FROST & SULLIVAN

60 Years of Growth, Innovation and Leadership

2025年中国合成数据解决方案发展洞察

2025年9月

A Frost & Sullivan
White Paper

FROST & SULLIVAN
沙利文

执行摘要

本报告聚焦**合成数据（Synthetic Data）解决方案**，分析其发展现状、技术路径、市场格局及未来趋势。合成数据是通过算法、仿真或其他方法人工生成的数据，能够模仿现实世界数据的结构、特征和统计属性，但不受现实世界数据的限制。当前，大模型技术和生成式AI的突破正推动人工智能范式由“以模型为中心”向“以数据为中心”转型。合成数据解决方案能够系统性地解决AI开发到落地的多重数据瓶颈，已经从空白真实数据的简单替代升级为驱动AI变革的核心战略资产，并在自动驾驶、具身智能和工业场景展现出巨大的价值潜力。

研究目的

本白皮书的研究目的在于全面梳理合成数据解决方案的发展历程、现状、核心价值、产业链图谱及其在全球的市场规模和地区渗透情况，并探讨合成数据解决方案未来的发展趋势。合成数据解决方案为模型的训练和开发以及AI应用的落地提供了高质量、高可用性、低成本、可用于AI消费的数据来源，已在自动驾驶、具身智能、工业等应用场景展现出巨大潜力，我们期望为相关领域的研究者、开发者以及企业提供有价值的参考信息，促进技术进步和产业发展。

目录

目录

6 Key Findings
关键发现

大模型技术和生成式AI的突破正推动人工智能范式由“以模型为中心”向“以数据为中心”转型，合成数据已经从空白真实数据的简单替代升级为驱动AI变革的核心战略资产。

预计到2026年，由于数据隐私和安全问题，

约有 **75%** 的企业将使用生成式AI来生成合成客户数据。

超过 **53%**

公司将边缘案例测试列为合成数据的首要用例

预计到 **2030年**，

人工智能模型中合成数据的生成量将超过真实数据的使用量。

在工业场景或具身智能领域，未来的数据范式正朝着

1%人类数据 **+**

99%高效合成数据

的混合模式演进，其成功依赖于“ Human in Loop”（人在环）机制

在合成数据供应商中，专注解决方案型展现出更强的延展性与商业化潜力

深信科创以物理真实数据为“种子”，提供高价值、高物理精准性的合成数据资产，在中国合成数据解决方案提供商中

领先

第一章

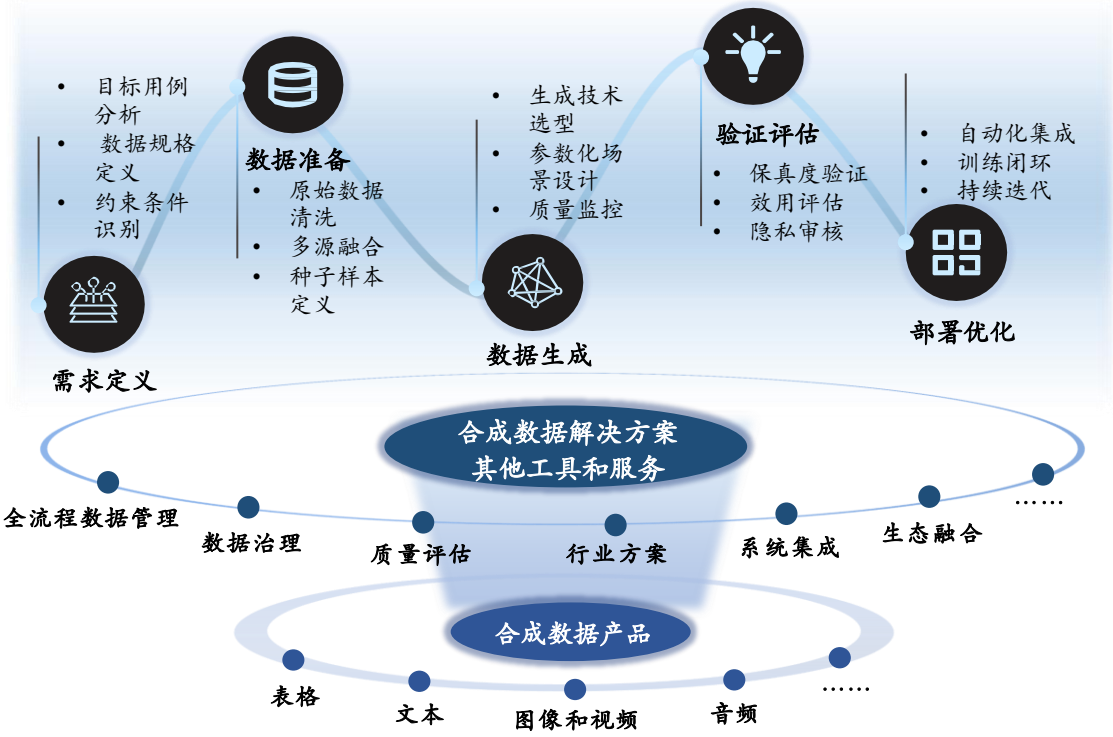
合成数据解决方案概述

合成数据解决方案定义

合成数据（Synthetic Data）是通过算法、仿真或其他方法人工生成的数据，能够模仿现实世界数据的结构、特征和统计属性，但不包含任何实际的现实世界信息。根据数据类型，合成数据可分为表格、文本、图像和视频、音频、时间序列和其他类型。合成数据的生成通常基于预定义的规则和模板、机器学习模型，或在仿真环境中生成，以提供模拟真实、符合隐私且可随时使用的数据集，且不受真实数据的限制。

合成数据解决方案面向AI时代模型训练和应用部署的数据需求，聚焦于解决真实数据稀缺、敏感、收集难度大等挑战，覆盖从**需求定义、数据准备、数据生成到数据评估、部署优化**的全流程闭环。合成数据解决方案在将合成数据本身作为一种资产的同时，还提供覆盖**全生命周期的数据管理、数据治理保障和质量评估体系**，并提供**系统集成、行业方案、生态融合**等核心服务，帮助企业完成以数据为中心的全流程价值交付。

合成数据解决方案的产品服务范围和技术流程



来源：弗若斯特沙利文

合成数据解决方案发展历程

1.0 填补空白的辅助工具

此阶段合成数据以随机分布、统计抽样和机理仿真为主，主要生成表格等**结构化数据**，聚焦于解决工业仿真、科学统计等领域真实数据的获取困境。然而，合成数据生成效率仅为真实数据采集的30%，且无法反映多变量动态交互。

2.0 AI落地的重要组件

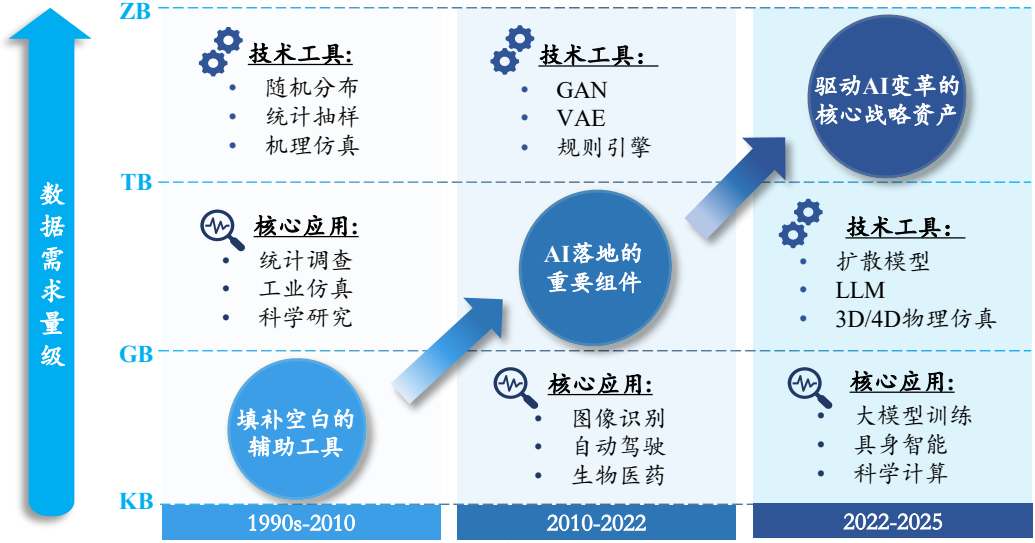
GAN、VAE等技术的突破使合成数据格式扩展到语音、图像和视频等，并广泛应用于图像识别、自动驾驶、生物医药等多个领域。同时，隐私和合规的需求升级，驱动合成数据成为AI落地的重要组件。

3.0 驱动AI变革的核心战略资产

大模型和生成式AI的突破正推动AI范式由“以模型为中心”向**“以数据为中心”**转型，合成数据展现出应对**大模型训练**与**具身智能进化**数据问题的巨大价值潜力。

- **互联网高质量文本**资源正接近枯竭，合成数据成为大模型训练的“可再生燃料”：合成数据已在OpenAI、Meta、英伟达等AI头部企业的大模型预训练与对齐阶段中使用，而大模型本身也能够生成合成数据。
- 合成数据是驱动AI从感知智能向具身智能跃迁的重要基础设施。具身智能训练所需的**物理交互数据**面临着千倍缺口的困境，而高保真物理仿真可将有限人类动作样本扩展至千倍规模，实现机器人零样本泛化。

合成数据解决方案发展历程



来源：弗若斯特沙利文

当前的数据模式在AI时代面临哪些挑战?

AI-Ready的数据是AI项目成功落地的基础，意味着高质量、高可用性、低成本、可用于AI消费的数据成为刚需。预计到2026年，将有60%的AI项目由于“数据未准备好”而被企业放弃。

一、数据可用性不足

在许多行业中，很多AI项目因数据不可用或不完整而受阻，数据收集成为主要障碍。研究发现，机器学习开发社区中用于训练模型的大多数数据集都被重复使用或借用，**缺乏针对性**。这导致项目目标不一致，最终产品不准确。同时，互联网公开训练数据面临枯竭瓶颈，行业面临“训练数据饥荒”，逼迫开发者探索新途径。

二、数据质量问题

制造业调研显示，高达87%的AI项目未能进入生产环境，其中主要原因是数据质量问题，如**缺失、不一致、错误标签**等。现实世界的数据集有时会受到不平衡的影响，收集有偏见的数据会导致AI/ML模型出现偏差和错误，在敏感应用中风险尤高，应当高度重视代表性与公平性问题。

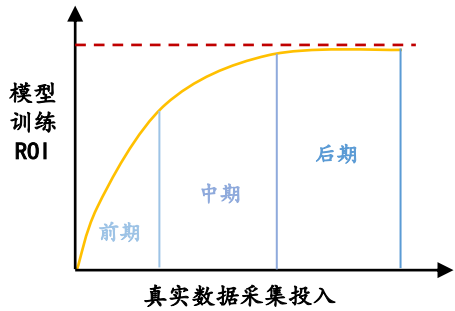
三、高成本和合规要求

真实数据的**收集、清理和维护**是一个昂贵且耗时的过程。团队必须投入大量资源进行人工标注、确保数据准确性、解决不一致问题并消除偏见。这些成本会导致项目延误，并降低数据驱动决策的效率。

随着欧洲GDPR和医疗领域HIPAA等法规出台，保护敏感的真实数据面临着越来越严格的要求。**数据共享**也变得更复杂，进一步限制了合作和创新的机会。

四、模型精度提升瓶颈

随着AI项目深入，需要覆盖更多复杂、罕见和边缘场景，拍摄、标注与质量控制成本急剧上升，真实数据采集重建的**边际成本不断增加**，但**模型训练的回报率逐渐降低**。当模型与数据覆盖度达到一定水平后，新增数据很难带来显著提升，以真实数据为主导的模型精度提升进入瓶颈阶段。

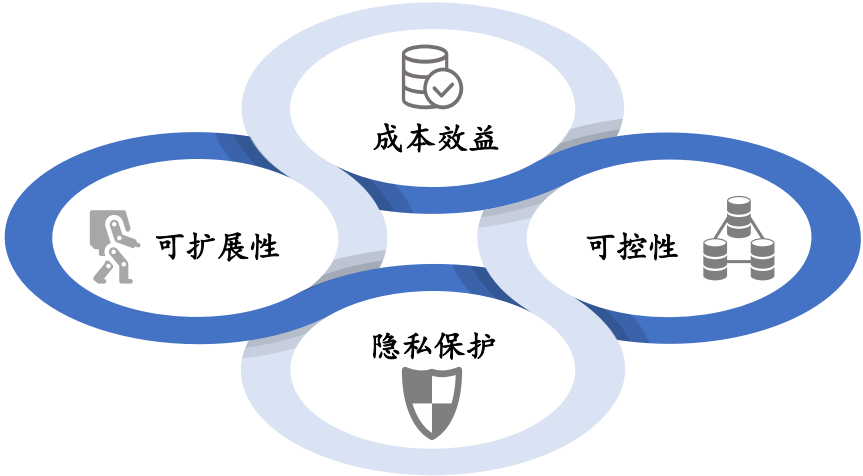


初期：少量真实数据就能显著提升模型精度与训练的ROI；**中期**：随着继续投入，数据带来的增量效益下降明显；**后期**：投入大量资源所换来的精度提升几乎停滞，边际回报趋近于零。

第二章

合成数据解决方案关键能力分析

合成数据解决方案的核心优势



可扩展性

合成数据支持**高效、灵活**的大规模数据生成，满足机器学习和AI模型对海量训练数据的需求。一旦生成环境搭建完成，便可以通过算法迭代，轻松产生**无限量**的数据变体，且**边际成本极低**。它允许企业按需创建大量、多样化的训练数据集，而无需投入相同的成本和精力。

例如，通过合成数据生成技术，可以快速生成数百万张在不同光照、天气和角度下的虚拟街道图像，其规模和多样性远超物理传感器所能捕获的极限。这种模式不仅加速了开发周期，还为测试和验证AI系统在无数假设情境下的表现提供了安全且经济的解决方案。

可控性

现实世界的数据可能存在偏差，或无法用于特定用例，从而限制了分析和机器学习模型的有效性。合成数据是填补数据集空白和解决代表性不足场景的有力工具，其生成技术允许研究人员精确控制数据分布、特征和异常值，从而减少真实数据中存在的偏见并提高模型鲁棒性。

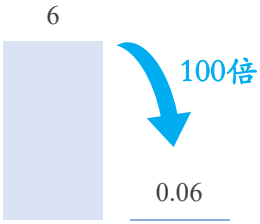
通过人为增加指定场景的数据量，合成数据可以确保模型看到**更平衡、更多样化**的示例集。因此，使用合成数据（或真实数据和合成数据的混合）进行训练实际上不仅可以**提高模型性能和公平性**，还能够显著提升其在极端情况下的安全性能和泛化能力。

来源：弗若斯特沙利文

成本效益

传统基于真实世界的数据解决方案需要成本高昂、耗时耗力，且逻辑复杂的数据采集、清洗和人工标注流程。而合成数据的生成无需调查、访谈或使用昂贵的传感器设备，从而大大降低了获取成本。其次，合成数据集本质上是干净且一致的，从而减少了数据预处理和验证所花费的大量时间。合成数据彻底改变了企业获取高质量训练数据的门槛，尤其适用于需要海量标注数据的计算机视觉项目。

真实与合成图像的成本对比

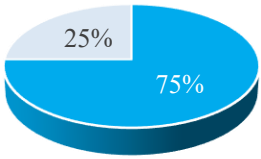


通过数据标注服务获得一张带注释的真实图像可能要花费6美元，而通过合成方式生成一张同等价值的带注释图像仅需约0.06美元。这意味着成本降低了**100倍**。

隐私保护

许多人工智能应用（例如金融或医疗保健领域的应用）依赖于受法规保护的敏感个人数据。使用真实的客户或患者数据来训练模型可能会引发隐私泄露和合规性问题。而合成数据提供了一种**低风险**的解决方法：由于它是人工生成的，不包含任何可识别个人身份的信息，因此可以自由使用而不会有泄露个人隐私的风险。这一特性使合成数据成为受严格监管的行业推动数据协作和AI创新的关键工具。

以AI技术生成客户数据的企业占比



预计到2026年，由于数据隐私和安全问题，约有**75%**的企业将使用生成式AI来生成合成客户数据。

合成数据解决方案有哪些应用价值?

合成数据解决方案是贯穿AI和ML workflow的多功能工具，能够支持模型复杂推理、帮助模型掌握领域知识、全面赋能测试验证与风险控制，并开发前沿领域的研究新范式。

提高认知与复杂推理能力

复杂推理被认为是模型的“北极星能力”。在实际训练中，合成数据能够**通过填补真实数据中缺失的逻辑链条与推理过程，显著提升了模型处理复杂问题的能力**。通过思维链（COT）技术，可将简单的“问题-答案”对扩展为包含完整推理步骤的“问题-思考过程-答案”合成数据。

例如，在数学推理领域，通过为数学问题自动生成详细的解题步骤和逻辑推导过程，模型能够学习到分解问题、逐步求解的推理模式；在医疗诊断场景，可合成包含症状分析、鉴别诊断和最终结论的完整推理链条，训练模型进行多步临床推理。

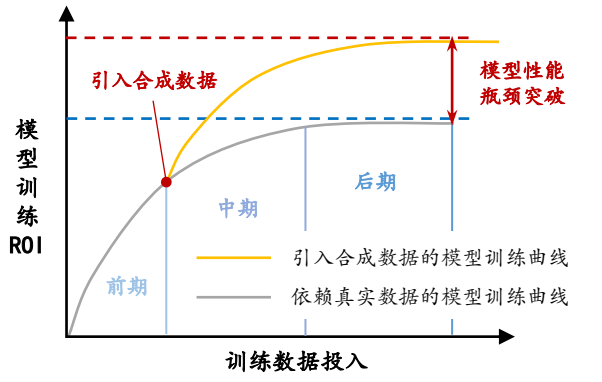
帮助模型掌握专业领域知识

领域里的专业理解是基础模型在产业中实际应用的最大门槛。各行各业都沉淀了大量非结构化的原始数据，如工业设备运行时序数据、医疗电子病历、科研论文图表等。但其格式复杂，模型难以直接学习。

而合成数据是将原始、庞杂的领域数据提炼为模型可直接吸收的结构化知识的关键工具，极大地降低了领域专业模型的应用门槛。利用大模型的理解能力将这些“生数据”转化为描述性文字或问答对话，可以合成高质量的领域特定训练数据集。**这为大模型在垂直领域的快速落地和专业化提供了可行路径。**

合成数据突破模型训练瓶颈

当模型性能提升进入平台期，单纯增加真实数据规模带来的边际效益递减，而合成数据能够提供更高阶的思维训练素材。引入合成数据不仅能显著提升模型处理复杂问题的能力，更能够**突破传统训练模式下的性能上限**，为模型实现更高层次的认知智能提供关键路径。



来源：弗若斯特沙利文

测试验证与风险控制

合成数据是**模型测试与验证阶段不可或缺的核心工具**，它通过模拟海量边缘案例和极端场景，为评估和提升AI系统的鲁棒性、安全性与可靠性提供了关键保障。在真实世界中，收集足够数量的罕见事件（如自动驾驶中的极端天气、工业设备中的罕见故障模式）数据进行测试，不仅成本高昂且极度危险。

合成数据完美解决了这一痛点。例如，自动驾驶公司通过合成数据模拟暴雨、传感器失效等 corner cases，在不上路的情况下对算法进行千万次压力测试；医疗AI厂商则利用合成的罕见病变影像，验证诊断模型的泛化能力，确保其在临床部署前的安全性。这极大地**降低了AI应用的实际风险，加速了其合规部署的进程**。

开辟前沿领域研究新范式

合成数据凭借其强大的场景模拟和生成能力，**为许多缺乏真实观测数据或进行实体实验成本高昂、风险巨大的前沿领域开辟了新的研究范式。**

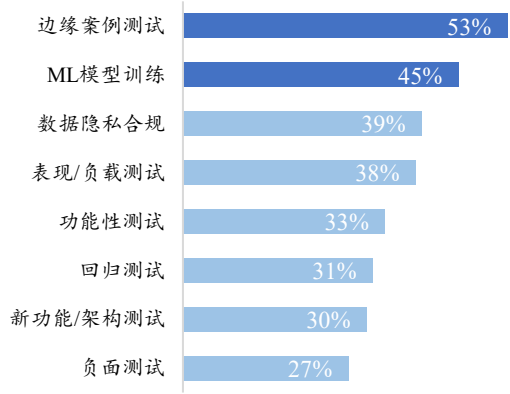
因此，合成数据不仅是AI模型的“可再生燃料”，更能成为推动科学发现和技术创新的“催化器”。

例如，在具身智能领域，采集真实数据需要搭建各种工作生活场景，耗时长成本高，使得技术研发速度严重滞后。现有的创新方式是通过人类佩戴头显等智能设备，采集人类真实运动数据用模拟框架做场景扩展，再用仿真工具做动作放大，**1次人类真实动作可以扩大到1000条量级的训练数据。**这样的方案可以低成本解决具身智能的数据稀缺，**增强模型的空间理解和动作能力**，已经在产业中广泛使用。

合成数据解决方案的主要用例有哪些？

软件测试和**ML模型训练**是合成数据最主要的两个用例。其他用例还包括：符合隐私的数据共享、产品设计和行为模拟等。据行业调查，超过53%的公司将边缘案例测试列为首要用例。

合成数据的主要用例



来源：弗若斯特沙利文

合成数据仍面临诸多局限性和挑战

一、真实性与公正性困境

合成数据的根本性挑战在于其与真实数据分布之间存在难以消除的**分布偏移**风险。尽管生成模型能复现宏观统计特征，但在高维数据的微观结构层面（如复杂特征交互、长尾分布模式等）仍存在显著差距。这种“过度清洁”的数据会导致模型在现实场景中出现**系统性的性能衰减**。

同时，合成数据在偏差控制方面可能面临**伦理困境**。生成算法不仅会继承原始数据中的偏差，更可能通过迭代生成过程强化这些偏见。基于有偏历史数据生成的合成样本可能产生较原数据更极端的分布偏差，使训练模型延续甚至加剧**歧视性决策模式**。

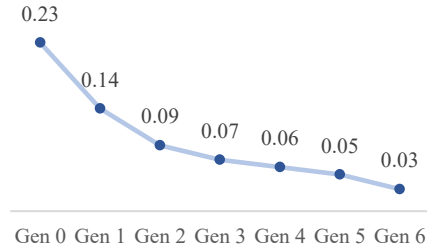
二、验证和采用难度

合成数据的有效性和可靠性面临着严峻的**验证挑战与信任壁垒**。验证合成数据并非简单比对统计指标，而需通过复杂的统计测试和领域专家评估来确保其不仅“形似”更“神似”真实数据。这一过程**成本高昂且缺乏行业标准**。因此，在医疗、金融等高风险、强监管的行业，决策者对采用合成数据持极度审慎的态度，严重阻碍了合成数据在关键任务中的推广应用。

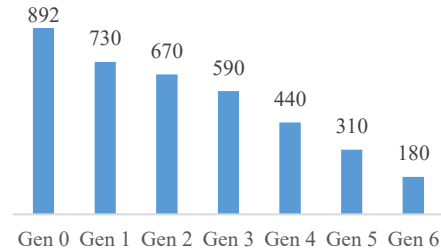
三、模型坍塌风险

长期使用合成数据训练可能导致**模型坍塌**，即**模型性能的渐进式退化**。其本质是由于生成过程中的**信息衰减**：当迭代使用前代模型的输出作为训练数据时，近似误差会不断累积，导致模型输出与真实分布产生显著偏离。模型坍塌的**早期阶段**表现为输出数据的复杂性和多样性下降；**晚期阶段**则出现系统性错误积累，模型开始曲解现实概念，泛化能力急剧恶化。

各代模型生成词汇的多样性趋势



各代模型生成的独特输出数量



实验结果显示，大型语言模型在迭代使用自身生成数据进行训练时，响应质量和多样性会持续下降。

如何控制合成数据的质量?

使用合成数据时，必须建立严格的质量监控机制和定期用真实数据“重新锚定”的更新策略，并结合完整的治理体系，否则既难以保证模型效能，还可能引发新的伦理危机。

合成数据的质量评估体系

评估合成数据在统计特性和数据结构上与真实数据的一致性与泛化能力。

- 统计相似性 / 分布相似性（例如直方图重叠、KS检验、Wasserstein 距离）：比对合成数据与真实数据在关键统计量和分布上的差异。
- 相关性保持：评估变量间的线性或非线性关系是否在合成数据中得到保留。
- 样本层级指标（如 α-Precision / β-Recall / Authenticity）:从生成样本多样性和泛化角度评估保真性与原创性。

评估合成数据在典型分析或模型任务中的效用是否可替代真实数据。

- TSTR与 TRTR对比：比较模型在两种训练数据上测试于真实集的表现差异，衡量实用性。
- FI Score：评估合成数据在特征重要性排序或权重方面，与真实数据是否一致。
- Q-Score：在合成与真实数据上运行多个聚合查询，检查结果一致性。

衡量合成数据在多大程度上可能暴露或复制真实样本，从而评估隐私泄露风险。

- DCR：衡量合成样本与真实样本的最小距离，距离越小可能存在复制风险。
- IMS：检查是否有真实记录被完全复制到合成数据中，理想值为 0。
- MIA风险：评估攻击者判断某条真实数据是否用于训练的可能性，是隐私风险的实证基准。

组织层面的合成数据治理体系

流程层

从需求分析、数据获取、模型生成到验证应用，明确责任主体、审批节点、决策机制，特别对高风险场景采用更严格的流程控制

技术层

选择合适的生成算法，设置参数，实施隐私增强措施，同时监控代表性、多样性与偏差控制等因素

评估层

持续开展统计验证、隐私测试、下游任务测试，确保合成数据满足预定义标准，形成常态化机制，并定期识别风险推动改进

反馈闭环

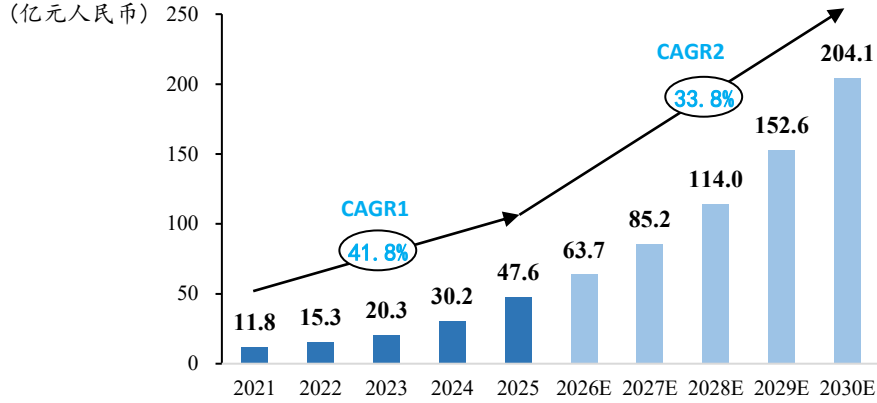
评估结果驱动策略调整、流程优化、模型迭代、权限重审，实现闭环治理，确保创新与合规并举

来源：弗若斯特沙利文

合成数据解决方案市场规模及渗透情况

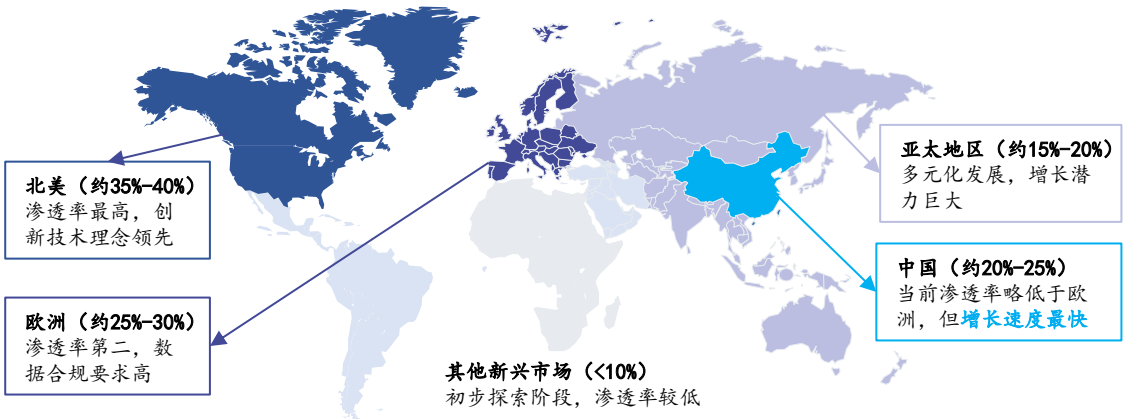
全球合成数据市场呈现爆发式增长态势。市场规模从2021年的11.8亿元人民币迅速扩张至2025年的47.6亿元人民币，期间年复合增长率高达41.8%。在AI技术迭代加速、数据安全要求提升以及成本效益优势凸显的多重驱动下，预计市场将保持强劲增长势头，2025-2030年复合增长率达33.8%，到2030年全球市场规模将突破200亿元人民币。

全球合成数据市场规模（按总收入计），2021年-2030年预测



得益于其成熟的技术生态、严格的数据法规以及早期积极的企业采纳，全球合成数据解决方案在**北美和欧洲的渗透率最高。中国市场增速最快**，由庞大的互联网用户基数、丰富的落地应用场景和强有力的政策支持驱动。亚太其他地区及新兴市场目前渗透率相对较低，但增长潜力巨大。

全球合成数据解决方案渗透率（按地区分类），2025



来源：弗若斯特沙利文

合成数据解决方案未来趋势

预计到2030年，AI模型中合成数据的生成量将超过真实数据的使用量。在工业场景或具身智能领域，未来的数据范式正朝着“人在环”的混合数据模式演进。

技术进步

新兴技术将彻底改变合成数据的生成，实现更高的**真实性、可扩展性和效率**。这些技术共同推动合成数据从“静态复制”向“动态演化”跃迁，极大拓展了其在复杂决策场景中的适用性。

- **先进AI模型**的进化不断提升生成数据的复杂度和规模，实现了跨领域（如医学影像、自动驾驶）的超现实数据合成；
- **量子计算**通过优化算法显著加速大规模数据生成过程，尤其在金融与物流等场景中增强了真实性与可扩展性；
- **数字孪生集成**则通过高保真模拟现实系统与环境，为预测建模与边缘场景测试提供了动态且精准的数据基础。

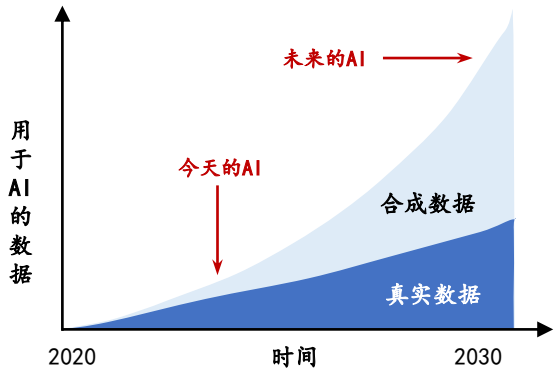
“人在环”的混合数据模式

当前，工业级AI训练严重依赖标注成本高昂的真实数据，且难以覆盖关键边缘案例。未来的数据范式正朝着**“1%人类数据+99%高效合成”**的混合模式演进：以少量高质量、经过严格标注的人类数据作为种子，驱动AI生成大规模、富含关键挑战性场景的合成数据。其成功依赖于**“Human in Loop”（人在环）**机制——领域专家通过介入数据筛选、规则定义与质量评估，确保合成数据的高价值与可信度。

最终，这种范式将构建一个远超纯人工标注规模与覆盖度的动态数据池，为核心行业的高可靠性AI训练提供关键解决方案。

合成数据用量将超过真实数据

鉴于对更高质量和更注重隐私保护的数据源的需求，**预计到2030年，人工智能模型中合成数据的生成量将超过真实数据的使用量**。未来，合成数据预计将支撑大多数AI分析项目，推动成熟领域和新兴领域的创新。



来源：弗若斯特沙利文

第三章

合成数据解决方案应用场景分析

合成数据解决方案应用场景总览

1 自动驾驶

- **动极端与长尾场景训练：**利用合成数据模拟暴雨、夜间、路口突发冲突等稀缺场景，帮助算法在极端条件下保持鲁棒性
- **大规模仿真测试验证：**通过批量生成虚拟路测场景，在不增加真实路测成本和风险的前提下，加速模型的验证迭代

2 具身智能

- **多模态交互学习：**基于合成的视觉、力觉、动作轨迹等数据，训练机器人完成抓取、搬运、平衡等复杂动作
- **长尾任务泛化：**通过仿真平台批量生成跨物体材质、不同摩擦系数和环境条件的交互数据，提升模型的泛化能力

3 工业

- **虚拟产线优化：**利用合成数据重建生产线，模拟不同产能配置与设备工况，优化装配、搬运等环节的效率与稳定性
- **危险工况模拟：**通过生成设备故障、操作失误等真实采集中难获得的数据，用于训练与测试质检与安全监测模型

4 金融

- **极端情景策略验证：**通过合成数据模拟极端行情，金融机构可测试交易策略、风控韧性，提高极端条件下的决策稳定性
- **反欺诈训练：**利用合成数据生成虚拟交易记录训练反欺诈模型，在无需接触真实客户数据的情况下，显著提升准确率

5 医疗

- **隐私保护下的AI训练：**基于合成电子健康记录，企业可以在不暴露真实患者信息前提下，训练出疾病AI模型
- **加速新药创新：**企业使用合成数据缩短新药临床试验设计周期

6 游戏

- **虚拟场景生成与扩展：**利用合成数据批量生成多样化的虚拟关卡、环境与NPC行为，支持游戏测试与AI对手训练，提升沉浸感与可玩性
- **玩家行为建模与对抗训练：**基于合成玩家行为数据，模拟不同策略、反作弊场景和极端玩法，从而优化平衡性与安全性

来源：弗若斯特沙利文

合成数据解决方案如何重塑垂直行业的未来?

合成数据解决方案弥补了真实数据的局限，推动垂直行业突破数据瓶颈

面对真实数据采集成本高、隐私风险大、极端场景稀缺等瓶颈，合成数据不仅能提供规模化的数据生产，也能通过真实性校验与经验流闭环，确保与真实世界保持一致，从而能够批量覆盖长尾与极端环境，同时兼顾高效迭代，助力垂直行业领域突破数据瓶颈，加快智能化发展。具体而言，重塑垂直行业体现在以下六个维度：

1

数据规模与覆盖度

真实采集往往覆盖不到这些极端情况（如严重事故、危险工况），合成数据可以批量生成大量样本，尤其是极端、稀缺、长尾的场景。

2

隐私与合规

医疗和金融等行业受隐私监管约束，不能随意共享真实数据，通过合成数据替代或补充真实敏感数据，降低隐私泄露和数据出境风险。

3

实时联动与数字底座

合成数据与实时仿真、数字孪生、标准化场景描述结合，不仅能做离线训练，还能实时推演、辅助决策，形成行业的数字底座。

4

效率与成本

很多行业的真实数据采集需要大量人力/时间/费用，效率瓶颈严重，通过并行仿真、自动标注、神经渲染等技术，高效生成训练数据，降低采集与标注成本。

5

数据新鲜度与闭环

传统依赖真实采集的数据体系往往是静态的、滞后的，难以跟上环境和业务的快速变化。合成数据通过闭环机制，可以形成动态更新、持续进化的训练数据。

6

安全韧性

现实中攻击与欺诈行为不断进化，必须提前用对抗性数据训练模型。用合成数据构造对抗性样本和复杂干扰场景，提高AI系统的稳健性。

来源：弗若斯特沙利文

合成数据解决方案行业应用场景分类

由于规模化采集高质量真实数据存在壁垒，合成数据是实体物理驱动应用场景的关键基础。

为更清晰理解合成数据对于垂直行业的价值，应用场景可以划分为**实体物理驱动**、**信息数据驱动**两大类。

在信息数据驱动应用场景中，合成数据通过生成安全、可控且符合逻辑分布的样本，既能保障敏感数据在合规框架下被充分利用，也能扩展虚拟环境中的创新能力。

在实体物理驱动场景中，合成数据能够在仿真环境中模拟物理规律与真实场景，并批量覆盖长尾与极端情况，为提升可靠性与性能提供基础。

总体而言，由于规模化采集高质量真实数据存在壁垒，实体物理驱动应用场景对合成数据的依赖更为显著；而在信息数据驱动应用场景中，合成数据的价值更多地体现在合规、安全和效率。

合成数据解决方案两类行业应用场景

实体物理驱动

以真实物理环境和多模态交互为核心的行业场景

- **特征：**交互复杂、采集困难、长尾稀缺
- **核心需求：**合成数据模拟物理规律与真实场景，覆盖长尾鱼极端情况

典型行业

自动驾驶　具身智能　工业

信息数据驱动

以数据本身的敏感性、合规性

- **特征：**隐私保护、共享受限、合规约束
- **核心需求：**合成数据生成符合逻辑替代样本，实现隐私保护与虚拟环境扩展

典型行业

金融　医疗　游戏

来源：弗若斯特沙利文

自动驾驶：安全是底线，数据是前提

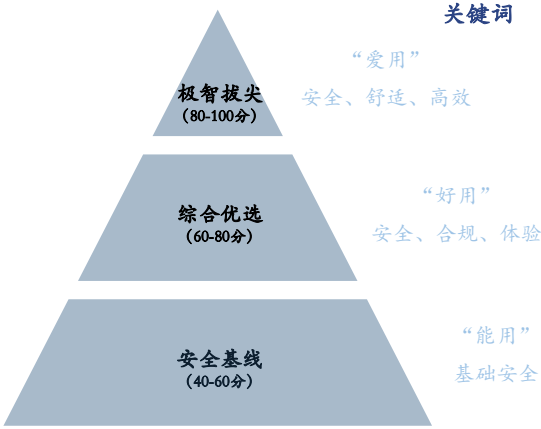
“事故风险哪怕只有1%，一旦发生，就是100%的伤害”

高阶自动驾驶落地面临的核心挑战是系统鲁棒性与安全保障，并在安全与性能之间取得平衡，从而提升驾驶体验。

具体而言，车辆必须能够在常规与极端事件下保持可靠驾驶。这关系到乘客和驾驶员、道路车辆和行人的生命安全，甚至对整体交通运行产生影响。同时，还需要兼顾驾驶体验，例如雨天安全变道时，既能保持驾驶顺畅，又避免被迫长时间跟随堵塞车流。

这要求自动驾驶算法形成多样化场景的适应能力，关键在于**高质量、大规模且覆盖全面**的数据支撑。

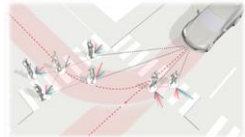
智能驾驶金字塔分级测试体系



代表性场景举例

极端天气下动态避障

- 强干扰环境中（如暴雨/大雾），多类型障碍物行为不可预测，且能见度骤降、路面摩擦突变等引发连锁反应。



无规则路口群体博弈

- 参与者遵循隐性社会规则导致多向交通流意图冲突，弱势道路使用者（如非机动车/行人）行动难以预判。

天气多变场景

- 当天气短时间出现晴天转暴雨再回晴天，路面条件等出现快速变化，容易带来感知偏差。

不同城市路况差异

- 城市间交通特征差异明显，常规采集数据覆盖有限。仅依赖局限场景训练，难以让模型在跨城市、多环境下保持稳定表现。

合成数据解决方案如何提升安全覆盖面?

通过持续与环境交互生成经验数据，并利用AI增强仿真，合成数据解决方案能以规模不足盲区，以泛化覆盖长尾场景。

目前，主机厂的竞争将取决于模型后训练与持续训练的质量，仅依赖真实数据的仿真模型训练存在两个方面的问题：

- **泛化能力受限：**在跨城市、跨区域、跨气候条件下，算法难以保持一致的安全表现
- **长尾风险难以暴露：**极端天气、不规则路口、弱势交通参与者等长尾场景，在有限真实数据中出现频率极低，难以充分训练与验证。

合成数据在自动驾驶场景中的核心价值，不在于取代真实数据，而在于**放大真实数据的价值、扩展人类示范的边界**，为高阶自动驾驶提供全面的安全保障。

一方面，合成数据解决方案能够以少量真实场景数据为起点，借助仿真与AI生成技术不断拓展出动态场景，并通过持续迭代形成多样化的训练样本，解决数据规模和覆盖面不足的问题。

另一方面，解决方案通过智能体与环境持续交互，将经验反馈融入仿真，并结合真实数据优进行化，从而实现场景参数化建模，增强仿真场景与数据的泛化价值。

合成数据解决方案在自动驾驶场景的模式

真实/实时数据采集	仿真泛化	自动生成	持续反馈
用实车、实地采集的数据，如暴雨路面	暴雨路面参数化扩展，如车速、水深、装载量、雾浓度	通过AI对抗等方式，主动制造冲突或长尾复杂场景与动态数据	实际运行数据反哺解决方案，不断丰富经验库

具身智能：成长在即，数据待补

借助合成数据扩展规模、提升质量、丰富多样性，是具身智能发展的必经之路。

具身智能的发展仍处在早期阶段，行业整体缺乏可供预训练的大规模数据。相比视觉或语言 AI，具身智能需要处理更复杂的物理与动作信息：既包含图像、语音指令等感知数据，也涵盖动作轨迹、力觉、接触反馈等多模态信号。

同时，不同构型的机器人（单臂、双臂、人形）在参数和动作方式上差异明显，使得通用数据集难以直接复用。

现实中采集数据不仅成本高昂，且场景复杂、难以覆盖，例如在公共空间中存在的人流密集、不规则交互，这些情况在真实采集里往往不可控。

即便采集到一定量的数据，也因规模有限、场景单一，难以满足算法对多样性和泛化的需求，因此很难仅依靠真实采集获得高质量数据。

具身智能当前面临的数据挑战

来源：弗若斯特沙利文

合成数据解决方案如何护航具身智能成长?

数据规模与多样性是基础，数据质量与真实一致性则是具身智能迭代升级的关键。

纯粹依赖合成数据并不能彻底解决问题：如果缺乏与真实场景的对照与校正，容易出现“Garbage In, Garbage Out”，生成的数据和真实需求脱节。

如果仅依赖单一真实采集或者纯合成的数据输入，就会形成封闭的自我循环，缺乏外部反馈与校正，难以真正推动具身智能能力的提升。

因此，合成数据需要与真实采集数据相结合，并通过人类专家、真实反馈和持续验证，不断校正和更新仿真环境及合成样本。

现阶段，为了护航具身智能的成长，合成数据解决方案关键在于提升数据的**真实性、新鲜度、规模、多样性与覆盖度**，以确保模型既能在真实物理规律下稳健表现，又能在复杂多变的长尾场景中保持泛化能力。

合成数据解决方案在具身智能场景的关键维度

数据真实性

作用：保证合成数据与真实世界一致，避免偏差累积

解决方案：

- Real to Sim to Real：从真实人类演示与传感器回放提取参数，再在仿真中扩展
- 真实性检验：对比真实分布，做真实性&有效性评估

数据新鲜度与纠错

作用：保持数据“新鲜”、持续校正偏差，形成动态经验库

解决方案：

- 经验流持续更新：智能体交互驱动，场景随算法动态调整，自动生成符合物理规律的新数据
- 人类监控标注：专家对难例和错误数据做人工审阅和校正

数据规模与效率

作用：解决真实采集成本高、进度慢问题，低成本快速产出大规模样本

解决方案：

- 经验流持续更新：智能体交互驱动，场景随算法动态调整，自动生成符合物理与时序规律的新数据
- 人类监控标注：专家对难例和错误数据做人工审阅和校正

数据多样性与覆盖度

作用：大幅拓展训练样本空间，提高模型泛化能力

解决方案：

- 仿真扩展：参数化生成跨物体材质、力学属性、环境条件的数据
- 多智能体对抗：制造冲突/边界条件，例如拥挤人流、抓取物体失败

来源：弗若斯特沙利文

工业行业：搭建数字孪生平台，加速数字化转型

合成数据是推动工业行业释放数字孪生价值、实现数字化转型的关键补充。

工业行业正通过**“数字孪生+仿真驱动”**的路线进行数字化转型，以提升运营效率、优化成本与增强安全保障：

- **制造业：**企业通过工厂数字孪生开展产能推演和工艺验证，提升效率并降低改造成本；
- **能源与电力行业：**核电、风电等场景可以通过仿真开展设备巡检、极端工况模拟和应急演练，减少人工风险与投入；
- **采矿行业：**矿山作业环境复杂且危险，传感器数据难以规模采集，合成数据和仿真能够训练无人矿卡和安全检测系统

数字孪生的价值的释放需要持续高质量的数据流。然而，仅依赖真实数据，会面临难以采集（如设备极限负荷）、实验无法落地（如危险事故演练）、数据稀缺（如新工艺早期数据）等局限。

因此，合成数据是必不可少的补充，其能够在虚拟产线中模拟**复杂工艺**与**极端工况**，批量生成可控、可复现的训练与验证样本，从而让数字孪生不局限于静态展示，而是能驱动工业优化迭代。

工业应用场景的真实数据为什么不够用？

长尾/极端工况稀缺

制造业难以采集罕见缺陷，能源与采矿则难以捕获极端天气、地质灾害或设备事故。

跨场景差异大

不同工厂、电厂、矿区的条件差异显著，导致真实数据迁移性差，需要频繁重新采集。

采集与安全成本高

在工业生产线、核电站或矿山中做停机或破坏性实验几乎不可行，风险和成本双高。

安全与合规

在危险工况、易燃易爆或辐照环境中进行真实采集有安全与合规红线，只能通过虚拟仿真补齐。

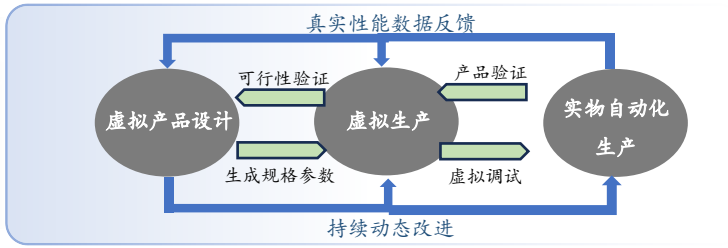
来源：弗若斯特沙利文

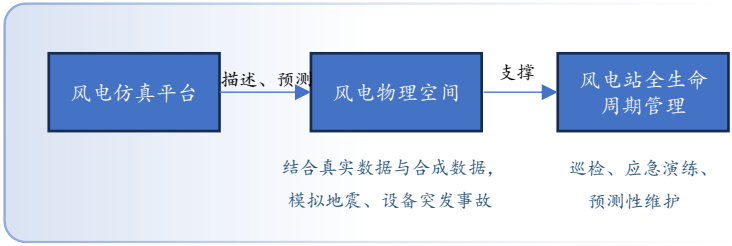
合成数据解决方如何加速工业数字化转型？

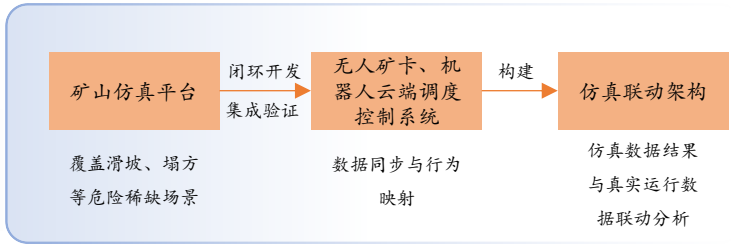
合成数据能为工业行业提供安全、低成本、可扩展和高效的虚拟数据工厂

在不同工业场景中，合成数据通过与数字孪生和仿真平台的结合，能加速数据流的生产与应用：

- **制造业：**合成数据补足工艺验证、质量检测和极端工况数据，使得产线升级和工艺改造能够在虚拟环境中提前验证，降低试错成本；
- **能源：**以风电为例，合成数据支持全生命周期管理，助力巡检、应急演练和预测性维护。通过极端天气和设备故障场景的生成，帮助企业在不增加安全风险情况下，提升可靠性和韧性；
- **采矿：**合成数据生成滑坡、塌方等危险稀缺场景，用于训练无人矿卡、巡检机器人和调度系统。在数据闭环联动架构下，仿真数据与真实运行数据相互检验，支撑矿山作业的智能化和安全。







趋势：自动驾驶为底，延伸至具身智能与工业场景

自动驾驶实践提供规模化的技术验证，具身智能与工业场景将承接并扩展其技术可迁移价值。

自动驾驶、具身智能和工业场景都属于实体物理驱动场景，需要在高复杂度环境中实现可靠决策。因此，对数据的要求具有一致性：

- 需大规模的长尾与极端场景覆盖；
- 需模态交互数据（视觉、力觉、语义等）；
- 需要模拟多主体协同（车辆、机器人、工业设备的写作）；
- 依赖仿真—真实的闭环验证，确保模型能迁移到现实

这种共性使得自动驾驶积累的经验能够向具身智能和工业进行迁移。

同时，在自动驾驶领域经过打磨的解决方案，具备高保真场景重建、多传感器精准同步、动态交互行为建模等技术能力。这些是支撑具身智能与工业中智能体或平台与环境持续互动的关键要素。更重要的是，自动驾驶场景商业化验证的成果，也为跨领域拓展显著降低了风险。

因此，做好自动驾驶合成数据，是进入具身智能的“必要不充分条件”。

合成数据解决方案核心技术模块复用

技术模块	自动驾驶	具身智能/工业	关联性
合成数据生成	生成极端天气、事故场景的图像与点云数据	生成机器人抓取不同材质工件的合成图像、矿山暴雨场景数据	均需解决真实数据稀缺问题
多传感器同步采集	同步摄像头、激光雷达、GPS数据	同步机器人的视觉传感器、力控传感器、关节编码器数据	均需保持数据有效性
多智能体行为建模	模拟其他车辆、行人的自然交互行为	模拟港口集卡与堆高机的协同作业、核电巡检机器人路径规划	均需模拟多主体协同
高保真仿真引擎	模拟道路、天气、交通流，测试车辆决策	模拟机器人操作空间、矿山地形	均需还原物理世界规律

从自动驾驶向具身智能转型的关键是什么？

精准模拟物理交互、丰富且高质量的合成数据是企业从自动驾驶向具身智能成功转型的关键。

企业从自动驾驶向具身智能切入时，其核心挑战是从一个规则相对明确、以“移动”为核心的封闭问题，跨越到一个以“交互”为核心的开放世界问题。企业必须克服从“第三人称”环境观测到“第一人称”具身交互的认知鸿沟，处理从大规模纯视觉数据到多模态物理交互数据的复杂性，并实现从特定任务驱动到通用认知与泛化的能力升级。

因此，**精准模拟物理交互、丰富且高质量的合成数据**已成为决定企业转型成败的关键基础设施，是从数据源头破解具身智能训练难题的核心支点。

综合来看，在从自动驾驶向具身智能转型过程中，企业需要关注以下核心能力的补充：

- **扩大合成数据能力范围：**不仅限于视觉模拟，还应包括触觉、力反馈、多模态融合等复杂交互要素，同时支持动态Agent学习过程
- **构建高保真的模拟环境：**采用高质量物理仿真引擎，支持柔性材料、复杂环境、动态变化场景模拟，减少与现实环境差距，保证Sim-to-Real 的有效迁移
- **支持高语义层次的数据注释：**例如关系推理、策略目标、因果场景说明等，以支撑认知推理能力的训练，弥补从自动驾驶到具身智能的认知和行为差距

合成数据解决方案的能力需求差异

技术模块	自动驾驶	具身智能/工业	转型需补充技能
几何与物理真实性	关注外观、天气、光照变化和基本障碍物位置，强调外观与环境的视觉保真	需精确生成物体的3D几何形状、质量、摩擦系数、刚度等物理属性	高保真物理仿真建模能力
多模态数据同步	主要关注视觉（摄像头、激光雷达点云）和毫米波雷达等	同步生成视觉、深度、力觉、触觉甚至音频模态的数据，并且需要多模态数据的时空对齐	多模态融合感知仿真
交互与动态模拟	主要关注交通参与者（车辆、行人等）的动态序列生成	不仅需要生成静态场景，更要能模拟复杂接触交互的动态序列	高保真接触力学模拟、人机交互知识
场景与语义多样性	仅模拟交通场景与其他相关长尾场景	需要生成海量多样化的长尾家居、办公、工业场景，且带有丰富的语义标签	大规模场景图谱与知识图谱构建

来源：弗若斯特沙利文

第四章

合成数据解决方案产业链分析

中国合成数据解决方案产业链图谱

上游：数据生产与基础支撑

上游环节涵盖了硬件与软件两大支撑领域：

- 传感器决定真实数据采集的精细度与可靠性，而芯片则是保障仿真模拟与数据生成的算力基础
- 数据管理、数据标注与数据安全构成了合成数据的治理底座

中游：合成数据解决方案

合成数据解决方案竞争特征在于技术迭代快、行业Know-how门槛高、生态兼容性要求高。这三个方面决定了供应商能否实现跨行业迁移与规模化落地。

其中，技术迭代速度决定了应对复杂多变、快速演进行业场景的能力；行业Know-how则影响着解决方案的迁移性与落地深度；生态兼容性不仅关系到规模化与商业化，更关乎供应链安全和稳定。

下游：垂直应用领域

在生成式AI快速发展与数字化转型驱动下，垂直行业对数据的需求不断凸显，规模化落地的潜力正加速释放。经过实践打磨与技术迭代与融合，合成数据解决方案在这趋势下迎来广阔商业化与应用前景

上游是合成数据的输入、算力与治理质量的关键基础

上游硬件与软件环节是影响合成数据解决方案的数据质量与治理效率的关键因素。

传感器与芯片，数据管理、标注与安全决定合成数据解决方案能否实现规模化落地的根基。

一方面，传感器与芯片决定合成数据的输入与算力基础。只有在“入口”和“算力”层面保证精度与效率，中游才能构建出足够逼真、可扩展的合成数据场景。

另一方面，数据管理、标注与安全决定了数据的治理与质量。如果这些环节薄弱，合成数据解决方案会面临数据源不足、治理效率低，同时也难以支撑其在下游垂直行业领域的拓展。

因此，具备充分软硬件兼容能力或整合能力的解决方案商，才能构建起稳定、高效、可扩展的合成数据体系，并真正支撑规模化落地。

上游技术对数据质量与治理效率的决定性作用

来源：弗若斯特沙利文

合成数据供应商主要有哪些发展路径？

三类供应商中，专注解决方案型展现出更强的延展性与商业化潜力

目前具备合成数据解决方案能力、能够全方位赋能用户构建、使用与优化合成数据的供应商有限。整体格局可以分为三类：专注解决方案型、硬件驱动型、仿真平台型。

相比于其他两类，**专注解决方案型供应商**不仅具备开放的软硬件兼容性能力，还基于自动驾驶的经验积累，打磨形成了具有行业迁移性的工具链和技术架构。

这一优势能够使其更容易突破场景限制，向具身智能、工业等更复杂、更高要求的应用场景延伸，形成从数据采集、仿真生成到应用优化的完整解决方案闭环。

专注解决方案型供应商

深信科创 SYNKROTRON

产品：Oasis Rover, Oasis Data, Oasis Sim，Oasis Bot

- 面向行业用户提供一体化工具链与软硬一体方案
- 从自动驾驶起步，解决方案已在众多主机厂落地
- 向具身智能与工业领域扩展，将推出基于新一代机器人操作系统Dora的全国产人形轮式机器人Oasis Bot
- **技术特点：**通过经验流闭环与持续学习，生成高物理保真精准性的合成数据；强调智能体持续与环境互动生成数据、支持场景与算法反馈实时耦合优化

光轮智能 Lightwheel AI

产品：基于Isaac Sim 与Omniverse的合成数据平台

- 基于上游软件，结合自身创新架构提供解决方案
- 起点在自动驾驶，逐步拓展至具身智能
- **技术特点：**基于“Real2Sim2Real + Realism Validation”架构，强调人在环与仿真结合，突出真实数据与合成数据的互补，同时提供合成数据的真实性评测与效用性评测平台

硬件驱动型供应商

NVIDIA

产品：Omniverse, Drive Sim, Isaas Sim

- 依托GPU硬件和CUDA生态，向下延伸至仿真、数据生成、模型训练，构建端到端方案
- 生态对硬件依赖度高，灵活性有限

仿真平台型供应商

松应科技 CGAAS

产品：ORCA Studio, ORCA仿真一体机

- 以工业场景起步构建仿真与虚拟训练场能力
- 支持不同厂商GPU硬件部署，并具备分布式多GPU协同运算，支撑仿真平台实时处理效率与大规模应用场景需求

下游应用商业规模化的关键

数据、迭代、平台能力是合成数据解决方案赋能垂直行业重塑工作流与价值链的关键。

合成数据解决方案下游应用场景的价值主要可以总结为预测与生成两个方面：

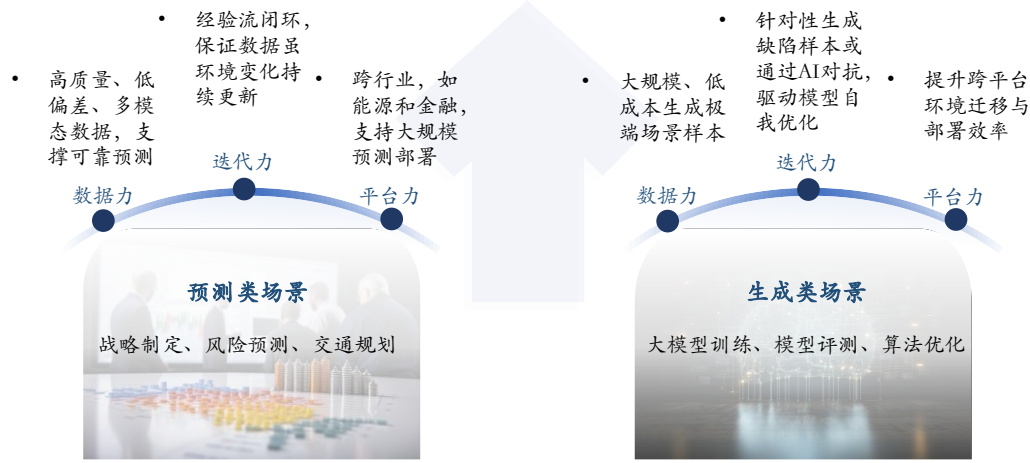
- **预测：**通过仿真与数据生成，模拟市场、工况环境等未来变化，价值在于定制与创新灵活度高、更新速度快和干扰与偏见少；

- **生成：**面向AI模型的训练与测试，生成大规模数据训练、测试评估和持续学习优化，价值在于数据体量大、降低数据获取难度、成本低。

推动预测与生成的价值实现商业化规模落地，合成数据解决方案需具备以下能力：

- **数据能力：** 覆盖长尾与极端场景，提升决策可靠性；确保仿真数据与真实世界高度一致，支持跨场景迁移；
- **迭代能力：**在“真实—仿真—真实”的闭环中不断引入新场景与错误样本；将人类示范与生成式 AI 相结合，减少偏差与过拟合；
- **平台能力：**构建标准化数据生产与交付流程，保证一致性；打通自动驾驶、具身智能、工业等行业场景的复用价值；提升生态适配能力，提升商业落地效率。

预测与生成场景落地需求：数据力×迭代力×平台力



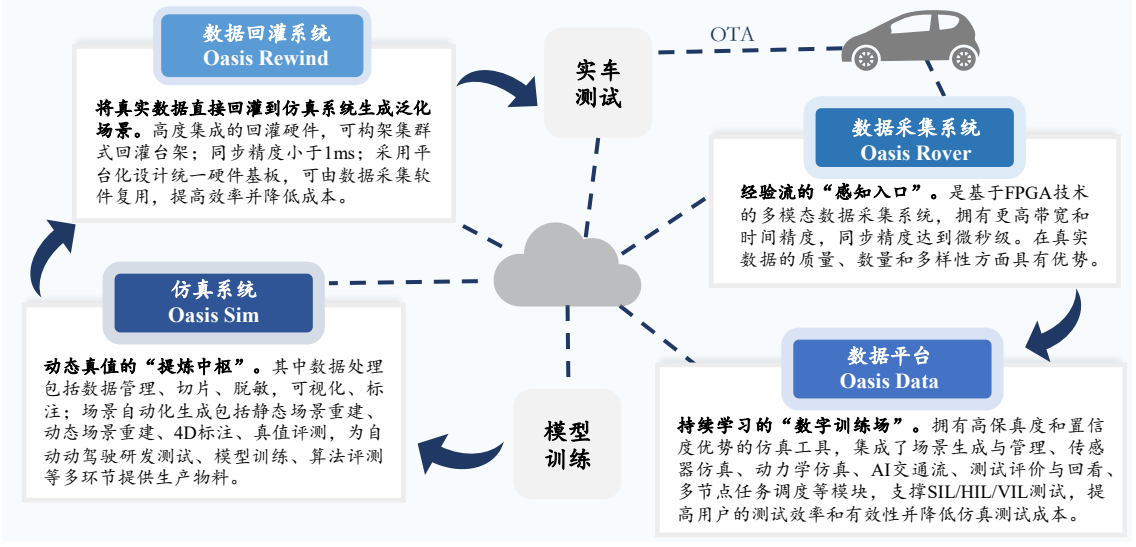
来源：弗若斯特沙利文

第五章

合成数据解决方案最佳实践

高价值的合成数据资产提供商——以物理真实数据为“种子”

核心产品的整体协同：从“数据闭环”到“持续学习闭环”



场景应用及成功案例：从自动驾驶向具身智能和工业场景拓展

自动驾驶

深信科创为国内外头部车企的自动驾驶系统测试验证提供极端场景数据支持和虚拟仿真测试平台，目前已经获得多家权威客户的认可。

长安汽车 CHANGAN AUTO　吉利汽车 GEELY AUTO　WEICHAI 潍柴　CHERY 奇瑞汽车　ZF　Volkswagen　理想　广汽集团 GAC GROUP

工业场景

为能源制造等工业企业提供高保真运维模拟支持和高风险作业虚拟仿真环境。

中广核 CGN　深信科创**工业数字孪生平台**已中标中广核项目，支撑核电站全生命周期管理。

具身智能

FUTUREWEI Technologies　深信科创与Futurewei合作出了**新一代基于数据流框架的机器人操作系统Dora**。

Dora自开源起已迅速引发国际AI开源社区的广泛关注，目前已经用于Hugging Face开发的LeRobot。此外，深信科创即将推出基于Dora的**全国产人形轮式机器人Oasis Bot。**

“Dora是对ROS（过去十几年国际最流行的开源机器人操作系统）极具竞争力的替代方案。”
——Hugging Face联合创始人兼首席科学家

Hugging Face

三者共享“经验流”“闭环学习”“数字孪生”的底层逻辑

来源：弗若斯特沙利文

光轮智能 Lightwheel AI

光轮智能

合成数据技术的创新应用者

核心技术栈

Real2Sim2Real ＋ **Realism Validation**

- **融合生成式AI与仿真引擎**，生成3D混合渲染、物理真实、高度交互的合成数据，并提供以数据为中心的端到端解决方案。
- 在数据**Pre-train（预训练）**与**Post-train（后训练）**阶段均有布局，既能支持基于模仿学习的模型训练，也能通过闭环仿真构建基于Self-play RL（自我对弈强化学习）的训练与评测新范式。

场景应用及成功案例

自动驾驶

为多家国内外头部主机厂和Tier 1供应商提供合成数据服务，包括为中国自主品牌出海提供数据支持，以及提供Corner Case（长尾场景）数据。

具身智能

为行业提供了高质量的具身智能训练资源，促进了行业技术共享与进步。

智元机器人

光轮智能与智元机器人达成战略合作，为其提供具备高保真物理属性的仿真资产，智元在此基础上构建并发布了**公开数据集Agibot Digital World**。

工业落地

全套合成数据支持

提供包括遥操作行为数据、高保真仿真场景与交互资产在内的全套合成数据

构建仿真环境

基于NVIDIA Isaac Sim与Omniverse平台，构建物理交互真实、场景多样化的仿真环境，模拟汽车工厂复杂任务场景

Real2Sim2Real技术架构

通过“人在环”的仿真遥操作生成数据，有效缩小了仿真环境与物理现实世界的差距，确保这些遥操数据能迁移至不同机器人本体

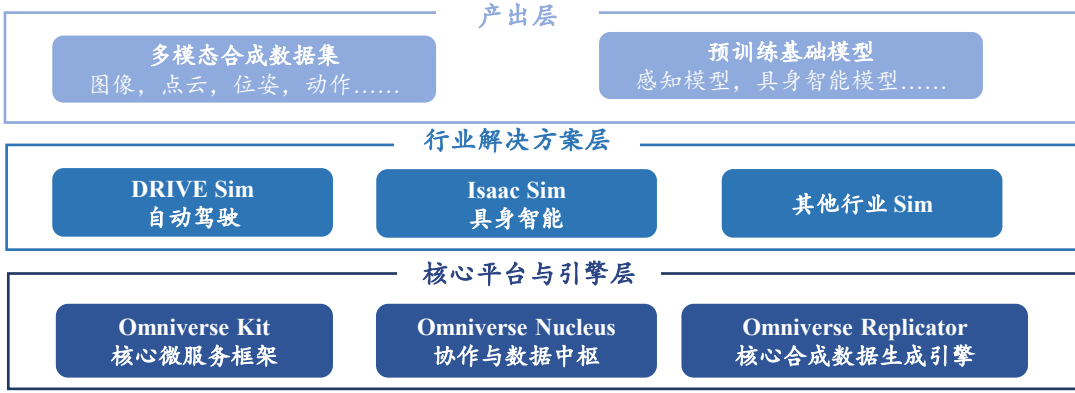
光轮智能为**英伟达GR00T N1人形机器人**模型提供全套合成数据支持，加速了全球首个通用人形机器人开源基础模型在汽车制造生产线的首次应用实例，展示了在复杂工业任务中的可靠性和适应性。

英伟达

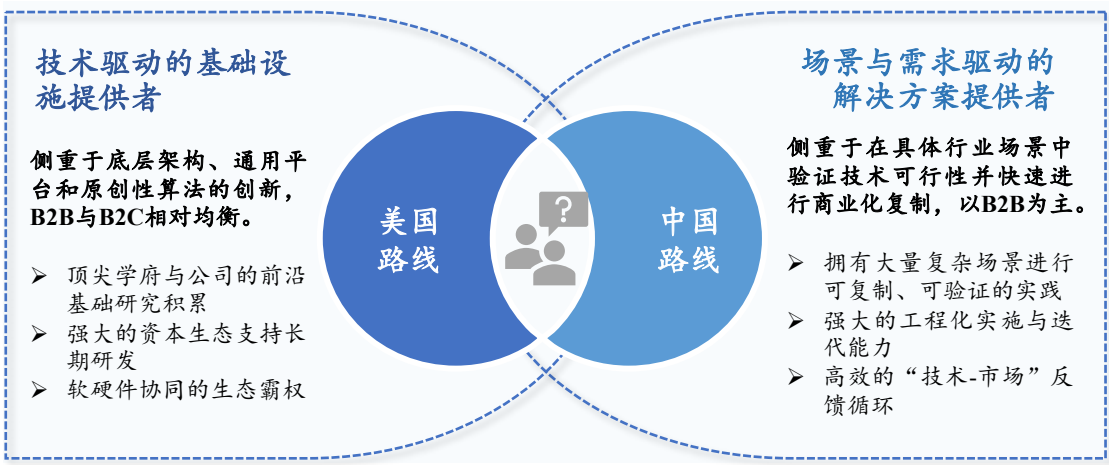
合成数据的重要基础设施提供商

产品与技术架构

英伟达构建了一个综合的**虚拟世界模拟和协作平台**（Omniverse），作为各种AI应用（如自动驾驶、机器人、工业数字化等）的底层基础设施。他们提供强大的工具和平台，吸引和赋能如光轮智能、深信科创等合作伙伴和客户在其之上开发具体的应用和解决方案。



合成数据解决方案的技术路线对比：美国 VS 中国



来源：弗若斯特沙利文

术语表

术语	解释
AI (人工智能)	让机器模拟人类智能行为的科学与工程。
ML (机器学习)	AI的一个子领域，计算机通过数据自动学习并改进，而无需显式编程。
LLM (大语言模型)	基于海量数据训练的、能够理解、生成和处理自然语言的巨大模型。
GAN (生成对抗网络)	一种通过生成器和判别器相互博弈来创造逼真数据的深度学习框架。
VAE (变分自编码器)	一种使用编码器-解码器结构来学习数据分布并生成新数据的生成式模型。
后训练 (Post-training)	在预训练模型基础上，针对特定任务进行额外训练（微调）以适应新任务。
持续训练 (Continuous Training)	模型部署后，持续用新数据更新以提高其性能和适应性的过程。
微调 (Fine-tuning)	一种迁移学习方法，通过调整预训练模型参数来快速适应新任务。
人在环 (Human in Loop)	将人类专家的判断和反馈纳入AI系统的训练或决策循环中，以提高可靠性。
Corner Case (corner case)	罕见但关键的极端场景，对自动驾驶等系统的安全性和鲁棒性构成巨大挑战。
数字孪生 (Digital Twin)	物理对象或系统的动态虚拟副本，通过实时数据同步和仿真来辅助决策。

术语表

术语	解释
GDPR & HIPAA	欧洲的《通用数据保护条例》和美国的医疗数据安全法规，规范数据隐私和处理。
DCR (Distance to Closest Record)	衡量合成数据与最近真实样本的相似度，距离过小暗示隐私泄露风险。
IMS (Identical Match Share)	检查合成数据中是否存在与真实数据完全相同的记录，理想值为零。
MIA (Membership Inference Attack)风险	评估攻击者能否推断出某条真实数据是否被用于模型训练的概率。
NNDR (Nearest Neighbor Distance Ratio)	通过计算与最近和次近真实样本的距离比值，来判定合成数据是否过于接近某个真实个体。
TSTR（Train on Synthetic, Test on Real）/TRTR （Train on Real, Test on Real）对比	核心评估方法：比较“用合成数据训练”与“用真实数据训练”的模型在真实数据上的表现差异。
FI Score (Feature Importance) 特征重要性	评估合成数据与真实数据在特征重要性排序或权重上是否一致。
Q-Score (查询一致性评分)	通过在合成与真实数据上运行相同查询来检验结果的一致性，衡量数据统计效用。

关于我们

弗若斯特沙利文是一家能够协助企业实现高速增长，优化投融资渠道，解决企业发展瓶颈的跨国咨询公司。我们凭借遍布全球 45 个办事处的增长顾问团队，精准识别数百个行业的增长机遇，积累了对全球价值链运作方式的深刻洞察。弗若斯特沙利文创新的市场进入策略和经过验证的最佳实践实施方案，已通过我们的“增长即服务”管道 (Growth Pipeline as a Service)，助力众多世界领先企业成功实现商业模式转型，使客户能够轻松创建并实施源源不断的增长机遇。

弗若斯特沙利文处于一个独特生态系统的中心，该生态系统融合了最佳实践辅导、高管同行支持社群和以增长为导向的内容。我们始终一心一意致力于通过有管理的增长重塑世界。