**中图法分类号**:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-22

**论文引用格式**:Weng Zihui, Zhang Quan, Xie Xiaohua, Lai Jianhuang. A survey on memorization and forgetting mechanisms in diffusion models[J/OL]. Journal of Image and Graphics, XXXX:1-22. DOI: 10.11834/jig.260095.(翁子辉, 张权, 谢晓华, 赖剑煌. 扩散模型中的记忆遗忘机制综述[J/OL]. 中国图象图形学报,XXXX:1-22. DOI: 10.11834/jig.260095.)[DOI:10.11834/jig.260095]

# 扩散模型中的记忆遗忘机制综述

翁子辉[1,2],张权[3],谢晓华[1,2],赖剑煌[1,2*]

1. 中山大学计算机学院　广州市　510006; 2. 广东省信息安全技术重点实验室(中山大学)　广州市　510006; 3. 中山大学系统科学与工程学院　广州市　510275

**摘　要**:扩散模型已迅速发展成为生成式视觉模型的主要范式。然而,模型所固有的记忆遗忘机制使得模型不自觉地记忆训练数据集中的敏感信息,进而加剧了图像领域的隐私安全和版权问题。尽管记忆与遗忘机制在语言模型领域已有深入研究,但在扩散模型这一视觉生成任务的核心技术中,仍缺乏系统性的综述。本综述旨在填补这一空白,具体从扩散模型的理论建模和模型架构的介绍,到模型记忆在非时序和时序扩散模型上的定义,对于模型记忆在扩散模型上的理解和对模型记忆在模型审计方和恶意攻击方两方面的量化方法,再到模型遗忘统一框架下差分隐私、提示词优化、模型遗忘的模型记忆缓解方法五个方面进行探讨,最后,本文展望了扩散模型中记忆-遗忘机制的未来发展方向,并重点指出了当前面临的关键挑战,包括隐私数据处理流水线和基准测试亟待规范,更符合扩散模型特性的模型记忆定义和模型遗忘算法,新学习场景的模型记忆-遗忘机制和垂直领域落地。

**关键词**:扩散模型;人工智能安全;隐私安全;模型记忆;模型遗忘

# A survey on memorization and forgetting mechanisms in diffusion models

Weng Zihui[1,2], Zhang Quan[3], Xie Xiaohua[1,2], Lai Jianhuang[1,2*]

*1. School of Computer Science and Engineering Sun Yat-sen University, Guangzhou 510006; 2. Guangdong Province Key Laboratory of Information Security Technology, Guangzhou 510006; 3. School of Systems Science and Engineering Guangzhou 510275*

**Abstract**: Diffusion models have rapidly become the dominant paradigm of generative visual models. However, the memorization-forgetting mechanisms inherent in these models make them unintentionally memorize sensitive information from training datasets, which further aggravates the privacy and copyright issues in the image domain. Although the memorization and forgetting mechanisms have been deeply studied in the field of language models, a systematic review on diffusion models, which are the core technology of visual generation tasks, is still lacking. Our survey aims to bridge this gap, and the existing works are reviewed from five aspects critically. First, we briefly introduce the theoretical modeling and architecture of diffusion models, including the formulations of denoising diffusion probabilistic models, score-based generative models, and flow matching, as well as the architectures of U-Net, DiT and MMDiT. Such theoretical and architectural foundations not only support the strong generative capability of diffusion models, but also provide the necessary preliminaries for understanding why and how memorization arises during the iterative denoising process. On this basis, the definitions of memorization in diffusion models are introduced from non-temporal and temporal perspectives. For non-temporal diffu-

**收稿日期**:2026-02-11;**修回日期**:2026-05-28

*通信作者:赖剑煌 stsljh@mail.sysu.edu.cn

**基金项目**:国家自然科学基金项目(12326618);广东省信息安全技术重点实验室项目(2023B1212060026)

**Supported by**: National Natural Science Foundation of China (12326618); Project of Guangdong Provincial Key Laboratory of Information Security Technology (2023B1212060026)

sion models, the memorization can be divided into global memorization and local memorization, where the former focuses on the global similarity between generated images and training images, and the latter is concerned with the partial replication of sensitive components such as faces, signatures and watermarks. For temporal diffusion models, the memorization can be further decomposed into content memorization and motion memorization, since the leakage in video diffusion models is reflected not only in the static appearance of objects but also in the dynamic motion patterns across frames. Once the manifestations of memorization are formally defined, a natural follow-up question is what factors give rise to such memorization behaviors and how they can be theoretically interpreted. Accordingly, the understanding of memorization in diffusion models is then discussed from the perspectives of model factor, data factor and theoretical view. The model factor is focused on the influence of over-parameterization and architectural components such as the cross-attention modules, while the data factor reveals that the long-tailed distribution and duplicated samples can significantly increase the memorization risk. The theoretical view is supported by the manifold memorization hypothesis and the geometry-adaptive harmonic representation, which explain why diffusion models tend to memorize specific samples from a geometric standpoint. Building upon these qualitative understandings, more quantitative tools are required for the practical assessment of memorization risks. To this end, the quantification methods of memorization are introduced from the perspectives of model auditor and malicious attacker. The auditor-based methods can be divided into proxy-based methods and replication-based methods. The proxy-based methods are designed to approximate the memorization score by influence functions, gradient variance, and geometric proxies such as the Hessian curvature. The replication-based methods utilize pixel-level, perceptual-level and semantic-level similarity metrics, such as SSIM, LPIPS, SSCD and CLIP, to detect the duplication between generated images and training images. The attacker-based methods can be divided into membership inference attacks and extraction attacks, which are designed to determine whether a sample is in the training dataset or to recover the original training samples directly. The threat models can be further categorized into white-box, gray-box and black-box settings according to the attacker's knowledge of the target model. Given that these quantification methods have empirically verified the severity of memorization-induced privacy leakage, how to effectively mitigate such risks becomes the next critical concern. Therefore, the memorization mitigation methods under a unified framework of machine unlearning are introduced, including 1) differential privacy, 2) prompt optimization, and 3) machine unlearning. The differential privacy based methods are typically applied in the pre-training stage and can provide formal privacy guarantees through input perturbation, output perturbation, or gradient perturbation such as DP-SGD, but at the cost of generative utility. The prompt optimization based methods intervene at the inference stage by perturbing the text embedding or attention scores to prevent the model from generating memorized content, which is lightweight but less rigorous. The machine unlearning based methods can be focused on fine-tuning the model parameters to redirect the memorized concepts to null concepts, locating and editing the memorization-related neurons, or applying closed-form parameter editing to remove the influence of specific samples or concepts in a post-hoc manner. These three families of methods are complementary to one another in terms of intervention stage, computational overhead and privacy guarantee, and their integration provides a more comprehensive perspective on memorization mitigation in diffusion models. To sum up, although a series of progress has been made along the above five aspects, several critical challenges still remain to be addressed in future research. 1) The privacy data processing pipeline and benchmarks of diffusion models are required to be standardized further, since the absence of unified datasets and evaluation protocols hampers fair comparison and reproducibility. 2) The memorization definitions and machine unlearning algorithms that are more compatible with the characteristics of diffusion models are required to be developed, since the existing methods are mostly inherited from classifiers or language models and cannot fully exploit the iterative and cross-modal nature of diffusion processes. 3) The memorization-forgetting mechanisms in novel learning scenarios such as test-time training, federated learning and continual learning are required to be explored, and the deployment in vertical domains such as medical imaging and financial generation is required to be accelerated for the cooperation of academia and industry. Furthermore, the data privacy policy-relevant ethical issues need to be considered for the responsible deployment of diffusion models in the future.

**Key words**: diffusion model; AI safety; privacy; memorization; forgetting

**论文引用格式：**[DOI:10.11834/jig.260095]

## 0 引 言

模型记忆(Wei等,2025)是一种模型存储并重现其训练集中的特定数据点的现象,与之对立的是模型遗忘,即给定遗忘数据集的条件,模型遗忘算法针对已训练的模型特定地消除某些数据样本的影响,无需从头训练使得模型的表现等同于或者趋近于遗忘数据集未参与训练。在有监督场景下,模型不自觉地记忆训练数据集中的敏感信息,进而存在因模型输出(Shokri等,2017)或者对抗性攻击,例如成员推断攻击、数据提取攻击而泄露的风险(Song等,2019)。

随着深度学习的发展(Feng等,2025;Ye等,2025;Zhang等,2023b;Xie等,2022),特别是缩放定律(scaling law)(Kaplan等,2020)的验证,大语言模型(large language models, LLMs)(Guo等,2025;Achiam等,2023)和视觉模型(Dosovitskiy等,2021;Peebles等,2023)的参数量和数据规模大幅增长,进一步放大了模型记忆现象。扩散模型(diffusion models, DM)作为当前视觉生成任务的主流范式,通过学习逆向去噪过程将高斯噪声还原为复杂数据分布,展现了卓越的生成能力(Peebles等,2023)并在视觉生成任务中得到了广泛的应用。尽管模型记忆和遗忘在大语言模型中已经得到较为充分的研究,并且有相关的大语言模型的模型记忆机制综述(Li等,2025b;Wei等,2025;Xiong等,2025;Hartmann等,2023),目前关于模型记忆-遗忘机制在扩散模型中的研究仍然缺乏系统性的研究。然而,其强大的生成能力同样伴随着对训练数据的记忆风险,如何理解扩散模型中的记忆现象并利用模型遗忘缓解其隐私安全隐患,已成为该领域亟待解决的关键课题。

本综述旨在从记忆-遗忘机制的角度,全面概述扩散模型中的记忆-遗忘机制,图1给出了本文综述总览。本文首先探讨了扩散模型中记忆-遗忘机制的不同理论理解,探索如何定义某些数据点被记忆。接着,本文对记忆评估和对应的模型遗忘方法的常见思路进行分类和分析,阐述它们的方法、优势和局限性。最后,本文介绍了其在隐私审计和医学影像中的下游应用,并讨论了在这个快速发展的领域中存在的开放性挑战和未来方向,强调了对视觉模型更加精确的记忆定义方式、更有效用的隐私保护方法、探索测试阶段的模型记忆-遗忘机制以及基准测试的统一等方面的展望。

本综述的主要贡献如下:

1)总结了在扩散模型中,非时序模型和时序模型场景下关于模型记忆的工作,并从模型,数据和理论视角介绍模型记忆的理解;从模型审计方和恶意攻击方的角度介绍模型记忆能力的量化方法。

2)从模型预训练、微调、后训练和推理阶段介绍了不同的模型记忆缓解方法,并给出了多项量化指标比较。

3)展望了扩散模型下的模型记忆-遗忘机制的三个未来方向:隐私数据流水线与基准测试的标准化;适配扩散模型特性的记忆定义与遗忘算法创新;新学习范式下的机制探索与垂直领域应用。

本综述的概览如图1所示:

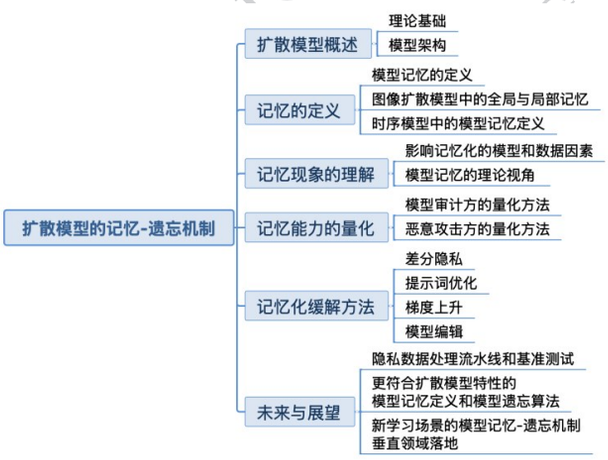



图1 本综述概览

Fig. 1 Overview of our survey

## 1 生成式扩散模型概述

扩散模型(Ho等,2020;Song等,2021;Croitoru等,2023;Lipman等,2023;Liu等,2023)已成为现代生成式建模的核心范式。从宏观层面看,它们构建了一个前向加噪过程,通过逐步向真实数据

添加噪声直至其转化为生成易于采样的分布(通常为各向同性高斯分布),随后学习一个反向生成过程过程,逐步消除噪声以从数据分布中重建样本。该设计能实现稳定优化与强可控性,使扩散模型成为图像生成领域的尖端技术,并在多模态与视

频生成场景中展现出日益显著的优势。本章节将从以下三个方面全面介绍扩散模型：理论基础、架构演进及向多模态与视频场景的扩展。

### 1.1 理论基础

扩散模型的理论建模经历了从去噪扩散概率模型到得分匹配和流匹配的转变，其建模方法如图2所示：

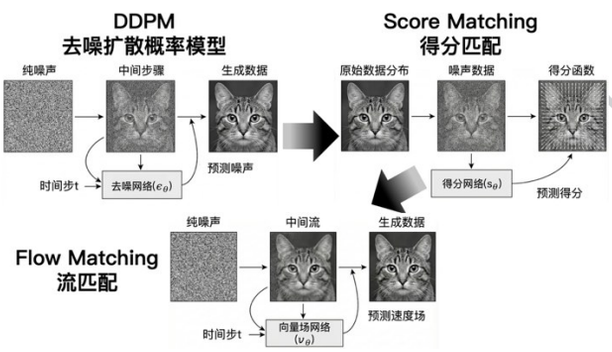



图2 扩散模型三种建模方法

Fig. 2 Three modeling of diffusion models

#### 1.1.1 基于噪声预测的建模：去噪扩散概率模型

去噪扩散概率模型（denoising diffusion probabilistic models，DDPM）（Ho等，2020）将扩散过程形式化为一对马尔可夫链，该模型遵循类似热力学扩散的过程，包含两个组成部分。前向过程从干净的数据样本$\boldsymbol{x}_0 \sim p_{\text{data}}$开始，逐步注入高斯噪声：

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \#(1)$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := N(\boldsymbol{x}_t; \sqrt{\alpha_t}\,\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}), \#(2)$$

式中，$q(\cdot|\cdot)$为前向过程的概率分布，$\boldsymbol{x}_t$是第$t$步的加噪结果，$N(\cdot;\sqrt{\alpha_t}\,\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})$表示均值为$\sqrt{\alpha_t}\,\boldsymbol{x}_{t-1}$，方差为$\beta_t \boldsymbol{I}$的正态分布，$\{\alpha_t\}$和$\{\beta_t\}$是预定义的均值和方差权重，通过选择较小的$\beta_t$，每一步仅对该模型进行平滑扰动确保数据到纯噪声的平滑路径。而DDPM的关键优势在于$\boldsymbol{x}_t$在$\boldsymbol{x}_0$的条件下其边缘分布可以通过累乘和重参数化得到其闭式解：

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = N\left(\sqrt{\overline{\alpha}_t}\,\boldsymbol{x}_0, \left(1-\overline{\alpha}_t\right)\boldsymbol{I}\right), \quad (3)$$

式中，$\alpha_t = 1 - \beta_t$以及$\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$，能够通过原始样本$\boldsymbol{x}_0$和高斯噪声的线性组合获得$\boldsymbol{x}_t$：

$$\boldsymbol{x}_t = \sqrt{\overline{\alpha}_t}\,\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \boldsymbol{I}). \#(4)$$

反向过程旨在从噪声中逐步重建数据，该噪声被建模为参数化马尔可夫链转移的序列，其数学表达式为：

$$p_\theta(\boldsymbol{x}_{0:T}) := p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \#(5)$$

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := N\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)\right), \#(6)$$

式中，$p(\boldsymbol{x}_T)$和$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$分别表示初始高斯分布和以$\theta$为模型参数的模型学习到的条件概率分布，其均值和方差分别为$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$和$\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$。

该正向-反向框架的优化目标是通过证据下界逼近法最小化负对数似然函数：

$$\mathbb{E}\left[-\log p_\theta(\boldsymbol{x}_0)\right] \leqslant \mathbb{E}_q\left[-\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] = \mathbb{E}_q\Bigg[-\log p(\boldsymbol{x}_T) - \sum_{t\geqslant 1}\log \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\Bigg] =: L. \#(7)$$

该式子可以分解成：

$$L = \mathbb{E}_q[KL(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T)) - \log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) + \sum_{t=2}^{T} KL(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))], \#(8)$$

式中，$KL$表示KL散度（Kullback-Leibler divergence）。通过展开此公式并重新参数化，最终将目标函数简化为采样噪声与预测噪声之间的均方误差：

$$L_\vartheta(\theta) = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}[w_t \parallel \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\parallel_2^2], \#(9)$$

式中，$\boldsymbol{x}_t = \sqrt{\overline{\alpha}_t}\,\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\,\boldsymbol{\epsilon}$，$\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$表示以$\boldsymbol{x}_t$和时间$t$为输入所预测的噪声，$w_t$表示时间步依赖的权重系数。

#### 1.1.2 基于得分的建模：得分匹配

基于得分匹配的生成模型（score matching）提供了更普遍的连续时间视角：若能估计噪声数据分布在每个时间点的对数密度梯度——即评分函数的梯度（Song等，2021），则能构建预测该向量场的模型，将噪声逆转为原始样本。如下展示了基于得分的生成模型被表述为正向时间随机微分方程：

$$d\boldsymbol{x} = f(\boldsymbol{x}, t)\,dt + g(t)\,d\boldsymbol{w}_t, \#(10)$$

及其唯一对应的逆向时间随机微分方程：

$$\mathrm{d}\boldsymbol{x} = (f(\boldsymbol{x}, t) - g(t)^2 \nabla_x \log p_t(\boldsymbol{x}))\mathrm{d}t + g(t)\,\mathrm{d}\overline{\boldsymbol{w}}_t, \#(11)$$

式中，$f(x, t)$和$g(t)$分别代表漂移系数和扩散系数，$\boldsymbol{w}_t$是标准布朗运动，$\overline{\boldsymbol{w}}_t$是对应的逆向时间布朗运动。$\nabla_x$表示针对$\boldsymbol{x}$计算指定函数的散度。该随机微分方程诱导出一个受扰分布族$\{p_t(\boldsymbol{x})\}$并表明该生成过程仅需要知道得分函数：

$$s^*(\boldsymbol{x},t)=\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}).\#(12)$$

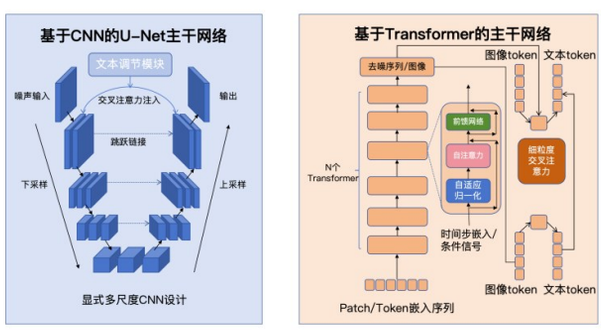
因此，学习$s_\theta(\boldsymbol{x},t)\approx s^*(\boldsymbol{x},t)$是采样的充分条件，从而得到基于得分的建模的另外一种损失函数构建方式：

$$L_\eta(\theta)=\frac{1}{2}\mathbb{E}_{p_d(\boldsymbol{x})}\left[\left\|s_m(\boldsymbol{x};\theta)-s_d(\boldsymbol{x})\right\|^2\right],\#(13)$$

其中，$s_d(\boldsymbol{x})$和$s_m(\boldsymbol{x};\theta)$分别表示真实的得分函数和模型预测的得分函数，$p_d(x)$为样本$\boldsymbol{x}$的真实分布。

1.1.3　基于速度场的建模：流匹配和校正流

流匹配（flow matching）（Lipman等，2023）提供了一种更直接的替代方案，它直接学习确定性生成常微分方程（ordinary differential equation，ODE）的时间依赖性向量场$v_\theta(\boldsymbol{x},t)$，而不是隐式推导速度场：




$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}=v_\theta(\boldsymbol{x},t),\qquad t\in[0,1],\#(14)$$

通过回归连接$\boldsymbol{x}_0\sim p_0$到$\boldsymbol{x}_1\sim p_1$作为选定条件概率路径的速度场$\boldsymbol{x}_t=\psi_t(\boldsymbol{x}_0,\boldsymbol{x}_1)$，可以得到如下的目标函数：

$$L_\kappa(\theta)=\mathbb{E}_{t,x_0,x_1}[\|v_\theta(\boldsymbol{x}_t,t)-\dot{\boldsymbol{x}}_t\|_2^2],\#(15)$$

式中，$\dot{\boldsymbol{x}}_t=\partial_t\psi_t(\boldsymbol{x}_0,\boldsymbol{x}_1)$，$\boldsymbol{x}_t=\psi_t(\boldsymbol{x}_0,\boldsymbol{x}_1)$。该方法实现了连续归一化流的训练，并提供了高效的确定流推理。校正流（Liu等，2023）可理解为一种特别简单且采样效率高的流匹配的特殊化形式，其能促进直线运输轨迹。一种常见的实现方法是采用线性插值法：

$$\boldsymbol{x}_t=(1-t)\boldsymbol{x}_0+t\boldsymbol{x}_1,\#(16)$$

使得条件速度变为

$$\dot{\boldsymbol{x}}_t=\boldsymbol{x}_1-\boldsymbol{x}_0.\#(17)$$

训练过程可简化为直接将速度回归至该恒定方向：

$$L_{\mathrm{RF}}=\mathbb{E}_{t,x_0,x_1}[\|v_\theta(\boldsymbol{x}_t,t)-(\boldsymbol{x}_1-\boldsymbol{x}_0)\|_2^2].\#(18)$$

这种“直线且快速”的设计理念源于以下观察：更直的常微分方程轨迹能够实现仅需极少函数评估即可生成高质量结果的效果。近期对校正流的改进进一步表明，通过优化重整流（rectified flow，ReFlow）方法，可以大幅降低函数调用次数，并提升一步或少步生成质量。

从如上的三种建模方式中可以看出，这种引导纯噪声走向原始数据分布的向量场路径，虽然保证了生成的保真度，但也为模型精准退火到特定训练样本（即记忆发生）提供了确定性的理论可能。

## 1.2　架构变化

扩散模型的架构经历了从U-Net到Transformer作为主干网络的过程，图3总结了扩散模型的两种主流架构的设计。

1.2.1　U-Net

早期高性能扩散模型主要依赖于U-Net（Ronneberger等，2015）主干网络，如图3左图所示，数据输入经过多尺度上采样和下采样，并在这些卷积模块中加入跳跃连接机制，使得U-Net能够同时捕捉全局结构和局部细节，以实现多分辨率去噪。而U-Net主干网络的一项重大改进是针对文本到图像任务的条件模块，由stable diffusion（Rombach等，2022）模型所提出。

图3 扩散模型两种主干网络

Fig. 3 Two backbone of diffusion models

1.2.2　DiT和MMDiT

从U-Net到DiT（diffusion Transformer）和MMDiT（multimodal diffusion Transformer）的过渡主要体现在去噪骨干网络的两大重构以及扩散模型中条件信息的整合方式的改变。

随着可扩展性成为核心问题，如图3右图所示，DiT（Peebles等，2023）用基于词元（token）的Transformer去噪器取代了多尺度卷积神经网络（convolutional neural networks，CNN）作为骨干网络：与ViT（vision Transformer）类似，数据输入通过分块或者词元（patch/token）嵌入映射到序列，堆叠的自注意力模块作为主要的噪声建模模块，时间步长和条件信号则通过Transformer模块内的归一化过程注入，从而将U-Net显式的多尺度设计转向了更统一、可扩展的表示范式。

MMDiT（Esser等，2024）在DiT的基础上，通过结构性地增强条件进入去噪器的方式，引入更明确的图像/文本分支设计，交叉注意力在选定层进行更细粒度的融合，得到更可控且语义一致的多模态交

互框架。

从显式的多尺度CNN转向高度可扩展的Transformer,带来了参数量指数级的爆发。正是这种过参数化(over-parameterization)赋予了模型足够的“冗余空间”来存储长尾和重复的训练样本。

1.2.3 扩散模型面临的数据泄露挑战

综上所述,扩散模型通过分阶段的去噪过程或者向量场建模过程实现了对复杂视觉分布的建模。然而,Kadkhodaie等人的研究(2024)解释了扩散模型的去噪器实际上在做基变换和收缩的操作,并在潜空间中寻找最优重构路径。当这种重构能力在模型缩放定律下往往会记忆一些特殊的训练数据。这种现象不仅是生成性能的副产物,更构成了严重的隐私漏洞。因此,深入探讨扩散模型中的模型记忆机制,量化其对敏感数据的记忆程度,已成为当前评估视觉生成模型安全性的首要任务。然而,要在这样一个高维连续的生成空间中准确评估并缓解数据泄露,首要挑战在于构建一个严谨且适配扩散模型特性的“模型记忆”数学定义。如下将介绍在模型记忆定义的演化和在扩散模型上的模型记忆的定义分类。

## 2 记忆的定义

由于模型记忆在作用客体、应用场景等方面的不同,本章节从如下的几类维度比较各种模型记忆的定义。

### 2.1 模型记忆的定义

对模型记忆的第一个通用定义主要集中于监督学习场景(Feldman, 2020a; Feldman 和 Zhang, 2020b)。在监督学习场景中,一条数据是否被模型记忆被定义为在有无该训练样本的情况下,模型成功预测样本标签的概率之差。形式上,对于一个数据集 $\boldsymbol{S}=\left(\boldsymbol{x}_i,\boldsymbol{y}_i\right)_{i\in[n]}$ 和 $i\in[n]$ 模型记忆定义的公式如下:

$$\boldsymbol{M}(A,\boldsymbol{S},i):=\left|Pr_{h\sim A(\boldsymbol{S})}\left[h\left(\boldsymbol{x}_i\right)=\boldsymbol{y}_i\right]-Pr_{h\sim A\left(\boldsymbol{S}^{\backslash i}\right)}\left[h\left(\boldsymbol{x}_i\right)=\boldsymbol{y}_i\right]\right|,\#(19)$$

式中,$\boldsymbol{S}^{\backslash i}$ 表示从 $\boldsymbol{S}$ 中移除 $\left(\boldsymbol{x}_i,\boldsymbol{y}_i\right)$ 后的数据集,$A$ 表示在给定数据集下的模型参数的分布,$h$ 表示从该分布中采样得到的一个具体、已完成训练的模型实例。然而这种模型记忆的定义涉及重新训练新模型而导致计算消耗极高,因此,这种模型记忆得分的计算将通过影响函数(influence function)得到改进。

由于上述定义仍然侧重于监督学习场景,仍然缺乏在视觉生成模型中记忆的精确定义。因此,Van等人(Van等,2021)提出了以下基于留一法的生成模型记忆定义,对应公式如下:

$$\boldsymbol{M}(A,\boldsymbol{D},i)=log\frac{P_A\left(\boldsymbol{x}_i|\boldsymbol{D}\right)}{P_A\left(\boldsymbol{x}_i|\boldsymbol{D}_{[n]\backslash\{i\}}\right)},\#(20)$$

式中,$P_A(\boldsymbol{x}|\boldsymbol{D})\approx\frac{1}{T}\sum_{t=1}^{T}p\left(\boldsymbol{x}|\boldsymbol{D},a_t\right)$,(21)表示模型 $\boldsymbol{A}$ 在数据集 $\boldsymbol{D}$ 上训练后,赋予观测值 $x\in X$ 的后验概率。$\boldsymbol{D}_{[n]\backslash\{i\}}$ 表示从数据集 $\boldsymbol{D}$ 中移除第 $i$ 个样本的数据集。它衡量的是,当观测值包含在训练集中时,其出现的可能性比未包含在训练集时的概率差值。

由于研究者们观察到过参数化的模型即使在图像分类任务中进行了显式正则化,也很容易拟合任意被污染的训练数据(Zhang等,2021)(例如,将狗的照片打上猫的标签);而随后关于学习动态中记忆的研究工作(Arpit等,2017)表明深度神经网络首先学习训练数据中的较为简单的模式,之后再对训练数据进行记忆。在此之前,传统统计学习理论得出的结论是,模型过拟合训练集会导致测试时误差显著增加,这可能表明记忆训练数据不利于模型的泛化能力(Hastie等,2009)。然而,这一结论是基于欠参数模型的假设,即模型参数的数量小于训练数据的数量。当模型参数数量达到或超过训练数据的数量级时,过拟合现象会变得良性(Cao等,2022; Bartlett等,2020; Tsigler等,2023; Kou等,2023):其训练误差和泛化误差都很小,即所谓的双下降现象(Nakkiran等,2021)。这表明记忆现象与过拟合本质上不同,因而在深度学习理论和应用中得到了更广泛的研究。

如上关于模型记忆的定义通常都受限于传统深度学习场景或者大语言模型场景(Schwarzschild等,2024; Carlini等,2023a; Zhang等,2023a),而由于生成式视觉模型,尤其是扩散模型主要基于从加噪到去噪的学习过程而不是固有的自回归机制,并且扩散模型的输出为高维的图像,传统的模型记忆定义难以精确计算后验概率之间的差异,因此对于扩散模型的模型记忆的定义研究更偏向考虑图像本身的特性:按照是否部分产生图像复制分成的额全局记忆和局部记忆定义,按照视频这类时序数据的特点

所给出的时序记忆等。下文将介绍扩散模型中模型记忆的现有定义。

### 2.2 图像扩散模型中的全局与局部记忆

Chen等人(2024a)的研究指出，扩散模型的记忆检测方法均基于生成结果与相应训练数据的全局相似性。当模型仅记住了训练集中的特定敏感组件(例如某个人脸特征、特定的艺术签名或水印)，并在生成时将其与全新的背景环境进行融合时，全局记忆检测方法容易产生假阴性结果，即本应被归类为真阳性结果的样本，其生成图像中只有部分组件与训练数据中的组件相同，而非全部组件。因此提出了局部记忆的概念，通过对预测噪声进行加权平均来量化记忆程度。具体而言，本研究观察到一种“bright ending”现象：当模型记忆训练数据时，最后一个令牌在最终去噪步骤中表现出异常高的交叉注意力得分，而未记忆的样本则具有接近于零的交叉注意力得分。因此，检测局部记忆的更佳指标是将此注意力分数用作掩码，以进行元素级乘法运算，公式如下：

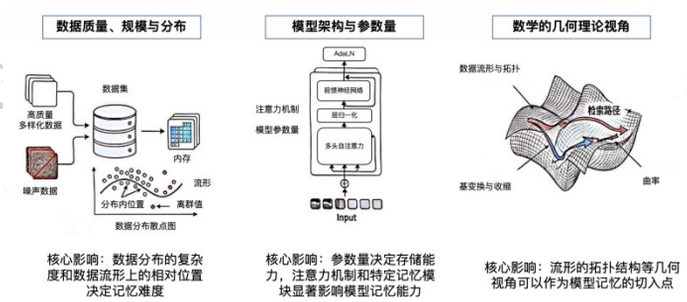
$$d = \frac{1}{T}\frac{\Lambda}{\Omega}, \#(22)$$

式中，

$$\Lambda = \sum_{t=1}^{T}\left\|\left(\boldsymbol{\varepsilon}_\theta\left(\boldsymbol{x}_t,\boldsymbol{e}_p\right) - \boldsymbol{\varepsilon}_\theta\left(\boldsymbol{x}_t,\boldsymbol{e}_\phi\right)\right)\circ \mathbf{m}\right\|_2, \quad (23)$$

$$\Omega = \sum_{i=1}^{N} m_i, \quad (24)$$

$N$是掩码$\boldsymbol{m}$中的元素个数。$\boldsymbol{\varepsilon}_\theta\left(\boldsymbol{x}_t, \boldsymbol{e}_p\right)$表示在状态$\boldsymbol{x}_t$和提示$\boldsymbol{e}_p$条件下的预测噪声，$\boldsymbol{e}_\phi$表示空提示，$\circ$表示逐元素相乘。




### 2.3 时序模型中的模型记忆定义

近期研究开始系统地考察视频扩散模型中的记忆问题，强调训练数据复制的风险不仅限于2D图像，在时空环境下的视频数据则更为复杂。在时序模型中，隐私数据的泄露途径发生了升维：模型可能并未完全复制原视频的静态物理表象(如人物外貌)，但却精准复刻了原视频中独有的连续动作轨迹或运动动力学特征。Chen等人(2024b)的研究认为，视频扩散模型的模型记忆应分解为内容记忆和运动记忆，并采用适用于无条件、文本条件和图像条件生成场景的隐私保护定义，从而解决早期视频扩散模型分析中混淆因素或局限于狭窄场景的局限性。

为了实现这种分离，作者将图像域的复制检测方法应用在时间维度，并提出了一种广义的基于帧的内容度量方法(generalized SSCD，GSSCD)。结果表明，与以往评估内容记忆的方法相比，GSSCD与人工标注的一致性更好。

对于运动记忆，他们引入了一种直接相似性度量，即光流相似性(optical flow score-k，OFS-k)，其定义为生成视频和训练视频的连续帧对之间具有较高的光流相似性，并进一步应用自然运动滤波(natural motion filtering，NMF)来排除不太可能构成隐私风险的相机平移和近乎静态的帧，从而提高运动记忆检测的可靠性。

## 3 记忆现象的理解

在明确扩散模型记忆现象的多维度定义后，进一步探究其底层驱动机制是实现准确量化与有效模型遗忘的理论前提。与传统模型不同，扩散模型在拟合高维复杂数据分布时，其记忆行为的产生是外部实证因素与内部数学机理共同作用的结果。一方面，从实证角度分析，模型参数规模的扩张(过参数化)与训练数据分布的特性(如长尾分布与高重复率)直接影响了模型对特定样本的记忆倾向。另一方面，从底层机理角度探讨，扩散去噪过程在数据流形上的拓扑特征与网络的局部几何收缩机制，为记

忆的发生提供了内在的数学解释。基于此逻辑,本节将依次从影响记忆的模型与数据要素(3.1节),以及解释记忆现象的流形理论视角(3.2节)展开系统性论述,如图4所示。

图4 模型记忆现象的因素和理论理解

Fig. 4 Factors and theoretical understanding of memorization

### 3.1 影响模型记忆的因素

#### 3.1.1 模型因素

过度参数化的模型是影响模型记忆的一大因素,研究者们聚焦于不同参数数量与训练数据集规模比例下的过参数化神经网络,以捕捉记忆效应与模型容量之间的关系。Nakkiran等人(2021)的研究表明,当参数数量超过训练数据规模时,过拟合现象对低泛化误差的模型影响较小,因此模型拥有足够的空间在存储单个训练样本。相反,参数量较小的模型虽然可能避免记忆效应,但面对大量数据时也可能难以有效泛化。Zhang等人(2021)的研究指出,深度神经网络具有记忆整个数据集随机标签的能力,这表明模型的容量足以通过纯粹的记忆来拟合数据,但在真实数据上,模型倾向于优先学习简单的模式,而非强行记忆。早期的研究将模型记忆的范围限定在单个神经网络模块中,例如卷积自编码器中下采样算子的影响(Radhakrishnan等,2019)、全连接网络和卷积神经网络(Zhang等,2020)。Nakkiran等人(2021)的研究指出,随着模型复杂度的增加,测试误差先下降后上升,但在跨过插值阈值进入过参数化区域后,测试误差会再次下降。

模型架构则是影响模型记忆的另一大因素,Maini等人(2023)的研究通过留一法,提出了模型记忆可以在模型架构中被定位的结论。Transformer及其衍生架构在处理记忆任务时表现出近乎最优的参数效率。Kajitsuka和Sato(2025)的最新理论研究严格证明,在序列生成任务中,Transformer仅需极少的参数量(理论下界为$\tilde{O}\left(\sqrt{N}\right)$)即可完美记忆$N$个独立样本的映射关系。在这种极致的记忆效率背后,注意力机制凭借其架构特性,扮演了极度高效的上下文识别与寻址路由角色,而实际的记忆内容绑定与容量瓶颈则落在后续的前馈网络或特定的值投影矩阵上。在条件扩散模型的架构中,文本提示词等条件信号被映射为键和值矩阵,而携带噪声的图像潜变量则作为查询。当扩散模型为了降低极端的局部重构误差而对特定训练样本发生记忆时,其数学本质表现为:面对特定敏感数据的查询,注意力权重计算$\mathrm{softmax}\left(\boldsymbol{QK}^{\top}/\sqrt{d_k}\right)$(式中$\boldsymbol{Q}$和$\boldsymbol{K}$分别表示注意力矩阵,$d_k$代表矩阵维度)在极少数对应的Key标记上发生了概率坍缩(即激活分布退化为趋近于独热向量)。这种注意力坍缩使得去噪网络放弃了生成分布中应有的特征泛化与平滑插值,转而直接通过该特定的键精确提取出被深层网络过拟合的高保真图像特征。基于这种理论推演,后续的研究则着重于定位模型记忆在模型架构中的主要发生位置。Ren等人(2024)的研究结果表明,在以U-Net为主干网络的扩散模型中,交叉注意力模块,尤其是较深的交叉注意力模块对扩散模型的模型记忆起主要作用。

#### 3.1.2 数据因素

数据分布以及单个数据在该分布下的相对位置会影响到模型对于该样本的记忆。Feldman(2020a)的研究提出了样本影响得分的概念,即一条数据是否被模型记忆被定义为在有无该训练样本的情况下,模型成功预测样本标签的概率之差;实验研究表明当对稀有和非典型样本(统计上的异常值)进行预测时,分类模型的准确率很低,而这些样本在任何长尾分布数据集中都普遍存在。Feldman和Zhang的同期研究(2020b)提供了一种样本影响得分的快速估计方法,可以有效地选择疑似被记忆的样本。由于非典型样本是一个特定分布上的相对位置的概念,Cailini等人(2022b)的研究提出了隐私洋葱效应来描述这样一种现象:如果移除分布中最外层的所有非典型样本之后,原本不被记忆的倒数第二层的非典型样本则更容易被模型记忆,进而导致泄露的风险。这表明,移除非典型样本数据等简单的数据移除方法实际上会增加分布中其他潜在敏感数据的隐私泄露风险。

训练集中高度重复或近乎重复的图像的存在会显著增加被记忆和被攻击成功提取的概率。Lee等人(2022)和Kandpal等人(2022)的研究指出被记忆的样本的重现概率与重复次数之间存在正相关。Huang等人(2024)的研究提出了一个在受控环境下研究逐字记忆的框架,得到相当数量的重复是逐字记忆发生的必要条件的结论。训练样本重复的现象

在各种扩散模型的数据集中得到验证，而去重已成为避免重复的常用技术)(Batifol等，2025)。

### 3.2 模型记忆的理论视角

为了更加深入理解扩散模型的模型记忆现象，不同于如上经由大量实验得到的经验结论，研究者们开始探讨如何在理论上建模扩散模型的记忆过程，数据分布固有的复杂性和几何结构也会影响记忆。这方面的研究主要基于图像训练数据的流形假设(Brown等，2023)，根据定义，流形假设认为在真实世界的高维图像数据$\boldsymbol{x}\in\boldsymbol{R}^D$是嵌入在$\boldsymbol{R}^D$中的一个低维流形$\boldsymbol{M}$的点，且$\boldsymbol{M}$的内在维度(local intrinsic dimension，LID)$d$满足$d\ll D$，在扩散模型的模型记忆研究中，其内在维度衡量了数据点的复杂性。Ross等人(2025)提出了流形记忆假设，该框架从几何角度出发，对于任意数据点$\boldsymbol{x}$，定义了真实数据流形$\boldsymbol{M}_*$和模型学习流形$\boldsymbol{M}_\theta$，即：真实数据分布$p_*(\boldsymbol{x})$所在的流形和生成模型$p_\theta(\boldsymbol{x})$实际学习到的流形，并在此基础上真实数据在该点附近的内在自由度$LID_*(\boldsymbol{x})$和模型在该点附近学到的内在自由度$LID_\theta(\boldsymbol{x})$；论文认为，当模型学到的流形维度过低($LID_\theta(\boldsymbol{x})$很小)时，会触发扩散模型的模型记忆。其中，当$LID_\theta(\boldsymbol{x})<LID_*(\boldsymbol{x})$时，论文认为是由过拟合驱动的模型记忆；而当$LID_\theta(\boldsymbol{x})\approx LID_*(\boldsymbol{x})$，且$LID_*(\boldsymbol{x})$较小时，论文认为是有数据驱动的记忆。利用针对$LID$的高效估计算法(Kamkari等，2024；Stanczuk等，2024)可以衡量扩散模型是否对单一样本点记忆。根据上述的理论框架，解释了分类器无关引导(classifier-free guidance，CFG)所得到的向量的范数越大，通常意味着分数函数的模长$\|s_\theta\|$越大。根据扩散模型的几何性质，分数模长越大，意味着该点距离流形越远，或者流形在该处极其陡峭，因而强CFG会降低$LID_\theta$，从而迫使模型输出记忆样本。

Kadkhodaie等人(2024)同样从几何角度提出了不同的视角：论文通过实验发现，当训练数据量$N$较小时(例如$N=10$或$100$)，模型处于高方差的状态。此时模型不仅记住了训练集，而且如果用两个不重叠的数据集分别训练两个模型，它们学到的得分函数是完全不同的；同时该论文通过分析扩散模型去噪网络的雅可比矩阵，证明了去噪器实际上在做基变换和收缩的操作，即将图像投影到一组特定的基向量上，对应的系数变小的过程。然而这组基向量是几何自适应(geometry-adaptive harmonic bases，GAHBs)而非固定的，论文指出，扩散模型能够记忆和泛化的核心原因是扩散模型的归纳偏置符合GAHB的几何结构。

然而，现有针对扩散模型的模型记忆的理论推导大多建立在相对简单的假设(强假设，例如数据分布呈现正态分布等)或者小规模的模型(例如早期的DDPM等模型)之上，并且其实验使用了清晰度较低的图像数据集作为验证，显然这无法为目前在海量的高清晰度数据集下训练的大型扩散模型提供相同的结论；并且利用这些研究提出的度量无法给模型记忆的量化提供精确的参考。因此理论层面的白盒分析在更大的模型上的未知结论，构成了当前扩散模型安全隐私研究的一大矛盾：尽管有一些白盒分析方法，为了真实评估并量化扩散模型的隐私泄露风险，当前的评估范式必然且只能转向基于经验主义的实证方法——即通过设计各种启发式的统计代理指标，或者直接模拟模型审计方与恶意攻击方的视角，发起对抗性测试(如成员推断攻击与数据提取攻击)，通过观测模型的输出行为来反向推算记忆的下界。第四节将具体阐述这些方法。

## 4 模型记忆能力的量化

第三节从模型、数据和几何三个方面分析了记忆是过度拟合数据流形的表现，那么在实际场景中，模型审计方和攻击方可以利用此特性来捕获记忆样本。鉴于对于扩散模型模型记忆的定义多样，并且通过定义来计算模型记忆的计算成本很高，通常需要重新训练模型。因此，本章将分成模型审计方和恶意攻击方两个角度介绍模型记忆的常见量化方法。其核心思想是针对模型记忆分数的定义提出近似计算方法作为代理，或者通过隐私攻击计算模型泄露的占比，进而提供模型记忆分数的下界。在本章节介绍这些量化方法，表1列出了这些量化方法的思路差异，并且在计算消耗维度进行比较。

### 4.1 模型审计方的量化方法

#### 4.1.1 通过代理变量的量化方法

在第二部分介绍了在不同学习场景下扩散模型记忆的几种定义。根据模型记忆定义计算记忆得分的方法基于留一法，即比较扩散模型在特定训练数据存在和不存在两种情况下的性能。这会导致需要

重新训练影子模型(shadow model),而当模型规模较大时,针对每一个训练样本重新训练影子模型是不可行的。对应的改进方法是使用部分训练数据而非全部训练数据进行近似计算。例如,Feldman和Zhang(2020b)利用在保留数据上通过随机选择独立训练的影子模型的聚合估计器来计算一个训练样本的影响得分(influence score),该方法已被证明具有良好的一致性和快速的收敛速度。Jiang等人(2021)在该研究的基础上,提出了一种被称为一致性得分(consistency score, C-score)的指标,对于一个样本$x, y$,其在训练集大小为$n$时的一致性得分$C_{P,n}(\boldsymbol{x},\boldsymbol{y})$定义为:在从数据分布$P$中采样大小为$n$的训练集$\boldsymbol{D}$(不包含该样本)上训练模型$f$,该模型能正确预测$\boldsymbol{y}$的期望概率:

$$C_{P,n}(\boldsymbol{x},\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{D}\sim P^n}[\boldsymbol{P}(f(\boldsymbol{x};\boldsymbol{D}) = \boldsymbol{y})]. \#(25)$$

为了得到一个单一的标量指标,作者对所有可能的训练集大小$n$取期望,得到C-score:

$$C(\boldsymbol{x},\boldsymbol{y}) = \mathbb{E}_n[C_{P,n}(\boldsymbol{x},\boldsymbol{y})]. \#(26)$$

上述的方法在扩散模型中可以将标签$\boldsymbol{y}$理解成对应的文本,而正确预测$\boldsymbol{y}$则表示预测文本和实际文本对应的嵌入很接近。Agarwal等人(2022)提出了在训练过程中跟踪梯度方差,而不仅仅是考虑最终更新的参数,解决了影响函数方法需要对Hessian矩阵进行近似会导致较高的误差的问题。Van等人(2021)提出了一种K折交叉验证和平均方法来估计生成模型的记忆得分。

如3.2节所述,近期研究已证实记忆现象与损失函数几何形状之间存在理论联系,而损失函数几何形状可作为评估工具。其核心思想是将扩散模型的模型记忆与训练中一些变量的几何特征联系起来。

目前对于模型记忆的几何代理有如下三类:一种是分析损失函数的曲率,即Hessian矩阵,高曲率样本通常对应于长尾、错误标记或冲突的实例,这些实例更容易被记忆。Ravi等人(2024)研究了将模型记忆与输入损失曲率和差分隐私联系起来的理论理解,其中记忆得分可以通过损失曲率来界定。Kim等人的研究(2023)指出,锐度感知最小化(sharpness aware minimization, SAM)在非典型数据上实现泛化的方式甚至可能带来更高的数据泄露。Garg等人(2024)的研究则将记忆样本识别为损失函数曲面中的尖锐极小值,并提出了一种基于损失函数Hessian矩阵迹的替代方法:

$$\tau(\boldsymbol{x}_i) = \mathrm{tr}\left(\nabla_\theta^2 L\left(\boldsymbol{x}_i,\hat{\theta}\right)\right), \#(27)$$

式中,$tr(\cdot)$表示矩阵的迹。

另一类几何代理则考虑扩散模型的概率密度曲面。Jeon等人(2025)用扩散模型对于单个样本的对数概率密度Hessian矩阵对应的大负特征值来表征概率曲面的尖锐性,这些特征值可以有效地检测和量化记忆。最后一类则是流形分析,如3.2节所述。Ross等人提出的框架(2025)利用了流形假设,通过比较真实数据流形和模型学习到的流形的维度来分析记忆。

**表1 模型记忆各类量化方法**

**Table 1 Comparison of Different Memorization Mitigation Approaches**

| 量化方法 | 核心思路 | 评估视角与方法类别 | 适用阶段 | 计算消耗 |
|---|---|---|---|---|
| 影响函数法(Feldman等,2020a, 2020b) | 利用独立训练的影子模型的聚合估计量作为影响得分。 | 审计方、黑盒 | 预训练 | 大 |
| 梯度方差(Agarwal等,2022) | 跟踪梯度方差以捕捉训练期间的记忆行为。 | 审计方、白盒 | 预训练 | 中等 |
| 几何代理变量(Ravi等,2024) | 利用模型记忆和训练中的一些变量的几何特征相联系 | 审计方、白盒 | 预训练/后训练 | 中等 |
| 图像复制指标(Zhang等,2018) | 像素层面/结构层面/语义层面相似性捕获 | 审计方、黑/白盒 | 推理阶段 | 中等 |
| 隐私攻击法(Hu等,2022) | 利用隐私攻击诱使模型输出隐私数据 | 攻击方、黑/白盒 | 推理阶段 | 大 |

4.1.2 基于复制指标的量化方法

除了对扩散模型的模型记忆的定义的代理指标进行量化以外,图像的复制指标也能为估计扩散模型的记忆能力提供下界。这一研究方向始于Xu等

人(2018)和Meehan等人(2020)的研究,该研究旨在解决双样本非参数假设检验问题。Bhattacharjee等人(2023)提出了一个统一的数据复制定义框架,解决了Meehan等人(2020)的研究中数据复制定义存在的一些缺陷。检测模型输出相似性的研究可以分为,像素级相似性指标、感知级相似性指标和语义级相似性指标。

当目标图像被扩散模型近乎完美地复制,即具有全局记忆时,基于结构的度量方法来捕捉像素级相似性,例如Wang等人(2004)提出的结构相似性指标(structural similarity index, SSIM)通过比较图像的平均灰度值、标准差和协方差判断两幅图像质检的结构性相似度。然而,由于扩散模型具有风格迁移等局部模型记忆,在认为设置的判断阈值下,这些生成的图像可能逃过结构化像素度量的检测,因此,基于学习的度量方法成为解决这一问题的另一研究方向。其主要流程是训练自监督模型并获得高质量的潜表示,以此比较两幅图像之间的潜在表示的相似性。这类基于学习的度量方法则可以按照模态分成如下两类:一是图像模态上的感知级相似性指标,Zhang等人(2018)提出的图像感知相似性指标(learned perceptual image patch similarity, LPIPS)指标将两张比较图像放进同一个预训练的神经网络中,例如VGG-Net(visual geometry group network),ResNet(residual network)等,提取网络中间层的特征图并计算这些特征图之间的距离;Pizzi等人(2022)提出的SSCD(self-supervised copy detection)度量则利用对比学习方法和特定的正则化项来学习不同增强方式之间的相似性;最近Wang等人(2024)提出的ICDiff(image copy detection for diffusion models)利用最先进的扩散模型和更细粒度的嵌入方法构建图像复制对,从而实现更灵敏的复制检测。另一类则是语言模态的语义级相似性指标,通过将比较图像经过多模态理解模型转换成对应文本嵌入之间的相似度比较。Radford等人(2021)的指标被广泛用于评估语义级别的图像相似性。与关注像素结构或局部纹理的指标不同,CLIP(contrastive language-image pretraining)利用在大规模图文对上预训练的图像编码器,将图像映射到一个语义的潜在空间。计算生成图像与训练数据在CLIP空间中的余弦相似度,能够有效捕捉模型在概念层级上的记忆行为,即使生成图像在像素和风格上与原图存在显著差异,CLIP仍能识别出深层的语义重合。

综合上述多层级的图像复制指标,可以看出当前扩散模型的模型记忆度量面临着表征解耦带来的度量困境。扩散模型的高维连续潜空间允许特征在不同语义层级进行重组。单一的像素级指标(如SSIM)或结构级指标(如SSCD)极易被模型生成的微小扰动或随机噪声掩盖,从而产生高昂的假阴性。而语义级指标(如CLIP)虽然能跨越像素差异捕获概念级记忆,却又难以区分模型是泛化了某一类风格还是精准记忆了某一张图片。这种度量维度的内在割裂表明,在缺乏底层统一物理定义的情况下,仅依赖生成输出的黑盒相似度比对,无法为扩散模型的隐私泄漏提供具备严格数学保证的上界。

### 4.2 恶意攻击方的量化方法

量化模型记忆能力的另一种方法是采用对抗性方式进行测试,即隐私攻击,该方法能实证性地提供记忆能力的下界,本小节将介绍主流隐私攻击技术并探讨其局限性。

#### 4.2.1 成员推断攻击

成员推断攻击(membership inference attack, MIA)(Hu等,2022)旨在确定特定数据样本是否属于模型的训练数据集。早期的研究利用传统模型的过拟合现象来设计成员推断攻击,即利用模型在过拟合数据上的输出显示出高置信度或者明显的低损失作为判断标准。攻击者通常依赖上述方法来区分训练数据和非训练数据。此外,后续研究还探索了更复杂的算法设计方法,例如Shokri等人提出的影子建模,通过训练代理模型来模仿目标模型的行为(Shokri等,2017; Carlini等,2022a),和Jia等人(2019)提出的对抗扰动方法以放大成员信号。同时,考虑到单次攻击的准确率和成功率无法真正反映攻击算法的性能和目标模型的脆弱性,Carlini等人(2022a)也探讨了检测指标的改进,选择在1%的假阳性下的真阳性水平作为成员的判断标准。

成员推理攻击的本质在于寻找模型在记忆样本和非记忆样本之间行为的差异。因此,设计合适的检测指标至关重要。从AI安全角度来看,可将检测指标分为如下三种设置:白盒(Pang等,2025)、灰盒(Dubinski等,2024)和黑盒(Li等,2025a;Matsumoto等,2023),分别代表攻击方能够获得整个模型、模型的中间结果以及仅模型输出的信息。

黑盒设置的隐私攻击通常利用4.1.2节中所述

的复制指标，通过预定义的阈值来判断样本是否被记忆。Pang等人（2023）的研究提出了一种黑盒成员攻击扩散模型的流水线，利用已获得或者经过视觉语言模型生成的图像文本标注作为提示输入扩散模型，并且生成多张图片，并通过预训练的图像编码器得到相应的嵌入，并以此计算原图嵌入和生成图像的嵌入之间的余弦相似度，在给定阈值的条件下判断该训练图像是否被记忆。

基于灰盒知识的威胁模型假设攻击者可以完全访问中间结果。在扩散模型中，这一假设指的是攻击者可以访问每个推理步骤的中间结果，例如DDPM建模下所预测的噪声或者流匹配建模下所预测的向量场。Duan等人（2023）提出了一种基于查询的隐私攻击SecMI（step-wise error comparing membership inference），该攻击利用扩散前向过程不同时间步的后验估计误差。具体而言，SecMI计算时间步$t$的近似后验估计误差，公式如下：

$$\tilde{l}_{t,\boldsymbol{x}_0}=\left\|\psi_\theta\left(\phi_\theta\left(\tilde{\boldsymbol{x}}_t,t\right),t\right)-\tilde{\boldsymbol{x}}_t\right\|^2,\#(28)$$

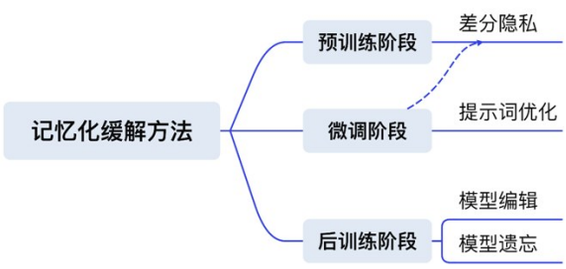
式中，$\tilde{\boldsymbol{x}}_t=\boldsymbol{\Phi}_\theta\left(\boldsymbol{x}_0,t\right)$是确定性逆向结果，$\psi_\theta$和$\phi_\theta$分别表示确定性逆向和去噪过程。研究表明，扩散模型中的文本条件会导致模型过拟合一些熟悉的生成模式，包括记忆样本。通过这一观察，Zhai等人（2024）进一步研究了条件扩散模型中的文本差异，并以此设计相应的灰盒攻击。

白盒攻击则利用模型参数的额外信息设计攻击方法。白盒攻击中的一大常见方法是是在训练过程中计算特定样本对应的梯度，Pang等人（2025）的研究表明这些梯度在范数下被认为是异常值或相对较大，进而可以作为成员推断攻击的判断标准。尽管梯度等一阶信息暗示了分布中的相对位置，但零阶信息（即模型参数）也可能在推理阶段造成数据泄露。因此另一大常见的方法和知识编辑中的神经元定位技术相关，即根据模型参数的分布来判断一个训练样本是否被记忆。例如，Hintersdorf等人（2024）提出了一种基于简单激活分布外检测的神经元级隐私检测方法，该方法通过定位负责记忆训练数据点的神经元来实现。他们通过因果干预来选择和优化U-Net中每个交叉注意力层关键矩阵中的责任神经元群，并通过实验验证了其有效性。另一项同期工作（Chavhan等，2024）则基于模型特定参数用于记忆训练样本的假设，采用Wanda（pruning by weights and activations）剪枝的方法（Sun等，2024）计算记忆神经元，研究了扩散模型建模中的前馈神经网络的神经元。

然而，成员推断攻击的有效性面临着挑战。尽管之前大量的研究从经验和理论层面提出了不同的MIA策略来检测基础模型的训练数据，但这些攻击对视觉生成模型和大型语言模型的有效性仍然存在疑问。如Zhang等人（2024b），Das等人（2025）的研究中所述，成员推理攻击无法提供生产机器学习模型训练中数据使用的可靠证据，因为由于无法满足数据集分布相同的要求，因此无法限制攻击的假阳性率。

#### 4.2.2 提取攻击

提取攻击（Extraction Attack，EA）本质上是一种更加强力的成员推理攻击，这种攻击方法使用不同的记忆检测指标，例如k-可提取性（k-extractable）（Carlini等，2023b）等。这类隐私攻击旨在直接从模型的训练数据中恢复原始样本或敏感信息。Carlini等人（2023b）的研究利用实验观察发现，将被记忆的图像文本标注作为提示时，会触发与常规提示输入不一样的生成过程，即生成过程更加具有确定性，以此作为触发条件设计了如下的提取攻击流程：根据不同随机种子下生成的图像的相似性（使用$l_2$距离）来识别记忆的样本。实验结果表明这类提取攻击能够很好的恢复被扩散模型记忆的图像。




## 5 模型记忆缓解方法

正如上几个章节所述，由于大模型的记忆机制导致了可能的敏感数据的泄露；而且，第四章阐述了现有的成员推断攻击等对抗性攻击方法能够重现模型记忆的训练样本；因此，如何从预训练的大模型中消除特定敏感数据的影响已经成为大模型隐私安全

的核心议题。由于社区目前缺乏能够完全解耦和透视十亿级参数（如DiT、MMDiT架构）隐空间记忆机制的统一理论框架，现有的模型遗忘方法尚未触及记忆消除的本质，而更多表现为一种启发式的工程防御。从粗粒度的全局噪声注入（差分隐私），到浅层的输入端规避（提示词优化），再到基于经验梯度的局部权重掩盖（概念消除与模型编辑），这些方法试图在效用-隐私权衡中寻找平衡。然而，正是由于底层理论的缺失，这些“工程补丁”往往只实现了表层的“概念隐藏”而非真正的“记忆擦除”。这直接导致了现有遗忘算法在面对精心构造的对抗性攻击时表现出极大的脆弱性。而为了应对诸如“被遗忘权”（Voigt等，2017）等隐私数据保护需求，以及训练数据当中通常只有一部分是敏感信息的情况，再加上测试时训练的概念的兴起，允许用户完全删除其指定数据的个性化需求产生了新的隐私保护范式——模型遗忘。形式上，可将模型遗忘问题定义如下：给定一个在数据集$\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$上预训练的模型$A$，模型遗忘问题的目标是设计一个高效的模型遗忘算法$U$，使其能够有效地处理数据集$\boldsymbol{X}$的一个子集$\boldsymbol{U}=\{\boldsymbol{u}_1,\cdots,\boldsymbol{u}_m\}\subset\boldsymbol{X}$，其中$m$表示模型遗忘算法的性能与在$X\setminus U$上预训练的扩散模型几乎相同，并且能够抵御各种隐私攻击。

图5 模型记忆缓解方法分类

Fig. 5 Categories of memorization mitigation

如上的定义给出了目前模型遗忘的两种主流方法：一种是单纯使用已删除敏感信息的数据集从头开始训练或者微调大模型，显然，这种方法不管是在预训练场景还是流式数据场景下都会造成极大的计算消耗；因此更加常用的方法是利用梯度上升等方法来达到近似遗忘而不是精确遗忘。从更高的观点来说，这些记忆化缓解的方法本质上都是属于模型遗忘（machine unlearning，MU）这个框架。如图5所示，根据模型遗忘方法在干预的阶段和粒度的不同，介绍从预训练、微调和后训练以及推理阶段的常见模型记忆缓解方法方法：差分隐私、提示词优化、模型遗忘，并且给出了常见方法在扩散模型上的性能比较，如表2所示。

### 5.1 差分隐私

差分隐私（differential privacy，DP）（Dwork等，2006a）是模型遗忘在预训练过程中的常见方法。差分隐私是衡量模型隐私泄露的一大标准，直观地说，其在模型训练当中是一种扰动策略，使得在略微不同的输入下（例如输入在Hamming距离下为1），模型能够以高概率输出相同的答案，而不会因为简单的差分导致敏感数据的泄露。在模型中应用差分隐私通常可以分为三种主要的方法：输入扰动（Kasiviswanathan等，2011；Warner等，1965）是最为直接的加噪方式，核心处理方法为针对原始数据集中的每一个样本或者特征添加随机噪声（通常该噪声服从Laplace分布或者Gaussian分布），这种做法在本地差分隐私（Arachchige等，2020）场景中尤为常见，因为需要考虑是否信任中心节点和传输节点。而另一种加噪的方式是输出扰动，即模型在干净的敏感数据上进行训练，在生成最终结果之后，计算该算法的全局敏感度$f$，并根据该敏感

度和隐私预算$\epsilon$，并向输出结果添加校准之后的噪声（Chaudhuri等，2011；Dwork等，2006b）。然而，这两类扰动方法具有同样的模型效用大幅下降的问题。而在深度学习当中，更为常用的是将差分隐私应用在训练算法上，其本质是针对训练时计算得到的梯度进行扰动，例如，Abadi等人提出的DP-SGD方法（Abadi等，2016）则是通过每一轮训练中针对单样本梯度进行梯度裁剪、梯度求和以及注入噪声（利用moments accountant方法计算总的隐私消耗），得到经过处理的梯度并做参数更新；该算法以及后续的改进都能够适用于基于梯度下降训练的神经网络模型，而且能够有效抵御当时发布的成员推断攻击和数据提取攻击，也成为了纪念城隐私模型训练的基准方法。在扩散模型的场景下，Liu等人（Liu等，2024）提出了在预训练阶段利用公共数据训练自编码器，而在微调阶段冻结自编码器，针对不同下游应用的私有数据应用DP-SGD（differential private stochastic gradient descent）算法。该方法可大幅减少训练参数的数量减少。实验证明，只微调注意力模块比微调整个模型或ResNet块效果更好，能更有效地捕捉数据分布的迁移，同时减少DP-SGD引入的噪声干扰，表2中第一栏给出了该算法在隐私预算$\epsilon=1$和$\epsilon=10$时stable diffusion v1.4在CelebA数据集上的性能表现，可以看出当隐私预算变小时，对应的图像生成质量和多样性（Fréchet inception distance，FID）大幅下降。

然而，差分隐私是针对整个训练数据集的统计隐私，而非个体隐私，是参数扰动的粗略版本，因此

表 2　各类模型记忆缓解方法量化结果对比

Table 2　Comparison of Memorization Mitigation Approaches by Defense Stage

| 阶段 | 方法 | 数据集 | 模型 | FID | SSCD | CLIP |
|---|---|---|---|---|---|---|
| 预训练阶段 | | | | | | |
| | DP-LDM ($\epsilon$ = 1)(Liu等, 2024) | CelebA | SD v1.4 | 21.1 ± 0.2 | – | – |
| | DP-LDM ($\epsilon$ = 10)(Liu等, 2024) | CelebA | SD v1.4 | 14.3 ± 0.1 | – | – |
| 推理阶段 | | | | | | |
| | 提示词优化(Somepalli等, 2023a,2023b) | LAION | SD v1.4/2.0 | – | 0.3 ± 0.1 | 0.28 ± 0.03 |
| | 注意力缩放 (Ren等,2024) | LAION | SD v1.4/2.0 | 15.0 ± 0.2 | 0.30 | **0.22** |
| | Token 扰动 (Wen等,2024) | LAION | SD v2.1 | 18.0 ± 1.0 | 0.5 ± 0.1 | – |
| 微 调/后 训练阶段 | | | | | | |
| | 神经元检测(Hintersdor等, 2025) | LAION | SD v1.4 | 16.5 ± 0.5 | **0.1**　**0.07** | 0.31 ± 0.02 |
| | ESD(Gandikota等,2023) | LAION | SD v1.4 | 13.68 | – | – |
| | 概念消除(Kumari等,2023) | – | SD v1.4 | 16.99 ± 0.2 | 0.30 | 0.60 ± 0.02 |
| | SalUn(Fan等,2024) | Imagenette | SD v1.4 | **1.22** | – | – |
| | SISS(Alberti等,2025) | CelebA-HQ | SD v1.4 | 20.1 | 0.32 | – |

注:SD 为 stable diffusion 的缩写;“–”表示原论文未提供该数据或不适用;黑色字体表示最优结果

在差分隐私约束下训练后,模型性能同样会显著下降。此外,由于DP-SGD 中的裁剪操作(差分隐私训练的核心),差分隐私训练会产生巨大的计算消耗,这对于大规模训练来说是难以承受的。因此,一系列研究工作致力于在不同的学习场景中开发更高效的审计算法,以及跳开差分隐私的框架做更加

适应大模型的隐私保护方法(Steinke等,2023;Nasr等,2023;Cebere等,2025;Panda等,2025)。差分隐私也可以用于推理阶段,如图5中的虚线所示,但由于后续模型记忆缓解方法的提出和差分隐私的计算消耗,扩散模型隐私领域逐渐少用这类方法。

### 5.2　提示词优化

提示词优化是模型遗忘在推断过程中的常见方法。Somepalli等人(2023a,2023b)的研究表明,被记忆的提示词中的某些词语或标记会对训练数据泄露产生显著影响,提示词优化则是在训练和推理阶段针对模型的输入进行偏移,使得模型无法将经过修改的输入和被记忆的训练数据所对应,使得生成的图像和被记忆的训练图像有差别,达到了模型遗忘敏感数据的效果。算法希望通过扰动提示嵌入来防止敏感部分的输出,同时保持与原始提示相似的语义。具体来说,给定一个包含N个标记的提示$p$的提示嵌入$\boldsymbol{e}$,定义其最小化问题的目标如下:

$$L(\boldsymbol{x}_t,\boldsymbol{e}) = \left\| \epsilon_\theta(\boldsymbol{x}_t,\boldsymbol{e}) - \epsilon_\theta(\boldsymbol{x}_t,\boldsymbol{e}_\varnothing) \right\|_2, \#(29)$$

式中,$\boldsymbol{e}$表示图像文本对样本中的文本的嵌入,$\boldsymbol{e}_\varnothing$则表示零嵌入。

Somepalli等人的研究(2023a,2023b)提出一种直接的方法,将这些触发词或标记与随机词元交替使用,但这种方法也导致了模型生成的结果相比起原生的生成结果要差很多。Wen等人(2024)尝试通过提示词优化来缓解记忆泄漏,具体来说,该方法寻找一个受扰动的嵌入向量$\boldsymbol{e}^*$,使得文本条件噪声预测的幅度最小化,并且为了防止优化后的嵌入向量偏离原始含义太远(导致生成的图和提示词无关),

作者引入了一个目标损失阈值$l_{\text{target}}$。一旦损失达到这个值，就停止优化，以此来平衡遗忘和语义。Chen等人(2024a)的后续工作进一步研究了局部记忆的概念，以更好地捕捉记忆泄漏，并利用该概念增强提示优化。表2第二栏给出了如上提及的提示词优化方法在FID，SSCD和CLIP三个指标下的表现，可以看出提示词优化在推理阶段的模型记忆缓解方法中达到了最佳性能。

尽管提示词优化在推理阶段展现出了极低的计算开销与较好的初步防御效用，但本质上仅仅是阻断了特定文本条件与敏感数据之间的显式映射路径。然而，敏感数据的底层几何特征依然作为无条件视觉先验被隐式地编码在去噪网络(如U-Net的残差块或DiT的前馈网络)的冗余参数中。这种隐蔽式安全防御机制极易被恶意攻击者通过构造对抗性提示词，或在连续潜空间中寻找替代触发路径所绕过。因此，单纯在输入端进行的访问控制，无法替代对模型参数空间的深层干预。

### 5.3 模型遗忘

模型遗忘(Nguyen等，2025；Cao等，2015)是针对已训练模型进行特定概念消除的后处理方法。传统的防御思路(如差分隐私)虽然能在预训练阶段提供统计学保证，但极大地破坏了生成效用。因此，研究重点必然转向干预成本更低、针对性更强的后训练阶段——即模型遗忘。模型遗忘旨在从预训练的模型权重中精准擦除特定数据或概念的影响，确保存储的模型参数不再包含非预期的知识(如版权图像等)，且无需从头重新训练模型。在扩散模型的场景下，现有的模型遗忘方法主要可以归纳为基于重定向的微调、基于显著性的参数定位以及基于闭式解的快速编辑三类策略。

基于重定向的微调是最早被应用于扩散模型遗忘的主流策略。该类方法的核心思想是通过微调模型参数，将目标概念的生成路径重定向至无关的空概念或随机输出。例如，Gandikota等人(2023)提出的ESD(erased stable diffusion)方法通过微调U-Net中的交叉注意力模块，利用负向引导将目标概念的预测噪声推向空文本对应的分布，从而在无需重新训练的情况下移除特定艺术风格或物体。Zhang等人(2024a)提出的forget-me-not和Kumari等人(2023)提出的概念消除法也采用了类似的微调逻辑，通过惩罚与目标概念相关的生成行为来抑制敏感内容的输出。Heng等人(2023)的研究从持续学习的观点，将模型遗忘的任务放入灾难性遗忘的框架当中，具体来说，作者给出了一个通用的贝叶斯优化框架，首先对于提示词中敏感的概念进行优化生成一个替代概念；其次加入了利用Fisher信息的正则项并且模型利用本身生成的伪数据进行记忆的选择性保留。

然而，全局微调可能会对模型的通用生成能力造成破坏。为了解决这一问题，研究者提出了基于显著性的参数定位方法。模型编辑是针对模型参数而不针对输入或者输出修改。该策略认为模型中仅有少部分关键权重负责特定概念的生成。先前Maini等(2023)和Chen等人(2024c)关于记忆定位的研究表明模型记忆发生在模型架构中的特定位置，但并非局限于单个层，而是一种局限于模型各层中一小部分神经元的现象。上述的研究只考虑了监督学习的场景，进一步的研究探索了扩散模型中的记

忆是否可以定位，以及如何有效地定位这些神经元。

Fan等人(2024)提出的SalUn以及Wu等人(2024)提出的ScissorHands利用梯度信息计算权重显著性，精准定位与待遗忘概念高度相关的参数子集进行更新，从而在实现遗忘的同时最大程度保留模型在其他任务上的性能。此外，为了进一步提升遗忘的效率，Gandikota等人(2024)提出了UCE(unified concept editing)，这是一种基于闭式解的参数编辑方法，通过在闭式形式下直接编辑模型投影矩阵，实现了对扩散模型概念的快速修改，避免了昂贵的迭代优化过程。Hintersdorf等人(2024)的研究应用基于阈值的策略，利用SSIM得分识别扩散模型的交叉注意力组件所有value层中具有分布外激活的候选记忆神经元。该选择过程包括初筛和细筛，细筛过程会移除不负责记忆的神经元。实验表明，仅移除百级神经元即可显著降低记忆泄漏。另一篇同期发表的论文Chavhan等人(2024)的研究应用了一种Wanda剪枝(Sun等，2024)的高效剪枝方法，用于扩散模型前馈网络所有线性层中的记忆神经元检测。实验也证明了该方法对提取攻击的有效抵抗力，表明其具有良好的记忆保留能力。

与上述的概念遗忘不同，研究者们开始考虑样本层面上的遗忘以满足更加个性化删除特定数据的

实际需求。Alberti等人(2025)提出的SISS(subtracted importance sampled scores)方法,利用重要性采样将目标函数重构成全量数据训练的目标函数与遗忘集的目标函数之差,使得模型能够在不需要额外锚点样本的情况下,精准地遗忘特定的训练图片(例如特定的人脸或甚至只有一张的样本),同时保持模型的生成能力不退化。

表2第三栏列出了本节所述模型编辑与概念遗忘方法的性能指标。然而,由于各方法所采用的数据集不同,且缺乏在FID、SSCD和CLIP等指标上的统一评估,导致直接的横向性能比较不可行。关于建立更加规范化基准测试的必要性,将于未来与展望章节中进一步讨论。

尽管上述方法在非对抗环境下表现出了良好的遗忘效果,但近期的安全性研究(Zhang等,2024c;Chin等,2024;Pham等,2024)指出,现有的遗忘算法在面对对抗性提示攻击时仍表现出显著的脆弱性。攻击者可以通过在输入提示中添加微小的对抗性扰动(例如在词元嵌入空间进行优化),成功绕过遗忘机制,诱导"已遗忘"的模型重新生成原本已被消除的不安全内容。这表明现有的模型遗忘方法主要实现了防御性的概念隐藏,而非彻底的知识清除,如何在对抗环境下实现鲁棒的模型遗忘成为了当前该领域亟待解决的新挑战(Zhang等,2024d)。

## 6 未来与展望

随着扩散模型在视觉生成领域的飞速发展,当前针对模型记忆与隐私遗忘的研究虽然取得了一定进展,但现有的技术在数据规范、算法机理及应用范式上仍存在显著的局限性。从这些问题出发,可以展望该领域未来的发展趋势:1)**隐私数据处理流水线和基准测试亟待规范**。高质量的数据治理与统一的评估标准是构建安全可信生成模型的基石。虽然研究者已经尝试构建了一些用于遗忘的数据集,但当前已有的数据无论在场景真实性还是评估维度的覆盖面上都有很大的发展空间。一方面,当前的数据治理缺乏全生命周期的视野,大多数防御仅停留在模型微调或者后训练阶段,在数据预处理方面仅限于一些预处理技巧而缺乏规范化的预处理流程,因此,有必要建立一套从敏感数据检测到标准化遗忘请求处理的完整流水线,实现大模型隐私安全的前置防御。另一方面,当前领域缺乏公认的测试基准,导致不同防御机制难以在同一维度下公平比较。现有的评估多依赖简单的数据集,难以反映真实攻击场景下的隐私泄露风险。因此,未来的研究需要构建涵盖多维度的统一评估基准,包括但不限于隐私保护率、图像生成质量和隐私防御成功率等,特别是构建模拟真实世界攻击(如成员推断攻击、模型反演)的高保真数据集。如何设计能够量化效用-隐私权衡的标准化指标,如何构建覆盖医疗等多场景的真实评测集,都是需要进一步研究的内容。2)**更符合扩散模型特性的模型记忆定义和模型遗忘算法亟待提出**。当前基于扩散模型的遗忘技术仍处于探索阶段,尚未形成统一的理论框架。主流的研究大多借鉴大语言模型的微调技术,将针对离散数据的编辑方法生搬硬套到连续的视觉流形上,忽略了扩散模型独特的去噪机制与时序特性。由于视觉生成任务与文本生成任务在记忆存储方式上存在本质差异,直接迁移的方法往往导致生成质量严重下降或记忆清除不彻底。因此,有必要研究符合扩散模型物理特性的原生算法:首先,需要重构视觉记忆的定义,建立从像素级别到语义级别的视觉模态的记忆量化理论;其次,未来的算法应更加深入扩散过程的内部机理,类似模型编辑方法,例如利用扩散模型马尔可夫链的时序建模特性和不同的建模方法,探索在特定时间步或分数函数层面进行干预。此外,模型架构的改进也是一个重要方向,当前的U-Net或DiT架构导致模型遗忘算法无法完全解耦遗忘的敏感信息,未来如何通过引入混合专家系统(mixtral of experts, MoE)或模型剪枝技术实现架构层面的记忆解耦,从而实现更高效的遗忘,是极具前景的研究课题。3)**新学习场景的模型记忆-遗忘机制与垂直领域落地**。当前的研究多聚焦于静态模型在通用场景下的隐私保护,而忽略了深度学习范式转变带来的新挑战以及高敏感领域的特殊需求。一方面,随着测试时训练(test-time training, TTT)(Behrouz等,2025a, 2025b)逐渐成为提升模型泛化能力的新范式,模型则处于动态更新的数据流中。现有的静态防御机制在面对在线学习时容易失效,这带来了动态隐私保护的新机遇:如何设计即时遗忘机制,在模型实时摄入新数据的过程中动态监测并清除有害记忆,是构建自适应安全系统的关键。另一方面,尽管通用模型的遗忘已有探索,但在金融、医疗影像等垂

直领域的落地仍面临巨大鸿沟。这些领域对隐私有着极高的合规要求，例如在医疗生成中，如何在严格消除患者生物特征的同时完美保留病理特征，是通用算法难以解决的难题。因此，针对垂直领域的专用遗忘技术，如何确保金融和医疗在合成数据和医学影像中隐私信息的擦除，将是未来推动隐私保护技术落地的核心方向。

## 7 结 论

本文综述了扩散模型中模型记忆-遗忘机制的研究进展。具体而言，本文从对扩散模型的理论建模和架构，模型记忆在时序和非时序扩散模型上的定义和理解，量化方法以及模型记忆缓解方法进行了全面的概述，并且在本文最后一章总结了当前扩散模型在记忆－遗忘机制方面所面临的主要开放性问题与挑战。

## 参考文献(References)

Abadi M, Chu A, Goodfellow I, McMahan H B, Mironov I, Talwar, et al. 2016. Deep learning with differential privacy//ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria: ACM: 308-318[DOI: 10.1145/2976749.2978318]

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, et al. 2023. GPT-4 technical report[EB/OL].[2024-3-4]. https://arxiv.org/pdf/2303.08774

Agarwal C, D'souza D and Hooker S. 2022. Estimating example difficulty using variance of gradients//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. New Orleans, LA, USA: IEEE: 10358 - 10368[DOI: 10.1109/CVPR52688.2022.01012]

Alberti S, Hasanaliyev K, Shah M and Ermon S. 2025. Data unlearning in diffusion models//The Thirteenth International Conference on Learning Representations. Singapore: OpenReview: 1 - 17

Arachchige P C M, Bertok P, Khalil I, Liu D, Camtepe S and Atiquzzaman M. 2020. Local differential privacy for deep learning. IEEE Internet of Things Journal, 7(7): 5827 - 5842[DOI: 10.1109/JIOT.2019.2952146]

Arpit D, Jastrzębski S, Ballas N, Krueger D, Bengio E, Kanwal M S, et al. 2017. A closer look at memorization in deep networks//International Conference on Machine Learning. Sydney, Australia: PMLR: 233 - 242[DOI: 10.5555/3305381.3305406]

Bartlett P L, Long P M, Lugosi G and Tsigler A.2020. Benign overfitting in linear regression. Proceedings of the National Academy of Sciences, 117 (48): 30063 - 30070 [DOI: 10.1073/pnas.1907378117]

Batifol S, Blattmann A, Boesel F, Consul S, Diagne C, Dockhorn T, et al. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space [EB/OL]. [2025-6-24] https://arxiv.org/pdf/2506.15742

Ali Behrouz, Meisam Razaviyayn, Peilin Zhong and Vahab Mirrokni. 2025a. It's All Connected: A Journey Through Test-Time Memorization, Attentional Bias, Retention, and Online Optimization [EB/OL]. [2025-4-17]. https://arxiv.org/pdf/2504.13173

Behrouz A, Zhong P and Mirrokni V. 2025b. Titans: Learning to memorize at test time//The Thirty-ninth Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS: 1-38

Bhattacharjee R, Dasgupta S and Chaudhuri K. 2023. Data-copying in generative models: a formal framework//International Conference on Machine Learning. Honolulu, USA: PMLR: 2364 - 2396 [DOI: 10.5555/3618408.3618509]

Brown B C, Caterini A L, Ross B L, Cresswell J C and Loaiza-Ganem G. 2023. Verifying the union of manifolds hypothesis for image data//International Conference on Learning Representations. Kigali, Rwanda: OpenReview: 1 - 24

Cao Y, Chen Z, Belkin M and Gu Q. 2022. Benign overfitting in two-layer convolutional neural networks//Advances in Neural Information Processing Systems, Curran Associates: 35: 25237 - 25250

Cao Y and Yang J. 2015. Towards making systems forget with machine unlearning//2015 IEEE Symposium on Security and Privacy. San Jose, USA: IEEE: 463 - 480[DOI: 10.1109/SP.2015.35]

Carlini N, Chien S, Nasr M, Song S, Terzis A and Tramer F. 2022a. Membership inference attacks from first principles//2022 IEEE Symposium on Security and Privacy, San Francisco, CA, USA: IEEE: 1897 - 1914[DOI: 10.1109/SP46214.2022.9833649]

Carlini N, Hayes J, Nasr M, Jagielski M, Sehwag V, Tramer F, et al. 2023a. Extracting training data from diffusion models//32nd USENIX Security Symposium. Anaheim, USA: USENIX Association: 5253 - 5270[DOI: 10.5555/3620237.3620531]

Carlini N, Ippolito D, Jagielski M, Lee K, Tramer F and Zhang C. 2023b. Quantifying memorization across neural language models// The Eleventh International Conference on Learning Representations, 2023, Kigali, Rwanda: OpenReview: 1 - 19

Cebere T I, Bellet A and Papernot N. 2025. Tighter privacy auditing of DP-SGD in the hidden state threat model//The Thirteenth International Conference on Learning Representations. Singapore: OpenReview: 1 - 23

Chaudhuri K, Monteleoni C and Sarwate A D. 2011. Differentially private empirical risk minimization. Journal of Machine Learning Research, 12(3): 1069-1109 [DOI: 10.5555/1953048.2021036]

Chavhan R, Bohdal O, Zong Y, Li D and Hospedales T. 2024. Memo-

rized images in diffusion models share a subspace that can be located and deleted[EB/OL].[2024-01-06]. https://arxiv.org/pdf/2406.18566

Chen C, Liu D, Shah M and Xu C. 2024a. Exploring local memorization in diffusion models via bright ending attention//The Thirteenth International Conference on Learning Representations 2025. Vienna, Austria: OpenReview: 1 - 15

Chen C, Liu E, Liu D, Shah M and Xu C. 2024b. Investigating memorization in video diffusion models[EB/OL].[2025-04-25]. https://arxiv.org/pdf/2410.21669

Chen R, Hu T, Feng Y and Liu Z. 2024c. Learnable privacy neurons localization in language models//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Bangkok, Thailand: Association for Computational Linguistics: 256 - 264 [DOI: 10.18653/v1/ 2024. acl-short.25]

Chin Z Y, Jiang C M, Huang C C, Chen P Y and Chiu W C. 2024. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts//Proceedings of 41st International Conference on Machine Learning. Vienna, Austria: PMLR: 8468 - 8486[DOI:10.5555/3692070.3692406]

Croitoru F A, Hondru V, Ionescu R T and Shah M. 2023. Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9): 10850 - 10869[DOI: 10.1109/TPAMI.2023.3261988]

Das D, Zhang J and Trantèr F. 2025. Blind baselines beat membership inference attacks for foundation models//IEEE Security and Privacy Workshops. San Francisco, CA, USA: IEEE: 118 - 125 [DOI: 10.1109/SPW67851. 2025.00016]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale//9th International Conference on Learning Representations. 2021. Vienna, Austria: OpenReview: 1 - 22

Duan J, Kong F, Wang S, Shi X and Xu K. 2023. Are diffusion models vulnerable to membership inference attacks? //International Conference on Machine Learning (ICLR). Honolulu, USA: PMLR: 8717 - 8730[DOI: 10.5555/3618408.3618757]

Dubiński J, Kowalczuk A, Pawlak S, Rokita P, Trzciński T and Morawiecki P. 2024. Towards more realistic membership inference attacks on large diffusion models//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA: IEEE: 4848 - 4857 [DOI: 10.1109/WACV57701. 2024.00479]

Dwork C. 2006a. Differential privacy//International Colloquium on Automata, Languages, and Programming. Berlin, Heidelberg: Springer: 1 - 12 [DOI: 10. 1007/11787006_1]

Dwork C, McSherry F, Nissim K and Smith A. 2006b. Calibrating noise to sensitivity in private data analysis//Theory of Cryptography Conference. Berlin, Heidelberg: Springer: 265 - 284[DOI: 10.1007/11681878_14]

Esser P, Kulal S, Blattmann A, Entezari R, Müller J, Saini H, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis//Proceedings of 41st International Conference on Machine Learning. Vienna, Austria: JMLR: 1 - 28 [DOI: 10.5555/3692070.3692573]

Fan C, Liu J, Zhang Y, Wong E, Wei D and Liu S. 2024. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation//The Twelfth International Conference on Learning Representations. Vienna, Austria: OpenReview: 1 - 31

Feldman V. 2020a. Does learning require memorization? a short tale about a long tail//Proceedings of 52nd Annual AUM SIGACT Symposium on Theory of Computing. Chicago, IL, USA: ACM: 954 - 959[DOI:10.1145/3357713.3384290]

Feldman V and Zhang C. 2020b. What neuralnetworks memorize and why: Discovering the long tail via influence estimation//Advances in Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates: 33: 2881-2891 [DOI: 10.5555/3495724.3495966]

Feng Z X, Lai J H, Yuan Z, Huang Y L, Lai P J.2025. Advancing universal person reidentification: a survey on the applications of large-scale pretraining models for identifying individuals. Journal of Image and Graphics, 30(6): 1638-1660 (冯展祥, 赖剑煌, 袁藏, 黄宇立, 赖培杰. 2025. 走向通用行人重识别:预训练大模型技术在行人重识别的应用综述. 中国图象图形学报, 30(6): 1638-1660) [DOI:10.11834/jig.240426]

Gandikota R, Materzynska J, Fiotto-Kaufman J and Bau D. 2023. Erasing concepts from diffusion models//Proceedings of the IEEE/CVF International Conference on Computer Vision . Paris, France: IEEE: 2426 - 2436[DOI: 10.1109/ICCV51070.2023.00230]

Gandikota R, Orgad H, Belinkov Y, Materzyńska J and Bau D. 2024. Unified concept editing in diffusion models//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 5111 - 5120 [DOI: 10.1109/WACV57701.2024.00503]

Garg I, Ravikumar D and Roy K. 2024. Memorization through the lens of curvature of loss function around samples//Proceedings of 41st International Conference on Machine Learning. Vienna, Austria: PMLR: 15083 - 15101[DOI: 10.5555/3692070.3692674]

Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. Nature, 645: 633-638[DOI: 10.1038/s41586-025-09422-z]

Hartmann V, Suri A, Bindschaedler V, Evans D Tople S and West R. 2023. Sok: Memorization ingeneral-purpose large language models [EB/OL].[2023-10-24]. https://arxiv.org/pdf/2310.18362

Hastie T, Tibshirani R and Friedman J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer: 1 - 764

Heng A and Soh H. 2023. Selective amnesia: Acontinual learning approach to forgetting in deep generative models//Advances in Neural Information Processing Systems, Curran Associates: 36: 17170 - 17194 [DOI: 10.5555/ 3666122.3666873]

Hintersdorf D, Struppek L, Kersting K, Dziedzic Aand Boenisch F. 2024. Finding nemo: Localizing neurons responsible for memorization in diffusion models//Advances in Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates: 37: 88236 - 88278[DOI: 10.5555/3737916.3740716]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Advances in Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates: 33: 6840 - 6851 [DOI: 10.5555/ 3495724.3496298]

Hu H, Salcic Z, Sun L, Dobbie G, Yu P S and Zhang X. 2022. Membership inference attacks on machine learning: A survey//ACM Computing Surveys, New York, NY, USA: Association of Computing Machinery: 54(11s): 1 - 37 [DOI: 10.1145/3523273]

Huang J, Yang D and Potts C. 2024. Demystifying verbatim memorization in large language models//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, Florida, USA: Association for Computational Linguistics: 10711 - 10732 [DOI:10.18653/v1/2024.emnlp-main.598]

Jeon D, Kim D and No A. 2025. Understanding and mitigating memorization in generative models via sharpness of probability landscapes// Proceedings of the 42nd International Conference on Machine Learning. Vancouver, BC, Canada: PMLR: 267: 27091-27112[DOI: 10.5555/3780338.3781404]

Jia J, Salem A, Backes M, Zhang Y and Gong N Z. 2019. Memguard: Defending against black-box membership inference attacks via adversarial examples//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. London, UK: ACM:259-274[DOI: 10.1145/3319535.3363201]

Jiang Z, Zhang C, Talwar K and Mozer M C. 2021. Characterizing structural regularities of labeled data in overparameterized models// Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR: 5034 - 5044

Kadkhodaie Z, Guth F, Simoncelli E P and Mallat S. 2024. Generalization in diffusion models arises from geometry-adaptive harmonic representations//The Twelfth International Conference on Learning Representations. Vienna, Austria: OpenReview: 1 - 25

Kajitsuka T and Sato I. 2025. On the Optimal Memorization Capacity of Transformers// The Thirteenth International Conference on Learning Representations. Singapore, OpenReview: 1 - 42

Kamkari H, Ross B L, Hosseinzadeh R, Cresswell J C and Loaiza-Ganem G. 2024. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models// Advances in Neural Information Processing Systems, Curran Associates: 37: 38307 - 38354

Kandpal N, Wallace E and Raffel C. 2022. Deduplicating training data mitigates privacy risks in language models//Proceedings of the 39th International Conference on Machine Learning (ICML). Baltimore, USA: PMLR: 10697 - 10707

Kaplan J, McCandlish S, Henighan T, Brown T B, Chess B, Child R, Gray S, Radford A, Wu J and Amodei D. 2020. Scaling laws for neural language models. [EB/OL].[2020-01-23]. https://arxiv.org/pdf/2001.08361

Kasiviswanathan S P, Lee H K, Nissim K, Raskhodnikova S and Smith A. 2011. What can we learn privately? SIAM Journal on Computing, 40(3): 793 - 826 [DOI: 10.1137/ 090756090]

Kim Y I, Agrawal P, Royset J O and Khanna R. 2023. On memorization and privacy risks of sharpness aware minimization. [EB/OL] [2026-1-28] https://arxiv.org/pdf/2310.00488

Kou Y, Chen Z, Chen Y and Gu Q. 2023. Benign overfitting in two-layer relu convolutional neural networks//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: 17615 - 17659[DOI: 10.5555/3618408.3619135]

Kumari N, Zhang B, Wang S Y, Shechtman E, Zhang R and Zhu J Y. 2023. Ablating concepts in text-to-image diffusion models//2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22634-22645 [DOI: 10.1109/ICCV51070.2023.02074]

Lee K, Ippolito D, Nystrom A, Zhang C, Eck D, Callison-Burch C and Carlini N. 2022. Deduplicating training data makes language models better//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics: 8424 - 8445 [DOI:10.18653/v1/ 2022.acl-long.577]

Li J, Dong J, He T and Zhang J. 2025a. Towards black-box membership inference attack for diffusion models//ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy. Singapore: OpenReview: 1 - 19

Li Q, Luo X, Chen Y and Bjerva J. 2025b. Trustworthy machine learning via memorization and the granular long-tail: A survey on interactions, tradeoffs, and beyond. [EB/OL]. [2025-03-10]. https://arxiv.org/pdf/2503.07501

Lipman Y, Chen R T Q, Ben-Hamu H, Nickel M and Le M. 2023. Flow matching for generative modeling//The Eleventh International Conference on Learning Representations. Kigali, Rwanda: OpenReview: 1 - 28

Liu M F, Lyu S, Vinaroz M and Park M. 2024. Differentially private latent diffusion models//Privacy Regulation and Protection in Machine Learning, OpenReview, 1 - 24

Liu X, Gong C and Liu Q. 2023. Flow straight and fast: Learning to generate and transfer data with rectified flow//The Eleventh Interna-

tional Conference on Learning Representations. Kigali, Rwanda: OpenReview: 1 - 33

Maini P, Mozer M C, Sedghi H, Lipton Z C, Kolter J Z and Zhang C. 2023. Can neural network memorization be localized? //International Conference on Machine Learning. Honolulu, Hawaii, USA: JMLR: 23536 - 23557[DOI: 10.5555/3618408.3619391]

Matsumoto T, Miura T and Yanai N. 2023. Membership inference attacks against diffusion models//2023 IEEE Security and Privacy Workshops. San Francisco, CA, USA: IEEE: 77 - 83[DOI: 10.1109/SPW59333.2023.00013]

Meehan C, Chaudhuri K and Dasgupta S. 2020. A Three Sample Hypothesis Test for Evaluating Generative Models//Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. Palermo, Italy: PMLR: 108: 3546--3556

Nakkiran P, Kaplun G, Bansal Y, Yang T, Barak B and Sutskever I. 2021. Deep double descent: Where bigger models and more data hurt. Journal of Statistical Mechanics: Theory and Experiment, 2021(12):124003[DOI: 10.1088/1742-5468/ac3a74]

Nasr M, Hayes J, Steinke T, Balle B, Tramèr F, Jagielski M, Carlini N and Terzis A. 2023. Tight auditing of differentially private machine learning//Proceedings of the 32nd USENIX Security Symposium. Anaheim, CA, USA: USENIX Association: 1631 - 1648[DOI: 10.5555/3620237.3620329]

Nguyen T T, Huynh T T, Ren Z, Nguyen P L, Liew A W C, Yin H and Nguyen Q V H. 2025. A survey of machine unlearning//ACM Transactions on Intelligent Systems and Technology, New York, NY, USA: Association of Computing Machinery: 16(5): 1 - 46[DOI: 10.1145/3749987]

Panda A, Tang X, Choquette-Choo C A, Nasr M and Mittal P. 2025. Privacy auditing of large language models//The Thirteenth International Conference on Learning Representations. Singapore: OpenReview: 1 - 17

Pang Y and Wang T. 2023. Black-box membershipinference attacks against fine-tuned diffusion models[EB/OL]. [2024-09-05]. https://arxiv.org/pdf/2312.08207

Pang Y, Wang T, Kang X, Huai M and Zhang Y. 2025. White-box membership inference attacks against diffusion models//Proceedings on Privacy Enhancing Technologies, 2: 398 - 415[DOI: 10.56553/popets-2025-0068]

Peebles W and Xie S. 2023. Scalable diffusion models with transformers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 4172 - 4182[DOI: 10.1109/ICCV51070.2023.00387]

Pham M, Marshall K O, Cohen N, Mittal G and Hegde C. 2024. Circumventing concept erasure methods for text-to-image generative models//The Twelfth International Conference on Learning Representations. Vienna, Austria: OpenReview: 1 - 26

Pizzi E, Roy S D, Ravindra S N, Goyal P and Douze M. 2022. A self-supervised descriptor for image copy detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE: 14512 - 14522[DOI: 10.1109/CVPR52688.2022.01413]

Radhakrishnan A, Belkin M and Uhler C. 2019. Memorization in overparameterized autoencoders//ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena. Long Beach, NY, USA: PMLR: 1 - 14

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark Jet al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR: 8748 - 8763

Ravikumar D, Soufleri E, Hashemi A and Roy K. 2024. Unveiling privacy, memorization, and input curvature links//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: JMLR: 42192 - 42212[DOI: 10.5555/3692070.3693786]

Ren J, Li Y, Zeng S, Xu H, Lyu L, Xing Y and Tang J. 2024. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention//Computer Vision - ECCV 2024: 18th European Conference. Milan, Italy: Springer-Verlag: 340 - 356[DOI: 10.1007/978-3-031-72980-5_20]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA: IEEE: 10674 - 10685[DOI: 10.1109/CVPR52688.2022.01042]

Ronneberger O, Fischer P and Brox T. 2015. U-net: Convolutional networks for biomedical image segmentation//International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234 - 241[DOI: 10.1007/978-3-319-24574-4_28]

Ross B L, Kamkari H, Wu T, Hosseinzadeh R, Liu Z, Stein G, Cresswell J C and Loaiza-Ganem G. 2025. A geometric framework for understanding memorization in generative models//International Conference on Learning Representations. Singapore: OpenReview: 1 - 11

Schwarzschild A, Feng Z, Maini P, Lipton Z and Kolter J Z. 2024. Rethinking LLM memorization through the lens of adversarial compression//Advances in Neural Information Processing Systems, Curran Associates: 37: 56244 - 56267

Shokri R, Stronati M, Song C and Shmatikov V. 2017. Membership inference attacks against machine learning models//IEEE Symposium on Security and Privacy. San Jose, CA, USA: IEEE: 3 - 18[DOI: 10.1109/SP.2017.41]

Somepalli G, Singla V, Goldblum M, Geiping J and Goldstein T. 2023a. Diffusion art or digital forgery? investigating data replication in diffusion models//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada: IEEE: 6048 - 6058[DOI: 10.1109/CVPR52729.2023.00586]

Somepalli G, Singla V, Goldblum M, Geiping J and Goldstein T. 2023b. Understanding and mitigating copying in diffusion models// Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA: Curran Associates: 36: 47783 - 47803[DOI: 10.5555/3666122.3668193]

Song L, Shokri R and Mittal P. 2019. Membership inference attacks against adversarially robust deep learning models//2019 IEEE Security and Privacy Workshops. San Francisco, CA, USA: IEEE: 50 - 56[DOI: 10.1109/SPW.2019.00021]

Song Y, Sohl-Dickstein J, Kingma D P, Kumar A, Ermon S and Poole B. 2021. Score-based generative modeling through stochastic differential equations//International Conference on Learning Representations. Vienna, Austria: OpenReview: 1 - 36

Stanczuk J P, Batzolis G, Deveney T and Schönlieb C B. 2024. Diffusion models encode the intrinsic dimension of data manifolds//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria:JMLR.[DOI:10.5555/3692070. 3693958]

Steinke T, Nasr M and Jagielski M. 2023. Privacy auditing with one (1) training run //Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA: Curran Associates: 36: 49268 - 49280 [DOI: 10.5555/ 3666122. 3668265]

Sun M, Liu Z, Bair A and Kolter J Z. 2024. A simple and effective pruning approach for large language models//The Twelfth International Conference on Learning Representations. Vienna, Austria: OpenReview: 1 - 23

Tsigler A and Bartlett P L. 2023. Benign overfitting in ridge regression. Journal of Machine Learning Research, 24(123): 1 - 76[10.5555/ 3648699.3648822]

Voigt P and Von dem Bussche A. 2017. The EUGeneral Data Protection Regulation (GDPR): A Practical Guide. 1st ed. Cham: Springer International Publishing: 10 - 555[DOI: 10.5555/3152676]

Wang W, Sun Y, Tan Z and Yang Y. 2024. Image copy detection for diffusion models//Proceedings of the 38th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada: Curran Associates: 37: 14417 - 14456 [DOI: 10.5555/ 3737916. 3738376]

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity//IEEE Transactions on Image Processing, 13(4): 600 - 612[DOI: 10. 1109/TIP.2003.819861]

Warner S L. 1965. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309): 63 - 69 [DOI: 10.2307/ 2283137]

Wei J, Zhang Y, Zhang L Y, Ding M, Chen C, Ong K L, Zhang J and Xiang Y. 2025. Memorization in deep learning: A survey//ACM Computing Surveys, New York, NY, USA: Association of Computing Machinery: 58(4): 1 - 35[DOI: 10.1145/3769076]

Wen Y, Liu Y, Chen C and Lyu L. 2024. Detecting, explaining, and mitigating memorization in diffusion models//The Twelfth International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview. 1-16

Wu J and Harandi M. 2024. Scissorhands: Scrub data influence via connection sensitivity in networks//Computer Vision - ECCV 2024: 18th European Conference. Milan, Italy: Springer-Verlag: 367 - 384 [DOI: 10.1007/ 978-3-031-72970-6_21]

Xie X H, Bian J T, Lai J H. 2022. Review on face liveness detection. Journal of Image and Graphics, 27(1): 63-87 (谢晓华, 卞锦堂, 赖剑煌. 2022. 人脸活体检测综述. 中国图象图形学报, 27(1): 63-87) [DOI:10.11834/jig.210470]

Xiong A, Zhao X, Pappu A and Song D. 2025. The landscape of memorization in LLMs: Mechanisms, measurement, and mitigation. [EB/OL]. [2025-12-12]. https://arxiv.org/pdf/2507.05578

Xu Q, Huang G, Yuan Y, Guo C, Sun Y, Wu F and Weinberger K. 2018. An empirical study on evaluation metrics of generative adversarial networks.[EB/OL].[2018-08-17]. https://arxiv.org/pdf/1806.07755

Ye B H, Kang D Q, Xie X H, Lai J H. 2025. Review of vision-based surface defect inspection methods with incomplete annotations. Journal of Image and Graphics, 30(6): 1661-1689 (叶标华, 康丹青, 谢晓华, 赖剑煌. 2025. 基于视觉的非完全标注表面缺陷检测综述. 中国图象图形学报, 30(6): 1661-1689) [DOI: 10. 11834/jig.240434]

Zhai S, Chen H, Dong Y, Li J, Shen Q, Gao Y, Su H and Liu Y. 2024. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy//Advances in Neural Information Processing Systems, Vancouver, BC, Canada: Curran Associates: 74122 - 74146[DOI: 10.5555/3737916.3740274]

Zhang C, Bengio S, Hardt M, Mozer M C and Singer Y. 2020. Identity crisis: Memorization and generalization under extreme overparameterization//The Eighth International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview: 1 - 39

Zhang C, Bengio S, Hardt M, Recht B and Vinyals O. 2021. Understanding deep learning (still) requires rethinking generalization// Communications of the ACM, New York, NY, USA: Association for Computing Machinery: 64 (3): 107 - 115 [DOI: 10.1145/ 3446776]

Zhang C, Ippolito D, Lee K, Jagielski M, Tramèr F And Carlini N. 2023a. Counterfactual memorization in neural language models// Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA: Curran Associates: 36: 39321 - 39362 [DOI: 10.5555/ 3666122.3667830]

Zhang G, Wang K, Xu X, Wang Z and Shi H. 2024a. Forget-me-not: Learning to forget in text-to-image diffusion models//2024 IEEE/ CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA: IEEE: 1755 - 1764 [DOI: 10. 1109/ CVPRW63382.2024.00182]

Zhang J, Das D, Kamath G and Tramèr F. 2024b. Membership inference attacks cannot prove that a model was trained on your data// 2025 IEEE Conference on Secure and Trustworthy Machine Learning, Copenhagen, Denmark: IEEE: 333-345 [DOI: 10.1109/SaTML64287.2025.00025]

Zhang Q, Lai J H, Xie X H, Chen H X. 2023b. A summary on group re-identification. Journal of Image and Graphics, 28(5): 1225-1241 (张权, 赖剑煌, 谢晓华, 陈泓栩. 2023. 小股人群重识别研究进展. 中国图象图形学报, 28(5): 1225-1241) [DOI:10.11834/jig.220697]

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE: 586 - 595 [DOI: 10.1109/CVPR.2018.00068]

Zhang Y, Chen X, Jia J, Zhang Y, Fan C, Liu J, Hong M, Ding K and Liu S. 2024c. Defensive unlearning with adversarial training for robust concept erasure in diffusion models// Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates: 36748 - 36776 [DOI: 10.5555/3737916.3739074]

Zhang Y, Jia J, Chen X, Chen A, Zhang Y, Liu J, Ding K and Liu S. 2024d. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now//Computer Vision - ECCV 2024: 18th European Conference. Milan, Italy: Springer-Verlag: 385 - 403 [DOI: 10.1007/978-3-031-72998-0_22]

## 作者简介

翁子辉,男,硕士研究生,主要研究方向为计算机视觉。E-mail: wengzh7@mail2.sysu.edu.cn

赖剑煌,通信作者,男,教授,博士生导师,主要研究方向为计算机视觉与模式识别。E-mail:stsljh@mail.sysu.edu.cn

张权,男,博士后,主要研究方向为计算机视觉与模式识别。E-mail: zhangq689@mail.sysu.edu.cn

谢晓华,男,教授,博士生导师,主要研究方向为计算机视觉与模式识别。E-mail:xiexiaoh6@mail.sysu.edu.cn