	0.5074	0.3448
	Flux.1	0.5567	0.3611
	PixArt-Σ	0.5521	0.3833
	Inf	0.5586	0.3337
	Show-O2	0.4772	0.3503
PAE (Mo et al., 2024)	SD2.1	0.5885	0.6165
NoxEye (Ours)	SD2.1	0.6524	0.6370
	CoMat2.1	0.6443	0.6569
	SD3	0.6992	0.7023
	Flux.1	0.6958	**0.7231**
	PixArt-Σ	0.7160	0.7113
	Inf	0.6948	0.6617
	Show-O2	0.6407	0.6536

Table 9. Full Evaluation results on Hard level AP Bench. Bold values indicate **best**.

User type		Novice			Expert		
Method	Model	Acc Mean ↑	Acc Std ↓	Align. ↑	Acc Mean ↑	Acc Std ↓	Align. ↑
Promptist (Hao et al., 2023)	SD2.1	0.1356	0.2848	0.6336	0.1424	0.2953	0.3786
	CoMat2.1	0.1209	0.2667	0.6516	0.1209	0.2667	0.3826
	SD3	0.1921	0.3242	0.6845	0.1774	0.3184	0.4179
	Flux.1	0.1882	0.3263	0.7185	0.1389	0.2847	0.4437
	PixArt	0.1772	0.3279	0.6745	0.1191	0.2663	0.4504
	Inf	0.1690	0.3083	0.6523	0.0991	0.2499	0.4392
	Show-O2	0.1647	0.3141	0.6336	0.1206	0.2642	0.4009
PAG (Yun et al., 2025)	SD2.1	0.1194	**0.2603**	0.3778	0.0918	0.2403	0.1900
	CoMat2.1	0.1213	0.2670	0.3660	0.0978	0.2460	0.1886
	SD3	0.1774	0.3125	0.3951	0.1137	0.2699	0.2174
	Flux.1	0.1925	0.3299	0.4854	0.1021	0.2517	0.2140
	PixArt-Σ	0.1520	0.2993	0.4437	0.0927	0.2449	0.2211
	Inf	0.1645	0.3031	0.3993	0.0804	0.2337	0.2113
	Show-O2	0.1499	0.2973	0.3850	0.0955	0.2478	0.1942
PAE (Mo et al., 2024)	SD2.1	0.1232	0.2660	0.6576	0.1228	0.2729	0.4978
NoxEye (Ours)	SD2.1	0.1498	0.3031	0.6840	0.1107	0.2560	0.5590
	CoMat2.1	0.1216	0.2663	0.6833	0.1203	0.2667	0.5599
	SD3	**0.2011**	0.3273	0.7220	**0.1668**	0.3143	0.5868
	Flux.1	0.1902	0.3241	**0.7718**	0.1410	0.2874	0.6351
	PixArt-Σ	0.1675	0.3153	0.6981	0.0952	**0.2286**	**0.6662**
	Inf	0.1837	0.3320	0.6958	0.1140	0.2656	0.6157
	Show-O2	0.1369	0.2912	0.6417	0.1037	0.2399	0.6011

Table 10. Evaluation results about CLIP Classification Accuracy (Mean ↑) on AP Bench. Bold values indicate **best**, and underlined values show second-best.

Method	SD2.1	CoMa2.1t	SD3	Flux.1	PixArt-Σ	Inf	Show-O2
Promptist (Hao et al., 2023)	0.3480	0.3655	0.3977	0.3751	0.3619	0.3731	0.3317
PAG (Yun et al., 2025)	0.2920	0.3019	0.3257	0.3060	0.2933	0.3033	0.2737
PAE (Mo et al., 2024)	0.3193	–	–	–	–	–	–
NoxEye (Mistral)	0.3184	0.3321	0.3868	0.3569	0.3496	0.3668	0.3213
NoxEye (Llama)	**0.3681**	**0.3822**	0.4180	0.3940	**0.3829**	0.3921	0.3405
NoxEye (Qwen)	0.3321	0.3438	0.3803	0.3592	0.3503	0.3640	0.3068

Table 11. Evaluation results about CLIP Classification Accuracy (standard deviation ↓) on AP Bench. Bold values indicate **best**, and underlined values show second-best.

Method	SD2.1	CoMa2.1t	SD3	Flux.1	PixArt-Σ	Inf	Show-O2
Promptist (Hao et al., 2023)	0.3783	0.3818	0.3834	0.3777	0.3758	0.3827	0.3681
PAG (Yun et al., 2025)	0.3726	0.3749	0.3849	0.3721	**0.3673**	0.3774	0.3593
PAE (Mo et al., 2024)	**0.3639**	–	–	–	–	–	–
NoxEye (Mistral)	0.3690	**0.3718**	**0.3765**	**0.3655**	0.3687	**0.3764**	0.3601
NoxEye (Llama)	0.3784	0.3803	0.3797	0.3744	0.3757	0.3795	0.3626
NoxEye (Qwen)	0.3744	0.3780	0.3817	0.3743	0.3730	0.3807	**0.3560**

Table 12. Full Evaluation results on GenEval (Ghosh et al., 2023). Bold values indicate **best**.

Method	**single Obj. ↑**	**Two Obj. ↑**	**Counting ↑**	**Colors ↑**	**Position ↑**	**Color Attri ↑**	**Overall**
Stable Diffuison 2.1 (Rombach et al., 2022)							
Original prompts	96.56%	50.76%	40.00%	**84.31%**	7.00%	15.00%	0.48938
Promptist (Hao et al., 2023)	96.25%	47.98%	41.25%	81.38%	7.25%	**16.25%**	0.48394
PAG (Yun et al., 2025)	94.69%	45.45%	34.06%	74.73%	4.50%	11.50%	0.44156
PAE (Mo et al., 2024)	85.62%	39.14%	23.12%	68.09%	6.50%	9.00%	0.38579
NoxEye (Ours)	**97.50%**	**51.52%**	**43.44%**	78.99%	**11.00%**	13.00%	**0.4924**
Stable Diffusion 3 (Esser et al., 2024)							
Original prompts	**99.38%**	**86.87%**	**63.12%**	**87.50%**	31.00%	**61.50%**	**0.71561**
Promptist (Hao et al., 2023)	99.06%	86.87%	59.69%	85.90%	**33.25%**	59.25%	0.7067
PAG (Yun et al., 2025)	98.75%	83.59%	52.50%	84.31%	22.00%	53.25%	0.65732
NoxEye (Ours)	**99.38%**	86.62%	61.88%	82.98%	29.25%	55.50%	0.69266
Flux.1 [schnell] (Labs, 2024)							
Original prompts	**100.00%**	86.87%	55.31%	74.47%	24.50%	45.50%	0.64442
Promptist (Hao et al., 2023)	96.25%	81.06%	50.31%	77.66%	**29.75%**	**46.00%**	0.63505
PAG (Yun et al., 2025)	97.50%	80.05%	46.56%	**80.85%**	22.75%	38.75%	0.61077
NoxEye (Ours)	99.06%	**91.41%**	**58.13%**	78.99%	29.25%	44.50%	**0.6689**
PixArt-Σ (Saharia et al., 2022)							
Original prompts	**99.38%**	64.14%	43.75%	83.24%	12.50%	**27.75%**	0.55127
Promptist (Hao et al., 2023)	95.00%	58.08%	**45.94%**	79.52%	10.00%	25.00%	0.52257
PAG (Yun et al., 2025)	97.81%	63.38%	40.94%	81.12%	10.25%	21.00%	0.52417
NoxEye (Ours)	97.50%	**67.42%**	42.50%	**85.64%**	**16.25%**	27.50%	**0.56135**

B.4. Human Evaluation Details

To complement quantitative evaluations with a human-centered perspective, we conducted a user study comparing our method with existing approaches. We first sampled a set of prompts at random and applied different optimization methods to obtain model-specific refined prompts, which were then used to generate images. A total of 20 volunteers were recruited

from diverse educational backgrounds. In each trial, participants were presented with either a pair of images or a pair of prompts and were asked to select the image they found more visually appealing or the prompt they preferred. As shown in Figure 5, participants most frequently selected images generated from prompts optimized by our method, indicating its superior effectiveness in aligning with human preference.

C. More Quantitative Results

We present more visual results between the images generated with different prompts. As illustrated in the Figure 10, when a prompt simultaneously mentions both *fish* and *cat*, the generated images may omit the fish, indicating **informational sparsity**, where overly concise user prompts encourage LLM-based expansion to hallucinate unintended attributes and deviate from user intent. In addition, anomalous fish–cat amalgamations are observed, revealing **lexical perturbation and noise sensitivity**, in which minor word-level disturbances can mislead the text encoder and result in inconsistent or even contradictory generations.

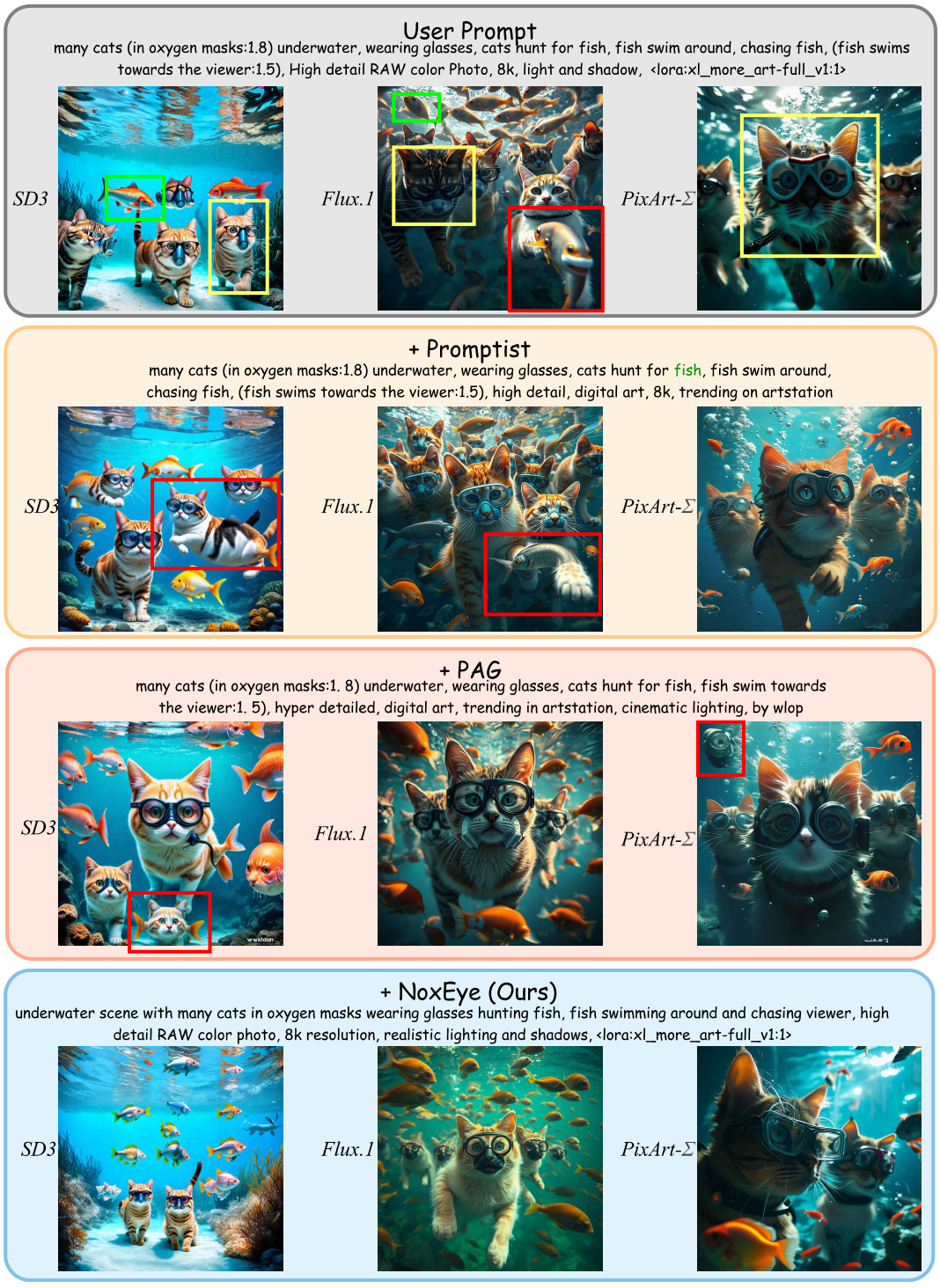


Figure 10. Informational sparsity stems from user prompts, while lexical perturbation and noise sensitivity arise in other prompt optimization approaches.