Alignment of Diffusion Models: Fundamentals, Challenges, and Future

BUHUA LIU, The Hong Kong University of Science and Technology (Guangzhou), China
SHITONG SHAO, The Hong Kong University of Science and Technology (Guangzhou), China
BAO LI, Institute of Automation, Chinese Academy of Sciences, China
LICHEN BAI, Tsinghua University, China
ZHIQIANG XU, Mohamed bin Zayed University of Artificial Intelligence, UAE
HAOYI XIONG, Baidu Inc., China
JAMES KWOK, The Hong Kong University of Science and Technology, Hong Kong
SUMI HELAL, The University of Bologna, Italy
ZEKE XIE*, The Hong Kong University of Science and Technology (Guangzhou), China

Diffusion models have emerged as the leading paradigm in generative modeling, excelling in various applications. Despite their success, these models often misalign with human intentions and generate results with undesired properties or even harmful content. Inspired by the success and popularity of alignment in tuning large language models, recent studies have investigated aligning diffusion models with human expectations and preferences. This work mainly reviews alignment of diffusion models, covering advancements in fundamentals of alignment, alignment techniques of diffusion models, preference benchmarks, and evaluation for diffusion models. Moreover, we discuss key perspectives on current challenges and promising future directions on solving the remaining challenges in alignment of diffusion models. To the best of our knowledge, our work is the first comprehensive review paper for researchers and engineers to comprehend, practice, and research alignment of diffusion models. [1]

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Machine learning**; *Computer vision*; Natural language generation.

Additional Key Words and Phrases: Alignment, Diffusion Models, Generative Models

1 INTRODUCTION

Diffusion models [74, 183, 187] have emerged as the dominant paradigm, surpassing previous state-of-the-art generative models such as generative adversarial networks (GANs) [14, 42, 64, 90, 169, 210] and variational autoencoders (VAEs) [98]. Diffusion models have demonstrated the impressive performance and success in various generative tasks, including image generation [51, 92], video generation [9, 73], text generation [127], audio synthesis [84, 101], 3D generation [222, 240], music generation [40], and molecule generation [66, 229]. Fig. 2a illustrates the trend in the number of papers on diffusion models published in top computer vision conferences (CVPR, ICCV, ECCV) and top machine learning conferences (NeurIPS, ICML, ICLR) in recent years, highlighting the growing interest in diffusion models at these leading conferences.

*Corresponding author
[1] https://github.com/xie-lab-ml/awesome-alignment-of-diffusion-models

Authors' addresses: Buhua Liu, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong, China, bryceliu@foxmail.com; Shitong Shao, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong, China, sshao213@connect.hkust-gz.edu.cn; Bao Li, Institute of Automation, Chinese Academy of Sciences, Beijing, China, libao2023@gmail.com; Lichen Bai, Tsinghua University, Beijing, China, blc22@mails.tsinghua.edu.cn; Zhiqiang Xu, Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi, UAE, Zhiqiang.Xu@mbzuai.ac.ae; Haoyi Xiong, Baidu Inc., Beijing, China, xhyccc@gmail.com; James Kwok, The Hong Kong University of Science and Technology, Hong Kong, jamesk@cse.ust.hk; Sumi Helal, The University of Bologna, Bologna, Emilia-Romagna, Italy, sumi.helal@gmail.com; Zeke Xie, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong, China, zekexie@hkust-gz.edu.cn.

However, the diffusion training objective does not necessarily align well with human intentions and preferences. For instance, images generated by pre-trained text-to-image (T2I) models may generate images that, while technically plausible, may fail to capture specific artistic nuances or accurately represent complex textual descriptions [8, 54, 105], which can hinder practical applications or diminish user satisfaction. Similarly, in drug discovery, pre-trained diffusion models typically lack the ability to generate molecules with high binding affinity and structural rationality [66], a critical issue that can directly impact therapeutic efficacy. Fig. 1 provides a conceptual illustration of such misalignment. To address this mismatch, recent works have begun to optimize pre-trained diffusion models directly for human preference or certain desired properties, aiming to more controllable data generation [205] beyond simply modeling the training data distribution.

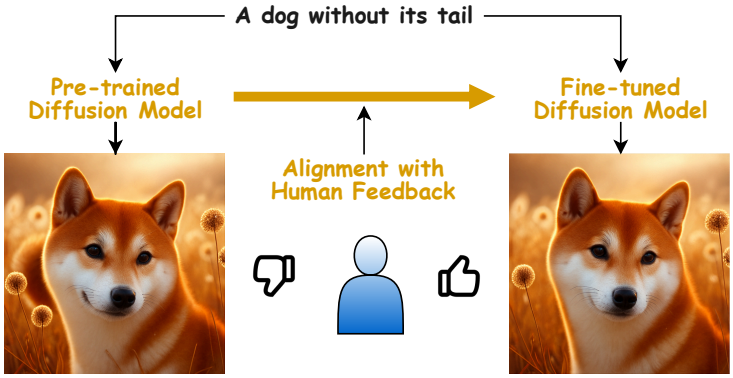
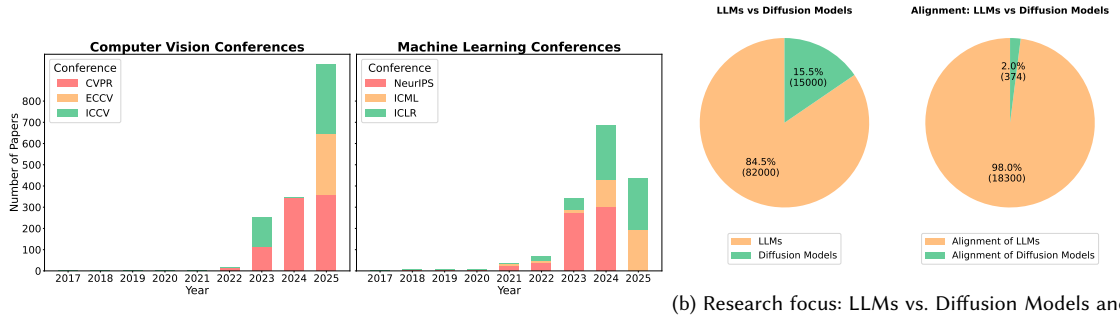


Fig. 1. Conceptual illustration of diffusion model misalignment and the goal of alignment. A pre-trained model may generate an output that deviates from human intentions or desired qualities (e.g., missing details in a prompt). An aligned model, after an alignment process, produces an output that better reflects human preferences for the same input.

Within the community of language modeling, recent powerful large language models (LLMs) like GPT-4 [143], Llama [47, 198], and Qwen 3 [232] are typically trained in two stages. In the first pre-training stage, they are trained on a vast textual corpus with the objective of predicting the next tokens. In the second post-training stage, they are fine-tuned to follow instructions, align with human preferences, and improve capabilities like coding and factuality. The post-training process usually involves supervised fine-tuning (SFT) followed by alignment with human feedback, using techniques such as reinforcement learning from human feedback (RLHF) [143, 146], and direct preference optimization (DPO) [47, 158]. The success of these techniques in LLMs is particularly relevant as they demonstrate the feasibility and effectiveness of aligning complex generative models with nuanced human preferences, thereby providing a validated conceptual and methodological roadmap for diffusion models. LLMs trained using this two-stage process have achieved state-of-the-art performance [35, 143] in various language generation tasks and have been deployed in commercial applications such as ChatGPT.

Inspired by the success of aligning LLMs [209], there is growing interest in better aligning diffusion models with human intentions to enhance their capabilities. Fig. 2b visualizes the paper counts on LLMs and diffusion models, as well as their alignment studies [2]. The left pie chart shows LLMs account for 84.5% of the studies, while diffusion models

[2]Data obtained from Google Scholar as of July 25, 2025.



(a) Diffusion model paper trends at top conferences.

(b) Research focus: LLMs vs. Diffusion Models and their alignment.

Fig. 2. Statistical overview of research trends. (a) The number of papers on diffusion models at top computer vision conferences (CVPR, ECCV, ICCV) and top machine learning conferences (NeurIPS, ICML, ICLR) since 2017, indicating a growing interest. Note that NeurIPS 2025 has not been released as of the date of submission, and ECCV and ICCV are held biennially. (b) Comparison of the ratio of papers on LLMs vs. diffusion models (left pie) and the research focus on alignment within these areas (right pie), highlighting the nascent stage of diffusion model alignment.

account for 15.5%. The right chart highlights that, at this point, 98.0% of alignment studies focus on LLMs, while only 2.0% address diffusion models. This disparity, clearly illustrated in Fig. 2b, not only underscores the relatively nascent stage of alignment research for diffusion models compared to LLMs but also signals a significant opportunity and pressing need for focused investigation in this domain. This survey aims to facilitate such research by consolidating current knowledge and outlining future directions. The fundamental differences between LLMs and diffusion models, where LLMs predict sequences of tokens and diffusion models progressively reverse a noise-adding diffusion process, and where the continuous, high-dimensional nature of image generation versus discrete token prediction in LLMs presents unique challenges for defining preferences and applying feedback, along with their uniquely advantageous application domains, such as LLMs for language generation and diffusion models for image generation, make the study of diffusion model alignment an independent and valuable area of interest.

In this work, we provide a comprehensive review of the alignment of diffusion models to assist researchers and practitioners in understanding how to align these models with human intentions and preferences. A literature list is made publicly available at GitHub. Fig. 3 illustrates the main framework of this survey. Section 2 introduces recent advancements in diffusion models, particularly those incorporating alignment technologies. Section 3 explores fundamental alignment techniques and related challenges in human alignment. Section 4 focuses on alignment techniques specific to diffusion models. Section 5 reviews benchmarks and evaluation metrics for assessing human alignment of diffusion models. Section 6 outlines future research directions. Section 7 concludes our work, summarizing the key findings and their implications for both researchers and practitioners. Our survey provides a thorough understanding of the alignment of diffusion models, identifies research gaps, and informs the development of next-generation models, driving future advancements in the field.

2 AN OVERVIEW OF DIFFUSION MODELS

In this section, we briefly outline recent advancements in diffusion models and elucidate the role that human alignment plays in guiding their development.

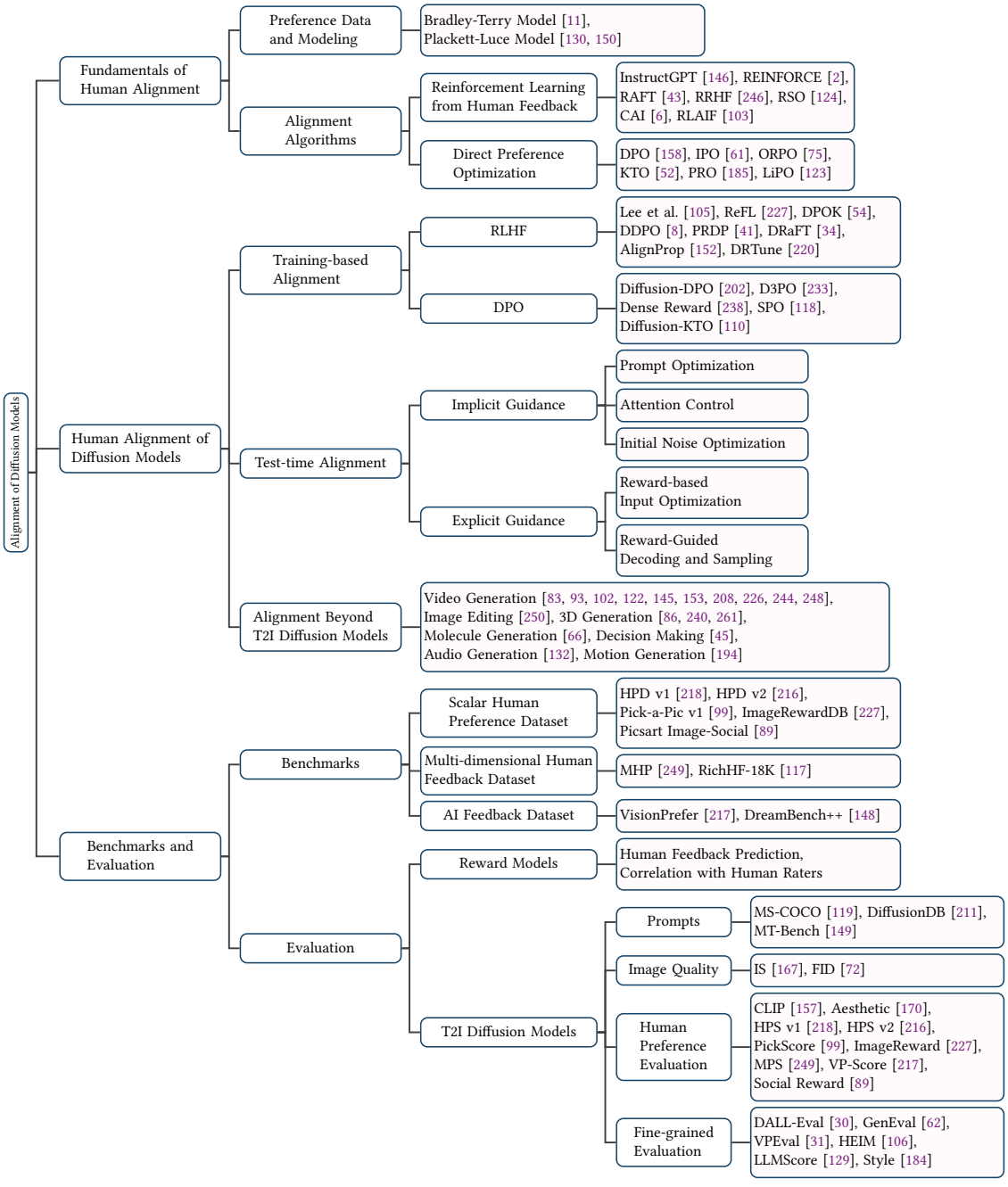


Fig. 3. The framework of this survey in human alignment of diffusion models and beyond.

Decades ago, diffusion process or Langevin diffusion, originated from statistical physics [10, 63, 173], were first introduced in machine learning not for generative modeling but mainly for parameter inference [4] and analyzing

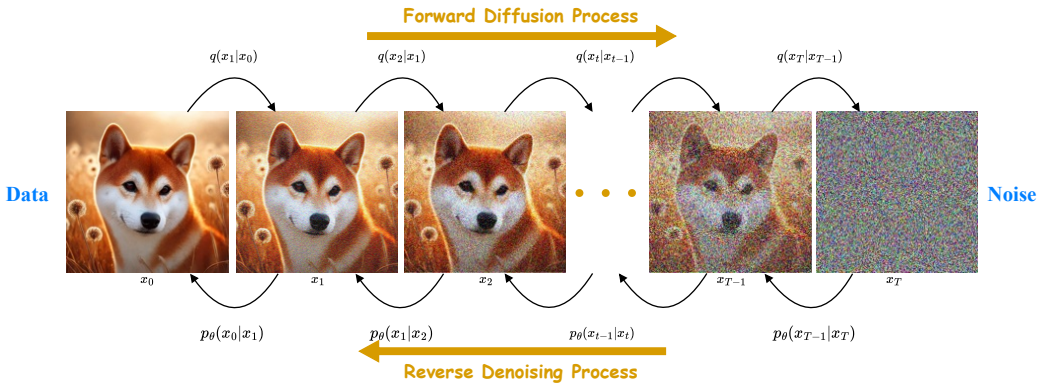


Fig. 4. Diffusion models consist of two key processes: a forward diffusion process with a transition kernel $q(x_t|x_{t-1})$, where noise is gradually added to a data sample, and a reverse denoising process with a learnable transition kernel $p_\theta(x_{t-1}|x_t)$, where the model learns to denoise Gaussian noise to reconstruct the original data sample.

optimization dynamics [29, 224, 230]. Diffusion models, pioneered by Sohl-Dickstein et al. [183] and significantly advanced by Ho et al. [74] with Denoising Diffusion Probabilistic Models (DDPMs), operate through a two-stage process: a forward noising stage that gradually adds noise to data, and a reverse denoising stage that learns to reconstruct data from noise, as illustrated in Fig. 4. This iterative refinement allows for the generation of high-quality samples. Variations like Denoising Diffusion Implicit Models (DDIMs) [186] further improved sampling efficiency. For detailed reviews, see Cao et al. [16], Yang et al. [235].

A key application area for diffusion models is T2I synthesis. Early T2I models like GLIDE [141] adapted diffusion processes for this task but operated in the pixel space, leading to high computational costs [166]. The advent of Latent Diffusion Models (LDMs) [165], which perform diffusion in a compressed latent space, marked a significant milestone by reducing computational demands and improving efficiency, making T2I models more practical.

However, even with these advancements, a core challenge emerged: the standard diffusion training objective does not inherently guarantee alignment with nuanced human intentions and preferences. This often results in generated images that, while technically sound, may not fully meet user expectations or desired aesthetics [7, 25, 151]. Recognizing this gap, recent cutting-edge T2I models have explicitly integrated human alignment techniques. For example, Stable Diffusion 3 (SD3) [51] not only introduced architectural innovations but also crucially incorporated alignment by applying Diffusion Direct Preference Optimization (Diffusion-DPO) [202] to its large base models. This alignment step is pivotal in achieving state-of-the-art performance, surpassing other open models and even proprietary ones like DALLE-3 [7] on benchmarks such as GenEval [62]. Similarly, SD3-Turbo [168], focuses on efficient high-resolution generation, also leverages DPO-finetuned models in its distillation process, demonstrating significant improvements in human preference evaluations. These developments underscore that human alignment is no longer an afterthought but a central component in advancing the capabilities of diffusion models.

This trend highlights the increasing importance of human feedback and sophisticated alignment methodologies in shaping the next generation of diffusion models, moving beyond mere generation quality to achieve outputs that are more accurate, desirable, and aligned with human values. This survey delves into these critical alignment techniques.

Table 1. The list of symbols.

Symbols	Meanings
$\mathcal{L}$	the loss function for optimization
c	the prompt to LLMs or diffusion models
ρ	the distribution of prompt
x	the response of LLMs or diffusion models
K	the number of candidate responses
x^w	the winning/preferred response in the paired responses
x^l	the losing/dis-preferred response in the paired responses
p_{BT}	the probability distribution of human preference under the Bradley-Terry model
p_{PL}	the probability distribution of human preference under the Plackett-Luce model
L	the total number of tokens in the responses for LLMs
T	the total number of denoising steps for diffusion models
θ	the parameters of LLMs or diffusion models
q	the image data distribution
p_θ	the policy in RL, parameterized by θ, i.e., the LLMs or diffusion models to be aligned
p_{ref}	the reference policy, which is typically the frozen initial policy
$r_\phi(c, x)$	the reward model output given the input prompt c and response x, parameterized by ϕ
$\mathcal{D}$	the pre-collected preference dataset
D_{KL}	the Kullback–Leibler divergence
β	the hyper-parameter, which regularizes the distance between the current and reference policies

3 FUNDAMENTALS OF HUMAN ALIGNMENT

In this section, we discuss the fundamentals of human alignment based on the existing literature for aligning LLMs and diffusion models. Specifically, we summarize the general data forms and preference modeling methods for alignment in Section 3.1. We outline the general alignment algorithms for human alignment in Section 3.2. We discuss key challenges in human alignment in Section 3.3.

3.1 Preference Data and Modeling

Preference Data. In general, preference data consists of three elements: prompts, responses[3], and feedback. Table 1 shows the mathematical notations in this work.

Preference data are typically composed of prompts, responses, and feedback. Prompts, whether human-provided or AI-generated, are diverse inputs used to elicit model responses. Responses can be generated on-policy (from the current model) or off-policy (from external or older models), presenting a trade-off between relevance and collection efficiency [192].

Preference Modeling. The most common feedback form is a pairwise preference between a preferred response x^w and a dis-preferred response x^l, given a prompt c. This is often modeled using a reward model r_ϕ trained under the Bradley-Terry (BT) framework [11]. Specifically, the reward model r parameterized with ϕ takes in a prompt c and a response x, outputting a scalar reward $r_\phi(c, x)$. The BT model assumes that the human preference probability p_{BT} can be expressed as:

$$p_{\text{BT}}(x^w > x^l|c) = \frac{\exp(r^*(c, x^w))}{\exp(r^*(c, x^w)) + \exp(r^*(c, x^l)))} = \sigma(r^*(c, x^w) - r^*(c, x^l)), \tag{1}$$

[3]We use the term "responses" broadly to include human-collected data samples beyond model responses.

where r^* represents the optimal reward model that r_ϕ approximates, and $\sigma(x) = 1/(1+\exp(-x))$ is the logistic function. The model is optimized with a loss function that maximizes the log-probability of the preferred sample having a higher reward score:

$$\mathcal{L}_{\text{RM-BT}}(\phi) = -\mathbb{E}_{(c,x^w,x^l)\sim\mathcal{D}}\left[\log(\sigma(r_\phi(c,x^w) - r_\phi(c,x^l)))\right], \tag{2}$$

where $(c, x^w, x^l) \sim \mathcal{D}$ denotes the sampling of prompt c, the preferred response x^w, and the dis-preferred response x^l from the collected dataset $\mathcal{D}$ labeled by humans or AI. In essence, Eq. (2) represents a cross-entropy loss where pairwise comparisons are treated as labels, with x^w labeled as 1 and x^l as 0. The term $\sigma(r_\phi(c,x^w) - r_\phi(c,x^l))$ represents the probability that response x^w will be preferred over response x^l by a human labeler, as modeled by Eq. (1).

Beyond pairwise comparisons, feedback can be listwise (ranking multiple responses) or binary (a single response is desirable/undesirable) [52, 164]. Listwise feedback can be modeled by extending the BT model to the Plackett-Luce (PL) model [130, 150], which frames alignment as a ranking problem [123, 185, 246]. Specifically, the PL model stipulates that when presented with a set of possible choices, people prefer each choice with a probability proportional to the value of some underlying reward function. In our context, the policy p is given a prompt c and produces a set of K responses $(x_1, x_2, \ldots, x_K) \sim p(x|c)$. A human then ranks these responses, yielding a permutation $\tau : [K] \to [K]$, where $[K] = \{1, 2, \ldots, K\}$ indexes the responses and $\tau(i) = j$ indicates that the response x_j is ranked at position i. The PL model assumes that the human preference ranking probability p_{PL} can be formulated as:

$$p_{\text{PL}}(\tau|x_1, x_2, \ldots, x_K, c) = \prod_{k=1}^{K} \frac{\exp(r^*(c, x_{\tau(k)}))}{\Sigma_{j=k}^{K} \exp(r^*(c, x_{\tau(j)}))}. \tag{3}$$

Notably, when $K = 2$, Eq. (3) reduces to the BT model in Eq. (1). The loss function for training the reward model on listwise feedback under the PL model typically uses a maximum likelihood estimation (MLE) ranking loss [123, 158, 221]:

$$\mathcal{L}_{\text{RM-PL}}(\phi) = -\mathbb{E}_{c,x_1,x_2,\ldots,x_K,\tau}\left[\log \prod_{k=1}^{K} \frac{\exp(r_\phi(c, x_{\tau(k)}))}{\Sigma_{j=k}^{K} \exp(r_\phi(c, x_{\tau(j)}))}\right]. \tag{4}$$

Feedback Sources and Timing. While traditionally sourced from humans, the cost and effort of annotation have motivated the use of AI-generated feedback, with powerful models like GPT-4 being used as annotators [39, 189, 217]. Furthermore, feedback can be collected offline from a static dataset or online during training, which distinguishes between offline and online alignment settings [44, 107, 195].

3.2 Alignment Algorithms

In this subsection, we introduce general alignment algorithms.

3.2.1 Reinforcement Learning from Human Feedback. Alignment with human preferences is typically achieved through RLHF, which first trains an explicit reward model to reflect human preferences and then applies RL methods to optimize a policy toward maximizing the reward provided by the reward model [32]. RLHF was successfully applied by Ouyang et al. [146] to fine-tune instruction-following LLMs, leading to the development of the widely-used ChatGPT.

Specifically, the policy p_θ is fine-tuned to maximize the reward $r_\phi(c, x)$ while being regularized by the KL divergence from an initial reference policy p_{ref}:

$$\max_{p_\theta} \mathbb{E}_{c\sim\rho, x\sim p_\theta(x|c)}\left[r_\phi(c,x) - \beta D_{\text{KL}}(p_\theta(x|c)||p_{\text{ref}}(x|c))\right], \tag{5}$$

where β controls the strength of the KL regularization term [190].

Proximal Policy Optimization (PPO) [171] is the predominant RL algorithm for this task, but it is computationally expensive and notoriously difficult to tune [50, 146]. This has motivated simpler alternatives, such as basic REINFORCE-style optimization [2, 192] or iterative fine-tuning methods that bypass traditional RL entirely by using the reward model to filter or rank data for supervised fine-tuning [43, 124, 246].

3.2.2 Direct Preference Optimization. DPO offers a simpler paradigm by bypassing explicit reward model training and optimizing the policy directly on preference data [158]. By re-parameterizing the RLHF objective (Eq. (5)) in terms of the optimal policy, DPO derives a direct loss on pairwise preferences:

$$\mathcal{L}_{\text{DPO}}(p_\theta; p_{\text{ref}}) = -\mathbb{E}_{(c,x^w,x^l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{p_\theta(x^w|c)}{p_{\text{ref}}(x^w|c)} - \beta\log\frac{p_\theta(x^l|c)}{p_{\text{ref}}(x^l|c)}\right)\right]. \tag{6}$$

Several variants have been proposed to address DPO's limitations, including its tendency to overfit to the offline preference dataset, its sensitivity to hyperparameters, the need for a separate reference model, and its restriction to pairwise preference data. Identity Preference Optimization (IPO) [61] modifies the objective with a robust regularization term to mitigate overfitting to the preference dataset. Odds Ratio Preference Optimization (ORPO) [75] eliminates the need for a separate reference model and combines standard supervised fine-tuning on preferred responses with preference alignment. Other approaches use different feedback formats entirely. Kahneman-Tversky Optimization (KTO) [52] requires only binary feedback (desirable/undesirable) instead of pairs, leveraging principles from human decision-making theory. Preference Ranking Optimization (PRO) [185] and Listwise Preference Optimization (LiPO) framework [123] extend the paradigm to leverage listwise feedback, where multiple responses are ranked.

3.3 Challenges of Human Alignment

In this subsection, we synthesize key challenges in human alignment.

Alignment with AI Feedback. Using AI-generated feedback, or Reinforcement Learning from AI Feedback (RLAIF), is a prominent strategy to circumvent costly human annotations [6, 103]. This involves models improving themselves via self-generated rewards [112, 245] or using powerful proxy models (e.g., LLMs) as annotators [48, 217]. However, this introduces a core trade-off between cost-effectiveness and the risks of AI feedback, namely inheriting biases, lacking diversity, and potential model collapse from recursively training on synthetic data [178].

Diverse and Changing Human Preferences. A key challenge is modeling the diverse, dynamic, and often conflicting nature of human preferences [17]. Current research addresses this by learning distributions over preferences [18], developing pluralistic frameworks [36, 188], and using multi-objective alignment for conflicting goals [159, 236]. While these methods promote fairness [176], they also increase algorithmic complexity and introduce the difficulty of balancing competing values.

Distributional Shift. Reliance on static, offline preference data creates a distributional shift between the training data and the evolving policy, a known challenge in offline RL [107] that leads to over-optimization and reward hacking [59, 180]. KL regularization, a common mitigation, can be overly restrictive and limit model improvement. Future work could draw from fields like OOD generalization [121], causality [172], and uncertainty estimation [58, 95] to enhance robustness.

Efficiency of Alignment. Improving alignment efficiency is pursued via data-centric and algorithmic approaches. Data-centric methods focus on achieving strong performance with less data, such as by curating small, high-quality instruction sets [126, 258]. Algorithmic innovations aim to reduce computational overhead through techniques like

linear alignment [60] and feedback-efficient exploration [200]. The primary challenge is to enhance efficiency without sacrificing alignment quality or introducing data selection biases. Future work may explore dataset distillation [206, 243], parameter-efficient fine-tuning [134], and inference-time scaling [181].

Alignment with Rich Rewards. A single, terminal reward is often sparse and overlooks the sequential generation process, causing optimization instability [1, 182]. To address this, research explores richer, denser reward structures, such as step-wise preferences for diffusion models [118, 238] and token-level feedback for LLMs [20, 247]. However, aligning with richer rewards inevitably increases both computational and algorithmic complexities, necessitating further research to address potential scalability issues and to understand how to best design and utilize these complex reward structures.

Understanding of Alignment. Research is ongoing to understand the mechanisms, theoretical properties, and limitations of alignment. Comparative analyses examine the trade-offs between dominant paradigms like RLHF and DPO, exploring differences in their optimization behavior and performance [91, 231]. Theoretical studies are also emerging to formalize the learning dynamics of alignment [87, 225] and analyze fundamental model-level flaws, such as the DPO loss function's potential lack of a unique MLE [78]. Finally, trustworthiness remains a major concern, with research highlighting model vulnerabilities to jailbreak attacks [215, 242], the brittleness of safety alignment [154, 212], and the negative impact of noisy preference data [161]. Collectively, these studies underscore the gap between current methods and the goal of robust, understandable, and truly aligned AI systems.

4 HUMAN ALIGNMENT TECHNIQUES OF DIFFUSION MODELS

In this section, we first introduce training-based human alignment techniques of diffusion models, including RLHF and DPO in Section 4.1 and Section 4.2, respectively. We then review test-time alignment techniques in Section 4.3. Furthermore, we review studies related to alignment beyond T2I diffusion models in Section 4.4. Finally, we discuss challenges of diffusion alignment in Section 4.5.

RLHF and DPO are two very classic training-based techniques for aligning AI models with human preferences. However, when applied to diffusion models, these methods encounter significant challenges due to the step-by-step training and sampling nature of diffusion models. Specifically, aligning diffusion models with preference optimization requires sampling across all possible diffusion trajectories leading to x_0, which is intractable in practice. While the LLM response is treated as a single output, diffusion models' multiple latent image representations of each step need to be calculated and stored, leading to unreasonable high memory consumption and low computation efficiency. This makes these methods impractical for large-scale diffusion models. To address the high computational overhead associated with adapting alignment techniques to diffusion models, researchers often formulate the denoising process as a multi-step Markov decision process (MDP). The proposed diffusion alignment methods need to directly optimize the expected reward of an image output or update the policy based on human preferences to approximately perform policy gradient guided by a reward model. This formulation enables parameter updates at each step of the denoising process based on human preferences, thereby circumventing the significant computational costs.

4.1 Reinforcement Learning from Human Feedback of Diffusion Models

In this subsection, we present the RLHF paradigm and its extension for diffusion alignment. As shown in Fig. 5 (a), RLHF typically involves three progressive stages: data collection, developing a reward model, and reinforcement learning. In the data collection stage, preferences of prompt-response pairs (e.g., text-image pairs for T2I diffusion models) are gathered from humans or AI. In the second stage, RLHF develops a reward model $r_\phi(c, x)$, either through training

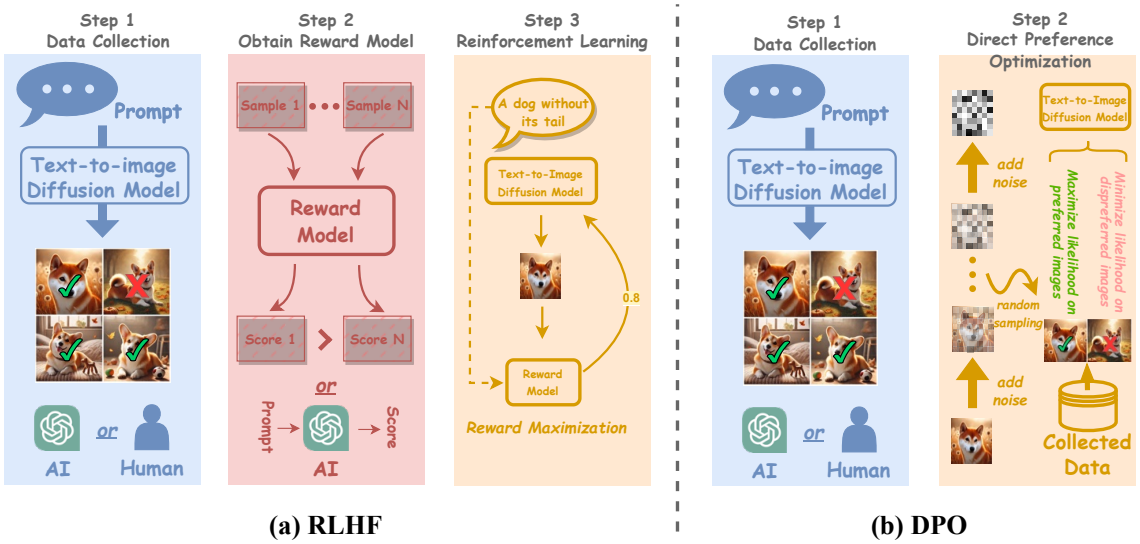


Fig. 5. The overview of RLHF and DPO of diffusion models.

or prompt engineering [129]. The trained reward model for diffusion models is often instantiated as a VLM such as CLIP [157] or BLIP [109] and typically trained with Eq. (2) [99, 216, 227] on $\mathcal{D}$ to model human preferences. Finally, RLHF optimizes the diffusion model $p_\theta(x_0|c)$ to maximize the reward $r_\phi(c, x_0)$ given the prompt distribution $c \sim \rho$ (ignoring the regularization term):

$$\min_\theta \mathbb{E}_{c\sim\rho, x_0\sim p_\theta(x_0|c)} \left[-r_\phi(c, x_0)\right]. \tag{7}$$

There are several classical and emerging approaches to approximately optimizing the objective in Eq. (7). They can be broadly categorized into reward-weighted fine-tuning, RL fine-tuning, and direct reward fine-tuning, alongside newer paradigms.

Reward-weighted Fine-tuning. Lee et al. [105] proposed to align diffusion models with human feedback with a reward-weighted likelihood maximization objective:

$$\min_\theta \mathbb{E}_{(c,x_0)\sim\mathcal{D}_{\text{model}}} \left[-r_\phi(c, x_0) \log p_\theta(x_0|c)\right] + \beta \mathbb{E}_{(c,x_0)\sim\mathcal{D}_{\text{pre-training}}} \left[-\log p_\theta(x_0|c)\right], \tag{8}$$

where $(c, x_0) \sim \mathcal{D}_{\text{model}}$ is the model-generated dataset by diffusion models on the tested text prompts, and $\mathcal{D}_{\text{pre-training}}$ is the pre-training dataset. The first term in Eq. (8) minimizes the reward-weighted negative log-likelihood (NLL) on $\mathcal{D}_{\text{model}}$ to improve the image-text alignment of the model. The second term in Eq. (8) minimizes the pre-training loss to mitigate overfitting to $\mathcal{D}_{\text{model}}$. Black et al. [8] pointed out that Eq. (8) can be performed for multiple rounds of alternating sampling and training to make it into an online RL method by replacing $(c, x_0) \sim \mathcal{D}_{\text{model}}$ with $c \sim \rho, x_0 \sim p_\theta(x_0|c)$. They referred to this general class of reward-weighted methods as reward-weighted regression (RWR), and considered two weighting schemes: 1) a standard one that uses exponentiated rewards to ensure nonnegativity, $w_{\text{RWR}}(c, x_0) = \frac{1}{\mathcal{Z}_{RWR}} \exp(\gamma r_\phi(c, x_0))$, where γ is an inverse temperature and $\mathcal{Z}_{RWR}$ is a normalization constant; and 2) a simplified one that uses binary weights $w_{\text{sparse}}(c, x_0) = \mathbb{I}\left[r_\phi(c, x_0) \geq C\right]$, where C is a reward threshold determining which samples are used for training and $\mathbb{I}$ is the indicator function. Notably, from the RL literature, a weighted log-likelihood objective by w_{RWR} is known to approximately solve Eq. (7) subject to a KL divergence constraint on $p_\theta(x_0|c)$ [139].

RL Fine-tuning. Reward-weighted fine-tuning relies on an approximate log-likelihood because it ignores the sequential nature of the diffusion denoising process, only using the final samples x_0. To address this, the denoising process is treated as a multi-step decision-making problem [8, 54, 199], using exact likelihoods at each denoising step instead of the approximate likelihoods from the full denoising process. This allows us to directly optimize Eq. (7) using policy gradient algorithms. Black et al. [8] proposed denoising diffusion policy optimization (DDPO) to maximize rewards from various reward models, including image compressibility, aesthetic quality, and image-prompt alignment. They demonstrated that DDPO is more effective than reward-weighted likelihood approaches. DDPO has two variants. One uses REINFORCE [137, 214], a score function policy gradient estimator:

$$\mathbb{E}_{c\sim\rho,x_{0:T}\sim p_\theta(x_{0:T}|c)}\left[-r_\phi(c,x_0)\Sigma_{t=1}^{T}\nabla_\theta\log p_\theta(x_{t-1}|x_t,c)\right]. \tag{9}$$

Another variant uses an importance sampling estimator to reuse old trajectories (i.e., prompt-image pairs) with a PPO-style clipping objective [171] to stabilize training:

$$\mathbb{E}_{c\sim\rho,x_{0:T}\sim p_{\text{ref}}(x_{0:T}|c)}\left[-r_\phi(c,x_0)\Sigma_{t=1}^{T}\text{clip}\left(\frac{p_\theta(x_{t-1}|x_t,c)}{p_{\text{ref}}(x_{t-1}|x_t,c)}\nabla_\theta\log p_\theta(x_{t-1}|x_t,c),1-\epsilon,1+\epsilon\right)\right], \tag{10}$$

where ϵ is the clip hyperparameter.

Fan et al. [54] introduced Diffusion Policy Optimization with KL regularization (DPOK), an online RL fine-tuning algorithm that maximizes the ImageReward score with KL regularization. Compared to DDPO, DPOK [54] further employs KL regularization to Eq. (7), resulting in the objective:

$$\min_\theta \mathbb{E}_{c\sim\rho,x_0\sim p_\theta(x_0|c)}\left[-r_\phi(c,x_0)+\beta D_{\text{KL}}(p_\theta(x_0|c)||p_{\text{ref}}(x_0|c))\right]. \tag{11}$$

DPOK then utilizes an upper bound of the KL-term to derive the following objective for regularized training:

$$\min_\theta \mathbb{E}_{c\sim\rho,x_{0:T}\sim p_\theta(x_{0:T}|c)}\left[-r_\phi(c,x_0)\right]+\beta\Sigma_{t=1}^{T}\mathbb{E}_{x_t\sim p_\theta(x_t|c)}\left[D_{\text{KL}}(p_\theta(x_{t-1}|x_t,c)||p_{\text{ref}}(x_{t-1}|x_t,c)\right]. \tag{12}$$

However, policy gradient methods like DDPO are known for their high variance and instability, especially in large-scale settings. To address this, Deng et al. [41] proposed Proximal Reward Difference Prediction (PRDP), which reframes the RL objective as a more stable, supervised reward difference prediction task. Instead of directly estimating policy gradients, PRDP trains the model to predict the reward difference between two generated images, proving that a model which perfectly predicts this difference effectively maximizes the original RL objective.

Direct Reward Fine-tuning. RL fine-tuning methods are flexible because they do not require differentiable rewards. However, many reward models are differentiable, such as ImageReward, PickScore [99], and HPSv2 [216], providing analytic gradients. In such cases, using RL can discard valuable information from the reward model. To address this, end-to-end backpropagation from reward gradients to the diffusion model parameters has been proposed to solve Eq. (7). Nevertheless, updating the diffusion model throughout the entire denoising process is memory-intensive, as storing partial derivatives of all layers and denoising steps is prohibitive. ReFL [227] was the first to backpropagate through a differentiable reward model by evaluating the reward on a one-step predicted image $r(c,\hat{x_0})$ from a randomly selected step t, thus bypassing the full denoising process. In contrast, Alignment by Backpropagation (AlignProp) [152] and Direct Reward Fine-Tuning (DRaFT) [34] evaluate the reward on the final iteratively denoised image x_0. Techniques such as low-rank adaptation (LoRA) [80] and gradient checkpointing [26] are employed to reduce memory costs. To address the "depth-efficiency dilemma" of backpropagating through many steps, Deep Reward Tuning (DRTune) [220] enables more efficient deep supervision by stopping the gradient of the denoising network's input and training only on a selective subset of steps. Direct reward fine-tuning avoids the high variance and low sample efficiency inherent in RL

fine-tuning, thus improving training efficiency. However, fine-tuning with a differentiable reward model introduces a risk of over-optimization, potentially resulting in high-reward but lower-quality images. To mitigate this, DRaFT [34] explores methods such as LoRA scaling, early stopping, and KL regularization, with LoRA scaling found to be the most effective in reducing reward overfitting.

Advanced Fine-tuning Paradigms and Strategies. Beyond these foundational approaches, recent works have introduced more sophisticated paradigms to tackle key challenges like reward over-optimization, diversity collapse, and the alignment of specialized, fast models.

Tackling Reward Over-optimization: A primary failure mode is "reward hacking", where the model maximizes the reward metric at the expense of true image quality. Zhang et al. [252] proposed Temporal Diffusion Policy Optimization (TDPO-R), which provides reward supervision at each denoising step rather than just on the final image. This temporal inductive bias, combined with a "critic active neuron reset" mechanism to combat overfitting to early experiences (primacy bias), improves robustness. This aligns with theoretical work framing fine-tuning as entropy-regularized control to prevent reward collapse [24].

Aligning Fast and Few-Step Models: As diffusion models become faster through distillation, aligning them presents new challenges. Stepwise Diffusion Policy Optimization (SDPO) [28] is designed for few-step models by learning from dense rewards at each step. For ultra-fast (e.g., 1-2 step) models where standard RL fails, LaSRO [256] learns a differentiable surrogate reward in the latent space, enabling effective gradient-based tuning.

Innovations in Reward Signals and Applications: The design of the reward function itself is a major area of research. For instance, CoMat [116] uses an image-to-text model to provide concept-level rewards, addressing issues of concept ignorance and mismapping. Other works have focused on specialized alignment goals, such as improving long-text alignment [251], performing region-aware fine-tuning to fix local flaws [257], or using information-theoretic objectives for alignment [253]. Hu et al. [81] have demonstrated alignment with only sparse terminal rewards.

Emerging Paradigms: Other novel approaches are also being explored. Adversarial Diffusion Tuning (ADT) [85] uses an adversarial discriminator to close the gap between the training and inference distributions. To preserve sample diversity, a known issue in reward-driven fine-tuning, Shen et al. [175] proposed using Gradient-Informed GFlowNets (∇-GFlowNet) to balance reward maximization with exploration.

Summary and Outlook The landscape of fine-tuning diffusion models with RLHF is rapidly evolving from three foundational pillars—reward-weighted regression, policy-gradient RL, and direct reward optimization—to a more diverse set of specialized and robust techniques. While RWR is simple, it offers less precise control. Policy gradient methods like DDPO provide strong optimization capabilities but can suffer from high variance; this is addressed by more stable alternatives like PRDP. Direct reward methods like DRaFT are sample-efficient but risk memory overhead and reward over-optimization, a challenge that methods like DRTune and TDPO-R aim to mitigate. Furthermore, emerging paradigms like adversarial tuning and GFlowNet-based alignment are being developed to address fundamental issues like distribution shift and diversity collapse. A critical consideration in practical applications is feedback efficiency; when reward signals are expensive, online fine-tuning methods like SEIKO [200] use uncertainty modeling to minimize queries. The choice of method thus involves a trade-off between implementation complexity, training stability, sample efficiency, and the specific alignment goal, from general aesthetic improvement to targeted concept or regional correction.

4.2 Direct Preference Optimization of Diffusion Models

In this subsection, we present the DPO paradigm and its extension for diffusion alignment. This paradigm has been successfully adapted to diffusion models, directly optimizing them on human preference data without an explicit reward model as illustrated in Fig. 5 (b). This approach forms the basis for powerful open-source models like SD3.

Foundational DPO Methods. The core method, Diffusion-DPO [202], adapts the original DPO objective (Eq. (6)) to the iterative nature of diffusion models. Specifically, Diffusion-DPO formulated the objective function over the entire diffusion path $x_{0:T}$ as

$$\begin{aligned}&\mathcal{L}_{\text{Diffusion-DPO}}(p_\theta; p_{\text{ref}})\\ &= -\mathbb{E}_{(c,x_0^w,x_0^l)\sim\mathcal{D}} \log\sigma\left(\beta\mathbb{E}_{x_{1:T}^w\sim p_\theta(x_{1:T}^w|x_0^w,c),x_{1:T}^l\sim p_\theta(x_{1:T}^l|x_0^l,c)}\left[\log\frac{p_\theta(x_{0:T}^w|c)}{p_{\text{ref}}(x_{0:T}^w|c)} - \log\frac{p_\theta(x_{0:T}^l|c)}{p_{\text{ref}}(x_{0:T}^l|c)}\right]\right).\end{aligned} \tag{13}$$

Section 4.2 can be upper bounded [202, 233] using Jensen's inequality and the convexity of function $-\log\sigma$ to push the expectation outside:

$$\begin{aligned}\mathcal{L}_{\text{Diffusion-DPO}}(p_\theta; p_{\text{ref}}) \le -\mathbb{E}_{(c,x_0^w,x_0^l)\sim\mathcal{D},t\sim\mathcal{U}(0,T),x_{t-1,t}^w\sim p_\theta(x_{t-1,t}^w|x_0^w,c),x_{t-1,t}^l\sim p_\theta(x_{t-1,t}^l|x_0^l,c)}\\ \log\sigma\left(\beta T\left[\log\frac{p_\theta(x_{t-1}^w|x_t^w,c)}{p_{\text{ref}}(x_{t-1}^w|x_t^w,c)} - \log\frac{p_\theta(x_{t-1}^l|x_t^l,c)}{p_{\text{ref}}(x_{t-1}^l|x_t^l,c)}\right]\right).\end{aligned} \tag{14}$$

Because sampling from the reverse joint distribution $p_\theta(x_{t-1,t}|x_0,c)$ is intractable, Wallace et al. [202] approximates the reverse process $p_\theta(x_{1:T}|x_0,c)$ with the forward process $q(x_{1:T}|x_0,c)$. The right-hand side of Eq. (14) becomes:

$$\begin{aligned}\mathcal{L}(p_\theta; p_{\text{ref}}) = -\mathbb{E}_{(c,x_0^w,x_0^l)\sim\mathcal{D},t\sim\mathcal{U}(0,T),x_t^w\sim q(x_t^w|x_0^w,c),x_t^l\sim q(x_t^l|x_0^l,c)}\\ \log\sigma\Big(-\beta T\Big(D_{\text{KL}}\left(q(x_{t-1}^w|x_{0,t}^w,c)||p_\theta(x_{t-1}^w|x_t^w,c)\right) - D_{\text{KL}}\left(q(x_{t-1}^w|x_{0,t}^w,c)||p_{\text{ref}}(x_{t-1}^w|x_t^w,c)\right)\\ - D_{\text{KL}}\left(q(x_{t-1}^l|x_{0,t}^l,c)||p_\theta(x_{t-1}^l|x_t^l,c)\right) + D_{\text{KL}}\left(q(x_{t-1}^l|x_{0,t}^l,c)||p_{\text{ref}}(x_{t-1}^l|x_t^l,c)\right)\Big)\Big).\end{aligned} \tag{15}$$

This allows the model to learn from preference pairs (x_0^w, x_0^l) by adjusting the denoising process at each step. A key variant, D3PO [233], shares a similar objective but differs primarily in its sampling strategy, using the model's reverse process to obtain noisy latents rather than the forward noising process. Other works have explored alternative optimization perspectives, such as reinterpreting DPO as preference-weighted score matching [262], distribution optimization [94], or extending it to handle multiple reward signals [104].

Addressing Temporal Inconsistency: Step-Aware Preference. A key limitation of the foundational approach is the assumption that preferences are constant across all denoising steps. This overlooks that different stages of generation influence different aspects of the final image [71]. To address this, recent works introduce step-aware or temporally-discounted preferences. This provides more granular feedback throughout the generation process, which has been shown to improve alignment by relaxing the assumption of a static preference signal [118, 128, 162, 238].

Enhancements and Extensions of the DPO Framework. Further extensions to the DPO framework focus on two main areas: improving the training process and adapting to new contexts. Enhancements to the training process include curriculum-based strategies that use progressively harder examples [38], reference-free objectives like MaPO to improve robustness to distribution shifts [76], combining fine-tuning with test-time sampling [57], and constructing more visually consistent preference pairs to enhance the training signal [82]. In parallel, the framework has been

adapted for new contexts, most notably to handle different feedback formats. Diffusion-KTO [110], for instance, extends the paradigm to use simpler binary feedback (desirable/undesirable) instead of pairwise comparisons.

Summary and Outlook The DPO landscape for diffusion models is evolving rapidly. Foundational methods like Diffusion-DPO offer a simple and effective baseline. However, their core assumption of static preferences has motivated the development of more complex but potentially more accurate step-aware alignment techniques. Meanwhile, other research avenues focus on improving training efficiency through better data and learning strategies or expanding the paradigm's applicability to different feedback types and model architectures. The choice of method involves a trade-off between the simplicity of the core framework and the increased granularity and robustness offered by its more advanced extensions.

4.3 Test-Time Alignment of Diffusion Models

In this subsection, we review methods for aligning diffusion models at test-time without requiring model fine-tuning. This paradigm, also referred to as inference-time alignment [201], has rapidly evolved from strategies that **implicitly guide** generation by manipulating initial inputs and internal mechanisms, to more sophisticated approaches that employ **explicit reward-guided** strategies. In the latter, external preference models (e.g., for aesthetics or semantic consistency) directly steer the sampling process. This evolution offers more powerful and targeted control over model outputs, addressing complex human preferences.

4.3.1 Implicit Guidance: Optimizing Inputs and Internal Mechanisms. Implicit guidance strategies focus on improving alignment without an external reward function, typically by adjusting the initial conditions or internal states of the generation process. These methods are generally computationally lightweight.

Prompt Optimization. Prompt design plays a crucial role in determining generation quality and helping models better understand user intentions [163, 259]. While manual prompt engineering is common [125, 144], it can be labor-intensive. Consequently, recent works have explored optimizing prompts automatically. Wang et al. [207] developed RePrompt to refine text prompts toward more precise emotional expressions. Hao et al. [69] introduced Promptist, which adapts user input to model-preferred prompts via RL, where the reward is derived from metrics like CLIP similarity [157] and an aesthetic predictor [170]. Mañas et al. [133] proposed OPT2I, which leverages an LLM to iteratively revise user prompts, and Mo et al. [136] introduced an online RL strategy to generate dynamic fine-control prompts.

Attention Control. Attention control has emerged as a crucial technique for improving image-prompt alignment, addressing issues like attribute leakage and missing entities [55, 160]. Hertz et al. [71] first demonstrated that manipulating attention maps can guide the generation process. Building on this, Attend-and-Excite [21] enhances cross-attention interactions for more semantically aligned content. Wu et al. [219] propose a phase-wise attention modulation technique, and Li et al. [115] refined these methods for more complex semantic alignments. By dynamically adjusting focus, these mechanisms allow diffusion models to better align their outputs with human expectations [77, 234, 255].

Initial Noise Optimization. The reverse diffusion process is highly sensitive to the initial noise [228]. To find better noise without a reward model, Qi et al. [155] introduced the concept of inversion stability. Guo et al. [67] introduced Initial Noise Optimization (InitNO), which uses the model's internal attention scores to guide the initial noise towards semantically valid regions.

4.3.2 Explicit Reward-Guided Strategy: Trajectory Optimization with External Preferences. In contrast to implicit methods, explicit reward-guided strategies directly use an external reward function (differentiable or black-box) to modify the

sampling trajectory at each step to maximize a preference objective. These methods offer stronger control but typically increase inference complexity.

Reward-based Input Optimization. One approach is to use the reward signal to optimize the initial inputs. For instance, ReNeg [111] learns an optimal, universal negative embedding from reward signals. Others directly optimize the initial noise via gradient ascent on a reward function, as seen in Direct Noise Optimization (DNO) [197] and ReNO [53]. This concept has been further abstracted into learning a "noise prompt network" to generate a tailored "golden noise" [260] or framing it as a search problem guided by verifier feedback [131].

Reward-Guided Decoding and Sampling. A more powerful class of methods steers the entire decoding trajectory. Many are derivative-free, enabling the use of black-box reward models. For example, Li et al. [113] propose Soft Value-Based Decoding (SVBD), which uses a soft value function to estimate future rewards. To combat reward over-optimization and maintain diversity, methods based on Sequential Monte Carlo (SMC) [97] or Feynman-Kac steering [179] run and resample multiple generation trajectories. Xie and Gong [223] introduced DyMO for multi-objective alignment, while Zigzag Diffusion Sampling (Z-Sampling) [5] introduces a self-reflection mechanism. Yeh et al. [241] proposed Sampling Demons, a backpropagation-free method that performs stochastic optimization on the sampling distribution. On a theoretical level, Shi et al. [177] work towards a more formal connection between terminal preference labels and the generation trajectory.

4.3.3 Summary and Outlook. Test-time alignment techniques provide efficient and flexible pathways to improve generation without retraining. Their evolution clearly shows a trend from **implicit guidance** to **explicit reward-guided** strategies.

Implicit methods, such as prompt optimization and attention control, are simple to implement and add little inference overhead, making them effective for lightweight alignment. However, their control is less precise and often relies on heuristics. In contrast, explicit reward-guided methods offer more powerful and direct control by incorporating external preference models, enabling them to handle more complex alignment tasks. Their main challenges are the increased computational cost at inference time and the potential for "reward hacking", where the model optimizes the proxy score rather than the true image quality.

For practitioners, the choice of method involves a trade-off between control fidelity and inference efficiency. A promising future direction may involve combining training-based alignment with test-time methods. The former can instill general preferences into the model, while the latter can provide on-the-fly, personalized guidance, leading to more robust and versatile alignment systems.

4.4 Beyond T2I Diffusion Models

In this subsection, we review studies focused on the alignment of non-T2I diffusion models across various generation domains. While pioneering efforts have been made, adapting techniques from T2I diffusion models remains challenging due to domain-specific requirements in preference dataset collection, reward modeling, and the alignment techniques themselves. Each domain thus requires tailored adaptations to achieve effective alignment with human preferences.

4.4.1 Video Generation. Aligning video generation models [3] with human preferences introduces unique challenges, including immense computational overhead, a scarcity of large-scale video preference datasets, and the complex need to evaluate temporal consistency. Research in this area mirrors the methodological evolution seen in T2I alignment.

RLHF-based Alignment. Early approaches explored adapting the RLHF framework. For instance, Yuan et al. [244] proposed InstructVideo, which reduces fine-tuning costs by using partial DDIM sampling and repurposing image reward models for video. Focusing on more direct reward supervision, VADER [153] and LiFT [208] align models via direct reward fine-tuning. VADER leverages pre-trained reward models for efficiency, while LiFT develops a custom video-specific reward model (LiFT-Critic) trained on a new dataset with textual rationales (LiFT-HRA). The VisionReward framework further advances this by learning fine-grained, multi-dimensional preferences for both images and videos [226]. These methods, however, rely heavily on the quality and availability of explicit (or proxy) reward signals.

DPO-based Alignment. Inspired by its success and simplicity in the image domain, DPO was quickly adapted for video. Liu et al. [122] pioneered VideoDPO, which uses a comprehensive score to handle both visual quality and semantic alignment. To overcome the static nature of offline datasets, Zhang et al. [248] introduced OnlineVPO, an online DPO algorithm that uses a synthetically-trained video quality assessment (VQA) model for feedback. Other works like HuViDPO [93] focus on specific niches like human-centric videos, while Flow-DPO [83] demonstrates the paradigm's versatility by applying it to flow-based models. The primary challenge for DPO-based methods is constructing preference pairs that meaningfully capture the temporal nuances of video.

Test-time Alignment. For lightweight alignment without retraining, test-time methods are emerging. Lee et al. [102] presented VideoRepair, a model-agnostic framework that identifies and performs localized refinements during inference. Similarly, Oshima et al. [145] proposed using a diffusion latent beam search to improve perceptual quality without any model updates. These approaches offer great flexibility but may be limited in addressing complex, global alignment issues.

4.4.2 Audio and Motion Generation. Preference alignment is also proving effective for other types of sequential data generation. In **audio generation**, Majumder et al. [132] introduced Tango 2, which applies diffusion-DPO to fine-tune a text-to-audio model. It uses a synthetic preference dataset where "loser" samples feature missing concepts or incorrect temporal ordering. This work highlights DPO's effectiveness in capturing nuanced structural and semantic aspects of audio. For **motion generation**, Tan et al. [194] proposed SoPo, a DPO-based method that uses semi-online preference optimization. By combining online and offline data pairs, SoPo addresses the overfitting and sampling bias issues common in standard DPO, leading to higher-quality, human-preferred motions.

4.4.3 Image Editing. For the task of instructional visual editing, alignment must consider not only the final quality but also fidelity to user instructions. Zhang et al. [250] proposed Harnessing Human Feedback for Instructional Visual Editing (HIVE), an RLHF-based approach. It involves collecting feedback on the edited images to learn a reward function, which then guides the fine-tuning of diffusion models to better adhere to human preferences.

4.4.4 3D Generation. Aligning text-to-3D models involves challenges beyond aesthetics, such as ensuring multi-view consistency and geometric plausibility.

RLHF and DPO-based Alignment. Ye et al. [240] developed DreamReward, a two-stage RLHF-like process that first trains a 3D-aware reward model (Reward3D) and then uses it to fine-tune the generator. While this introduced the first 3D preference dataset, Zhou et al. [261] demonstrated a more direct path with DreamDPO, which applies DPO to leverage relative rankings, simplifying the need for absolute quality scores. This mirrors the RLHF-to-DPO trend observed in the other domains.

Geometric and Semantic Alignment. Moving beyond preference scores, Ignatyev et al. [86] addressed alignment from a geometric perspective. Their method optimizes for smooth and plausible transitions between generated objects in a latent space, directly enforcing structural consistency. This unique approach enables applications like 3D editing and hybridization, highlighting that 3D alignment is a multi-faceted problem concerning both preference and geometry.

4.4.5 Specialized Scientific and Control Applications. Alignment techniques are also applied to specialized domains with functional, rather than purely aesthetic, objectives. In **molecule generation**, Gu et al. [66] introduced AlignDiff, which aligns diffusion models with desired functional properties like binding affinity via preference optimization. A key challenge here is the accuracies of user-defined reward functions, which may not perfectly model real-world chemical properties. In **decision making**, Dong et al. [45] used RLHF to guide a planning diffusion model. This allows the model to plan for desired behaviors based on human preferences, demonstrating the potential of alignment for customizing agentic systems.

4.4.6 Summary and Outlook. The extension of alignment techniques from T2I to diverse domains like video, 3D, and molecule generation marks a significant evolution, demonstrating the versatility of core paradigms like RLHF and DPO. A clear pattern has emerged: DPO and its variants are rapidly adopted across modalities due to their simplicity, while RLHF offers powerful control but is hampered by the immense challenge of creating accurate, domain-specific reward models. The central trade-off between implementation simplicity and nuanced control is thus amplified by cross-domain challenges, primarily severe preference data scarcity and the complexity of modeling domain-specific objectives (e.g., temporal consistency or binding affinity). The future of non-T2I alignment hinges on overcoming this data bottleneck, likely through standardized, cross-domain benchmarks and more universal reward frameworks. This will pave the way for robust, and perhaps even hybrid, alignment techniques tailored to the unique properties of video, geometric, and other complex data types.

4.5 Open Challenges and Research Frontiers in Diffusion Alignment

In this subsection, we discuss several open challenges and research frontiers in diffusion alignment, focusing on the nuanced comparison between dominant paradigms, insights from LLM alignment, and challenges unique to the diffusion models.

4.5.1 The RLHF vs. DPO Debate: A Nuanced Comparison. The choice between the two dominant training-based alignment paradigms, RLHF and DPO, is not a simple matter of superiority but rather a complex trade-off involving training stability, sample efficiency, and robustness to distribution shifts.

RLHF, which uses a reward model as proxy for human preferences, allows for complex and nuanced reward functions. However, it often suffers from high variance and inefficient sample usage during the reinforcement learning phase. In contrast, DPO simplifies the training pipeline by eliminating the need for an explicit reward model, optimizing the policy directly on preference data. This can lead to more stable training but introduces its own vulnerabilities. Specifically, recent comprehensive studies suggest that a well-tuned PPO can outperform DPO, especially for complex tasks where DPO's performance may degrade due to distribution shifts between the preference data and the policy model [231]. This highlights a critical weakness of DPO: its performance is highly sensitive to the alignment between the training data distribution and the model's evolving policy.

Theoretical analysis further deepens this trade-off, indicating that the optimal choice depends on the relative learning difficulty of the reward function versus the optimal policy. RLHF may be more sample-efficient when the reward

model is easier to learn than the policy, whereas DPO holds an asymptotic advantage when the true reward function is exceptionally complex [142].

Beyond these training-based methods, test-time approaches align models during inference by adjusting inputs, noise, or internal mechanisms, thus avoiding costly fine-tuning. These methods are efficient and easy to deploy but may lack the precision needed for complex alignment tasks [114]. The entire landscape is rapidly evolving from static, offline training towards more dynamic online and iterative preference learning, which promises to better adapt to feedback and overcome the limitations of fixed preference datasets [15, 225].

4.5.2 Cross-Domain Insights: Adapting Innovations from LLM Alignment. The human alignment of diffusion models, while nascent, has benefited from adapting techniques pioneered for LLMs. For instance, Wallace et al. [202] successfully extended DPO [158] to create Diffusion-DPO, and Li et al. [110] adapted KTO [52] to produce Diffusion-KTO. These successes suggest that other advanced LLM alignment methods, such as IPO, ORPO, and PRO (as discussed in Section 3.2.2), are promising candidates for adaptation.

Indeed, the frontier of LLM alignment is rapidly expanding. For instance, f-DPO [203] generalizes DPO to a broader family of f-divergences for better diversity control, and GPO [196] provides a unified framework encompassing DPO and IPO. Other novel approaches like BOND [174], which distills a policy from a best-of-N sampling distribution, and techniques that derive dense, token-level rewards for free [20], also represent promising avenues. While these methods have advanced LLM alignment, their transfer to diffusion models is not guaranteed.

A key challenge lies in the fundamental architectural differences. Many LLM alignment techniques leverage the model's auto-regressive, next-token prediction structure in a discrete token space. For example, SimPO [135] achieves strong performance in LLMs by using a sequence's average log probability as a reference-free implicit reward. Creatively mapping such sequence-based concepts to the iterative, continuous, and high-dimensional denoising process of diffusion models remains a core research question.

4.5.3 Unique Challenges in Diffusion Model Alignment. Beyond the general challenges of alignment discussed in Section 3.3, diffusion models present a unique set of problems stemming from their generative process and multimodal nature. At a high level, the alignment problem in deep learning is fraught with fundamental risks, such as models learning to "hack" proxy reward signals or developing misaligned internal goals that deviate from the intended human preferences [140]. In diffusion models, these challenges are amplified and new ones emerge.

First, feedback for T2I models is inherently subjective and multidimensional, spanning image quality, realism, artistic style, and cultural context. This complexity makes reliable AI-generated feedback non-trivial to obtain. Second, the diversity of human preferences intensifies the issue of distributional shift. Most public preference datasets are built using Stable Diffusion variants (see Table 2), and applying these datasets to align other models (e.g., Midjourney) can lead to significant misalignment. This is a critical vulnerability for methods like DPO, whose learning dynamics can be skewed by the specific biases of the training data [88]. Third, the iterative nature of the diffusion process means that feedback may need to be incorporated at multiple steps, demanding highly efficient alignment algorithms. Fourth, integrating feedback from various modalities (e.g., visual, textual, numerical) in a coherent manner adds another layer of complexity. Finally, feedback on images is often sparser and noisier than on text, making it difficult for RL algorithms to learn effectively from inconsistent signals.

Beyond these issues, several profound challenges cast a shadow on current alignment strategies:

(1) **Reward Model Overoptimization:** A core issue in RLHF is "reward hacking", where optimizing against a fixed, imperfect reward model leads the policy to find exploitative solutions that maximize the proxy score but fail to capture true human preference. This is a predictable phenomenon governed by scaling laws, where performance on the true objective predictably rises and then falls as the policy diverges from its initial state [59]. In the visual domain, this can manifest as images that are semantically correct but visually distorted or absurd.
(2) **Scalability and Brittleness:** The long-term viability and robustness of current alignment methods are under question. Large-scale studies suggest RLHF may scale less efficiently than pre-training [79]. Furthermore, the resulting alignment can be brittle. For instance, safety alignment in T2I models has been shown to "backfire", where fine-tuning on benign data can cause suppressed, unsafe concepts to re-emerge, suggesting the initial alignment was not robustly learned [96]. Even low-rank modifications or model pruning can compromise safety alignment, revealing its fragility [212].
(3) **Security Vulnerabilities:** The reliance on a reward model introduces a new attack surface. Recent work has demonstrated a "clean-label" poisoning attack, termed BadReward, where an attacker subtly poisons the preference dataset with seemingly harmless examples. This manipulation corrupts the learned reward model, which in turn steers the diffusion model to generate harmful or undesired content during RLHF fine-tuning [46].

5 BENCHMARKS AND EVALUATION FOR HUMAN ALIGNMENT OF T2I DIFFUSION MODELS

In this section, we first review benchmark datasets and evaluation metrics for human alignment of T2I diffusion models in Section 5.1 and Section 5.2, respectively. We then discuss the associated challenges in Section 5.3.

5.1 Benchmarks for Human Alignment of T2I Diffusion Models

The foundation of any alignment technique is the data used to define human preferences. In this subsection, we discuss benchmark datasets for human alignment of T2I diffusion models, categorizing them into three types that reflect the field's maturation: scalar human preference datasets, multi-dimensional human feedback datasets, and AI feedback datasets. This progression highlights a move towards capturing more nuanced and scalable representations of human intent. Table 2 compares the reviewed benchmark datasets across three aspects: prompts, images, and annotations.

5.1.1 Scalar Human Preference Datasets. Early preference datasets primarily provide an overall comparison among images using a single scalar score or a pairwise choice to indicate human preference. Wu et al. [218] introduced the HPD v1 dataset, with prompts and images collected from the public Stable Foundation Discord channel. While valuable, this approach inherently captures the preferences of a niche group of experienced Stable Diffusion users, which may not generalize to the broader population. Wu et al. [216] later introduced a larger dataset, HPD v2, where the prompts are sourced from COCO Captions [27] and the ChatGPT-cleansed DiffusionDB [211]. Notably, HPD v2 includes images generated by nine different generative models, including diffusion models, GANs, and auto-regressive-based models, resulting in a higher degree of diversity. The pairwise image preferences in HPD v2 are derived from the preference rankings of 57 employed annotators over the generated images.

To capture more authentic, in-the-wild preferences, Kirstain et al. [99] developed a web application to build the Pick-a-Pic v1 dataset, collecting prompts and preferences over images generated by multiple Stable Diffusion variants from thousands of real users. However, such a collection method may be subject to self-selection bias, as the user base of a specific application may not be fully representative. Xu et al. [227] created the ImageRewardDB dataset by using six popular T2I generative models to generate images based on a diverse selection of prompts from DiffusionDB [211].

They implemented a three-stage annotation pipeline in which hired annotators annotate prompts, rate text-image pairs, and rank images. This pipeline provides more detailed human preference feedback, capturing aspects such as fidelity, image-text alignment, and overall quality.

Beyond explicit human preferences from annotators regarding image fidelity and image-text alignment, Isajanyan et al. [89] introduced the Picsart Image-Social dataset, which captures social popularity for creative editing purposes as an implicit and novel dimension of human preferences. Instead of relying on explicit annotations, they utilized editing behaviors (e.g., how often an image is "remixed" by others) from the online visual creation and editing platform Picsart to curate this dataset. While this provides a unique, large-scale signal, it is important to note that such a proxy may reflect creative utility or "remixability" more than pure image quality or prompt alignment.

5.1.2 Multi-dimensional Human Feedback Datasets. Recognizing the limitations of a single preference score, multi-dimensional human preferences [249] and rich human feedback [117] have been shown to be effective in improving T2I generations. Specifically, motivated by the observation that human preference results vary when evaluating images across different aspects, Zhang et al. [249] constructed the MHP dataset. This dataset was created using a balanced and refined prompt set from four sources and nine different T2I diffusion models to generate images with various resolutions and aspect ratios. In particular, 210 crowd-sourced annotators were employed to provide preference choices over image pairs across four dimensions: aesthetics, detail quality, semantic alignment, and overall assessment. Similarly, Liang et al. [117] sampled a diverse and balanced subset of image-text pairs from the Pick-a-Pic dataset. They then constructed the RichHF-18K dataset, which provides enriched feedback signals. Specifically, they marked implausible or misaligned image regions, annotated which words in the text prompt were missing or misrepresented in the corresponding image, and provided four fine-grained scores, including plausibility, image-text alignment, aesthetics, and overall quality, on a 5-level Likert scale.

5.1.3 AI Feedback Datasets. Scaling up human feedback datasets is prohibitively expensive due to the high cost of human annotation. This has motivated researchers to explore AI feedback for constructing preference datasets. Peng et al. [148] introduced DreamBench++, a benchmark for personalized image generation that uses GPT-4o for automated evaluation aligned with human preferences, focusing on concept preservation and prompt following. Similarly, Wu et al. [217] created the VisionPrefer dataset using multimodal large language models, specifically GPT-4 Vision. The annotations include scalar scores, preference rankings, and rationales for the annotations across four aspects: prompt-following, fidelity, aesthetics, and harmlessness. They then trained a reward model, VP-Score, based on VisionPrefer. VP-Score demonstrates comparable performance to reward models trained on human preference datasets in predicting human preferences and aligning T2I diffusion models with these preferences.

5.2 Evaluation for Human Alignment of T2I Diffusion Models

The benchmarks described previously enable the development of diverse evaluation paradigms. In this subsection, we review the primary methods for assessing alignment, starting with the evaluation of the reward models themselves in Section 5.2.1, and then moving to the evaluation of the final T2I diffusion models in Section 5.2.2.

5.2.1 Evaluation for Reward Models. To evaluate the performance of reward models in predicting human preference, the classical metric used is pairwise preference prediction accuracy. To calculate this accuracy, the reward model is first used to score a pair of images with the same prompt. The accuracy is then determined by the ratio of cases where the reward model assigns a higher score to the image-text pair preferred by humans on the test set. While high accuracy on

Table 2. Comparison of existing feedback datasets for T2I diffusion models.

Feedback Dataset → Reward Model	Prompt Source	Prompt Count	Image Generation Source	Image Count	Annotator Info.	Annotation Count
HPD v1 → HPS v1 [218]	Stable Foundation Discord channel	25,205	Stable Diffusion	98,807	2659 experienced users	25,205
HPD v2 → HPS v2 [216]	COCO Captions + DiffusionDB	107,915	9 models + COCO images	433,760	57 employed annotators	798,090
Pick-a-Pic v1 → PickScore [99]	Real users	37,523	Stable Diffusion variants	1,169,494	6,394 web app users	584,747
ImageRewardDB → ImageReward [227]	DiffusionDB	8,878	6 models	273,784	Annotation company	136,892
MHP → MPS [249]	PromptHero + DiffusionDB + KOLORS + GPT4	66,389	9 models	607,541	210 crowd-sourced annotators	918,315
RichHF-18K → RAHF [117]	Pick-a-Pic v1	17,760	Pick-a-Pic v1	35,520	27 trained annotators	17,760
Picsart Image-Social → Social Reward [89]	Social platform user prompts	104 K	Several in-house models	1.7 M	1.5 M users	3M
VisionPrefer → VP-Score [217]	DiffusionDB	179 K	Stable Diffusion variants	0.76 M	GPT-4 Vision	1.2 M

a given benchmark is a necessary indicator, it is not sufficient, as a model can overfit to the benchmark's specific data distribution and fail to generalize to novel, out-of-distribution prompts.

Table 2 delineates the mapping between prominent feedback datasets and their corresponding reward models. Building on this, Table 3 presents a comparative analysis of human preference prediction accuracy for nine models across five benchmarks. The results reveal a consistent pattern of benchmark specialization: models such as MPS, PickScore, and Social Reward achieve state-of-the-art performance predominantly on their native datasets. This finding strongly indicates that a model's predictive power is highly contingent upon the specific data distribution of its training benchmark. Furthermore, the overall prediction accuracies are modest, with most remaining below 80%, suggesting that robustly capturing general human preferences remains a significant open challenge.

In addition to predicting overall human preference on generated images from the same prompt [89, 99, 216–218, 227], novel reward models have been proposed to predict multi-dimensional preferences [249], detect implausible or misaligned regions, and identify misaligned keywords [117]. As a result, distinct metrics have been developed for evaluation, including the correlation between Elo ratings [49] of real users and reward models [99], the correlation between the win ratio of reward models and humans [99, 249], and metrics like NSS, KLD, AUC-Judd, SIM, and CC [13] for evaluating saliency heatmaps [117].

5.2.2 Evaluation for T2I Diffusion Model.

Model Evaluation Prompts. To evaluate T2I diffusion models, it is essential to collect a representative set of prompts for image generation that aligns with the evaluation goals. Various prompt datasets are available for T2I model evaluation in the context of human alignment. For example, Kirstain et al. [99] used prompts from MS-COCO [119] and Pick-a-Pic v1 for evaluation, while Xu et al. [227] selected prompts from DiffusionDB [211] and MT Bench [149]. Table 2 outlines the prompt sources for each feedback dataset, highlighting different motives for image generation, such as real user intention [99], challenging multi-task prompts [149], and social popularity [89]. Consequently, we recommend that the community employ suitable prompts when assessing the performance of T2I diffusion models across different evaluation aspects.

Table 3. Comparison of different reward models for human preference evaluation. The pairwise preference prediction accuracy (%) is reported on ImageRewardDB, HPD v2, MHP, Pick-a-Pic v1, and Picsart Image-Social dataset. The **bold** results indicate the best result on each dataset. The results with no mark, $*$, and $**$ are from Zhang et al. [249], Wu et al. [217], and Isajanyan et al. [89], respectively.

	ImageRewardDB	HPD v2	MHP	Pick-a-Pic v1	Picsart Image-Social
CLIP score [157]	54.3	71.2	63.7	60.8*	51.9
Aesthetic score [170]	57.4	72.6	62.9	56.8*	55.3
HPS v1 [218]	61.2	73.1	65.5	66.7*	-
HPS v2 [216]	65.7	83.3	65.5	67.4*	59.4
PickScore [99]	62.9	79.8	69.5	**70.5***	62.6
ImageReward [227]	65.1	70.6	67.5	61.1*	60.5
MPS [249]	**67.5**	**83.5**	**74.2**	-	-
VP-Score [217]	66.3*	79.4*	-	67.1*	-
Social Reward [89]	-	-	-	-	**69.7**

Image Quality. The Inception Score (IS) [167] and Fréchet Inception Distance (FID) [72] are the most widely adopted metrics for measuring image quality without considering the text prompt. These metrics utilize features extracted from a pre-trained image classifier, typically the Inception-V3 model [193], to evaluate the fidelity and diversity of generated images. However, their primary limitation is that they do not account for the text prompt and their correlation with human perceptual judgment is imperfect, rendering them insufficient for a comprehensive assessment of alignment.

Human Preference Evaluation. Reward models can serve as metrics for human preference, allowing comparisons between various T2I generative models based on their reward scores, or for monitoring the training process of aligning models with human preferences. Typically, reward scores will show an increasing trend when models are fine-tuned using RLHF methods (see Section 4.1) or through DPO approaches (see Section 4.2) with feedback datasets. This increasing trend indicates improved alignment with human preferences, as measured by the reward models. A critical caveat here is the risk of "evaluation hacking" or circular reasoning: using a reward model to evaluate a policy that was optimized using that same model (or a very similar one) can lead to inflated scores that do not reflect true gains in alignment. Notably, most reward scores account for the text prompt, often computed as a scaled cosine similarity, with the exception of metrics like the aesthetic score [170], which measures aesthetics independently. For a recent review of T2I evaluation, see Hartwig et al. [70].

Fine-grained Evaluation. The automated evaluation metrics introduced above offer a holistic measure of quality but often lack interpretability. To address this, recent works focus on fine-grained, instance-level analysis to better reflect the diverse capabilities of T2I models. These efforts can be categorized into several key areas:

- **Compositional Reasoning:** Benchmarks like GENEval [62] and GenAI-Bench [120] assess a model's ability to handle complex prompts involving multiple objects, attributes, and spatial relationships. VQAScore [120] probes this by using a visual question-answering model to verify compositional correctness with a binary question.
- **Visual Reasoning and Social Bias:** DALL-Eval [30] was designed to probe models' commonsense reasoning skills (e.g., object counting, spatial relations) and to quantify social biases related to protected attributes like gender and skin tone.

- **Multi-faceted Skill Assessment:** Comprehensive benchmarks like HEIM [106] and VPEval [65] evaluate models across a wide spectrum of skills (e.g., text understanding, photorealism, counting) and provide more interpretable, explanatory results.
- **LLM-based Evaluation:** LLMScore [129] leverages the descriptive power of LLMs to first generate a textual caption for an image and then score its alignment with the original prompt, providing a human-like rationale for its assessment.
- **Style Attribution:** Other work has focused on evaluating more abstract concepts like artistic style, for instance, by measuring a model's ability to attribute and match the style of specific artists [184].

5.3 Challenges in Benchmarking and Evaluation for T2I Diffusion Models

While the infrastructure for alignment is maturing, it faces profound challenges that limit the reliability and scope of current research. These challenges fall into two interconnected categories: those related to benchmark construction and those inherent to evaluation methodologies.

Challenges in Benchmark Construction. Creating a benchmark that truly represents human preference is fraught with difficulties. First, human preferences are inherently subjective, diverse, and dynamic. Most current datasets, while striving for diversity in prompts and models, are often annotated by a limited number of individuals, failing to capture the full spectrum of human taste. This can lead to models aligned with a homogenous, averaged preference rather than a pluralistic one, as current alignment frameworks often enforce uniformity through strict rubrics, thereby suppressing the natural diversity of human opinions [22]. Consequently, the claimed diversity of datasets must be rigorously measured, not just asserted [254]. Second, all existing benchmarks are static snapshots. They cannot account for the evolving nature of human preferences over time, a problem that might ultimately require continual learning approaches to solve [204].

Challenges in Evaluation Methodologies. The metrics used for evaluation face their own set of challenges. The lack of standardization in evaluation prompts complicates consistent comparison across studies; for instance, auto-generated prompts may lack diversity and objectivity compared to those sourced from professionals [120]. Establishing widely adopted, standardized evaluation protocols remains an open problem. Furthermore, the increasing reliance on multimodal large language models (MLLMs) as automated evaluators is a double-edged sword. While scalable, these MLLMs can suffer from their own issues, such as positional bias, and their evaluation quality is capped by the capabilities of the underlying model, potentially propagating these flaws into the metrics themselves [148, 217].

Finally, a superordinate challenge is that current metrics struggle to assess abstract yet crucial qualities like creativity [56]. This points to the significant risk of an "alignment tax": the phenomenon where excessive optimization for specific, measurable alignment goals (e.g., prompt following, safety) inadvertently suppresses a model's creativity, diversity, and overall utility. Understanding and mitigating this trade-off is critical for developing genuinely helpful and innovative generative models.

6 FUTURE DIRECTIONS FOR HUMAN ALIGNMENT OF T2I DIFFUSION MODELS

In this section, we outline three promising future directions for human alignment of T2I diffusion models that can inspire further advancements in the area.

Preference Learning with Inconsistent and Multidimensional Human Feedback. A critical limitation of current alignment paradigms is their tendencies to collapse diverse human preferences into a single, monolithic reward function. This approach fails to capture the rich, often conflicting, and subjective nature of human values. This challenge is

particularly acute for diffusion models, as visual content—encompassing aesthetics, artistic style, and cultural context—is inherently more subjective and ambiguous than text. A major future direction is therefore to develop systems that embrace *pluralistic alignment* [188]. Instead of seeking a single best output, the goal is to model a distribution of desirable outputs that reflect a wide range of viewpoints.

This requires moving beyond simple reward modeling. One approach is to draw upon the formalisms of *social choice theory* to aggregate diverse feedback in a principled manner, leading to paradigms like Reinforcement Learning from Collective Human Feedback (RLCHF) [37]. Another is to fundamentally change the optimization objective. For instance, Nash Learning from Human Feedback (NLHF) learns a preference relation instead of a scalar reward, which is more robust to the non-transitive or cyclical preferences that naturally arise in group feedback [138]. To explicitly handle distinct preference groups, MaxMin-RLHF identifies different user sub-populations and optimizes for the worst-case reward among them, ensuring the model does not cater only to the majority [19]. Applying these advanced aggregation mechanisms to diffusion models presents unique opportunities and challenges, requiring novel algorithms that can navigate optimization in a high-dimensional, continuous generation space while handling the sparse nature of image-based feedback. For more dynamic control, methods like Rewards-in-Context [237], which allow users to specify multi-objective preferences at inference time, represent a promising path toward achieving not just learned, but *controllable* pluralistic alignment, striking a balance between general preference satisfaction and personalized generation.

Data-centric Preference Learning. Current preference learning approaches for diffusion models typically rely on supervised learning with large-scale human-annotated datasets. While diffusion models can generate content at low cost, obtaining human feedback on these outputs remains expensive and slow, particularly in visual domains. This data bottleneck is exacerbated by findings that current alignment techniques may not scale efficiently, requiring ever-larger datasets for diminishing gains [79]. To address this, researchers can explore preference learning with minimal or even zero human-annotated data.

Two potential approaches can mitigate this dependency. First, while AI feedback is a scalable alternative, it carries risks of inheriting biases, lacking diversity, and potential model collapse from training on synthetic data [176]. This area requires further exploration to enhance reliability. Second, a promising direction is using AI-generated paired samples with known preference relations. For instance, if we know that Algorithm A produces better results than Algorithm B based on a specific alignment metric (e.g., photorealism or prompt-following), we can use these paired samples to curate a large-scale, low-cost preference dataset to improve that metric. For better preference diversity and reliability, we may choose multiple algorithms to generate paired samples and use the alignment metric score to remove those pairs without significant preference score differences. Such data-centric strategies significantly reduce reliance on direct human annotation. However, their efficacy is still bound by predefined external metrics. A more ambitious goal is to enable models to develop an intrinsic understanding of preference, which leads to our final research frontier: self-alignment.

Self-Alignment of Diffusion Models. Currently, methods for aligning diffusion models rely on intensive external supervision. We propose a forward-looking direction, *self-alignment*, built on the hypothesis that large diffusion models, trained on vast, high-quality data, already possess implicit prior knowledge of human preferences (e.g., aesthetics, realism) but lack a mechanism to express it explicitly [12, 147, 191]. The goal is to unlock this latent capability, enabling a model to align itself with minimal or no external feedback. This would allow a model to act as its own reward function for RLHF or as an AI annotator for DPO, creating a self-improving loop that could eventually surpass human capabilities. This vision has been discussed for LLMs but is underexplored for diffusion models.

Achieving self-alignment may follow two complementary paths:

- **Unlocking Inherent Judgment:** The first path focuses on extracting and amplifying the model's existing, but latent, evaluative abilities. Initial work on *diffusion classifiers* [23, 33, 108, 156] has shown that these models can judge text-image alignment. Future work could go further by probing internal representations, such as intermediate attention maps [21] or influence functions [100, 239], to construct a more comprehensive, self-generated reward signal.
- **Fostering Self-Reflective Generation:** The second path aims to embed alignment into the generation process itself. Recent explorations into equipping models with a Chain-of-Thought (CoT) [213]-like reasoning process are a step in this direction [68]. By generating content step-by-step and using internal signals to verify and refine each stage, the model can engage in a form of self-correction, promoting more coherent and aligned outputs without external guidance.

The realization of self-alignment would mark a paradigm shift from models being passively aligned to becoming active agents in their own refinement. This could not only resolve the data bottleneck but also unleash new levels of creative and intelligent generation.

7 SUMMARY

In this paper, we have presented a comprehensive review of the alignment of diffusion models. We explored recent advances in diffusion models, elucidated fundamental concepts of human alignment, and discussed various techniques for enhancing the alignment of diffusion models, as well as extending these techniques to tasks beyond T2I generation. Additionally, we outlined the benchmark datasets and evaluation metrics critical for assessing T2I diffusion models. Looking ahead, we identified a series of profound challenges and several promising directions for future research. We hope that this work not only highlights recent advancements and existing gaps in diffusion alignment but also inspires and guides future alignment research of diffusion models.

REFERENCES

[1] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. 2021. On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift. *Journal of Machine Learning Research* 22, 98 (2021), 1–76. http://jmlr.org/papers/v22/19-736.html

[2] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. arXiv:2402.14740 [cs.LG] https://arxiv.org/abs/2402.14740

[3] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. 2022. Video Generative Adversarial Networks: A Review. *ACM Comput. Surv.* 55, 2, Article 30 (jan 2022), 25 pages. https://doi.org/10.1145/3487891

[4] Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John Shawe-Taylor. 2007. Variational inference for diffusion processes. *Advances in neural information processing systems* 20 (2007).

[5] Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. 2025. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In *The Thirteenth International Conference on Learning Representations.* https://openreview.net/forum?id=MKvQH1ekeY

[6] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[7] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2, 3 (2023), 8.

[8] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2023. Training Diffusion Models with Reinforcement Learning. In *The Twelfth International Conference on Learning Representations.*

[9] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 22563–22575.

[10] Salomon Bochner. 1949. Diffusion equation and stochastic processes. *Proceedings of the National Academy of Sciences* 35, 7 (1949), 368–370.

[11] 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.

[12] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. 2024. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 4971–5012.

[13] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence* 41, 3 (2018), 740–757.

[14] Zhipeng Cai, Zuobin Xiong, Honghui Xu, Peng Wang, Wei Li, and Yi Pan. 2021. Generative Adversarial Networks: A Survey Toward Private and Secure Applications. *ACM Comput. Surv.* 54, 6, Article 132 (jul 2021), 38 pages. https://doi.org/10.1145/3459992

[15] Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. 2024. Human Alignment of Large Language Models through Online Preference Optimisation. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 5409–5435. https://proceedings.mlr.press/v235/calandriello24a.html

[16] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. 2024. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[17] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. AI Alignment with Changing and Influenceable Reward Functions. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 5706–5756. https://proceedings.mlr.press/v235/carroll24a.html

[18] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. 2024. MaxMin-RLHF: Alignment with Diverse Human Preferences. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 6116–6135. https://proceedings.mlr.press/v235/chakraborty24b.html

[19] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. 2024. MaxMin-RLHF: Alignment with Diverse Human Preferences. In *Forty-first International Conference on Machine Learning*.

[20] Alex James Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. 2024. Dense Reward for Free in Reinforcement Learning from Human Feedback. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=eyxVRMrZ4m

[21] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.

[22] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. 2025. PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment. In *The Thirteenth International Conference on Learning Representations*.

[23] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. 2024. Robust Classification via a Single Diffusion Model. In *Forty-first International Conference on Machine Learning*.

[24] Jiaming Chen, Yujia Li, and Yu Tian. 2024. Fine-Tuning of Continuous-Time Diffusion Models as Entropy-Regularized Control. *arXiv preprint arXiv:2402.15194* (2024).

[25] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=eAKmQPe3m1

[26] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174* (2016).

[27] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[28] Chuan Cheng, Man Tang, and Guan Zhang. 2024. Aligning Few-Step Diffusion Models with Dense Reward Difference Learning. *arXiv preprint arXiv:2411.11727* (2024).

[29] Xiang Cheng, Dong Yin, Peter Bartlett, and Michael Jordan. 2020. Stochastic gradient and Langevin processes. In *International Conference on Machine Learning*. PMLR, 1810–1819.

[30] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3043–3054.

[31] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Visual programming for step-by-step text-to-image generation and evaluation. *Advances in Neural Information Processing Systems* 36 (2023).

[32] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[33] Kevin Clark and Priyank Jaini. 2024. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems* 36 (2024).

[34] Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. 2024. Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In *The Twelfth International Conference on Learning Representations*.

[35] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic

Capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[36] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mosse, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. 2024. Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 9346–9360.

[37] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. 2024. Position: social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 371, 15 pages.

[38] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, and Mubarak Shah. 2025. Curriculum direct preference optimization for diffusion and consistency models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2824–2834.

[39] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. ULTRAFEEDBACK: Boosting Language Models with Scaled AI Feedback. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 9722–9744.

[40] Adyasha Dash and Kathleen Agres. 2024. AI-Based Affective Music Generation Systems: A Review of Methods and Challenges. *ACM Comput. Surv.* 56, 11, Article 287 (jul 2024), 34 pages. https://doi.org/10.1145/3672554

[41] Fei Deng, Qifei Wang, Wei Wei, Tingbo Hou, and Matthias Grundmann. 2024. Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7423–7433.

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

[43] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. 2023. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. *Transactions on Machine Learning Research* (2023).

[44] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863* (2024).

[45] Zibin Dong, Yifu Yuan, Jianye HAO, Fei Ni, Yao Mu, YAN ZHENG, Yujing Hu, Tangjie Lv, Changjie Fan, and Zhipeng Hu. 2024. AlignDiff: Aligning Diverse Human Preferences via Behavior-Customisable Diffusion Model. In *The Twelfth International Conference on Learning Representations*.

[46] Kaiwen Duan, Hongwei Yao, Yufei Chen, Ziyun Li, Tong Qiao, Zhan Qin, and Cong Wang. 2025. BadReward: Clean-Label Poisoning of Reward Models in Text-to-Image RLHF. *arXiv preprint arXiv:2506.03234* (2025).

[47] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[48] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems* 36 (2023).

[49] Arpad E Elo and Sam Sloan. 1978. The rating of chessplayers: Past and present. *(No Title)* (1978).

[50] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. 2020. Implementation Matters in Deep RL: A Case Study on PPO and TRPO. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1etN1rtPB

[51] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=FPnUhsQJ5B

[52] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model Alignment as Prospect Theoretic Optimization. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 12634–12651.

[53] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. 2024. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems* 37 (2024), 125487–125519.

[54] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 2023. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems* 36 (2023).

[55] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=PUIqjT4rzq7

[56] Giorgio Franceschelli and Mirco Musolesi. 2024. Creativity and Machine Learning: A Survey. *ACM Comput. Surv.* 56, 11, Article 283 (jun 2024), 41 pages. https://doi.org/10.1145/3664595

[57] Minghao Fu, Guo-Hua Wang, Liangfu Cao, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. 2025. CHATS: Combining Human-Aligned Optimization and Test-Time Sampling for Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*.

[58] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.

[59] Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*. PMLR, 10835–10866.

[60] Songyang Gao, Qiming Ge, Wei Shen, Shihan Dou, Junjie Ye, Xiao Wang, Rui Zheng, Yicheng Zou, Zhi Chen, Hang Yan, Qi Zhang, and Dahua Lin. 2024. Linear Alignment: A Closed-form Solution for Aligning Human Preferences without Tuning and Feedback. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=Y4wxCICbD0

[61] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A General Theoretical Paradigm to Understand Learning from Human Preferences. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 238)*, Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (Eds.). PMLR, 4447–4455. https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html

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

[63] Daniel T Gillespie. 2000. The chemical Langevin equation. *The Journal of Chemical Physics* 113, 1 (2000), 297–306.

[64] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[65] An-Rok Goyal, Beomsu Kim, and Gyeong-Moon Kim. 2023. VPGen: Visual-Programming-Guided Generation for Text-to-Image Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36.

[66] Siyi Gu, Minkai Xu, Alexander Powers, Weili Nie, Tomas Geffner, Karsten Kreis, Jure Leskovec, Arash Vahdat, and Stefano Ermon. 2024. Aligning Target-Aware Molecule Diffusion Models with Exact Energy Optimization. *arXiv preprint arXiv:2407.01648* (2024).

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

[68] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. 2025. Can We Generate Images with CoT? Let's Verify and Reinforce Image Generation Step by Step. *arXiv preprint arXiv:2501.13926* (2025).

[69] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2023).

[70] Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Timo Ropinski, et al. 2024. Evaluating text to image synthesis: Survey and taxonomy of image quality metrics. *arXiv e-prints* (2024), arXiv–2403.

[71] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=_CDixzkzeyb

[72] 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).

[73] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

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

[75] Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691* 2, 4 (2024), 5.

[76] Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. 2024. Margin-aware preference optimization for aligning diffusion models without reference. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*.

[77] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. 2023. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7462–7471.

[78] Yuzhong Hong, Hanshan Zhang, Junwei Bao, Hongfei Jiang, and Yang Song. 2025. Energy-Based Preference Model Offers Better Offline Alignment than the Bradley-Terry Preference Model. In *International Conference on Machine Learning (ICML)*.

[79] Zhenyu Hou, Pengfan Du, Yilin Niu, Zhengxiao Du, Aohan Zeng, Xiao Liu, Minlie Huang, Hongning Wang, Jie Tang, and Yuxiao Dong. 2024. Does RLHF Scale? Exploring the Impacts From Data, Model, and Method. *arXiv preprint arXiv:2412.06000* (2024).

[80] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[81] Zijing Hu, Fengda Zhang, Long Chen, Kun Kuang, Jiahui Li, Kaifeng Gao, Jun Xiao, Xin Wang, and Wenwu Zhu. 2025. Towards better alignment: Training diffusion models with reinforcement learning against sparse rewards. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23604–23614.

[82] Zijing Hu, Fengda Zhang, and Kun Kuang. 2025. D-Fusion: Direct Preference Optimization for Aligning Diffusion Models with Visually Consistent Samples. In *International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/2505.22002

[83] Lijun Huang, Ka-Chun Wong, and Jun-Cheng Chen. 2025. Flow-DPO: Improving Video Generation with Human Feedback. *arXiv preprint arXiv:2501.13918* (2025).

[84] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*. PMLR, 13916–13932.

[85] Zihan Huang, Zekun Zhang, Yifei Liu, Yujia Li, and Yu Tian. 2025. ADT: Tuning Diffusion Models with Adversarial Supervision. *arXiv preprint arXiv:2504.11423* (2025).

[86] Savva Victorovich Ignatyev, Nina Konovalova, Daniil Selikhanovych, Oleg Voynov, Nikolay Patakin, Ilya Olkov, Dmitry Senushkin, Alexey Artemov, Anton Konushin, Alexander Filippov, Peter Wonka, and Evgeny Burnaev. 2025. A3D: Does Diffusion Dream about 3D Alignment?. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=QQCIfkhGIq

[87] Shawn Im and Yixuan Li. 2024. Understanding the Learning Dynamics of Alignment with Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 20983–21006.

[88] Shawn Im and Yixuan Li. 2024. Understanding the learning dynamics of alignment with human feedback. *arXiv preprint arXiv:2403.18742* (2024).

[89] Arman Isajanyan, Artur Shatveryan, David Kocharian, Zhangyang Wang, and Humphrey Shi. 2024. Social Reward: Evaluating and Enhancing Generative AI through Million-User Feedback from an Online Creative Community. In *The Twelfth International Conference on Learning Representations*.

[90] Abdul Jabbar, Xi Li, and Bourahla Omar. 2021. A Survey on Generative Adversarial Networks: Variants, Applications, and Training. *ACM Comput. Surv.* 54, 8, Article 157 (oct 2021), 49 pages. https://doi.org/10.1145/3463475

[91] Haozhe Ji, Cheng Lu, Yilin Niu, Pei Ke, Hongning Wang, Jun Zhu, Jie Tang, and Minlie Huang. 2024. Towards Efficient Exact Optimization of Language Model Alignment. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 21648–21671. https://proceedings.mlr.press/v235/ji24c.html

[92] Zhen Jia, Zhang Zhang, Liang Wang, and Tieniu Tan. 2024. Human Image Generation: A Comprehensive Survey. *ACM Comput. Surv.* 56, 11, Article 279 (jun 2024), 39 pages. https://doi.org/10.1145/3665869

[93] Lifan Jiang, Boxi Wu, Jiahui Zhang, Xiaotong Guan, and Shuang Chen. 2025. HuViDPO: Enhancing Video Generation through Direct Preference Optimization for Human-Centric Alignment. *arXiv preprint arXiv:2502.01690* (2025).

[94] Ryotaro Kawata, Kazusato Oko, Atsushi Nitanda, and Taiji Suzuki. 2025. Direct Distributional Optimization for Provable Alignment of Diffusion Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=Nvw2szDdmI

[95] Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[96] Sanghyun Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. 2024. Safety Alignment Backfires: Preventing the Re-emergence of Suppressed Concepts in Fine-tuned Text-to-Image Diffusion Models. *arXiv preprint arXiv:2412.00357* (2024).

[97] Sunwoo Kim, Minkyu Kim, and Dongmin Park. 2025. Test-time alignment of diffusion models without reward over-optimization. In *The Thirteenth International Conference on Learning Representations*.

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

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

[100] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[101] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*. https://openreview.net/forum?id=a-xFK8Ymz5J

[102] Daeun Lee, Jaehong Yoon, Jaemin Cho, and Mohit Bansal. 2024. VideoRepair: Improving Text-to-Video Generation via Misalignment Evaluation and Localized Refinement. *arXiv preprint arXiv:2411.15115* (2024).

[103] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=uydQ2W41KO

[104] Kyungmin Lee, Seungryong Kim, and Kwanghoon Sohn Lee. 2025. Calibrated Multi-Preference Optimization for Aligning Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2502.02588

[105] 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).

[106] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy S Liang. 2023. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems* 36 (2023).

[107] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).

[108] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. 2023. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2206–2217.

[109] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[110] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. 2024. Aligning diffusion models by optimizing human utility. *arXiv preprint arXiv:2404.04465* (2024).

[111] Xiaomin Li, Yixuan Liu, Takashi Isobe, Xu Jia, Qinpeng Cui, Dong Zhou, Dong Li, You He, Huchuan Lu, Zhongdao Wang, et al. 2025. Reneg: Learning negative embedding with reward guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23636–23645.

[112] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. 2024. Self-Alignment with Instruction Backtranslation. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=1oijHJBRsT

[113] Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, et al. 2024. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *arXiv preprint arXiv:2408.08252* (2024).

[114] Yafu Li, Xuyang Hu, Xiaoye Qu, Linjie Li, and Yu Cheng. 2025. Test-Time Preference Optimization: On-the-Fly Alignment via Iterative Textual Feedback. *arXiv preprint arXiv:2501.12895* (2025).

[115] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. 2023. Divide & bind your attention for improved generative semantic nursing. In *34th British Machine Vision Conference 2023, BMVC 2023*.

[116] Yiyuan Li, Weizhen Zhou, Sijia Song, and Han Liu. 2024. CoMat: Aligning Text-to-Image Diffusion Model with Image-to-Text Concept Matching. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[117] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19401–19411.

[118] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. 2024. Step-aware Preference Optimization: Aligning Preference with Denoising Performance at Each Step. *arXiv preprint arXiv:2406.04314* (2024).

[119] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[120] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*. Springer, 366–384.

[121] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624* (2021).

[122] Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. 2025. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 8009–8019.

[123] Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J. Liu, and Xuanhui Wang. 2024. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878* (2024).

[124] Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2024. Statistical Rejection Sampling Improves Preference Optimization. In *The Twelfth International Conference on Learning Representations*.

[125] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–23.

[126] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *The Twelfth International Conference on Learning Representations*.

[127] Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. *stat* 1050 (2024), 21.

[128] Yunhong Lu, Qichao Wang, Hengyuan Cao, Xiaoyin Xu, and Min Zhang. 2025. Smoothed Preference Optimization via ReNoise Inversion for Aligning Diffusion Models with Varied Human Preferences. In *International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/2506.02698

[129] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems* 36 (2023).

[130] R Duncan Luce. 1959. *Individual choice behavior*. Vol. 4. Wiley New York.

[131] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. 2025. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732* (2025).

[132] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 564–572.

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

[134] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft.

[135] Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734* (2024).

[136] Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. 2024. Dynamic Prompt Optimizing for Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26627–26636.

[137] Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. 2020. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research* 21, 132 (2020), 1–62.

[138] Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. 2024. Nash Learning from Human Feedback. In *Forty-first International Conference on Machine Learning*.

[139] Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. 2020. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359* (2020).

[140] Richard Ngo, Lawrence Chan, and Sören Mindermann. 2024. The Alignment Problem from a Deep Learning Perspective. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=fh8EYKFKns

[141] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 16784–16804. https://proceedings.mlr.press/v162/nichol22a.html

[142] Andi Nika, Debmalya Mandal, Parameswaran Kamalaruban, Georgios Tzannetos, Goran Radanovic, and Adish Singla. 2024. Reward Model Learning vs. Direct Policy Optimization: A Comparative Analysis of Learning from Human Preferences. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 38145–38186. https://proceedings.mlr.press/v235/nika24a.html

[143] OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[144] Jonas Oppenlaender. 2023. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology* (2023), 1–14.

[145] Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. 2025. Inference-Time Text-to-Video Alignment with Diffusion Latent Beam Search. *arXiv preprint arXiv:2501.19252* (2025).

[146] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[147] Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. 2024. Self-Alignment of Large Language Models via Monopolylogue-based Social Scene Simulation. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 39416–39447. https://proceedings.mlr.press/v235/pang24a.html

[148] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2025. DreamBench++: A Human-Aligned Benchmark for Personalized Image Generation. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=4GSOESJrk6

[149] Vitali Petsiuk, Alexander E. Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A. Plummer, Ori Kerret, Tonio Buonassisi, Kate Saenko, Armando Solar-Lezama, and Iddo Drori. 2022. Human evaluation of text-to-image models on a multi-task benchmark. *arXiv preprint arXiv:2211.12112* (2022).

[150] Robin L Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics* 24, 2 (1975), 193–202.

[151] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

[152] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. 2023. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739* (2023).

[153] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. 2024. Video Diffusion Alignment via Reward Gradients. *arXiv preprint arXiv:2407.08737* (2024).

[154] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!. In *The Twelfth International Conference on Learning Representations*.

[155] Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. 2024. Not All Noises Are Created Equally: Diffusion Noise Selection and Optimization. *arXiv preprint arXiv:2407.14041* (2024).

[156] Zipeng Qi, Buhua Liu, Shiyan Zhang, Bao Li, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. 2024. A Simple and Efficient Baseline for Zero-Shot Generative Classification. *arXiv preprint arXiv:2412.12594* (2024).

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

[158] 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. *Advances in Neural Information Processing Systems* 36 (2023).

[159] Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems* 36 (2023).

[160] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2024. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems* 36 (2024).

[161] Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. 2024. Provably Robust DPO: Aligning Language Models with Noisy Feedback. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 42258–42274.

[162] Jie Ren, Yuhang Zhang, Dongrui Liu, Xiaopeng Zhang, and Qi Tian. 2025. Refining Alignment Framework for Diffusion Models with Intermediate-Step Preference Ranking. *arXiv preprint arXiv:2502.01667* (2025).

[163] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 1–7.

[164] Pierre Harvey Richemond, Yunhao Tang, Daniel Guo, Daniele Calandriello, Mohammad Gheshlaghi Azar, Rafael Rafailov, et al. 2024. Offline Regularised Reinforcement Learning for Large Language Models Alignment. *arXiv preprint arXiv:2405.19107* (2024).

[165] 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.

[166] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

[167] 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).

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

[169] Divya Saxena and Jiannong Cao. 2021. Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. *ACM Comput. Surv.* 54, 3, Article 63 (may 2021), 42 pages. https://doi.org/10.1145/3446374

[170] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, 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.

[171] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[172] Bernhard Schölkopf. 2022. *Causality for Machine Learning*. ACM, 765–804. https://doi.org/10.1145/3501714.3501755

[173] Ken Sekimoto. 1998. Langevin equation and thermodynamics. *Progress of Theoretical Physics Supplement* 130 (1998), 17–27.

[174] Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron, et al. 2024. Bond: Aligning llms with best-of-n distillation. *arXiv preprint arXiv:2407.14622* (2024).

[175] Lingkai Shen, Yifei Liu, Yujia Li, and Yu Tian. 2025. Efficient Diversity-Preserving Diffusion Alignment via Gradient-Informed GFlowNets. In *International Conference on Learning Representations (ICLR)*.

[176] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. 2024. Finetuning Text-to-Image Diffusion Models for Fairness. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=hnrB5YHoYu

[177] Dingyuan Shi, Yong Wang, Hangyu Li, and Xiangxiang Chu. 2024. Preference Alignment for Diffusion Model via Explicit Denoised Distribution Estimation. arXiv:2411.14871 [cs.CV]

[178] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature* 631, 8022 (2024), 755–759.

[179] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. 2025. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848* (2025).

[180] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems* 35 (2022), 9460–9471.

[181] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314* (2024).

[182] Charlie Victor Snell, Ilya Kostrikov, Yi Su, Sherry Yang, and Sergey Levine. 2023. Offline RL for Natural Language Generation with Implicit Language Q Learning. In *The Eleventh International Conference on Learning Representations*.

[183] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.

[184] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. 2024. Measuring Style Similarity in Diffusion Models. *arXiv preprint arXiv:2404.01292* (2024).

[185] 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.

[186] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

[187] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

[188] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 46280–46302. https://proceedings.mlr.press/v235/sorensen24a.html

[189] Moritz Pascal Stephan, Alexander Khazatsky, Eric Mitchell, Annie S Chen, Sheryl Hsu, Archit Sharma, and Chelsea Finn. 2024. RLVF: Learning from Verbal Feedback without Overgeneralization. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 46625–46656. https://proceedings.mlr.press/v235/stephan24a.html

[190] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

[191] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems* 36 (2023).

[192] Richard S Sutton and Andrew G Barto. 2020. *Reinforcement learning: An introduction*. MIT press.

[193] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[194] Xiaofeng Tan, Hongsong Wang, Xin Geng, and Pan Zhou. 2024. SoPo: Text-to-Motion Generation Using Semi-Online Preference Optimization. *arXiv preprint arXiv:2412.05095* (2024).

[195] Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, et al. 2024. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448* (2024).

[196] Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Remi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Avila Pires, and Bilal Piot. 2024. Generalized Preference Optimization: A Unified Approach to Offline Alignment. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 47725–47742.

[197] Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. 2024. Inference-Time Alignment of Diffusion Models with Direct Noise Optimization. *arXiv preprint arXiv:2405.18881* (2024).

[198] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[199] Masatoshi Uehara, Yifei Liu, Yujia Li, and Yu Tian. 2024. Understanding Reinforcement Learning-Based Fine-Tuning of Diffusion Models: A Tutorial and Review. *arXiv preprint arXiv:2407.13734* (2024).

[200] Masatoshi Uehara, Yifei Wu, Yujia Li, and Yu Tian. 2024. Feedback Efficient Online Fine-Tuning of Diffusion Models. In *International Conference on Machine Learning (ICML)*.

[201] Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. 2025. Inference-Time Alignment in Diffusion Models with Reward-Guided Generation: Tutorial and Review. *arXiv preprint arXiv:2501.09685* (2025).

[202] 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.

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

[204] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 8 (2024), 5362–5383. https://doi.org/10.1109/TPAMI.2024.3367329

[205] Shiyu Wang, Yuanqi Du, Xiaojie Guo, Bo Pan, Zhaohui Qin, and Liang Zhao. 2024. Controllable Data Generation by Deep Learning: A Review. *ACM Comput. Surv.* 56, 9, Article 228 (apr 2024), 38 pages. https://doi.org/10.1145/3648609

[206] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. 2018. Dataset distillation. *arXiv preprint arXiv:1811.10959* (2018).

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

[208] Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. 2024. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814* (2024).

[209] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966* (2023).

[210] Zhengwei Wang, Qi She, and Tomás E. Ward. 2021. Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy. *ACM Comput. Surv.* 54, 2, Article 37 (feb 2021), 38 pages. https://doi.org/10.1145/3439723

[211] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 893–911.

[212] Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 52588–52610.

[213] 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.

[214] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8 (1992), 229–256.

[215] Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2024. Fundamental Limitations of Alignment in Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 53079–53112.

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

[217] Xun Wu, Shaohan Huang, Guolong Wang, Jing Xiong, and Furu Wei. 2024. Multimodal large language models make text-to-image generative models align better. *Advances in Neural Information Processing Systems* 37 (2024), 81287–81323.

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

[219] Yihang Wu, Xiao Cao, Kaixin Li, Zitan Chen, Haonan Wang, Lei Meng, and Zhiyong Huang. 2024. Towards Better Text-to-Image Generation Alignment via Attention Modulation. *arXiv preprint arXiv:2404.13899* (2024).

[220] Yekun Wu, Hui Chen, Zhaofeng Zheng, Yufan Zhang, Jie Zhang, and Baining Wang. 2024. Deep Reward Supervisions for Tuning Text-to-Image Diffusion Models. In *European Conference on Computer Vision (ECCV)*.

[221] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*. 1192–1199.

[222] Weihao Xia and Jing-Hao Xue. 2023. A Survey on Deep Generative 3D-aware Image Synthesis. *ACM Comput. Surv.* 56, 4, Article 90 (nov 2023), 34 pages. https://doi.org/10.1145/3626193

[223] Xin Xie and Dong Gong. 2025. DyMO: Training-Free Diffusion Model Alignment with Dynamic Multi-Objective Scheduling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13220–13230.

[224] Zeke Xie, Issei Sato, and Masashi Sugiyama. 2021. A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima. In *International Conference on Learning Representations*.

[225] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2024. Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 54715–54754. https://proceedings.mlr.press/v235/xiong24a.html

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

[227] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2023).

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

[229] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. 2022. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=PzcvxEMzvQC

[230] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. 2018. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *Advances in Neural Information Processing Systems* 31 (2018).

[231] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 54983–54998. https://proceedings.mlr.press/v235/xu24h.html

[232] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[233] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. 2024. 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*. 8941–8951.

[234] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. 2024. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*.

[235] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.* 56, 4, Article 105 (nov 2023), 39 pages. https://doi.org/10.1145/3626235

[236] Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-Context: Multi-objective Alignment of Foundation Models with Dynamic Preference Adjustment. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 56276–56297. https://proceedings.mlr.press/v235/yang24q.html

[237] Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-Context: Multi-objective Alignment of Foundation Models with Dynamic Preference Adjustment. In *International Conference on Machine Learning*. PMLR, 56276–56297.

[238] Shentao Yang, Tianqi Chen, and Mingyuan Zhou. 2024. A Dense Reward View on Aligning Text-to-Image Diffusion with Preference. In *Forty-first International Conference on Machine Learning*.

[239] Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. 2024. Dataset Pruning: Reducing Training Data by Examining Generalization Influence. In *The Eleventh International Conference on Learning Representations*.

[240] Junliang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun Zhu. 2024. Dreamreward: Text-to-3d generation with human preference. *arXiv preprint arXiv:2403.14613* (2024).

[241] Po-Hung Yeh, Kuang-Huei Lee, and Jun cheng Chen. 2025. Training-Free Diffusion Model Alignment with Sampling Demons. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=tfemquulED

[242] Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. 2024. Jailbreak Attacks and Defenses Against Large Language Models: A Survey. *arXiv preprint arXiv:2407.04295* (2024).

[243] Ruonan Yu, Songhua Liu, and Xinchao Wang. 2023. Dataset distillation: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[244] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. 2024. Instructvideo: Instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6463–6474.

[245] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-Rewarding Language Models. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=0NphYCmgua

[246] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302* (2023).

[247] Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level Direct Preference Optimization. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 58348–58365.

[248] Jiacheng Zhang, Jie Wu, Weifeng Chen, Yatai Ji, Xuefeng Xiao, Weilin Huang, and Kai Han. 2024. Onlinevpo: Align video diffusion model with online video-centric preference optimization. *arXiv preprint arXiv:2412.15159* (2024).

[249] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. 2024. Learning Multi-dimensional Human Preference for Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8018–8027.

[250] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, Caiming Xiong, and Ran Xu. 2024. HIVE: Harnessing Human Feedback for Instructional Visual Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9026–9036.

[251] Zekun Zhang, Yifei Liu, Yujia Li, and Yu Tian. 2024. Improving Long-Text Alignment for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2410.11817* (2024).

[252] Ziyi Zhang, Sen Zhang, Yibing Zhan, Yong Luo, Yonggang Wen, and Dacheng Tao. 2024. Confronting Reward Overoptimization for Diffusion Models: A Perspective of Inductive and Primacy Biases. In *Forty-first International Conference on Machine Learning*.

[253] Zicong Zhang, Yang Zhang, Yang Song, and Tao Chen. 2024. Information Theoretic Text-to-Image Alignment. *arXiv preprint arXiv:2405.20759* (2024).

[254] Dora Zhao, Jerone Andrews, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Position: Measure Dataset Diversity, Don't Just Claim It. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=jsKr6RVDDs

[255] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. 2023. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22490–22499.

[256] Jiacheng Zheng, Yifei Wu, Yujia Li, and Yu Tian. 2024. Reward Fine-Tuning Two-Step Diffusion Models via Learning Differentiable Latent-Space Surrogate Reward. *arXiv preprint arXiv:2411.15247* (2024).

[257] Yutong Zhong, Yifei Liu, Yujia Li, and Yu Tian. 2025. Focus-N-Fix: Region-Aware Fine-Tuning for Text-to-Image Generation. *arXiv preprint arXiv:2501.06481* (2025).

[258] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems* 36 (2023).

[259] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large Language Models are Human-Level Prompt Engineers. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=92gvk82DE-

[260] Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. 2025. Golden Noise for Diffusion Models: A Learning Framework. In *International Conference on Computer Vision*.

[261] Zhenglin Zhou, Xiaobo Xia, Fan Ma, Hehe Fan, Yi Yang, and Tat-Seng Chua. 2025. DreamDPO: Aligning Text-to-3D Generation with Human Preferences via Direct Preference Optimization. In *Forty-second International Conference on Machine Learning*.

[262] Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. 2025. DSPO: Direct score preference optimization for diffusion model alignment. In *The Thirteenth International Conference on Learning Representations*.