Sem-DPO: Mitigating Semantic Inconsistency in Preference Optimization for Prompt Engineering

Anas Mohamed[1,*] **Azal Ahmad Khan**[1,*]
Xinran Wang[1] **Ahmad Faraz Khan**[2] **Shuwen Ge**[3]
Saman Bahzad Khan[4] **Ayaan Ahmad**[5] **Ali Anwar**[1]
[1]University of Minnesota [2]Virginia Tech [3]Xi'an University of Technology
[4]Lahore University of Management Sciences [5]UC, Santa Cruz [*]Equal contributions
{moha1325, khan1069, wang8740, aanwar}@umn.edu, ahmadfk@vt.edu
shuwen8681@gmail.com, 27100111@lums.edu.pk, ayahmad@ucsc.edu

Abstract

Generative AI can now synthesize strikingly realistic images from text, yet output quality remains highly sensitive to how prompts are phrased. Direct Preference Optimization (DPO) offers a lightweight, off-policy alternative to RL for automatic prompt engineering, but its token-level regularization leaves semantic inconsistency unchecked as prompts that win higher preference scores can still drift away from the user's intended meaning.

We introduce Sem-DPO, a variant of DPO that preserves semantic consistency yet retains its simplicity and efficiency. Sem-DPO adjusts the DPO loss using a weight based on how different the winning prompt is from the original, reducing the impact of training examples that are semantically misaligned. We provide the first analytical bound on semantic drift for preference-tuned prompt generators, showing that Sem-DPO keeps learned prompts within a provably bounded neighborhood of the original text. On three standard text-to-image prompt-optimization benchmarks and three language models, Sem-DPO achieves 8–12% higher CLIP similarity and 5–9% higher human-preference scores (HPSv2.1, PickScore) than DPO, while also outperforming state-of-the-art prompt optimization baselines as well as several DPO variants. These findings suggest that strong flat baselines augmented with semantic weighting should become the new standard for prompt-optimization studies and lay the groundwork for broader, semantics-aware preference optimization in language models.

1 Introduction

Recent advances in Generative AI have democratized creative expression, enabling users to generate high-quality images from textual input prompts (Ramesh et al., 2021; Saharia et al., 2022; Yu et al., 2022; Rombach et al., 2022; Ramesh et al., 2022). However, the quality of these outputs remains highly dependent on the design of input prompts, which must precisely represent styles, and contexts to guide generative models. Developing effective prompts is a non-trivial task, and users often resort to trial-and-error or manual engineering, a process that is labor-intensive and model-specific (Reynolds and McDonell, 2021). As a result, recent work has investigated the use of Large Language Models (LLMs) to automate prompt engineering by paraphrasing and stylistic augmentation. This has led to the development of strategies for fine-tuning LLMs to generate prompts that optimize proxy metrics such as aesthetic score, using reinforcement learning (RL)-based approaches (Hao et al., 2022; Cao et al., 2023). However, these methods suffer from poor human alignment, since maximizing surrogate metrics does not necessarily reflect user preferences, as prompts optimized for aesthetic score, for instance, might produce overly stylized images that deviate from the user's intended meaning. Moreover, they can incur significant computational costs due to on-policy sampling (e.g., RLAIF) (Lee et al., 2023; Lindström et al., 2024; Agarwal et al., 2024).

Challenges. To address the challenges of human alignment and cost of RL-based approaches, Direct Preference Optimization (DPO) has emerged as a compelling off-policy alternative (Rafailov et al., 2023). DPO aligns the policy with preference data by maximizing a contrastive likelihood objective over preferred and dispreferred samples, typically from human or proxy annotator comparisons. While DPO has shown strong empirical performance in dialogue and instruction tuning tasks (Khan et al., 2024; Saeidi et al., 2024; Dong et al., 2024; Jung et al., 2025), its application to prompt optimization reveals a critical weakness. As shown in Figure 1, DPO improves human preference alignment, i.e., it generates outputs more liked by humans, but often reduces semantic consistency, where outputs stray from the original

Findings of the Association for Computational Linguistics: ACL 2026, pages 23656–23674

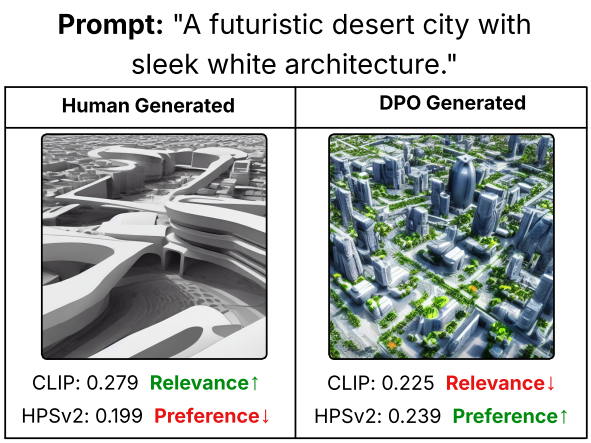


Figure 1: An illustration of semantic drift in prompt optimization. Note how the DPO generated image, while achieving a higher human preference score, deviates semantically from the original prompt's intent (e.g. emphasizing lush greenery despite the desert descriptor). This results in a lower CLIP score, showing the preference vs fidelity tradeoff that Sem-DPO addresses.

prompt's meaning (as measured by CLIP). This suggests that the DPO-optimized prompts, though more aligned with human preferences, drift semantically from the intended meaning. We attribute this to the structure of the DPO gradient, which is driven by token-level likelihood ratios and provides no signal for preserving the semantic distance between the input and the preferred output (Razin et al., 2024; Yang et al., 2024a). Recent work shows that even sophisticated prompt optimizers (Zhu et al., 2025) and pair-reweighted DPO (Amini et al., 2024) still overlook semantic drift. In compound AI systems, where a prompt optimizer feeds a downstream generator, unconstrained preference tuning can propagate semantic drift through the pipeline, so preserving input meaning is a key reliability concern.

Prior Works. Several variants of DPO have been proposed to improve optimization stability and alignment robustness, such as KL-regularized DPO (Wang et al., 2023), T-REG (Zhou et al., 2024b), β-DPO (Wu et al., 2024) and two-stage filtering pipelines (Shan et al., 2025). More recently, semantic-aware Kernel-based Preference Optimization (KTO; Ethayarajh et al. (2024)) has been introduced. KTO compares model outputs using divergence-based similarity metrics, enabling the model to capture richer preference structures beyond simple log-likelihood ratios and helping preserve semantic content during training. However, these approaches still primarily operate within the token-probability or distributional space, lacking direct integration of semantic meaning into the loss function itself. Critically, they do not penalize preference pairs in which the selected output is semantically misaligned with the input, thereby implicitly rewarding stylistic overfitting.

This motivates our central question. *How can we adapt DPO to optimize prompts so the outputs are both highly preferred by humans and remain semantically faithful to the original input?*

Our Approach. We propose *Sem-DPO* (Semantic-DPO), a variant of DPO that explicitly incorporates semantic consistency into the preference alignment process. While current methods focus primarily on optimizing token-level log-likelihood ratios, Sem-DPO introduces an importance weighting mechanism that adjusts the influence of each training sample based on its semantic alignment with the original prompt. Specifically, for each triplet (x, y_w, y_l), we compute a semantic consistency weight $W(x, y_w) = \exp(-\alpha \cdot d(e_\varphi(x), e_\varphi(y_w)))$, where $e_\varphi(\cdot)$ is an embedding function and $d(\cdot, \cdot)$ is a cosine distance metric. This weight downscales the gradient contribution of preference pairs where the preferred output semantically diverges from the input, thereby mitigating the semantic inconsistency often observed in prompt optimization. Crucially, the weighting is computed offline and integrated to the loss function, preserving the computational efficiency and training simplicity of DPO. Sem-DPO thus fills the gap of DPO framework to enforce preference alignment and semantic consistency, making it better for tasks where preserving input intent is critical.

Contributions. This work makes *four decisive advances* toward trustworthy prompt optimization for text-to-image models. **(C1) Problem Formulation.** We are the first to cast *semantic drift* as a core failure of preference-based prompt optimization, turning an often-ignored side-effect into a formal learning objective. **(C2) Semantically-Weighted DPO Objective.** We introduce *Sem-DPO*, a lightweight, powerful variant of DPO that injects semantic-aware importance weights into the loss, seamlessly combining human preference alignment with semantic fidelity, *without* sacrificing DPO's performance and simplicity. **(C3) Theoretical Analysis.** We provide the *first* theoretical guarantees that explicitly bound semantic drift

under preference optimization, offering provable insight into when and why Sem-DPO outperforms prior work. **(C4) Experimental Analysis.** Extensive experiments on three benchmark datasets and five language models show Sem-DPO consistently offers the best trade-off between human preference and semantic alignment, outperforming baselines across automated metrics and human evaluation (82.2% preference in direct comparison), setting a new state of the art in prompt optimization.

2 Related Works

Preference alignment has evolved from RLHF, which relies on reward models and costly policy optimization (Bai et al., 2022; Ouyang et al., 2022), to DPO, which reformulates alignment as a closed-form contrastive classification problem (Rafailov et al., 2023). DPO has since been extended across domains (Dong et al., 2024; Yang et al., 2024b; Sun et al., 2024; Lu et al., 2024; Khaki et al., 2024; Wallace et al., 2024) with contrastive and entropy-based adjustments (Xiao et al., 2024; Omura et al., 2024). Recent work introduces regularization via entropy penalties (Omura et al., 2024), calibrated objectives (Xiao et al., 2024), feature-level constraints (Yin et al., 2024), and adversarial self-play (Tang et al., 2025). Semantic-aware approaches couple probability-space and embedding-space objectives (Das et al., 2025; Cai et al., 2025; Zhao et al., 2023), though many remain modality-specific or annotation-heavy (Shekhar et al., 2024; Jinnai and Honda, 2024), while reweighting schemes focus only on preference margins (Zhou et al., 2024a). Existing methods either treat semantics as post-hoc filters or rely on heuristics, without directly integrating semantic consistency into the preference loss or offering guarantees against prompt-level drift. A complete discussion is in Appendix A.

3 Proposed Approach

This section begins by providing a theoretical background of DPO. Following this, we introduce Sem-DPO, our proposed variant of DPO that explicitly incorporates semantic consistency while retaining DPO's simplicity and efficiency.

Preliminaries. Reinforcement Learning from Human Feedback (RLHF) aligns models with human preferences using a dataset of triplets $\mathcal{D} = \{(x, y_w, y_l)\}$, where a response y_w is preferred over y_l for a given input x. While traditional RLHF first trains a reward model and then uses it to optimize a policy, DPO streamlines this into a single-stage training process. DPO achieves this by establishing a direct analytical link between the reward function and the optimal policy π^*. Specifically, the reward function is reparameterized in terms of the policy as:

$$r^*(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)$$

By substituting this relationship into the Bradley-Terry preference model, DPO derives a simple maximum likelihood objective that directly optimizes the language model policy π_θ on the preference data. This avoids the complexities of training a separate reward model by instead minimizing the following negative log-likelihood loss:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \quad (1)$$

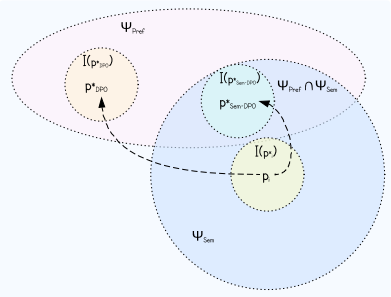


Figure 2: **Concept Figure**. DPO steers prompts into the preference region, while the ideal target lies at the intersection of preference and semantic regions. Sem-DPO is designed to reach that overlap

3.1 Motivation

The goal of automated prompt optimization is to refine an initial user prompt, p_i, to generate a superior image. An ideal prompt must satisfy two conditions: it must produce a high-quality image preferred by humans, thus lying in a high-preference region (Ψ_{Pref}), and it must preserve the core meaning of the original request, remaining within a semantic consistency space (Ψ_{Sem}), as shown in Figure 2. While powerful methods like DPO are effective at navigating the prompt space to find solutions (p^*_{DPO}) that score high on preference metrics, they suffer from a critical flaw. Since DPO operates at a token level without an intrinsic understanding of meaning, its optimization path often exits the semantic consistency space. This "Semantic Drift"

results in a final prompt that, although optimized for preference, generates an image that is semantically inconsistent from the user's intent.

To address this, we propose Sem-DPO, a method designed to find an optimal prompt, $p^*_{\text{Sem-DPO}}$, that resides in the intersection of both the preference and semantic spaces ($\Psi_{\text{Pref}} \cap \Psi_{\text{Sem}}$). Our approach augments the standard DPO objective with a semantic-aware weighting mechanism. This mechanism calculates the semantic distance between candidate prompts and the initial prompt p_{i}, penalizing any updates that cause significant semantic drift. This penalty acts as a corrective force, constraining the optimization search to remain within the semantic consistency space while simultaneously seeking higher-preference solutions. As a result, Sem-DPO produces a prompt that is not only preferred but also semantically faithful, ensuring the final generated image successfully aligns with the user's intent.

Problem Formulation The task of preference-based prompt optimization is to fine-tune a policy model, π_θ to generate improved prompts. Standard DPO achieves this by minimizing Equation 1.

Definition 1 (Semantic Drift) *Let $e_\varphi : \mathcal{X} \to \mathbb{R}^\delta$ be a frozen pre-trained embedding model and $d_{cos}(\cdot,\cdot)$ a cosine distance measure. For an initial prompt x and an optimized prompt y, we define the* semantic drift $d(x, y) = d_{cos}\big(e_\varphi(x),\, e_\varphi(y)\big)$. *The prompt y exhibits significant* semantic drift *whenever $d(x, y) \geq \tau$ for a user-chosen threshold.*

The central issue we address is that the DPO loss operates at the token level and includes no term to penalize semantic drift. Consequently, a DPO-finetuned model may generate a prompt that produces a stylistically superior yet semantically inconsistent image with the original prompt, a phenomenon we observe empirically (Figure 1).

3.2 *Sem-DPO*: Semantic Direct Preference Optimization

To mitigate the semantic drift in the standard DPO framework, we propose Semantic Direct Preference Optimization (Sem-DPO). Our approach modifies the DPO objective by introducing a per-sample semantic weight that is a function of the semantic similarity between the input prompt (x) and the preferred output (y_w). Specifically, we define a semantic consistency weight $W_\alpha(x, y_w)$ as:

$$W_\alpha(x, y_w) = \exp\Big(-\alpha \cdot d_{\cos}\big(e_\varphi(x), e_\varphi(y_w)\big)\Big) \tag{2}$$

where, $\alpha \geq 0$ is a hyperparameter that controls the strength of the semantic weighting. This weight is then incorporated directly into the DPO loss function, yielding the Sem-DPO objective:

$$\mathcal{L}_{\text{Sem-DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}$$
$$W_\alpha(x, y_w) \cdot \log\left(\sigma\left(\beta\Big(\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\Big)\right)\right)$$

This weighting scheme (Equation 2) down-weights training samples where the preferred prompt semantically drifts from the input. Pairs with high semantic similarity receive higher weights, preserving their gradient influence, while those with low similarity are exponentially suppressed. This encourages outputs that are both aesthetically and semantically aligned, ensuring the optimized prompt stays within the target region $\Psi_{\text{Pref}} \cap \Psi_{\text{Sem}}$. Since weights are computed offline using a frozen encoder, this incurs no additional overhead and retains the original DPO's efficiency.

4 Theoretical Analysis

We present a theoretical analysis of Sem-DPO, establishing two key results that justify its design. First, we show that the exponential weighting provides a smooth, stable approximation to hard semantic filtering. Second, we prove that by controlling semantic drift, Sem-DPO bounds the drift between the generated image and the input prompt. The log-odds difference used in DPO is:

$$\Delta_\theta(x, y_w, y_l) = \beta\left(\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right) \tag{3}$$

$$\ell(\Delta) = -\log\sigma(\Delta) \tag{4}$$

Smooth Filtering Mechanism. Sem-DPO employs a smooth exponential kernel to down-weight semantically divergent samples, avoiding the discontinuities introduced by hard filtering based on a fixed threshold. Our first proposition shows that this yields a stable relaxation of hard filtering, with bounded approximation error.

Proposition 1 (Uniform Deviation Bound) *Sem-DPO's exponential importance-weighting, $W_\alpha(d) = \exp(-\alpha d)$, serves as a smooth relaxation of hard semantic filtering, with a bounded approximation error.*

Remark (distance). Sem-DPO measures semantic drift using cosine distance in a frozen CLIP-style embedding space. For results that require a proper metric—most notably Proposition 2, where we invoke the triangle inequality—we instead work with the Euclidean distance between ℓ_2-normalized embeddings:

$$d_{\text{metric}}(x, y) = \|e_\phi(x) - e_\phi(y)\|_2.$$

For unit-normalized embeddings u, v, cosine similarity and Euclidean distance are monotonically equivalent via

$$\|u - v\|_2^2 = 2\big(1 - \cos(u, v)\big).$$

Thus, when embeddings are normalized, using cosine-based drift for weighting and Euclidean distance for the metric argument is consistent up to a monotone transformation.

Proof. Let hard semantic filtering be defined by an objective function that uses an indicator function $1_{\{d \le \tau\}}$ to discard examples where the semantic distance d exceeds a threshold $\tau > 0$:

$$\mathcal{L}_\tau(\theta) = \mathbb{E}_\mathcal{D}[1_{\{d \le \tau\}} l(\Delta_\theta)]$$

where $l(\Delta_\theta) = -\log \sigma(\Delta_\theta)$ is the standard DPO loss term. The threshold τ is used only to define an idealized hard-filter objective for analysis (Proposition 1). Sem-DPO training itself does not require choosing or tuning τ.

Sem-DPO replaces this discontinuous indicator function with the smooth exponential kernel $W_\alpha(d)$, yielding the objective:

$$\mathcal{L}_{\text{Sem-DPO}}(\theta) = \mathbb{E}_\mathcal{D}[W_\alpha(d) l(\Delta_\theta)]$$

Assuming the loss term is bounded, such that $|l(\cdot)| \le M$ for some constant M, the absolute deviation between the Sem-DPO objective and the hard-filtered objective is bounded as follows:

$$\begin{aligned} \big|\mathcal{L}_{\text{Sem-DPO}}(\theta) - \mathcal{L}_\tau(\theta)\big| &= \Big|\mathbb{E}_\mathcal{D}\big[(W_\alpha(d) - \mathbf{1}_{d \le \tau})\, \ell(\Delta_\theta)\big]\Big| \\ &\le \mathbb{E}_\mathcal{D}\big[\big|W_\alpha(d) - \mathbf{1}_{d \le \tau}\big|\, \big|\ell(\Delta_\theta)\big|\big] \\ &\le M\, \mathbb{E}_\mathcal{D}\big[\big|W_\alpha(d) - \mathbf{1}_{d \le \tau}\big|\big]. \end{aligned}$$

The maximum pointwise difference between the weighting function $W_\alpha(d)$ and the indicator function occurs at the discontinuity point $d = \tau$ and is bounded by $1 - e^{-\alpha\tau}$. Thus, the expected deviation is also bounded:

$$|\mathcal{L}_{\text{Sem-DPO}}(\theta) - \mathcal{L}_\tau(\theta)| \le M(1 - e^{-\alpha\tau})$$

which completes the proof.

Interpretation. Proposition 1 shows that the exponential weighting used by Sem-DPO is a smooth surrogate for an idealized hard semantic filter. The bound

$$|L_{\text{Sem-DPO}}(\theta) - L_\tau(\theta)| \le M\big(1 - e^{-\alpha\tau}\big)$$

makes explicit how τ and M affect the approximation. Here, τ is the reference hard-filter threshold used only for analysis: larger τ makes the hard filter less selective and correspondingly increases the worst-case deviation upper bound, approaching M as $\tau \to \infty$. The constant M is a conservative worst-case bound on the magnitude of the per-example DPO loss term $|\ell(\Delta)|$. In practice, the effective deviation can be much smaller when logits or observed loss values remain in a moderate range during training.

Guarantee for Semantic Consistency. We now present our main theoretical result, which provides a formal guarantee that Sem-DPO achieves its ultimate goal of ensuring the final, high-preference image is semantically aligned with the original user prompt. This is predicated on a reasonable assumption about the quality of the underlying text-to-image generator.

Assumption 1 (T2I Consistency Error) *For any prompt y and a high-quality text-to-image generator I, the consistency error $d_{T2I}(y) = \|e_(y) - E_{Img}((I(y))\|$, which measures the distance between the prompt's embedding and its corresponding image's embedding in a shared space, is bounded by a small constant ϵ.*

$$\|e_{(y)} - E_{\text{Img}}(I(y))\| \le \epsilon$$

This implies the generated image remains semantically close to its prompt, within an ϵ-bounded neighborhood in embedding space.

Proposition 2 (Bounded Semantic Drift) *The semantic distance between the original prompt x, optimized prompt y and the final image $I(y)$, denoted $d_{T2I\text{-}Drift}(x, y)$, is upper-bounded by the sum of the semantic drift and the T2I consistency error of the generator.*

Proof. Let $e(x)$ and $e(y)$ be the text embeddings of prompts x and y, and let $E_{\text{Img}}(I(y))$ be the image embedding of the generated image from prompt y. The text-to-image drift is defined as $d_{\text{T2I-Drift}}(x, y) = \|e_(x) - E_{\text{Img}}(I(y))\|$.

By applying the triangle inequality property of vector norms, we can establish an upper bound:

$$\|e(x) - E_{\text{Img}}(I(y))\| \leq \|e(x) - e(y)\| + \|e(y) - E_{\text{Img}}(I(y))\|$$

This inequality can be expressed using our distance notations:

$$\underbrace{d_{\text{T2I-Drift}}(x, y)}_{\text{total drift}} \leq \underbrace{d_{\text{Semantic-Drift}}(x, y)}_{\text{prompt-level drift}} + \underbrace{d_{\text{T2I}}(y)}_{\text{generator-level drift}}$$

Here $d_{\text{Semantic-Drift}}(x, y)$ denotes the metric drift $\|e_\phi(x) - e_\phi(y)\|_2$ in the shared embedding space. Applying our T2I Consistency Error Assumption, where $d_{\text{T2I}}(y) \leq \epsilon$, we arrive at the final bound:

$$d_{\text{T2I-Drift}}(x, y) \leq d_{\text{Semantic-Drift}}(x, y) + \epsilon$$

which completes the proof.

Interpretation. In this decomposition, $d_{\text{Semantic-Drift}}(x, y)$ is the *prompt-level* drift induced by the prompt optimizer (the term Sem-DPO is designed to reduce), whereas $d_{\text{T2I}}(y)$ is the *generator-level* inconsistency of the downstream T2I model (assumed $\leq \epsilon$). Thus $d_{\text{T2I-Drift}}(x, y)$ measures the end-to-end drift of the full text-to-image system. More generally, the same argument extends to all compound AI pipelines composed of multiple generative modules.

5 Evaluation

5.1 Experimental Setup

Datasets and Baselines. To assess the efficacy of Sem-DPO, we conduct experiments on three publicly available datasets. For in-domain testing, we utilize DiffusionDB and Lexica. For out-of-domain generalization assessment, we use the COCO dataset. Following prior work, the prompts for our experiments are sourced from Hao et al. (2022), with 200 prompts selected from each dataset. Our evaluation spans five prompt-generation backbones: Qwen1.5-1.8B, Llama3.2-1b, GPT-2, Qwen3-0.6b and Qwen3-1.7b. We compare Sem-DPO across a set of baselines, including **Human Input**, **SFT**, and a vanilla **DPO**. *To align with earlier prompt-optimization studies that build on older models (GPT-2), we additionally repeat all experiments with a GPT-2 backbone* and, in that setting, report results for **Promptist** (Hao et al., 2022) and **BeautifulPrompt** (Cao et al., 2023) alongside GPT-2 versions of SFT and DPO. Additionally, we compare each model against contemporary modern DPO variants including KTO (Ethayarajh et al., 2024), ORPO (Hong et al., 2024), and β-DPO (Wu et al., 2024) using the different model backbones.

Metrics. We evaluate the generated images using a combination of metrics to assess semantic consistency and human preference alignment. These include the CLIP Score, to measure the semantic relevance between the generated image and the input prompt, where higher scores indicate better semantic alignment. We also use PickScore (Kirstain et al., 2024), an image quality metric trained on human preferences reflecting the likelihood a human would prefer a given image, and HPSv2.1 (Wu et al., 2023), a Human Preference Score also trained to predict human judgments of image quality and alignment. These metrics provide quantitative measures to assess the quality and relevance of the generated images based on different criteria.

Preference Labels and Finetuning. For 50k inputs, candidate prompts were generated with a SFT model. Each candidate was then rendered into an image using Stable Diffusion v1.4 with the DDIM scheduler. The images were scored by our preference model (ImageReward (Xu et al., 2024b)) and assigned "chosen" or "rejected" labels, which served as the training signal for both DPO and Sem-DPO. Such use of AI models to generate preference labels is common in large reward modeling pipelines, including RLHF and RLAIF, where proxy or AI feedback is used in place of direct human annotations. Because the same preference labels are shared across all methods, the comparison remains fair and free from labeling bias. We add further experimental details in Appendix E.

5.2 Experimental Results

Performance and Semantic Consistency. The efficacy of Sem-DPO was rigorously evaluated against standard and state-of-the-art prompt optimization methods, leveraging both automated alignment metrics (CLIP Score) and human preference metrics (HPSv2.1 and PickScore). As shown in our main results (Figure 3) and detailed in the Appendix (Tables 5, 7, & 6), Sem-DPO consistently demonstrated superior performance across all three backbones. Additional validation on Qwen3-0.6B/1.7B and multiple encoders (CLIP, Jina CLIP v4, BLIP-ITM) confirms these improvements (Appendix B). Sem-DPO especially excels in enhancing human preference alignment while maintaining strong semantic consistency across diverse datasets, thereby outperforming previous

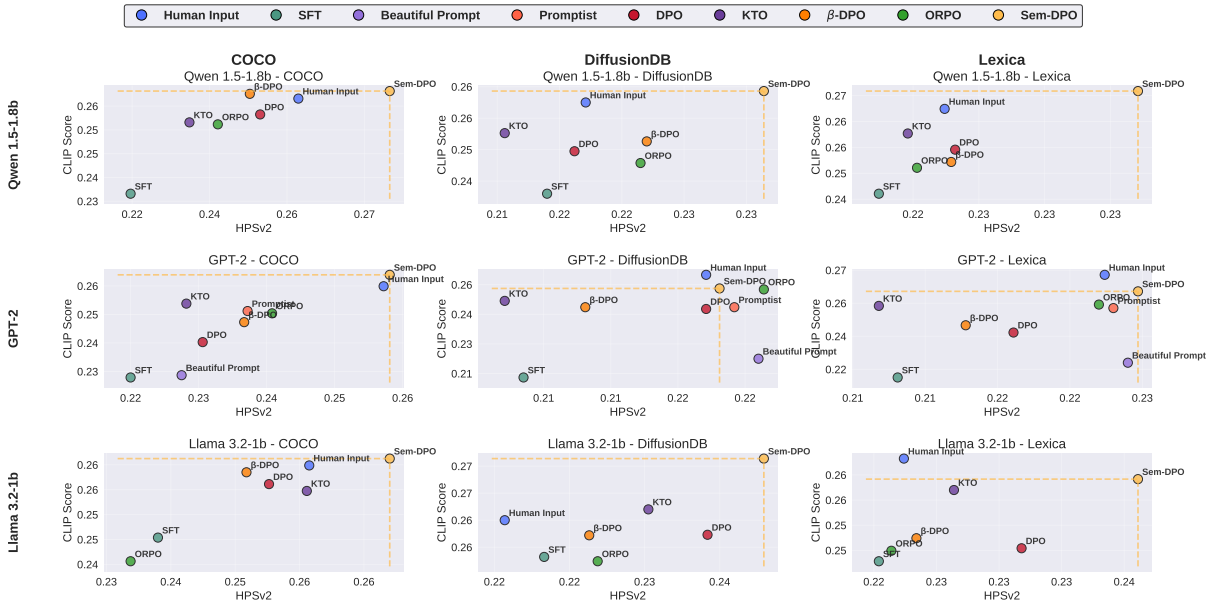


Figure 3: **Semantic–preference landscape of prompt-optimization methods across models and datasets.** Scatter plots chart CLIP Score against human preference (HPSv2.1) on three datasets with Qwen 1.5-1.8b (top), GPT-2 (middle), and Llama 3.2-1b (bottom) generators. Points closer to the upper-right indicate prompts excelling in both metrics. Figures comparing CLIP score against a second human preference metric (PickScore) are in the Appendix.

baselines of prompt optimization in text-to-image generation. As depicted in Figure 1, standard DPO improved human alignment over Original Prompt but frequently resulted in lower CLIP scores, indicative of semantic inconsistency where generated prompts diverged from the user's original intent. Sem-DPO addresses this limitation by explicitly incorporating semantic consistency into the optimization process, and does so without excess verbosity, as shown in Tables 1 and 2.

Analysis Across Datasets. A detailed analysis of the results, presented in Figure 3, underscores Sem-DPO's advancements across in-domain (DiffusionDB & Lexica) and out-of-domain (COCO) datasets. For DiffusionDB, Sem-DPO achieved a better HPSv2.1 score and PickScore compared to other baselines in most of the scenarios, while maintaining a strong CLIP Score. On the Lexica dataset, Sem-DPO exhibited the highest CLIP Score, HPSv2.1 score, and PickScore among all evaluated methods, highlighting its robustness in an in-domain setting. In out-of-domain, Sem-DPO continued to excel, yielding the highest HPSv2.1 score and a PickScore, further affirming its generalization capabilities and superior human preference alignment. Although the absolute gains in PickScore appear modest, they translate into noticeably improved visual fidelity in the generated images, also shown in prior works.

Method	N	Mean	Std	Median	Max
Human Input	600	18.10	20.49	10.0	211
SFT	600	33.64	17.82	28.5	77
DPO	600	54.61	20.98	59.0	104
Sem-DPO	600	35.10	17.70	30.0	75

Table 1: Overall token length statistics of rewritten prompts.

Dataset	Human	SFT	DPO	Sem-DPO
COCO	8.0	25.5	71.0	29.0
DiffusionDB	19.0	34.0	56.0	37.5
Lexica	13.0	28.0	55.0	29.0

Table 2: Per-dataset median token length of rewritten prompts.

Head-to-head Comparison. Figure 4 reports how often a Sem-DPO-generated prompt wins, ties, or loses when pitted one-by-one against outputs from Human Input, base DPO, and SFT under each metric. Against DPO, Sem-DPO wins on 53.5% of prompts for CLIP, 57.2% for PickScore, and 67.3% for HPS v2.1, with losses dropping to roughly one-third of the cases. The margin is even larger over SFT (e.g., 63.2% wins on CLIP and 75.0% on HPS), while gains over raw Human Input are positive but smaller, reflecting that Sem-DPO occasionally prefers stylistic tweaks that do not always boost semantic alignment. Overall, the

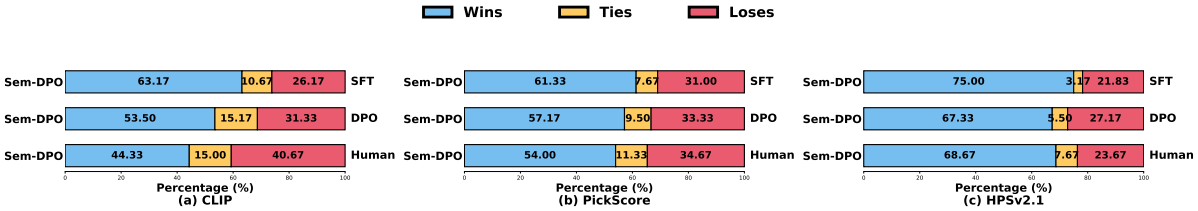


Figure 4: Head-to-head comparison of Sem-DPO against three baselines (SFT, Base DPO, and Human Input) across three evaluation metrics. Results averaged across datasets using Qwen 1.5-1.8B.

Method	CLIP	Jina v4	BLIP-ITM
Human Input	0.2599	0.7740	0.8762
DPO	0.2569	0.7359	0.7104
Sem-DPO	**0.2610**	**0.7739**	**0.8907**
Human Input	0.2600	0.6456	0.8615
DPO	0.2573	0.6293	0.7747
Sem-DPO	**0.2714**	**0.6597**	**0.9134**
Human Input	0.2679	0.6244	0.8682
DPO	0.2465	0.5983	0.7479
Sem-DPO	**0.2630**	**0.6402**	**0.9507**

Table 3: CLIP Score comparisons along with two additional embedding-based evaluators (Jina v4 and BLIP-ITM) for the **Llama 3.2-1B** model. Colors represent different datasets: COCO, DiffusionDB, and Lexica.

chart shows a clear majority-win pattern, confirming that Sem-DPO is consistently preferred to both implementation baselines and the original human prompts across all three evaluation axes.

Embedding Generalization. We test whether Sem-DPO generalizes beyond the embedding space used for training. While semantic weights are computed using a frozen CLIP encoder during optimization, we evaluate the resulting images with three independent text–image matchers: CLIP, Jina CLIP v4, and BLIP-ITM. This provides an out-of-distribution check across different embedding families on COCO, DiffusionDB, and Lexica (Llama 3.2-1B prompt optimizer). As shown in Table 3, Sem-DPO consistently achieves the best semantic alignment across all evaluators.

Hyperparameter Analysis. Varying the semantic-weighting coefficient α reveals a trade-off between semantic fidelity and preference alignment (Figure 5). For small values ($\alpha = 1$–2), the exponential weight $W_\alpha \approx 1$, and the loss behaves like standard DPO, yielding modest gains in CLIP, HPSv2.1, and PickScore. Increasing α to 4 sharpens the weighting, filtering out semantically drifting samples while retaining a large effective batch. This yields the best semantic consistency (CLIP ≈ 0.272) along with strong preference alignment. At $\alpha = 8$, further regularization

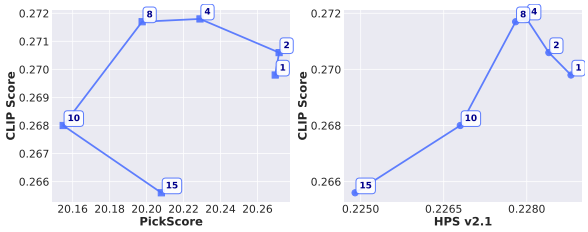


Figure 5: **Impact of semantic-weighting coefficient α on alignment metrics using Qwen 1.5-1.8B** on the Lexica dataset. Each point corresponds to a different α value (shown in labels), tracing the trade-off between CLIP Score and human-preference metrics: PickScore (left) and HPS v2.1 (right).

Method	Total Preferences
DPO	32
Sem-DPO	**148**

Table 4: Human preference counts comparing DPO vs. Sem-DPO on 20 randomly sampled prompts (180 total pairwise comparisons).

slightly reduces both CLIP (≈ 0.27) and PickScore (≈ 20.20). However, $\alpha \geq 10$ over-suppresses informative gradients, shrinking the batch and reducing all metrics ($\alpha = 15$: CLIP ≈ 0.266). Correlation tests support this: CLIP and HPSv2.1 are strongly correlated ($r = 0.84$, $p = 0.034$), while CLIP–PickScore shows weaker correlation ($r = 0.35$). In practice, $\alpha \approx 4$ is recommended for optimal performance, but this may depend on the requirements for semantic fidelity.

5.3 Human Validation

To validate that the semantic alignment improvements measured by CLIP translate to actual human perceptions, we conducted user studies with 9–57 participants across multiple rounds. For each case, participants saw the original prompt and two images generated from prompts optimized with DPO and Sem-DPO respectively, and selected which better matched the original prompt. Images were presented in randomized order, with participants

blinded to the generating method.

Table 4 shows preference counts across 20 randomly sampled test cases (9 participants). Sem-DPO was preferred in 148 out of 180 pairwise comparisons (82.2%). To further validate these results, we reanalyzed the data using statistical tests appropriate for human evaluation in NLP (Schuff et al., 2023). A per-participant signed-rank test revealed that all 9/9 participants individually preferred Sem-DPO on the majority of their judgments, with preference rates ranging from 70% to 95% (median 80%). A per-prompt test showed that Sem-DPO was preferred on 18/20 prompts (median preference rate 89%). An earlier round with 57 participants and 5 test cases yielded similar results (95.8% preference for Sem-DPO; see Appendix). These results strongly validate that prompts optimized with Sem-DPO have markedly increased semantic fidelity with respect to the original prompt. Sample test cases are documented in Appendix Table 10.

6 Discussion

How can Sem-DPO enable more reliable Compound AI Systems? Compound AI systems combine multiple models in pipelines (e.g., prompt optimizer $\rightarrow$ image generator). Fine-tuning each component with DPO is common, but standard DPO can introduce semantic drift that propagates downstream. By directly enforcing semantic consistency, Sem-DPO ensures each model remains faithful to its intended role while aligning to human preferences. This reliability makes Sem-DPO particularly valuable for compound systems, where errors in one stage can cascade across the pipeline.
Can AI-generated preference labels introduce bias into alignment training? While this concern is valid, automated preference annotation remains the only scalable way to realize RLHF- and RLAIF-style pipelines at the scale needed for modern alignment research. In our setup, all compared methods share the same proxy preferences, ensuring a fair and controlled comparison. We thus view automated preference labeling not as a compromise, but as a necessary step toward scalable alignment.

7 Conclusion

DPO maximizes human preference but leaves *semantic drift* unchecked, allowing optimized outputs to deviate from the user's intent. We introduce *Sem-DPO*, which scales the DPO loss by an exponential weight derived from the distance to the original input, down-weighting semantically mismatched pairs while preserving DPO's simplicity.

Limitations

While Sem-DPO offers significant improvements, several limitations warrant consideration for future work. The current implementation relies on a fixed, pre-trained, external frozen embedding model (e_φ) to compute semantic similarity. The choice and robustness of this embedding model can directly influence the effectiveness of the semantic weighting. Exploring adaptive or task-specific embedding functions could potentially further enhance performance. Additionally, the hyperparameter α, which controls the strength of the semantic weighting, was set to 4 after manual tuning. While this yielded strong results, the optimal value of α may vary depending on the dataset characteristics or the specific task, suggesting a need for more robust hyperparameter tuning strategies or automated methods for its determination. Finally, while Sem-DPO addresses semantic inconsistency, future research could address other potential issues in prompt optimization, such as stylistic overfitting not directly tied to semantic drift.

Ethical Considerations

Semantic-DPO's efficiency may lower the barrier for malicious or deceptive image generation, so any deployment should couple it with rigorous content-safety filters and provenance tools. Because the semantic weight inherits biases from CLIP embeddings and human-preference data, practitioners must audit and diversify these sources to avoid reinforcing stereotypes. Our semantic-drift guarantee rests on cosine distance accurately reflecting meaning, in domains where that proxy fails, human oversight is required to catch residual errors.

Acknowledgments

The work of Azal Ahmad Khan was supported in part by the Amazon Machine Learning Systems Fellowship and the UMN GAGE Fellowship. The work of Xinran Wang was supported by the 3M Science and Technology Graduate Fellowship. The work of Ali Anwar was supported by the Samsung Global Research Outreach Award and the National Science Foundation Privacy-Preserving Data Sharing in Practice (PDaSP) program under grant number 2452817.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.

Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Miaomiao Cai, Simiao Li, Wei Li, Xudong Huang, Hanting Chen, Jie Hu, and Yunhe Wang. 2025. Dspo: Direct semantic preference optimization for real-world image super-resolution. *arXiv preprint arXiv:2504.15176*.

Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. 2023. Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. *arXiv preprint arXiv:2311.06752*.

Amitava Das, Suranjana Trivedy, Danush Khanna, Rajarshi Roy, Gurpreet Singh, Basab Ghosh, Yaswanth Narsupalli, Vinija Jain, Vasu Sharma, Aishwarya Naresh Reganti, et al. 2025. Dpo kernels: A semantically-aware, kernel-enhanced, and divergence-rich paradigm for direct preference optimization. *arXiv preprint arXiv:2501.03271*.

Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Self-play with execution feedback: Improving instruction-following capabilities of large language models. *arXiv preprint arXiv:2406.13542*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2022. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

Yuu Jinnai and Ukyo Honda. 2024. Annotation-efficient preference optimization for language model alignment. *arXiv preprint arXiv:2405.13541*.

Sunghee Jung, Donghun Lee, Shinbok Lee, Gaeun Seo, Daniel Lee, Byeongil Ko, Junrae Cho, Kihyun Kim, Eunggyun Kim, and Myeongcheol Shin. 2025. Diatool-dpo: Multi-turn direct preference optimization for tool-augmented large language models. *arXiv preprint arXiv:2504.02882*.

Saeed Khaki, JinJin Li, Lan Ma, Liu Yang, and Prathap Ramachandra. 2024. Rs-dpo: A hybrid rejection sampling and direct preference optimization method for alignment of large language models. *arXiv preprint arXiv:2402.10038*.

Azal Ahmad Khan, Sayan Alam, Xinran Wang, Ahmad Faraz Khan, Debanga Raj Neog, and Ali Anwar. 2024. Mitigating sycophancy in large language models via direct preference optimization. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1664–1671. IEEE.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2024. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2024. Ai alignment through reinforcement learning from human feedback? contradictions and limitations. *arXiv preprint arXiv:2406.18346*.

Junru Lu, Siyu An, Min Zhang, Yulan He, Di Yin, and Xing Sun. 2024. Fipo: Free-form instruction-oriented prompt optimization with preference dataset and modular fine-tuning schema. *arXiv preprint arXiv:2402.11811*.

Motoki Omura, Yasuhiro Fujita, and Toshiki Kataoka. 2024. Entropy controllable direct preference optimization. *arXiv preprint arXiv:2411.07595*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. 2024. Unintentional unalignment: Likelihood displacement in direct preference optimization. *arXiv preprint arXiv:2410.08847*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

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

Amir Saeidi, Shivanshu Verma, Md Nayem Uddin, and Chitta Baral. 2024. Insights into alignment: Evaluating dpo and its variants across multiple tasks. *arXiv preprint arXiv:2404.14723*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, 29(5):1199–1222.

Zhao Shan, Chenyou Fan, Shuang Qiu, Jiyuan Shi, and Chenjia Bai. 2025. Forward kl regularized preference optimization for aligning diffusion policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14386–14395.

Shivanshu Shekhar, Shreyas Singh, and Tong Zhang. 2024. See-dpo: Self entropy enhanced direct preference optimization. *arXiv preprint arXiv:2411.04712*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Baohua Dong, Ran Lin, and Ruohui Huang. 2024. Conifer: Improving complex constrained instruction-following ability of large language models. *arXiv preprint arXiv:2404.02823*.

Xiaohang Tang, Sangwoong Yoon, Seongho Son, Huizhuo Yuan, Quanquan Gu, and Ilija Bogunovic. 2025. Game-theoretic regularized self-play alignment of large language models. *arXiv preprint arXiv:2503.00030*.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. 2023. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024. Beta-dpo: Direct preference optimization with dynamic beta. *Advances in Neural Information Processing Systems*, 37:129944–129966.

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

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G Honavar. 2024. Cal-dpo: Calibrated direct preference optimization for language model alignment. *arXiv preprint arXiv:2412.14516*.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. 2024a. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024b. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Chenxu Yang, Ruipeng Jia, Naibin Gu, Zheng Lin, Siyuan Chen, Chao Pang, Weichong Yin, Yu Sun, Hua Wu, and Weiping Wang. 2024a. Orthogonal finetuning for direct preference optimization. *arXiv preprint arXiv:2409.14836*.

Jiuding Yang, Weidong Guo, Kaitong Yang, Xiangyang Li, Yu Xu, and Di Niu. 2024b. Enhancing and assessing instruction-following with fine-grained instruction variants. *arXiv preprint arXiv:2406.11301*.

Qingyu Yin, Chak Tou Leong, Hongbo Zhang, Minjun Zhu, Hanqi Yan, Qiang Zhang, Yulan He, Wenjie Li, Jun Wang, Yue Zhang, et al. 2024. Direct preference optimization using sparse feature-level constraints. *arXiv preprint arXiv:2411.07618*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. 2024a. Wpo: Enhancing rlhf with weighted preference optimization. *arXiv preprint arXiv:2406.11827*.

Wenxuan Zhou, Shujian Zhang, Lingxiao Zhao, and Tao Meng. 2024b. T-reg: Preference optimization with token-level reward regularization. *arXiv preprint arXiv:2412.02685*.

Zixiao Zhu, Hanzhang Zhou, Zijian Feng, Tianjiao Li, Chua Jia Jim Deryl, Mak Lee Onn, Gee Wah Ng, and Kezhi Mao. 2025. Rethinking prompt optimizers: From prompt merits to optimization. *arXiv preprint arXiv:2505.09930*.

A Related Works

Preference Alignment. Early alignment research framed the problem as RLHF, where a reward model trained on pairwise human preferences guides policy optimization with PPO or variants (Bai et al., 2022; Ouyang et al., 2022). Although effective, RLHF imposes a heavy engineering overhead and a large on-policy sampling cost. DPO eliminates the explicit reward model by casting alignment as a contrastive classification problem whose optimal policy is available in closed form (Rafailov et al., 2023), matching or exceeding RLHF while being simpler and compute-efficient. DPO has demonstrated strong empirical results in dialogue and instruction tuning tasks, and its simplicity and efficiency have led to its adoption in various domains (Dong et al., 2024; Yang et al., 2024b; Sun et al., 2024; Lu et al., 2024; Khaki et al., 2024). Since then, DPO has been developed for multiple specialized axes like domain generalization (Wallace et al., 2024), contrastive refinements (Xiao et al., 2024; Omura et al., 2024), and theoretical analyses (Park et al., 2024).

Regularization. Recent work introduces explicit regularizers to the DPO loss to restrain over-confident or reward-hacking behaviours. Entropy-based penalties adjust the exploration–exploitation balance during fine-tuning (Omura et al., 2024), while calibrated objectives match the scale of ground-truth rewards (Xiao et al., 2024). Feature-level constraints restrict updates to sparse or disentangled subspaces to preserve fluency (Yin et al., 2024), and game-theoretic self-play adds adversarial robustness (Tang et al., 2025). Recent efforts edge toward semantic regularization, *DPO-Kernels* couples probability-space losses with embedding-space kernels (Das et al., 2025), *DSPO* injects instance-level semantic guidance for real-image super-resolution (Cai et al., 2025), and Hallucination-Aware DPO penalises vision-language mismatch (Zhao et al., 2023). Many existing regularization techniques are tailored to specific modalities or require additional annotation, limiting their scalability and general applicability (Shekhar et al., 2024; Jinnai and Honda, 2024). Further re-weighting schemes such as Weighted-PO adjust gradients according to preference margins alone, leaving semantic drift unaddressed (Zhou et al., 2024a). *Therefore, existing regularizers either treat semantics as a post-hoc filter or rely on ad-hoc domain heuristics and none integrate semantic consistency directly into the core preference loss or offer guarantees against prompt-level drift.*

B Additional Model Evaluation

Tables 5 & 6 give the complete metric values for every language model and dataset pair. Tables 5 and 7 (Qwen 1.5-1.8 B, Qwen 3-0.6b, Qwen 3-1.7b and Llama 3.2-1b) show Sem-DPO topping across all three metrics for all the datasets. Table 6 (GPT-2) repeats the pattern, with Sem-DPO reclaiming the lead even against specialized prompt-engineering baselines. These full numbers support the main-text claim that Sem-DPO yields state-of-the-art preference and alignment across models and domains.

Method	*Metrics*			
	CLIP Score	**HPS v2.1**	**PickScore**	**Avg. Score**
Human Input	0.2599	0.2572	21.4127	0.786
SFT	0.2359	0.2247	20.5758	0.000
DPO	0.2559	0.2498	21.0490	0.572
KTO	0.2539	0.2361	20.8834	0.403
β-DPO	0.2611	0.2478	21.1933	0.671
ORPO	0.2534	0.2416	20.9463	0.453
Sem-DPO	**0.2618**	**0.2749**	**21.6433**	**1.000**
Human Input	0.2600	0.2171	19.7258	0.595
SFT	0.2368	0.2140	19.5800	0.112
DPO	0.2476	0.2162	19.6595	0.368
KTO	0.2522	0.2106	19.5243	0.197
β-DPO	0.2501	0.222	19.7626	0.599
ORPO	0.2446	0.2215	19.6836	0.439
Sem-DPO	**0.2629**	**0.2314**	**19.8464**	**1.000**
Human Input	0.2679	0.2224	20.0578	0.424
SFT	0.2417	0.2174	20.0973	0.111
DPO	0.2553	0.2232	20.1131	0.371
KTO	0.2603	0.2196	20.0036	0.233
β-DPO	0.2515	0.2229	20.1429	0.361
ORPO	0.2497	0.2203	20.1226	0.274
Sem-DPO	**0.2734**	**0.2371**	**20.2857**	**1.000**

Table 5: Comparison of prompt optimization methods and variants of DPO with the **qwen 1.5-1.8b** model in Text-to-Image Generation across different metrics. Colors represent different datasets: COCO , DiffusionDB , and Lexica . The average score is calculated with all scores normalized into the range [0,1]. The highest-performing optimization is bolded.

Method	*Metrics*			
	CLIP Score	**HPS v2.1**	**PickScore**	**Avg. Score**
Human Input	0.2599	0.2572	21.4127	0.950
SFT	0.2279	0.2200	20.3911	0.000
Beautiful Prompt	0.2287	0.2275	20.4893	0.105
Promptist	0.2512	0.2372	21.0774	0.587
DPO	0.2403	0.2306	20.5972	0.274
KTO	0.2538	0.2282	20.6063	0.381
β-DPO	0.2473	0.2367	20.9107	0.493
ORPO	0.2504	0.2408	20.9403	0.567
Sem-DPO	**0.2639**	**0.2581**	**21.4289**	**1.000**
Human Input	**0.2600**	0.2171	19.7258	0.867
SFT	0.2080	0.2035	19.2212	0.024
Beautiful Prompt	0.2175	0.2210	19.6115	0.601
Promptist	0.2436	0.2192	**19.8338**	0.862
DPO	0.2427	0.2171	19.6598	0.725
KTO	0.2468	0.2021	19.4863	0.398
β-DPO	0.2436	0.2081	19.6539	0.572
ORPO	0.2526	**0.2214**	19.7766	**0.928**
Sem-DPO	0.2531	0.2181	19.7519	0.860
Human Input	**0.2679**	0.2224	20.0578	0.844
SFT	0.2213	0.2081	19.7741	0.024
Beautiful Prompt	0.2280	0.2240	20.0134	0.554
Promptist	0.2528	0.2230	**20.2028**	0.860
DPO	0.2417	0.2161	20.0635	0.544
KTO	0.2538	0.2068	19.9086	0.337
β-DPO	0.2450	0.2128	20.0942	0.530
ORPO	0.2544	0.2220	20.1853	0.84
Sem-DPO	0.2604	**0.2247**	20.1341	**0.893**

Table 6: Comparison of prompt optimization methods and variants of DPO with the **GPT-2** model in Text-to-Image Generation across different metrics. Colors represent different datasets: COCO , DiffusionDB , and Lexica . The average score is calculated with all scores normalized into the range [0,1]. The highest-performing optimization is bolded.

Method	*Metrics*			
	CLIP Score	**HPS v2.1**	**PickScore**	**Avg. Score**
Human Input	0.2599	0.2572	**21.4127**	0.874
SFT	0.2483	0.2384	20.6495	0.112
DPO	0.2569	0.2522	20.9886	0.577
KTO	0.2558	0.2569	21.2431	0.714
β-DPO	0.2588	0.2494	20.8701	0.534
ORPO	0.2445	0.235	20.7622	0.049
Sem-DPO	**0.261**	**0.2672**	21.4047	**0.997**
Human Input	0.2600	0.2171	19.7258	0.278
SFT	0.2532	0.2213	19.546	0.086
DPO	0.2573	0.2387	19.7881	0.560
KTO	0.2620	0.2324	19.666	0.469
β-DPO	0.2572	0.2261	19.5188	0.193
ORPO	0.2524	0.227	19.6828	0.249
Sem-DPO	**0.2714**	**0.2447**	**19.9409**	**1.000**
Human Input	**0.2679**	0.2224	20.0578	0.503
SFT	0.2434	0.2204	19.942	0.074
DPO	0.2465	0.2318	19.969	0.315
KTO	0.2604	0.2264	19.9888	0.428
β-DPO	0.2489	0.2234	19.8077	0.123
ORPO	0.2459	0.2214	20.0824	0.201
Sem-DPO	0.2630	**0.2411**	**20.4131**	**0.933**

Table 7: Comparison of prompt optimization methods and variants of DPO with the **llama 3.2-1B** model in Text-to-Image Generation across different metrics. Colors represent different datasets: COCO, DiffusionDB, and Lexica. The average score is calculated with all scores normalized into the range [0,1]. The highest-performing optimization is bolded.

Method	*Metrics*			
	CLIP Score	**HPS**	**PickScore**	**Avg. Score**
Human Input	0.2626	0.2322	20.3987	0.8094
SFT	0.2284	0.2149	19.7858	0.000
DPO	0.2602	0.2340	20.3285	0.7782
Sem-DPO (Ours)	**0.2688**	**0.2404**	**20.4653**	**1.000**

Table 8: Comparison of CLIP Score, HPS, and PickScore on **Qwen3-0.6B** used as a prompt-optimization model. Sem-DPO achieves the highest overall average score (calculated with scores normalized into [0, 1]), demonstrating stronger alignment with human preferences while preserving semantic relevance. Results are averaged across the three benchmarks (DiffusionDB, Lexica, and COCO).

Method	*Metrics*			
	CLIP Score	**HPS**	**PickScore**	**Avg. Score**
Human Input	0.2626	0.2322	20.3987	0.9027
SFT	0.2124	0.2086	19.5343	0.000
DPO	0.2563	0.2300	20.1695	0.7512
Sem-DPO (Ours)	**0.2686**	**0.2376**	**20.3975**	**0.9995**

Table 9: Comparison of CLIP Score, HPS, and PickScore on **Qwen3-1.7B** used as a prompt-optimization model. Sem-DPO achieves the highest overall average score (calculated with scores normalized into [0, 1]), demonstrating stronger alignment with human preferences while preserving semantic relevance. Results are averaged across the three benchmarks (DiffusionDB, Lexica, and COCO).

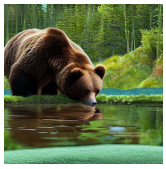
DPO	Sem-DPO
Prompt: cybernetic grey werewolf with power armor.	
	
Prompt: brown bear drinking water from a tap.	
	
Prompt: a lone gnarled tree clinging to a scree slope.	
	
Prompt: bronze fossilized cicada moth in a large cage made of orange nylon wire and beeswax.	
	
Prompt: Cinematic beautiful jungle tree house.	
	

Table 10: Side-by-side generations from the same original prompt, comparing DPO vs. Sem-DPO.

Table 10 illustrates qualitative differences in image generations using DPO optimized prompts versus Sem-DPO optimized prompts, showcasing a substantial increase in semantic alignment with the original prompt.

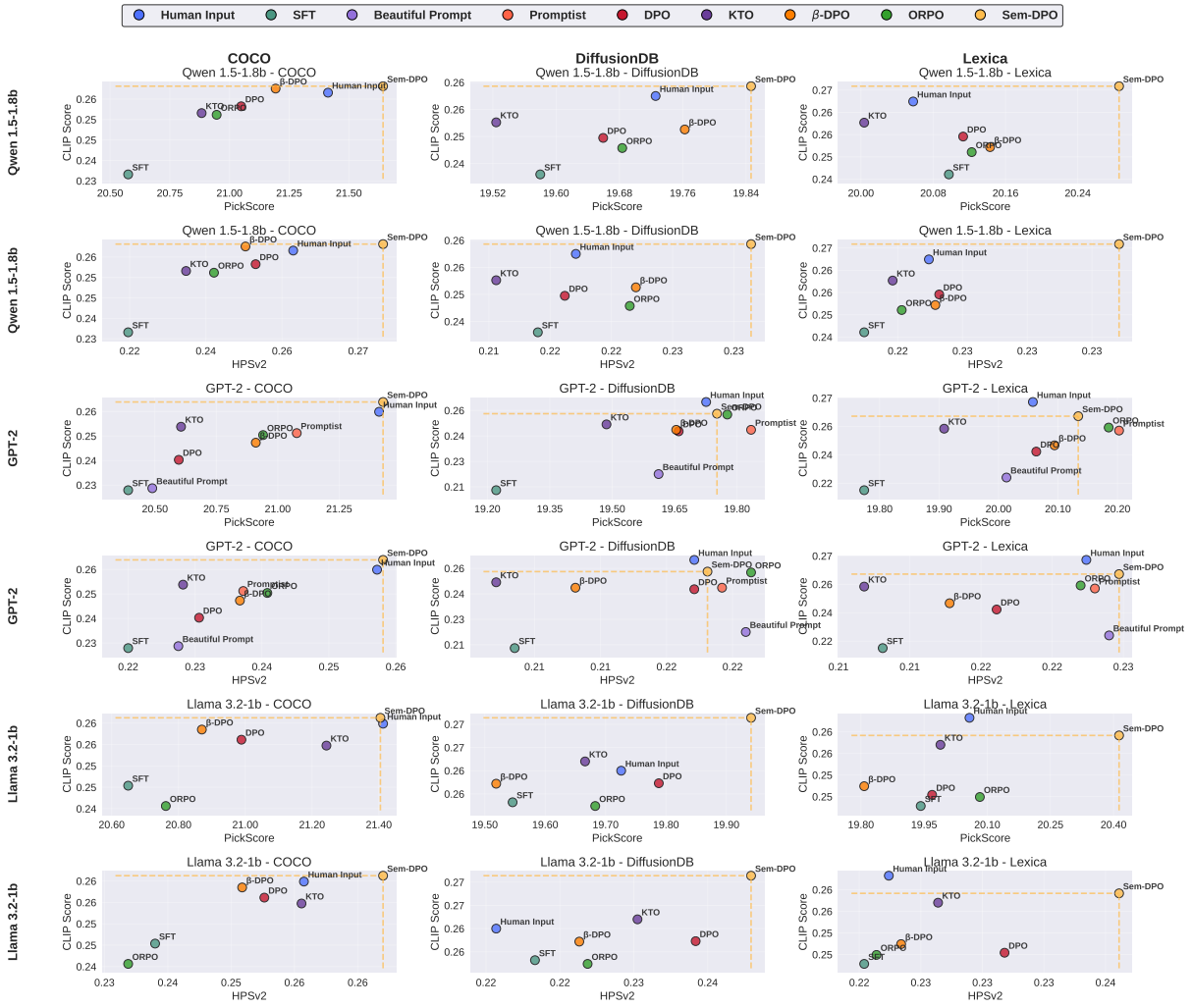


Figure 6: **Semantic–preference landscape of prompt-optimization methods across models and datasets.** Scatter plots chart CLIP Score against human preference (HPSv2.1 and PickScore) on three datasets with Qwen 1.5-1.8b (top), GPT-2 (middle), and Llama 3.2-1b (bottom) generators. Points closer to the upper-right indicate prompts excelling in both metrics. Sem-DPO generally shows a stronger performance across models and datasets.

C Human Validation

A previous round of human validation was also conducted, with 57 participants and 5 test cases. The results are shown in table 11, with a 95.8% preference for Sem-DPO across data points.

D Sem-DPO vs. Additive Instruction-Following Rewards

A natural alternative to Sem-DPO is to add an explicit instruction-following (text-fidelity) reward to the optimization objective similar to VisionReward (Xu et al., 2024a). Our work is fundamentally different in *where* it intervenes in the text-to-image pipeline and *which* source of semantic drift it targets. In our setting, generation proceeds as a pipeline $x \to y \to I(y)$, where a prompt optimizer maps the user instruction x to an optimized prompt y, and a fixed text-to-image generator produces an image $I(y)$. Under the metric formulation used in Proposition 2, the end-to-end text–image drift can be decomposed

Method	Total Preferences
DPO	12
Sem-DPO	**273**

Table 11: Human preference counts across 5 test prompts (n=57 participants per prompt). Participants selected which image better matched the original prompt.

Dataset	Method	CLIP Score	HPSv2	PickScore
COCO	Sem-DPO (Ours)	0.2618	0.2749	21.6433
	Additive Reward (VisionReward)	0.2484 (-5.1%)	0.2482 (-9.7%)	21.0414 (-2.8%)
DiffusionDB	Sem-DPO (Ours)	0.2629	0.2314	19.8464
	Additive Reward (VisionReward)	0.2481 (-5.6%)	0.2264 (-2.2%)	19.8764 ($+0.2\%$)
Lexica	Sem-DPO (Ours)	0.2734	0.2371	20.2857
	Additive Reward (VisionReward)	0.2456 (-10.2%)	0.2264 (-4.5%)	20.2063 (-0.4%)

Table 12: Comparison between Sem-DPO and an additive instruction-following baseline that adds VisionReward's text-fidelity score as an extra reward term. Percentages (in parentheses) report the relative change of the additive baseline compared to Sem-DPO for each metric.

into (i) a *prompt-level* drift term measuring how far y moves away from x in embedding space, and (ii) a *generator-level* consistency term capturing how faithfully the generator follows a given prompt. Reward-model-based instruction-following signals (including text-fidelity dimensions) primarily operate at the *image level* by encouraging $I(y)$ to match y (and to satisfy other axes such as aesthetics/realism). In contrast, Sem-DPO explicitly controls the *prompt-level* drift by reweighting preference pairs according to the semantic distance between the original instruction x and the preferred optimized prompt y, thereby discouraging updates that improve reward at the cost of drifting away from the user intent.

To empirically test whether Sem-DPO reduces to simply adding an instruction-following reward, we construct an *Additive Reward (VisionReward-based)* baseline that treats VisionReward's text-fidelity score as an additional instruction-following reward component. Concretely, we keep the same preference signal as our main setup (e.g., ImageReward) and add VisionReward's text-fidelity score as an extra scalar term (with a single normalization coefficient) inside the training objective, while keeping the data, backbone (Qwen-1.5-1.8B), and the underlying text-to-image generator fixed. Table 12 compares this baseline against Sem-DPO on three benchmarks using CLIP Score, HPSv2, and PickScore. Across all datasets, Sem-DPO yields stronger semantic alignment (CLIP) and comparable or better preference metrics (HPSv2/PickScore), indicating that semantics-aware weighting provides a more favorable preference–semantics trade-off than simply adding a text-fidelity reward term.

These results also clarify the scope of our contribution relative to VisionReward-style methods. VisionReward's primary novelty lies in reward-model design (a multi-dimensional visual reward) and its use for aligning image/video generators at the image level. Our contribution is complementary: we assume a fixed reward model is already available and show that DPO-style prompt optimization can still induce prompt-level semantic drift even under strong rewards that include instruction-following components. Sem-DPO addresses this failure mode by integrating semantic information as *pairwise importance weights* in DPO, behaving like a soft constraint on drift: heavily drifted prompts are strongly downweighted regardless of how rewarding they may be on other axes.

E Training Details

Training and Experimental Details. Completing this end-to-end workflow required approximately 26 hours on a single NVIDIA RTX 4090 GPU. Model fine-tuning (both DPO and Sem-DPO) was carried out on a single NVIDIA A100 GPU, with the DPO baseline requiring about 1 hour and Sem-DPO only ten minutes longer.

Training Configuration. All models were trained with a batch size of 8, learning rate of 5e-6, and fine-tuned for one epoch. Training was performed on a single NVIDIA A100 GPU, requiring approximately 1 hour for DPO and 70 minutes for Sem-DPO. The semantic weighting parameter α was set to 4 based on

validation performance (see Section 5.2 for ablation analysis).

Notation. We use β for the DPO temperature parameter, $Z(x)$ for the partition function, $\sigma(\cdot)$ for the logistic sigmoid, $\ell(\cdot)$ for the loss term, $e_\phi(\cdot)$ for the embedding function, and $I(\cdot)$ for the text-to-image generator.

Experimental Details. Our experiments utilize the Stable Diffusion v1.4 text-to-image model. To expedite the image sampling process, we employ the Denoising Diffusion Implicit Models scheduler(Song et al., 2020). The number of denoising steps was set to 20 for all image generation tasks. For both DPO and Sem-DPO, the β parameter was set to 0.05. For Sem-DPO, the semantic weighting parameter α was set to 4 after manual tuning (see Appendix for more details).