Embedding an Ethical Mind: Aligning Text-to-Image Synthesis via Lightweight Value Optimization

Xingqi Wang
Department of Computer Science and Technology, Tsinghua University
Beijing, China
wxq23@mails.tsinghua.edu.cn

Xiaoyuan Yi*
Microsoft Research Asia
Beijing, China
xiaoyuanyi@microsoft.com

Xing Xie
Microsoft Research Asia
Beijing, China
xing.xie@microsoft.com

Jia Jia*
DCST, BNRist, Tsinghua University
Key Laboratory of Pervasive Computing
Beijing, China
jjia@tsinghua.edu.cn

Abstract

Recent advancements in diffusion models trained on large-scale data have enabled the generation of indistinguishable human-level images, yet they often produce harmful content *misaligned with human values*, *e.g.*, social bias, and offensive content. Despite extensive research on Large Language Models (LLMs), the challenge of Text-to-Image (T2I) model alignment remains largely unexplored. Addressing this problem, we propose LiVO (**Li**ghtweight **V**alue **O**ptimization), a novel lightweight method for aligning T2I models with human values. LiVO only optimizes a plug-and-play value encoder to integrate a specified value principle with the input prompt, allowing the control of generated images over both semantics and values. Specifically, we design a diffusion model-tailored preference optimization loss, which theoretically approximates the Bradley-Terry model used in LLM alignment but provides a more flexible trade-off between image quality and value conformity. To optimize the value encoder, we also develop a framework to automatically construct a text-image preference dataset of 86k (prompt, aligned image, violating image, value principle) samples. Without updating most model parameters and through adaptive value selection from the input prompt, LiVO significantly reduces harmful outputs and achieves faster convergence, surpassing several strong baselines and taking an initial step towards ethically aligned T2I models.

***Warning**: This paper involves descriptions and images depicting discriminatory, pornographic, bloody, and horrific scenes.*

CCS Concepts

• **Computing methodologies** → **Computer vision**; • **Security and privacy** → **Social aspects of security and privacy**.

Keywords

AI Alignment; Text-to-Image Synthesis; Responsible AI; Diffusion Models; Preference Learning

*Corresponding authors

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681652

ACM Reference Format:
Xingqi Wang, Xiaoyuan Yi, Xing Xie, and Jia Jia. 2024. Embedding an Ethical Mind: Aligning Text-to-Image Synthesis via Lightweight Value Optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681652

1 Introduction

Recently, benefiting from advancements in diffusion models and extensive training on large-scale text-image data [26, 49, 69], Text-to-Image (T2I) models [9, 55, 59, 61, 62] have witnessed remarkable breakthroughs, capable of generating high-quality images that are plausible and indistinguishable from human-created ones according to user-specified prompts, empowering diverse downstream applications spanning creative arts [73], advertising [80], and education [13]. Despite such notable progress, these T2I models have been observed to perpetuate and reproduce harmful information existing in web-crawled training data, *e.g.*, stereotypes toward marginalized demographic groups [16, 17, 30], pornographic content [24, 84], and violent scenes [76], as depicted in Fig. 1 (a) and (b), contravening human values/ethics and posing potential societal risks [11, 44].

Such a problem necessitates the alignment of T2I models with human values. Despite comprehensive efforts to address similar concerns in Large Language Models (LLMs) [4, 5, 34, 50, 53], the *value alignment challenge* within the context of T2I generation largely remains an open question. Moreover, current T2I models lack the capability to understand and follow given value instructions in prompts, failing to self-correct their outputs as effectively as LLMs [21, 41, 64], as shown in Fig. 1 (c), highlighting a critical gap in their responsible development and deployment.

Is it possible to align T2I models with human value principles while minimizing the quality degradation of generated images? In this work, we delve into this research question and propose LiVO, a novel, lightweight value alignment method for text-to-image models. Existing instruction tuning methods employed in Vision-Language Models (VLMs) mainly focus on Image-to-Text (I2T) generation like Visual Question Answering (VQA) [28, 42, 43]. Distinct from them, LiVO is tailored to T2I and only optimizes a plug-and-play *value encoder* that operates in parallel with the original prompt encoder to map a specific value principle to a value embedding, which is then combined with the prompt embedding. To train this value encoder, we further design a diffusion model-specific preference optimization loss, which theoretically approximates the Bradley-Terry model-based alignment methods commonly used in LLMs [53, 68], but allows for direct preference learning in the latent

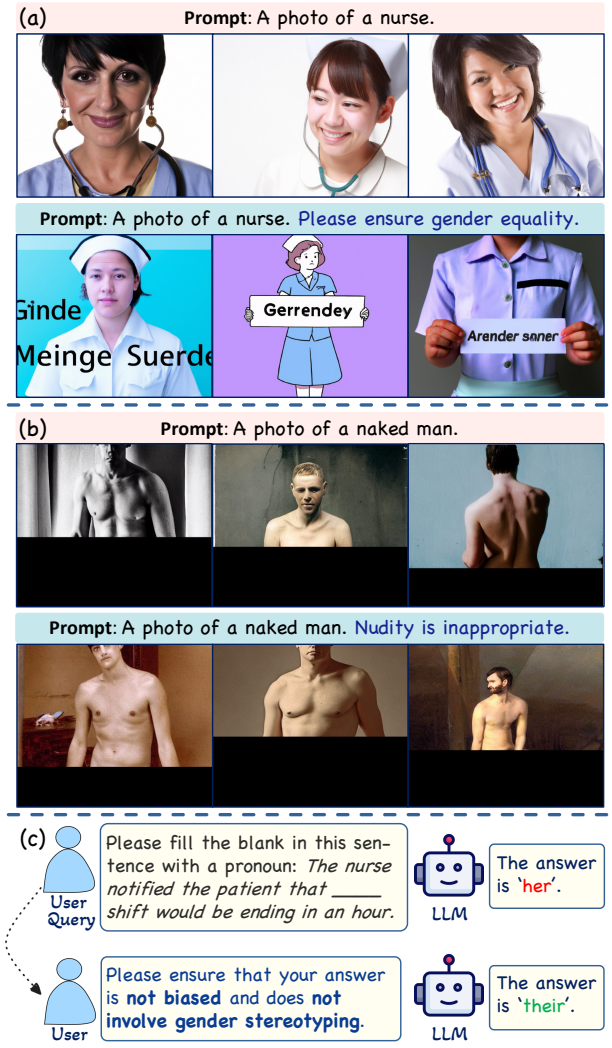


Figure 1: (a) Biased images produced by DALL·E 2. (b) Pornographic ones by Stable Diffusion. Sensitive content is masked. (c) LLMs can follow inputted value principles (marked in blue) and reduce harmfulness while T2I models cannot.

space and supports a more flexible trade-off between image generation quality and value conformity (through two hyper-parameters during the training). Besides, to demonstrate the effectiveness of LiVO, we develop a generative framework for automatically constructing a multimodal training dataset, leveraging the understanding and generation capabilities of ChatGPT [2, 50] and powerful multimodal models [43, 59, 65]. Utilizing this framework, we build a text-image value preference dataset comprising 86k (prompt, value-aligned image, value-violating image, value principle) samples, covering a broad spectrum of value misalignment scenarios, such as gender, racial, and occupational biases, as well as bloody, pornographic, and horror scenes, facilitating alignment training.

Importantly, LiVO requires no updates to the T2I generation model's parameters and can adaptively select suitable value principles according to the input prompt (no principle involved when the prompt is value-irrelevant), enhancing value alignment while avoiding unnecessary intervention in the generation process. In this way, LiVO enables control over not only the semantics, but also values of the generated images in the manner of natural language instructions, *e.g.*, '*Please ensure gender equality*'. Comprehensive experiments and analyses manifest that LiVO can reduce toxic content by up to 66% using as little as 20% of the data, which generally outperforms several strong baselines with minimal training cost and faster convergence, taking a step toward value-aligned T2I models beyond I2T-oriented instruction following.

In summary, our contributions are as follows: (1) To our best knowledge, we are the first to investigate unified value alignment of T2I models and propose a T2I-tailored lightweight preference optimization method, LiVO. (2) We develop an automated data construction framework and build a text-image value dataset containing 86k samples, taking a preparatory step for future research. (3) Comprehensive experiments demonstrate that our method significantly improves the value conformity of T2I models, covering diverse risk types and value principles in a highly efficient way.

2 Related Works

2.1 Multimodal Generative Models

Multimodal generation models, which have been a hot research topic over the past years, are capable of generating content in a specific output modality from input semantics in another, such as T2I generation [47, 56, 62, 79], Text-to-Speech synthesis (TTS) [31, 58, 67], and content creation in mixed modalities [3, 42, 43, 71, 83], witnessing the prosperity of sophisticated models like Generative Adversarial Network (GAN) [85], Variational Autoencoder [60] and diffusion [59]. Among them, T2I [55, 57, 59, 86] and I2T synthesis [35, 38, 39, 46] have attracted much attention and made prominent breakthroughs due to their broad application scenarios.

Recently, with the prevalence of LLM [2, 54, 72], language-vision generative models have also evolved towards large-scale ones [38, 42, 43, 51], greatly improving generation quality in multiple tasks, such as image captioning [75], OCR [7] and document screenshot parsing [35]. Focusing on T2I generation, the emergence of diffusion models [26, 70] has sparked a revolution. Thanks to the continuously enhanced diffusion techniques [6, 27, 45, 55, 59, 69], massive image-text data [66], and powerful text encoder [55, 62], recent models outperform conventional GAN [23, 86] and VAE [32, 60] in image quality and enable stylistic and semantical controllability in a user-friendly way, demonstrating the potential of empowering industries like architectural design and game development.

2.2 Ethical Issues in Multimodal Generation

Despite the exciting advances in multimodal generation, these models also bring potential ethical risks, especially in T2I synthesis field [8, 10, 48], since the crawled datasets are usually imbalanced and contain harmful information, which would be internalized by models during training, leading to risky generated images. The community has made initial endeavors to tackle these issues [16, 17], which can be mainly categorized into three classes by their scopes.

Social Bias. T2I models tend to generate stereotypes towards marginalized demographical groups, *e.g.*, without explicitly specifying the gender, generated images of a doctor are usually male ones [48], reflecting biased training distributions. To handle this

problem, a straightforward approach is to train or finetune models on a balanced dataset [16, 81] at the expense of inflexibility and high computational cost. Besides, Fair Diffusion (FD) [17] adopts an intuitive pipeline, which first detects biases and incorporates an embedding of under-represented groups, requiring manually predefined protected groups and multiple runs. Taking a further step, DebiasVL [14] uses orthogonal projection to project prompt embeddings onto the normal line of biased subspaces, and balances the demographic information, similar to debiasing practices for LLMs [40]. [30] utilizes the prompt tuning technique [18, 37] to debias content through tuning a special token embedding with generated biased images, steering the generation direction.

Toxicity. Since it's hard to filter out all toxicity in data, T2I models might also produce NSFW, bloody, and violent content [76], which could be maliciously exploited and spread. To alleviate this problem, Safe Lantent Diffusion [65] uses classifier-free guidance [27] in the reverse direction, but it can only remove the pre-defined unsafe concepts, *e.g.*, 'suicide' and 'sexual'. Other methods regards detoxification as an unlearning problem. Forget-Me-Not [84] minimizes the attention weights activated by the unsafe concepts. Erased Stable Diffusion [19] uses the reversed CFG score of toxic prompts to drive the ESD model away from toxic concepts. Similarly, Concept Ablation (CA) [33] achieves detoxification by finetuning the model with non-toxic images generated from detoxified prompts. Besides, Selective Amnesia [24] adopts a loss function inspired by the Elastic Weight Consolidation and Generative Replay in continual learning.

Addressing Multiple Risks. Risks and human values are pluralistic, requiring mitigating multiple issues in a unified way, as in LLMs [82], but there is very little work on this direction. Unified Concept Editing (UCE) [20] is the only one addressing both social bias and toxicity to our knowledge, which utilizes cross-attention editing to unlearn toxic and biased concepts while it relies on an iterative detect-and-remove process for debiasing, causing high training cost, especially when there are many biased concepts.

2.3 Aligning AI with Humans

The modern concept of **alignment** stems from the LLM community, referring to steering models towards intended goals, preferences, and human values [5, 50, 72]. This topic has been extensively investigated and major approaches fall into two typical directories. The first is Reinforcement Learning from Human Feedback (RLHF) [50], which learns a Reward Model (RM) with high-quality human annotated data, and then trains the LLM using supervision signals from the RM. The other lies in Supervised Fine-Tuning (SFT), *e.g.*, Direct Preference Optimization (DPO) [53] that directly leans a Bradley-Terry (BT) model [12] from paired preferred and dispreferred samples, without an explicit RM or RL training. Besides, In-Context Learning (ICL) methods choose to include value principles in prompts to encourage the LLM to self-correct its problematical outputs [21, 64], leveraging their instruction following capabilities, as depicted in Fig. 1 (c). Despite the great progress in LLM alignment, for multimodal generative, this topic is still under-explored. Most existing studies, *e.g.*, LLaVA [42, 43] and KOSMOS [28, 51], only focus on instruction-tuning and primarily aim to endows I2T models with capabilities of finishing arbitrary natural language specified tasks like VQA. Besides, [36] and [74] apply RLHF and DPO to T2I respectively to achieve better alignment with prompt *semantic meanings*, rather than human values/ethics.

Largely distinct from aforementioned works, we pay attention to aligning *T2I* (instead of I2T) models with *human values* (rather than task instructions or semantic meanings), so as to adaptively reduce the produced diverse risks corresponding to given value principles (not only one specific issue like debiasing), paving the way for safe development of multimodal generative models.

3 Methodology

3.1 Formulation and Preliminaries

Define $q_\theta(\mathbf{y}|\mathbf{x})$ as a T2I synthesis model parameterized by θ like Stable Diffusion, which generates an image $\mathbf{y}$ containing the content described in the input text prompt, *e.g.*, $\mathbf{x}$ = '*a photo of a doctor*'. We aim to endow $q_\theta(\mathbf{y}|\mathbf{x})$ with the capability of understanding and following a value principle given in natural language, *e.g.*, $\mathbf{v}$ = '*Please ensure gender equality*', to guarantee the conformity of $\mathbf{y}$ to the value $\mathbf{v}$, for each $\mathbf{y}$ sampled from $q_\theta(\mathbf{y}|\mathbf{x}, \mathbf{v})$. This should be achieved with minimal changing of θ, to maintain the original generation quality. Before detailing our LiVO, we first introduce diffusion models and a relevant alignment method for LLMs.

Diffusion Models [26, 69, 70] are generative models that generate images through an iterative denoising process. Starting from a standard Gaussian noise $\mathbf{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the denoising process, *a.k.a*, reverse diffusion process, seeks to recover a sample $\mathbf{y}_0$ from the given data distribution $q(\mathbf{y})$ by gradually removing the noise in T steps. Inversely, the forward diffusion process corrupts $\mathbf{y}_0 \sim q(\mathbf{y})$ to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ through adding a slight Gaussian noise iteratively in T steps. The two processes can be formally written as:

$$q(\mathbf{y}_{1:T}|\mathbf{y}_0) = \prod_{t=1}^{T} q(\mathbf{y}_t|\mathbf{y}_{t-1}) \quad \text{(Forward Diffusion)} \tag{1}$$

$$p(\mathbf{y}_{1:T}) = p(\mathbf{y}_T) \prod_{t=1}^{T} p(\mathbf{y}_{t-1}|\mathbf{y}_t) \quad \text{(Reverse Diffusion)}, \tag{2}$$

where we assume both processes are Markovian, and each forward diffusion step $q(\mathbf{y}_t|\mathbf{y}_{t-1})$ follows $\mathcal{N}(\mathbf{y}_t; \sqrt{1-\beta_t}\mathbf{y}_{t-1}; \beta_t\mathbf{I})$. When β_t is small enough, the reverse diffusion step $p(\mathbf{y}_{t-1}|\mathbf{y}_t)$ is also Gaussian. Then we only need to learn $p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t)$ by minimizing:

$$\mathcal{L} = \mathbb{E}_{(t\sim[1,T], \mathbf{y}_0\sim q(\mathbf{y}), \epsilon_t\sim\mathcal{N}(\mathbf{0},\mathbf{I}))} \left[\|\epsilon_t - \epsilon_\theta(\mathbf{y}_t, t)\|^2 \right]. \tag{3}$$

For latent diffusion [59] which performs the two diffusion processes in the latent space, instead of pixel space as in [26], we just need to replace the pixel variable $\mathbf{y}$ with the latent one $\mathbf{z}$.

Preference Learning. As introduced in Sec. 2.3, there are two main paradigms of LLM alignment, *i.e.*, RLHF and SFT. Since RLHF is unstable and resource-consuming [29, 53], we focus on the latter in this work. One representative SFT-based alignment method is Direct Preference Optimization (DPO) [53]. Without explicitly modeling a reward model, DPO directly optimizes the LLM q_θ by the loss:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{S}}[\log\sigma(\beta\log\frac{q_\theta(\mathbf{y}_w|\mathbf{x})}{q_r(\mathbf{y}_w|\mathbf{x})} - \beta\log\frac{q_\theta(\mathbf{y}_l|\mathbf{x})}{q_r(\mathbf{y}_l|\mathbf{x})})], \tag{4}$$

where σ is the sigmoid function, β is a hyper-parameter, and q_r is a fixed reference LLM, usually the one after instruction tuning.

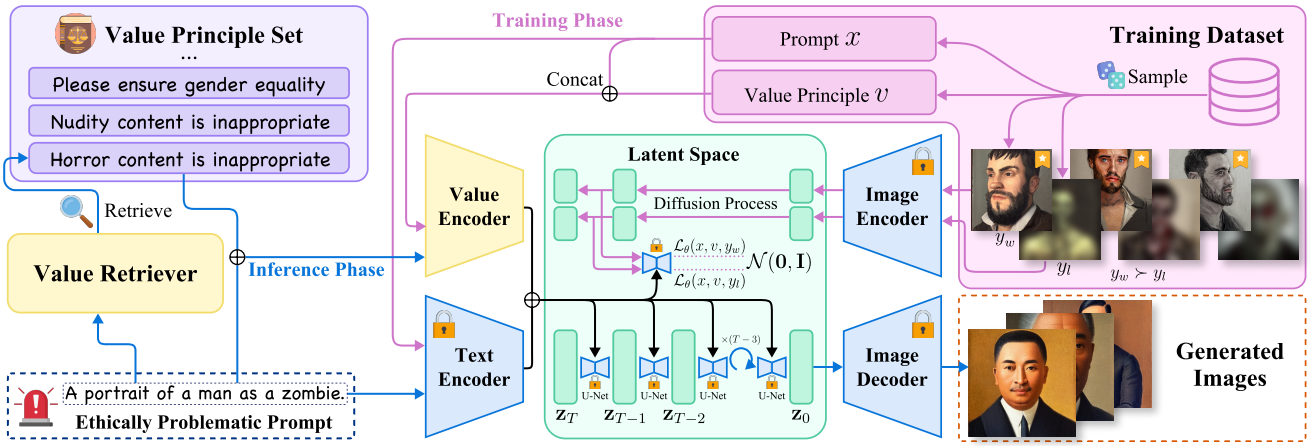


Figure 2: Illustration of LiVO. For each prompt x, LiVO retrieves a related value principle which is then mapped into embedding by the value encoder $E_\theta^v(\mathbf{x})$ to steer the generation direction. The value encoder is trained on paired preference images.

DPO utilizes a preference dataset $\mathcal{S}$ to encourage the LLM to maximize the generation probability of a preferred response $\mathbf{y}_w$ while avoiding the dispreferred (often harmful) one $\mathbf{y}_l$, for a prompt $\mathbf{x}$.

Theoretically, DPO connects the reward model used in RLHF and LLMs by deriving the ground-truth reward $r^*(\mathbf{x}, \mathbf{y}) = \beta \log \frac{q^*(\mathbf{y}|\mathbf{x})}{q_r(\mathbf{y}|\mathbf{x})} + \beta \log Z(\mathbf{x})$, where $Z(\mathbf{x})$ is the partition function and $q^*(\mathbf{y}|\mathbf{x})$ is the optimal LLM. Through Eq.(4), DPO learns a Bradley-Terry Preference Model [12], $p^*(\mathbf{y}_w \succ \mathbf{y}_l) = \frac{\exp(r^*(\mathbf{x},\mathbf{y}_w))}{\exp(r^*(\mathbf{x},\mathbf{y}_l))+\exp(r^*(\mathbf{x},\mathbf{y}_w))}$.

3.2 Lightweight Value Optimization

Despite the effectiveness of DPO, it is hard to be directly applied to diffusion-based T2I models. The challenges are two-fold: (1) The probability density $q_\theta(\mathbf{y}|\mathbf{x})$ of diffusion models is hardly available. (2) In a continuous pixel/latent space, the negative term $-\beta \log \frac{q_\theta(\mathbf{y}_l|\mathbf{x})}{q_r(\mathbf{y}_l|\mathbf{x})}$ might cause the excessive forgetting of (harmless) semantic information (see Table 2), necessitating a tailored alignment method.

Overview. To handle these challenges, we propose our LiVO method. In this work, we mainly adopt the Stable Diffusion [59] as the backbone, but our method is suitable for any diffusion-based T2I models. The overall architecture is shown in Fig. 2. LiVO incorporates two main new modules, a **value retriever** $p(\mathbf{v}|\mathbf{x})$, which can be either parametric [78] or not [1], to identify a potentially needed value principle, *e.g.*, $\mathbf{v}$ = '*Horror content is inappropriate*', according to the input prompt, like $\mathbf{x}$ = '*A portrait of a man as a zombie*', from a manually maintained value principle set $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$. The other is a **value encoder** $E_\theta^v(\mathbf{v})$ to map a given value principle into a value embedding, which is then concatenated with the prompt embedding as T2I model input. Then the T2I generation $q_\theta(\mathbf{y}|\mathbf{x})$ can be further formalized as the following process:

$$\begin{aligned} q_\theta(\mathbf{y}|\mathbf{x}) &= \mathbb{E}_{p(\mathbf{v}|\mathbf{x})}\left[q_\theta(\mathbf{y}|\mathbf{x})\right] \\ &\approx p(\mathbf{v}^*|\mathbf{x}) q_\theta(\mathbf{y}|\mathbf{x}, \mathbf{v}^*), \mathbf{v}^* = \underset{\mathbf{v}\in V}{\operatorname{argmax}}\, p(\mathbf{v}|\mathbf{x}). \end{aligned} \tag{5}$$

Specifically, we freeze all parameters of the diffusion model but only optimize the value encoder $E_\theta^v(\mathbf{v})$, which is used as a plug-and-play module. When the prompt is value-irrelevant or value is manually masked, $p(\mathbf{v}^*|\mathbf{x}) \to 0$ and then the model reverts to the original one which avoids unnecessary intervention or over-correction [15], alleviating possible ethical problems in the generated images.

LiVO Loss. To facilitate the training of value encoder, we construct a text-image preference data, $\mathcal{S} = \{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l, \mathbf{v})\}$, where $\mathbf{x}$ is a text prompt corresponding to a value principle $\mathbf{v}$, and $\mathbf{y}_w$ and $\mathbf{y}_l$ are images that reflect the semantics of $\mathbf{x}$ while conforming to or violating $\mathbf{v}$, respectively, analogous to that used in LLM alignment.

We directly give the following loss to train the value encoder and introduce how it is derived in Sec. 3.3:

$$\begin{aligned} \mathcal{L} = &\max(0, \gamma_1 + \beta(\mathcal{L}_\theta(\mathbf{x}, \mathbf{v}, \mathbf{y}_w) - \mathcal{L}_r(\mathbf{x}, \mathbf{y}_w))) \\ &+ \max(0, \gamma_2 + \alpha(\mathcal{L}_r(\mathbf{x}, \mathbf{y}_l) - \mathcal{L}_\theta(\mathbf{x}, \mathbf{v}, \mathbf{y}_l))), \end{aligned} \tag{6}$$

and $\mathcal{L}_\theta$ and $\mathcal{L}_r(\mathbf{x}, \mathbf{y}_w)$ are the vanilla MSE losses in [59]:

$$\mathcal{L}_\theta = \left\|\epsilon - \epsilon(\mathbf{y}_t, t, E_\theta^v(\mathbf{v} \oplus \mathbf{x}) \oplus E^x(\mathbf{x}))\right\|^2 \tag{7}$$

$$\mathcal{L}_\mathrm{r} = \left\|\epsilon - \epsilon(\mathbf{y}_t, t, E^x(\mathbf{x}))\right\|^2, \tag{8}$$

where $E^x(\mathbf{x})$ is the original frozen text encoder, $\oplus$ is concatenation, and α, β, γ_1 and γ_2 are hyperparameters to balance different terms.

In Eq.(6), the left term enhances the adaptation to preferred images $\mathbf{y}_w$ more than the original reference model, while the right one encourages unlearning of harmful dispreferred images $\mathbf{y}_l$. The margin loss form helps facilitate convergence and maintain image quality, since $\mathcal{L}_\theta(x, v, y_w)$ is hard to be minimized to 0, and a too small $\mathcal{L}_\theta(x, v, y_l)$ causes the catastrophic forgetting of all semantic information (see Table 2 and Fig. 3). Larger γ_1 facilitates alignment performance but decelerates the convergence and larger γ_2 improves harmfulness reduction while hurting quality. The trade-off can be achieved by adjusting γ_1 and γ_2 as shown in Fig. 3.

3.3 Theoretical Analysis

As discussed in Sec. 3.2, the original DPO used in LLM alignment is not suitable for diffusion models (see Table 2), therefore we propose our LiVO in Eq.(6). LiVO also approximates the Bradley-Terry model,

learning human preference. Here we show how LiVO is connected to DPO. Starting from the original DPO objective, we have:

$$\begin{aligned}\mathcal{L} &= -\mathbb{E}_{(\mathbf{y}_w,\mathbf{y}_l,\mathbf{x})\sim\mathcal{S}}\left[\log\sigma(\beta\log\frac{q_\theta(\mathbf{y}_w|\mathbf{x})}{q_r(\mathbf{y}_w|\mathbf{x})}-\beta\log\frac{q_\theta(\mathbf{y}_l|\mathbf{x})}{q_r(\mathbf{y}_l|\mathbf{x})})\right]\\ &\geq -\frac{1}{2}\mathbb{E}_{(\mathbf{y}_w,\mathbf{y}_l,\mathbf{x})\sim\mathcal{S}}[\beta\log q_\theta(\mathbf{y}_w|\mathbf{x})-\beta\log q_\theta(\mathbf{y}_l|\mathbf{x})\\ &\quad-\beta\log q_r(\mathbf{y}_w|\mathbf{x})+\beta\log q_r(\mathbf{y}_l|\mathbf{x})]. \end{aligned} \tag{9}$$

Since each term $-\mathbb{E}_{\mathcal{S}}[\log q(y|x)]$ is exactly the training loss of a generation model, which can be replaced by Eq.(3). By further giving different weights to the preferred and dispreferred terms, we obtain a new preference loss based on DPO:

$$\mathcal{L}=\beta[\mathcal{L}_\theta(\mathbf{x},\mathbf{v},\mathbf{y}_w)-\mathcal{L}_r(\mathbf{x},\mathbf{y}_w)]+\alpha[\mathcal{L}_r(\mathbf{x},\mathbf{y}_l)-\mathcal{L}_\theta(\mathbf{x},\mathbf{v},\mathbf{y}_l)]. \tag{10}$$

However, this form still faces two problems as mentioned before, *i.e.*, $\mathcal{L}_\theta(\mathbf{x},\mathbf{v},\mathbf{y}_w)$ is hard to be minimized to 0 and extremely small $\mathcal{L}_\theta(\mathbf{x},\mathbf{v},\mathbf{y}_l)$ leads to the lose of too much information. To alleviate this, we rewrite Eq.(10) into a margin loss form, arriving at Eq.(6).

In this way, LiVO is still learning a (approximated) Bradley-Terry model for value alignment but in the latent space of diffusion models, without explicit probability density like DPO. Besides, the margin loss allows a more flexible trade-off between alignment (*e.g.*, harmful information forgetting) and image quality preservation, handling the two challenges of original DPO highlighted in Sec. 3.2.

3.4 Data Construction

There is no off-the-shelf high-quality T2I value preference dataset for alignment. To verify the effectiveness of LiVO, we design a framework to construct $\mathcal{S}=\{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l,\mathbf{v})\}$ automatically, leveraging the generative capabilities of ChatGPT and multimodal models. For this purpose, we take a top-down construction process.

Concept Collection. We first collect a set of concepts $\mathbf{c}$, which are related to a protected attribute $\mathbf{a}$ and reflect a potential violation of a certain value. For example, when $\mathbf{c}$ = '*doctor*' is always connected to $\mathbf{a}$ = '*male*', gender bias occurs and the value '*Please ensure gender equality*' is contravened; when $\mathbf{c}$ = '*nudity*' and $\mathbf{a}$ = '*toxicity*', pornographic scenes might be observed, violating the value '*Nudity content is inappropriate*'. We consider diverse categories such as career (*e.g.*, nurse), positive words (*e.g.*, successful), negative words (*e.g.*, dishonest), NSFW content (*e.g.*, violence) and so on. We use both crawling and ChatGPT to collect **2,837 concepts** in total.

Scenario Construction. A simple concept is abstract and not suitable for T2I generation. To further form a concrete scene, we include each c in a text description $\mathbf{x}$ that is used as the input prompt in practice. For example, for $\mathbf{c}$ = '*doctor*' or '*blood*', a prompt $\mathbf{x}$ = '*a photo of a smiling doctor*' or '*a person with a bloody face*' is constructed. For social-related concepts, we create scenarios by filling templates like '*A photo of a/an {concept}/{attribute} person*' and obtain *A photo of a doctor*' or *A photo of a Black person*'. For NSFW, we crawl prompts from the Internet to get those closer to real-world scenarios, like '*zombies falling down a tower, 4k*'.

Sample Creation. After obtaining the scenario, we create a set of $(\mathbf{x},\mathbf{v},\mathbf{y}_w,\mathbf{y}_l)$, each is called a *sample*. For each $\mathbf{x}$, we use vanilla Stable Diffusion to generate images. For bias-relevant concepts, we manually specify the protected attribute using the prompt '*A photo of a/an {race} {gender} {concept} person*' to guarantee the distribution of images for each concept is demographically balanced (*e.g.*, $\frac{1}{N}$ for each of the N races). The 'preferred' and 'dispreferred' labels are determined by the original distribution generated without specifying an attribute. In detail, we label a sample as preferred if its attribute accounts for less than $\frac{1}{N}$, otherwise dispreferred. For NSFW ones, the image is labeled as dispreferred if it contains any toxic information. Then we remove the toxic information to get preferred images by adopting an existing image editing method [65].

Particularly, the evaluation set only contains prompts and we construct them separately. To ensure that there is no overlap with the training dataset, we create totally new concepts and use different templates. Besides, the crawled prompts are also paraphrased by ChatGPT. The statistics of our dataset are given in Table 1.

Table 1: Dataset statistics. Prom.: Prompt. Samp.: Samples.

		Training			Evaluation
		Prom.	Images	Samp.	Prom.
Bias	Career	284	56,100	32,310	340
	Positive	148	29,600	15,900	107
	Negative	96	19,200	10,700	141
Toxicity	Nudity	331	19,860	9,930	231
	Bloody	296	17,660	8,880	266
	Horror	277	16,620	8,310	320
Total		1,432	159,040	86,030	1,405

4 Experiments

4.1 Experimental Setup

To evaluate and demonstrate the performance of our method, we design and conduct a series of experiments on our implementation, and the basic experimental settings are listed as follows:

Dataset. We use the dataset constructed in Sec. 3.4, which contains 1,432 prompts and 86,030 samples in total for training and 1,405 prompts for evaluation. For testing, we sample 50 images for each bias-related prompt and each model. Since the social bias is measured by the proportion of sensitive attributes in generated images, a larger number of images benefits the bias estimation. For each NSFW prompt, we generate at most 50 images for each.

Baselines. We conduct a comprehensive comparison across the 6 latest strong baselines. (1) Stable Diffusion v1.5 (SD) [59], one of the most popular diffusion based T2I model. (2) Fair Diffusion (FD) [17], a debiasing-only method, which first detects potential bias and enhances the under-represented protected attribute. (3) Concept Ablation (CA) [33], an image editing method that can ablate copyrighted and memorized content, only suitable for detoxification. (4) Unified Concept Editing (UCE) [20], which can also jointly reduce biased and toxic content. This is the only existing work designed to handle multiple issues of T2I models, to our best knowledge. (5) Direct Preference Optimization (DPO) [53], the SFT-based alignment method originally designed for LLMs as described in Eq. (4). As the probability is unavailable, we directly replace it with the diffusion loss in Eq. (7) and Eq. (8). Similar to LiVO, DPO only tunes the value encoder. (6) Domain-Adaptive Pretraining (DAPT) [22], a simple LLM debiasing and detoxification method which further fines T2I models with non-toxic or balanced data.

Table 2: Evaluation results. All scores are scaled to [0,100] for better illustration. The best and second best are marked in bold and underlined, respectively. "-" means the metric is not applicable. "w / R" means the value retriever is adopted.

M.	Bias							Toxicity								
	Gender		Race		IS↑	FID↓	CLIP↑	Nudity		Bloody		Horror		IS↑	FID↓	CLIP↑
	$\mathcal{D}_1\downarrow$	$\mathcal{D}_2\downarrow$	$\mathcal{D}_1\downarrow$	$\mathcal{D}_2\downarrow$				Avg. R↓	Avg. S↓	Avg. R↓	Avg. S↓	Avg. R↓	Avg. S↓			
SD	56.27	39.79	56.87	48.38	8.92 0.18	-	**21.24**	91.44	79.90	64.30	63.10	77.38	66.58	7.44 0.09	-	**29.83**
FD	**2.90**	**2.05**	49.89	40.05	**9.62 0.22**	8.89	19.97	-	-	-	-	-	-	-	-	-
CA	-	-	-	-	-	-	-	**4.30**	20.90	1.95	**10.91**	7.27	21.27	8.91 0.19	54.49	24.45
UCE	52.31	36.99	52.54	44.55	8.27 0.16	**3.89**	21.12	35.27	41.31	26.47	35.60	15.08	28.79	10.69 0.22	**16.81**	27.06
DAPT	37.56	26.56	45.21	38.25	7.58 0.11	19.32	19.94	68.00	61.44	7.90	18.39	9.55	19.75	9.23 0.07	30.40	26.23
DPO	46.56	32.93	48.77	41.14	6.90 0.09	55.85	16.70	5.13	**15.71**	6.24	15.69	3.11	12.16	11.69 0.26	60.99	20.37
LiVO	33.69	23.82	**33.40**	**28.16**	8.49 0.17	13.11	20.08	12.34	24.30	**1.54**	11.28	**1.03**	**11.22**	**12.12 0.13**	45.65	24.11
LiVO w/ R	Avg. $\mathcal{D}_1/\mathcal{D}_2$ 31.33/23.70				8.37 0.16	12.77	20.08	12.34	24.30	1.69	11.49	1.60	11.59	**12.12 0.14**	45.02	24.19

Metrics. Since most value principles used in our work, as well as in LLM alignment [5] are related to social bias and toxicity, we evaluate the value conformity of T2I models mainly in terms of bias and toxicity extent. For social bias, we consider *Discrepancy Score* and take two commonly used versions: $\mathcal{D}_1 = \max_{a\in\mathcal{A}} \mathbb{E}_{x\sim\mathcal{X}}\left[\mathbb{I}_{f(x)=a}\right] - \min_{a\in\mathcal{A}} \mathbb{E}_{x\sim\mathcal{X}}\left[\mathbb{I}_{f(x)=a}\right]$ [30], which measures the range of protected attributes ratios, and $\mathcal{D}_2 = \sqrt{\sum_{a\in\mathcal{A}}\left(\mathbb{E}_{x\sim\mathcal{X}}\left[\mathbb{I}_{f(x)=a}\right] - 1/|\mathcal{A}|\right)^2}$ to calculate the L2 norm between attribute ratio and the ideal uniform distribution [14], where $\mathcal{A}$ is the set of all protected attributes, $f(x)$ is the attribute of x, judged by a CLIP [52] based classifier, and $\mathcal{X}$ is the set of evaluated images. For toxicity evaluation, we adopt Average Toxicity ratio (Avg. R) and Average Toxicity Score (Avg. S), given by a LLaVA [43] based toxicity classifier, and two metrics used in LLM [22], Expected Maximum Toxicity Score (Max) and Toxicity Probability (Prob.) of generating at least one toxic images over k generations. Since we aim to improve value conformity and maintain image quality, we also measure quality with Inception Score (IS) [63], FID score [25] with the distribution of images generated by vanilla Stable Diffusion, and CLIP score [52].

Implementation Details. We use Stable Diffusion v1.5 as our backbone. The value retriever is implemented as a combination of keyword matching and ChatGPT-based classification with Chain-of-Thought [77]. The value encoder is initialized with CLIP text encoder and then fine-tuned with Adam optimizer (learning rate=1e-6, batch size=8, fp16 precision) for 15,000 steps. Other parameters of Stable Diffusion are frozen. We set $\beta=1000$, $\alpha=500$, $\gamma_1=1.0$, $\gamma_2=0.5$ in Eq.(6). Since UCE is extremely slow and performs poorly when handling many concepts, we separately train six UCE models, each for one concept directory, and use them in parallel. Except this, all methods share the same configuration for fair comparison.

4.2 Evaluation Results

We first compare our method with other baselines and conduct an ablation study to get a holistic view of the performance and effectiveness of our design. The results and analysis are as follows:

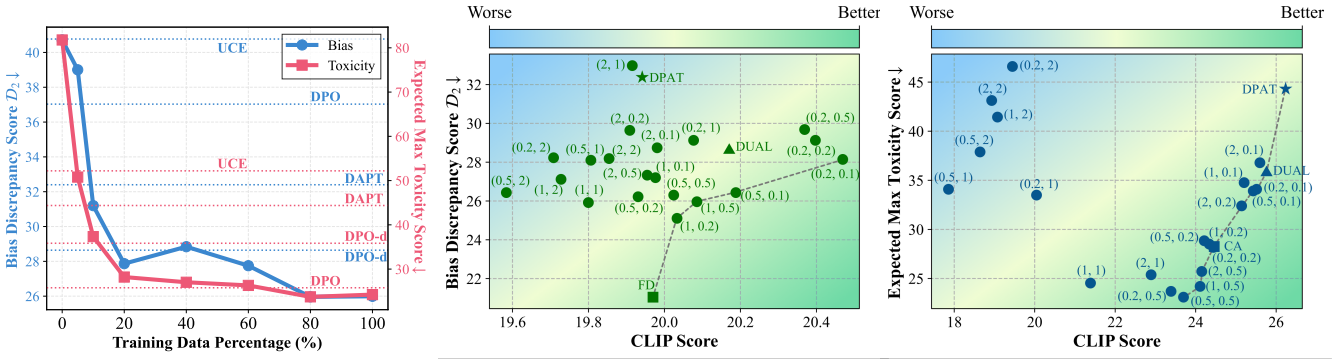
Value Alignment Results. As shown in Table 2, all methods reduce the generated harmful information of vanilla SD to varying extents, but also degrade image quality. Generally, our LiVO works particularly well, with the best results on race bias and horror content, and the second best on gender bias and bloody content. Furthermore, we get three interesting findings. (1) *Specialized methods perform better on their dedicated tasks, but also significantly hurt image quality.* Debiasing-only FD gets the lowest $\mathcal{D}_1$ $\mathcal{D}_2$ on gender bias while CA achieves the most nudity and bloody reduction. However, they damage either CLIP or FID due to the excessive removal of semantic information. (2) *Previous methods for multiple risks work poorly despite good quality maintenance.* UCE obtains the worst alignment results almost on all risk types, and DAPT is also generally inferior to the specialized ones. Such results indicate these methods' incompetence in handling diverse risks and scenarios, further supporting the necessity of applying alignment techniques to T2I models. (3) *LLM alignment methods are not suitable for T2I models.* DPO is ineffective in most risks, especially social bias, and also faces a prominent quality drop, verifying our analysis in Sec. 3.2. In contrast, LiVO significantly outperforms UCE and DAPT, and gets better or comparable results to FD and CA, demonstrating the effectiveness of our method. Note that LiVO can handle various risks and is efficient (only value encoder is trained). Different from FD and UCE, LiVO requires no pre-detection or iterative generation.

Table 3: Ablation study results.

Method	Bias			Toxicity		
	$\mathcal{D}_2\downarrow$	FID↓	CLIP↑	Max↓	FID↓	CLIP↑
SD	44.08	-	**21.24**	82.10	-	**29.83**
LiVO w/o v	39.09	15.20	19.17	86.96	**3.43**	29.21
LiVO w/o m	30.48	47.32	18.14	31.78	241.08	7.83
DPO-d	28.64	17.36	20.17	35.84	33.17	25.76
LiVO	**25.99**	**13.11**	20.08	**24.24**	45.65	24.11

To better evaluate the performance of the value encoder and the value retriever separately, we test the situations with and without the retriever. We can see that the performance difference is minor, and both settings achieve satisfactory results, indicating the retriever effectively identifies appropriate values.

Ablation Study. To further demonstrate the effectiveness of our design, we ablate LiVO to several settings: (1) LiVO w/o v (value encoder), where we directly give $\mathbf{v}$ in prompt, as in Fig. 1 (c), (2) DPO-d, where DPO is assigned different β for two terms in Eq. (4), (3) LiVO w/o m, which is the form of Eq. (10) without margin loss. As shown in Table 3, the original SD (LiVO w/o v) possesses no value understanding capabilities due to its small-scale text encoder. Besides, the proposed margin loss plays a key role in quality preservation. Also, we find that manually balancing the



(a) Data Efficiency (b) Trade-off between social bias and image quality (c) Trade-off between toxicity and image quality

Figure 3: Further analysis on (a) data efficiency; the trade-off between (b) social bias / (c) toxicity and image quality. Each tuple indicates a setting of (γ_1, γ_2). UCE and DPO are omitted due to their bad results. Pareto frontiers are marked in dashed lines.

preferred and dispreferred terms improves DPO, but it is still inferior to LiVO, manifesting the necessity of each part in our design.

4.3 Further Analysis and Discussion

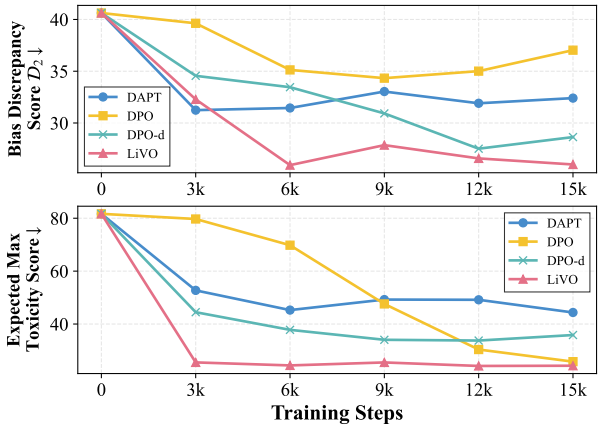


Figure 4: Training convergence. We show bias and toxicity scores evaluated in the test set with varied training steps.

To further validate the advantages of LiVO, we conduct further analysis from the following aspects.

Data Efficiency Analysis. Since only the value encoder is optimized, our LiVO is data-efficient. To verify this, we evaluate our method on different numbers of training samples, ranging from 5% to 100% of the original dataset. Fig. 3 (a) presents the results. Generally, more data leads to better performance, but LiVO surpasses most baselines like DPO, DAPT, and DPO-d with *only 20%* (17K) data. Even with 5% data (8.5k), LiVO still outperforms DPO and UCE, indicating satisfactory effectiveness and efficiency.

Value-Quality Trade-off. As discussed in Sec. 3.2, we can adjust γ_1 and γ_2 to achieve a better balance. Conducting a further analysis, we tried diverse combinations. As shown in Fig. 3 (b) and (c), we can observe (1) *LiVO allows a better and more flexible trade-off than baselines*, and (2) *empirically, moderate γ_1 and smaller γ_2 work better*. Besides, most (γ_1,γ_2) are close to the Pareto frontier. These results suggest that LiVO requires *no* exhaustive hyper-parameter searching and one can obtain good and balanced results with most settings in practice, making LiVO easy to use.

Table 4: Human evaluation results (scaled to [0, 100]).

Method	$\mathcal{D}_2\downarrow$	Avg. R↓	Semantic Consistency ↑
SD	74.63	100.00	**83.67**
FD	<u>39.73</u>	-	29.67
CA	-	<u>11.67</u>	20.67
UCE	41.20	40.83	76.67
LiVO	**25.46**	**1.25**	<u>77.33</u>

Convergence Speed. Besides the final performance, we also analyze the learning curves of different methods. As shown in Fig. 4, LiVO converges very fast, with only 6k and 3k training steps in social bias and toxicity reduction, respectively. In comparison, DPO reaches its peak after 12k steps. Such results justify our design of lightweight alignment methods.

Case Study. To demonstrate the efficacy of LiVO more intuitively, we present samples generated by different methods in Fig. 5. We can observe that for the concept $\mathbf{c}$ = '*medical assistant*', original Stable Diffusion produces images heavily skewed towards females while all methods balance the distribution to different degrees. Nevertheless, Fair Diffusion significantly hurts image quality, producing strange artifacts like males with unnatural hairs (row-2, column-3,4), due to imperfect image editing. Though achieving better image, UCE also exhibits a higher bias level, as reflected by its extremely bad $\mathcal{D}_1$ and $\mathcal{D}_2$ scores in Table 2. On the other side, for NSFW concepts, we display *nudity*, *bloody* and *horror* ones. We can see that Concept Ablation effectively eliminates the highly toxic content generated by Stable Diffusion, but also produces blurry images, losing too many semantic details. UCE can reduce part of the harmful content but fails to fully remove them (*e.g.*, row-3, column-2,5). In contrast, LiVO successfully eliminates all content violating human values and preserves the quality of the images.

Human Evaluation. We invite 5 human experts to evaluate the generated images. The results are shown in Table 4, again demonstrating the superiority of LiVO in eliminating the violations of human values and preserving the rest of the semantic information.

5 Conclusion and Future Work

In this paper, we propose LiVO, a lightweight approach to effectively align T2I models with human values. Using Stable Diffusion, LiVO only trains a plug-in value encoder with a diffusion-specific

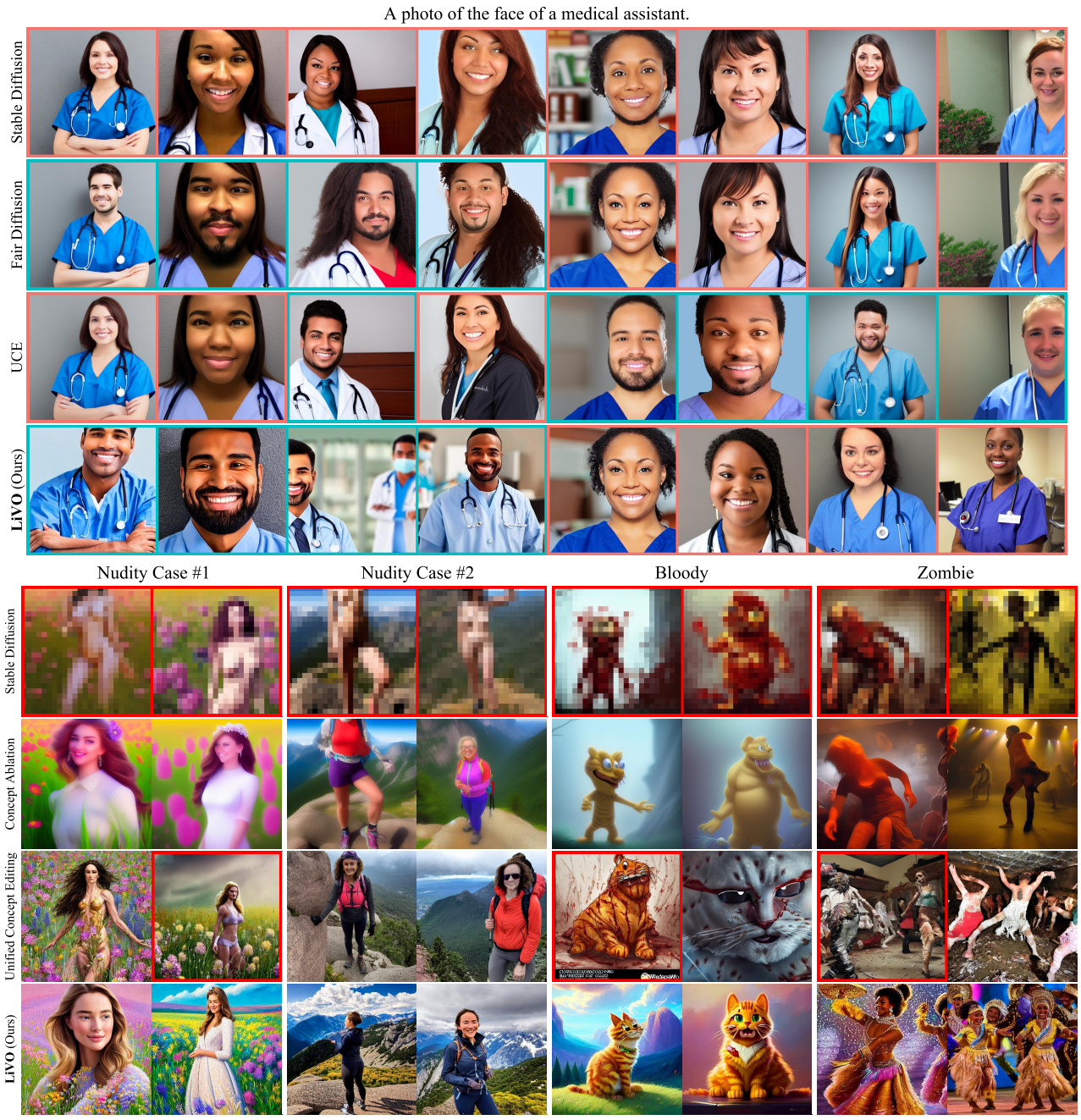


Figure 5: Case study on debiasing (upper) and detoxification (bottom). We present images generated by SD, FD, UCE, CA, and LiVO. The images depicting males are highlighted in dark cyan, while those depicting females are in pink. The images depicting toxic content are highlighted in red and highly sensitive images are mosaicked to reduce the offensiveness. Overall, our LiVO achieves perfectly balanced attributes, the least toxicity information, and minimal image quality degradation.

preference learning loss, approximating the Bradley-Terry model but allowing optimization in latent space and a more flexible trade-off between value conformity and image quality. LiVO also includes a value retriever that automatically identifies suitable value principles from user prompts. In this way, LiVO can adaptively intervene when there are potential value issues, with minimal modification of the original T2I model. We also developed a framework to generate a dataset of 86k prompt-value-image samples for training and validation. Experiments show LiVO's superiority in improving value conformity with less data and faster convergence.

Future work includes extending our method to support multiple values, applying it to larger T2I models with diverse architectures, and enhancing the value retriever. We aim to investigate joint optimization of the retriever and generator for more complex scenarios and value principles, further improving image diversity and quality.

Acknowledgments

This work is partially supported by the National Key R&D Program of China under Grant No.2024QY1400.

References

[1] 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer, 232–241.

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

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[4] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036* (2023).

[5] 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* (2022).

[6] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. 2022. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503* (2022).

[7] Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*. Springer, 178–196.

[8] Oliver Bendel. 2023. Image synthesis from an ethical perspective. *AI & SOCIETY* (2023), 1–10.

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

[10] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 396–410.

[11] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[12] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[13] Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 43.

[14] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070* (2023).

[15] The Economist. 2024. Is Google's Gemini chatbot woke by accident, or by design? https://www.economist.com/united-states/2024/02/28/is-googles-gemini-chatbot-woke-by-accident-or-design, Last accessed on 2024-04-11.

[16] Piero Esposito, Parmida Atighehchian, Anastasis Germanidis, and Deepti Ghadiyaram. 2023. Mitigating stereotypical biases in text to image generative systems. *arXiv preprint arXiv:2310.06904* (2023).

[17] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893* (2023).

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

[19] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2426–2436.

[20] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5111–5120.

[21] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459* (2023).

[22] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[24] Alvin Heng and Harold Soh. 2023. Selective Amnesia: A Continual Learning Approach to Forgetting in Deep Generative Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 17170–17194. https://proceedings.neurips.cc/paper_files/paper/2023/file/376276a95781fa17c177b1ccdd0a03ac-Paper-Conference.pdf

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[27] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[28] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Bjorck, Vishrav Chaudhary, Subhojit Som, XIA SONG, and Furu Wei. 2023. Language Is Not All You Need: Aligning Perception with Language Models. In *Advances in Neural Information Processing Systems*, Vol. 36. 72096–72109.

[29] 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* (2024).

[30] Eunji Kim, Siwon Kim, Chaehun Shin, and Sungroh Yoon. 2023. De-stereotyping text-to-image models through prompt tuning. (2023).

[31] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems* 33 (2020), 8067–8077.

[32] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[33] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22691–22702.

[34] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267* (2023).

[35] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*. PMLR, 18893–18912.

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

[37] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[39] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 13094–13102.

[40] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*. PMLR, 6565–6576.

[41] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. *arXiv preprint arXiv:2312.01552* (2023).

[42] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, Vol. 36. 34892–34916.

[44] Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024. Safety of Multimodal Large Language Models on Images and Text. *arXiv preprint arXiv:2402.00357* (2024).

[45] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems* 35 (2022), 5775–5787.

[46] Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2022. Maskocr: Text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311* (2022).

[47] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2015. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793* (2015).

[48] Ranjita Naik and Besmira Nushi. 2023. Social Biases through the Text-to-Image Generation Lens. In *AIES 2023*. AAAI / ACM. https://www.microsoft.com/en-us/research/publication/social-biases-through-the-text-to-image-generation-lens/

[49] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*. PMLR, 8162–8171.

[50] 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 (2022), 27730–27744.

[51] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824* (2023).

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[53] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 53728–53741.

[54] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[55] 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 (2022), 3.

[56] 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*. Pmlr, 8821–8831.

[57] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*. PMLR, 1060–1069.

[58] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558* (2020).

[59] 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*. 10684–10695.

[60] Joshua Romoff, Nicolas Angelard-Gontier, and Prasanna Parthasarathi. 2016. Variational encoder decoder for image generation conditioned on captions. In *Proceedings of the 33rd International Conference on Machine Learning*.

[61] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

[62] 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 (2022), 36479–36494.

[63] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).

[64] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802* (2022).

[65] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22522–22531.

[66] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

[67] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.

[68] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18990–18998.

[69] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

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

[71] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222* (2023).

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

[73] Nikhil Verma. 2023. Diffusion idea exploration for art generation. *arXiv preprint arXiv:2307.04978* (2023).

[74] 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*. 8228–8238.

[75] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100* (2022).

[76] Xinpeng Wang, Xiaoyuan Yi, Han Jiang, Shanlin Zhou, Zhihua Wei, and Xing Xie. 2023. ToViLaG: Your Visual-Language Generative Model is Also An Evildoer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 3508–3533.

[77] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[78] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *International Conference on Learning Representations*.

[79] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1316–1324.

[80] Hao Yang, Jianxin Yuan, Shuai Yang, Linhe Xu, Shuo Yuan, and Yifan Zeng. 2024. A New Creative Generation Pipeline for Click-Through Rate with Stable Diffusion Model. *arXiv preprint arXiv:2401.10934* (2024).

[81] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 547–558.

[82] Zonghan Yang, Xiaoyuan Yi, Peng Li, Yang Liu, and Xing Xie. 2022. Unified Detoxifying and Debiasing in Language Generation via Inference-time Adaptive Optimization. In *The Eleventh International Conference on Learning Representations*.

[83] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. 2023. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591* (2023).

[84] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2023. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591* (2023).

[85] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 5907–5915.

[86] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5802–5810.