从博弈论视角解构去噪扩散概率模型的视觉概念生成机制

刘超一[1],耿浩棒[1],葛亚维[2],林 晗[2],侯 娜[3],赵二虎[1],黄礼泊[1],徐勇军[1]
(1. 中国科学院计算技术研究所,北京 100190;2. 军事科学院,北京 100091;3. 中国人民解放军32801部队,北京 100082)

摘 要: 去噪扩散概率模型(Denoising Diffusion Probabilistic Models,DDPMs)作为当前生成式AI领域的核心技术,在高质量图像合成任务中实现了革命性突破,但其内在工作机制长期被视为"黑箱",严重制约了其在医疗影像、自动驾驶等高可信度要求场景中的规模化应用. 现有研究多聚焦于对逆向去噪过程的宏观行为分析,缺乏对潜空间中不同语义区域间动态交互机制的细粒度解构,导致模型可解释性与精准操控能力之间存在显著鸿沟. 本研究从视觉概念生成解耦的新视角,探索了去噪扩散概率模型的可解释性. 该发现不仅从理论角度解释了局部性在DDPMs上的表现,还在下游应用中实现了细粒度的图像操控. 受博弈论启发,本文提出采用沙普利值来评估区域间的交互作用. 然而,单纯按传统定义计算沙普利值将面临时间复杂度上的可行性问题. 为此,本文进一步提出一个定理及配套采样策略,将时间复杂度降至$O(KC)$,其中K代表区域数,C为采样数. 定性定量实验表明,采用本方法进行真实图像处理时,对比现有方法本文提出的方法在局部操控方面性能提升30%~55%. 实际应用中,用户可针对性修改特定视觉概念而不会干扰其他区域. 通过博弈论与DDPM的深度融合,不仅在理论上首次阐明了局部性在扩散模型中的数学本质与实现路径,更在实践中构建了首个具备语义解耦能力的可解释DDPM框架.

关键词: 计算机科学;人工智能;大模型;可解释;去噪扩散概率模型
基金项目: 北京市自然科学基金(No.4244098)
中图分类号: TP18;TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2025)11-3910-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250716

Disentangling the Visual Concept Generation of Denoising Diffusion Probabilistic Model from a Game-Theoretic View

LIU Chao-yi[1], GENG Hao-bang[1], GE Ya-wei[2], LIN Han[2],
HOU Na[3], ZHAO Er-hu[1], HUANG Li-bo[1], XU Yong-jun[1]
(1. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*;
2. *Academy of Military Science, Beijing* 100091, *China*; 3. *The People's Liberation Army of China* 32801, *Beijing* 100082, *China*)

Abstract: Denoising diffusion probabilistic models (DDPMs), as a core technology in the current generative AI field, have achieved revolutionary breakthroughs in high-quality image synthesis tasks. However, their internal working mechanisms have long been regarded as a "black box", severely restricting their large-scale application in high-trust scenarios such as medical imaging and autonomous driving. Existing research mostly focuses on the macroscopic behavior analysis of the reverse denoising process, lacking fine-grained deconstruction of the dynamic interaction mechanisms among different semantic regions in the latent space, resulting in a significant gap between model interpretability and precise control ability. This study explores the interpretability of denoising diffusion probabilistic models from a new perspective of decoupled visual concept generation. The findings not only explain the manifestation of locality in DDPMs from a theoretical standpoint but also enable fine-grained image manipulation in downstream applications. Inspired by game theory, we propose to use Shapley values to evaluate the interactions between regions. However, calculating Shapley values according to the traditional definition would face feasibility issues in terms of time complexity. Therefore, we further propose a theorem and an accompanying sampling strategy to reduce the time complexity to $O(KC)$, where K represents the number of regions and C is the number of samples. Qualitative and quantitative experiments show that our method, when applied to real image process-

收稿日期:2025-08-15;录用日期:2025-11-13;责任编辑:覃怀银

ing, achieves a 30%~55% performance improvement in local manipulation compared with existing methods. In practical applications, users can modify specific visual concepts without interfering with other regions. Through the deep integration of game theory and DDPM, not only has the mathematical essence and implementation path of locality in diffusion models been theoretically clarified for the first time, but also the first interpretable DDPM framework with semantic decoupling capability has been constructed in practice.

Key words: computer science; artificial intelligence; large models; interpretability; DDPMs

Foundation Item(s): Beijing Municipal Natural Science Foundation (No.4244098)

1 引言

去噪扩散概率模型(Denoising Diffusion Probabilistic Models,DDPMs)近期在图像生成任务中取得显著突破,涵盖从零生成[1~5]到图像修复[6~9]等多个领域. 尽管DDPMs表现出色,但其在每个去噪步骤中的内部机制(即DDPMs的可解释性)尚未得到充分理解. 这不仅对理论研究至关重要,也对实际应用中有效且高效地操控其生成图像具有重要意义[10~14].

因此,关于扩散概率模型可解释性的研究开始受到关注. 其中一个方向尝试借助生成过程的提示引导来提升模型的可解释性. 类别引导方法[9,12]开发了能合成指定类别图像的模型. 后续研究[14~16]将类别提示扩展至自然语言,实现了基于输入文本的更灵活图像操控. 然而这些方法常面临提示与图像内多样视觉概念间对齐粗糙的问题,限制了细粒度图像操控的实现. 另一类研究[7,11,17]则更具实用主义倾向,它们通过精心设计推理过程,在保持非目标区域不受干扰的前提下实现感兴趣区域的精准修改. 尽管具备应用价值,这些方法对理解扩散概率模型内部机制的贡献仍然有限.

本文提出了一种全新的DDPMs可解释性研究视角,该方法在概念上更为简洁优雅. 提出的技术方案能够在DDPM训练过程中直接实现区域化视觉概念的解耦生成.

直观来看,扩散概率模型在合成过程中输入的高斯噪声与生成图像具有相同维度,由此可以预期输入高斯噪声的某些区域会自然对应生成图像的最终输出. 例如,在生成人脸图像的DDPMs中,当修改输入高斯噪声中对应面部鼻子的区域时,最终输出图像应仅呈现不同的鼻子,而其他无关区域保持不变. 然而,这种直观的可解释性特性并非原生存在于原始DDPMs中.

为实现区域解耦,我们将DDPMs的多步去噪过程视为整体,约束输出图像与输入高斯噪声间的关联关系. 具体而言,首先定义与目标解耦视觉概念相对应的区域;其次量化输入高斯噪声上不同区域对输出图像目标区域的贡献程度;最终通过添加额外损失函数项来最小化跨区域交互作用. 该损失函数确保生成图像上的目标区域完全且仅由输入高斯噪声对应区域内的信号控制,即实现局部性特性.

选择这样的区域是可解释DDPM目标的首要考量. 直观地说,若区域过大(例如接近整幅图像),实际上就难以实现有效的特征解耦. 另一方面,若区域过小(例如将单个像素视为一个区域),则优化将变得困难,因为这本质上抛弃了相邻像素间的局部相关性. 即便采用固定的规则区域(如矩形补丁),这一结论依然成立,因为此类补丁可能会在其边界处强行切断语义关联. 因此,本文根据最终生成图像的语义来选择区域[18,19]. 例如在人脸领域,可选取眼睛、鼻子、嘴巴等部位. 通过这种方式划分的区域,本质上代表了生成过程中被解耦的视觉概念.

另一个重大挑战在于如何量化输入高斯噪声中不同区域对输出图像目标区域的贡献度. 为此,受博弈论启发,本方法采用沙普利值[20]来评估输入噪声中各区域的贡献. 沙普利值源于多人合作博弈理论,作为一种满足优良性质的公平无偏贡献度计算方法[21],其评估结果比其他归因方法更具可信度[22~24]. 然而,若按传统定义直接计算输入噪声每个独立区域对输出图像各目标区域的沙普利值,其时间复杂度将难以承受. 针对此问题,研究人员提出一个定理,通过高效计算所有其他区域的集体贡献来显著加速运算. 此外,本方法采用抽样策略[25]进一步提升效率,在DDPM单次反向传播步骤中,将时间复杂度从$O\left(K(K-1)2^K\right)$降至$O(KC)$,其中K为区域数量,C为采样次数.

在定性分析方面,通过局部语义区域内交换输入高斯噪声并重复采样这些区域的噪声,直观展示了生成图像的修改过程. 定量评估则涉及修改生成图像中的视觉概念,并评估改变的局部化特性. 受Ruiz等人[26]和Yang等人[27]的技术启发,本文还将该方法拓展至真实图像的操作.

本文的贡献可概括如下:

(1)提出一种创新的博弈论方法,将传统去噪扩散概率模型(DDPMs)转化为可解释版本. 每个输入的高斯噪声片段对应输出图像上的特定视觉概念区域,从而增强可解释性.

(2)提出一种经理论验证的高效时序方法,显著降低学习可解释DDPM的计算成本.

(3)该方法支持用户修改生成图像中的精确语义视觉概念,对多种下游任务具有实用价值. 实验结果验证了其有效性与实用性.

2 相关工作

2.1 深度神经网络的解释性

总体而言,深度神经网络可解释性的实现路径可分为两大框架:一是在模型开发阶段强制实现自我解释性[28~30],二是在模型开发后提供事后解释[23,31~34]. 自我解释模型的共同目标是学习解耦表征,而事后可解释性方法旨在为模型预测提供解释和合理性证明.

本文聚焦于自解释模型的研究. 构建自解释模型的方法多种多样,具体取决于目标模型架构的可解释性需求. 对于判别式模型,根据先前研究所示[30,35],可通过改造使其成为特征提取器,在每个滤波器中嵌入语义概念. 在基于变分自编码器(Variational Auto-Encoder, VAE)的模型中,β-VAE[28]及其变体[29,36,37]通过在损失函数中添加额外正则化项来实现潜在变量解耦. 对于生成对抗网络(Generative Adversarial Network, GAN)模型,InfoGAN[30,38]采用辅助分类器来解缠潜在编码,从而捕捉生成图像中的语义特征变化. 然而考虑到去噪扩散概率模型固有的马尔可夫过程特性,直接应用上述自解释方法可能会引发预期之外的棘手障碍. 与现有方法不同,本文提出通过解耦不同视觉概念的生成过程来增强DDPM的可解释性,这是通过利用输入高斯噪声与输出图像维度匹配的固有特性来实现的.

2.2 扩散模型的解耦

先前的研究聚焦于从不同视角学习去噪扩散概率模型的解耦表征. 分类引导方法[5,12]通过显式分类器引导图像生成朝向特定类别,成功解耦了与类别相关的语义. 语言引导研究[13,14,39]将分类引导生成的概念扩展至单模态方法之外,提出了融合语言引导的新型图像编辑方法. 这些技术允许通过操控伴随的文本描述来修改生成图像,但其高度依赖预训练的CLIP(Contrastive Language Image Pre-training)模型,因其需要借助该模型理解与解释文本指令的能力.

此外,通过在模型架构中融入交叉注意力层,部分研究[2,9,16]强化了文本提示与图像之间的对齐关系,从而实现对编辑过程的更精准控制. 交叉注意力图使模型能根据文本输入聚焦图像相关区域,确保生成输出与提示中指定的修改要求保持一致. 提示到提示编辑技术[40]进一步探索了注意力图的内在机制,提出通过直接操控注意力图来实现对生成过程的更精确调控. 然而,这些对齐方法仍存在粒度粗糙的问题.

为突破这一局限,近期研究方法[26,41,42]致力于微调生成模型,将目标物体的视觉信息与特定标签关联,从而实现对文生图过程的精准控制. 该技术使模型能将关联标记视为常规语言字符,并借此添加控制信息. DiffEdit[43]则利用两个修改后文本提示间的差异,自动生成图像编辑任务所需的蒙版. 这种方法不仅能实现编辑区域与背景的融合,相较交互式蒙版生成技术更具多重优势.

目前,大多数关于解耦去噪扩散概率模型的研究都集中在文本引导的图像生成方法上. 这些方法旨在实现特定文本标记与目标视觉概念之间更精确的对齐,以获得更好的编辑效果. 然而,探索无需显式文本引导的条件生成是一个重要但常被忽视的研究方向. 为此,扩散自编码器[44]通过引入额外编码器来学习语义概念,旨在发现高层语义. 尽管该方法能成功解耦头发、微笑、年龄和姿势等特定语义概念,但在捕捉更全面、更细微的语义概念时可能存在局限. Locatello等人[45]已证明,在无监督设置下,若缺乏适当的归纳偏置,学习解耦特征是不可能实现的.

与先前方法不同,本文的方法引入了一种新颖的高斯噪声划分方式,这种方式融合了强大的先验知识,有助于语义概念的解耦分离.

3 算法设计

在本节中,我们的目标是将传统DDPM改造为可解释的端到端模型. 所谓可解释DDPM,是指在去噪过程中,输入高斯噪声的特定区域能独立生成输出图像对应区域,并保持严格的一维对应关系. 具体而言,为最大限度保持相邻像素间的局部相关性,方法中将输入噪声区域划分为不同语义视觉概念(如人脸的眼睛、鼻子和嘴巴),使其对应输出图像的特定语义区域. 通过这种方式,研究者期望特定语义区域的生成内容完全归因于输入高斯噪声的对应区域,而不受其他无关区域影响. 为此,本文基于沙普利值[20]提出了一种新颖的博弈论方法,用以量化这种区域归因效果.

3.1 基本知识

3.1.1 去噪扩散概率模型

DDPM[1]的总体目标是通过对输入的高斯噪声进行迭代去噪来生成逼真的图像. 具体而言,其前向过程是逐步向图像添加噪声,按以下方式实现.

$$q\left(\boldsymbol{x}_t\middle|\boldsymbol{x}_{t-1}\right)=\mathcal{N}\left(\boldsymbol{x}_t;\ \sqrt{\alpha_t}\,\boldsymbol{x}_{t-1},\left(1-\alpha_t\right)\boldsymbol{I}\right) \tag{1}$$

其中,$q\left(\boldsymbol{x}_t\middle|\boldsymbol{x}_{t-1}\right)$表示时间戳$t$处的高斯变换;$\boldsymbol{x}_t$表示时间步$t$的加噪图像;$\boldsymbol{x}_T$是当$t=T$时的最终各向同性高斯噪声;$\{\alpha_t\}(t=1,2,\cdots,T)$表示方差调度序列. 通过这种方式,其逆向过程按如下方式进行.

$$p_\theta\left(\boldsymbol{x}_{0:T}\right)=p\left(\boldsymbol{x}_T\right)\prod_{t=1}^{T}p_\theta\left(\boldsymbol{x}_{t-1}\middle|\boldsymbol{x}_t\right) \tag{2}$$

其中,θ表示DDPM的参数. 在实现过程中,通常通过噪声预测器$\epsilon_\theta(\boldsymbol{x}_t,t)$来学习$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$,该预测器旨在恢复$\boldsymbol{x}_t$最初添加的噪声. 通过以下迭代过程生成图像:

$$\boldsymbol{x}_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\boldsymbol{x}_t,t)\right)+\sigma_t z_t \tag{3}$$

其中,$z_t\sim\mathcal{N}(0,1)$且$\bar{\alpha}_t=\prod_{t=1}^{T}\alpha_i\cdot\sigma_t^2$. 根据DDPM的设定,$z_t$通常取值为$1-\alpha_t$.

3.1.2 沙普利值

近年来,沙普利值[20]被广泛用于深度神经网络的特征归因分析. 该指标可作为可视化工具,通过模型输出来突显输入图像不同区域对结果的贡献度. 与其他归因方法[23,33,34]相比,沙普利值基于合作博弈论,能为每个输入特征提供具有坚实理论基础的唯一归因值,从而确保其公平性与可信度. 具体而言,给定参与博弈v的n个输入玩家集合$N=\{1,2,\cdots,n\}$,这些玩家可获得分数$v(N)$. 博弈v将任何参与玩家映射为一个数值,当无玩家参与博弈时的分数记为$v(\varnothing)$. n个玩家获得的总收益可表示为$v(N)-v(\varnothing)$. 当玩家i加入任意潜在子集$S\subseteq N\backslash\{i\}$时,其沙普利值$\varphi(i|N)$计算公式如下:

$$\varphi_v(i|N)=\sum_{S\subseteq N\backslash\{i\}}\frac{|S|!(n-1-|S|)!}{n!}\left(v(S\cup\{i\})-v(S)\right) \tag{4}$$

值得注意的是,沙普利值是满足四项特性[21]的,用于计算贡献度的唯一无偏方法.

(1)虚拟玩家特性. 若对于任意子集$S\subseteq N\backslash\{i\}$满足$v(S\cup\{i\})-v(S)=0$,则参与者$i$被视为虚拟玩家. 其贡献值计算公式为$\varphi_v(i|N)=v(\{i\})-v(\varnothing)$. 该特性意味着参与者$i$对博弈没有实质贡献.

(2)线性特性. 假设存在三个博弈u、v、w,其中w由u和v组合而成且满足$w(S)=u(S)+v(S)$,那么参与者i在博弈w中的沙普利值可通过其在博弈u和v中的沙普利值相加获得,即$\varphi_w(i|N)=\varphi_u(i|N)+\varphi_v(i|N)$.

(3)对称特性. 若参与者i与j对博弈的贡献相同,则他们在博弈v中的沙普利值应当相等. 具体而言,若对任意子集$S\subseteq N\backslash\{i,j\}$满足$v(S\cup\{i\})-v(S\cup\{j\})=0$,则有$\varphi_v(i|N)=\varphi_v(j|N)$.

(4)效率特性. 所有参与者的沙普利值之和等于全体参与者获得的总收益,即$\sum_i\varphi_v(i|N)=v(N)-v(\varnothing)$.

3.2 问题描述

设$\{x_t\in R^D\}(t=0,1,\cdots,T)$表示DDPM逆向过程中的输出图像序列. 基于$\boldsymbol{x}_0$的语义分割结果,对$\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_T$进行相同的语义区域划分. 令$\boldsymbol{f}_{k_t}(\cdot)\in\{0,1\}^K$表示独热向量,该向量用于指示$\boldsymbol{x}_t$的第$k_t$个语义区域($k_t\in\mathcal{K}_T$,其中$\mathcal{K}_T=\{1,2,\cdots,K\}$,$K$表示所有语义区域的数量). 由此可得$\boldsymbol{f}_{k_0}(x_0)=\boldsymbol{f}_{k_1}(x_1)=\cdots=\boldsymbol{f}_{k_T}(x_T)\Leftrightarrow k_0=k_1=\cdots=k_T$. 从沙普利值的归因角度出发,研究人员期望可解释DDPM满足以下条件:

$$\boldsymbol{f}_{k_0}(\boldsymbol{x}_0)=I\left(\varphi_{v_{k_0}}(\cdot|\boldsymbol{x}_T,\mathcal{K}_T)\neq 0\right) \tag{5}$$

其中,$I(\cdot)$表示指示函数;$\varphi_{v_{k_0}}(\cdot|\boldsymbol{x}_T,\mathcal{K}_T)\in R^K$表示$\boldsymbol{x}_T$为在$\boldsymbol{x}_0$上生成第$k_0$个区域所对应的归因图,该图基于沙普利值计算得出. 例如,$\varphi_{v_{k_0}}(k_T|\boldsymbol{x}_T,\mathcal{K}_T)\in R$表示输入高斯噪声$\boldsymbol{x}_T$上第$k_T$个区域的标量沙普利值. 博弈$v_{k_0}$指代DDPM为在$\boldsymbol{x}_0$上生成第$k_0$个语义区域的整体逆向过程,可设定其返回$\boldsymbol{x}_0$上第$k_0$个语义区域的均值. 换言之,对于$\boldsymbol{x}_0$上的任何第$k_0$个语义区域(眼睛,嘴巴,鼻子),研究人员期望此类生成结果仅由$\boldsymbol{x}_T$上对应的第k_T个语义区域贡献,其中满足$k_T=k_0$的关系.

3.3 直接方法

DDPM遵循马尔可夫过程. 因此可以确保,在从时间t到$t-1$的每一步去噪过程中,$\boldsymbol{x}_{t-1}$上的第k_{t-1}个语义区域都独立受控于$\boldsymbol{x}_t$上的第k_t个语义区域($\forall k_{t-1}=k_t$). 由此可得:

$$\boldsymbol{f}_{k_{t-1}}(\boldsymbol{x}_{t-1})=I\left(\varphi_{v_{k_{t-1}}}(\cdot|\boldsymbol{x}_t,\mathcal{K}_t)\neq 0\right),\forall t\in\{1,2,\cdots,T\} \tag{6}$$

通过这种方式,在从时间t到$t-1$的每一步去噪过程中,针对在x_{t-1}上生成第k_{t-1}个语义区域的任务,算法旨在实现x_t归因图的以下特性:对应的第k_t个区域(即$k_t=k_{t-1}$)预期具有较高的绝对归因值,而其他区域(即$k_t\neq k_{t-1}$)的归因值应尽可能接近零. 此外,考虑到DDPM中的原始逆向过程,研究人员采用去噪扩散隐式模型(DDIM)[4]中的采样方法来消除z_t的影响. 在这种情况下,生成过程完全由输入的高斯噪声决定. 如此一来,博弈函数v_{k_t}可以等效地设置为返回预测噪声(即$\epsilon_\theta(x_t,t)$)上第k_t个语义区域的平均值. 通过这种方式,本文提出如下可解释性损失函数:

$$\mathcal{L}_I=E_t\left[\sum_{k_{t-1}}\sum_{k_t\neq k_{t-1}}\left|\varphi_{v_k t-1}(k_t|\boldsymbol{x}_t,\mathcal{K}_t)\right|\right] \tag{7}$$

这种损失函数可以直观地理解为:约束$\boldsymbol{x}_t$上的第k_t个区域仅影响$\boldsymbol{x}_{t-1}$上满足$k_{t-1}=k_t$的第k_{t-1}个区域,而对其他$k_{t-1}\neq k_t$的区域不产生任何影响.

3.4 时间效率优化方法

令人困扰的是,上述的损失函数的计算在DDPM的每个反向步骤中都需要$O\left(K(K-1)2^K\right)$的时间复杂度. 具体而言,该方法需要计算每个输入语义区域对每个输出语义区域的沙普利值,这种计算成本高昂且在

实际应用中不适用. 因此,基于提出的定理1,本文提出了一种等效的高效时序方法来简化计算.

定理1 若对于任意子集$S\subseteq N\backslash\{i\}$满足$v(S\cup i)=v(i)$且$v(S)=v(\varnothing)$,则认为$N\backslash\{i\}$中的所有参与者皆为虚拟参与者.

证明 通过对集合N基数的归纳法来证明定理1.

基础步骤:当$|N|=2$且$|S|=1$时,对于$j\in S$,有$v(\{j\}\cup\{i\})=v(\{i\})$且$v(\{\{j\}\})=v(\{\{j\}\cup\varnothing\})=v(\varnothing)$,这满足参与者$j$的虚拟性条件. 因此,$S$中唯一的参与者可视为虚拟参与者.

归纳步骤:假设当$|N|=n\geqslant 2$时,定理1的假设成立. 那么当$N=n+1$时,随机选取一个参与者$j\in N\backslash\{i\}$,并将$\{i,j\}$视为单个参与者. 这样,对于任意子集$S\subseteq N\backslash\{i,j\}$,将得到:

$$\begin{aligned}v(S\cup\{i,j\})&=v(\{S\cup\{j\}\}\cup\{i\})\\&=v(\{i\})//\{S\cup\{j\}\}\subseteq N\backslash\{i\},\\v(\{i,j\})&=v(\{j\}\cup\{i\})\\&=v(\{i\})//\{j\}\subseteq N\backslash\{i\}.\end{aligned}$$

那么就有

$$v(S\cup\{i,j\})=v(\{i,j\}).$$

此外,还有

$$\begin{aligned}v(S)&=v(\varnothing)//S\subseteq N\backslash\{i,j\},\\&\quad //\text{因此}S\subseteq N\backslash\{i\}.\end{aligned}$$

通过这种方式,可以基于归纳假设得出$N\backslash\{i,j\}$中的所有参与者都是虚拟参与者. 更进一步,由于参与者j是从$N\backslash\{i\}$中任意选择的,通过考虑所有可能的参与者j,可以得出$N\backslash\{i\}$中的所有参与者都是虚拟参与者.

证毕.

通过将某些沙普利值约束为零,3.3节中的损失函数本质上可以视为强制要求:当$k_t=k_{t-1}$时,$\boldsymbol{x}_t$上的第k_t个语义区域对于生成$\boldsymbol{x}_{t-1}$上的第k_{t-1}个语义区域必须成为虚拟参与者. 这一目标可以通过定理1以更协同高效的方式实现. 具体而言,不再需要耗费$K(K-1)$次计算沙普利值,而是采用以下等效损失函数. 如下方函数所示,该方法将DDPM每个反向步骤的时间复杂度从$O(K(K-1)2^K)$降低至$O(K2^K)$:

$$\begin{aligned}\mathcal{L}_{\mathrm{IE}}=E_t\sum_{k_{t-1}}\sum_{S\subseteq\mathcal{K}_t/\{k_t\}}&\left(v_{k_{t-1}}(S\cup\{k_t\})\right.\\&\left.-v_{k_{t-1}}(\{\{k_t\}\})\right)^2+\left(v_{k_{t-1}}(S)-v_{k_{t-1}}(\varnothing)\right)^2\end{aligned}\tag{8}$$

根据定理1,这种损失函数预期会迫使$\mathcal{K}/\{k_t\}$集合中的每个参与者成为虚拟参与者,从而实现对DDPMs可解释性特性的设计目标. 然而,$\mathcal{L}_{\mathrm{IE}}$的精确计算仍需要每个时间步$O(K2^K)$的时间复杂度. 为进一步解决时间资源受限的问题,研究人员采用多项式采样算法[25]来加速沙普利值的计算,将计算复杂度简化为$O(KC)$. 此处C代表采样次数.

通过这种方式,最终用于训练可解释的DDPM的损失函数如下:

$$\mathcal{L}=\mathcal{L}_{\mathrm{DDPM}}+\lambda_0\mathcal{L}_{\mathrm{IE}}\tag{9}$$

其中,λ_0为正向权重;$\mathcal{L}_{\mathrm{DDPM}}$表示传统DDPM损失函数[1,3];$\mathcal{L}_{\mathrm{IE}}$用于确保输入高斯噪声上的特定语义区域能独立生成输出图像上的对应语义区域.

4 实验验证

4.1 实验设置

4.1.1 数据集

在实验中,研究人员将方法应用于最先进的DDPM模型[1,3],这些模型在分辨率为64×64的CelebA数据集[46]上训练;以及应用于在分辨率为256×256的CelebA-HQ数据集[47]上训练的潜在扩散模型(Latent Diffusion Model,LDM)[9].

4.1.2 评估细节

本文从定性和定量两个维度对可解释扩散模型(Interpretable DDPM)与可解释潜在扩散模型(Interpretable LDM)进行了系统评估. 在定性评估方面,通过重采样特定语义区域内的输入高斯噪声,可视化展示了生成图像中修改特定视觉概念的效果. 实验还验证了本方法能通过交换选定区域内的输入高斯噪声,实现生成图像间视觉概念的互换. 基于Kim等人[14]的方法,进一步将该技5术拓展至真实图像的视觉概念编辑. 定量评估方面,参照Li等人[48]的方法计算原始图像与修改图像的均方误差(Mean Squared Error,MSE)以评估修改的局部性,同时采用弗雷歇初始距离(Fréchet Inception Distance,FID)[49]量化生成图像的真实性. 实验结果表明,本文的方法能有效解耦不同视觉概念的生成过程.

4.1.3 实现细节

为保持生成图像的真实性,本文利用提出的损失函数(即算法部分中的$\mathcal{L}$)对预训练DDPM[1]和LDM[9]进行微调,以增强生成过程中视觉概念的解耦能力. 生成图像的视觉概念划分通过预训练人脸分割模型BiSeNet[19]实现,该模型将输入人脸图像分割为11个视觉概念:皮肤、眉毛、眼睛、耳朵、鼻子、嘴巴、嘴唇、颈部、衣物、头发与帽子. 通过对每个语义区域施加膨胀操作来增强模型的图像编辑能力,从而实现更广泛的泛化性. 为简化流程,在训练可解释扩散模型时,仅针对特定视觉概念与其对应背景区域进行解耦生成. 为此选取眼睛、鼻子和嘴巴3个视觉概念,分别训练3个

可解释DDPM模型,命名为可解释DDPM(鼻子)、可解释DDPM(嘴巴)与可解释DDPM(眼睛). 可解释LDM采用相同机制.

4.1.4 训练细节

遵循DDPM[1]与LDM[9]的原始设置,分别训练可解释DDPM和LDM. 学习率设为2×10^{-4},所有实验均采用衰减因子为0.999 9的指数移动平均(Exponential Moving Average,EMA). 训练可解释DDPM(鼻部)时,λ_0初始权重设为1×10^3并以指数衰减至10;对于可解释DDPM(眼部)和可解释DDPM(嘴部),λ_0初始权重设为1×10^3并以指数衰减至1. 模型在四块NVIDIA V100 GPU上训练36 h,确保高效收敛与优化.

4.2 定性评估

4.2.1 单独修改视觉概念

本文提出的可解释性LDM能够针对性地调整生成图像中的不同视觉概念. 通过对相关图像区域的高斯噪声进行随机重采样,可在保持整体结构和真实感的同时,实现特定语义区域的局部变化. 如图1所示,相较于原始LDM的重采样技术,本文的方法展现出更优异的局部编辑能力.

4.2.2 图像对之间的视觉概念交换

此外,该方法还能通过交换对应区域内的输入高斯噪声,实现图像间视觉概念的互换. 为评估修改的局部性,在Lab色彩空间[50]中计算的MSE映射图可直观呈现被修改区域. 图2展示了图像对之间有效的视觉概念交换效果. 基于Kim于2022年提出的方法,使用潜在扩散模型(LDM)对真实图像的眼睛(图2(a))和嘴巴(图2(b))进行替换. 在此过程中,我们将原始图像中的视觉概念替换为源图像中的对应部分. 均方误差热力图用于量化原始图像与修改后图像之间的差异. 与传统LDM相比,我们提出的可解释LDM成功实现了特定视觉概念的生成解耦,且不会干扰其他无关区域.

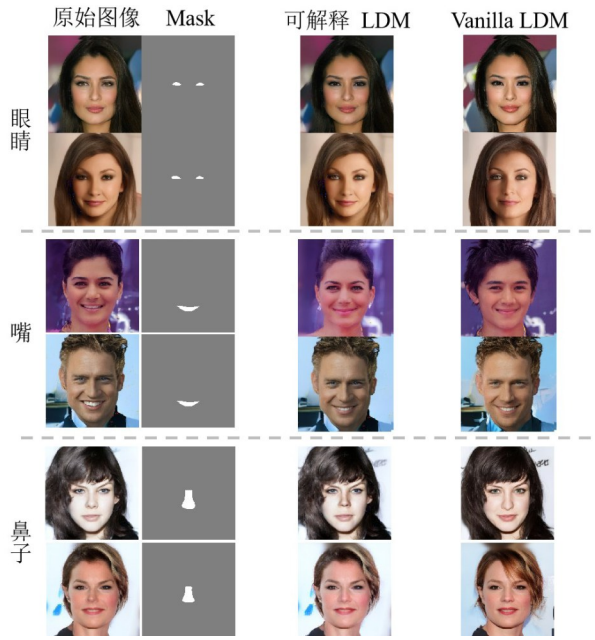


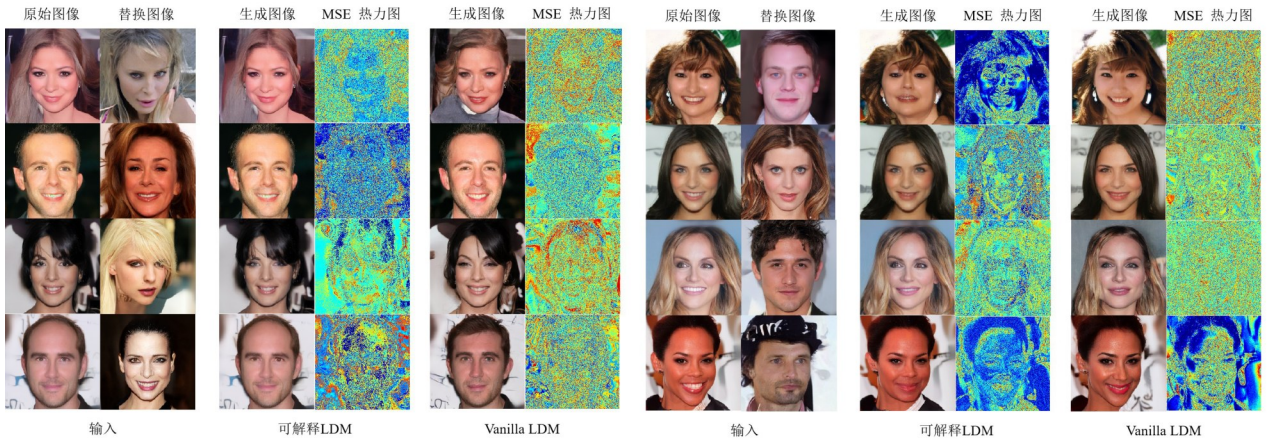
图1　编辑图像的局部视觉概念图,与原始潜在扩散模型的重采样结果进行对比

4.3 定量评估

4.3.1 局部性评估

首先,研究人员沿用Li等人[48]的方法对本方案的局部性进行量化评估,该方法用于评估修改特定视觉概念的局部性效果. 该局部性指标用于衡量特定区域内的高斯噪声是否能在图像中独立生成对应视觉概念,而不影响其他无关区域. 具体而言,生成100组图像对,通过交换输入高斯噪声上的对应区域来实现特定视觉概念的置换.

给定生成图像$\boldsymbol{x}_0\in R^D$和修改后图像$\boldsymbol{x}_0'\in R^D$,令$M_{k_0}(\boldsymbol{x}_0)$表示通过BiSeNet[19]获得的二值掩码(取值为$\{0,1\}^D$),其表征图像$\boldsymbol{x}_0$上第k_0个视觉概念的语义区域



(a) 替换眼睛　　(b) 替换嘴巴

图2　基于Kim于2022年提出的方法,使用潜在扩散模型(LDM)对真实图像的眼睛和嘴巴进行替换

$(k_0=1,2,\cdots,K)$. $M'_{k_0}(\boldsymbol{x}_0)=1-M_{k_0}(\boldsymbol{x}_0)$表示反向掩码. 内部均方误差定义为$\mathrm{MSE}(\mathrm{in},k_0)=\dfrac{M_{k_0}(\boldsymbol{x}_0)\cdot(\boldsymbol{x}_0-\boldsymbol{x}'_0)^2}{\left\|M_{k_0}(\boldsymbol{x}_0)\right\|_1}$. 外部均方误差定义为$\mathrm{MSE}(\mathrm{out},k_0)=\dfrac{M'_{k_0}(\boldsymbol{x}_0)\cdot(\boldsymbol{x}_0-\boldsymbol{x}'_0)^2}{\left\|M_{k_0}(\boldsymbol{x}_0)\right\|_1}$. 研究人员按照Li等人[48]的方法计算局部性指标$\dfrac{\mathrm{MSE}(\mathrm{out},k_0)}{\mathrm{MSE}(\mathrm{in},k_0)}$. 该指标值越小,表明在语义区域内对图像的修改越精确,且不影响其他无关区域. 表1展示了各视觉概念的局部性指标结果,本方案展现出卓越的定位能力,说明其他语义无关区域的改变微乎其微.

表1 局部性评估

模型名称	嘴	眼睛	鼻子
Stable Diffusion	0.128 4	0.230 7	0.170 7
RePaint	0.055 3	0.780 6	0.112 5
SEGA	0.861 7	0.695 6	1.018 9
Editing in Style	1.364 9	0.974 5	2.942 3
IGAN	0.060 6	0.080 2	0.099 5
LDM	0.051 1	0.086 7	0.350 6
DDPM	0.086 9	0.083 7	0.203 5
可解释LDM	0.035 7	0.064 0	0.319 1
可解释DDPM	**0.036 9**	**0.029 5**	**0.068 3**

注:最佳结果以粗体显示,次佳结果以下划线标注.

表格中展示了与以下方法比较的结果:Stable Diffusion[9]、RePaint[7]、SEGA[51]、Editing in Style[50]、IGAN[48]、LDM[9]以及DDPM[1].

4.3.2 真实性评估

为实现对图像的真实性评估,本文计算了生成图像与真实图像分布之间的FID[49]. FID通过特征空间量化图像集之间的相似度,得分越低表明相似度越高、图像质量越好. 表2展示了真实图像与采用DDPM采样方法生成的1万张图像之间的FID结果. 结果表明,引入解构的区域化的视觉概念可能会影响图像的真实感,可解释的DDPM相较于原版生成的图像FID值都有了一定的上涨,这在之前的研究中同样出现过[35],之前的工作[45]已经证明提升模型可解释性的同时不可避免地会影响性能.

表2 真实性评估

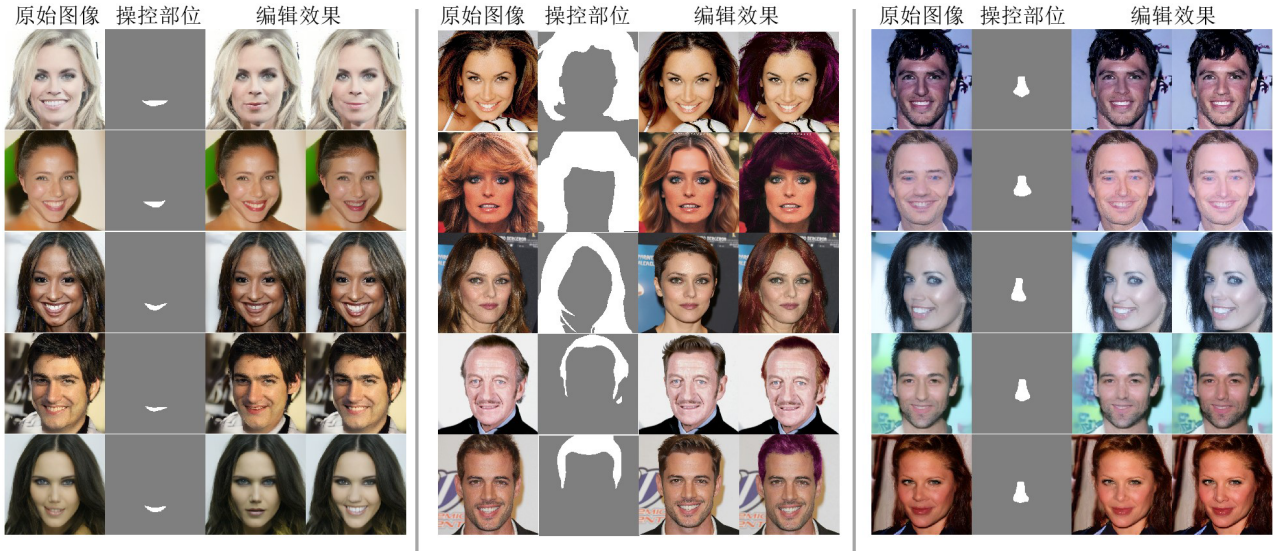
模型名称	FID
DDPM	12.86
可解释DDPM(鼻子)	18.81
可解释DDPM(眼睛)	19.42
可解释DDPM(嘴)	18.44

4.4 真实图像编辑

受DiffusionCLIP[14]启发,本文的方法亦可扩展至真实图像编辑领域. 具体而言,按以下步骤执行前向扩散过程:

$$\boldsymbol{x}_{t+1}=\sqrt{\alpha_{t+1}}\left(\frac{\boldsymbol{x}_t-\sqrt{1-\bar{\alpha}_t}\int_\theta(\boldsymbol{x}_t,t)}{\sqrt{\alpha_t}}\right)+\sqrt{1-\alpha_{t+1}}\int_\theta(\boldsymbol{x}_t,t) \tag{10}$$

通过这种方式,操纵最终的高斯噪声$\boldsymbol{x}_T$并逆转DDPM过程以生成修改后的图像. 如图3所示,本文的方法通过改变对应区域内的高斯输入噪声,实现了对真实图像中视觉概念的修改. 图中我们分别展示了对图像嘴、头发和鼻子的编辑结果,每个部位均展示了两种不同的高斯噪声输入的生成结果. 图像第1列为原



注:第2列表示要修改的选定组件.第3和第4列显示使用不同采样噪声修改后的图像.

图3 修改原始图像的特定视觉概念

始图像,第2列为选定的编辑部位,第3、4列为两种不同的高斯噪声输入所生成的结果. 可视化结果凸显了本方法在改进真实图像局部修改方面的卓越能力. 在使用两种不同的高斯噪声作为输入的情况下,可解释的DDPM均准确地对目标区域进行了合理的编辑. 通过对比两种高斯噪声产生的结果和原始图像可以发现,在任一输入下未被选定编辑的部分均忠实的还原了原图的样貌. 这样的结果展示了可解释DDPM在真实图像编辑领域的强大性能.

5　结论

本文聚焦于提升扩散模型的可解释性,通过在图像去噪场景中促进解耦视觉概念的生成. 研究人员运用博弈论设计定制化的损失函数,使输入高斯噪声中对应区域能精确操控视觉概念. 同时提出降低计算复杂度的定理,使可解释性优化具有可行性. 实验验证了该方法在保持图像质量的前提下有效解耦视觉概念,并成功拓展至真实图像编辑应用. 研究表明,本文的方法揭示了噪声与视觉概念间的交互机制,阐明了视觉概念的生成过程. 更重要的是,该方法阐明了去噪扩散概率模型(DDPMs)处理复杂视觉抽象的路径,有望推动扩散模型可解释性领域的未来研究.

致谢　感谢杨传广研究员,以及安竹林研究员给本文提出的参考意见.

参考文献

[1] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[EB/OL]. (2020-12-16)[2025-11-11]. https://arxiv.org/abs/2006.11239.

[2] NICHOL A, DHARIWAL P, RAMESH A, et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models[EB/OL]. (2022-03-08)[2025-11-11]. https://arxiv.org/abs/2112.10741.

[3] SONG Y, ERMON S. Generative modeling by estimating gradients of the data distribution[C]//Neural Information Processing Systems. Curran Associates Inc.: Red Hook, 2019: 11918 - 1193.

[4] SONG J M, MENG C L, ERMON S. Denoising diffusion implicit models[EB/OL]. (2022-10-05)[2025-11-11]. https://arXiv.org/abs/2010.02502.

[5] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[EB/OL]. (2021-06-01) [2025-11-11]. https://arXiv.org/abs/2105.05233.

[6] SONG Y, ERMON S. Improved techniques for training score-based generative models[EB/OL]. (2020-10-23) [2025-11-11]. https://arXiv.org/abs/2006.09011.

[7] LUGMAYR A, DANELLJAN M, ROMERO A, et al. RePaint: Inpainting using denoising diffusion probabilistic models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 11451-11461.

[8] CHOI J, KIM S, JEONG Y, et al. ILVR: Conditioning method for denoising diffusion probabilistic models[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 14347-14356.

[9] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 10674-10685.

[10] NICHOL A, DHARIWAL P. Improved denoising diffusion probabilistic models[EB/OL]. (2021-02-18)[2025-11-11]. https://arXiv.org/abs/2102.09672.

[11] GUO Z L, LEI C T, FANG L, et al. A gray-box attack against latent diffusion model-based image editing by posterior collapse[EB/OL]. (2024-09-20) [2025-11-11]. https://arXiv.org/abs/2408.10901.

[12] HO J, SALIMANS T. Classifier-free diffusion guidance[EB/OL]. (2022-07-26) [2025-11-11]. https://arxiv.org/abs/2207.12598.

[13] KWON M, JEONG J, UH Y. Diffusion models already have a semantic latent space[EB/OL]. (2023-03-29)[2025-11-11]. https://arXiv.org/abs/2210.10960.

[14] KIM G, KWON T, YE J C. DiffusionCLIP: Text-guided diffusion models for robust image manipulation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 2416-2425.

[15] RAMESH A, DHARIWAL P, NICHOL A, et al. Systems and methods for hierarchical text-conditional image generation: U.S. Patent Application 18/419,675[P]. 2024-10-03.

[16] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with CLIP latents[EB/OL]. (2022-04-13)[2025-11-11]. https://arXiv.org/abs/2204.06125.

[17] WANG Y H, YU J W, ZHANG J. Zero-shot image restoration using denoising diffusion null-space model[EB/OL]. (2022-12-07)[2025-11-11]. https://arXiv.org/abs/2212.00490.

[18] LI X M, HOU X Y, LOY C C. When StyleGAN meets stable diffusion: A W_+ adapter for personalized image generation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 2187-2196.

[19] TSAI T H, TSENG Y W. BiSeNet V3: Bilateral segmenta-

tion network with coordinate attention for real-time semantic segmentation[J]. Neurocomputing, 2023, 532: 33-42.

[20] SHAPLEY L S .A value for n-person games[J/OL]. Annals of Mathematical Studies, 1953. DOI: 10.1017/CBO9780511528446.003.

[21] DUBEY P, WEBER R J. Probabilistic values for games[J]. Cowles Foundation for Research in Economics, 1977.

[22] SARKAR S, BABU A R, MOUSAVI S, et al. RL-CAM: Visual explanations for convolutional networks using reinforcement learning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2023: 3861-3869.

[23] ZHAO X, WANG L M, ZHANG Y F, et al. A review of convolutional neural networks in computer vision[J]. Artificial Intelligence Review, 2024, 57(4): 99.

[24] SHRIKUMAR A, GREENSIDE P, SHCHERBINA A, et al. Not just a black box: Learning important features through propagating activation differences[EB/OL]. (2017-04-11)[2025-11-11]. https://arXiv.org/abs/1605.01713.

[25] CASTRO J, GÓMEZ D, TEJADA J. Polynomial calculation of the Shapley value based on sampling[J]. Computers & Operations Research, 2009, 36(5): 1726-1730.

[26] RUIZ N, LI Y Z, JAMPANI V, et al. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 22500-22510.

[27] YANG B X, GU S Y, ZHANG B, et al. Paint by example: Exemplar-based image editing with diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 18381-18391.

[28] HIGGINS I, MATTHEY L, PAL A, et al. Beta-VAE: Learning basic visual concepts with a constrained variational framework[C]//International Conference on Learning Representations. Appleton: ICLR, 2017: 1-13.

[29] XIE R C, DU C, SONG P, et al. MUSE-VL: Modeling unified VLM through semantic discrete encoding[EB/OL]. (2025-07-28) [2025-11-11]. https://arXiv. org/abs/2411.17762.

[30] CHEN X, DUAN Y, HOUTHOOFT R, et al. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets[EB/OL]. (2016-06-12)[2025-11-11]. https://arXiv.org/abs/1606.03657.

[31] BAU D, ZHOU B L, KHOSLA A, et al. Network dissection: Quantifying interpretability of deep visual representations[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 3319-3327.

[32] MAHENDRAN A, VEDALDI A. Understanding deep image representations by inverting them[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 5188-5196.

[33] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2921-2929.

[34] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

[35] LIANG H Y, OUYANG Z H, ZENG Y Y, et al. Training interpretable convolutional neural networks by differentiating class-specific filters[C]//Computer Vision - ECCV 2020. Cham: Springer, 2020: 622-638.

[36] MBACKE S D, CLERC F, GERMAIN P. Statistical guarantees for variational autoencoders using PAC-Bayesian theory[EB/OL]. (2023-12-07)[2025-11-11]. https://arXiv.org/abs/2310.04935.

[37] KUMAR A, SATTIGERI P, BALAKRISHNAN A. Variational inference of disentangled latent concepts from unlabeled observations[EB/OL]. (2018-12-27)[2025-11-11]. https://arXiv.org/abs/1711.00848.

[38] TRAN L, YIN X, LIU X M. Disentangled representation learning GAN for pose-invariant face recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1283-1292.

[39] AVRAHAMI O, LISCHINSKI D, FRIED O. Blended diffusion for text-driven editing of natural images[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 18187-18197.

[40] HERTZ A, MOKADY R, TENENBAUM J, et al. Prompt-to-prompt image editing with cross attention control[EB/OL]. (2022-08-02) [2025-11-11]. https://arXiv. org/abs/2208.01626.

[41] KAWAR B, ZADA S, LANG O, et al. Imagic: Text-based real image editing with diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 6007-6017.

[42] VALEVSKI D, KALMAN M, MOLAD E, et al. UniTune: Text-driven image editing by fine tuning a diffusion model on a single image[J]. 2023, 42(4): 1-10.

[43] COUAIRON G, VERBEEK J, SCHWENK H, et al. DiffEdit: Diffusion-based semantic image editing with mask guidance[EB/OL]. (2022-10-20)[2025-11-11]. https://arXiv.org/abs/2210.11427.

[44] PREECHAKUL K, CHATTHEE N, WIZADWONGSA S, et al. Diffusion autoencoders: Toward a meaningful and decodable representation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 10609-10619.
[45] LOCATELLO F, BAUER S, LUCIC M, et al. Challenging common assumptions in the unsupervised learning of disentangled representations[EB/OL]. (2019-06-18)[2025-11-11]. https://arXiv.org/abs/1811.12359.
[46] LIU Z W, LUO P, WANG X G, et al. Deep learning face attributes in the wild[C]//2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2016: 3730-3738.
[47] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[EB/OL]. (2018-02-26) [2025-11-11]. https://arXiv.org/abs/1710.10196.
[48] LI C, YAO K L, WANG J, et al. Interpretable generative adversarial networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 1280-1288.
[49] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[EB/OL]. (2018-01-12)[2025-11-11]. https://arxiv.org/abs/1706.08500.
[50] COLLINS E, BALA R, PRICE B, et al. Editing in style: Uncovering the local semantics of GANs[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 5770-5779.
[51] BRACK M, FRIEDRICH F, HINTERSDORF D, et al. SEGA: Instructing text-to-image models using semantic guidance[EB/OL]. (2023-11-02) [2025-11-11]. https://arxiv.org/abs/2301.12247.

作者简介

刘超一　男,1998年5月出生于山东省济南市. 现为中国科学院计算技术研究所博士研究生. 主要研究方向为领域泛化.
E-mail: liuchaoyi22@mails.ucas.ac.cn

耿浩棒　男,2000年10月出生于河南省郑州市. 2024年6月毕业于中国科学院计算技术研究所. 主要研究方向为基于扩散模型的视觉生成.
E-mail: haobang.geng@kunlun-inc.com

葛亚维　男,1990年10月出生于山东省枣庄市. 现为军事科学院战略评估咨询中心助理研究员. 主要研究方向为军事评估与运筹决策.
E-mail: vvrues11@163.com

林　晗　男,1998年2月出生于福建省莆田市. 现为军事科学院战略评估咨询中心博士研究生. 主要研究方向为军事评估及因果推断技术.
E-mail: lh98cool@163.com

侯　娜　女,1982年8月出生于陕西省西安市. 现任中国人民解放军32801部队副研究员. 主要研究方向为人工智能、大数据.
E-mail: xiangku860101@163.com

赵二虎　男,1985年9月出生于河北省邢台市. 博士、高级工程师、硕士生导师,就职于中国科学院计算技术研究所,任装备智能系统研究中心智算平台研究组组长. 主要研究方向为嵌入式智能计算系统、专用计算机系统、芯片微系统结构.
E-mail: zhaoerhu@ict.ac.cn

黄礼泊　男,1992年7月出生于江西省吉安市. 现为中国科学院计算技术研究所助理研究员. 主要方向为机器学习与人工智能.
E-mail: huanglibo@ict.ac.cn

徐勇军　男,1979年7月出生于四川省成都市. 中国科学院计算技术研究所正高级工程师、研究员、博士生导师,现任该所专项技术研究中心主任、国防科工局“华罗庚”创新中心常务副主任. 主要研究方向为人工智能系统、大数据处理技术.
E-mail: xyj@ict.ac.cn