

Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing

Sihao Wu¹, Xiaonan Si², Chi Xing³, Jianhong Wang⁴, Gaojie Jin⁵,
Guangliang Cheng¹, Lijun Zhang², Xiaowei Huang¹,

¹University of Liverpool, ²Institute of Software Chinese Academy of Sciences

³University of Edinburgh, ⁴University of Bristol, ⁵University of Exeter

{sihao.wu, guangliang.cheng, xiaowei.huang}@liverpool.ac.uk, {sixiaonan, zhanglj}@ios.ac.cn,
s2682783@ed.ac.uk, jianhong.wang@bristol.ac.uk, g.jin@exeter.ac.uk

Abstract

The integration of preference alignment with diffusion models (DMs) has emerged as a transformative approach to enhance image generation and editing capabilities. Although integrating diffusion models with preference alignment strategies poses significant challenges for novices at this intersection, comprehensive and systematic reviews of this subject are still notably lacking. To bridge this gap, this paper extensively surveys preference alignment with diffusion models in image generation and editing. First, we systematically review cutting-edge optimization techniques such as reinforcement learning with human feedback (RLHF), direct preference optimization (DPO), and others, highlighting their pivotal role in aligning preferences with DMs. Then, we thoroughly explore the applications of aligning preferences with DMs in autonomous driving, medical imaging, robotics, and more. Finally, we comprehensively discuss the challenges of preference alignment with DMs. To our knowledge, this is the first survey centered on preference alignment with DMs, providing insights to drive future innovation in this dynamic area.

1 Introduction

Advancements in Generative AI have greatly enhanced image generation and editing with diffusion models (DMs) [Ramesh *et al.*, 2022; Saharia *et al.*, 2022], addressing key challenges across applications such as media enhancement [Zhou *et al.*, 2024], driving simulation [Guan and *et al.*, 2024], and virtual reality [Qin and Weaver, 2024]. This progress broadens visual media capabilities and optimizes the production and integration of digital content on diverse platforms. Despite these advances, images generated or edited with DMs often encounter issues such as text-image misalignment, deviations from human aesthetic preferences, and propagation of content that may include toxic or biased content.

To address these challenges, scholars have employed reinforcement learning (RL) strategies, notably using reward models (RM) to infer human preferences from expert-annotated output. A key instance is the Large Language

Model (LLM), which applies reinforcement learning with human feedback (RLHF) to steer the intention towards human values and preferences [Achiam *et al.*, 2023]. In image generation and editing, many initiatives adopt various preference alignment strategies to produce images that better align with textual descriptions, aesthetics, and human values. However, the various forms of integration between DMs and preference alignment strategies pose challenges for novices in understanding the intersection of these domains. Therefore, this paper addresses the gap by systematically reviewing how DMs and RL integrate for image generation and editing, offering a concise overview and anticipating future developments.

Although there are numerous surveys on diffusion-based image generation [Zhang *et al.*, 2023] and image editing [Huang *et al.*, 2024a], few have provided a systematic and comprehensive review that thoroughly integrates DMs with RL in the field of image generation and editing. Several studies [Uehara *et al.*, 2024b; Du *et al.*, 2023; Cao *et al.*, 2023] have reviewed the integration of text-to-image models with RL; however, they tend to primarily concentrate on the applications of RL to diffusion processes or offer overarching assessments of generative AI applications. They do not systematically explore the specific integration of DMs and RL in image generation and editing, nor adequately summarize the prevailing challenges and prospective direction in this field. To address this gap, this paper provides a comprehensive review of preference alignment-based DMs, examining their theoretical foundations, developmental progress, and practical implementations in the realms of image generation and editing. Building on these insights, it identifies key challenges and outlines future research directions, offering a forward-looking perspective on potential advancements in this rapidly evolving field.

Our contributions include: i) systematically reviewing optimization techniques such as RLHF, DPO, and others, highlighting their role in the preference alignment with DMs (Sec. 3); ii) thoroughly exploring applications of DMs in autonomous driving, medical imaging, robotics, and more (Sec. 4); iii) comprehensively discussing challenges of preference alignment with DMs (Sec. 5). To our knowledge, this paper is the first survey specifically focusing on preference alignment with DMs in image generation and editing, aiming to help researchers deepen their understanding and drive further innovation.

2 Preliminaries

2.1 Diffusion Model

DDPMs employ a forward diffusion process that iteratively adds Gaussian noise to the data x_0 over T time steps. This process forms a Markov chain, where the data at each time step x_t depend only on the data from the immediately preceding step x_{t-1} [Ho *et al.*, 2020]. The forward process is governed by the following equation:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I),$$

with noise schedules $(\beta_1, \dots, \beta_T)$ determining how quickly the original data degrades. The distribution at any step t can also be computed directly as following form.

$$q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I),$$

During the denoising phase, the reverse process starts from pure noise at time step T , uses a learnable model to iteratively denoise and reconstruct the original signal x_0 , which is formulated as below:

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)),$$

The model is trained by minimizing a variational lower bound on the negative log-likelihood of the data.

$$\mathbb{E}[-\log p_\theta(x_0)] \leq \mathbb{E}_q \left[-\log p(x_T) - \sum_{t>1} \log \frac{p_\theta(x_{t-1} | x_t)}{q(x_t | x_{t-1})} \right].$$

DDIMs is a more efficient and deterministic alternative, replacing the Markov chain with a non-Markovian process [Song *et al.*, 2020]. This approach enables faster generation, accurate reconstruction, and supports applications such as image and video editing. The backward denoising step can be computed as follows:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \frac{x_t - \sqrt{1 - \alpha_t} \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} + \sqrt{1 - \alpha_{t-1}} \epsilon_\theta(x_t, t).$$

2.2 Reinforcement Learning

The preference fine-tuning pipeline typically begins with a supervised fine-tuning (SFT) stage, where the diffusion model is trained to accurately map textual prompts to images. During SFT, the focus is on improving the model’s ability to produce high-quality images that closely reflect the prompt. Once the model is sufficiently proficient at generating images, it is further refined via reinforcement learning (RL). Proximal Policy Optimization (PPO) [Schulman *et al.*, 2017] is commonly employed for this step, owing to its stability and robustness during training. The following section provides a detailed overview of PPO in this context.

PPO. In each update timestep, the PPO algorithm collects a batch of transition samples using a rollout policy $\pi_{\theta_{\text{old}}}(a_t | s_t)$ and optimizes a clipped surrogate objective, as follows:

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t [\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t)],$$

where $r_t(\theta)$ represents the probability ratio, defined as $r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}$. The term A_t is the advantage in the timestep t , estimated using Generalized Advantage Estimation (GAE) [Schulman and *et al.*, 2015], and ϵ is a hyperparameter that controls the width of the clipping interval. The

value function is trained by minimizing the following loss function:

$$L^V(\theta) = \hat{\mathbb{E}}_t \left[\frac{1}{2} \max \left((V_\theta - V_t^{\text{targ}})^2, (\text{clip}(V_\theta, V_{\text{old}} - \epsilon, V_{\text{old}} + \epsilon) - V_t^{\text{targ}})^2 \right) \right],$$

where V_t^{targ} denotes the bootstrapped value function target. Additionally, an entropy bonus is often incorporated to encourage sufficient exploration such that

$$L^H(\theta) = \hat{\mathbb{E}}_t [H[\pi_\theta](s_t)],$$

where $H[\pi_\theta](s_t)$ is the entropy of a policy π_θ at state s_t . Incorporating the above loss functions, the overall optimization objective of PPO is defined as:

$$L^{\text{PPO}}(\theta) = -L^{\text{CLIP}}(\theta) + \lambda_V L^V(\theta) - \lambda_H L^H(\theta),$$

where λ_V and λ_H are weighting coefficients to balance the importance of the value function loss and the entropy bonus. The algorithm alternates between collecting trajectory data using the policy and optimizing the collected data based on this loss function, until convergence.

3 Preference Alignment on DMs

3.1 Framework

Generally, the preference alignment process for generative models can be formulated as an RL problem, which includes the following components: a policy model, a reward model, an action space, and an environment. The policy model π_θ is a generative model taking an input as prompt x , and producing either a sequence of outputs or probability distributions y . Furthermore, we categorize all preference alignment approaches into the following taxonomy, as shown in Table 1.

Sampling. In RL, methods are categorized by how data is sampled and used to train or obtain rewards: online and offline human alignment. Online alignment refers to generative models acting as policies that collect data through interaction with their environments, acquiring reward values either from a predefined reward model or from the policies themselves. Online methods are further divided into on-policy learning with identical behavior and target policies and off-policy learning with different behavior and target policies. Offline alignment means that generative models use collected human demonstrations in prior for training.

Training Strategy. Training strategies for preference alignment in DMs include RL, DPO, and others. RL techniques, such as PPO, allow models to align with human preferences by maximizing a reward function. DPO directly optimizes the reward signal without relying on RL steps, providing a computationally efficient alternative that will be detailed later. We discuss GFlowNets-based and self-play methods in others.

Reward Feedback. Data collection strategies can be divided into two main genres: human annotation and AI-generated annotation. Human annotation means collecting preference data through manual evaluations and comparisons, which provide high-quality but often limited datasets. AI annotation leverages pre-trained models to generate large-scale annotations automatically, serving as a scalable and cost-effective

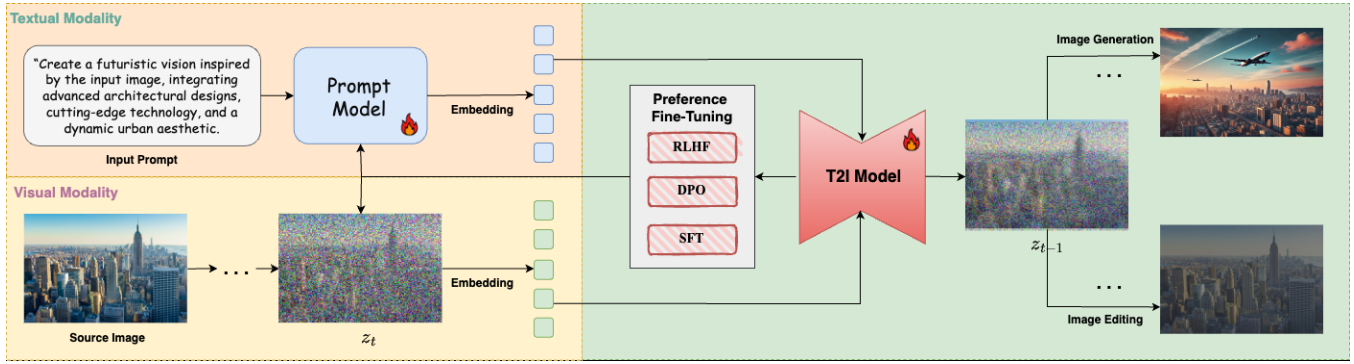


Figure 1: Preference Alignment on DM Framework: The process begins with a prompt model encoding a prompt into an embedding, and a source image is transformed into a latent representation z_t . DMs fine-tuned with RLHF, DPO, and SFT, process these embeddings to generate new images or edit existing ones while preserving structural integrity, ensuring outputs align with user preferences.

Method	Training Strategy			Task		Reward Feedback		Modality	
	DPO	RLHF	Others	Generation	Editing	Human	AI Feedback	Textual	Visual
Online Methods									
DDPO [Black <i>et al.</i> , 2023]		✓		✓		✓			✓
DPOK [Fan <i>et al.</i> , 2024]		✓		✓		✓			✓
HPS [Wu <i>et al.</i> , 2023]		✓		✓		✓			✓
ReFL [Xu <i>et al.</i> , 2024]		✓		✓		✓			✓
VersaT2I [Guo <i>et al.</i> , 2024]		✓		✓		✓			✓
Parrot [Lee <i>et al.</i> , 2025]		✓		✓		✓			✓
Diffusion-MVP [He <i>et al.</i> , 2023]		✓		✓		✓			✓
DRTun [Wu <i>et al.</i> , 2025]		✓		✓		✓			✓
LRLD [Zhang <i>et al.</i> , 2025]		✓		✓		✓			✓
InstructRL4Pix [Li <i>et al.</i> , 2024b]		✓			✓	✓			✓
HIVE [Zhang <i>et al.</i> , 2024c]		✓			✓	✓			✓
DRLDiv [Miao and <i>et al.</i> , 2024]		✓			✓	✓			✓
LfVoid [Gao and <i>et al.</i> , 2023]		✓			✓	✓		✓	✓
RAHF [Liang <i>et al.</i> , 2024]		✓		✓			✓		✓
LVLm-REFL [Wen <i>et al.</i> , 2023]		✓		✓			✓		✓
PAE [Mo <i>et al.</i> , 2024]		✓		✓				✓	
NegOpt [Ogezi and Shi, 2024]		✓		✓		✓		✓	
TextCrafter [Li <i>et al.</i> , 2024c]		✓		✓		✓		✓	
OPT2I [Mañas and <i>et al.</i> , 2024]		✓		✓			✓	✓	
RPG [Yang <i>et al.</i> , 2024a]		✓		✓			✓	✓	
DiffChat [Wang <i>et al.</i> , 2024a]		✓		✓			✓	✓	
DRaFT [Clark <i>et al.</i> , 2023]		✓		✓		✓			✓
AlignProp [Prabhudesai and <i>et al.</i> , 2024]		✓		✓		✓			✓
ELEGANT [Uehara <i>et al.</i> , 2024a]		✓		✓		✓			✓
IPO [Yang <i>et al.</i> , 2025]	✓			✓		✓			✓
OnlineVPO [Zhang <i>et al.</i> , 2024b]	✓			✓		✓			✓
DAG [Zhang <i>et al.</i> , 2024a]			✓	✓		✓			✓
RLCM [Oertel <i>et al.</i> , 2024]			✓	✓		✓			✓
SPIN-Diffusion [Yuan <i>et al.</i> , 2024a]			✓	✓		✓			✓
Offline Methods									
Diffusion-DPO [Wallace and <i>et al.</i> , 2024]	✓			✓		✓			✓
HF-T2I [Lee <i>et al.</i> , 2023]	✓			✓		✓		✓	✓
temporal-DPO [Yang <i>et al.</i> , 2024b]	✓			✓		✓			✓
PRDP [Deng <i>et al.</i> , 2024]	✓		✓	✓		✓		✓	✓
D3PO [Yang and <i>et al.</i> , 2024]	✓			✓		✓			✓
Diffusion-KTO [Li <i>et al.</i> , 2024a]	✓			✓		✓			✓
TailorPO [Ren <i>et al.</i> , 2025]	✓			✓		✓			✓

Table 1: Preference Alignment methods, which are categorized based on the approaches under study, without restricting their applicability to other domains or modalities.

approach. These two methods are often combined to create hybrid datasets that balance quality and scalability. **Modalities.** The approaches reviewed are categorized based on their optimized modalities: textual, visual, and multi-modalities. Textual modalities focus on optimizing textual prompts for improved image generation. Visual modalities mean directly updating the DM. Multi-modalities approaches combine textual and visual modalities to create comprehensive frameworks that leverage cross-modal information, resulting in more cohesive and context-aware outputs.

3.2 RLHF

Generally, RLHF involves learning a reward function from human feedback and subsequently optimizing policies with respect to the reward [Christiano *et al.*, 2017]. The training process for RLHF consists of three stages. Firstly, the policy model π_θ interacts with the environment, and its parameters are updated using reinforcement learning techniques. Next, pairs of output segments generated by π_θ are selected and presented to human annotators for preference comparison. Finally, the model’s parameters are optimized based on a reward signal r , which is elicited from the human-provided comparisons.

As briefly introduced in Sec. 2.1, the formulation of text-to-image DMs can be described as a Markov Chain Process (MDP). There are two typical related works in RLHF for DMs, DDPO [Black *et al.*, 2023] and DPOK [Fan *et al.*, 2024]. Consider taking (i, x_i, c) as the state space, and define the action as the next hierarchy x_{i-1} to go to. Then, Eq. 1 naturally defines a stochastic policy: the stochasticity of the policy comes from $\sqrt{\beta_i z_i}$. Accordingly, the policy follows a Gaussian distribution $\mathcal{N}(s_\theta^*(i, x_i, c), \beta_i)$ with the mean determined by $s_\theta^*(i, x_i, c)$ and variance β_i :

$$\pi_\theta(x_{i-1} | x_i) \sim \mathcal{N}\left(\frac{1}{\sqrt{1 - \beta_i}}(x_i + \beta_i s_\theta(i, x_i, c)), \beta_i\right).$$

Given this formulation, DDPO [Black *et al.*, 2023] directly maximize the expected reward, $J_{\text{DDPO}} = \mathbb{E}_\theta[r(x_0, c)]$, using either the REINFORCE or PPO algorithms, without applying regularization.

$$\nabla_\theta J_{\text{DDPO}} = \mathbb{E} \left[\sum_{t=0}^T \nabla_\theta \log p_\theta(x_{t-1} | x_t, c) r(x_0, c) \right].$$

In contrast, DPOK [Fan *et al.*, 2024], optimizes the regularized reward objective as below. They proposed a clipped gradient algorithm inspired by the PPO objective and showed that incorporating regularization improves generation quality over unregularized versions.

$$J_{\text{DPOK}} = \mathbb{E}_\theta[r(x_0, c)] - \beta \mathbb{E}_{p(z)}[\text{KL}(p_\theta(x_0 | z) \| p_{\text{pre}}(x_0 | z))].$$

Moreover, the Human Preference Score (HPS) aligns T2I models with human preferences by leveraging a classifier trained on a large-scale dataset of human choices [Wu *et al.*, 2023]. Xu *et al.* [2024] applies Reward Feedback Learning (ReFL) for better synthesis, which trains on 137,000 expert comparisons. VersaT2I [Guo *et al.*, 2024] adopts a self-training approach with Low-Rank Adaptation (LoRA) to

address multiple image quality aspects, introducing a mixture of LoRA models for improved quality consistency. The Parrot framework [Lee *et al.*, 2025] employs a multi-reward optimization algorithm using reinforcement learning to balance aesthetics, alignment, sentiment, and human preferences through Pareto-optimal selection, achieving superior image quality. Diffusion-MVP integrates visual rewards into text-to-image generation including Visual Market, Visual Aesthetic, and CLIP relevance [He *et al.*, 2023]. To address the depth-efficiency trade-off, Deep Reward Tuning (DRTune) optimizes reward functions in DMs by controlling early denoising steps [Wu *et al.*, 2025]. LRLD [Zhang *et al.*, 2025] introduces a scalable RL framework that focuses on improving text-to-image alignment with human preferences, fairness, and object composition across a diverse set of reward functions.

On the other hand, recent efforts in fine-tuning DMs have expanded to **image editing**. InstructRL4Pix [Li *et al.*, 2024b] introduces a reinforcement learning framework for image editing, utilizing attention maps and Proximal Policy Optimization to enhance feature localization and preservation. Incorporating human preferences into image editing workflows, HIVE takes advantage of human feedback to train a reward function [Zhang *et al.*, 2024c]. To address the challenge of generative diversity, DRLDiv [Miao and et al., 2024] is a method that quantifies the overlap between generative and reference distributions. LfVoid exemplifies multi-modal integration by leveraging pre-trained T2I models to generate visual objectives for robotic learning tasks [Gao and et al., 2023]. By editing images based on natural language instructions, it guides RL agents without in-domain data, showcasing strong performance in simulated and real-world tasks and the potential of generative models in robotics.

The above methods underscore the growing importance of human-centered optimization strategies in T2I generation, while further advancements have introduced exploration into **AI feedback** mechanisms. Specifically, LLMs and LVLMs have demonstrated to function with minimal human intervention as the evaluation strategy. Liang *et al.* [2024] introduces RAHF, which provides automatic feedback on implausible regions and misaligned keywords, facilitating training data refinement and region inpainting. LVLM-REFL integrates LVLMs to evaluate alignment between generated images and input texts [Wen *et al.*, 2023]. Additionally, this method incorporates an iterative correction process during inference, applying image-editing algorithms guided by the LVLM.

Recently, there has been a growing surge of interest in the precise refinement of **text modality**. Prompt Auto-Editing (PAE) utilizes a two-stage training process combining supervised fine-tuning and online RL for dynamic prompts control [Mo *et al.*, 2024]. This approach enables the iterative refinement of prompts of the weights and injection time steps of each word. In contrast, NegOpt optimizes negative prompts to improve image aesthetics and fidelity [Ogezi and Shi, 2024]. TextCrafter [Li *et al.*, 2024c] focuses on the text encoder, leveraging reward functions to optimize performance without paired datasets. The integration of LLMs and VLMs as carriers or alternative representations of reward functions has introduced novel strategies for optimizing textual modality. OPT2I [Mañas and et al., 2024] exemplifies

this by employing automatic prompt optimization, where an LLM iteratively refines prompts to maximize a consistency score while preserving image quality and improving recall between generated and real data. Similarly, the RPG framework [Yang et al., 2024a] advances text-to-image generation by using multimodal LLMs to decompose complex prompts into detailed subprompts, allocate these across image regions via chain-of-thought planning, and apply complementary regional diffusion for precise image synthesis. DiffChat introduces an interactive image creation framework that leverages the InstructPE dataset and RL to integrate feedback on aesthetics, user preferences, and content integrity, enabling efficient generation of high-quality target prompts [Wang et al., 2024a]. These approaches underscore the growing importance of LLMs and VLMs in enhancing prompt optimization and aligning text-to-image models with user-driven goals.

Another common approach is direct reward backpropagation. Direct reward backpropagation simplifies implementation, especially with a pre-trained differentiable reward model, and accelerates training by directly backpropagating reward gradients. DRaFT propagates gradients directly from reward functions to update policies [Clark et al., 2023]. It has proven effective across a variety of reward functions, significantly enhancing the aesthetic quality of images generated by Stable Diffusion. Prabhudesai and et al. [2024] introduces AlignProp, reformulating denoising in text-to-image DMs as a differentiable recurrent policy. It uses randomized truncated backpropagation to mitigate over-optimisation, randomly selecting the denoising step for reward backpropagation. Furthermore, Uehara et al. [2024a] propose ELEGANT to directly optimize entropy-enhanced rewards with neural SDEs, effectively preventing excessive reward maximization.

3.3 DPO and its Variants

This section explains each preference alignment method like DPO and its variants. One notable drawback of RLHF is that the RL optimization step often demands significant computational resources [Winata et al., 2024], such as those required for PPO. To address this, DPO, recently proposed by [Rafailov et al., 2024], offers a promising alternative by bypassing the reward modelling stage and eliminating the need for reinforcement learning, which has garnered considerable attention. The core idea of DPO is based on the observation that, given a reward function $r(x, y)$, the DPO objective can be described as follows:

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

Wallace and et al. [2024] investigates human preference learning in text-to-image DMs and introduces Diffusion-DPO, a novel approach to aligning these models with user preferences. By adapting the DPO framework, Diffusion-DPO reformulates the objective of incorporating the evidence lower bound, enabling a differentiable and efficient optimization process. HF-T2I [Lee et al., 2023] proposes a reward-weighted likelihood maximization framework by collecting binary feedback. As for temporal rewards DPO, Yang et

al. [2024b] introduces a dense reward perspective with temporal discounting in DPO-style objectives, prioritizing early steps in the generation process to improve alignment efficiency. PRDP [Deng et al., 2024] reformulates RLHF as a supervised regression task, allowing stable fine-tuning in large prompt datasets. In response to the provided text prompts, the method selects two image candidates and trains the DM to predict reward differences between them based on their denoising trajectories. The model parameters are updated based on the prediction’s mean square error loss. In DMs, the storage of gradients for multiple latent image representations, which are substantially larger than word embeddings, imposes memory requirements that are frequently infeasible. In [Yang and et al., 2024] (D3PO), the denoising process is reframed as a multi-step MDP to reduce computational overhead and enable DPO-based fine-tuning of DMs with human feedback. Using a pre-trained model to represent the action value function Q , the DPO framework is extended to allow direct parameter updates at each denoising step, eliminating the need for a reward model and significantly reducing computational costs. Diffusion-KTO introduces a novel approach for aligning text-to-image models by using only per-image binary feedback, removing the need for costly pairwise preference data [Li et al., 2024a]. TailorPO [Ren et al., 2025] ranks intermediate noisy samples by step-wise reward and efficiently addresses gradient direction issues.

Unlike DPO, which operates on a fixed offline dataset, **Online DPO** dynamically constructs preference datasets by leveraging model outputs and various labelers, such as pre-trained reward models, LLM judges, or even the trained model itself via prompting. To address off-policy optimization and scalability challenges in existing video preference learning frameworks, Zhang et al. [2024b] propose an online DPO algorithm (OnlineVPO), which continually refines video generation quality by incorporating feedback in real time. Moreover, Iterative Preference Optimization (IPO) extends the concept of integrating human feedback to enhance generated video quality [Yang et al., 2025]. Specifically, IPO utilizes a critic model to evaluate and rank video outputs pairwise, similar to the approach used in DPO, thereby refining the generative process based on explicit preference signals.

3.4 Others

In addition to RL and DPO, some literature uses GFlowNets [Bengio et al., 2023], also coined as the consistency model, to achieve direct reward backpropagation and alignment. Regarding the GFlowNet-based algorithm, the formulation of DM is directly connected to the definition of GFlowNet MDP, as highlighted in [Zhang et al., 2022]. Importantly, the conditional distribution of the denoising process, $p_{\theta}(\mathbf{x}_{T-t-1} | \mathbf{x}_{T-t})$, aligns with the GFlowNet forward policy $P_F(s_{t+1} | s_t)$. Similarly, the conditional distribution of the diffusion process, $q(\mathbf{x}_{T-t} | \mathbf{x}_{T-t-1})$, corresponds to the GFlowNet backward policy $P_B(s_t | s_{t+1})$ [Lahlou et al., 2023]. Building on the above works, Zhang et al. [2024a] proposes the DAG to post-train DMs with black-box property functions. Further, they propose a KL-based objective for optimizing GFlowNets to offer improved sample efficiency. Moreover, RLCM proposes a framework for fine-tuning con-

sistency models via RL, framing the iterative inference process as an MDP [Oertel *et al.*, 2024]. This approach speeds up training and inference while maintaining the efficiency of consistency models, enhancing performance across metrics. However, as a caveat, this property might also pose challenges in effectively generating high-reward samples beyond the training data. This implies that these approaches may not be suitable when accurate reward feedback is readily available without learning. Hence, we generally recommend using them when reward functions are unknown. Innovative self-play methods like SPIN-Diffusion [Yuan *et al.*, 2024a] remove the need for human preference data, achieving superior performance in text-to-image tasks. By engaging the model in a self-improvement process, where the diffusion model engages in competition with its earlier versions, facilitating an iterative self-improvement process. SPIN-Diffusion ensures efficient fine-tuning that leads to improved alignment with target distributions.

4 Applications

This section underscores the critical role of preference alignment in diffusion models across various domains, including medical imaging, robotics, and autonomous driving. As shown in Figure 2, most applications rely on fine-tuning modules integrated into the denoising process, ensuring optimized generation conditioned on the specific requirements of each domain.

4.1 Medical Imaging

Segmentation Medical image segmentation partitions images into distinct regions of interest, playing a significant role in diagnosis and image-guided surgery. However, lesions and organs often exhibit ambiguity, making them difficult to be identified from the background. To this end, Wu and *et al.* [2022] proposed MedSegDiff to enhance the step-wise regional attention in DDPMs. Their approach optimizes the lesion’s feature map during the feature extraction phase in the diffusion encoder to enhance its visibility. Building on this work, Wu and *et al.* [2024] proposed MedSegDiff-V2, integrating diffusion processes and transformers with dual conditioning, anchor and semantic, guided by SS-Former to improve medical image segmentation. Furthermore, Rahman *et al.* [2023] utilized the stochastic sampling mechanism to generate a distribution of segmentation masks with cost-effective additional training.

Synthesis & Reconstruction The scarcity of public medical imaging datasets, combined with strict privacy regulations, underscores the urgent need for advanced image synthesis techniques. Daum and *et al.* [2024] presented latent DMs designed to generate synthetic images conditioned on medical attributes while protecting patient privacy through differentially private training. These models are pre-trained on public datasets and then fine-tuned with differential privacy using the UK Biobank dataset. Subsequently, Han *et al.* [2024] fused RLHF strategies and proposed CXRL to optimize the generation result. Moreover, preference fine-tuned DMs are widely used to reconstruct high-quality medical imaging data by integrating feature engineering into the denoising pro-

cess [Xie and Li, 2022], effectively addressing issues such as missing data, noise interference, and low resolution.

4.2 Robotics

Motion Generation DMs have enhanced robustness and adaptability in motion generation by unifying offline learning with preference alignment and physics-informed control. Offline diffusion learning, combined with online preference alignment, improves quadrupedal locomotion by ensuring stability, precision, and zero-shot generalization to real-world robots [Yuan *et al.*, 2024b]. These advancements showcase the power of DMs to capture complex motion dynamics in robotic motion learning. By generating diverse and realistic trajectories, these models enable robots to perform robust and adaptive behaviors, paving the way for more efficient and capable robotic systems.

Trajectory Planning DMs have significantly advanced trajectory planning in robotics by facilitating preference alignment and decision-making across complex, multi-objective scenarios. Janner *et al.* [2022] embedded the trajectory optimization process into the denoising process, enabling long-horizon planning while addressing alignment inconsistencies between conditions and generated trajectories. Preference-conditioned DMs further enhance multi-task decision-making by maximizing mutual information to improve trajectory alignment [Yu *et al.*, 2024].

Policy Learning & Others DMs are revolutionizing policy learning in robotics by tackling challenges like multi-task learning [Wang *et al.*, 2024b], data scarcity [Liu *et al.*, 2024b], and aligning policies with human preferences [Dong *et al.*, 2023]. Furthermore, preference fine-tuned DMs demonstrate remarkable potential not only in advancing the design and control of soft robots [Wang *et al.*, 2023], but also in the training of sophisticated world simulators [Yang *et al.*, 2023]. These simulators leverage DMs to generate high-fidelity, diverse, and contextually accurate virtual environments, providing a robust framework for training and testing autonomous systems in a wide range of scenarios.

4.3 Autonomous Driving

Data Generation Autonomous driving models require a large amount of high-quality and complex training data. However, real data are costly to collect with restricted coverage, so synthesizing data through generative techniques becomes an effective alternative. To improve data quality, specifically spatio-temporal consistency and resolution, RLHF, SFT, and DPO have all been shown to be remarkably effective methods. Several methods have been proved to generate training data for multi-view videos [Li *et al.*, 2023], panoramic videos [Wen and *et al.*, 2024], etc.

Decision Making Traditional auto-driving decision-making processes rely on predicting the future behavior of traffic participants, such as vehicles and pedestrians, and using deterministic or rule-based methods for path planning. However, these approaches are limited by predicting only a few possible scenarios, failing to account for the uncertainty in real-world traffics. To mitigate these issues, several methods are proposed to optimize the decision-making process

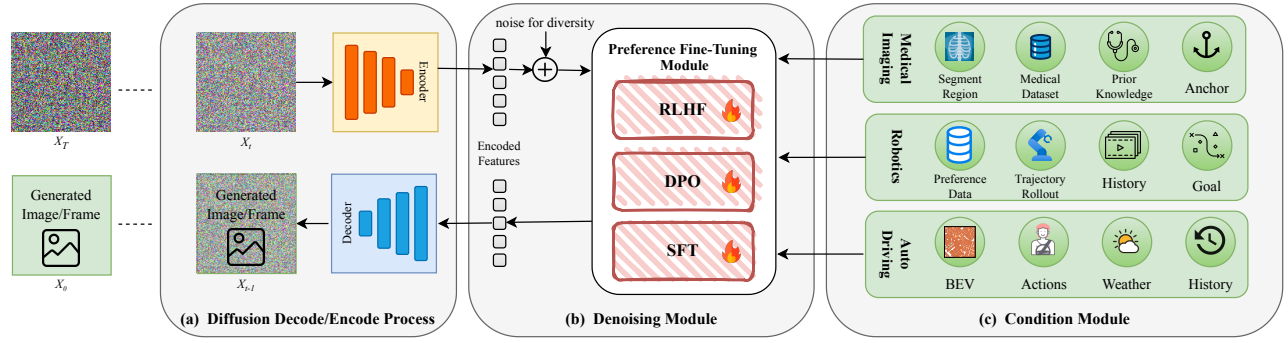


Figure 2: General paradigm of preference alignment for various **condition modules** on DM applications.

by RLHF [Huang *et al.*, 2024b]; to accelerate the decision-making process by shortening denoising process [Liao *et al.*, 2024]; and to increase the security of the decision-making process by merging RLHF and SFT [Liu *et al.*, 2024a].

World Model Similar to robotics, world models hold great promise in autonomous driving, where accurately predicting diverse movements is essential for making safe and effective driving decisions [Feng *et al.*, 2025]. Recently, generalized diffusion-based world models have attracted considerable attention [Chen *et al.*, 2024].

4.4 Others

The above discussion demonstrates that preference alignment on DMs exhibits remarkable potential in domains with a critical demand for large-scale training data. Beyond the above domains, in the field of biology, DMs can not only predict protein structures [Wu *et al.*, 2024] but also optimize their stability and functionality, expediting experimental workflows. In game simulation [Menapace *et al.*, 2024], these models generate realistic virtual worlds, simulate complex agent behaviors, and streamline gameplay mechanics, ultimately enhancing user immersion and engagement. Furthermore, in the content creation sphere, DMs facilitate the production of high-quality, preference-aligned outputs—spanning digital art, music, and other creative mediums—satisfying diverse aesthetic and stylistic requirements [Lin *et al.*, 2023]. Taken together, these applications illustrate how preference alignment on DMs can adapt across multidisciplinary domains, opening up new avenues for innovation and collaboration.

5 Challenges and Future Directions

The field currently faces several fundamental challenges. The predominant pressing is the computational burden of preference fine-tuning, particularly in RLHF approaches that require loads of resources for both training and inference. This is along with the challenge of collecting high-quality human preference data, which is both time-consuming and expensive. Besides, maintaining semantic consistency while optimizing for multiple diverse objectives—such as image quality, text alignment, and aesthetic appeal—remains difficult. Furthermore, the subjective nature of human preferences also makes it challenging to design effective reward functions that can reliably capture and reproduce these preferences in scale.

Looking ahead, several promising directions emerge for advancing the field. One key area is developing more efficient fine-tuning methods that can reduce computational requirements while maintaining performance. This includes exploring parameter-efficient approaches like LoRA and techniques that can minimize the need for extensive human feedback. Another crucial direction is improving the integration of multiple modalities, particularly in combining text and image understanding to achieve better alignment with human preferences. This could involve taking the use of LLMs to better interpret and implement human preferences in image generation. The field is also moving toward more practical applications. There is growing interest in adapting these techniques for specific domains like medical imaging, autonomous driving, and robotics. This application-oriented direction requires more robust evaluation metrics and reliable models for real-world scenarios. Also, there is an increasing emphasis on making fine-tuning more interpretable and controllable, for a better understanding of how preferences affect generation and enabling more fine-grained control over different aspects of the output, respectively.

As these models grow more powerful and widely deployed, safety and ethical considerations become increasingly important. This includes developing evaluation metrics and approaches, to ensure alignment with human values, preventing misuse, and maintaining privacy and security during the fine-tuning process. Success in these areas will be essential milestones toward the responsible advancement of preference fine-tuning in DMs.

6 Conclusion

This survey provides a comprehensive review of preference alignment methods for DMs, focusing on integrating RL-related techniques like RLHF and DPO to enhance alignment with human preferences. We categorize approaches into sampling, training strategies, data collection, and modalities, as well as highlight their applications in medical imaging, autonomous driving and robotics. Despite the temporary success, challenges such as computational efficiency and preference modelling still remain. Future research should address these issues, and continue to explore scalable algorithms, multimodal capabilities, and ethical considerations to expand the full potential of generative AI.

References

- Josh Achiam, Steven Adler, and et al. Gpt-4 technical report. *ArXiv*, 2023.
- Yoshua Bengio, Salem Lahlou, and et al. Gflownet foundations. *JMLR*, 2023.
- Kevin Black, Michael Janner, and et al. Training diffusion models with reinforcement learning. *ArXiv*, 2023.
- Yuanjiang Cao, Quan Z. Sheng, and et al. Reinforcement learning for generative AI: A survey. *Arxiv*, 2023.
- Yuntao Chen, Yuqi Wang, and et al. Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. *ArXiv*, 2024.
- Paul F Christiano, Jan Leike, and et al. Deep reinforcement learning from human preferences. *Neurips*, 2017.
- Kevin Clark, Paul Vicol, and et al. Directly fine-tuning diffusion models on differentiable rewards. *ArXiv*, 2023.
- Deniz Daum and et al. On differentially private 3d medical image synthesis with controllable latent diffusion models. *MICCAI Workshop on Deep Generative Models*, 2024.
- Fei Deng, Qifei Wang, and et al. Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. *CVPR*, 2024.
- Zibin Dong, Yifu Yuan, and et al. Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model. *ArXiv*, 2023.
- Hongyang Du, Ruichen Zhang, and et al. Beyond deep reinforcement learning: A tutorial on generative diffusion models in network optimization. *ArXiv*, 2023.
- Ying Fan, Olivia Watkins, and et al. Reinforcement learning for fine-tuning text-to-image diffusion models. *Neurips*, 2024.
- Tuo Feng, Wenguan Wang, and et al. A survey of world models for autonomous driving. *ArXiv*, 2025.
- Jialu Gao and et al. Can pre-trained text-to-image models generate visual goals for reinforcement learning? *Neurips*, 2023.
- Yanchen Guan and et al. World models for autonomous driving: An initial survey. *T-IV*, 2024.
- Jianshu Guo, Wenhao Chai, and et al. Versat2i: Improving text-to-image models with versatile reward. *ArXiv*, 2024.
- Woojung Han, Chanyoung Kim, and et al. Advancing text-driven chest x-ray generation with policy-based reinforcement learning. *MICCAI*, 2024.
- Huiguo He, Tianfu Wang, and et al. Learning profitable nft image diffusions via multiple visual-policy guided reinforcement learning. *ACMMM*, 2023.
- Jonathan Ho, Ajay Jain, and et al. Denoising diffusion probabilistic models. *Neurips*, 2020.
- Yi Huang, Jiancheng Huang, and et al. Diffusion model-based image editing: A survey. *Arxiv*, 2024.
- Zhiyu Huang, Xinshuo Weng, and et al. Gen-drive: Enhancing diffusion generative driving policies with reward modeling and reinforcement learning fine-tuning. *ArXiv*, 2024.
- Michael Janner, Yilun Du, and et al. Planning with diffusion for flexible behavior synthesis. *ArXiv*, 2022.
- Salem Lahlou, Tristan Deleu, and et al. A theory of continuous generative flow networks. *ICML*, 2023.
- Kimin Lee, Hao Liu, and et al. Aligning text-to-image models using human feedback. *ArXiv*, 2023.
- Seung Hyun Lee, Yinxiao Li, and et al. Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. *ECCV*, 2025.
- Xiaofan Li, Yifu Zhang, and et al. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *ArXiv*, 2023.
- Shufan Li, Konstantinos Kallidromitis, and et al. Aligning diffusion models by optimizing human utility. *ArXiv*, 2024.
- Tiancheng Li, Jinxiu Liu, and et al. Instructrl4pix: Training diffusion for image editing by reinforcement learning. *ArXiv*, 2024.
- Yanyu Li, Xian Liu, and et al. Textcraftor: Your text encoder can be image quality controller. *CVPR*, 2024.
- Youwei Liang, Junfeng He, and et al. Rich human feedback for text-to-image generation. *CVPR*, 2024.
- Bencheng Liao, Shaoyu Chen, and et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. *ArXiv*, 2024.
- Chen-Hsuan Lin, Jun Gao, and et al. Magic3d: High-resolution text-to-3d content creation. *CVPR*, 2023.
- Jiaqi Liu, Peng Hang, and et al. Ddm-lag: A diffusion-based decision-making model for autonomous vehicles with lagrangian safety enhancement. *ArXiv*, 2024.
- Songming Liu, Lingxuan Wu, and et al. Rdt-1b: a diffusion foundation model for bimanual manipulation. *ArXiv*, 2024.
- Oscar Mañas and et al. Improving text-to-image consistency via automatic prompt optimization. *ArXiv*, 2024.
- Willi Menapace, Aliaksandr Siarohin, and et al. Promptable game models: Text-guided game simulation via masked diffusion models. *TOG*, 2024.
- Zichen Miao and et al. Training diffusion models towards diverse image generation with rl. *CVPR*, 2024.
- Wenyi Mo, Tianyu Zhang, and et al. Dynamic prompt optimizing for text-to-image generation. *CVPR*, 2024.
- Owen Oertell, Jonathan Daniel Chang, and et al. RL for consistency models: Reward guided text-to-image generation with fast inference. *RLC*, 2024.
- Michael Ogezi and Ning Shi. Optimizing negative prompts for enhanced aesthetics and fidelity in text-to-image generation. *ArXiv*, 2024.
- Mihir Prabhudesai and et al. Aligning text-to-image diffusion models with reward backpropagation, 2024.

690	Xue Qin and Garrett Weaver. Utilizing generative ai for vr	Xiaoshi Wu, Keqiang Sun, and et al. Human preference score:	742
691	exploration testing: A case study. <i>ASEW</i> , 2024.	Better aligning text-to-image models with human prefer-	743
692	Rafael Rafailov, Archit Sharma, and et al. Direct preference	ence. <i>ICCV</i> , 2023.	744
693	optimization: Your language model is secretly a reward	Kevin E Wu, Kevin K Yang, and et al. Protein structure gener-	745
694	model. <i>Neurips</i> , 2024.	ation via folding diffusion. <i>Nature communications</i> , 2024.	746
695	Aimon Rahman, Jeya Maria Jose Valanarasu, and et al. Am-	Xiaoshi Wu, Yiming Hao, and et al. Deep reward supervi-	747
696	biguous medical image segmentation using diffusion mod-	sions for tuning text-to-image diffusion models. <i>ECCV</i> ,	748
697	els. <i>CVPR</i> , 2023.	2025.	749
698	Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu,	Yutong Xie and Quanzheng Li. Measurement-conditioned	750
699	and Mark Chen. Hierarchical text-conditional image gen-	denoising diffusion probabilistic model for under-sampled	751
700	eration with clip latents. <i>ArXiv</i> , 2022.	medical image reconstruction. <i>MICCAI</i> , 2022.	752
701	Jie Ren, Yuhang Zhang, and et al. Refining alignment frame-	Jiazheng Xu, Xiao Liu, and et al. Imagereward: Learning	753
702	work for diffusion models with intermediate-step prefer-	and evaluating human preferences for text-to-image gener-	754
703	ence ranking. <i>ArXiv</i> , 2025.	ation. <i>Neurips</i> , 2024.	755
704	Chitwan Saharia, William Chan, and et al. Photorealistic text-	Kai Yang and et al. Using human feedback to fine-tune diffu-	756
705	to-image diffusion models with deep language understand-	sion models without any reward model. <i>CVPR</i> , 2024.	757
706	ing. <i>Neurips</i> , 2022.	Mengjiao Yang, Yilun Du, and et al. Learning interactive	758
707	John Schulman and et al. High-dimensional continuous con-	real-world simulators. <i>ArXiv</i> , 2023.	759
708	trol using generalized advantage estimation. <i>ArXiv</i> , 2015.	Ling Yang, Zhaochen Yu, and et al. Mastering text-to-image	760
709	John Schulman, Filip Wolski, and et al. Proximal policy op-	diffusion: Recaptioning, planning, and generating with	761
710	timization algorithms. <i>ArXiv</i> , 2017.	multimodal llms. <i>ICML</i> , 2024.	762
711	Jiaming Song, Chenlin Meng, and et al. Denoising diffusion	Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense	763
712	implicit models. <i>ArXiv</i> , 2020.	reward view on aligning text-to-image diffusion with pref-	764
713	Masatoshi Uehara, Yulai Zhao, and et al. Fine-tuning of	erence. <i>ArXiv</i> , 2024.	765
714	continuous-time diffusion models as entropy-regularized	Xiaomeng Yang, Zhiyu Tan, and et al. Ipo: Iterative pref-	766
715	control. <i>ArXiv</i> , 2024.	erence optimization for text-to-video generation. <i>ArXiv</i> ,	767
716	Masatoshi Uehara, Yulai Zhao, and et al. Understanding rein-	2025.	768
717	forcement learning-based fine-tuning of diffusion models:	Xudong Yu, Chenjia Bai, and et al. Regularized condi-	769
718	A tutorial and review. <i>Arxiv</i> , 2024.	tional diffusion model for multi-task preference alignment.	770
719	Bram Wallace and et al. Diffusion model alignment using	<i>ArXiv</i> , 2024.	771
720	direct preference optimization. <i>CVPR</i> , 2024.	Huizhuo Yuan, Zixiang Chen, and et al. Self-play fine-tuning	772
721	Tsun-Hsuan Johnson Wang, Juntian Zheng, and et al. Dif-	of diffusion models for text-to-image generation. <i>ArXiv</i> ,	773
722	fusebot: Breeding soft robots with physics-augmented gen-	2024.	774
723	erative diffusion models. <i>Neurips</i> , 2023.	Xinyi Yuan, Zhiwei Shang, and et al. Preference aligned dif-	775
724	Jiapeng Wang, Chengyu Wang, and et al. Diffchat: Learning	fusion planner for quadrupedal locomotion control. <i>ArXiv</i> ,	776
725	to chat with text-to-image synthesis models for interactive	2024.	777
726	image creation. <i>ArXiv</i> , 2024.	Dinghuai Zhang, Ricky TQ Chen, and et al. Unifying gener-	778
727	Yixiao Wang, Yifei Zhang, and et al. Sparse diffusion policy:	ative models with gflownets and beyond. <i>ArXiv</i> , 2022.	779
728	A sparse, reusable, and flexible policy for robot learning.	Chenshuang Zhang, Chaoning Zhang, and et al. Text-to-	780
729	<i>ArXiv</i> , 2024.	image diffusion models in generative AI: A survey. <i>Arxiv</i> ,	781
730	Yuqing Wen and et al. Panacea: Panoramic and controllable	2023.	782
731	video generation for autonomous driving. <i>CVPR</i> , 2024.	Dinghuai Zhang, Yizhe Zhang, and et al. Improving	783
732	Song Wen, Guian Fang, and et al. Improving compositional	gflownets for text-to-image diffusion alignment. <i>ArXiv</i> ,	784
733	text-to-image generation with large vision-language mod-	2024.	785
734	els. <i>ArXiv</i> , 2023.	Jiacheng Zhang, Jie Wu, and et al. Onlinevpo: Align video	786
735	Genta Indra Winata, Hanyang Zhao, and et al. Preference tun-	diffusion model with online video-centric preference opti-	787
736	ing with human feedback on language, speech, and vision	mization. <i>ArXiv</i> , 2024.	788
737	tasks: A survey. <i>ArXiv</i> , 2024.	Shu Zhang, Xinyi Yang, and et al. Hive: Harnessing human	789
738	Junde Wu and et al. Medsegdiff: Medical image segmenta-	feedback for instructional visual editing. <i>CVPR</i> , 2024.	790
739	tion with diffusion probabilistic model. <i>ArXiv</i> , 2022.	Yinan Zhang, Eric Tzeng, and et al. Large-scale reinforce-	791
740	Junde Wu and et al. Medsegdiff-v2: Diffusion-based medical	ment learning for diffusion models. <i>ECCV</i> , 2025.	792
741	image segmentation with transformer. <i>AAAI</i> , 2024.	Pengyuan Zhou, Lin Wang, and et al. A survey on gener-	793
		ative ai and llm for video generation, understanding, and	794
		streaming. <i>arXiv</i> , 2024.	795