Automatic Prompt Engineering for Scalable Prompt Inversion in Text-to-Image Ad Generation

Zixin Ding[1*], **Qi Zeng**[2], **Boying Gong**[2], **Wenlong Deng**[3,⋆], **Bo Pan**[4,⋆], **Yuxin Chen**[1†]

[1]University of Chicago, [2]Meta, [3]University of British Columbia, [4]Emory University

{zixin,chenyuxin}@uchicago.edu

Abstract

While prompt engineering offers effective control over Text-to-Image (T2I) generation, it remains labor-intensive for large-scale production. We present PRISM-DUEL, a black-box framework that formalizes prompt optimization as Automatic Prompt Engineering (APE), motivated by advertising workflows requiring low-latency, diverse variants faithful to a human-designed ads. Since zero-shot LLMs are unreliable judges of image quality, PRISM-DUEL obtains label-free pairwise preferences and rationales from an LLM judge over pairs of generated images, then uses a dueling-bandit optimizer to optimize a prompt for generating controlled *variations* while matching the reference ad's visual content. By iteratively steering the prompt distribution towards higher-quality generations and improving posterior calibration, PRISM-DUEL preserves visual similarity and semantic faithfulness while increasing diversity. Experiments on PartiPrompts and DreamBooth across Gemini 2.5 Flash Image, FLUX.1, and Qwen-Image show consistent gains over strong baselines in visual faithfulness and prompt interpretability.

1 Introduction

Advertising platforms produce high-quality, brand-safe image variants under tight latency and computational cost constraints (Jiang et al., 2025a). While modern T2I models (Betker et al., 2023; Chen et al., 2025) simplify generation, production performance hinges on the reference images used to condition prompts. Zero-shot VLM captions often miss key product semantics, miscount elements (Liu et al., 2024c), ignore spatial relations (Wang et al., 2026), resulting in off-brand creatives (Jansen et al., 2023).

We formulate *scalable prompt inversion* (Mahajan et al., 2024) for ad creatives: given a pool of reference ads $\{r\}$, and a fixed T2I generator $\mathbf{G}$, the goal is to recover a structured prompt p that faithfully reproduce r with high fidelity. Our method therefore optimizes p to (i) preserve advertiser intent and constraints, (ii) align with reference image semantics, and (iii) yield visually faithful, aesthetically coherent variations.

Prior T2I prompt optimization typically relies on pointwise (single-image) feedback from VLMs (Mañas et al., 2024; He et al., 2025) or TIFA/VQA-style scores (Mrini et al., 2024), which are primarily weakly discriminative and do not directly measure reference preservation. Another line trains prompt rewriters from prompt–image data via supervised finetuning (Datta et al., 2024), reinforcement learning (Jiang et al., 2025b; Kong et al., 2024), or hybrids (Li et al., 2024; Yang et al., 2025; Wu et al., 2025b), but it requires campaign-specific retraining. In practice, ad catalogs change rapidly (new SKUs, layouts, seasonal styles), making per-campaign retraining slow and expensive (Mehrotra, 2025).

In this paper, we first empirically demonstrate that VLM-based judges give saturated and non-discriminative pointwise scores, even with two state-of-the-art VLM models (GPT-4o-mini and GPT-5-nano). This observation motivates a shift towards **pairwise** feedback: given a fixed $\mathbf{G}$, we leverage a VLM judge $\mathcal{J}$ to compare two images produced under competing prompts. To ensure compatibility with black-box T2I models and the rapid turnover of industrial ad catalogs, we adopt a training-free automatic prompt engineering (APE) setup and cast it as a **dueling-bandit** problem driven by reference-conditioned comparisons, extending label-free prompt optimization beyond text-only settings (Wu et al., 2025c; Xiang et al., 2025). Unlike LLM prompt optimization for classification or QA, where outputs admit a notion of correctness (often via "reasoning and final answer"), T2I optimization must handle stochastic multi-attribute

[*]Work done at Meta.

[†]Corresponding author

image quality, requiring judging over composition and visual faithfulness to a reference. Finally, to improve reference preservation, we incorporate a *low-fidelity* text-embedding prior to calibrate posterior estimates, effectively steering prompt candidates towards the reference caption derived by the VLM itself. We summarize our contributions as follows:

1. **Quantifying Judge Saturation:** We show that pairwise image comparisons yield more distinctive and reliable preferences than pointwise scoring, and that these preferences better correlate with perceptual similarity measures such as DINOv3 (Siméoni et al., 2025) and CLIP-I (Valerio et al., 2023).

2. **Copeland-UCB for non-transitive preferences:** Motivated by non-transitive VLM/LLM preferences (Liu et al.; Xu et al.; Liusie et al., 2024), we frame T2I prompt selection as finding a *Copeland winner* and introduce a Copeland-UCB dueling-bandit optimizer (Zoghi et al., 2015), regularized by a low-fidelity text-embedding prior that steers exploration toward the reference caption. Compared to Double Thompson search with quadratic pairwise scaling (Wu et al., 2025c), our UCB-style solver yields tighter convergence guarantees with near-linear dependence on the prompt set size.

3. **Extensive evaluation across models:** We validate PRISM-DUEL on PartiPrompts (Writing & Symbols) (Yu et al.) and DreamBooth (Ruiz et al., 2023) across three T2I generators and VLM judges, showing consistent improvements in text rendering, prompt interpretability, and single-/multi-reference faithfulness for both open and black-box models.

2 Related Work

Automatic Prompt Engineering. Our work builds on Automatic Prompt Engineering (APE), which typically optimizes prompts with *supervised* signals from labeled data or validation performance. While these methods excel in classification and reasoning tasks (Yuksekgonul et al., 2024; Ding et al., 2025; Zhou et al., 2022; Pryzant et al., 2023) where ground truth is available, industry-scale T2I prompt inversion lacks such supervision: evaluating image quality and alignment (e.g., DINOv3, CLIP) requires GPU-heavy encoding and large-scale generation (Oquab et al., 2023; Siméoni et al., 2025; Radford et al., 2021). This motivates *label-free* prompt optimization based on pairwise preferences from an LLM/VLM judge (Xiang et al., 2025; Wu et al., 2025c). Unlike SPO's (Xiang et al., 2025) greedy hill-climbing and PDO's (Wu et al., 2025c) dueling-bandit setup, PRISM-DUEL addresses the inherent *non-transitivity* of image preferences (e.g., competing trade-offs between identity preservation and stylistic alignment) where a Condorcet winner may not exist. By targeting the *Copeland* criterion via a Copeland-UCB solver, PRISM-DUEL focuses comparisons on near-tied cyclic regions, ensuring more stable prompt selection under limited industrial query budgets.

Bandit-Based Prompt Optimization. Our work contributes to a growing body of work on combining bandit algorithm and prompt selection. ProTeGi (Pryzant et al., 2023) couples beam search with a bandit-style best arm identification step while assuming labeled data and scalar task metrics; in contrast, large-scale T2I prompt optimization is generally *label-free* and judgments shall be *non-transitive*. Similarly, TRIPLE (Shi et al., 2024) studies budgeted best arm identification over a *static* prompt pool, whereas our industry setting involves an *evolving* prompt population where new candidates are continually generated and compared against existing ones, often without a pre-enumerated pool (Wang et al., 2023). OPTS (Ashizawa et al., 2025) applies Thompson Sampling (Agrawal and Goyal, 2012) to select among prompt-design strategies, but relies on scalar rewards computed from development-set scoring with labels. However, the notion of absolute correctness on images remains in ambiguity and preference-based feedback is non-transitive.

3 Recognizing Judge Patterns for T2I Models

We start from the widely used *pointwise* VLM for T2I prompt engineering given reference, instantiated by PRISM (He et al., 2024): given a generated image x and a reference image r, the judge assigns an absolute similarity score (e.g., a 0–10 Likert scale rating (Likert, 1932)) and the LLM as the optimizer seeks to maximize the scalar feedback.

Setup. We fix the VLM as $\mathcal{J}$ and score images generated by several T2I models (Qwen-Image (Wu et al., 2025a), Gemini 2.5 Flash Image (Comanici et al., 2025), and FLUX.1 (Black

Forest Labs, 2024)) on personalized image generation using DreamBooth (Ruiz et al., 2023) reference images. For each reference r, we run T=5 refinement iterations with a candidate prompt pool of size K_c=10 per iteration, yielding K_cT=50 generated images per $(r, \mathbf{G})$. We evaluate four VLM judges: two closed-source models, GPT-4o mini and GPT-5 nano, and two open-source models, LLaVA-NeXT-7B (Liu et al., 2024a) and InternVL3.5-8B (Chen et al., 2024).[1] We employ the SOTA DINOv3 ViT-S/16 backbone (`facebook/dinov3-vits16-pretrain-lvd1689m`) as a frozen feature extractor to calculate the object-sensitive image similarity between generated and reference images, following He et al. (2024); Ruiz et al. (2023). For each generated image per fixed r, we compute average pairwise cosine similarity between DINOv3 embeddings of r and generated images within the pool of candidates for PRISM. We additionally compute CLIP-I, the cosine similarity between CLIP image embeddings of the generated and reference images, and aggregate it across the K_cT samples per r and then the dataset.

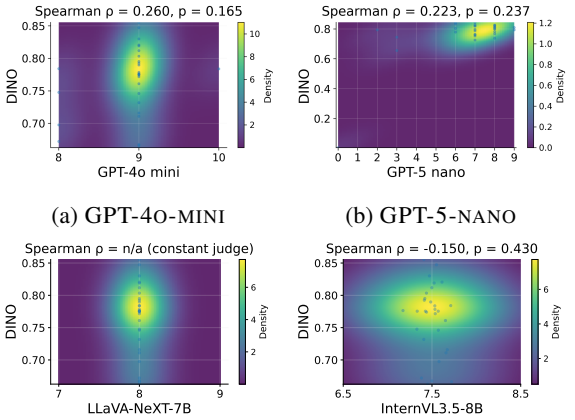
Results. Pointwise VLM judges give saturated, non-discriminative scores. Across three $\mathbf{G}$, single-image VLM scores exhibit strong ceiling effects on the 0–10 Likert scale (Likert, 1932) for both proprietary and open-source judges. Figure 1 shows that on FLUX.1 (Black Forest Labs, 2024), scores cluster near 9 for GPT-4o-mini, 7–8 for GPT-5-nano, and 7–8 for InternVL3.5-8B (Chen et al., 2024), while LLaVA-NeXT-7B (Liu et al., 2024b) degenerates to a constant judge, assigning 8 to nearly every image. None of the four judges correlate significantly with DINOv3 similarity (Spearman ρ=$0.26, 0.22, -0.15$ for GPT-4o-mini, GPT-5-nano, and InternVL3.5-8B respectively, all p>0.15; undefined for LLaVA-NeXT-7B due to zero variance). This compression—and in LLaVA-NeXT-7B's case, outright collapse—makes pointwise scores effectively low-resolution for ranking candidate prompts (Zheng et al., 2023). Figure 9 further shows per-iteration means with largely overlapping error bars and near-identical distributions across generators and judge families, explaining the flat optimization trajectories (Appendix B). Saturation persists across closed- and open-weight VLMs and is most severe for the smallest judge, suggesting the problem is intrinsic to pointwise Likert scoring rather than a model-specific artifact and motivating our pairwise evaluation given reference.

Pairwise evaluation yields a more distinctive measure of preference alignment. We extend the original single-image PRISM framework (He et al., 2025) from pointwise scoring to a pairwise setting (Liu et al.). Using the meta-prompt (P1 and P2) in Appendix A, we run a dueling variant of PRISM for 5 refinement iterations with 10 *duels* per round. All other settings are held *fixed*; the only change is replacing single-image scores with reference-conditioned pairwise comparisons. Figure 2 reports agreement between VLM pairwise preferences and DINOv3-based pairwise winners for FLUX.1 across two proprietary (GPT-4o-mini, GPT-5-nano) and two open-source (LLaVA-NeXT-7B, InternVL3.5-8B) judges. GPT-4o-mini shows the clearest separation, with 37/50 duels on the diagonal, strongly preferring images with higher DINOv3 similarity. Notably, InternVL3.5-8B achieves comparable alignment (35/50 on the diagonal), suggesting that capable open-source VLMs can rival proprietary judges in reference-based pairwise agreement. In contrast, GPT-5-nano is close to chance-level agreement (27/50) with counts distributed fairly evenly across cells, and LLaVA-NeXT-7B collapses to chance (25/50) while exhibiting a pronounced positional bias toward option A (selecting A in 34/50 duels regardless of DINOv3-preferred side). Overall, pairwise feedback improves discriminability relative to pointwise scoring but can still remain misaligned with reference-similarity metrics, motivating the need for stronger, comparison-driven optimization signals during APE. We report the full pairwise-preference results using CLIP-I and DINOv3 scores in Figures 7 and 8 (Appendix B).

4 Pairwise Preference Feedback: Setup and Limitations

With the comparison on pairwise and pointwise judgment in hand, we shift to pairwise comparisons that ask a VLM to choose which of two candidates better matches a reference image. This naturally induces a *dueling* feedback model: for any two candidate prompts $p_i, p_j \in \{p_1, \ldots, p_K\}$ and generated images, a comparison returns a binary outcome indicating whether p_i is prefered to p_j with fixed $\mathbf{G}$ and $\mathcal{J}$. Recent works introduce SPO (Xiang et al., 2025) and PDO (Wu et al., 2025c) use pair-

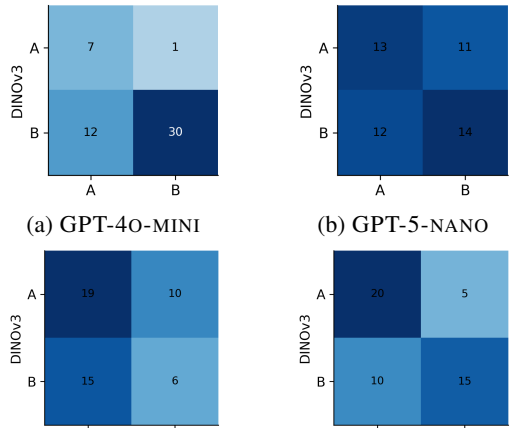
[1]We omit GPT-4V from the original paper setup because this model has been deprecated by OpenAI.



(a) GPT-4O-MINI (b) GPT-5-NANO

(c) LLAVA-NEXT-7B (d) INTERNVL-3.5B

Figure 1: 2D kernel density estimate (KDE) heatmap of single-image VLM scores vs. DINOv3 similarity with $\mathbf{G}$ as FLUX1 for (a) GPT-4o-mini (b) GPT-5-nano (c) LLaVA-NeXT-7B (d) InternVL3.5-8B; brighter regions indicate higher sample density.



(a) GPT-4O-MINI (b) GPT-5-NANO

(c) LLAVA-NEXT-7B (d) INTERNVL-3.5B

Figure 2: Confusion matrices comparing $\mathcal{J}$ pairwise preferences with DINOv3-based pairwise winners for FLUX.1: (a) GPT-4o-mini, (b) GPT-5-nano (c) LLaVA-NeXT-7B (d) InternVL3.5-8B.

wise comparisons primarily as a local accept/reject signal to update the current best prompt p_t.

Limitations of prior pairwise optimization. A closer look, however, reveals the dueling comparison for T2I models is not actually so straightforward. (i) VLM-based preferences are often *noisy* and can be *non-transitive* (Xu et al.): there may exist prompts p_i, p_j, p_k such that $\mathcal{J}(\mathbf{G}(p_i), \mathbf{G}(p_j)) = p_i$ and $\mathcal{J}(\mathbf{G}(p_j), \mathbf{G}(p_k)) = p_j$, yet $\mathcal{J}(\mathbf{G}(p_k), \mathbf{G}(p_i)) = p_k$, forming a preference cycle. Therefore local accept/reject updates may become brittle under cyclic preferences. (ii) simply concentrating comparisons on p_t and proposed new prompts matchups under-utilizes the comparison budget: many comparisons are spent on uninformative pairs, and there is no explicit mechanism to allocate queries to the most uncertain or decision-critical pairs. (iii) both SPO and PDO are inherently computationally costly and budget-limited: exhaustive pairwise comparisons scales quadratically with number of candidates prompts (Zheng et al., 2023). SPO sets a maximum rounds, and PDO stops after that rounds and returns the current leader, without a confidence-based criterion certifying the winner is statistically separated from close competitors.

5 Methodology

These observations motivate our design: (i) a Copeland-style objective that remains meaningful under non-transitive preferences, (ii) an uncertainty-aware (UCB) comparison strategy that allocates duels to the most informative pairs under a fixed budget, and (iii) a structured prior over prompt space to propose higher-quality prompts candidates with fewer computationally costly duels. We present PRISM-DUEL in Algorithm 1, building on PRISM (He et al., 2025) for T2I models but in *pairwise* fashion and specifically designed for non-transitive, budget-limited VLM judging. We defer the detailed complexity analysis of each submodule in Algorithm 1 to Appendix C.

Pairwise Preference Feedback. Given fixed $\mathbf{G}$, $\mathcal{J}$ and a candidate set $\{p_i\}_{i=1}^K$. Each duel (p_i, p_j) yields $y_{ij} \in \{p_i, p_j\}$ with $p_{ij} = \Pr(p_i \succ p_j)$ and $p_{ij} + p_{ji} = 1$. Specifically, a *Condorcet Winner* is an arm p_{i^*} such that $p_{i^*j} > 0.5$ for all $j \neq i^*$. Since VLM preferences may be non-transitive, we target a *Copeland Winner* i^* maximizing $\sum_{j \neq i^*} \mathbb{1}\{p_{ij} > \frac{1}{2}\}$. Given reference image r, we sample images $x_i \sim \mathbf{G}(p_i)$ and query $y_{ij} = \mathcal{J}(r, x_i, x_j)$. We maintain pairwise statistics for each ordered pair (p_i, p_j): n_{ij} is the number of duels between (p_i, p_j) and w_{ij} is the number of times p_i wins over p_j. We maintain pairwise counts (w_{ij}, n_{ij}). The empirical win rate is $\hat{p}_{ij} \triangleq \frac{w_{ij}}{\max(1, n_{ij})}$.

Scalable Copeland Selection with Embedding Priors. A naive Round-Robin tournament requires estimating the pairwise preference for all pairs, scaling quadratically as $O(K^2)$ (Wu et al., 2025c). In our industrial setting, where every comparison involves an expensive T2I generation and VLM evaluation, this complexity is prohibitive. To

Algorithm 1 PRISM-DUEL

Require: Reference r; rounds T; candidates/round K_c; Duels/round B; Prompt Engineer $\mathcal{F}$; text embedder Embed; prior strength λ; confidence δ; seed CRN; $\mathbf{G}$; $\mathcal{J}$.

Ensure: Optimized Prompt $p^\star$ and final image $\hat{x}^\star$.

$c_0 \leftarrow$ VLM-Caption(r); $e_0 \leftarrow \mathrm{Embed}(c_0)$
Initialize pool $\mathcal{P} \leftarrow \{c_0\}$; $w_{ij}, N_{ij} \leftarrow 0$; APE feedback $\mathsf{fb} \leftarrow \emptyset$
for $t = 1$ to T **do**
  $\mathcal{C} \leftarrow \mathcal{F}(r, p, \mathsf{fb}, K_c)$ ▷ K_c candidates
  $\mathcal{S}_k \leftarrow \mathcal{C} \cup \{p\}$; $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{C}$
  WarmStart: For each $p_i \in S_k$, calculate $s_i \leftarrow$ cosine_sim(Embed$(p_i), e_0$)
    $\forall p_j \in \mathcal{S}_k : \tilde{w}_{ij} = \lambda \cdot \sigma(s_i - s_j), \tilde{n}_{ij} = \lambda$
  $w_{ij}, n_{ij} \leftarrow 0$
  $N_{round} \leftarrow 0$
  while $N_{round} < B$ and $|S_k| > 1$ **do**
    UpdateConfidenceBounds(S_k, δ)
    $S_k \leftarrow$ PruneSuboptimal(S_k)
    $(p_i, p_j) \leftarrow$ SelectUncertainPair(S_k)
    if $(p_i, p_j) = \emptyset$ **then break**
    $y \leftarrow \mathcal{J}(\mathbf{G}(p_i), \mathbf{G}(p_j))$
    $w_{ij} \leftarrow w_{ij} + y$; $n_{ij} \leftarrow n_{ij} + 1$
    $N_{\text{round}} \leftarrow N_{\text{round}} + 1$
  $p \leftarrow \mathrm{argmax}_{p_i \in \mathcal{S}_k}$ CopelandScore(p_i) ▷ Eq. 3 in Appendix C
  $\mathsf{fb} \leftarrow$ SUMMARIZEFEEDBACK$(\mathcal{S}_k, p, w, n)$ ▷ Algorithm 3 in Appendix C.
$p^\star \leftarrow p$; $\hat{x}^\star \leftarrow \mathbf{G}(p^\star)$
return $p^\star, \hat{x}^\star$

resolve this, a necessity in high-fidelity industrial ad generation (Shi et al., 2024; Jiang et al., 2025a), we adopt *Scalable Copeland Bandits* (SCB) (Zoghi et al., 2015). Instead of exploring all pairs, we maintain confidence intervals $[L_{ij}, U_{ij}]$ for the pairwise win probabilities. We calculate optimistic $CS_{\max}$ and pessimistic $CS_{\min}$ Copeland scores for every candidates and aggressively prune any p_j where $CS_{\max}(j) < \max_i CS_{\min}(i)$. This transforms the process into an adaptive elimination tournament where "easy" comparisons against weak prompts are resolved with minimal samples ($O(1)$), allowing the algorithm to concentrate most of the budget B solely on distinguishing the top-tier candidates. In reality, this reduces sample complexity to $\tilde{O}(K \log K)$, significantly lowering the total GPU hours for inversion (Mokady et al., 2023).

Embedding-Guided Warm Start. To further accelerate convergence in "cold-start" scenarios, we inject domain knowledge via a text-embedding prior. We calculate the cosine similarity s_i between the candidate p_i and the reference caption c_0 using OpenAI embeddings. We then initialize the bandit with pseudo-counts:

$$\tilde{w}_{ij} = \lambda \cdot \sigma(s_i - s_j), \quad \tilde{n}_{ij} = \lambda \tag{1}$$

where λ is the prior strength and σ is the Sigmoid function. Simultaneously, we reset the observed counts w_{ij} and n_{ij} to zero. This configuration allows the embedding prior to dominate early exploration, steering the bandit toward semantically similar regions and avoiding computationally costly VLM duels on irrelevant prompt mutations.

Industrial Deployment: Early Termination. Our per-round budget B ensures predictable latency and computational costs, a prerequisite for production-grade ad-tech. Additionally, our SCB implementation naturally provides an early termination criterion: the loop breaks immediately when the set of "uncertain pairs" where $0.5 \in [L_{ij}, U_{ij}]$ becomes empty. In deployment, this allows the pipeline to bypass up to 40% of the budgeted VLM calls for clear-cut candidate sets.

6 Experiments

Implementation Details. We consider three $\mathbf{G}$: Qwen-Image (Wu et al., 2025a), Gemini-2.5 Flash Image (Comanici et al., 2025), and FLUX.1 (Black Forest Labs, 2024). For $\mathcal{J}$ we use GPT-5 nano and GPT-4o mini. We run PRISM-DUEL with a total query budget of $B = 100$ and report results using $K_c = 10$ candidates per round over $T = 10$ rounds. We evaluate our models on two benchmarks: PartiPrompts (specifically the Writing & Symbols category) (Yu et al.), totalling up to 91 prompts, and Dreambooth (Ruiz et al., 2023) (30 subjects, with 4-6 reference images per subject). Evaluation focuses on two critical industrial axes: (1) *Text Rendering Fidelity*, using the "Writing & Symbols" subset of PartiPrompts (Yu et al.). This benchmark stresses exact symbol reproduction, aligning with fidelity requirements for brand logos and slogans. Here, we mimic an "asset variation" workflow by performing self-inversion on images generated by the models themselves (Cui et al., 2025; Zhang et al., 2023). (2) *Subject Consistency*, using DreamBooth (Ruiz et al., 2023) to measure fidelity under diverse contexts. Following

VLM Judge ($\mathcal{J}$)	Method	PartiPrompts			DreamBooth		
		CLIP-I ↑	DINOv3 ↑	CLIP-T ↑	CLIP-I ↑	DINOv3 ↑	CLIP-T ↑
GPT-4o-mini	GPT-4o Caption (zero-shot)	0.801	0.534	0.313	0.750	0.512	0.302
	PRISM (He et al., 2024)	0.812	<u>0.569</u>	0.321	0.768	<u>0.532</u>	0.325
	SPO-T2I (Xiang et al., 2025)	<u>0.823</u>	0.562	0.323	0.762	0.529	0.331
	PDO-T2I (Wu et al., 2025c)	0.821	0.566	<u>0.331</u>	<u>0.772</u>	0.521	<u>0.341</u>
	PRISM-DUEL (ours)	**0.829**	**0.578**	**0.346**	**0.798**	**0.548**	**0.357**
GPT-5-nano	GPT-4o Caption (zero-shot)	0.792	0.542	0.308	0.744	0.503	0.305
	PRISM (He et al., 2025)	0.789	0.601	0.312	<u>0.742</u>	0.503	0.341
	SPO-T2I (Xiang et al., 2025)	<u>0.809</u>	<u>0.615</u>	<u>0.325</u>	0.703	0.509	0.339
	PDO-T2I (Wu et al., 2025c)	0.806	0.611	0.322	0.712	<u>0.513</u>	<u>0.346</u>
	PRISM-DUEL (ours)	**0.824**	**0.621**	**0.342**	**0.759**	**0.526**	**0.364**

Table 1: **Quantitative Evaluation for Image quality across baselines with G as FLUX.1.**

previous protocols by He et al. (2025), we generate targets for 30 subjects across 25 templates for all three generators, yielding a comprehensive test suite of 2,250 images.

Baselines. We compare against four classes of *training-free* baselines. **(i) Zero-shot captioning:** GPT-4o, a strong zero-shot image captioner (OpenAI, 2024; Cheng et al., 2025a,b). **(ii) Pointwise prompt optimization:** PRISM (He et al., 2025), which optimizes prompts using pointwise VLM similarity scores; for fair comparison we use the same total budget as our method (10 iterations, 10 candidates per iteration). **(iii) Dueling-bandit prompt optimization:** PDO (Wu et al., 2025c), adapted to reference-conditioned T2I. Each arm is a prompt/template instantiation; at each step PDO selects a prompt pair via Double Thompson Sampling (D-TS), generates two images with the same T2I generator $\mathbf{G}$ (matched random seeds), and queries a $\mathcal{J}$ to choose the better image conditioned on the reference creative. PDO updates pairwise preference posteriors and proposes new candidates via top-performer-guided template mutation (e.g., style, composition, negative constraints) while keeping reference semantics fixed; the final prompt is selected by its posterior Copeland score under the same image-generation budget. **(iv) Self-supervised iterative refinement:** SPO (Xiang et al., 2025) adapted to reference-conditioned T2I. SPO uses pairwise comparisons of generated outputs to guide prompt revisions; in our adaptation we execute each prompt with $\mathbf{G}$, use a $\mathcal{J}$ to perform reference-conditioned pairwise comparisons, and revise prompts via constrained template edits. We match the total image-generation budget and use matched random seeds per duel to reduce stochasticity.

Evaluation Metrics. Following PRISM (He et al., 2024), we evaluate reference faithfulness using CLIP image similarity (CLIP-I) (Valerio et al., 2023) and DINOv3 similarity (Siméoni et al., 2025) between each generated image and the reference image. Following DreamBooth (Ruiz et al., 2023), we also report CLIP text-image similarity (CLIP-T) between generated prompt and initial caption to measure prompt fidelity, computed as the cosine similarity between CLIP embeddings of the prompt text and the generated image. All metrics are averaged over each dataset.

Results. Table 1 benchmarks PRISM-DUEL against recent baselines (including PRISM, PDO-T2I, SPO-T2I, and zero-shot GPT-4o) using the FLUX.1 generator. We refer readers to Appendix B for parallel results on Gemini (Table 3) and Qwen-Image (Table 4) and Appendix E for qualitative results. PRISM-DUEL demonstrates notable improvements in both text-image alignment (CLIP-T) and image fidelity (CLIP-I, DINOv3). On the PartiPrompts dataset, when evaluated by GPT-4o-mini, PRISM-DUEL achieves a CLIP-T score of 0.346, outperforming the strongest baseline (PDO-T2I at 0.331). Furthermore, it yields the highest image fidelity scores (0.829 for CLIP-I and 0.578 for DINOv3). This trend continues on the DreamBooth dataset, where PRISM-DUEL effectively preserves subject identity while following complex prompts, achieving a peak CLIP-T of 0.357 compared to SPO-T2I's 0.341. The consistent margins over competitive baselines like PDO-T2I and SPO-T2I highlight the efficacy of our dual-pronged approach in generating high-quality, closely aligned images.

Method	API & Compute Complexity			Early Stopping
	$\mathbf{G}$	$\mathcal{J}$	**Embed** (Embed)	
GPT-4o Caption	1	1	0	None
PRISM	$\mathcal{O}(T \cdot K_c)$	$\mathcal{O}(T \cdot K_c)$	0	None (Pointwise)
SPO	$\mathcal{O}(T \cdot m)$	$\mathcal{O}(T \cdot m)$	0	None (Greedy)
PDO	$\mathcal{O}(K^2)$	$\mathcal{O}(K^2)$	0	Soft (Probabilistic)
PRISM-DUEL	$\mathcal{O}(T \cdot K_c)$	$\mathcal{O}(T \cdot \bar{B})$	$\mathcal{O}(T \cdot K_c)$	Hard (Confidence)

Table 2: **Computational Complexity per Prompt.** $\bar{B}$ is the effective pairwise evaluation budget where $\bar{B} \leq B$. Our Copeland Bandit formulation with pruning bounds the $\mathcal{O}(K^2)$ VLM computational cost typically associated with pairwise comparisons.

7 Computational Cost Analysis

For real-world deployment, we seek to minimize computational overhead. As shown in Table 2, while both PRISM-DUEL and PDO (Wu et al., 2025c) require $\mathcal{O}(T \cdot K_c)$ image generations, their evaluation computational costs diverge. PDO relies on exhaustive pairwise comparisons, forcing a quadratic VLM API complexity of $\mathcal{O}(T \cdot K_c^2)$. As the candidate pool K_c expands, this introduces prohibitive computational costs and latency bottlenecks. By utilizing a near-zero computational cost text-embedding prior (**WarmStart**), PRISM-DUEL actively prunes redundant VLM comparisons whose upper confidence bounds fall below the leader's lower bound. This restricts VLM calls to an effective budget $\bar{B} \ll K_c^2$, dropping evaluation complexity to $\mathcal{O}(T \cdot \bar{B})$.

8 Conclusion

We present PRISM-DUEL, a framework for reducing high latency and quadratic computational cost of traditional dueling-bandit prompt optimization, addressing large-scale image inversion for industrial deployment. By reformulating the selection process via SCB, we reduce the evaluation complexity from $O(K^2)$ to $O(K \log K)$, significantly lowering the number of expensive VLM judge queries and T2I generation calls per round. A text-embedding prior guides early exploration, and certified early stopping ends duels once a winner is statistically clear, yielding predictable runtimes. Overall, PRISM-DUEL achieves high-fidelity inversion while remaining robust to the non-transitive preferences common in ad-creative evaluation.

9 Acknowledgement

Z. Ding and Y. Chen acknowledge support from the National Science Foundation under Grant Nos. IIS-2313131 and CMMI-2037026.

References

Shipra Agrawal and Navin Goyal. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings.

Rin Ashizawa, Yoichi Hirose, Nozomu Yoshinari, Kento Uchida, and Shinichi Shirakawa. 2025. Bandit-based prompt design strategy selection improves prompt optimizers. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20799–20817, Vienna, Austria. Association for Computational Linguistics.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Black Forest Labs. 2024. Flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev. Model card on Hugging Face. Accessed: 2025-11-10.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. 2025a. CapArena: Benchmarking and analyzing detailed image captioning in the LLM era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14077–14094, Vienna, Austria. Association for Computational Linguistics.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. 2025b. Caparena: Benchmarking and analyzing detailed image captioning in the llm era (project website). https://caparena.github.io/. Accessed: 2026-01-26.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Weila Cui, Martin J Liu, and Ruizhi Yuan. 2025. Exploring the integration of generative ai in advertising agencies: A co-creative process model for human–ai collaboration. *Journal of Advertising Research*, pages 1–23.

Siddhartha Datta, Alexander Ku, Deepak Ramachandran, and Peter Anderson. 2024. Prompt expansion for adaptive text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3449–3476, Bangkok, Thailand. Association for Computational Linguistics.

Zixin Ding, Junyuan Hong, Jiachen T Wang, Zinan Lin, Zhangyang Wang, and Yuxin Chen. 2025. Scaling textual gradients via sampling-based momentum. *arXiv preprint arXiv:2506.00400*.

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

Yutong He, Alexander Robey, Naoki Murata, Yiding Jiang, Joshua Williams, George J Pappas, Hamed Hassani, Yuki Mitsufuji, Ruslan Salakhutdinov, and J Zico Kolter. 2024. Automated black-box prompt engineering for personalized text-to-image generation. *arXiv preprint arXiv:2403.19103*, 2(5).

Yutong He, Alexander Robey, Naoki Murata, Yiding Jiang, Joshua Nathaniel Williams, George J. Pappas, Hamed Hassani, Yuki Mitsufuji, Ruslan Salakhutdinov, and J Zico Kolter. 2025. Automated black-box prompt engineering for personalized text-to-image generation. *Transactions on Machine Learning Research*.

Tijmen Jansen, Mark Heitmann, Martin Reisenbichler, and David A Schweidel. 2023. Automated alignment: Guiding visual generative ai for brand building and customer engagement. *Available at SSRN*.

Daniel R Jiang, Alex Nikulkov, Yu-Chia Chen, Yang Bai, and Zheqing Zhu. 2025a. Improving generative ad text on facebook using reinforcement learning. *arXiv preprint arXiv:2507.21983*.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. 2025b. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*.

Weize Kong, Spurthi Hombaiah, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Prewrite: Prompt rewriting with reinforcement learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–601.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM Web Conference 2024*, pages 3367–3378.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llavanext: Improved reasoning, ocr, and world knowledge.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*.

Yufang Liu, Tao Ji, Changzhi Sun, Yuanbin Wu, and Aimin Zhou. 2024c. Investigating and mitigating object hallucinations in pretrained vision-language (clip) models. *arXiv preprint arXiv:2410.03176*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th conference of the European chapter of the Association for Computational Linguistics (volume 1: long papers)*, pages 139–151.

Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. 2024. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6808–6817.

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

Udit Mehrotra. 2025. Leveraging generative ai in e-commerce for catalog enrichment.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047.

Khalil Mrini, Hanlin Lu, Linjie Yang, Weilin Huang, and Heng Wang. 2024. Fast prompt alignment for text-to-image generation. *arXiv preprint arXiv:2412.08639*.

OpenAI. 2024. Gpt-4o system card. Accessed 2026-01-26.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510.

Chengshuai Shi, Kun Yang, Zihan Chen, Jundong Li, Jing Yang, and Cong Shen. 2024. Efficient prompt optimization through the lens of best arm identification. *Advances in Neural Information Processing Systems*, 37:99646–99685.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, and 1 others. 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.

Rodrigo Valerio, Joao Bordalo, Michal Yarom, Yonatan Bitton, Idan Szpektor, and Joao Magalhaes. 2023. Transferring visual attributes from natural language to verified image generation. *arXiv preprint arXiv:2305.15026*.

Yunlong Wang, Shuyuan Shen, and Brian Y Lim. 2023. Reprompt: Automatic prompt editing to refine ai-generative art towards precise expressions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–29.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. 2026. Everything in its place: Benchmarking spatial intelligence of text-to-image models. In *The Fourteenth International Conference on Learning Representations*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025a. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Mingrui Wu, Lu Wang, Pu Zhao, Fangkai Yang, Jianjin Zhang, Jianfeng Liu, Yuefeng Zhan, Weihao Han,

Hao Sun, Jiayi Ji, and 1 others. 2025b. Reprompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning. *arXiv preprint arXiv:2505.17540*.

Yuanchen Wu, Saurabh Verma, Justin Lee, Fangzhou Xiong, Poppy Zhang, Amel Awadelkarim, Xu Chen, Yubai Yuan, and Shawndra Hill. 2025c. Llm prompt duel optimizer: Efficient label-free prompt optimization. *arXiv preprint arXiv:2510.13907*.

Jinyu Xiang, Jiayi Zhang, Zhaoyang Yu, Xinbing Liang, Fengwei Teng, Jinhao Tu, Fashen Ren, Xiangru Tang, Sirui Hong, Chenglin Wu, and Yuyu Luo. 2025. Self-supervised prompt optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9017–9041, Suzhou, China. Association for Computational Linguistics.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. 2025. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3024–3034. IEEE.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. Investigating non-transitivity in llm-as-a-judge. In *Forty-second International Conference on Machine Learning*.

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

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, and 1 others. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*.

Weichen Yu, Ziyan Yang, Shanchuan Lin, Qi Zhao, Jianyi Wang, Liangke Gui, Matt Fredrikson, and Lu Jiang. 2024. Is your text-to-image model robust to caption noise? *arXiv preprint arXiv:2412.19531*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. 2023. Real-world image variation by aligning diffusion inversion chain. *Advances in Neural Information Processing Systems*, 36:30641–30661.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*.

Masrour Zoghi, Zohar S Karnin, Shimon Whiteson, and Maarten De Rijke. 2015. Copeland dueling bandits. *Advances in neural information processing systems*, 28.

A Meta Prompt

Below is the system prompt template for $\mathcal{J}$ for pairwise comparison for personalized T2I generation.

P1 Pairwise comparison meta-prompt

Pairwise Image Evaluation

Instruction

You are an image judge. You will see a GOAL image, then two candidates A and B. Choose which candidate better matches the GOAL. Be decisive. Respond *STRICT JSON* with keys: `{"winner": 0 or 1, "conf": number in [0,1], "scoreA": number, "scoreB": number}`. Use a 0..10 scale for `scoreA`/`scoreB` (higher = closer). Avoid giving identical scores unless they are truly indistinguishable; if so, still pick a winner.

Input placeholder

`{Goal Image} {Image Candidate A} {Image Candidate B}`

For reference, we also provide the original system prompt template for single image evaluation (He et al., 2024).

P2 Single-image evaluation meta-prompt

Single Image Evaluation

Instruction

You are a helpful prompt engineer assistant. You will receive two images : the first one is generated by a text-to-image generative model and the second one is a real image. Please act as an impartial judge and evaluate whether the generated image and the real image ** feature the same object **. Your evaluation should only consider the main object featured in the images and ignore all irrelavant factors such as the background, lighting, environment, camera angles, the pose of the object and style, etc . Be as objective as possible. Rate the response on a scale from 0 to 10. A rating of 0 signifies two images with completedly different and unrelated objects featured in them. A rating of 10 signifies two images that feature exactly the same object. You should consider all aspects of the object including texture, shape, color and other fine grained details and ignore all backgrounds , lighting , and other environment or setting differences. Pay attention to the details and be as critical as possible. Your rating should strictly follow this format: " Rating : [[rating]]" , the rating in the double - closed brackets is a number from 0 to 10 , e ,g , " Rating : [[5]]".

Input placeholder

`{Image Candidate} {Goal Image}`

B Additional Experiments

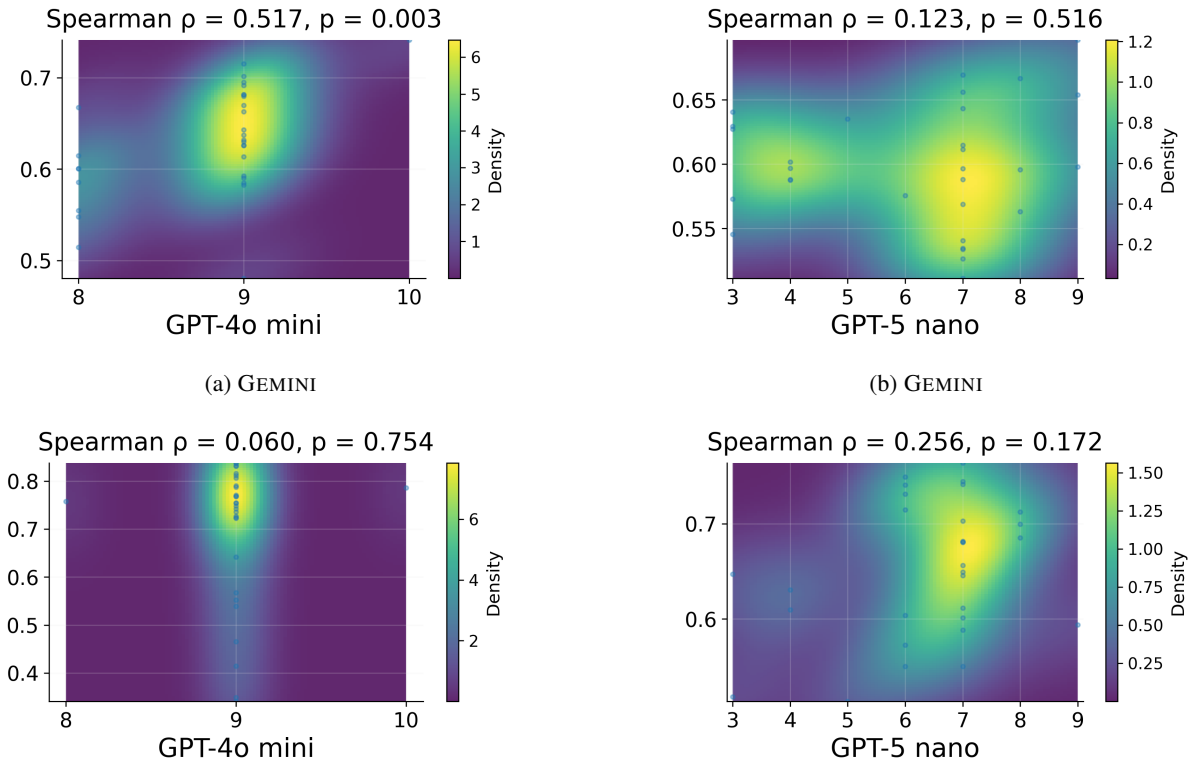
We provide additional analyses of pointwise VLM judging. Figure 5 reports correlations between VLM scores and DINOv3, and Figure 6 reports correlations with CLIP-I. We further visualize VLM judge behavior over optimization by plotting per-iteration mean scores for PRISM and the full score distributions: Figure 9 shows histograms (or KDEs), and Figure 10 shows violin plots.

As an example, we also present additional CLIP-I score trends on the PartiPrompts dataset. Figure 4 reports the average CLIP-I trajectory over optimization iterations on PartiPrompts across three T2I generators (FLUX.1, Gemini, Qwen-Image) and two VLM judges (GPT-4o-mini, GPT-5-nano). Across all six settings, PRISM-DUEL achieves the strongest or near-strongest performance, typically improving earlier and converging to the best final-iteration CLIP-I compared to PRISM, SPO-T2I, and PDO-T2I. Notably, PRISM-DUEL shows robust gains under both judges (i.e., improvements persist when switching from GPT-4o-mini to the stricter GPT-5-nano) and remains competitive across $\mathbf{G}$, indicating that its advantage is not tied to a single model or judge. Overall, the curves suggest PRISM-DUEL is the most reliable optimizer among all baselines, showcasing higher end performance with less sensitivity to $\mathbf{G}/\mathcal{J}$ choice.

C Additional PRISM-DUEL Descriptions

Algorithm 2 shows the implementation details of the sub-routines used in Algorithm 1. These components are designed to minimize VLM calls and provide semantic guidance for prompt mutation.

Theoretically Complexity Bound. PRISM-DUEL improves the standard Copeland Bandits as PDO (Wu et al., 2025c) $O(K^2 \log T)$ by pruning candidates that mathematically cannot become the Copeland winner. The theoretical sample complexity (i.e., how many API calls/duels are needed to find the best prompt with high probability) to iden-



(a) GEMINI (b) GEMINI

(c) QWEN (d) QWEN

Figure 3: KDE of single-image VLM scores versus DINOv3 similarity for **G** (a) Gemini 2.5 Flash Image and (b) Qwen-Image with $\mathcal{J}$ as GPT-4o-mini and GPT-5-nano.

tify the Copeland winner with probability $1-\delta$ is $O(\frac{K \log K}{\delta^2} \log(\frac{1}{\delta}))$ where K is number of candidates (size of pool S_k) and δ as the gap of distinctiveness between the best prompt and the second best.

The PRISM-DUEL algorithm adapts the *Scalable Copeland Bandits* (SCB) framework (Zoghi et al., 2015) to the problem of discrete prompt optimization. Unlike standard dueling bandits that minimize cumulative regret, our objective is *Best Arm Identification* (BAI) under a fixed query budget B.

Complexity Reduction. A naive Round-Robin approach requires estimating the pairwise preference $P(p_i \succ p_j)$ for all $K(K-1)/2$ pairs, leading to a sample complexity of $O(K^2)$. PRISM-DUEL reduces this to $\tilde{O}(K \log K)$ by maintaining confidence bounds $[L_{ij}, U_{ij}]$ for the preference probabilities.

PruneSuboptimal At each time step t, we maintain an active set S_k. A candidate p_i is removed from S_k if it is dominated by the current leader. Specifically, let $CS_{min}(i)$ be the count of pairs where p_i is statistically significant to win ($L_{ij} > 0.5$), and $CS_{max}(i)$ be the count where p_i is not yet statistically proven to lose ($U_{ij} > 0.5$). We discard any prompt p_j such that:

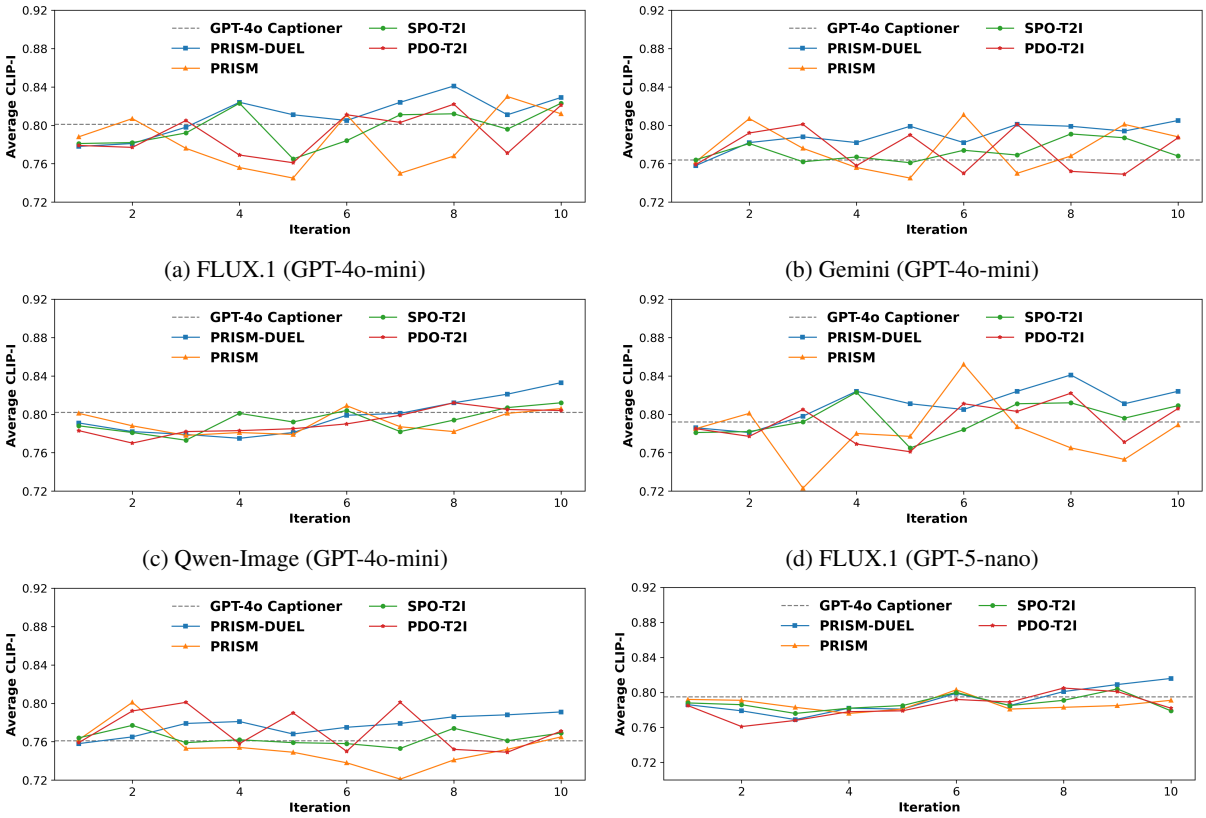
$$CS_{max}(j) < \max_{p_i \in S_k} CS_{min}(i) \tag{2}$$

This ensures that we stop allocating budget to prompt mutations that mathematically cannot become the Copeland winner, effectively focusing the remaining budget B on resolving the "hard" comparisons between top-tier candidates.

Copeland Score. Here, we provide the definition of CopelandScore (CS) in Algorithm 1 as the number of other candidates in the current active set S_k that p_i is estimated to beat (win probability > 0.5).

We define

$$\text{CS}(p_i) = \sum_{p_j \in S_k, j \neq i} \mathbb{1}\left(\frac{w_{ij} + \tilde{w}_{ij}}{n_{ij} + \tilde{n}_{ij}} > 0.5\right) \tag{3}$$



(a) FLUX.1 (GPT-4o-mini)

(b) Gemini (GPT-4o-mini)

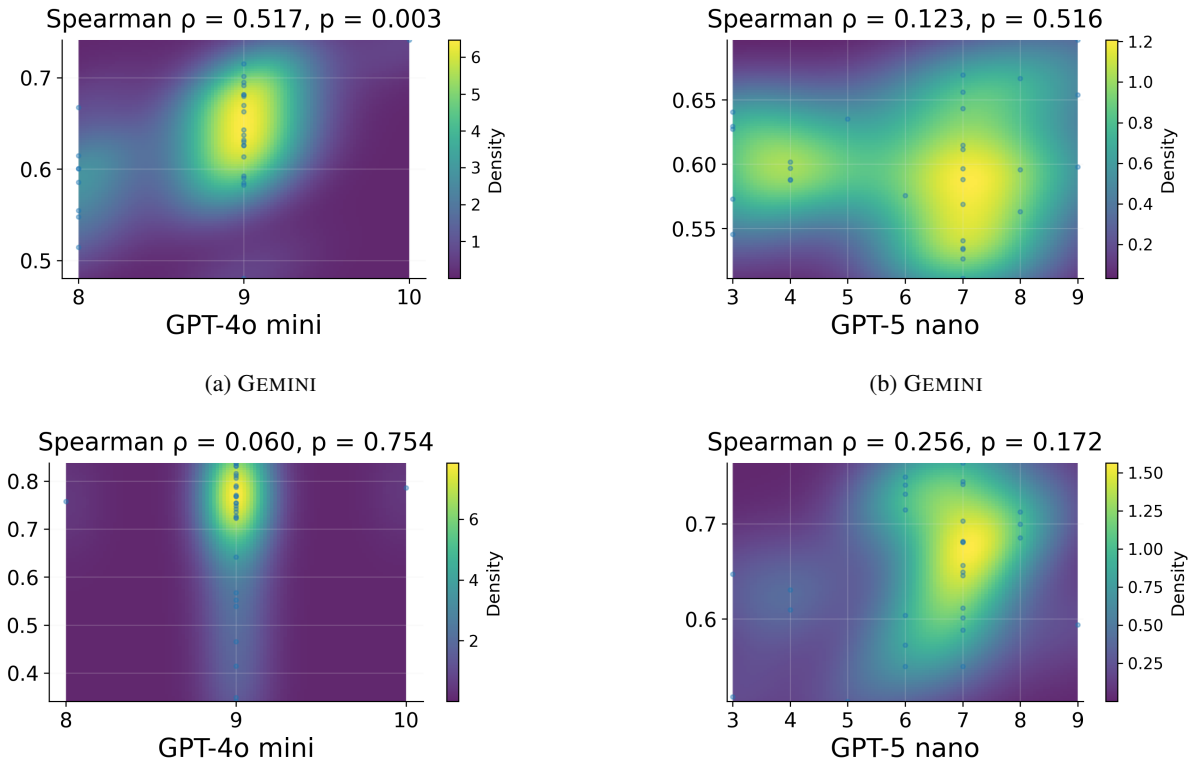
(c) Qwen-Image (GPT-4o-mini)

(d) FLUX.1 (GPT-5-nano)

(e) Gemini (GPT-5-nano)

(f) Qwen-Image (GPT-5-nano)

Figure 4: Average CLIP-I with reference ads trends over generated images using all methods for PartiPrompts with 3 T2I generators and 2 VLM judges.



(a) GEMINI

(b) GEMINI

(c) QWEN

(d) QWEN

Figure 5: KDE of single-image VLM scores versus DINOv3 similarity for **G** (a) Gemini 2.5 Flash Image and (b) Qwen-Image with $\mathcal{J}$ as GPT-4o-mini and GPT-5-nano.

Algorithm 2 Subroutines for PRISM-DUEL

Function UpdateConfidenceBounds(S_k, δ)
for $p_i, p_j \in S_k, p_i \neq p_j$ **do** ▷ Combine Real VLM preferences win (w, n) with Embedding Prior $(\tilde{w}, \tilde{n})$.
  $\hat{\mu}_{ij} \leftarrow (w_{ij} + \tilde{w}_{ij})/(n_{ij} + \tilde{n}_{ij})$
  $Rad_{ij} \leftarrow \sqrt{2\ln(t \cdot |S_k|^2/\delta)/(n_{ij} + \tilde{n}_{ij})}$
  $U_{ij} \leftarrow \hat{\mu}_{ij} + Rad_{ij}; \quad L_{ij} \leftarrow \hat{\mu}_{ij} - Rad_{ij}$
End Function

Function PruneSuboptimal(S_k)
  ▷ Count how many rivals i is guaranteed to beat $(L > 0.5)$
$CS_{min}(i) \leftarrow \sum_{j \neq i} \mathbb{I}(L_{ij} > 0.5)$
▷ Count how many rivals i could possibly beat $(U > 0.5)$
$CS_{max}(i) \leftarrow \sum_{j \neq i} \mathbb{I}(U_{ij} > 0.5)$
$BestScore \leftarrow \max_{i \in S_k} CS_{min}(i)$ **return** $\{i \in S_k \mid CS_{max}(i) \geq BestScore\}$
End Function

Function SelectUncertainPair(S_k)
$\mathcal{U} \leftarrow \{(i,j) \in S_k^2 \mid L_{ij} \leq 0.5 \leq U_{ij}\}$
if $\mathcal{U} \neq \emptyset$ **then return** $\text{argmax}_{(i,j) \in \mathcal{U}}(U_{ij} - L_{ij})$ ▷ Pick most uncertain
elsereturn $\emptyset$
End Function

C.1 LLM Feedback Generation

Similar to TextGrad (Yuksekgonul et al., 2024), the SummarizeFeedback utilizes an LLM to perform semantic summarization of dueling round, providing descriptive feedback for next round of mutations.

Algorithm 3 SummarizeFeedback for $\mathcal{F}$

Require: Candidate set $\mathcal{S}_k$; stats w, n; reference ad r.
HardPairs $\leftarrow \{(i,j) \mid n_{ij}$ is high and $|w_{ij}/n_{ij} - 0.5| < \epsilon\}$
Failures $\leftarrow$ Top 3 candidates with lowest Copeland score
fb $\leftarrow$ LLM.generate("Analyze why " + Failures+" lost compared to leader and ad "+ r)
return fb

D Complexity Analysis over all methods

Cost model. In black-box T2I prompt optimization, wall-clock time is dominated by (i) generator **G** calls and (ii) judge $\mathcal{J}$ calls. Let C_{gen} be the cost of one T2I generation and C_{judge} the cost of

VLM Judge ($\mathcal{J}$)	Method	PartiPrompts			DreamBooth		
		CLIP-I ↑	DINOv3 ↑	CLIP-T ↑	CLIP-I ↑	DINOv3 ↑	CLIP-T ↑
GPT-4o-mini	GPT-4o Caption (zero-shot)	0.764	0.508	0.341	0.712	0.594	0.305
	PRISM (He et al., 2024)	0.788	0.513	0.365	0.702	0.592	0.312
	SPO-T2I (Xiang et al., 2025)	0.768	0.498	0.332	0.780	0.592	0.311
	PDO-T2I (Wu et al., 2025c)	0.787	0.510	0.359	0.782	0.602	0.331
	PRISM-DUEL (ours)	**0.805**	**0.576**	**0.388**	**0.790**	**0.613**	**0.351**
GPT-5-nano	GPT-4o Caption (zero-shot)	0.761	0.511	0.343	0.718	0.589	0.299
	PRISM (He et al., 2025)	0.765	0.524	0.363	0.735	0.601	0.342
	SPO-T2I (Xiang et al., 2025)	0.769	0.517	0.351	0.744	0.608	0.337
	PDO-T2I (Wu et al., 2025c)	0.771	0.528	0.366	0.746	0.604	0.347
	PRISM-DUEL (Ours)	**0.791**	**0.599**	**0.387**	**0.782**	**0.628**	**0.369**

Table 3: **Quantitative Evaluation for Image quality across baselines with G as Gemini 2.5 Flash Image.**

VLM Judge ($\mathcal{J}$)	Method	PartiPrompts			DreamBooth		
		CLIP-I ↑	DINOv3 ↑	CLIP-T ↑	CLIP-I ↑	DINOv3 ↑	CLIP-T ↑
GPT-4o-mini	GPT-4o Caption (zero-shot)	0.802	0.505	0.314	0.721	0.534	0.303
	PRISM (He et al., 2024)	0.806	0.523	0.328	0.742	0.577	0.317
	SPO-T2I (Xiang et al., 2025)	0.812	0.542	**0.368**	0.772	0.582	0.341
	PDO-T2I (Wu et al., 2025c)	0.804	0.519	0.321	0.768	0.579	0.346
	PRISM-DUEL (ours)	**0.833**	**0.549**	0.366	**0.789**	**0.598**	**0.352**
GPT-5-nano	GPT-4o Caption (zero-shot)	0.795	0.485	0.318	0.727	0.529	0.301
	PRISM (He et al., 2024)	0.791	0.461	0.355	0.762	0.542	0.322
	SPO-T2I (Xiang et al., 2025)	0.779	0.465	0.344	0.759	0.562	0.318
	PDO-T2I (Wu et al., 2025c)	0.782	0.463	0.349	0.758	0.549	0.329
	PRISM-DUEL (ours)	**0.816**	**0.533**	**0.361**	**0.785**	**0.579**	**0.355**

Table 4: **Quantitative Evaluation for Image quality across baselines with G as Qwen-Image.**

one pairwise (or pointwise) judge query. We report complexity primarily in the *number of such calls*, and separately note the (typically smaller) CPU overhead for selection/bookkeeping.

D.1 PRISM-DUEL (ours): Copeland-UCB/SCB with a text-embedding prior

Setup. We run for T rounds. In each round we propose K_c new prompts and duel within the round set $S_k = C \bigcup \{p\}$ (size $K_c + 1$). We maintain a global pool $\mathcal{P}$ by accumulating proposed prompts across rounds. Each duel produces two images and one pairwise preference label, updating win counts (w_{ij}, n_{ij}).

Call complexity. With a duel budget of B comparisons per round, PRISM-DUEL incurs

$$T_{\text{calls}}^{\text{PRISM-DUEL}} \leq T \cdot (K_c \cdot C_{\text{gen}} + B \cdot C_{\mathcal{J}}) \quad (+ \text{ optional final evaluation}).$$

In our industrial deployment using Gemini 2.5 Flash ($\approx$ \$0.039/image) and GPT-4o mini/GPT-5-nano as judge, removing the factor of 2 from the generation cost (via caching) and capping B (via SCB pruning) reduces total round cost by approximately 40–60% compared to a naive round-robin tournament.

Embedding-prior warm start. The warm-start procedure requires embedding K_c new prompts per round. The cost is $O(T \cdot K_c \cdot C_{\text{embed}})$. Given that $C_{\text{embed}} \ll C_{\text{gen}}$ (e.g., `text-embedding-3-small` costs \$0.02/1M tokens vs. \$30+/1M tokens for VLM generation), this overhead is negligible ($<$ 0.1% of total cost). Mathematically, the pseudo-counts $(\tilde{w}, \tilde{n})$ provide $O(1)$ "virtual samples," reducing the number of real VLM calls n_{ij} required to shrink the confidence radius Rad_{ij} below the pruning threshold.

Selection/CPU overhead and memory. By adopting Scalable Copeland Bandits, we avoid the $O(K^2)$ complexity of full tournaments. The pruning mechanism reduces the sorting complexity to $\tilde{O}(K \log K)$ in the number of active candidates. Regarding memory, we only store a sparse adjacency matrix of active comparisons, requiring $O(B)$ space, which is significantly more efficient than the dense $O(K^2)$ storage required by full Double Thompson Sampling (Wu et al., 2025c).

D.2 PRISM (He et al., 2024)

PRISM runs K_c parallel streams for T refinement iterations, sampling a prompt, generating an image, scoring it, and updating the prompt distribution each iteration, then re-evaluates a top subset at the end (He et al., 2024). Concretely, PRISM is an *iterative sampling* loop repeated for a predetermined budget of iterations/streams and then re-evaluates the top prompts.

Call complexity. PRISM's main loop cost is

$$T_{\text{calls}}^{\text{PRISM}} = (K_c \cdot T) \cdot (C_{\text{gen}} + C_{\text{judge}}) + M \cdot (C_{\text{gen}} + C_{\text{judge}}),$$

where C_{judge} is the cost of pointwise similarity VLM judge scorer and M is the number of top prompts re-evaluated. Thus PRISM scales linearly in the total number of iterations across streams, with no K^2 pairwise matrix.

Remarks. PRISM optimizes via *distribution refinement and pointwise scoring* rather than explicit dueling-based identification; its time is primarily controlled by $(K_c \cdot T)$ and final re-evaluation.

D.3 SPO (Xiang et al., 2025): Pairwise Comparison

Setup. SPO performs *sequential* improvement with an incumbent prompt p (i.e. the current best prompt). At iteration t, a LLM-based optimizer or the prompt engineer under PRISM setting $\mathcal{F}$, proposes a new prompt p', conditioned on textual feedback, then the method *executes* both prompts and uses a judge to decide whether to accept p' as the new incumbent (Xiang et al., 2025).

Vanilla SPO primarily studies on text-only tasks where a prompt returns a (nearly) deterministic answer with decoding temperature as 0. T2I generation is inherently stochastic: a fixed prompt shall introduce a distribution over images due to random seeds (Xu et al., 2025), caption noise (Yu et al., 2024) and decoder noise. Adapting SPO to T2I requires estimating the prompt's *expected* preference, rather than relying on a single draw.

Therefore, to form a lower variance estimate of whether the proposed new prompt p' improves over the current best prompt p_t, we evaluate prompts by generating images under a small set of randomized seeds. Let m denote the number of seeds used per iteration to reduce judge variance.

Call complexity (T2I adaptation). At each iteration $t = 1, \ldots, T$: (i) generate m images for the proposed new prompt p'; (ii) compare the outputs of p' against the incumbent outputs using m pairwise judgements with average win-rate; (iii) accept/reject and update the textual feedback for the next proposal. We assume generated images for each iteration are cached and reused across iterations, only the proposed new prompt p' per iteration needs new image generations.

With caching of past prompts' images, SPO's dominant cost over T iterations is

$$T_{\text{calls}}^{\text{SPO}} = T \cdot \Big(m\, C_{\text{gen}} + m\, C_{\text{judge}} \Big) \quad (+\ \text{LLM prompt-update overhead}).$$

Without caching (regenerating both sides each iteration), the worst-case generator term doubles:

$$T_{\text{calls}}^{\text{SPO,worst}} = T \cdot \Big(2m\, C_{\text{gen}} + m\, C_{\text{judge}} \Big).$$

Thus SPO scales *linearly* with the iteration budget K and the per-iteration replication m.

CPU overhead and memory. CPU bookkeeping is $O(1)$ per iteration (accept/reject and feedback update). Memory is $O(m)$ if caching only the incumbent images for m seeds, or $O(Tm)$ if all historical outputs are retained. Unlike dueling-bandit pool methods, SPO does not maintain a dense $O(T^2)$ pairwise table.

Stopping. SPO is *budgeted*: given an iteration budget T, it runs for exactly T iterations and returns the final best prompt

$$p^\star := p_T,$$

where p_T denotes the selected prompt after iteration T (Xiang et al., 2025).

D.4 PDO (Wu et al., 2025c): Double Thompson Sampling + mutation

PDO frames prompt optimization as a dueling bandit problem solved via Double Thompson Sampling (D-TS). Unlike all previous round-based appraoch, PDO maintains a dynamic candidate pool $\mathcal{P}$ initialized with a substantial number of candidates and requires a relatively large starting pool $|\mathcal{P}_0|$, typically $|\mathcal{P}_0| = 50$ (Wu et al., 2025c)), and iteratively selects pairs to duel based on their posterior probability of being the Condorcet winner. Notably, in practical T2I prompt optimization scenarios, practitioners rarely have access to a large initial pool of high-quality prompts (Hao et al., 2023).

Call complexity. PDO operates under a global query budget B_{total}. The total call is dominated by the number of comparisons allocated to explore the pool:

$$T_{\text{calls}}^{\text{PDO}} \approx B_{\text{total}} \cdot C_{\text{judge}} + N_{\text{mutations}} \cdot C_{\text{gen}}$$

Because D-TS probabilistically samples candidates rather than aggressively pruning them (as PRISM-DUEL does), it typically requires a larger B_{total} to converge to a Copeland winner compared to elimination-based methods. Specifically, theoretically regret bounds for D-TS scale as $O(K^2 \log T)$, implying that the judging budget must grow quadratically with the pool size K to ensure reliable optimization.

CPU overhead and memory. The primary bottleneck for PDO in an industrial setting is the quadratic state requirement:

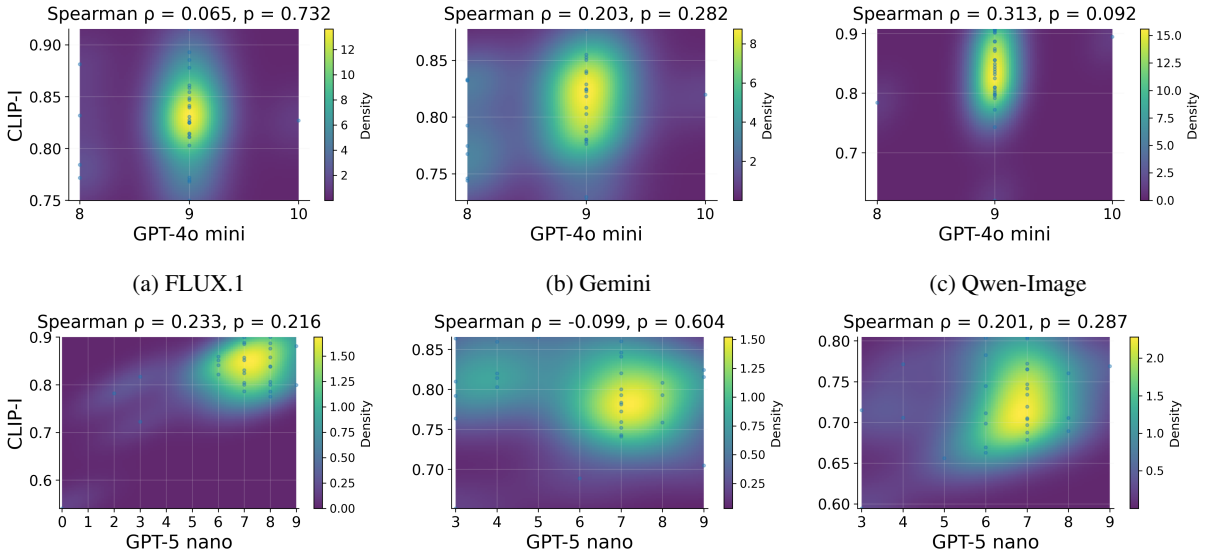
- **Memory:** PDO maintains Beta posteriors (W_{ij}, W_{ij}) for all pairwise combinations in the pool, where W_{ij} and W_{ji} be the current number of wins of p_i over p_j, requiring $O(K^2)$ space.
- **Compute:** To select a single duel, D-TS requires sampling from the posterior of every candidate in the pool and solving a maximization problem. This incurs a selection cost of $O(K^2)$ per iteration, which becomes computationally significant as the pool expands during the mutation phase.

D.5 Summary

PRISM-DUEL distinguishes itself by acting as the Pareto-optimal solution for industrial T2I prompt optimization for image inversion: it adopts the ranking benefits of pairwise bandits (unlike PRISM) without incurring the quadratic scaling of standard bandit algorithm (unlike PDO).

E Additional Qualitative Results

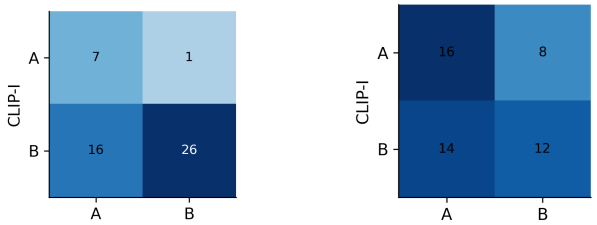
In Figure 11, we present additional qualitative examples for subject-driven personalized T2I generation.



(a) FLUX.1 (b) Gemini (c) Qwen-Image

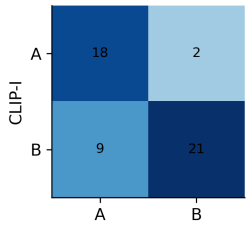
(d) FLUX.1 (e) Gemini (f) Qwen-Image

Figure 6: KDE of single-image VLM scores versus CLIP-I similarity across 3 T2I models and two VLM judges. Each panel shows the joint distribution of scores (color indicates density, points are individual images) together with the Spearman rank correlation ρ between VLM and CLIP-I scores. Correlations are weak and statistically insignificant in most cases, with substantial variation in CLIP-I at each discrete VLM score and clear saturation of VLM scores at the high end of the scale. Single-image VLM ratings are noisy and only weakly aligned with CLIP-I scores.

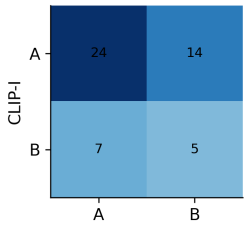


(a) FLUX.1(GPT-4o-mini)

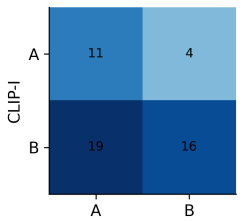
(b) FLUX.1(GPT-5-nano)



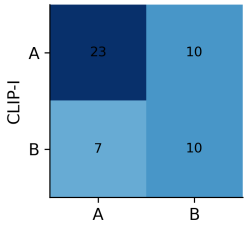
(c) QWEN(GPT-4o-mini)



(d) QWEN(GPT-5-nano)

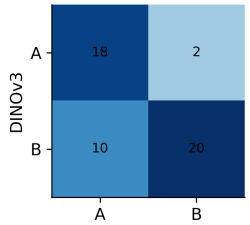


(e) GEMINI(GPT-4o-mini)

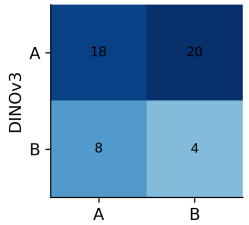


(f) GEMINI(GPT-5-nano)

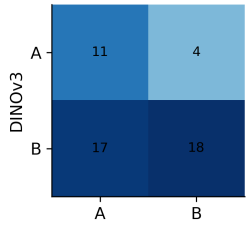
Figure 7: Confusion matrix for CLIP-I scores with $\mathbf{G}$ as FLUX.1, QWEN-Image and Gemini 2.5 Flash Image and $\mathcal{J}$ as GPT-4o-mini and GPT-5-nano.



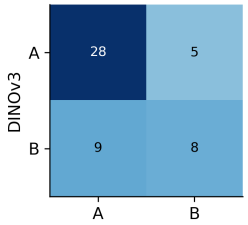
(a) QWEN(GPT-4o-mini)



(b) QWEN(GPT-5-nano)

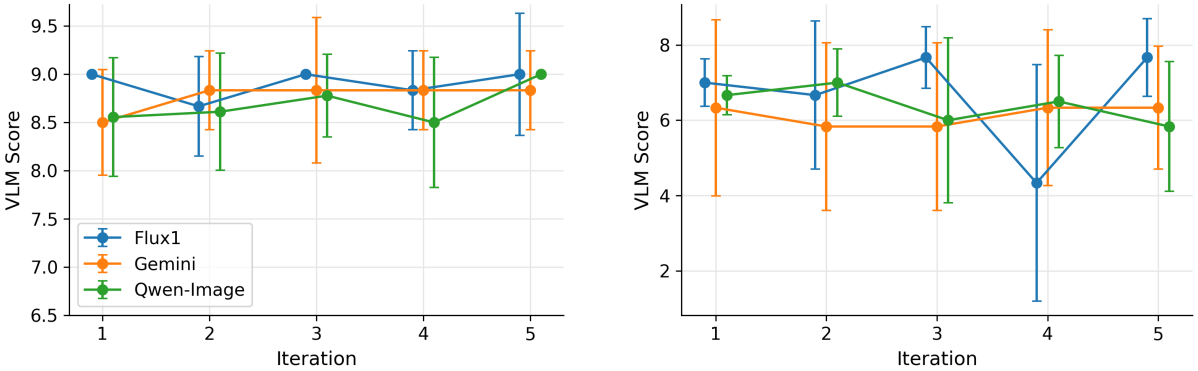


(c) GEMINI(GPT-4o-mini)



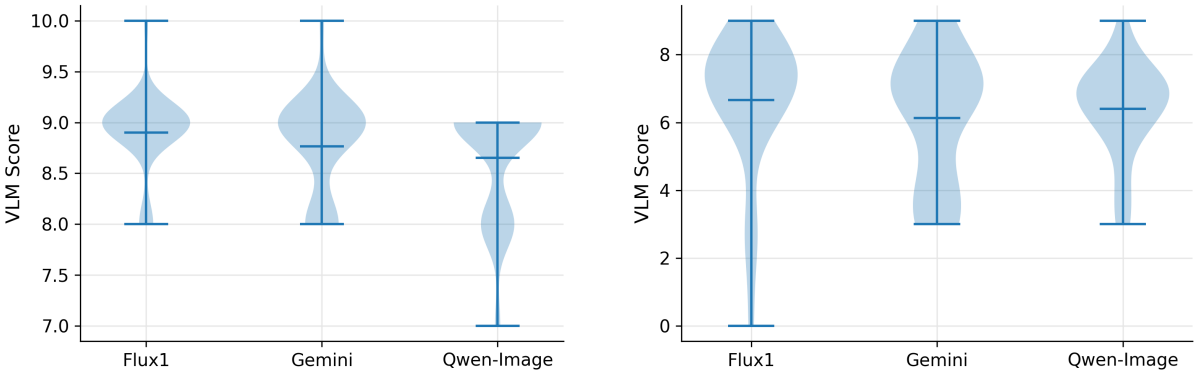
(d) GEMINI(GPT-5-nano)

Figure 8: Confusion matrix for DINOv3 scores with $\mathbf{G}$ as Qwen-Image and Gemini 2.5 Flash Image and $\mathcal{J}$ as GPT-4o-mini and GPT-5-nano.



(a) GPT-4o mini (b) GPT-5 nano

Figure 9: Per-iteration of mean VLM judges scores for PRISM. Mean scores for all T2I models lie in a narrow band for each judge with largely overlapping error bars, suggesting indifferentiation of single image evaluation scores.



(a) GPT-4o mini (b) GPT-5 nano

Figure 10: Violin plots of VLM score distributions of T2I models for GPT-4o-mini (a) and GPT-5-nano (b). Both VLM models have scores centering around 9 for GPT-4o-mini and 7 for GPT-5-nano.

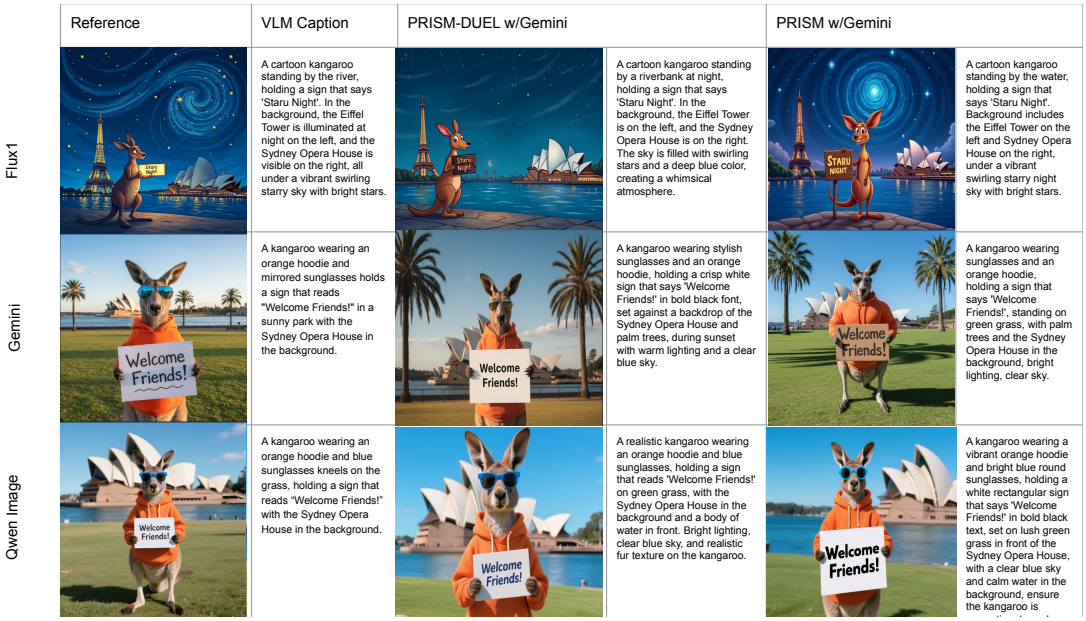


Figure 11: Pointwise vs. Pairwise Judgements. Qualitative results of PartiPrompts using Gemini as the T2I Generator $\mathbf{G}$ with VLM as GPT-4o mini. PRISM-DUEL, which updates prompts via pairwise duels and textual embeddings, produces images that are more faithful to the reference subject(identity-perserving details and overall appearance) than standard PRISM.

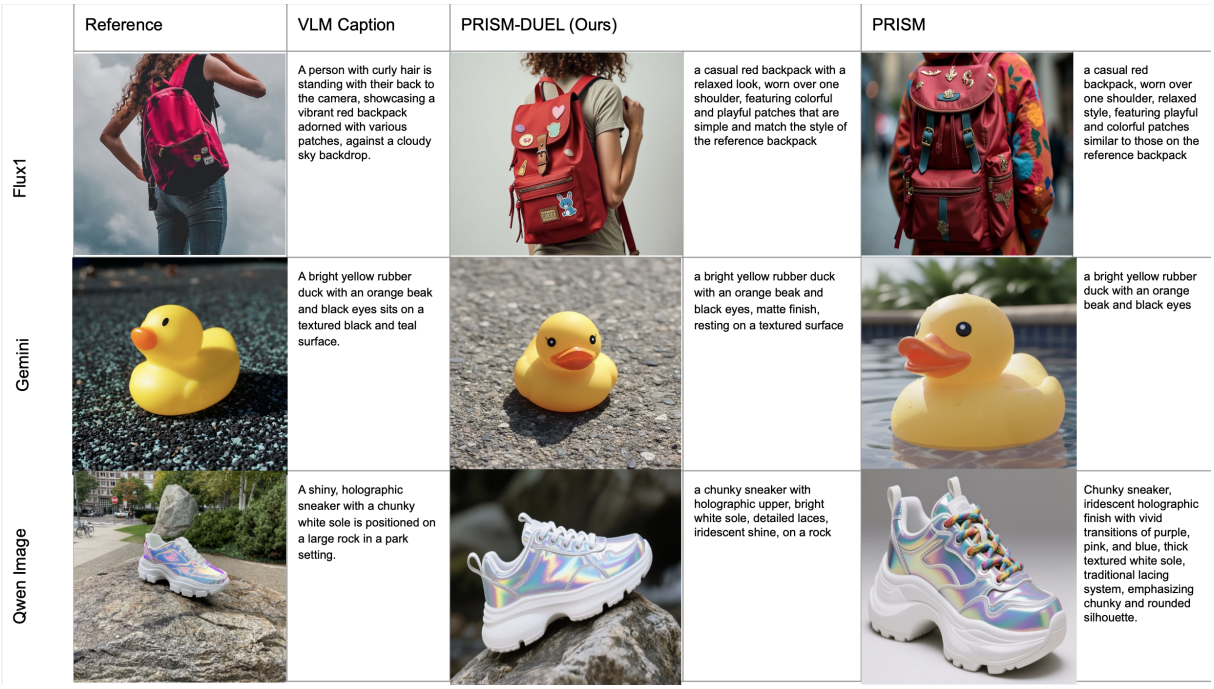


Figure 12: Pointwise vs. Pairwise Judgements. Qualitative personalization results on DreamBooth. For each reference image (left), we compare the initial VLM caption, the final prompt and a generated image by PRISM-DUEL (pairwise duels + textual embedding guidance), and the final prompt and generated image by PRISM. Across backbones (FLUX.1/Gemini/Qwen-Image), PRISM-DUEL better preserves fine-grained instance attributes (e.g., woman with the curly hair wearing the backpack, matte rubber-duck appearance on a textured surface instead of zero content on the surface as PRISM selected prompt does, sneaker on the rock).