This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DyMO: Training-Free Diffusion Model Alignment with Dynamic Multi-Objective Scheduling

Xin Xie, Dong Gong*
University of New South Wales (UNSW Sydney)
{xin.xie3, dong.gong}@unsw.edu.au

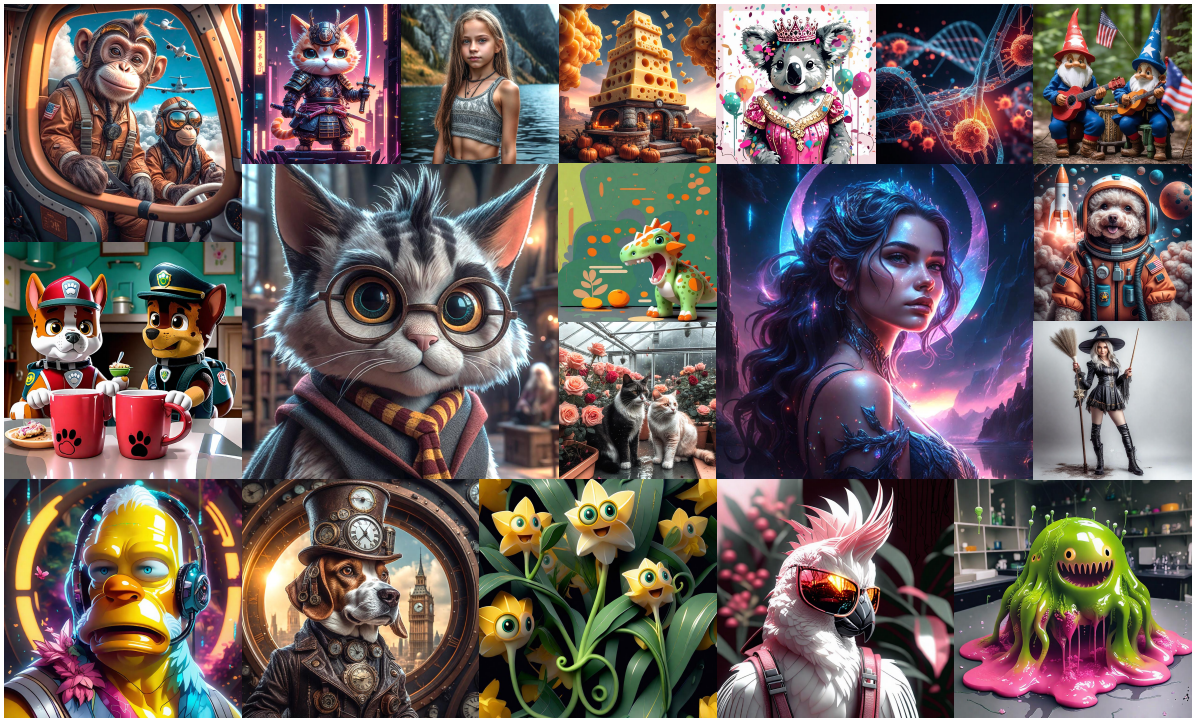


Figure 1. Sample images generated by our method based on SDXL backbones. Under the guidance of our method, the generated images not only achieve a high alignment with text prompt and human preferences, but also exhibit visually attractive and stunning aesthetics.

Abstract

Text-to-image diffusion model alignment is critical for improving the alignment between the generated images and human preferences. While training-based methods are constrained by high computational costs and dataset requirements, training-free alignment methods remain underexplored and are often limited by inaccurate guidance. We propose a plug-and-play training-free alignment method, DyMO, for aligning the generated images and human preferences during inference. Apart from text-aware human preference scores, we introduce a semantic alignment objective for enhancing the semantic alignment in the early stages of diffusion, relying on the fact that the attention maps are effective reflections of the semantics in noisy images. We propose dynamic scheduling of multiple objectives and intermediate recurrent steps to reflect the requirements at different steps. Experiments with diverse pre-trained diffusion models and metrics demonstrate the effectiveness and robustness of the proposed method. The project page: dymo.github.io.

1. Introduction

Text-to-image (T2I) diffusion models [3, 19, 32, 33, 39] have demonstrated state-of-the-art effectiveness in image generation, transforming textual prompts into contextually rich visuals. Despite achieving impressive image quality and fidelity, the generative data distribution trained directly

*D. Gong is the corresponding author. This project was partially supported by an ARC DECRA Fellowship (DE230101591) to D. Gong.

from diverse data (*e.g.*, web images) often diverges from user preferences (*e.g.*, aesthetically pleasing images and images aligning the text instructions). To enhance usability, diffusion model *alignment* [25] is investigated to enhance the consistency between the generated images and human user preferences, including the visual preference and intention in the text prompts, through *training-based* [13, 15, 23, 24, 45, 50, 51] or *training-free* approaches [7, 11, 41, 44].

T2I diffusion model alignment can be achieved through training-based methods [10, 29, 32, 34], via direct two-stage fine-tuning on customized datasets [20, 47] that better represent user preferences with high-quality image-text pairs or preference information. Similarly, relying on text-aware human preference scoring models trained on the preference datasets [20, 47], some methods [6, 30] directly tune the model to increase the differentiable score. Rather than straightforward fine-tuning based on static and pre-defined datasets or scores, more advanced and adaptive approaches [4, 12] are proposed by incorporating Reinforcement Learning from Human Feedback (RLHF), which allows the diffusion models to learn and refine outputs through reward signals iteratively [20, 35, 47, 49]. To relieve the complexity of learning a reward model, Direct Preference Optimization (DPO) [31] that implicitly estimates a reward model is applied to diffusion model alignment [13, 24, 45, 50]. Despite successful applications, it is challenging and resource-intensive to learn a universal model fulfilling diverse preferences or requirements.

Training-free methods align generated images with specific objectives by applying differentiable rewards to adjust the generation process of a pre-trained diffusion model during inference [41, 52, 54, 55]. To enhance text-image alignment, classifier guidance [8, 39] uses the gradient of the pre-trained image classifier. The idea is extended to alignment with different types of conditional signals [46, 52, 54], and some training-free methods [7, 41, 44] also consider aesthetic improvement through optimizing related objectives. Unlike training-based methods, most training-free methods are mainly restricted to the alignment of specific conditions (*e.g.*, class). To achieve more general applications, in this paper, we consider the *training-free alignment of general preference* in inference. The learned text-aware human preference scores [20, 24, 47] measure both text-image alignment and human visual preferences. We can perform training-free alignment with the differentiable scores in score-based diffusion models (SBDMs) [38, 39]. Considering that the preference scores perform on clean images by default, specific designs are required to incorporate the guidance to the denoising process in diffusion models. It needs retraining noise-aware or step-dependent score estimator [2, 18, 26, 28] or full-chain backpropagation from the output to intermediate steps [7, 41, 44], which is data or time demanding or time expensive, respectively. For memory and time efficiency, we use one-step clean image approximation from noisy image [36, 52, 54], causing underestimated guidance. Especially, compared to the visual characteristics, the semantic contents (*e.g.*, entities and layout) tend to be determined in early noisy steps. Although the text-aware preference scores can reflect the text-image semantic alignment, it is more challenging to guide the contents due to the blurred predicted image from noisy samples.

To address these limitations above, we propose a training-free diffusion model alignment approach relying **Dy**namic scheduling of **M**ultiple **O**bjectives (DyMO). We aim to align the generated images with the user preferences in terms of both the intention in the text input and the appealing visual quality. Relying on the pre-trained differentiable text-aware preference scores as an *alignment objective*, we guide the denoising process with the gradient computed for the intermediate noisy images in an SBDM formulation. Since the one-step approximation is used for efficiency and the contents are not sufficiently reflected in early-stage noisy images, the preference model cannot provide effective and accurate guidance. We thus propose a semantic *alignment objective* relying on a discovery that the text-image attention maps are an indirect reflection of the semantic contents (*e.g.*, the entities and layout). The semantic alignment objective/guidance is used to minimize the discrepancies between the attention-map-based contents and the semantic meanings extracted from the text based on a large-language model (LLM). Considering that the two objectives perform with different importance for different stages, we dynamically schedule two objectives to generate detailed content while keeping the layout. Additionally, we propose a dynamic recurrent strategy to improve the guidance by automatically deciding the number of iterations at different stages.

The main contributions are summarized as:

- We propose a plug-and-play training-free diffusion model alignment method, DyMO, that can effectively align the generated images and human preferences in terms of both user intentions in text and appealing visual quality.
- Apart from text-aware human preference scores, we introduce a semantic alignment objective to mitigate the ineffectiveness of the guidance in the early stages of diffusion. We propose dynamic scheduling of multiple objectives and intermediate recurrent steps to reflect the requirements at different steps.
- We conduct validation of the proposed DyMO with diverse pre-trained diffusion models, *e.g.*, SD V1.5, SDXL, etc. DyMO outperforms different pre-trained baseline models and other state-of-the-art training-based and training-free methods significantly on different metrics, demonstrating effectiveness and superiority.

2. Related Work

2.1. Diffusion Model Fine-Tuning

Recently, diffusion models have rapidly advanced in image generation, with various methods emerging to improve quality and align with human preference. One simple yet effective way [6, 9, 30, 48, 49] is to use differentiable reward functions as the objective, enabling model optimization through gradient backpropagation. To tackle the limitations of non-differentiable reward functions and unstable training dynamics, some RL-based methods [4, 12, 22] are proposed for better image-text alignment. Following the success of Direct Preference Optimization (DPO) [31] by eliminating the need for explicit reward models, D3PO [50] fine-tuned the model on the preferred and dispreferred image pairs based on human evaluators. Diffusion-DPO [45] reformulated DPO for model optimization on the preference dataset[20]. SPO [24] further generalized DPO paradigm at each denoising step to ensure accurate step-aware preference alignment. DenseReward [51] introduced a temporal discounting factor to prioritize early steps for better preference alignment. NCPPO [13] emphasized preference optimization in perceptual feature space with more information. Diffusion-RPO [15] made contrastive weights for image pairs and Diffusion-KTO [23] defined the alignment objective as human utility maximization. Despite notable progress in diffusion model alignment, discrepancies persist between model performance and human preferences, along with large computational demands.

2.2. Training-Free Guidance

Training-free methods can be likened to "navigating uncharted waters with a compass", developing various interesting technologies to advance image fidelity. Prafulla *et al.* [8] firstly proposed classifier-based guidance to enable conditional image synthesis. Any loss functions can be the classifer, numerous approaches [2, 44, 46, 52, 54] are proposed to align with different conditional signals by setting related objectives. For example, previous works control the contents by relying on the manipulation on the noise map or attention maps, which are restricted to pre-defined and additionally given layout [40, 54] or oversimplified the semantics [55], limiting them to simple text prompts and restricted usage cases. Additionally, there are some methods [11, 16, 41] designed to optimize injected noise vectors, increasing the differentiable reward score. Similarly, Deckers *et al.* [7] conducted prompt embedding manipulation to better align with users' intentions. Besides, recent works have paid more attention on guidance accuracy and efficiency. Shen *et al.* [36] utilized random augmentation [56] to alleviate the adversarial gradient and Polyak step size to accelerate the convergence. Dreamguider [27] eliminated compute-heavy backpropagation through the diffusion network by regulating the gradient flow via a time-varying factor. These studies show impressive image generation without any training but are limited to the computational cost and inaccurate guidance, leading to suboptimal outcomes.

3. Preliminaries

3.1. Diffusion Models

Diffusion models [19, 32] are composed of forward and reverse processes. Given a clean data distribution $p(\mathbf{z}_0)$, the forward process incrementally transforms the data $\mathbf{z}_0$ into Gaussian noise over the time interval from 0 to T:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}_t, \tag{1}$$

where $\mathbf{z}_t$ is noisy data, $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$, β_t is a increasing linearly noise scheduling function of t, $\sigma_t = \sqrt{1-\bar{\alpha}_t}$ and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$ represents random noise. Conversely, the reverse process reconstructs the original data by denoising manipulation from time T back to 0. Diffusion models adopt a neural network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ to predict the noise at each step:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z}_t, \boldsymbol{\epsilon}_t, t}\left[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, t) - \boldsymbol{\epsilon}_t\|_2^2\right] \\ &= \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z}_t, \boldsymbol{\epsilon}_t, t}\left[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, t) + \sigma_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t)\|_2^2\right], \end{aligned} \tag{2}$$

where $p(\mathbf{z}_t) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{z}_0, \sigma_t^2\mathbf{I})$ is the distribution of z_t. The reverse process can be expressed as a score-based ODE in the variance-preserving setting [39], where the score estimator $s(\mathbf{z}_t, t) \approx \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t)$:

$$\mathbf{z}_{t-1} = (1 + \frac{1}{2}\beta_t)\mathbf{z}_t + \beta_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t) + \sqrt{\beta_t}\,\boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is randomly sampled Gaussian noise.

3.2. Diffusion Guidance

For conditional diffusion models, the target is to generate the data that meets the specified condition $\mathbf{y}$. The reverse process can be reformulated as:

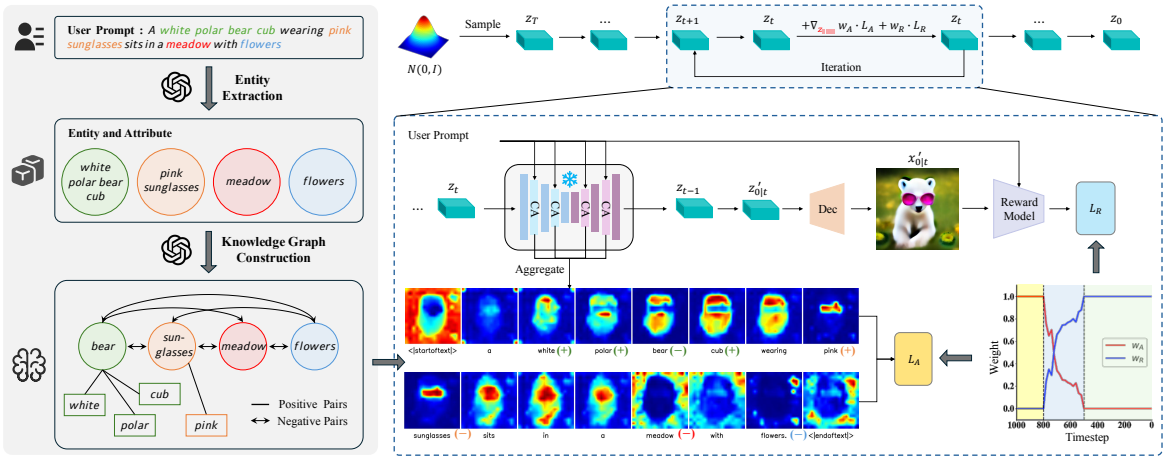
$$\mathbf{z}_{t-1} = (1 + \frac{1}{2}\beta_t)\mathbf{z}_t + \beta_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{y}) + \sqrt{\beta_t}\,\boldsymbol{\epsilon}, \tag{4}$$

where $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{y})$ is the conditional score function. Many recent works, including DPS[5], ΠGDM[37], FreeDoM [54], UGD [2], decompose a conditional score function into the unconditional score function and the loss-based term:

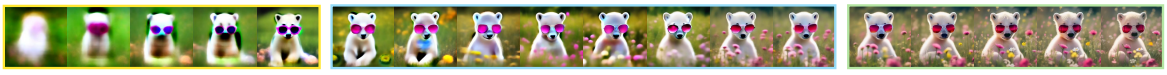
$$\begin{aligned} \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{y}) &= \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t) + \nabla_{\mathbf{z}_t} \log p(\mathbf{y}|\mathbf{z}_t) \\ &= \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t) - \lambda \nabla_{\mathbf{z}_t} \mathcal{L}_t(\mathbf{z}_t, \mathbf{y}), \end{aligned} \tag{5}$$

where λ is a scalar coefficient. With Eq. (4) and Eq. (5), the conditional sampling can be represented as:

$$\mathbf{z}_{t-1} = \mathbf{m}_t - \eta_t \nabla_{\mathbf{z}_t} \mathcal{L}_t(\mathbf{z}_t, \mathbf{y}), \tag{6}$$



(a) Our pipeline, using the example "*A white polar bear cub wearing pink sunglasses sits in a meadow with flowers*"



(b) Demonstration of the importance of different sampling stages

Figure 2. The framework of our method. (a) Given a user prompt, we use the LLMs to identify the entities and corresponding attributes for knowledge graph construction. Then we design a semantic alignment objective via cross attention map alignment based on graph, cooperating with a pre-trained preference model to dynamically guide the denoising process for high-quality image generation. (b) The entire denoising process of one-step predicted clean images under the guidance of our method.

where $\mathbf{m}_t = (1 + \frac{1}{2}\beta_t)\mathbf{z}_t + \beta_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t) + \sqrt{\beta_t}\,\boldsymbol{\epsilon}$ and η_t is a scaling factor that determines the strength of guidance. In practical applications, the gradient of the last term in Eq. (6) is obtained via backpropagation through both the guidance network and the diffusion backbone, accommodating various loss functions.

4. Methodology

Overview. We propose a training-free framework to align generated contents with contextual semantics and human preferences. We utilize a preference loss $\mathcal{L}_R(\mathbf{x}'_{0|t}, \mathbf{c}, t)$ to estimate the text-aware human preference score [20, 24, 49] of denoised image $\mathbf{x}'_{0|t}$ generated by decoding the one-step predicted latent feature $\mathbf{z}'_{0|t}$ from $\mathbf{z}_t$. We compute the gradient of the reward score to guide the denoising process like Eq. (6). Estimating $\mathcal{L}_R$ and propagating its guidance to early steps is ineffective, where the images are highly noisy or the predicted outputs are highly blurred. To guide image content in the early steps, we leverage text-image attention maps to capture semantic information in the noisy images, as the content is shaped by the attention. We propose a semantic align objective (Sec. 4.1), through minimizing the discrepancies between image content semantics (as reflected in attention maps) and text semantics extracted via LLMs. We design the tailored guidance via dynamically scheduling two objectives (Sec. 4.2), generating detailed content while preserving the layout. Besides, we propose a dynamic time-travel strategy to offer better guidance for diffusion models (Sec. 4.3). The whole pipeline is demonstrated in Fig. 2 and Algorithm 1.

4.1. Semantic Alignment Guidance

Attention map exploration. During the initial chaotic stage, the basic image structure takes shape quietly. However, the sample z_t exhibits a high level of noise, causing that the approximately predicted image $x'_{0|t}$ is extremely blurred. Consequently, training-free guidance is hard to make anything reasonable, providing ineffective semantic supervision. Previous works [17, 43, 55] show attention maps reflect the semantic contents, $M = QK^T/\sqrt{d}$, where d is the feature dimension, queries Q and keys K are the linear projection of the intermediate image feature and text embedding of prompt $\mathbf{c}$, respectively. In T2I Diffusion models, each text token u is injected into the image via cross-attention with an attention map M_u. We use M_u to supervise early-stage generation by aligning image semantics with the text prompt.

Semantic capture from text prompts. Upon input of a complex scene prompt $\mathbf{c}$ from the user, we leverage LLMs (*e.g.*, GPT-4 [1] and Llama [42]) to analyze the words likely to present in the final image and explicitly extract the entities and their corresponding attributes: $\{E_i\}_{i=1}^n, \{A_i\}_{i=1}^n$, where n is the number of the entities identified in the given prompt using LLM and E_i represents i-th entity, A_i denotes the attribute *set* associated with entity node E_i, encompassing features like color, shape, texture, etc. Each entity pos-

sesses a unique set of attributes in varying quantities, which are determined dynamically by LLM-driven heuristics.

Text semantic graph construction. In order to achieve semantic alignment, we construct a semantic graph $\mathbf{G} = (\mathbf{N}, \mathbf{S})$ to find the internal relationships within the prompts for manipulating attention maps, where $\mathbf{N} = \{E_i\}_{i=0}^{n} \cup \{\tilde{A}_i\}_{i=0}^{n}$ is the set of nodes in the graph $\mathbf{G}$, where $\tilde{A}_i$ denotes the attribute elements in A_i. For the set of edges $\mathbf{S}$, we propose two kinds of criteria to determine all possible relationships. We hypothesize that each entity exists independently within the image, occupying its own distinct position and space. The relationships between any two entities are *negative*. The relationships between the entities and their corresponding attributes are *positive*, for correct binding between entities and attributes:

$$\mathbf{S}_{\text{pos}} = \{(E_i, A_{ij}) \mid \forall i = 1, \cdots, n, \forall j = 1, \cdots, |A_i|\}, \tag{7}$$

$$\mathbf{S}_{\text{neg}} = \{(E_i, E_m) | i \neq m, \forall i, m = 1, \cdots, n\}, \tag{8}$$

where $\mathbf{S}_{\text{pos}}$ and $\mathbf{S}_{\text{neg}}$ represent positive and negative pair sets, respectively. A_{ij} denotes the j-th attirbute of the i-th entity, with a slight abuse of notations. The set of edges is obtained by $\mathbf{S} = \mathbf{S}_{\text{pos}} \cup \mathbf{S}_{\text{neg}}$.

Semantic alignment objective. Each entity or attribute corresponds to an attention map, denoted by M_u. We design a semantic alignment loss to align the image contents (M_us' relationships) with the text semantic graph represented by ($\mathbf{S}_{\text{pos}}$ and $\mathbf{S}_{\text{neg}}$), aiming to achieve better layout composition and precise attribute binding in early steps:

$$\begin{aligned}\mathcal{L}_A = &- \frac{1}{|\mathbf{S}_{\text{pos}}|} \sum\nolimits_{(s,l)\in\mathbf{S}_{\text{pos}}} f(M_s, M_l) \\ &+ \frac{1}{|\mathbf{S}_{\text{neg}}|} \sum\nolimits_{(s,l)\in\mathbf{S}_{\text{neg}}} f(M_s, M_l),\end{aligned} \tag{9}$$

where (s, l) denotes the edges in $\mathbf{S}$ and $f(\cdot)$ is the cosine similarity function. If more than one tokens are generated from a single object (*e.g.* u_1, u_2), this case would be detected, and the attention maps M_{u_1}, M_{u_2} would share the correspondences to the same set of attributes. The proposed semantic alignment loss mitigates the catastrophic object neglect problem by isolating each entity, which prevents the overlap among different objects and helps reduce semantic information loss. Meanwhile, it enhances accurate attribute association by encouraging higher alignment of cross attention maps between objects and related attributes.

4.2. Multi-Objective Dynamic Scheduling

In this section, we dynamically guide the entire denoising process and design tailored guidance based on the characteristics of diffusion models to gradually synthesize the high-quality image from T to 0.

Preference alignment. After the overall layout of the generated image is roughly established, we incorporate human feedback to guide conditional image generation, aligning the output with human preference in intermediate generation steps. Specifically, we adopt a pre-trained preference model [20, 24] as the reward function $\mathcal{L}_R(\mathbf{x}'_{0|t}, \mathbf{c}, t) = \exp(\tau \cdot f_{\text{V}}(\mathbf{x}'_{0|t}, t)^T f_{\text{T}}(\mathbf{c}))$ to update the $\mathbf{z}_{t-1}$ as Eq. (6), where τ is the temperature and $f_{\text{V/T}}$ represents vision/text encoder of a CLIP-style model trained on preference data. However, directly applying the preference loss can unintentionally alter the layout because the predicted image $\mathbf{x}'_{0|t}$ remains blurred, which leads to inaccurate feedback for the preference model. To address this issue, we propose a solution by adaptively combining the semantic alignment loss and preference alignment loss:

$$\begin{aligned} w &= 1 - e^{-k(\|\mathbf{z}'_{0|t} - \mathbf{z}'_{0|t+1}\| / \|\mathbf{z}'_{0|t+1}\|)}, \\ \mathcal{L} &= w_A \cdot \mathcal{L}_A(M) + w_R \cdot \mathcal{L}_R(\mathbf{x}'_{0|t}, \mathbf{c}, t), \end{aligned} \tag{10}$$

where $w_A = w$ and $w_R = 1 - w$ are dynamical weights and $\mathbf{z}'_{0|t}$ denotes the predicted latent features of $\mathbf{x}'_{0|t}$. As the image content becomes increasingly stable, the value of weight w decreases and the effect of human feedback increases. The adaptive mechanism for balancing these two losses maintains the crucial image structure while improving the final image quality.

Algorithm 1 Our method + Dynamic Time-Travel Straregy

Input: prompt $\mathbf{c}$, noise predictor $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot, t)$, human preference evaluator $\mathcal{L}_R(\cdot, \mathbf{c})$, semantic alignment loss function $\mathcal{L}_A(\cdot)$, timesteps T, decoder D, guidance strength η_t and pre-defined parameters β_t, $\bar{\alpha}_t$, h_t, k.

$\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\boldsymbol{\epsilon}_1 \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $\boldsymbol{\epsilon}_1 = 0$.
  $\tilde{\boldsymbol{\epsilon}}_t, M = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, t)$
  $\mathbf{z}_{t-1} = (1 + \frac{1}{2}\beta_t)\mathbf{z}_t + \beta_t \tilde{\boldsymbol{\epsilon}}_t + \sqrt{\beta_t}\boldsymbol{\epsilon}_1$
  $\mathbf{z}'_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{z}_t + (1 - \bar{\alpha}_t)\tilde{\boldsymbol{\epsilon}}_t)$
  $\mathbf{x}'_{0|t} = D(\mathbf{z}'_{0|t})$
  $\begin{cases} w_A = 1, w_R = 0 & \textbf{if } t \geq 800 \\ w = 1 - e^{-k(\frac{\|\mathbf{z}'_{0|t} - \mathbf{z}'_{0|t+1}\|}{\|\mathbf{z}'_{0|t+1}\|})} & \textbf{if } 800 > t \geq 500 \\ w_A = w, w_R = 1 - w & \\ w_A = 0, w_R = 1 & \textbf{if } 500 > t \geq 1 \end{cases}$
  $\mathcal{L} = w_A \cdot \mathcal{L}_A(M) + w_R \cdot \mathcal{L}_R(\mathbf{x}'_{0|t}, \mathbf{c}, t)$
  $\boldsymbol{g}_t = \nabla_{\mathbf{z}_t}\mathcal{L}$
  $\mathbf{z}_{t-1} = \mathbf{z}_{t-1} - \eta_t \cdot \frac{\|\tilde{\boldsymbol{\epsilon}}_t\|}{\|\boldsymbol{g}_t\|_2^2} \cdot \boldsymbol{g}_t$
  $r_t = h_t \cdot \|\boldsymbol{g}_t\|$ ▷ Compute once at each timestep
  for $i = r_t, ..., 1$ **do** ▷ Iterate r_t times
    $\boldsymbol{\epsilon}_2 \sim \mathcal{N}(0, \mathbf{I})$
    $\mathbf{z}_t = \sqrt{1 - \beta_t}\mathbf{z}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_2$
    Repeat from step3 to step16
return x_0

Table 1. Comparison of AI feedback on SD V1.5-based methods.

Methods	PickScore	HPSv2	ImageReward	Aesthetics
SD V1.5	20.73	0.2341	0.1697	5.337
DNO	20.05	0.2591	-0.3212	5.597
PromptOpt	20.26	0.2490	-0.3366	5.465
FreeDom	21.96	0.2605	0.3963	5.515
AlignProp	20.56	0.2627	0.1128	5.456
Diffusion-DPO	20.97	0.2656	0.2989	5.594
Diffusion-KTO	21.15	0.2719	0.6156	5.697
SPO	21.46	0.2671	0.2321	5.702
SD V1.5+Ours	**23.07**	**0.2755**	**0.7170**	**5.831**

Table 2. Comparison of AI feedback on SDXL-based methods.

Methods	PickScore	HPSv2	ImageReward	Aesthetics
SDXL	21.91	0.2602	0.7755	5.960
DNO	22.14	0.2725	0.9053	6.042
PromptOpt	21.98	0.2708	0.8671	5.881
FreeDom	22.13	0.2719	0.7722	5.908
SDXL+Ours	**24.90**	**0.2839**	**1.074**	**6.138**
Diffusion-DPO	22.30	0.2741	0.9789	5.891
Diff-DPO+Ours	**24.46**	**0.2836**	**1.049**	**6.116**
SPO	22.81	0.2778	1.082	**6.319**
SPO+Ours	**23.85**	**0.2821**	**1.166**	6.278
SD V3.5	21.93	0.2726	0.9697	5.775
FLUX	22.04	0.2760	1.011	6.077

Detail refinement. In the late stage, as changes in the generated results become minimal, we apply the preference loss $\mathcal{L}_R(\mathbf{x}'_{0|t}, \mathbf{c}, t)$ to focus on enhancing fine details, such as texture refinement and subtle feature improvements. These nuanced adjustments boost realism and fidelity, ensuring that the generated output achieves a higher level of texture details and visual aesthetics.

4.3. Improving Training-free Guidance

Polyak Step Size. We dynamically schedule the step size of the update of the latent variable via Polyak step size [36] $\mathbf{z}_{t-1} = \mathbf{z}_{t-1} - \eta_t \cdot \frac{\|\tilde{\epsilon}_t\|}{\|\boldsymbol{g}_t\|_2^2} \cdot \boldsymbol{g}_t$, where η_t is the guidance strength, $\tilde{\epsilon}_t$ is the output of $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, t)$ and $\boldsymbol{g}_t = \nabla_{\mathbf{z}_t}\mathcal{L}$ represents the gradient of loss function in each stage.

Dynamic Time-Travel Straregy. After rethinking the efficient time-travel strategy from FreeDom [54], we discover the fixed iteration steps hinder convergence efficiency since the abilities of each denosing step in diffusion models are different. We propose dynamic recurrent scheme to dynamically schedule the number of iterations at each denosing step in diffusion models are different. We compute the iteration count $r_t = h_t \cdot \|\boldsymbol{g}_t\|$ only once at the start of each step where h_t is a pre-defined parameter. When $\|\boldsymbol{g}_t\|$ is larger, more iterations are required to promote convergence.

5. Experiments

In this section, different experiments are conducted on a wide range of generative models to verify the effectiveness and flexibility of our method. We validate the performance of our method through comparing with various generative backbones and existing state-of-the-art approaches. We also delve deeper into the role of each component for the guidance during the denoising process, supported by ablation studies for further analysis.

5.1. Experimental Setting

Implementation Details. Our method requires no training and is compatible with various generative models. We employ the GPT-4 [1] as the base LLMs to construct semantic graph. Additionally, we adopt the pre-trained step-aware preference model from SPO [24] to facilitate human feedback, developed from PickScore [20] model and fine-tuned on the human preference pairs dataset, Pick-a-Pic [20]. All our experiments uses an NVIDIA A100 GPU for SDXL-based methods or V100 GPU for SD V1.5-based methods.

Datasets and Metrics. We utilized three datasets to validate the effectiveness of our method: 500 unique prompts of the Pick-a-Pic validation set, 500 prompts from HPSv2 benchmark [47] and 1000 prompts from Partiprompt [53]. We choose four AI feedback models to assess the image quality: PickScore [20] (general human preference), HPSv2 [47] (prompt alignment), ImageReward [49] (general human preference), Aesthetic Predictor [35] (non-text-based visual appeal). *For all metrics, higher values indicate better performance*. All test images are generated using 50 denoising steps during inference.

5.2. Comparison with Existing Methods

To verify the effectiveness of our approach, we compare our method with training-based and training-free models. The former comprises human preference learning methods, including AlignProp [30], Diffusion-DPO [45], Diffusion-KTO [23], SPO [30]. The latter includes DNO [41], PromptOpt [7], FreeDom [54]. For SDXL-based frameworks, we make a comparison with the latest models with advanced network architectures (FLUX [21] and SD V3.5 [32]).

5.2.1. Quantitative Analysis

To objectively evaluate the performance of our method, we conduct a quantitative comparison on the Pick-a-Pic dataset, with the results organized in Tab. 1 for SD V1.5-based backbones and Tab. 2 for SDXL-based backbones, respectively. Under the guidance of our method, models including SD V1.5 [32], SDXL [29], Diffusion-DPO [45], and SPO [24] achieve notable improvements, surpassing even advanced models like FLUX [21] and SD V3.5 [32]. Specifically, the training-free methods (DNO [41], PromptOpt [7], FreeDom [54]) suffer from underestimation in early steps, resulting in the inferior outcomes. AlignProp [30] introduces the dependency bias via truncated backpropagation. Diffusion-DPO [45] and Diffusion-KTO [23] produce suboptimal results due to trajectory-level preference supervision, caus-

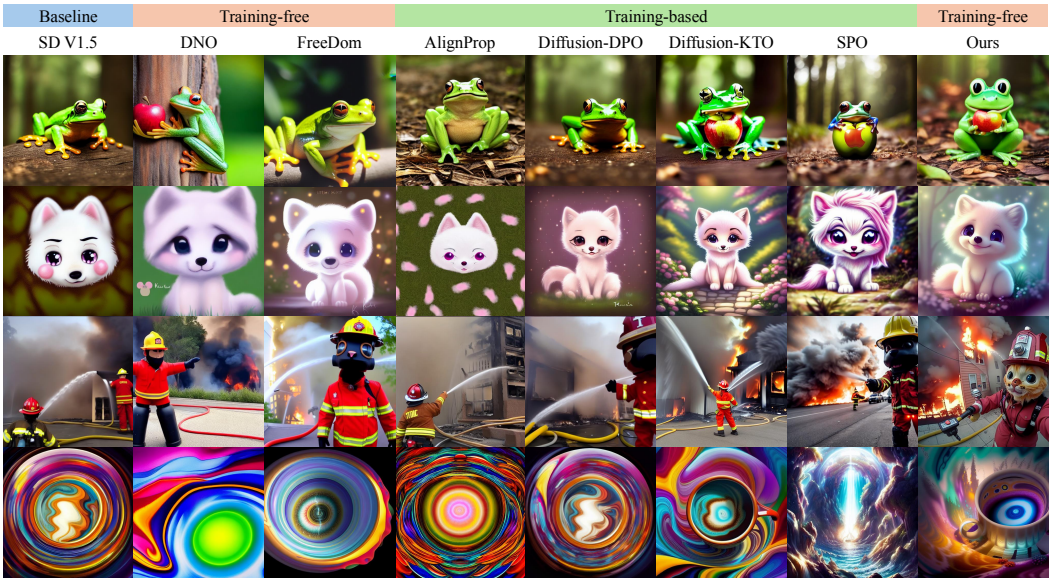


Figure 3. Qualitative comparison based on SD V1.5 backbones.

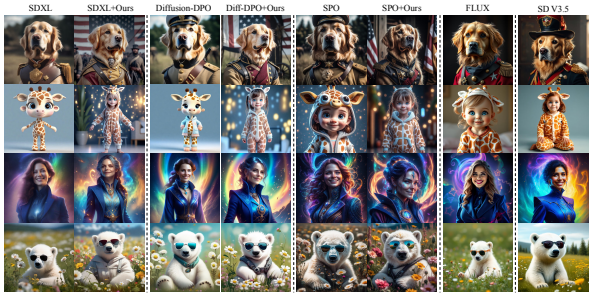


Figure 4. Qualitative comparison based on SDXL backbones.

ing inaccurate alignment. SPO [24] lose the layout and semantic information in early steps. In contrast, our training-free method provides tailored guidance for different denoising steps, generating high-quality images with better layout composition and human preference alignment.

5.2.2. Qualitative Evaluation

We compare our method against a baseline model (SD V1.5 [32]), training-free models (DNO [41], FreeDom [54]) and training-based models (AlignProp [30], Diffusion-DPO [45], Diffusion-KTO [23], SPO [24]), as illustrated in Fig. 3. It is observed that our method contains coherent layout and rich semantic information aligned with the prompt, while achieving overall visual appeal and aesthetic quality. Additionally, we apply our framework to SDXL [29], Diffusion-DPO [45], SPO [24] and compare some sample image generation results in Fig. 4. The results show significant improvement in image quality before and after using our method. Specifically, our method enhances the color vibrancy and detail richness, bringing out more defined textures on fur and fabric as well as vivid lighting effects.

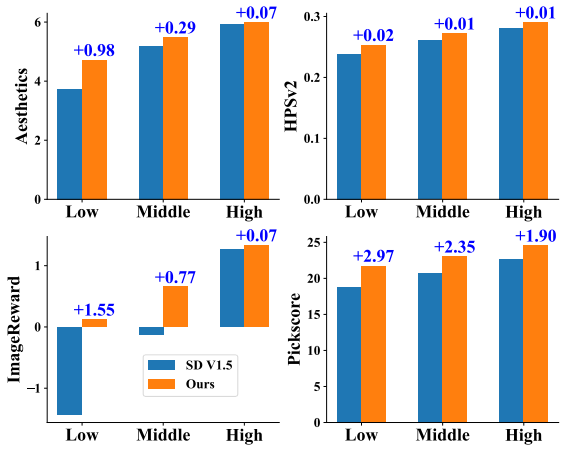
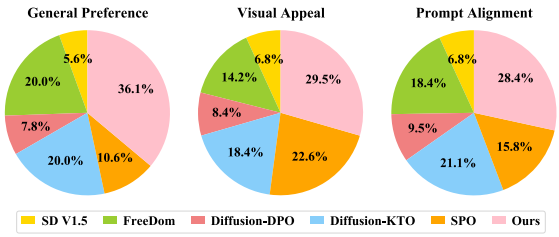


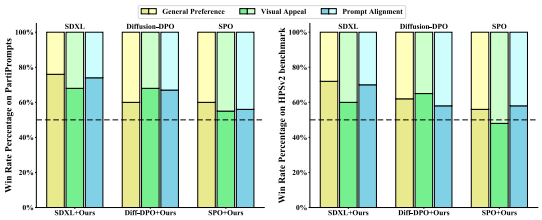
Figure 5. The case comparison of improvements between baseline model and our proposed method.

5.2.3. Case Performance

We compare the performance of our method with baseline model across different cases, as shown in Fig. 5. The test dataset is divided into low-, middle-, and high-reward subsets based on the scores achieved in Aesthetics, HPSv2, ImageReward, and PickScore. For aesthetic, our method consistently surpasses the baseline with gains of +0.98, +0.29, and +0.07 across three subsets, respectively. HPSv2 shows slight improvements across all subsets. ImageReward shows a notable +1.55 gain in the low-reward subset, with moderate increases in the other subsets. For PickScore, our method achieves substantial gains of +2.97, +2.35, and +1.90 across the reward levels. These results demonstrate our method boosts the performance on low-reward cases, particularly in layout and visuals, while maintaining competitive results across middle- and high-reward cases.



(a) User preference distribution on HPSv2 benchmark.



(b) Human evaluation w/ & w/o our method on PartiPrompts & HPSv2.

Figure 6. User study results.

5.2.4. User Study

We conduct a user study to compare our method (based-on SD V1.5) against SD V1.5 [32], FreeDom [54], Diffusion-DPO [45], Diffusion-KTO [23], SPO [24]. We randomly sample 100 unique prompts from HPSv2 [47] benchmark and synthesize images with aforementioned methods. We invite 100 participants and request them to compare all generated images from three different aspects: Q1 General Preference (Which image do you prefer given the prompt?), Q2 Visual Appeal (Which image is more visually appealing?), Q3 Prompt Alignment (Which image better fits the text description?). Fig. 6a shows the approval percentage of each method in three aspects, which demonstrates our method outperforms the previous preference learning models on human feedbacks. Additionally, we conduct a human evaluation on 100 prompts from the PartiPrompts [53] dataset and 100 prompts from the HPSv2 [47] benchmark, comparing SDXL [29], Diffusion-DPO [31], and SPO [24], both with and without our method. The win-rate percentage results are reported in the Fig. 6b, further verifying the effectiveness of our method.

5.3. Ablation Study

All ablation experiments are conducted on SD V1.5 [32] and Pick-a-Pic [20] dataset if not specified.

Effect of semantic alignment loss $\mathcal{L}_A$. The structural formation of image content occurs early in the denoising process, as verified by Tab. 3. The inferior performances are due to the lost of semantic components in the prompt.

Effect of preference alignment loss $\mathcal{L}_R$. The results in Tab. 3 highlight the importance of the human feedback. Accurate preference guidance enhances the image-prompt alignment and improves overall image quality. Additionally, we use the PickScore [20] to replace the step-aware preference model [24], achieving improvements over the baseline model, which further demonstrates the effectiveness and versatility of our method.

Table 3. Ablation study results.

Methods	PickScore	HPSv2	ImageReward	Aesthetics
w/o $\mathcal{L}_A$	22.07	0.2546	0.6413	5.686
w/o $\mathcal{L}_R$	20.37	0.2418	0.1230	5.426
w/o w	22.34	0.2656	0.6748	5.708
w/o Polyak step	20.61	0.2640	0.2838	5.470
w PickSocre	**23.38**	0.2746	0.5463	5.694
Ours (Llama-3.3)	22.58	**0.2829**	0.7048	5.802
Ours (GPT-4)	23.07	0.2755	**0.7170**	**5.831**

Table 4. Effect of the iteration count in time-travel straregy.

Iteration	PickScore	HPSv2	ImageReward	Aesthetics	Avg Time
1	20.41	0.2423	0.6191	5.652	40s
5	21.27	0.2610	0.6679	5.661	170s
10	22.85	0.2677	0.7055	5.706	295s
Dynamic	**23.07**	**0.2755**	**0.7170**	**5.831**	190s

Effect of adpative weight w. As shown in Tab. 3, the adaptive mechanism for balancing these two losses not only preserves essential image structure in the initial steps but also enhances final image fidelity, supporting both content alignment and visual quality.

Effect of Polyak step size. We adopt the Polyak step size [36] to achieve near-optimal convergence rates for our method. Tab. 3 demonstrate it is effective to guide the generation process by bridging the substantial gap between unconditional generation and specified condition.

Effect of open-source LLMs. DyMO's LLM-based text prompt pre-analysis is moderate for LLMs and does not require advanced models (LLM-agnostic). We validate it with open-source Llama3.30-70B (Tab. 3), achieving strong performance (22.58)—better than others, slightly below GPT-4-based results (23.07).

Effect of dynamic recurrent strategy. We conduct experiments with fixed iteration counts set to 1, 5, and 10 as presented in Tab. 4. Our approach achieves the better performance with a moderate time cost, demonstrating the effectiveness of dynamic recurrent strategy.

6. Conclusion

In this work, we propose a novel framework to achieve better diffusion model alignment via dynamic multi-objective scheduling. Apart from text-aware human preference score, we introduce a semantic alignment objective to mitigate the inaccurate estimation for blurred denoised image in early denoising steps. We utilize two objectives in a dynamical coordinated manner to guide model outputs aligned with contextual semantics and human preferences. Extensive experiments show that our algorithm is plug-and-play and enables effective alignment, producing high-quality images with fine layout structure and captivating aesthetics. The performance of our method on different metrics outperforms pre-trained baseline models and surpasses both training-based and training-free approaches.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 6

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 2, 3

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

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 2, 3

[5] Hyungjin Chung, Dohoon Ryu, Michael T McCann, Marc L Klasky, and Jong Chul Ye. Solving 3d inverse problems using pre-trained 2d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22542–22551, 2023. 3

[6] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 2, 3

[7] Niklas Deckers, Julia Peters, and Martin Potthast. Manipulating embeddings of stable diffusion prompts. *arXiv preprint arXiv:2308.12059*, 2023. 2, 3, 6, 13, 15

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3

[9] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023. 3

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 14

[11] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *arXiv preprint arXiv:2406.04312*, 2024. 2, 3

[12] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[13] Alexander Gambashidze, Anton Kulikov, Yuriy Sosnin, and Ilya Makarov. Aligning diffusion models with noise-conditioned perception. *arXiv preprint arXiv:2406.17636*, 2024. 2, 3

[14] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 14

[15] Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-rpo: Aligning diffusion models through relative preference optimization. *arXiv preprint arXiv:2406.06382*, 2024. 2, 3

[16] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9380–9389, 2024. 3

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 4

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[20] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. 2, 3, 4, 5, 6, 8

[21] Black Forest Labs. Flux.1-schnell, 2024. Accessed: 2024-08-17. 6, 14

[22] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 3

[23] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *arXiv preprint arXiv:2404.04465*, 2024. 2, 3, 6, 7, 8, 13

[24] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2024. 2, 3, 4, 5, 6, 7, 8, 13

[25] Buhua Liu, Shitong Shao, Bao Li, Lichen Bai, Haoyi Xiong, James Kwok, Sumi Helal, and Zeke Xie. Alignment of diffusion models: Fundamentals, challenges, and future. *arXiv preprint arXiv:2409.07253*, 2024. 2

[26] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 289–299, 2023. 2

[27] Nithin Gopalakrishnan Nair and Vishal M Patel. Dreamguider: Improved training free diffusion-based conditional generation. *arXiv preprint arXiv:2406.02549*, 2024. 3

[28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 6, 7, 8, 13

[30] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 2, 3, 6, 7, 13

[31] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 8

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 6, 7, 8, 13

[33] 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. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[34] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*, 2024. 2

[35] Christoph Schuhmann. Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. Accessed: 2023 - 11- 10. 2, 6

[36] Yifei Shen, Xinyang Jiang, Yezhen Wang, Yifan Yang, Dongqi Han, and Dongsheng Li. Understanding and improving training-free loss-based diffusion guidance, 2024. 2, 3, 6, 8

[37] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. 3

[38] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2, 3

[40] Zhenhong Sun, Junyan Wang, Zhiyu Tan, Daoyi Dong, Hailan Ma, Hao Li, and Dong Gong. Eggen: Image generation with multi-entity prior learning through entity guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6637–6645, 2024. 3, 12

[41] Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Tuning-free alignment of diffusion models with direct noise optimization. *arXiv preprint arXiv:2405.18881*, 2024. 2, 3, 6, 7, 13, 15

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[43] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 4

[44] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7280–7290, 2023. 2, 3

[45] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 2, 3, 6, 7, 8, 13

[46] Hao Wang, Yongsheng Yu, Tiejian Luo, Heng Fan, and Libo Zhang. Magic: Multi-modality guided image completion. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[47] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2, 6, 8

[48] Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. *arXiv preprint arXiv:2405.00760*, 2024. 3

[49] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4, 6

[50] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, 2024. 2, 3

[51] Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. *arXiv preprint arXiv:2402.08265*, 2024. 2, 3

[52] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *arXiv preprint arXiv:2409.15761*, 2024. 2, 3

[53] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 6, 8

[54] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023. 2, 3, 6, 7, 8, 12, 13, 15

[55] Yasi Zhang, Peiyu Yu, and Ying Nian Wu. Object-conditioned energy-based attention map alignment in text-to-image diffusion models. In *European Conference on Computer Vision*, pages 55–71. Springer, 2025. 2, 3, 4, 12

[56] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in neural information processing systems*, 33:7559–7570, 2020. 3