多模态嵌入的全局对齐增强下的基于强化学习的扩散模型

尤昊辰[1]*，刘宝静[2]
[1]哥伦比亚大学，人文与科学研究生院，纽约市，10025
[2]河北传媒学院，人工智能学院，石家庄市，051430
hy2854@columbia.edu
liubj@hebic.edu.cn

摘要

扩散模型作为新一代生成模型，在文本引导图像生成任务中展现出卓越性能。然而，现有预训练扩散模型的训练目标通常无法直接对齐用户偏好或下游任务需求，导致其生成结果难以兼顾图文语义一致性与主观美学质量。为此，近年来研究者提出将强化学习引入扩散微调过程，使模型在奖励信号引导下优化生成策略，代表性方法如策略优化扩散模型与去噪扩散策略优化已取得显著成果。然而，此类方法所依赖的奖励函数多为黑盒式打分器，难以捕捉生成图像与输入文本之间的结构性语义关系，缺乏对模态间对齐结构的显式建模。为解决上述问题，本文提出一种融合强化学习与结构对齐正则的文本引导扩散模型微调方法**GARD**（*Geometry-Aligned Reinforced Diffusion*）。该方法在强化学习微调框架下，引入一种基于嵌入空间几何结构的对齐正则项，即通过计算图像与文本嵌入向量构成的平行多面体体积，衡量其语义对齐程度，并与奖励信号与散度正则共同构成统一优化目标，从而在提升生成质量的同时增强多模态语义一致性。实验结果表明，**GARD** 在多个公开数据集上相较于现有方法在语义一致性、审美得分与训练稳定性等方面均实现显著提升，验证了本文方法在多模态结构对齐建模与强化学习微调融合方面的有效性与通用性。

关键词： 大语言模型；强化学习；生成模型；多模态对齐；扩散模型

Reinforcement Learning-based Diffusion Model Enhanced by Global Alignment of Multimodal Embeddings

Haochen You[1], Baojing Liu[2]
[1]Graduate School of Arts and Sciences, Columbia University, New York, 10025
[2]School of Artificial Intelligence, Hebei Institute of Communications, Shijiazhuang, 051430
hy2854@columbia.edu
liubj@hebic.edu.cn

Abstract

As a new generation of generative models, diffusion models have demonstrated remarkable performance in text-guided image generation tasks. However, existing pretrained diffusion models are typically optimized with training objectives that do not directly align with user preferences or downstream task requirements. As a result, the generated outputs often struggle to balance semantic consistency between text and image with subjective aesthetic quality. To address this limitation, recent research has introduced reinforcement learning into the diffusion fine-tuning process, allowing the generation policy to be optimized under the guidance of reward signals. Representative approaches

*通讯作者，Corresponding Author

第二十四届中国计算语言学大会论文集，第285页-第297页，济南，中国，2025年8月11日至14日。

such as policy optimization for diffusion models and denoising diffusion policy optimization have achieved promising results. Nevertheless, the reward functions used in such methods are often black-box scorers, making it difficult to capture the structural semantic relationships between generated images and input text. These approaches also lack explicit modeling of alignment structures across modalities. To tackle the above challenges, this paper proposes a novel text-guided diffusion model fine-tuning framework that integrates reinforcement learning with structural alignment regularization, named **GARD** (*Geometry-Aligned Reinforced Diffusion*). Under the reinforcement learning fine-tuning paradigm, **GARD** introduces a geometry-based alignment regularization term in the embedding space. Specifically, it measures semantic alignment by computing the volume of the parallelotope formed by image and text embedding vectors. This alignment loss is jointly optimized with the reward signal and a divergence regularizer, forming a unified objective that enhances both generation quality and multimodal semantic consistency. Experimental results show that **GARD** significantly outperforms existing methods on multiple public datasets in terms of semantic alignment, aesthetic score, and training stability, validating the effectiveness and generalizability of our proposed approach in modeling multimodal structural alignment and reinforcement learning-based fine-tuning.

Keywords: Large Language Models , Reinforcement Learning , Generative Models , Multimodal Alignment , Diffusion Models

1 引言

近年来，基于扩散过程的生成模型已迅速发展为图像、音频、视频等多模态领域的主流建模范式 (Croitoru et al., 2023; Yang et al., 2023)。扩散模型通过构造逐步加噪的正向过程与逐步去噪的逆向过程，将数据生成任务建模为从标准高斯噪声中恢复真实样本的多步预测问题 (Chen et al., 2023a; Li et al., 2023)。该类模型不仅在训练上具备高度稳定性，还能生成视觉质量极高的样本，已逐渐成为替代对抗生成网络的新一代生成框架 (Gandikota et al., 2023)。

在此基础上，条件生成模型，特别是文本引导的扩散模型，将自然语言描述作为生成条件嵌入到扩散网络中，实现了从指令到图像的端到端语义映射能力 (Zhang et al., 2023; Chen et al., 2025a)。通过结合强大的语言编码器与图像生成结构 (Li et al., 2025)，生成模型已广泛应用于图像创作、智能设计、艺术生成等任务中 (Zhu et al., 2023; Bie et al., 2024)。

然而，现有预训练的文本引导扩散模型主要依赖最小均方误差或对数似然近似进行训练，其目标仅是拟合真实图像分布，而非对齐用户偏好或任务需求 (Wallace et al., 2024)。在实际应用中，用户往往希望模型生成图像能具备更高的主观美感、更强的语义对齐度，或满足更明确的任务控制需求 (Shekhar and Zhang, 2025)。例如，在艺术生成中，用户关注风格一致性；在广告设计中，则可能关注审美得分与主题吻合度 (Liu et al., 2024)。

为突破上述限制，近年来研究者开始尝试引入强化学习机制对预训练扩散模型进行微调，以更直接地优化目标指标。该类方法借鉴了基于人类反馈的强化学习在大规模语言模型中的成功实践，典型方法如策略优化扩散模型（DPOK） (Fan et al., 2023) 与去噪扩散策略优化（DDPO） (Black et al., 2023)提出将扩散采样过程建模为一个具有有限步数的马尔可夫决策过程。在此框架下，生成模型被视为一个策略函数，每一步的去噪操作被看作一个动作，最终图像的偏好打分作为奖励信号，利用策略梯度方法实现策略参数的优化更新 (Yang et al., 2024)。

这类方法打破了传统监督微调必须依赖有标签数据的限制，使得模型能够直接对齐如美学评分器、语义相似度判别器等反馈信号所表达的任务目标 (Shekhar et al., 2024)。此外，为防止微调过程产生策略漂移或样本分布崩塌等问题，相关工作还引入了与原始模型之间的散度正则项作为稳定机制，从而在优化主观偏好的同时保留生成的稳定性与真实性 (Tang, 2024)。

尽管上述工作取得了显著进展，但现有方法中所使用的奖励函数多来源于黑盒式的打分模型，缺乏可解释性，且大多数奖励信号仅在样本级别进行评分，难以对生成图像与文本之间的

跨模态语义结构建立更深层的约束 (Jiang et al., 2023)。这种弱语义对齐的监督信号容易导致模型仅提升表面得分，而忽略更深层次的语义共现关系与模态一致性 (Niu et al., 2024)。

为解决上述问题，本文引入了一种结构化的几何正则项。该方法通过计算图像与文本在共享嵌入空间中所张成的平行多面体体积，来衡量其多模态嵌入之间的结构对齐程度，具有明确的几何意义、良好的可微性与可扩展性，能够直接建模多模态之间的空间结构关系。

在本文中，我们将该结构对齐机制作为正则项嵌入至强化学习微调的损失函数中，与奖励信号和散度约束一同组成统一优化目标。基于此思想，我们提出了一种新的文本引导扩散模型微调方法，融合了强化学习优化与多模态几何对齐的双重机制，旨在提升生成图像的语义一致性、美学质量以及整体表达能力。

本文的主要贡献如下：

- 我们提出了一种基于强化学习优化的扩散微调方法**GARD**（*Geometry-Aligned Reinforced Diffusion*）以提升生成图像在语义精度与主观质量上的表现。整体框架如图 1 所示。
- 我们在KL正则项的基础上引入多模态几何对齐正则，统一建模奖励优化与结构对齐目标，增强图文模态在嵌入空间中的结构一致性。
- 我们在多个公开数据集上进行充分实验，验证本文方法在奖励提升、语义对齐度与收敛稳定性等方面均显著优于现有主流方法。

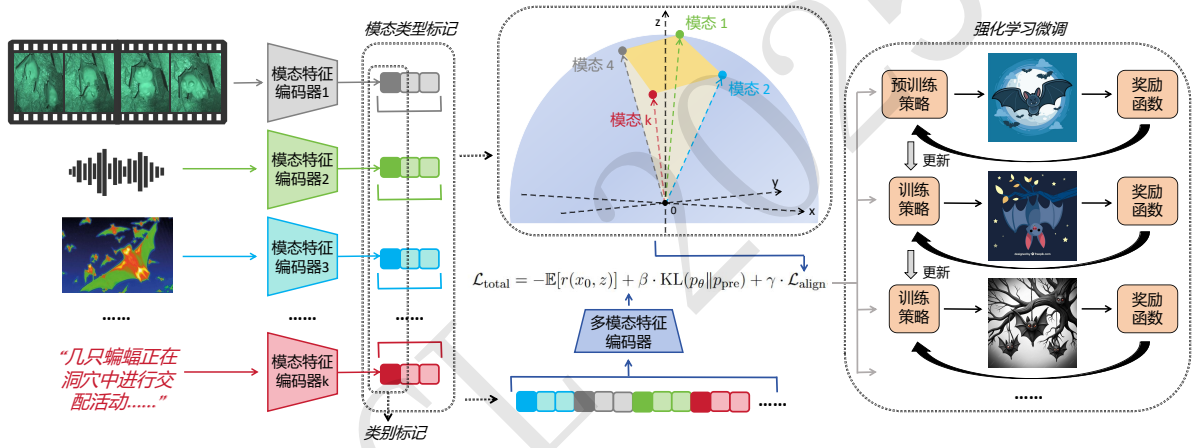

Figure 1: 模型整体框架的示意图。每个输入模态的类别标记参与构建空间多面体，其体积作为模态之间语义一致性的表征，与多模态编码器在损失函数中结合以增强预测效果。而在在线强化学习微调中，模型从预训练策略出发，使用来自先前训练模型的新样本对参数进行更新。

2 相关工作

文本引导的扩散模型 扩散模型已成为近年来生成建模领域的主流方法之一，其通过构造前向随机扰动过程与反向去噪生成过程，实现了从高斯噪声中逐步采样恢复出高质量样本的能力 (Higham et al., 2023; Guan et al., 2024)。最早的DDPM提出了基于高斯过程的逐步重建框架，并以其稳定训练与生成效果成为强大的替代GAN的生成模型(Ho et al., 2020)。在条件生成任务中，文本引导的扩散模型进一步取得突破性进展 (Liu et al., 2025)。许多等方法通过结合大型语言模型与视觉特征建模，实现了语义丰富、图文高度一致的图像生成 (Zhao et al., 2023; Li et al., 2024a)。尽管这些预训练扩散模型在大规模数据上表现出良好的生成能力，但其训练目标通常为MSE损失或对数似然近似，并不能直接对齐下游任务需求或人类偏好 (Lin and Yang, 2023)。此外，在特定任务（如医疗图像合成、艺术风格控制等）中，用户更希望模型生成结果能满足明确的主观偏好和语义意图，这也促使研究者引入强化学习机制对预训练扩散模型进行微调优化 (Ma et al., 2024b; Chen et al., 2025b)。

基于强化学习的人类偏好建模 强化学习近年来在大语言模型中的应用，尤其是“基于人类反馈的强化学习”，已成为提升生成质量与安全性的关键路径 (Wang et al., 2023)。它的核心思想是引入人类偏好（如对多个生成结果的排序），通过训练奖励模型并在此基础上使用强化学习来优化策略生成行为，使生成模型更符合用户期待 (Wallace et al., 2024)。代表性方法如PPO-based RLHF，使用最大策略优化对语言模型输出概率进行更新 (Christiano et al., 2017)；DPO直接从偏好数据中构建分类目标，替代奖励模型 (Rafailov et al., 2023)；RLAIF采用大模型作为教师信号代替人工偏好，提升训练效率 (Wang and Klabjan, 2024; Yu et al., 2024)。这些方法普遍将生成模型视作一个策略函数，输出可被奖励信号引导的序列，从而将生成任务建模为一个序列决策问题 (Yang et al., 2024)。这类思想为将强化学习引入图像生成与扩散模型提供了理论与实践基础。

强化学习微调扩散模型 将扩散模型建模为强化学习问题，是近年来生成模型可控性研究中的一项重要进展 (Rafailov et al., 2023)。该类方法的基本思路是：将扩散过程视为一个多步马尔可夫决策过程，将每一步去噪操作视为智能体的一次动作，从而以策略优化的方式直接优化生成行为 (Uehara et al., 2024)。其中，DPOK首次明确提出将扩散模型的去噪过程视为RL策略，并引入KL正则项缓解奖励导致的策略退化问题 (Fan et al., 2023)。DDPO则进一步将该思路系统化，提出两种策略梯度估计方式，支持多prompt并行训练，并在奖励函数多样性方面做出扩展（如美学评分、压缩率、语义相似性等） (Black et al., 2023)。除上述方法外，DRaFT等研究也尝试将行动者-评估者、离线强化学习等机制引入扩散模型训练，通过不同形式的奖励结构提升图像表达质量 (Hansen-Estruch et al., 2023; Fang et al., 2024)。

多模态对齐与表示学习 多模态对齐是指在共享语义空间中建立来自不同模态（如图像、文本、音频等）的表示之间的一致性关系，是多模态理解与生成任务的核心问题之一 (Cao et al., 2024; Wang et al., 2024)。近年来对比学习成为主流技术路径，其通过最大化正对之间的相似度、最小化负对之间的相似度，实现跨模态语义映射的高效学习 (Hager et al., 2023)。典型方法如CLIP (Radford et al., 2021)、ALIGN (Jia et al., 2021)、FILIP (Yao et al., 2021)等，均在大规模图文对中采用对比损失进行双塔式编码器训练，取得了良好性能。然而，这类方法通常仅捕捉了模态间的“点对点对齐”信息，难以建模模态内部与模态间的全局几何结构 (Cicchetti et al., 2024)。为此，后续研究提出了一系列结构增强机制，如：利用共享注意力机制显式建模模态间对齐路径 (Ma et al., 2024a; Li et al., 2024b)，采用中心损失或分布对齐方式约束嵌入聚类结构等 (Wang et al., 2025)。

3 问题设定

3.1 去噪扩散概率模型

去噪扩散概率模型（Denoising Diffusion Probabilistic Models, DDPM）是一类通过逐步去噪操作实现数据生成的概率模型。其基本思想是，将原始图像数据分布逐步“扩散”成一个多维高斯噪声分布，并学习其逆过程，从而从噪声中逐步恢复出高质量图像。本文以条件生成场景为背景，考虑给定文本提示z 的条件扩散模型$p_\theta(x_0 \mid z)$。

- 前向扩散过程

  设原始数据分布为$q_0(x_0)$，扩散过程由如下马尔可夫链定义：

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \quad q(x_t \mid x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t \cdot \mathbf{I}), \tag{1}$$

  其中$\{\beta_t\}_{t=1}^{T}$为噪声方差调度序列。基于该构造可推得任意时间步x_t与原始样本x_0的关系：

$$x_t = \sqrt{\bar{\alpha}_t} \cdot x_0 + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{2}$$

  其中$\alpha_t = 1 - \beta_t$，$\bar{\alpha}_t = \prod s = 1^t \alpha_s$。

- 反向生成过程

  反向过程的目标是从标准正态噪声$x_T \sim \mathcal{N}(0, \mathbf{I})$恢复出原始样本$x_0$。设条件文本为$z$，条件扩散模型学习如下形式的反向转移概率：

$$p_\theta(x_{t-1} \mid x_t, z) = \mathcal{N}(\mu_\theta(x_t, t, z), \sigma_t^2 \cdot \mathbf{I}), \tag{3}$$

第二十四届中国计算语言学大会论文集，第285页-第297页，济南，中国，2025年8月11日至14日。

其中μ_θ是参数化的均值函数（由深度神经网络建模），σ_t^2通常为预设常数。

- 训练目标

为了简化优化，通常采用预测噪声的方式训练模型，优化的目标函数为：

$$\mathcal{L}_{\text{DDPM}}(\theta) = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, z)\|^2\right], \tag{4}$$

其中$\epsilon \sim \mathcal{N}(0, \mathbf{I})$，$\epsilon_\theta$是模型对噪声的预测。

3.2 去噪过程的强化学习建模

尽管DDPM可生成高质量图像，但在实际任务中，其生成质量往往难以精确满足用户偏好或下游评价指标（如图文对齐度、美学评分等）。因此，近年来研究者尝试将扩散采样过程建模为一个马尔可夫决策过程（MDP），并使用强化学习方法进行微调，使生成模型能在奖励信号指导下优化行为策略。

在本文中，我们将反向采样过程视为一个T步的有限时域MDP，其各组成部分定义如下：

- 状态$s_t = (z, x_t)$：表示当前的文本提示z 以及第t 步的图像状态x_t；
- 动作$a_t = x_{t-1}$：对应去噪模型在该步输出的图像；
- 初始状态分布：$x_T \sim \mathcal{N}(0, \mathbf{I})$，即从高斯噪声开始采样；
- 策略函数：$\pi_\theta(a_t \mid s_t) = p_\theta(x_{t-1} \mid x_t, z)$，即扩散模型本身；
- 状态转移函数：$P(s_{t+1} \mid s_t, a_t) = \delta_{x_{t+1}=a_t}$，为确定性转移；
- 奖励函数$r(s_t, a_t)$：仅在$t = 0$ 时给予一次奖励$r(x_0, z)$，其余时间步为0。

强化学习目标为最大化终点图像的期望奖励，具体目标函数表示为：

$$\max_\theta \ \mathbb{E}_{z\sim p(z)}\ \mathbb{E}_{x_{0:T}\sim p_\theta(x_{0:T}|z)}\left[r(x_0, z)\right]. \tag{5}$$

为了优化上述目标，可使用策略梯度方法。考虑模型生成轨迹$x_{0:T}$的联合概率：

$$p_\theta(x_{0:T} \mid z) = p(x_T) \cdot \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, z). \tag{6}$$

则策略梯度可表示为：

$$\nabla_\theta \mathbb{E}_{x_{0:T}}\left[r(x_0, z)\right] = \mathbb{E}_{x_{0:T}}\left[r(x_0, z) \cdot \sum_{t=1}^{T} \nabla_\theta \log p_\theta(x_{t-1} \mid x_t, z)\right]. \tag{7}$$

该表达式即为经典的REINFORCE算法（无基线项），可用于直接更新扩散模型的参数。

而为了防止模型过拟合奖励函数、生成偏离原始分布的图像（如非自然图像、颜色失真等），常引入KL散度作为正则项。具体而言，引入一个与预训练模型$p_{\text{pre}}(x_0 \mid z)$ 的KL距离约束，构成正则化目标：

$$\mathcal{L}_{\text{total}} = -\mathbb{E}_{x_0\sim p_\theta}[r(x_0, z)] + \beta \cdot \text{KL}(p_\theta(x_0 \mid z) \| p_{\text{pre}}(x_0 \mid z)). \tag{8}$$

由于边缘分布$p_\theta(x_0 \mid z)$ 不可显式计算，我们进一步推导出该KL距离的上界可写为每一步条件分布间KL的累加：

$$\text{KL}(p_\theta(x_0 \mid z) \| p_{\text{pre}}(x_0 \mid z)) \le \sum_{t=1}^{T} \mathbb{E}_{x_t}\left[\text{KL}(p_\theta(x_{t-1} \mid x_t, z) \| p_{\text{pre}}(x_{t-1} \mid x_t, z))\right]. \tag{9}$$

该上界便于在强化学习微调中进行高效估计与优化，且可与奖励一并构成最终训练目标。

第二十四届中国计算语言学大会论文集，第285页-第297页，济南，中国，2025年8月11日至14日。

4 研究方法

4.1 扩散模型微调框架

在我们的整体框架下，扩散模型的反向采样过程（即从高斯噪声x_T恢复至最终图像x_0）被建模为一个T步的马尔可夫决策过程。由此，整个扩散生成过程$x_T, x_{T-1}, \ldots, x_0$被看作策略$\pi_\theta$控制下的一个轨迹$\tau$，其最终输出图像$x_0$将由外部奖励模型进行评估（如ImageReward、VLM-based reward等），并以此反馈更新策略。

表达式 7 揭示了：在强化学习视角下，对扩散模型的优化可通过奖励加权的对数概率梯度累加实现。每轮迭代中，模型根据当前策略采样生成一批图像轨迹，评估每个最终样本x_0的奖励后，反向传播其对当前策略的梯度，逐步提升高质量图像的生成概率。

在具体实现中，直接使用当前策略采样并计算上述期望自然是最自然的思路，但为了提高样本利用率，也可重用上一轮采样并引入重要性权重：

$$\nabla_\theta \mathbb{E}\left[r\left(x_0, z\right)\right] \approx \mathbb{E}_{x_{0:T}\sim p_{\theta_{\text{old}}}}\left[\omega\left(x_{0:T}\right) \cdot r\left(x_0, z\right) \cdot \sum_t \nabla_\theta \log p_\theta\left(x_{t-1} \mid x_t, z\right)\right], \tag{10}$$

其中$\omega(x_{0:T})$为采样轨迹的比值权重，通常通过引入截断或置信域来控制其变化范围。

4.2 多模态嵌入对齐

尽管通过人类反馈定义的奖励函数已可对生成图像的质量和偏好性进行一定程度的优化，但其本质仍是基于黑盒打分器的弱监督信号。为了进一步提升生成图像与输入文本在语义空间中的对齐程度，增强跨模态的一致性结构表达，本文引入一种结构化的多模态嵌入正则项，旨在通过几何方式直接刻画图像与文本在嵌入空间中的对齐关系。核心思想是利用高维空间中多模态嵌入向量所张成平行多面体的体积来衡量对齐程度，体积越小则代表对齐越好。

设图像生成模型在生成最终图像x_0后，我们使用图像嵌入函数$\phi_I(\cdot)$和文本嵌入函数$\phi_T(\cdot)$，将图像与文本分别投射到n维共享嵌入空间中：

$$m_1 = \phi_T(z), \quad m_2 = \phi_I(x_0), \tag{11}$$

其中z为输入的文本提示词，x_0为生成图像，$m_1, m_2 \in \mathbb{R}^n$为归一化后的嵌入向量（即$|m_1| = |m_2| = 1$）。若存在更多模态（如音频、视频帧、深度图等），亦可定义对应的嵌入$m_3, \ldots, m_k$。

我们记所有模态的嵌入向量为$v_1, v_2, \ldots, v_k$，其中$v_i \in \mathbb{R}^n$，通过归一化处理，所有向量端点落在单位球面上。高维空间中由k个模态嵌入向量张成的k维平行多面体的体积可以表示为：

$$\text{Vol}(v_1, \ldots, v_k) = \sqrt{\det G(v_1, \ldots, v_k)}, \tag{12}$$

其中格拉姆矩阵$G \in \mathbb{R}^{k\times k}$定义为$G(v_1, \ldots, v_k) = A^\top A = [\langle v_i, v_j\rangle]_{i,j=1}^k$。其中$A = [v_1, v_2, \ldots, v_k] \in \mathbb{R}^{n\times k}$，即将各模态嵌入向量按列拼接组成的矩阵。由于体积为嵌入之间夹角与长度的函数，若多个模态语义高度一致，则其嵌入向量方向接近，体积趋于0；若模态差异较大，则体积增大。因此，$\text{Vol}(\cdot)$ 可以作为多模态对齐程度的自然度量，如图 2 所示。

我们将上述几何体积作为一个结构性正则项添加至总训练目标中。记：

$$\mathcal{L}_{\text{align}} = \sqrt{\det G(m_1, m_2, \ldots, m_k)} \tag{13}$$

为多模态对齐损失项，其中k 表示参与对齐的模态个数。在我们以文本与生成图像为主的情形中，$k = 2$，该体积可进一步简化为$\mathcal{L}_{\text{align}} = \sqrt{1 - \langle m_1, m_2\rangle^2}$。该表达式实质上是$\sin(\theta)$，其中$\theta$为两个嵌入向量之间的夹角。

我们将该体积视作惩罚项加入强化学习微调目标函数中，构建如下联合损失函数：

$$\mathcal{L}_{\text{total}} = -\mathbb{E}[r(x_0, z)] + \beta \cdot \text{KL}(p_\theta \| p_{\text{pre}}) + \gamma \cdot \mathcal{L}_{\text{align}}, \tag{14}$$

其中γ 为对齐正则项的权重超参数，用于平衡语义一致性与生成质量之间的关系。

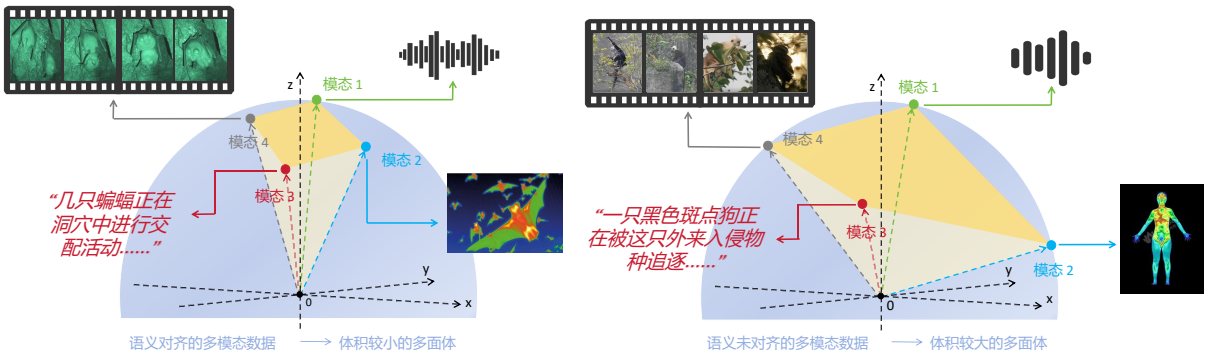

Figure 2: 基于嵌入向量构成的空间多面体体积的多模态嵌入的可视化说明。左图中，来自已经进行了较为恰当的语义对齐的多模态数据的嵌入向量构成了一个体积较小的空间多面体；而在右图中，各模态之间缺乏足够良好的语义对齐，所构成的空间多面体体积较大。

相比于传统的成对余弦相似度方法，该方法提供了一种天然可扩展到$k \geq 2$ 个模态的对齐度量方式。其几何意义明确、可导且在实际训练中开销极小（仅需计算$k \times k$ 的矩阵行列式），同时对不同模态之间的互信息具有更高的敏感性和全局表达能力。

我们的方法也与CLIP-guided diffusion 和contrastive learning-based alignment 存在本质差异。后者通常通过逐对模态进行最大/最小余弦相似度约束，其限制在于：（1）仅构建点对点对齐，不具备建模全局模态共线性的能力；（2）无法有效捕捉多个模态嵌入共同张成的结构性信息；（3）当扩展到$k \geq 3$ 模态时需引入额外权重聚合策略或独立损失项，影响模型稳定性。相比之下，GARD 的体积正则天然具备处理多模态间协同对齐的表达能力，更适合大规模语义结构建模场景。

4.3 整体训练目标

在前两节中，我们分别建立了基于强化学习的扩散模型微调框架以及结构化的多模态嵌入对齐正则项。本节将对这些模块进行整合，提出完整的训练目标函数，并详细说明训练流程。

综合奖励函数最大化目标、KL 正则项与多模态嵌入对齐正则项，本文提出以下联合损失函数作为最终优化目标：

$$\mathcal{L}_{\text{total}} = -\mathbb{E}_{x_{0:T}\sim p_\theta}\left[r(x_0, z)\right] + \beta \cdot \sum_{t=1}^{T} \mathbb{E}_{x_t}\left[\text{KL}(p_\theta(x_{t-1} \mid x_t, z) \| p_{\text{pre}}(x_{t-1} \mid x_t, z))\right] + \gamma \cdot \mathcal{L}_{\text{align}}(x_0, z), \tag{15}$$

其中θ为当前扩散模型的可学习参数；$r(x_0, z)$为外部奖励模型对最终生成图像与文本的评分；$\text{KL}(\cdot|\cdot)$为当前模型与预训练模型在每一步采样分布间的KL 散度；$\mathcal{L}_{\text{align}}(x_0, z)$为多模态嵌入空间对齐正则项；$\beta$ 与γ 为可调超参数，控制正则项的权重。

该联合目标函数较好地体现了我们方法的三个优化方向：最大化人类偏好：通过奖励函数提升图像质量和提示词相关性；约束分布偏移：通过KL 项限制模型行为不偏离原始训练分布；强化语义对齐：通过正则项约束图像与文本在嵌入空间中的一致性。具体的训练过程采用在线强化学习策略优化框架，每轮迭代中，模型根据当前策略生成样本，计算各项损失并更新模型参数。训练流程如算法 1 所述。

需要强调的是，在重要性采样部分，为提升采样效率，奖励函数和KL 项的梯度可基于旧策略采样，通过引入重要性权重稳定优化；对于正则系数动态调整，为避免训练初期奖励函数崩溃或嵌入偏移，β, γ 可随迭代轮次增大；在嵌入器冻结或微调环节，ϕ_T, ϕ_I 可以选择来自预训练CLIP / BLIP 模型，也可在后期联合微调；而扩散模型参数可采用低秩适配（LoRA）或选择性微调，以降低训练成本。

Algorithm 1 基于多模态对齐正则的强化学习微调扩散模型

Input: 预训练模型p_{pre}, 奖励函数$r(\cdot,\cdot)$, 多模态编码器ϕ_T, ϕ_I, 提示词集合$\{z_i\}_{i=1}^{N}$, 超参数β, γ, 学习率η, 总迭代次数K, 批次大小B
Output: 微调后的扩散模型$p_\theta(x_{t-1} \mid x_t, z)$

for $k = 1$ to K **do**
  从提示词分布中采样一批$z_1, \ldots, z_B$
  for $i = 1$ to B **do**
    使用当前模型p_θ 对z_i 生成图像轨迹$x_T \to \cdots \to x_0^{(i)}$
    计算奖励：$r_i \leftarrow r(x_0^{(i)}, z_i)$
    计算KL 正则：$\mathrm{KL}_i \leftarrow \sum_{t=1}^{T} \mathrm{KL}(p_\theta(x_{t-1} \mid x_t, z_i) \,\|\, p_{\text{pre}}(x_{t-1} \mid x_t, z_i))$
    计算多模态对齐项：$\mathcal{L}_{\text{align}}^{(i)} \leftarrow \sqrt{1 - \left\langle \phi_T(z_i), \phi_I(x_0^{(i)}) \right\rangle^2}$
    组合损失函数：$\mathcal{L}^{(i)} \leftarrow -r_i + \beta \cdot \mathrm{KL}_i + \gamma \cdot \mathcal{L}_{\text{align}}^{(i)}$
  end for
  梯度更新：$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \left(\frac{1}{B} \sum_{i=1}^{B} \mathcal{L}^{(i)} \right)$
end for

5 实验

5.1 数据集

为了评估我们的模型框架，我们选取了如下几个经典的公开数据集进行实验:

- **PartiPrompts-50**: 一个包含超过1600条提示的人工构造英文提示词集，包含具有可控性和可验证性的简单描述 [0]。该数据集常用于测试模型在少量训练条件下的定向控制能力与训练稳定性，适合微调实验的初步对比与消融分析。
- **PartiPrompts-398**: 一个覆盖面更广的中等规模提示词数据集，包含上百条不同风格的英文文本提示，涵盖人物、物体、场景等多种类型 [1]。该数据集适合用于训练通用文本引导扩散模型，并检验模型在多样条件下的泛化与鲁棒性。
- **COCO Captions**: 计算机视觉领域最广泛使用的图像标注数据集之一，其验证集包含约330,000 张图像及对应的文本描述 [2]。本文采用val2014 子集中的图文对，用于评估生成图像与输入提示词在语义嵌入空间中的对齐程度，如CLIPScore 或BERTScore。
- **LAION-Aesthetics**: LAION-Aesthetics v2 是一个包含图像与美学评分的公开数据集 [3]。其评分由基于CLIP的模型预测，广泛用于图像生成任务中的审美评分训练或评估。本文采用该数据集中的评分模型对生成图像的美观度进行自动评价。

5.2 模型效果比较

我们首先评估所有方法在可压缩性、不可压缩性和美学质量任务上的表现，因为这些任务能够将强化学习方法的有效性与奖励函数相关的考虑因素区分开来。我们选择的对比模型包括DDPO (Black et al., 2023)、DPOK (Fan et al., 2023)、MMD (Miao et al., 2024)和CaPO (Lee et al., 2025)，它们都是近两年间生成模型领域同方向中的最优性能模型。

因为奖励评估在许多实际应用中已成为限制因素，我们绘制了奖励函数查询次数与获得的奖励之间的关系。我们在图 3 中提供了所有方法的定量比较数据。实验结果表明，我们提出的**GARD**模型在所有任务上都明显优于同类方法，表明了将去噪过程建模为多步骤马尔科夫决策过程并直接估计策略梯度的优越性。在权重方案的性能相当时，由于其简单性和较低的资源需求，稀疏权重方案在这类任务中更受欢迎。

[0] https://sites.research.google/parti/
[1] https://github.com/jannerm/ddpo
[2] https://github.com/tylin/coco-caption
[3] https://laion.ai/blog/laion-aesthetics/

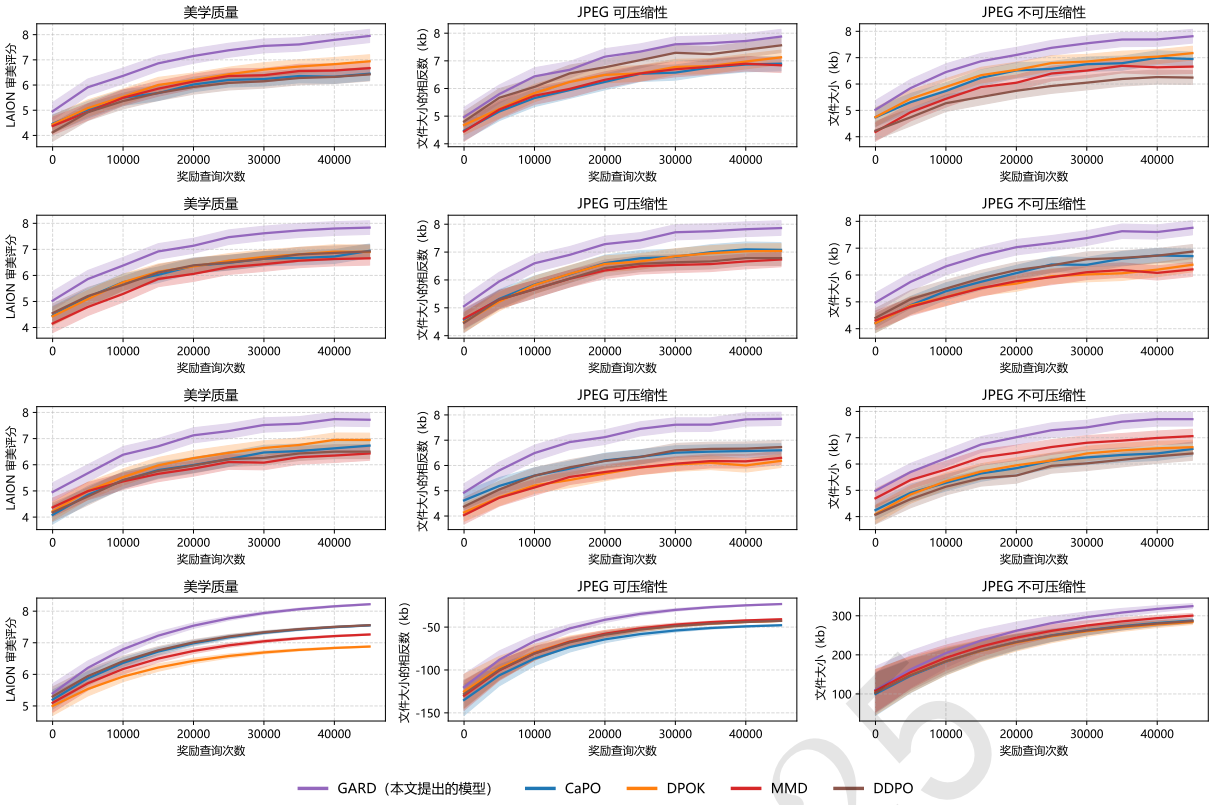

Figure 3: 强化学习微调有效性的展示。我们展示了不同的生成模型在各个奖励函数上的相对有效性。折线上下的有色阴影代表置信区间。从上到下四列分别代表在四个数据集上的结果。

为了更清晰地展示不同方法在主要评价指标上的整体表现，我们在表 1 中列出了GARD与DDPO、DPOK 在四个公开数据集上的对比结果。可以观察到，GARD 在语义一致性、美学得分和训练稳定性三项指标上均显著优于其他方法，验证了本文方法在优化图文对齐与生成质量方面的有效性与泛化能力。

5.3 消融实验

为了评估我们在 4.2 小节中提出的核心正则项$\mathcal{L}_{\text{align}}$在模型性能中的具体贡献，我们设计了一系列消融实验，将其替换为多种多模态表示学习中常见的对齐目标。我们选择的方案有：

- 余弦相似度：许多经典多模态模型使用的对齐度量$\cos(\theta) = \frac{\langle m_i, m_j \rangle}{\|m_i\| \|m_j\|}$ (Radford et al., 2021)。该方案只适用于双模态的情形，若涉及更多模态则需调整为两两余弦相似度聚合。
- 带锚模态的对比损失:在对比学习框架中引入锚定模态，并通过该模态与其他模态之间的配对关系进行对齐建模 (Jeong et al., 2024)。
- 特征融合：先用简单加权或多层感知机融合不同模态，再与锚模态计算损失 (Chen et al., 2023b)。该方案的可解释性较差，并可能破坏模态结构。
- 基于矩阵的相关性度量：通过最大化两个模态投影后的相关系数来实现对齐，衡量两个模态嵌入之间的线性相关性，如典型相关分析 (Andrew et al., 2013)。

在以上方案外，我们还考虑了有或无KL正则项的对比设置。最终在不同数据集上的对比展示结果如图 4 所示，直观展现了我们提出的模型**GARD**在微调场景下对多模态对齐的显著改进。我们的模型在各个设置下均高于其他的对比方法，并且有KL正则项的输出比无KL正则项的输出也有明显提升。这一结果表明了以统一的方式建模多模态潜在表征空间的重要性，以及多种模态的贡献对于检索正确数据至关重要。我们的模型被预训练以在语义上对齐所有模态，从而形成一个信息量更高、代表性更强且更有生成意义的潜在空间。

Table 1: 不同方法在各数据集上的语义一致性、审美得分与训练稳定性对比

方法	数据集	语义一致性↑	审美得分↑	稳定性↑
DDPO	Parti-50	0.721	6.24	0.88
DPOK	Parti-50	0.743	6.41	0.91
GARD	Parti-50	**0.782**	**6.67**	**0.94**
DDPO	Parti-398	0.715	6.17	0.86
DPOK	Parti-398	0.738	6.34	0.90
GARD	Parti-398	**0.775**	**6.59**	**0.93**
DDPO	COCO Captions	0.702	5.89	0.84
DPOK	COCO Captions	0.725	6.02	0.87
GARD	COCO Captions	**0.768**	**6.31**	**0.91**
DDPO	LAION-Aesthetics	0.686	6.48	0.82
DPOK	LAION-Aesthetics	0.704	6.73	0.85
GARD	LAION-Aesthetics	**0.741**	**6.96**	**0.89**

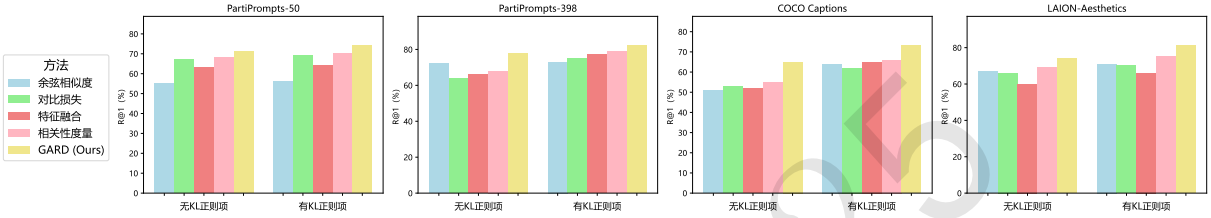

Figure 4: 不同正则化方案下的多模态微调在Recall@1（R@1）上的对比结果展示（%）。

5.4 训练效率与部署可行性

为了评估GARD 方法的实际计算开销，我们对比了DDPO、DPOK 与GARD 在Parti-50 数据集上的训练时间与显存占用。实验在同一NVIDIA A100 GPU（80GB）环境下进行，batch size 设为64，学习率为1e-5，优化器均为Adam。

结果表明，GARD 在每个epoch 上相较于DDPO 增加约3.1% 的训练时间（DDPO 平均epoch time 为1.42h，GARD 为1.46h），显存使用增加约1.8GB，主要来源于对齐正则的嵌入缓存与Gram 矩阵计算。但整体GPU 使用率保持在92% 以上，未出现明显的I/O 等待或显存溢出问题。此外，由于体积正则项不依赖于推理阶段采样，因此模型部署时与DDPO、DPOK 的推理速度完全一致，不会影响实际生成延迟，具备较好的工业落地潜力。

6 结论

本文提出了一种结合结构对齐正则项与强化学习优化机制的文本引导扩散模型微调方法**GARD**。该方法策略梯度框架基础上，引入几何对齐体积作为多模态结构一致性的度量，将其作为正则项融入整体损失函数中，构建出一套同时优化奖励目标与语义对齐结构的统一微调策略。通过将图像与文本在嵌入空间中的结构关系显式建模为平行多面体体积，本文的方法有效强化了跨模态语义一致性，并通过对奖励项与KL 项的联合优化保证模型收敛稳定与生成质量。在多个任务与数据集上的实验结果表明，**GARD** 不仅在Aesthetic Score、CLIPScore 等指标上显著优于现有方法，同时在嵌入对齐度与奖励提升效率方面也表现出更强的鲁棒性与泛化能力。未来工作可进一步探索该结构对齐机制在更多模态及下游任务（如图文检索、可控生成）中的扩展潜力，并结合更多结构感知奖励设计构建更具解释性与精度的生成优化系统。同时，本文方法计算开销小，易于在现有扩散模型生成服务框架中部署，具备较强的工程可落地性与应用推广价值。

参考文献

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR.

Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A Clifton, et al. 2024. Renaissance: A survey into ai text-to-image generation in the era of large model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. 2024. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15710–15719.

Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. 2023a. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19830–19843.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023b. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866.

Hang Chen, Qian Xiang, Jiaxin Hu, Meilin Ye, Chao Yu, Hao Cheng, and Lei Zhang. 2025a. Comprehensive exploration of diffusion models in image generation: a survey. *Artificial Intelligence Review*, 58(4):99.

Yuxin Chen, Devesh K Jha, Masayoshi Tomizuka, and Diego Romeres. 2025b. Fdpp: Fine-tune diffusion policy with human preference. *arXiv preprint arXiv:2501.08259*.

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.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. 2024. Gramian multimodal representation learning and alignment. *arXiv preprint arXiv:2412.11959*.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 2023. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885.

Linjiajie Fang, Ruoxue Liu, Jing Zhang, Wenjia Wang, and Bing-Yi Jing. 2024. Diffusion actor-critic: Formulating constrained policy iteration as diffusion noise regression for offline reinforcement learning. *arXiv preprint arXiv:2405.20555*.

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

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. 2024. Decompdiff: diffusion models with decomposed priors for structure-based drug design. *arXiv preprint arXiv:2403.07902*.

Paul Hager, Martin J Menten, and Daniel Rueckert. 2023. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23924–23935.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. 2023. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*.

Catherine F Higham, Desmond J Higham, and Peter Grindrod. 2023. Diffusion models for generative artificial intelligence: An introduction for applied mathematicians. *arXiv preprint arXiv:2312.14977*.

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

Minoh Jeong, Min Namgung, Zae Myung Kim, Dongyeop Kang, Yao-Yi Chiang, and Alfred Hero. 2024. Anchors aweigh! sail for optimal unified multi-modal representations. *arXiv preprint arXiv:2410.02086*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Zutao Jiang, Guian Fang, Jianhua Han, Guansong Lu, Hang Xu, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. 2023. Realigndiff: Boosting text-to-image diffusion model with coarse-to-fine semantic re-alignment. *arXiv preprint arXiv:2305.19599*.

Kyungmin Lee, Xiaohang Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. 2025. Calibrated multi-preference optimization for aligning diffusion models. *arXiv preprint arXiv:2502.02588*.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. 2023. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17535–17545.

Ruijun Li, Weihua Li, Yi Yang, Hanyu Wei, Jianhua Jiang, and Quan Bai. 2024a. Swinv2-imagen: Hierarchical vision transformer diffusion models for text-to-image generation. *Neural Computing and Applications*, 36(28):17245–17260.

Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, and Dongmei Jiang. 2024b. Alignmamba: Enhancing multimodal mamba with local and global cross-modal alignment. *arXiv preprint arXiv:2412.00833*.

Ziqiang Li, Jun Li, Lizhi Xiong, Zhangjie Fu, and Zechao Li. 2025. A comprehensive survey on visual concept mining in text-to-image diffusion models. *arXiv preprint arXiv:2503.13576*.

Shanchuan Lin and Xiao Yang. 2023. Diffusion model with perceptual loss. *arXiv preprint arXiv:2401.00110*.

Kendong Liu, Zhiyu Zhu, Chuanhao Li, Hui Liu, Huanqiang Zeng, and Junhui Hou. 2024. Prefpaint: Aligning image inpainting diffusion model with human preference. *Advances in Neural Information Processing Systems*, 37:30554–30589.

Mushui Liu, Yuhang Ma, Zhen Yang, Jun Dan, Yunlong Yu, Zeng Zhao, Zhipeng Hu, Bai Liu, and Changjie Fan. 2025. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5523–5531.

Feipeng Ma, Yizhou Zhou, Zheyu Zhang, Shilin Yan, Hebei Li, Zilong He, Siying Wu, Fengyun Rao, Yueyi Zhang, and Xiaoyan Sun. 2024a. Ee-mllm: A data-efficient and compute-efficient multimodal large language model. *arXiv preprint arXiv:2408.11795*.

Zhiyuan Ma, Yuzhu Zhang, Guoli Jia, Liangliang Zhao, Yichao Ma, Mingjie Ma, Gaofeng Liu, Kaiyan Zhang, Jianjun Li, and Bowen Zhou. 2024b. Efficient diffusion models: A comprehensive survey from principles to practices. *arXiv preprint arXiv:2410.11795*.

Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. 2024. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853.

Xuexiang Niu, Jinping Tang, Lei Wang, and Ge Zhu. 2024. Bridging the gap: Aligning text-to-image diffusion models with specific feedback. *arXiv preprint arXiv:2412.00122*.

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

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*, 36:53728–53741.

Shivanshu Shekhar and Tong Zhang. 2025. Rocm: Rlhf on consistency models. *arXiv preprint arXiv:2503.06171*.

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

Wenpin Tang. 2024. Fine-tuning of diffusion models via stochastic control: entropy regularization and beyond. *arXiv preprint arXiv:2403.06279*.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. 2024. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *arXiv preprint arXiv:2407.13734*.

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

Shuyang Wang and Diego Klabjan. 2024. An ensemble method of deep reinforcement learning for automated cryptocurrency trading. In *2024 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pages 461–463. IEEE.

Yuanhao Wang, Qinghua Liu, and Chi Jin. 2023. Is rlhf more difficult than standard rl? a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:76006–76032.

Fei Wang, Liang Ding, Jun Rao, Ye Liu, Li Shen, and Changxing Ding. 2024. Can linguistic knowledge improve multimodal alignment in vision-language pretraining? *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(12):1–22.

Xinpeng Wang, Rong Zhou, Han Xie, Xiaoying Tang, Lifang He, and Carl Yang. 2025. Clusmfl: A cluster-enhanced framework for modality-incomplete multimodal federated learning in brain imaging analysis. *arXiv preprint arXiv:2502.12180*.

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 Computing Surveys*, 56(4):1–39.

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*, pages 8941–8951.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. 2023. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739.

Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, and Cong Yao. 2023. Conditional text image generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14235–14245.