Design 设计, 2024, 9(6), 109-115
Published Online December 2024 in Hans. https://www.hanspub.org/journal/design
https://doi.org/10.12677/design.2024.96672

Hans汉斯

文生视频大模型设计的安全风险及其矫治

陈　钲，陈　靖

南京林业大学，人文社会科学学院，江苏　南京

收稿日期：2024年10月8日；录用日期：2024年12月2日；发布日期：2024年12月10日

摘　要

本文深入探讨了文生视频大模型设计中的安全风险及其矫治策略。随着人工智能技术的快速发展，文生视频大模型如Sora和PixelDance等，已经能够根据文本描述生成视频内容，为影视、广告、教育等行业带来了革命性的变化。然而，这些技术进步也伴随着隐私泄露、数据安全、道德价值偏离等安全风险。本文分析了训练数据、提示词注入攻击、电信欺诈、道德价值偏离和人机交互等方面的风险，并介绍了差分隐私和联邦学习等风险治理策略。

关键词

文生视频大模型，安全风险

Security Risks and Remediation of the Design of Text-to-Video Generative Models

Zheng Chen, Jing Chen

College of Humanities and Social Sciences, Nanjing Forestry University, Nanjing Jiangsu

Received: Oct. 8th, 2024; accepted: Dec. 2nd, 2024; published: Dec. 10th, 2024

Abstract

This article delves into the safety risks and rectification strategies in the design of text-to-video generative models. With the rapid advancement of AI technology, models such as Sora and PixelDance can generate video content based on textual descriptions, revolutionizing industries like film, advertising, and education. However, these technological leaps also come with safety risks, including privacy breaches, data security, and moral value deviations. The article analyzes risks in training data, prompt injection attacks, telecommunication fraud, moral value shifts, and human-computer interaction, and proposes risk governance strategies like differential privacy and federated learning to ensure the healthy development of technology and the harmonious stability of society.

文章引用: 陈钲, 陈靖. 文生视频大模型设计的安全风险及其矫治[J]. 设计, 2024, 9(6): 109-115.
DOI: 10.12677/design.2024.96672

Keywords

Text-to-Video Generative Models, Security Risks

1. 引言

在人类凭借自身尚未充分发掘奥秘的生物大脑，设计并开发出同样蕴含深邃原理的大型深度学习模型的过程中，我们见证了科学与技术的非凡融合。这些模型，依托海量的训练数据与精妙设计的模型架构，在达到某一“临界点”时，不仅展现出了卓越的性能与前所未有的能力，也标志着人工智能领域的一次重要飞跃。然而，尽管这一突破令人振奋，其开发者对于模型内部机制的深刻理解与精准控制仍显不足。在此背景下，大型模型的安全风险管理与矫治设计成为了一个亟待解决的问题，尤其是随着文生视频大模型等新兴应用的涌现，其安全风险的复杂性与多样性更加凸显。

2024 年 3 月 5 日，第十四届全国人民代表大会第二次会议在北京隆重开幕，政府工作报告中首次提及“人工智能”，这一信号引发了人大代表对于“AI 大模型”及其安全风险矫治设计的广泛讨论与策略建议[1]。大模型的发展如同遵循“摩尔定律”的火箭，一路高歌猛进，甚至在学术界对其本质定义与边界尚存争议之时，其已迅速构建起了一个丰富多元、错落有致的应用生态。例如，在卷积神经网络(CNN)框架下的大型模型，凭借其出色的图像识别能力，成为了计算机视觉领域的佼佼者；而大型语言模型则通过深度学习，展现了对人类语言的深刻理解与生成能力，为自然语言处理开辟了新天地[2]。此外，依据应用场景的不同，预训练模型、自然语言处理模型等被广泛应用于各类任务处理，同时，针对特定行业如医疗(疾病诊断模型)、教育(知识大模型)等领域定制训练与优化的大模型也层出不穷，展现了 AI 技术的广泛渗透力与深远影响[3]。

在这些大模型的快速发展中，一个尤为值得关注的趋势是文生视频大模型的兴起。这类模型通过巧妙的算法设计，能够将文本转化为生动的视频内容，为媒体创作、教育娱乐等领域带来了革命性的变化。然而，随之而来的安全风险也不容忽视。如何设计有效的安全机制，以确保文生视频大模型在生成内容时的真实性、准确性与合规性，防止虚假信息的传播与不良内容的生成，成为了当前亟需解决的关键问题。因此，在大模型的设计与应用过程中，安全风险矫治设计应被视为一个不可或缺的重要组成部分，以保障技术的健康发展与社会的和谐稳定。

2. 视频大模型设计

2.1. 视频大模型设计的演变

在人工智能领域，视频大模型的发展经历了从初步探索到技术革新的演变过程，其技术基础涵盖了循环网络(RNN)、生成对抗网络(GAN)、自回归模型(autoregressive transformers)以及扩散模型(diffusion models)等多种深度学习架构[2]。其中，文生视频大模型(Text to Video Generation)作为一类能够根据文本输入生成相应视频内容的人工智能大模型，近年来备受关注。

文生视频大模型结合了自然语言处理和计算机视觉两大技术领域，是人工智能应用领域的代表性成果之一。然而，相较于文生文、文生图模型，文生视频大模型在技术上更为复杂，其发展前景及应用形

式仍有待进一步探索。在文生视频大模型发展的早期阶段，基于文本检索的视频片段拼接方法曾是主流，但这种方法的效果较为局限，无法根据文本语义生成原创的视频内容。

随着生成式对抗网络(GANs)的出现，文生视频模式迎来了新的转机[4]。GANs 通过两个神经网络的对抗训练，能够生成近似逼真的样本，为文生视频大模型提供了新的思路。2018 年前后，一些专门用于视频生成的模型如 MoCoGAN、TGAN 等相继出现，这些模型能够结合文本生成简单的视频内容，但文本与视频的关联程度仍远未达到理想状态[5]。

随后，研究者开始尝试大型的文本–视频预训练模型，以更好地关联起文本与视频之间的复杂关系。在这一阶段，扩散模型(Diffusion Models)取得了突破性进展。扩散模型通过“去噪过程”生成样本，其可逆的建模过程使得模型能够生成高质量、多样化的样本。在文本、音视频等多个领域，扩散模型已展现出了不俗的生成效果。然而，扩散模型也存在训练成本高昂、生成时间长等缺点。

与此同时，基于 Transformer 架构的大语言模型生成视频技术路线也成为了当下的主流路径之一。谷歌发布的 VideoPoet 模型便是基于大语言模型的生成式 AI 视频模型的代表。该模型通过接收输入的文字、图片、视频，对内容进行编码，并最终生成视频内容。然而，由于视频编码的复杂程度，该模型也需要耗费大量的计算资源，且依赖于训练数据的质量和多样性。

此外，多模态大模型(Multimodal Large Models)也展现出了强大的跨模态生成能力。这些模型能够基于大规模的多模态训练数据(文本、图像、视频)，通过自监督或半监督的方式学习跨模态的生成能力[4]。多模态大模型的出现，为视频大模型的发展提供了新的可能性和方向。

2.2. 视频大模型设计的现状

在视频大模型领域，当前应用较为广泛且表现突出的几类模型主要包括 Runway 视频生成模型、Pika 视频生成模型，以及 Open AI 最新推出的 Sora 视频生成模型。这些模型不仅代表了视频生成技术的最新进展，也揭示了该领域面临的挑战与机遇。

本文着重探讨 Open AI 的 Sora 模型，该模型在视频生成领域展现了卓越的能力。结合前文分析，我们可以看出，视频大模型生成的技术难度极高，涉及从文本到视频的转换，需要解决逻辑性、物理特性以及人类生物特征等一系列复杂问题。此外，强大的算力是视频生成的基础，但即便在解决了大部分技术难题后，当前多数视频模型仍只能生成几秒时长的视频内容。当创作者试图生成较长视频时，需不断利用提示词勾勒完整的文本内容，这往往导致视频上下文逻辑连贯性缺失，难以将素材完整串联[5]。

Open AI 的 Sora 模型在这一问题上取得了显著突破，不仅大幅延长了生成视频的时长，还保持了视频中主题和人物场景的一致性。此外，Sora 模型支持视频、图像、文本提示词等多形式的输入方式，这有助于提高生成内容的准确性。通过对 Open AI 官方文件的深入解读，我们发现 Sora 模型融合了扩散模型与 Transformer 架构的大语言模型，同时借助 ChatGPT 强大的语言文字理解能力和 DALLE3 文生图模型的图像标注功能，使得 Sora 模型能够更有效地学习和预测未训练数据的泛化能力，从而生成更加精准、有趣且符合逻辑的视频内容。

然而，即便是如此强大的 Sora 模型，也面临着当前视频生成大模型的共有缺点。首先，模型对数据存在过分依赖，导致输出内容质量不够稳定。其次，Sora 模型在某些方面仍具有局限性，如无法准确模拟许多基本交互作用的物理过程。此外，从已发布的视频来看，Sora 模型在理解人类世界内在逻辑方面仍存在不足，导致视频出现“局部合理”与“整体荒谬”的矛盾。这种矛盾反映了视频生成过程中的一种自觉、有意识的“再创作”过程，类似于抽象派画作，虽然局部可能合理，但整体却可能因缺乏深层次的逻辑连贯性而显得荒谬。

3. 视频大模型设计中的安全风险

3.1. 训练数据风险

文生视频大模型的训练依赖于海量的数据集，然而，数据采集、训练及部署的复杂性使得原始数据中的不可控因素成为安全隐患。具体而言，未经验证的数据可能潜藏恶意模式，导致模型在测试样本上表现异常。此外，攻击者可通过推断攻击等手段，暴露敏感信息的样本来源，进而提取公共或个人隐私信息[6]。这类信息在传输过程中亦存在被第三方恶意截获的风险，对使用者的隐私安全构成严重威胁。

随着大模型的广泛应用，个人及企业用户的隐私数据在模型使用过程中难免被涉及。一旦这些数据被泄露并用于恶意活动，将对用户造成严重的安全风险。同时，大模型的“黑箱”特性加剧了这一风险。由于模型结构和参数量庞大，使用者难以了解其内部运作方式及决策过程，导致模型可能产生“大模型幻觉”，即复现训练数据中的原始问题，生成不符合事实、带有偏见、政治敏感甚至违法违规的内容[7]。

随着 AI 大模型商业化的成熟，算法和训练数据逐渐从开源转向闭源模式，底层训练数据来源的不透明性进一步加剧了大模型的训练数据隐私问题及大模型幻觉的隐现。

3.2. 提示词注入攻击风险

提示词注入攻击是文生视频大模型面临的另一大风险。这种攻击类似于 SQL 注入，通过向模型输入非法或违规的提示词，实现对违规内容的生成或敏感数据的获取及修改[8]。对于生成式大语言模型而言，即使存在对敏感违规提示词的初始规避，攻击者仍可通过巧妙构造的提示词绕过这些规避措施，生成违法、暴力或色情等内容。

在文生视频模型中，由于视频内容相较于文字和图片更为复杂且直观，因此某些恶意提示词攻击的效度会被放大，可能引发更大的风险。例如，攻击者可通过提示词注入，使模型生成具有误导性或危害性的视频内容，进而对社会造成不良影响。

此外，攻击者还可通过近义词替换、单词或汉字拆分、混淆模糊的符号 token 等方法，将敏感关键词“脱敏化”，绕过安全对齐训练中风险 token 的引导[8]。这种“脱敏”后的关键词在模型中可能被视为无风险内容，从而加剧了提示词注入攻击的风险。

3.3. 电信欺诈风险

在自媒体等相对独立的传播环境中，生成式人工智能如 ChatGPT 等存在被误用或滥用的风险。个体的自由度和对特定利益的追求，使得 AIGC (人工智能生成内容)制造虚假信息变得更加容易[9]。例如，2023 年 2 月，一起关于杭州取消限行的虚假“新闻稿”事件，便是由一位小区业主在使用 ChatGPT 时尝试编写的，该文章随后被错误地传播为真实新闻，造成了广泛的影响。

这一事件凸显了技术强大内容生成能力的同时，也暴露了缺乏有效的监管机制的问题。在当前的算法框架下，缺乏对事实的核实能力，使得未经验证的信息容易被包装成“新闻信息”。部分社交媒体平台和自媒体出于利益需要，甚至利用算法扩散假新闻，进一步加剧了虚假信息的传播。当信息传播平台与内容生成工具的结合变得更加自由便捷时，一旦有 AI 造假信息事件被广泛传播，便可能轻易摧毁人们长久建立的信任体系。

Sora 文生视频大模型的出现，进一步加剧了此类风险。Sora 模型能够以文本、图片以及视频方式进行输入，将静态图片转变为动态视频，或对基础输入的视频进行扩展。这种逼真的图片或视频生成能力，可能被恶意使用者用于非法活动，带来极大的欺诈风险。特别是当自媒体创作者利用 Sora 等视频大模型进行视频内容创作时，“创作乱象”可能会被放大，从过去的文字“新闻造假”演变为更具欺骗性的“视

频造假”[9]。这种造假不同于简单的视频拼接或 AI 换脸，而是基于创作者提示生成的全新内容，使得普通信息接受者难以分辨其真伪。若不法分子利用 Sora 等视频大模型生成诈骗类内容，则极易造成他人的财产损失。

3.4. 道德价值偏离风险

文生视频大模型在道德价值方面也存在潜在风险。与现有的文、图、视频内容一样，大模型生成的视频可能受到数据本身刻板印象和偏见的影响[10]。这是开发人员努力避免但难以完全消除的问题之一。隐藏于数据中的歧视、偏见以及各类不和谐的声音，都可能以另一种形式被保存下来。

大模型的创造者对训练数据的选取态度，反映了其是否致力于消除这样的偏见。然而，在实际操作中，一些人工智能公司如 Open AI 等，在选取训练数据时可能面临挑战。例如，与 Reddit 等社交网站签订协议，使用其内容数据训练人工智能程序时，可能包含偏离传统价值观的言论或违规图片。这些内容作为训练数据，可能不符合公司的运作初衷。

理想状态是 AI 公司在价值标准中取得一个较为平衡的点，既不过度校准以避免“AI 幻觉”的产生，也不过度忽视以避免偏见的延续。然而，在实践中，这种平衡往往难以实现。例如，谷歌旗下的 Gemini 视频大模型在生成马斯克图片时出现了“黑人马斯克”的错误，这可能源于对“种族主义”、“性别歧视”等行为的“过度矫正”，导致 AI 生成的内容偏离了客观事实。这种偏离客观事实的“AI 幻觉”，虽然相较于传统的违规内容严重程度有所降低，但仍可能对用户造成不良影响。

3.5. 人机交互风险

正如 AI 公司需要平衡 AI 工具生成内容的平衡点，用户使用它们时，也存在一个隐性的“使用平衡点”。艺术家、设计师或是电影制作者使用 Sora 等工具创作时，作品中无法避免地存在一些裸露或暴力元素，这样的内容会被视作“对价值观的完全违背”吗？技术工具在很大程度上是较为中立的，而生成内容的道德评判标准一方面为使用者意图所左右，另一方面则取决于生成内容的客观事实影响[11]。

在信息技术领域，人工智能(AI)作为一种响应指令的工具，它的基本目标并非在于提供一个绝对客观中立的信息展现，而是在解读使用者的意图基础上，通过对数据的筛选和编排来满足其预期。这种机制在某些情况下可能导致信息的非中性处理。

以 ChatGPT 为例，当用户请求关于某个主体的信息时，AI 系统可能会倾向于采用正面的、甚至有时过度夸张的语言来构建回应。这种倾向性可能源自于 AI 系统的设计原则，即优先考虑满足用户的预期和喜好。当系统预测到某些回答可能不会获得用户的正面反馈时，它可能会尝试不同的信息组合以产生更受欢迎的回答，而这也正是 AI 虚构内容的一大重要原因。

训练数据的质量高低同样影响大模型的表现能力，训练数据质量不够或者训练流程的不完备，会导致大模型生成无意义或与提示词无关的内容，这样的情况被成为“大模型幻觉”，“大模型幻觉”可能会给用户带来理解偏差甚至产生错误决策。

与笔者前文所描述的直观偏离不同，Sora 等视频工具还存在与用户隐性的交互风险，在 Sora 最初的演示视频放出时，人们为其逼真的现实世界模拟画面震惊，但类似图片造假，若仔细甄别 Sora 生成的某些视频，便会发现其存在一些十分隐秘但又有细微问题的部分。如在一则蚂蚁穿过地下洞穴的视频中，该视频的光影展现和崎岖的洞壁十分逼真，几乎与现实世界无异，但仔细观察便会发现视频中的蚂蚁有四只脚，然而地球上并不存在四只脚的蚂蚁，如果这样的视频被长期、广泛地传播，那么则会造成隐性人机传播危机。马斯克则大胆地嘲讽：“这将会让整整一代孩子接受虚假的教育。”

与 ChatGPT 等 AI 工具诞生时面临的争议类似，许多人认为它们将会挤占人类更多的思考时间，让

传统的教育体系崩塌，在没有合适的解决办法之前，公司与学校又将如何看待并应用这些智能的工具呢？这一直是一个被永恒争论的议题。

4. 风险治理策略

在大数据与人工智能蓬勃发展的时代背景下，大模型训练数据的隐私保护问题日益凸显。从技术治理的角度出发，差分隐私(Differential Privacy)与联邦学习(Federated Learning)作为两种前沿的技术范式，为解决这一问题提供了有效的路径。

4.1. 差分隐私机制

差分隐私是一种旨在保护个体数据隐私的数据分析技术。其核心理念在于，在初始数据中引入适量的噪声，以掩盖单个数据点对统计结果的影响，从而在确保数据统计信息准确性的同时，实现对个体数据隐私的保护[12]。这种隐私保护的程度是可控的，主要取决于所添加噪声的量级。

在大模型隐私保护领域，差分隐私的应用可以从以下几个方面展开：首先，在训练数据输入模型之前，可以对数据集进行差分隐私噪声注入处理。对于数值型数据，可以采用拉普拉斯机制注入噪声；对于类别型数据，则可以使用指数机制进行随机响应，并辅以拉普拉斯噪声进行隐私保护。此外，还可以利用语义相似性对类别进行适度的泛化和归并，以降低隐私泄露的风险。对于结构化数据，可以基于其结构关系设计合适的数据合成(Data Synthesis)和机会抽样(Data Sub-sampling)策略。前者通过建立一个可以生成类似结构数据的模型，生成一批人工合成的噪声化数据，从而掩盖原始数据；后者则从原始数据中抽取子集记录，形成新的样本集，再对样本集加入噪声，实现数据的噪声化处理[12]。

通过对数据进行细致的分类，并选择恰当的差分隐私保护机制，可以制定出切实可行的隐私保护措施。类似的方法同样可以应用于模型更新以及模型输出的策略中，以全面保障大模型训练数据的隐私安全。

4.2. 联邦学习机制

联邦学习作为一种分布式的机器学习模式，为解决数据隐私保护问题提供了全新的思路[13]。在联邦学习框架下，中央服务器将全局模型的参数分发给各个联邦节点(即终端设备)，联邦节点再根据本地数据对全局模型进行局部训练与参数更新。重要的是，联邦节点将更新后的参数而非原始数据上传到服务器，以供全局模型的优化更新。

有学者指出，联邦学习不仅能够有效利用分布式数据资源，提高模型的训练效率和性能，更重要的是，它能够在不暴露用户私有数据的前提下，实现模型的全局优化[13]。这一特性使得联邦学习在保护用户隐私、降低数据泄露风险方面具有显著优势。

数据隐私保护层面，联邦学习通过避免将用户的私有原始数据上传到服务器，有效保护了数据的隐私。这对于大模型训练来说至关重要，因为大型模型通常需要大量数据来支持其复杂的结构和算法。通过联邦学习，可以在不牺牲数据隐私的前提下，充分利用分布式数据资源，提高模型的训练效果。

其次，联邦学习机制允许不同地域、不同设备上的数据参与模型训练，这有助于增加模型的多样性和泛化能力。由于不同设备上的数据往往具有不同的分布和特征，因此通过联邦学习得到的模型能够更好地适应各种场景和条件，从而提高其鲁棒性和稳定性。

再者，在大型模型训练过程中，数据泄露、模型被恶意攻击等安全风险时有发生。联邦学习通过将数据留在本地进行训练，并仅上传模型参数，大大降低了这些安全风险。即使某个联邦节点被攻击或泄露数据，也只会影响该节点的局部数据，而不会对整个全局模型造成严重影响。

最后但也很关键的是，虽然联邦学习在通信方面存在一定的开销，但通过优化算法和参数更新策略，可以显著降低通信成本。例如，可以采用稀疏更新、量化压缩等技术来减少传输的数据量，从而提高通信效率。这有助于在大规模分布式系统中实现更加高效和可持续的模型训练。

5. 结语

随着人工智能技术的飞速发展，我们正迈入一个潜力无限的新纪元，大型深度学习模型，尤其是文生视频大模型，以其变革性的力量，预示着生活与工作方式的深刻转变。然而，正如诸多学者所深刻洞察，技术的每一步前进都伴随着数据隐私、安全风险、道德价值偏离及人机交互等多重挑战。

为确保这些前沿技术能够健康演进，最大化其正面效益，我们必须采取多维度的策略加以应对。这包括强化数据隐私保护机制，构建有效的风险治理框架，提升模型的透明度与可解释性，以及加强公众的数字素养教育，使之能更准确地理解和评估 AI 生成的内容。正如伦理学家所强调的，技术的每一步都应与人类的价值观、伦理规范相契合，同时维护社会的多样性和包容性。

跨学科合作、政策引导与技术革新，是我们共同塑造一个安全、公正、可持续 AI 未来的关键。在这个过程中，技术开发者、政策制定者、教育工作者及广大用户，每个人都是不可或缺的参与者。

参考文献

[1] 本刊综合. 全国两会启航安全保密新征程[J]. 保密工作, 2024(3): 5-7.

[2] 腾讯研究院. 大模型安全与伦理研究报告[EB/OL]. https://www.tisi.org/27403/, 2024-09-05.

[3] 朱光辉, 王喜文. ChatGPT 的运行模式、关键技术及未来图景[J]. 新疆师范大学学报(哲学社会科学版), 2023, 44(4): 113-122.

[4] Pan, Y., Mei, T., Yao, T., Li, H. and Rui, Y. (2016) Jointly Modeling Embedding and Translation to Bridge Video and Language. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 4594-4602. https://doi.org/10.1109/cvpr.2016.497

[5] Tulyakov, S., Liu, M., Yang, X. and Kautz, J. (2018) MoCoGAN: Decomposing Motion and Content for Video Generation. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 526-1535. https://doi.org/10.1109/cvpr.2018.00165

[6] Wei, A., Haghtalab, N. and Steinhardt, J. (2024) Jailbroken: How Does LLM Safety Training Fail? arXiv: 2307.02483.

[7] Sun, H., Zhang, Z., Deng, J., *et al*. (2023) Safety Assessment of Chinese Large Language Models. arXiv: 2304. 10436.

[8] Liu ,Y., Deng, G., Xu, Z., *et al*. (2023) Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. arXiv: 2305. 13860.

[9] 胡泳, 刘纯懿. 大语言模型“数据为王”: 训练数据的价值、迷思与数字传播的未来挑战[J]. 西北师大学报(社会科学版), 2024, 61(3): 43-54.

[10] 许雪晨. ChatGPT 等大语言模型赋能数字时代金融业: 基于隐私保护, 算法歧视与系统风险[J]. 暨南学报(哲学社会科学版), 2024, 46(8): 108-122.

[11] Yang, H., Ma, X., Du, K., Li, Z., Duan, H., Su, X., *et al*. (2017) How to Learn Klingon without a Dictionary: Detection and Measurement of Black Keywords Used by the Underground Economy. 2017 *IEEE Symposium on Security and Privacy* (*SP*), San Jose, 22-26 May 2017, 751-769. https://doi.org/10.1109/sp.2017.11

[12] Yang, M., Guo, T., Zhu, T., Tjuawinata, I., Zhao, J. and Lam, K. (2024) Local Differential Privacy and Its Applications: A Comprehensive Survey. *Computer Standards & Interfaces*, **89**, Article ID: 103827. https://doi.org/10.1016/j.csi.2023.103827

[13] Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J. and Zhang, W. (2022) A Survey on Federated Learning: Challenges and Applications. *International Journal of Machine Learning and Cybernetics*, **14**, 513-535. https://doi.org/10.1007/s13042-022-01647-y