



集成电路大模型：驱动集成电路制造技术跨越发展

孙奇¹, 邓锦堦¹, 金谦¹, 李渝¹, 严哲雨¹, 高大为^{1,2,4}, 熊诗圣³, 吴振华¹, 卓成^{1*}, 许宁生³, 吴汉明¹

1. 浙江大学集成电路学院, 杭州 311200

2. 浙江创芯集成电路有限公司, 杭州 311200

3. 复旦大学纳米研究院, 上海 200433

4. 浙江省 CMOS 集成电路成套工艺与设计技术创新中心, 杭州 311200

* 通信作者. E-mail: czhuo@zju.edu.cn

收稿日期: 2025–10–05; 修回日期: 2026–01–05; 接受日期: 2026–03–24; 网络出版日期: 2026–09–08

国家自然科学基金 (批准号: W2412034, U25A20486, 62331009, 62306093)、国家重点研发计划 (批准号: 2024YFA1209000) 和浙江省自然科学基金 (批准号: D24F040002) 资助项目

摘要 随着先进工艺节点的持续演进, 集成电路制造工艺复杂性显著提升, 产业对跨层次、多模态和智能化制造能力提出了更高要求. 传统制造技术主要依赖专家经验、物理机理与数理模型, 在应对工艺迭代、流程优化以及跨场景技术迁移方面能力明显不足, 迫切需要具备泛化能力和产业适配性的智能化技术体系. 近年来, 大模型在多个领域展现出卓越的知识抽象、语义理解与泛化能力, 为推动集成电路制造技术智能化发展提供了新的契机. 然而, 集成电路制造通常伴随多模态异构数据、小样本、复杂工艺耦合, 以及效率与精度难以兼顾等复杂特性, 使得通用大模型难以直接适用. 因此, 亟需在模型设计中有机融合多模态数据、物理机理与领域知识, 构建面向制造环节的专用大模型体系. 本文聚焦集成电路制造的核心难点与关键挑战, 介绍集成电路大模型发展现状, 并展望未来发展路线.

关键词 集成电路, 设计制造协同, 智能制造, 人工智能, 大语言模型

1 引言

集成电路 (integrated circuits, IC) 制造技术是支撑现代信息产业发展的核心基石, 其先进性与智能化水平直接决定了芯片的性能、良率与生产效率. 随着工艺节点的持续演进和器件结构的不断复杂化, 制造环节面临一系列新挑战, 包括制程窗口逐渐收窄、工艺参数间耦合增强、缺陷模式快速演化、研发与量产成本持续攀升等问题 (参见文献 [1]). 同时, IC 制造流程链条极长, 覆盖晶圆准备、光刻、刻蚀、离子注入、薄膜沉积、化学机械抛光、金属互连、封装测试等多个关键环节, 各阶段之间存在高度的流程耦合与反馈闭环, 导致数据关联复杂、问题定位困难、全局优化空间巨大而不透明. 在此背景下, 亟需构建具备多阶段协同感知、预测与优化能力的新型智能制造体系.

引用格式: 孙奇, 邓锦堦, 金谦, 等. 集成电路大模型: 驱动集成电路制造技术跨越发展. 中国科学: 信息科学, 2026, 56: 2327–2341, doi: 10.1360/SSI-2025-0423

Sun Q, Deng J Y, Jin Q, et al. Large models for integrated circuits: driving leapfrogging advances in IC manufacturing technologies. Sci Sin Inform, 2026, 56: 2327–2341, doi: 10.1360/SSI-2025-0423

长期以来,传统的集成电路制造方法依赖专家经验和数理建模,通过仿真工具与统计方法实现参数调节与质量控制.然而,这一范式在面对多模态异构数据融合、跨工艺流程协同优化、缺陷根因追溯、工艺鲁棒建模等问题时表现出明显的瓶颈,难以支撑未来越发复杂的先进工艺发展.与此同时,制造过程中的知识碎片化、数据孤岛化与反馈延迟等问题,也限制了知识的有效迁移.近年来,大语言模型 (large language model, LLM) 技术在自然语言处理、图像理解、科学计算等多个领域展现出强大的跨模态理解、知识抽象与任务泛化能力,推动智能系统从感知智能向认知决策智能跃迁^[2].在此背景下,构建面向集成电路制造任务的大模型系统,融合产线工艺知识、制造物理机制与多源感知数据,成为极具潜力的技术发展方向.然而,集成电路制造过程具有多模态、高耦合、强物理约束和数据敏感等特征,对系统的确定性、可解释性、实时性与边缘部署适配能力提出了极高要求,通用大模型难以直接适配,面临以下关键挑战.

(1) 多模态、多物理场数据融合难.集成电路制造涉及版图 GDSII 文件、器件结构文件、脚本代码、扫描电镜图、量测数据、工艺参数、装备监控日志等多源异构数据,呈现强物理依赖、跨尺度耦合、多模态协同的特征.这些数据之间缺乏统一语义表示,传统多模态模型无法捕捉其工艺物理间的本征联系,难以实现有效的信息融合与过程建模,从而限制了大模型对制造过程的全面感知与精准建模能力.

(2) 小样本数据驱动的建模与泛化难.集成电路制造的关键数据 (如特殊缺陷图像、新材料特性、工艺窗口信息等) 普遍为小样本,关键异常数据样本量少,高标注成本与严格保密要求进一步限制了可用数据规模.现有大模型依赖大规模训练样本,在数据稀缺条件下常出现过拟合、生成幻觉或迁移失效等问题,难以胜任制造场景中高鲁棒性、高精度的建模与预测任务.

(3) 跨工艺、尺度与装备的知识迁移难.集成电路制造覆盖多个制程平台 (如 12 英寸/8 英寸)、装备厂商及不同工艺节点 (如 55, 28, 14, 7 nm),其底层材料系统、工艺配方与设备响应特性均存在显著差异.通用大模型缺乏面向工艺结构演化的可解释表征能力,模型迁移往往受到物理机制差异的严重制约,缺乏跨平台、跨工艺节点与工艺流程的泛化应用能力.

(4) 制造全流程任务的协同优化难.集成电路制造链条长,数千道工艺步骤间存在复杂的工艺依赖与反馈路径.例如,光刻误差将影响刻蚀均匀性,进而影响后续测试良率.大模型缺乏对制造链全流程结构的知识感知与阶段性记忆能力,难以建立跨环节协同优化模型,制约了其在缺陷预测、根因追溯、良率提升等多目标任务中的应用深度.

(5) 本地化实时推理难.许多制造场景 (如设备控制、缺陷筛选、产线分拣) 要求毫秒级或亚毫秒级响应,且需在边缘计算平台上运行.然而,现有大模型体量庞大、计算资源依赖重,剪枝、量化等轻量化方法仍难以满足工业实时控制的性能需求.缺乏面向制造现场需求的极致高效推理结构与异构硬件协同机制,成为大模型在集成电路制造场景下落地的主要技术瓶颈.

现有通用大模型在多模态理解、小样本学习、工艺迁移、流程优化、实时推理等方面均难以满足集成电路智能化、虚拟化制造的实际需求.集成电路大模型并非通用大模型在集成电路领域的简单迁移或垂直应用形式,而是一种面向复杂物理机制与专业知识体系、融合设计与制造全过程的全新智能建模范式.目前,国内外关于该方向的系统性研究仍处于起步阶段,尚缺乏统一的理论框架与工程实践路径.本文结合集成电路产业制造的核心特征与关键任务,系统性总结当前技术发展趋势与进展,梳理为“五层体系架构”,包括基础设施层、机理增强算法底座、专用基座大模型、智能交互接口和典型应用等,围绕大模型的落地实施路径进一步提出了“五阶段构建范式”:数据集构建与数据生成、虚实融合预训练、器件与工艺迁移、流程智能体协同、私有化部署与持续增强,为大模型在集成电路设计制造中的工程化部署提供方法论支撑.围绕当前亟需突破的关键技术瓶颈,本文进一步详细探讨了其在光刻仿真与光学邻近效应校正 (OPC) 优化、工艺计算机辅助设计技术 (TCAD) 仿真、良率建模与分析等典型应用中的技术路线.最后,本文对集成电路大模型未来的发展方向进行了前瞻性展望,旨在为该领域实现理论突破与产业落地提供系统参考与路径指导.

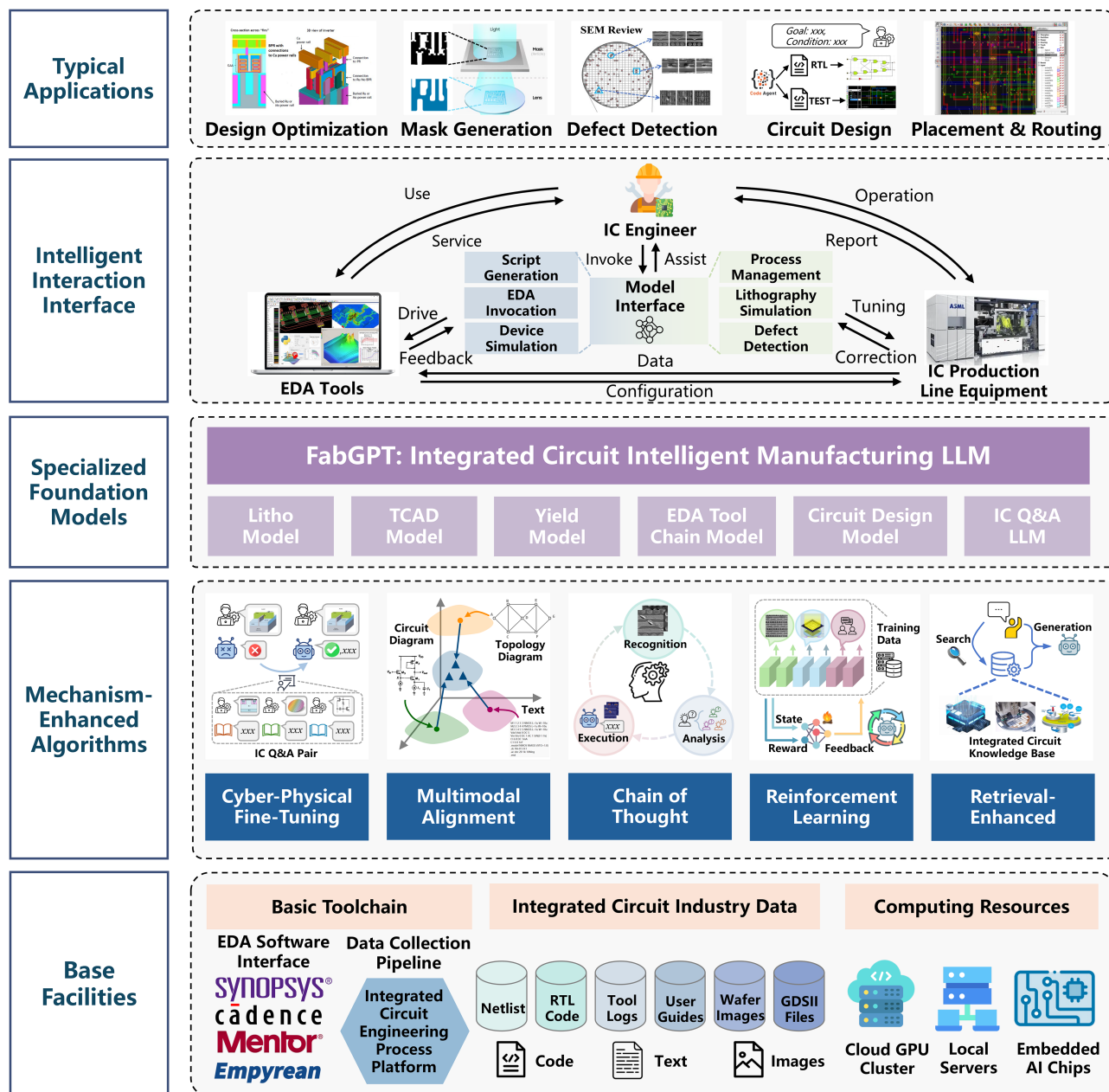


图 1 (网络版彩图) 集成电路大模型体系架构图。

Figure 1 (Color online) Overall architecture of the integrated circuit large model system.

2 集成电路大模型定义与体系架构

本文综合考虑当前技术进展, 构建了集成电路大模型的五层体系架构 (如图 1 所示), 从下至上分别为基础设施层、机理增强算法底座、专用基座大模型、智能交互接口、典型应用等。各层次紧密衔接, 形成从底层资源到上层应用的完整生态体系^[3], 下面逐层进行介绍。

2.1 基础设施层

基础设施层是 IC 大模型的底层支撑, 提供模型训练和运行所需的工具链、数据及计算资源。首先是基础工具链, 依托设计自动化 (electronic design automation, EDA) 软件, 构建大规模仿真数据集, 结合集成电路产线的

实际数据,实现虚实融合数据底座,保障设计与制造数据的高效对接和流转.其次是集成电路工业数据和知识,涵盖设计与制造过程中产生的各种数据模态,包括设计阶段的 RTL 代码、网表、EDA 脚本与日志,版图阶段的 GDSII 文件、版图图像,制造阶段的量测数据、工艺配方参数、过程控制数据、晶圆扫描电镜图等.这些多模态知识以数据库或知识图谱形式提供,构成模型学习的语料支撑.再次是计算资源,提供训练与推理服务器集群、嵌入式芯片等多层次算力设施,支持模型高效训练、实时推理与部署应用.这一层的目标是构建 IC 大模型训练与推理的基础设施,使上层算法可以高效获取所需的训练素材和计算资源.

2.2 机理增强算法底座

面向 IC 专业大模型构建,需要创新虚实融合的工艺建模、模型训练、迁移泛化技术,将集成电路专业机理知识与通用大模型算法相结合,形成专用大模型基座.主要包括以下几项典型技术.(1) 虚实融合的监督微调技术^[4]:融合仿真与产线实际数据,构建涵盖逻辑设计、版图实现、工艺优化等场景的高质量问答数据,通过虚实融合提升建模精度,在通用大模型基础上进行监督微调,使其能精准建模单点工艺,理解 IC 设计、制造中的专业问题.(2) 多模态特征优化及对齐技术^[3]:开发处理图像、拓扑图、文本等异构 IC 数据的编码与融合策略,定制面向异构数据的编码器,通过对比学习等方法对齐模态表示,提升模型对电路图、网表等数据的跨模态理解能力.(3) 思维链 (CoT) 分析技术^[5]:引入“理解-分解-执行”的显式思维链,对输入进行多步推理和约束检查,并结合链式提示词 (prompt) 提升模型的可解释性与复杂任务分解能力.(4) 强化学习技术^[6]:结合仿真工具与集成电路数据集向模型提供状态与奖励,在奖励的驱动下生成设计和制造的决策,并实时写回形成反馈.(5) 检索增强技术^[7]:构建海量的 IC 多模态知识库,并集成检索-生成框架,使模型先查找设计规范、图片、文档片段等,再融合结果生成可追溯的回答.通过以上五类核心技术,形成面向 IC 领域的大模型算法基座,为具体的应用模型提供共性支撑.

此外,小样本与数据不平衡是制约集成电路大模型工程化落地的核心问题之一,需要在算法底座层面进行针对性设计,常采用三类工程化路径:一是元学习/原型方法,其通过在嵌入空间构建判别性类别原型以实现快速适配,例如在晶圆图缺陷识别中基于原型网络并结合混合自注意力的方案在 5-way 5-shot 下可达约 92.3% 的识别率^[8].二是自监督预训练或少样本微调,先用大规模未标注数据学习鲁棒表征,再微调进行少样本适配,相较传统卷积神经网络 (convolutional neural network, CNN) 可带来约 11.5%~21.5% 的性能提升^[9],在更精细的纳米尺度缺陷检测场景中,引入跨模态预训练并利用文本提示进行参数高效微调也可带来性能的大幅提升^[10].三是生成式与增强式数据扩充方法,通过生成模型或重建预训练任务合成高质量伪样本,以扩充训练分布并提升模型泛化能力,如通过对抗训练生成多样化的缺陷模式以辅助特征学习,相比传统手工特征方法可带来约 10%~25% 的检索精度增益^[11].

2.3 专用基座大模型

基座大模型是集成电路大模型体系的核心,通过构建一系列专门针对集成电路设计与制造的大模型,不仅是语言模型,更是视觉模型、多模态模型等领域专用大模型,形成紧密协同的大模型体系.典型模型如下.(1) 光刻大模型^[12]:基于视觉感知和工艺数据的智能仿真平台,能够灵活适应不同工艺节点与光源条件,实现快速精准的光刻图形预测和跨节点通用化.(2) TCAD 大模型^[13]:融合物理机理与图学习技术,覆盖关键半导体工艺(如扩散、氧化、刻蚀等)及载流子输运过程,提供实时、高精度的工艺仿真与器件性能预测服务.(3) 良率大模型:面向先进工艺节点良率优化需求,整合设备运行数据、缺陷检测图像和多种传感器数据,构建精确的良率预测与异常预警体系.(4) EDA 串链大模型^[14]:面向 EDA 工具链的智能控制模型,通过任务分解与自动化脚本生成技术,实现 RTL 到 GDSII 全流程的智能设计优化.(5) 电路代码大模型^[15]:针对功能描述、RTL、网表与版图布局等多模态电路信息进行统一表征和深度学习,建立跨阶段的电路性能、功耗、面积预测与优化体系.(6) IC 知识问答大模型^[3]:结合视觉与文本交互能力,构建面向集成电路制造过程的实时知识辅助系统.以上模型通过合理的功能划分和协同机制,在小样本、低成本条件下,共同构成完整覆盖集成电路制造与设计领域的智能化基础模型体系.

表 1 五层体系架构的量化指标.

Table 1 Quantitative metrics for the five-layer architecture.

Architecture layer	Core metrics	Target thresholds	Engineering impact
Infrastructure layer	Compute & data capacity	Compute: 0.1~1 PFLOPS; multimodal data: >10 TB	Supports multimodal pre-training & continuous production iteration
Physics-informed layer	Consistency & efficiency	Convergence speedup: >30%; labeled data need: -30% to -50%	Accelerates convergence; significantly cuts data & simulation costs
Foundation model layer	Accuracy & inference speed	Key task acc./recall: >90%; inference speedup: 10~100×	Serves as an efficient surrogate meeting strict physical simulation standards
Interface layer	Latency & compatibility	API response: <500 ms; mainstream tool coverage: 100%	Enables seamless engineer interaction & closed-loop production scheduling
Application layer	Cost & efficiency yield	Defect prediction acc.: >95%; R&D cycle: -30% to -50%	Transforms model capabilities into direct economic value for semiconductors

2.4 智能交互接口

通过统一的模型适配接口, 将集成电路工程师、EDA 工具与产线装备紧密联结, 依托大模型技能 (skills)、上下文协议 (model context protocol, MCP) 等技术, 可以支持本地工具链的插件化调用^[14], 也兼顾云端协同管理, 主要工作包括如下内容. (1) 模型驱动的工具链集成: 依托串链大模型, 通过适配接口发起代码与脚本生成、EDA 调用与优化、工艺与器件仿真等请求, 驱动 Synopsys, Cadence 等 EDA 工具执行相应流程, 并将仿真结果与优化建议实时反馈给工程师. (2) 智能产线协同: 针对制造流程, 通过接口下达制造流程管理、光刻仿真与优化、缺陷检测与根因分析等指令. 大模型将调度装备执行任务, 并通过接口回传报表和调整建议, 以指导装备运行、工艺配方参数修正与优化. (3) 数据与配置同步: 通过工具与装备接口共享设计数据、运行日志与配置信息, 模型适配接口负责统一的数据格式转换和权限控制, 确保整个闭环流程中数据的及时互通与安全可追溯. 交互接口层通过“工程师-大模型-工具/装备”三方联动, 实现了设计与制造环节的无缝对接与智能协同, 极大提升了研发效率与生产灵活性.

2.5 典型应用

集成电路大模型落地推广, 需要结合 IC 设计制造的实际业务场景, 聚焦各类具体任务的解决方案. 大模型最终将以具体应用形态落地, 并真正产生业务价值. 例如, 在可制造性设计优化中^[12], 大模型在版图前端引入 DFM/DRC 规则库、工艺窗口等知识, 为版图提供可制造性改进建议; 光刻掩模生成^[4], 大模型参与 OPC 优化和光刻仿真, 加快光罩补偿设计并提高硅上成像准确度; 缺陷检测模型^[10]从海量晶圆检测数据中发现缺陷模式关联, 指导工艺改进; 逻辑电路设计^[16], 大模型自动合成多功耗域约束下的 RTL 架构, 插入形式化验证友好的时序修复逻辑, 并生成覆盖率导向的测试激励集, 显著缩短逻辑收敛与验证周期. 不同领域 (CPU/GPU/ASIC 设计公司、晶圆代工厂、EDA 工具厂商等) 都会在这一层关注自己的应用场景. 为了实现任务落地, 需要将认知应用服务与具体业务流程融合, 并辅以必要的人机交互和工程集成. 后续, 我们将详细讨论三个典型任务的落地情况和效果, 其反馈可驱动大模型持续迭代.

整体而言, 上述讨论的五层架构从底层的数据算力支持, 到中间算法、模型、接口, 再到顶层的任务落地, 构成了一个闭环生态: 下层为上层提供能力支撑, 上层应用产生的数据和反馈又可反哺模型改进. 这种架构有助于理清 IC 大模型建设所需的各方面要素, 以及各部分之间的关系, 为企业和研究者规划开发路线图提供了参考. 为使该架构具备工程可落地性, 各层级均需满足相应的核心性能标准, 表 1 总结了该架构在实际集成电路制造场景中的关键量化指标基准.

3 五阶段构建范式

构建一个满足 IC 设计制造协同需求的大模型, 不可能一蹴而就, 而需要分阶段、有步骤地推进. 结合通用

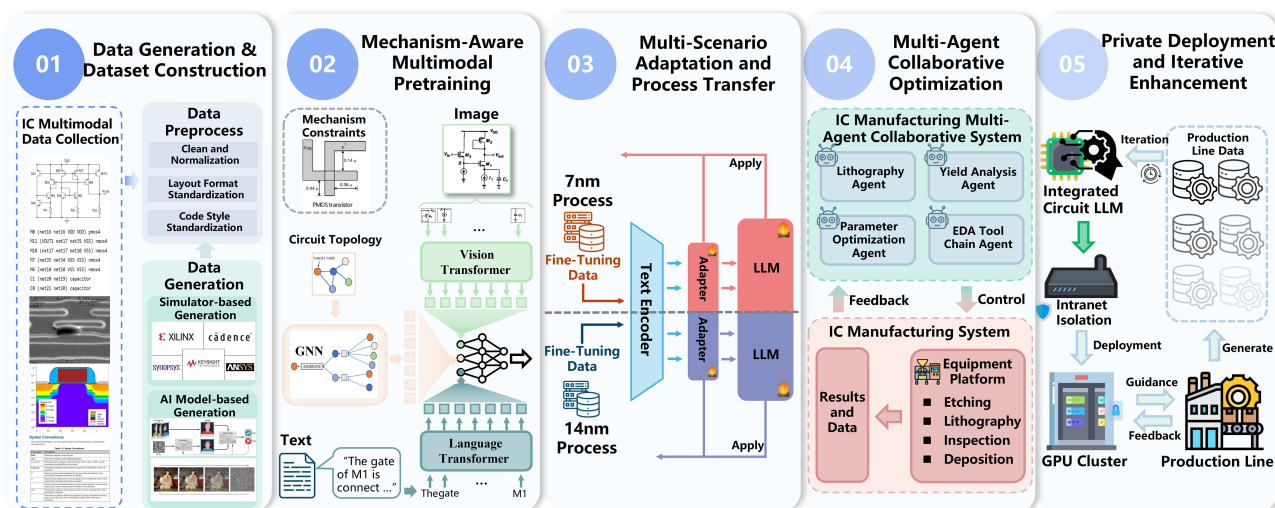


图 2 (网络版彩图) 五阶段构建范式。

Figure 2 (Color online) Five-stage construction paradigm.

大模型构建范式^[2], 本文总结性阐述了集成电路大模型的五阶段构建范式, 如图 2 所示, 下面逐一介绍每个阶段的目标与方法。

(1) 数据生成与数据集构建. 数据是大模型的血液, 而 IC 领域的准备数据尤为关键且困难. 在第一阶段, 需要围绕目标任务汇聚 IC 全流程数据与知识资源, 包括设计阶段的开源电路^[17]、仿真日志, 制造阶段的工艺与测试数据, 以及通过联邦学习^[18]等机制整合、构建大规模数据, 构建 IC 领域知识库; 收集行业标准、教材、论文、手册等资料, 提炼为知识图谱, 为大模型提供推理支持与检索增强能力. 随后需对原始数据进行清洗与标准化, 并对不同模态数据采用定制化处理策略, 如规范 RTL 代码风格、统一 GDSII 版图格式、增强图像样本等, 以确保数据质量与一致性. 此外, 由于真实数据有限, 需借助仿真和生成模型扩充数据量. 一方面, 可运行 EDA 仿真器产生模拟数据, 如生成不同工况下的电路时序功耗数据、不同工艺偏差下的光刻图形. 另一方面, 也可以通过生成对抗网络 GAN^[19]、扩散模型^[20]等生成式 AI (artificial intelligence) 生成版图样本、缺陷图案等逼真数据以丰富训练集。

(2) 机理感知多模态预训练. 在获取充足数据后, 第二阶段的目标是训练大模型的基础模型, 赋予其对 IC 多模态数据的理解和初步生成能力. 首先根据 IC 数据特点选取 Transformer^[21]、图神经网络 (GNN)^[22]等合适的架构, 采用多分支的方式处理不同模态输入并在高层融合. 再进行自监督预训练, 设计适合 IC 数据的预训练任务, 如电路网表的节点预测、版图的缺失图形重建、工艺文本说明的掩码预测等, 使模型学会提取各模态的特征表示. 随后进行多任务混合训练, 通过对比学习等训练手段将各个模态在语义空间中进行对齐. 在集成电路产业场景中, 将机理知识融入预训练过程至关重要, 通过引入显式的机理约束, 如将设计规则检查器的反馈作为辅助损失项, 从而提升模型对行业规范的理解能力. 经过充分预训练, 我们将得到一个 IC 大模型基座, 具备广泛的 IC 知识和数据表征能力。

(3) 多场景适配与工艺迁移. 获得通用预训练基座后, 第三阶段将聚焦于面向具体应用场景的模型定制化适配, 旨在释放模型在特定任务中的性能潜力, 并实现跨工艺节点、跨器件结构与跨装备平台的知识迁移. 该阶段的核心挑战在于, 不同工艺节点与装备平台在特征尺寸、材料体系及工艺窗口上的显著差异, 往往会导致模型在直接迁移时出现性能退化. 领域内指令微调与参数高效微调是实现跨工艺迁移的核心技术路径. 通过 LoRA 等低秩适配方法^[23], 模型仅需更新少量参数即可捕获新工艺节点的差异性规律. 已有研究从多个典型场景验证了该路线的有效性. 在光刻建模任务中, 从 10 nm 向 7 nm 工艺节点迁移时, 尽管光源与抗蚀剂材料发生显著变化, 相关研究^[24]仍可在仅使用约 1% 目标工艺数据的情况下通过迁移学习, 将关键尺寸均方根 (RMS) 误差由 2.5 nm 降至 1.5 nm, 相对误差降低超过 50%, 显著提升数据利用效率. 在先进节点的可靠性预测任务中, 模型从

16 nm 向 5 nm 节点迁移时,通过冻结通用特征层并微调任务相关层,可在有限样本条件下将预测误差降低约 33% [25]. 此外,跨器件结构的迁移能力同样得到了实证支持,针对从 FinFET 向全环绕栅场效应晶体管 (GAAFET) 的器件结构演进,通过引入在线学习机制可实现模型在新结构下的免重训适配,在失效检测任务中取得超过 15% 的性能提升 [26]. 这些结果共同表明,基于参数高效微调与在线适配的技术路径,为集成电路大模型在跨工艺、跨尺度场景下的工程落地提供了可靠支撑.

(4) 多智能体协同闭环优化. 第四阶段关注如何将多个大模型嵌入实际 IC 设计制造流程中,使其与工程师、其他智能体及 IC 工具相互协同,构成闭环反馈的整体. 大语言模型可作为反思智能体来自我验证和改进预测结果 (参见文献 [27]). 在制造流程中,光刻智能体、良率智能体、工艺控制智能体等专用智能体可以协同工作,形成跨工艺环节的闭环优化框架,从而实现从光刻仿真到制造执行的全流程智能化协同,提升整体制造效率和产品良率. 同时,针对多智能体系统的优化,当前其主要面临提示词敏感性、拓扑设计复杂性和组合搜索空间巨大等挑战. 常见的优化方法包括自动化设计提示词示例,以及使用蒙特卡罗树方法搜索更好的智能体拓扑结构 [28].

(5) 私有化部署与迭代增强. 集成电路制造企业面临严格的数据安全与合规要求,工艺配方、设备参数、良率数据等核心信息属于高度敏感的商业机密,这使得集成电路大模型必须采用私有化部署方式,确保敏感数据不离开企业内部环境. 私有化部署主要包括本地化环境搭建、安全隔离配置等核心环节. 企业需在内部数据中心部署专用的 GPU 集群和推理服务器,通过内网隔离确保模型服务的安全可控运行. 在部署中,可采用低精度量化、剪枝、软硬件协同优化等方法来降低硬件资源需求以及功耗,例如通过 1-bit 二值化量化结合专用加速器设计,在牺牲约 25% 准确率的情况下,可在边缘设备上实现 11 倍的推理加速和 2 倍以上的能效提升,验证了极致轻量化技术的实时推理可行性 [29]. 根据大模型推理加速的系统性调研 [30],在 7B 参数模型上采用 INT4 量化可实现 18~343 tokens/s 的吞吐量 [31,32],结合 50% 结构化剪枝,相比仅量化方案,性能可额外提升 1.2~1.5 倍 [33],而现场可编程门阵列 (FPGA) 等边缘硬件通过量化与剪枝协同优化可达到 3.6~333 tokens/s [33],这为集成电路产线本地化实时推理提供了可行的技术参考. 同时,由于通用大模型基于公开数据训练,往往缺乏对企业特定制造场景的深度理解,需要建立基于私有产线数据的持续学习机制. 当企业技术更新迭代、引入新工艺或新设备时,模型需要通过增量学习方法进行适配调优,以准确适应新的技术特性和业务需求. 通过建立从数据收集、模型训练、应用验证到反馈优化的完整闭环机制,企业能够实现模型能力的持续提升,确保其能够准确适应企业特有的制造场景和技术挑战.

从数据准备到预训练,再到任务定制、流程集成,最后到部署维护,五阶段构建范式为开发和应用 IC 大模型提供了清晰路径. 遵循这一范式,可以确保各环节衔接顺畅,在解决技术问题的同时兼顾工程落地需要.

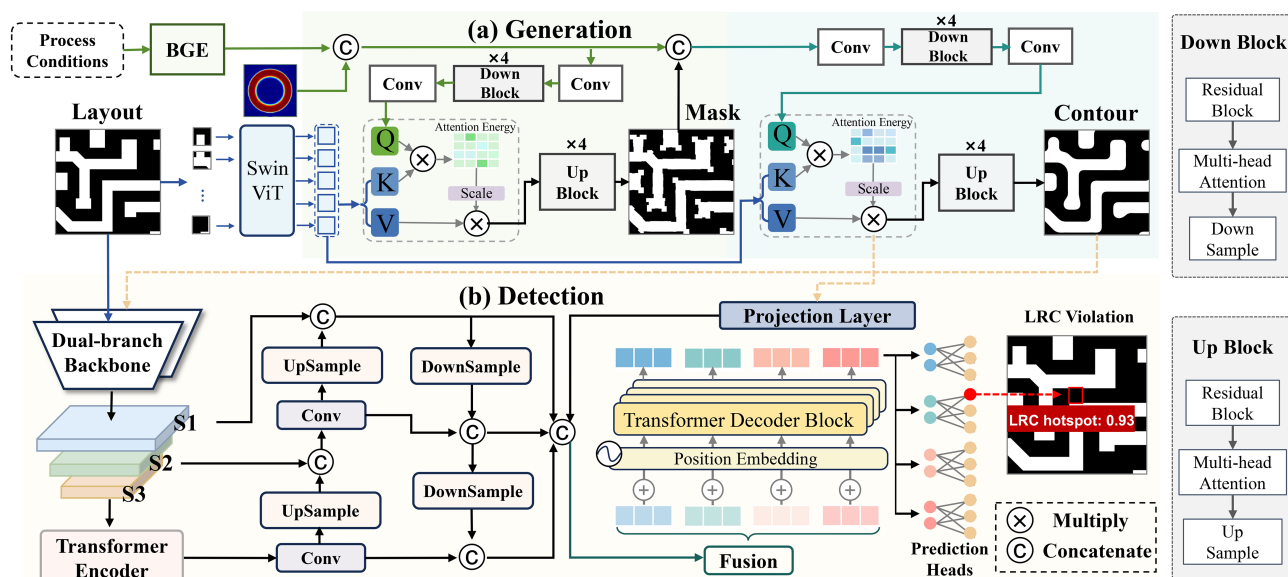
4 关键技术与典型应用

近年来,学术界和产业界在集成电路大模型领域取得了显著进展,FabGPT 作为首个集成电路专用大模型 [3],以及英伟达的 ChipNeMo [34] 和东京电子的 SemiKong [35] 的成功应用,都充分证明了大语言模型在半导体领域的巨大潜力. 为了进一步评估和展示集成电路大模型的潜力,本节选取了典型场景对技术路径进行详细阐述,并总结取得的技术亮点和性能提升.

4.1 光刻仿真与 OPC 优化

光刻是半导体制造中的关键步骤,负责将设计电路与布局精确转化为硅片上的物理结构. 随着先进制程节点的不断演进,传统光刻工艺在图形分辨率、图形保真度与工艺窗口等方面面临严峻挑战. 近年来,各界开始探索利用大模型增强计算光刻能力. 凭借其在文本理解、图形语义建模与视觉内容生成方面的优势,大模型已在亚分辨率辅助图形 (SRAF) 生成、OPC、光刻热点检测等关键任务中取得了非凡的效果 (参见文献 [4,36]).

在光刻优化过程中,为提高掩模优化过程中的保真度和可打印性,SRAF 被广泛用于增强孤立特征和半隔离型特征的可打印性与光刻工艺窗口. 现有的 SRAF 生成方法主要分为基于规则的方法、基于模型的方法以及基于机器学习的方法 [37,38]. 这些方法普遍缺乏统一、泛化且高效的生成能力,难以同时兼顾准确性、适应性与

图3 (网络版彩图) 基于Transformer的计算光刻大模型^[4].Figure 3 (Color online) Transformer-based computational lithography large model^[4].

制造可行性. 在此背景下, Li 等^[39]提出了一种基于大型语言模型的 SRAF 生成方法. 该方法通过引入语义提示, 将 SRAF 生成任务转化为自然语言处理问题. 实验表明, 该方法在工艺变化带 (PVB) 面积和边缘放置误差 (EPE) 精度方面优于现有主流方法, 且在运行速度上实现了 3 倍以上的提升, 展现出大型语言模型在复杂掩模优化中的显著潜力.

随着制程缩小, OPC 导致设计图形与曝光图形间偏差不断加剧. OPC 优化是先进制程中用于修正光学邻近效应的重要手段, 其目标是使曝光图形尽可能贴近设计图形. 传统方法依赖专家制定复杂的优化配方^[40], 需要在不同设计和工艺条件下反复试错与调整, 效率较低. 为提升芯片开发效率并减少试错过程, Chen 等^[41]提出了一种 LLM 与强化学习结合的智能 OPC 配方生成框架: 先由强化学习自动搜索并产出最优候选配方, 再借助大模型对候选进行结构归纳、特征标注与决策树整合, 生成可适配多场景的最终配方. 大模型在总结与泛化上发挥关键作用, 使方法同时兼具精度与效率, 大幅减轻了芯片级 OPC 配方设计的人力负担.

光刻热点是导致图形打印缺陷的关键风险结构, 准确识别热点对于保障工艺良率具有重要意义 (参见文献^[42]). 近年来, 基于卷积神经网络的光刻热点检测取得显著进展^[43], 但此类方法依赖图像形式的版图表示, 易引入冗余信息且下采样过程可能导致图像失真和特征损失. 针对这一挑战, Chen 等^[12]提出了基于语义编码的 LLM 热点检测模型, 将原始图形设计系统第二版 (GDSII) 文件无损转换为文本形式, 并利用精心设计的预训练自然语言模型在原始二进制数据上直接进行热点检测. 相比传统图像表示, 该方法更具灵活性, 避免了信息损失, 在 ICCAD2012^[44]金属层基准测试数据和更具挑战性的 ICCAD2020^[45]过孔层基准测试数据上均取得了优异性能, 展现了大模型在复杂几何结构识别任务中的可行性.

与此同时, 研究者们也致力于构建统一的多任务视觉大模型, 以端到端的方式应对整个计算光刻流程的挑战. 例如, Jin 等^[4]提出的 Unitho 框架, 通过一个统一的 Transformer 架构集成了掩模生成、光刻仿真和多规违例检测 (DRC/MRC/LRC) 等多个核心任务, 如图 3 所示. 该模型的核心创新在于利用交叉注意力机制深度融合版图特征与多变的工艺条件, 从而实现了针对不同工艺窗口的高效、精准建模. 该框架不仅将仿真速度提升至商业工具的 85 倍, 其生成的光刻轮廓精度也高达 99.76%, 为实现快速的设计-工艺协同优化提供了强有力的技术支撑.

大模型通过其强大的文本理解和语义建模能力, 为半导体制造中的光刻优化提供了新的技术路径. 从 SRAF 生成、OPC 配方优化到光刻热点检测, 大模型的应用展现了其在提升计算光刻效率、优化掩模设计及减少人力

负担方面的巨大潜力. 随着制程节点的进一步缩小, 基于大模型的光刻技术将为先进半导体制造带来更大的技术突破, 推动制造工艺的持续发展.

4.2 TCAD 仿真

TCAD 通过物理建模、数值求解器件物理方程来预测器件的电学、光学和热学特性, 是半导体器件开发的核心仿真工具. 传统 TCAD 方法依赖专家经验进行参数调节和模型校准, 存在建模复杂度高、计算资源消耗大、人工干预程度深等局限. 近年来, 凭借其强大的自然语言理解、知识推理和代码生成能力, 大模型在 TCAD 领域展现出巨大潜力, 能够以自然语言交互的方式, 实现参数调优、工艺校准、器件结构生成以及仿真自动化等复杂任务, 从而显著提升器件设计与制造的智能化水平.

随着器件复杂度提升, 工艺参数成为影响器件性能的关键因素. 传统参数调优需要工程师反复调整器件制造过程中的上百个工艺参数, 对应参数优化方法缺乏系统性指导, 过程烦琐且难以保证全局最优. 现有基于贝叶斯优化的方法虽然在一定程度上改善了这一问题, 但仍需要大量专业知识来设置优化目标、约束条件以及搜索空间^[46]. 针对以上挑战, Liu 等^[47] 构建了参数调整-仿真的智能体框架, 通过自然语言解析设计目标, 并基于 LLM 预训练的物理知识、优化经验以及历史参数配置, 在实验迭代中获取最优配置. 工艺参数校准过程与参数调优相近, 同样需要考虑多个工艺步骤的耦合效应, 且严重依赖工程师的经验和试错方法. 以 FinFET 工艺开发为例, 完整的校准流程需要优化浅沟槽隔离氧化、鳍刻蚀、栅极氧化等多个工艺步骤的参数配置, 传统方法往往需要多轮迭代才能获得收敛且准确的仿真结果. Jeong 等^[48] 将贝叶斯优化框架与循环神经网络代理模型结合, 成功实现了浅沟槽隔离刻蚀轮廓的自动校准. 基于 LLM 的方法在此基础上进一步提升了校准过程的智能化水平, 能够通过自然语言理解工程师的校准目标, 结合领域知识提供参数调整建议和物理机理解释, 大幅简化了校准流程的复杂性.

针对器件结构生成与仿真自动化, 传统 TCAD 建模依赖工程师手工编写 Sentaurus 代码, 存在学习成本高、易出错、难复用等问题. Lu 等^[13] 以纳米线器件数据为例, 成功生成了 7000 个包含 18 个关键参数的纳米线结构, 首次验证了使用大语言模型生成 TCAD 结构输入文件的可行性. 在仿真脚本自动生成方面, 基于在用户需求-仿真脚本的数据上进行的领域自适应训练, 英伟达的 ChipNeMo 能够理解工程师的高层次需求并生成相应的仿真脚本, 展示了 LLM 在仿真脚本自动生成中的潜力^[34].

为进一步增强大模型在 TCAD 领域的的能力, 本文提出两种方法: 一是基于多模态检索增强生成 (retrieval-augmented generation, RAG) 技术整合 TCAD 领域知识资源; 二是利用 CoT 推理提升复杂建模任务的推理能力, 如图 4 所示.

考虑到器件建模的最佳实践和参数配置分散在各类文档中, 缺乏系统化组织和高效检索手段的问题, 多模态 RAG 技术通过构建统一的知识表示空间, 能够有效整合包括文档、参数表格、仿真代码以及历史仿真结果等多种信息资源, 实现跨模态的语义检索和知识融合^[49]. 在器件结构建模场景中, 工程师可以构建包括 Sentaurus 器件结构代码库、器件结构示意图以及仿真结果数据等内容的知识库. 当用户通过自然语言描述器件需求时, 多模态 RAG 系统首先对查询进行语义理解, 并在知识库中进行相似性检索, 快速定位相关的代码示例、文档知识和历史优化经验为推理过程提供上下文支持. 同时, 当模型的输出不符合预期时, 系统还会重新检索更相关的内容以提供优化迭代的指导. 在多模态 RAG 提供丰富上下文信息的基础上, 思维链推理技术可进一步将复杂的 TCAD 建模过程分解为逻辑清晰的推理步骤, 确保生成结果的质量和可解释性^[15]. 以 SDE 器件结构代码生成为例, 该推理过程从解析用户自然语言描述入手, 逐步规划器件几何结构、定义材料区域、设计掺杂分布、配置电极接触及优化网格密度等环节. 进一步地, 若在结果中检测到器件结构错误, 系统可重新推理并修正材料区域定义, 实现结果的自我修正.

通过多模态 RAG 和思维链推理等技术的深度集成, TCAD 智能代理系统能够显著降低专业工具的使用门槛, 使更多工程师能够高效地进行器件设计和工艺优化. 这种技术路径有效提高了设计效率, 为半导体器件开发提供了强有力的智能化支撑.

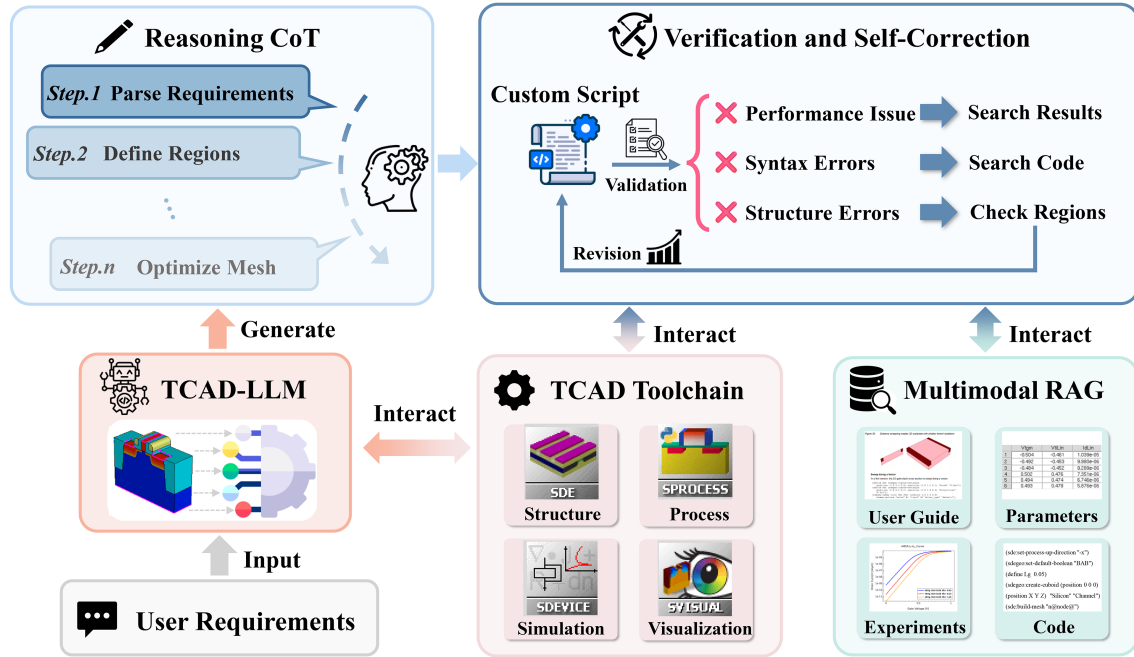


图 4 (网络版彩图) 基于多模态检索增强生成和思维链推理的 TCAD 智能体框架。

Figure 4 (Color online) TCAD intelligent agent framework based on multimodal RAG and CoT reasoning.

4.3 良率建模与分析

良率预测建模旨在根据工艺参数、设计特征和历史数据准确估计芯片的合格率。传统方法如泊松模型或负二项模型难以处理高维变量间复杂的交互关系,且预测精度受限于数据规模和模型假设。随着制造过程中多模态数据(如缺陷图像、光刻图、设备日志、传感器数据、测试记录等)的广泛产生,良率预测面临更高维度和更强语义理解的挑战。集成电路大模型^[3,5,50]依托其强大的特征提取与模式建模能力,能够从海量历史数据中学习深层语义表征,构建高精度良率预测模型。通过对实时生产数据的连续分析,大模型可智能识别良率波动、趋势异常或潜在风险,及时预警并定位至具体批次或工艺步骤,辅助工程师快速介入,降低系统性良率损失风险^[51,52]。

影响良率的根本因素往往来自产线上各类微观缺陷,准确检测并溯源是良率提升的关键。传统依赖人工目检和专家分析的流程不仅效率低,且难以适应复杂多样的缺陷类型。借助集成电路大模型对图像和工艺数据的多模态学习能力,可融合迁移学习与少样本学习方法,提升对细微缺陷模式的识别能力,甚至无需显式标注即可发现新型缺陷。Jin 等^[10]提出了基于大模型结构的多模态缺陷检测框架 SEM-CLIP,通过构建双分支 CLIP 结构,能够精确判别微小缺陷形貌,构建了可学习的提示词模块,结合缺陷产生的工艺步骤和位置,进一步提升检测性能,能够在 10 个样本等极低样本量条件下实现高达 99.8% 的缺陷分割精度和 88.4% 的分类精度。Jiang 等^[3]提出了 FabGPT 和 FabThink^[5],可以进一步结合工艺日志、设备状态、批次特征等多源信息进行图谱化建模与关联推理,可揭示传统方法难以识别的隐藏因子,通过与工程师的交互问答,实现良率下降的高效归因与预防控制,最终实现了高达 95.5% 的缺陷检测精度和 93% 的缺陷问答精度。

4.4 领域大模型对比分析

为从整体架构层面理解当前集成电路领域大模型的发展特征,并进一步明确本文提出体系的定位与必要性,表 2 对近年来国内外具有代表性的集成电路专用大模型及相关应用框架进行了对比,涵盖其模型架构、核心能力、适用场景及性能表现。从对比结果可以看到,现有工作大多围绕单一或局部流程展开:一类方法通过领域自适应预训练或微调,显著提升模型在特定设计或制造任务中的专业性能,如 ChipNeMo^[34], SemiKong^[35], FabGPT^[3], ChatEDA^[14], VeriGen^[53], RTLCode^[54]等;另一类方法则以通用大模型为基础,通过提示工程或

表 2 集成电路领域专用大模型及应用框架对比.

Table 2 Comparison of domain-specific large models and application frameworks in integrated circuits.

Model/tool name	Core architecture	Core functions	Application scenarios	Key performance metrics
ChipNeMo [34]	Llama2 (7B/13B/70B)	Engineering assistant, EDA script generation, bug analysis	Full chip design flow, verification, EDA tools	EDA script accuracy: 49%~77%; retrieval hit rate: +30%
SemiKong [35]	Llama3 (8B/70B)	Process knowledge QA, etching optimization, anomaly detection	Semiconductor manufacturing, process control	13% better performance than Claude-3.5; expert-level communication metric improved by 8.1%
FabGPT [3]	PandaGPT-7B	Wafer defect detection, root cause analysis, manufacturing QA	IC wafer manufacturing inspection, process optimization	QA accuracy 14.29% higher than GPT-4; AP improved by 8.16%
ChatEDA [14]	Llama2 (7B/13B)	Automated physical design flow from RTL to GDSII	Chip synthesis, placement and routing, physical design	Task excellence rate: 82%
VeriGen [53]	CodeGen (355M~16B)	Automatic Verilog code generation	Digital circuit design, FPGA development	Syntactic correctness: +41%; functional correctness: 18% higher than baseline; fast inference
RTLCoder [54]	Mistral-7B/DeepSeek-Coder 6.7B	Natural language to Verilog code generation	Digital IC design, RTL automated generation	Performance metrics improved by 31%~56% over GPT-3.5
ChipGPT [55]	GPT-3.5-based framework	Automatic Verilog generation, PPA space exploration	Logic design, front-end design optimization	Average design area reduced by 35%~47%
Chip-Chat [56]	GPT-4-based framework	Natural language to Verilog, multi-round refinement, collaborative design	Rapid prototyping, microprocessor design	Successfully taped out a 130nm 8-bit microprocessor, but complex modules required multiple iterations
GPT4AIGChip [57]	GPT-4-based framework	Automatic hardware generation	HLS code design, FPGA AI accelerator, GEMM operators	Code generation success rate improved by 5×; inference latency reduced by 2%~16%
DocEDA [58]	EDocNet + GAM-YOLO + LLM	Circuit document parsing, parameter extraction, netlist generation	Analog circuit design, technical document automation	Accuracy improved by 11.8% over YOLOv8; recall improved by 40%
LayoutCopilot [59]	GPT-4/Claude/GPT-3.5	Natural language to layout commands, layout optimization	Interactive analog IC layout design	Accuracy improved by 13.8%; performance improved by 115%

多智能体协同实现对既有 EDA 工具链的智能封装与调用, 如 ChipGPT [55], Chip-Chat [56], GPT4AIGChip [57], DocEDA [58], LayoutCopilot [59] 等. 这些研究在各自关注的任务维度取得了良好效果, 但整体上仍以“点状能力增强”为主, 模型、工具与产线之间的协同关系缺乏统一建模.

基于上述分析可以看出, 当前集成电路领域大模型研究虽然在模型能力和工程效率上不断取得进展, 但尚缺乏一个能够系统整合数据基础、机理增强算法、专用模型与交互接口的整体框架. 针对这一不足, 本文提出的

集成电路大模型体系不以单一模型或任务性能为目标,而是从工程化与产业落地视角出发,构建覆盖设计与制造全流程的分层架构与阶段化构建范式,从而为复杂场景下的大模型协同应用提供统一的技术组织方式。

5 未来展望

集成电路大模型作为一个新兴方向,已经展现出巨大的应用潜力,但距离在IC产业中广泛落地仍有一定距离。未来的发展需要产业界深入挖掘实际需求,通过更多实践来推动大模型能力提升。本文认为,从技术角度看,IC领域的特殊需求将驱动大模型朝着以下几个重点方向演进。

统一仿真基座。传统集成电路开发中电路仿真、热仿真、可靠性仿真等工具相互独立,形成了大量的信息孤岛。未来应建立统一的仿真基座,实现电、热、机械、可靠性等多物理场的集成与交互。这一统一基座将基于多任务学习与物理机理约束融合,使模型能够一次性完成多目标、多场景的综合优化,大幅提升研发效率与仿真精度。

黑盒优化范式。随着工艺日趋复杂,一些优化问题难以直接进行机理解析,未来将更多依赖基于强化学习等技术的大模型黑盒优化范式。这种方法无需深入机理,可充分利用不确定性估计机制,仅凭数据反馈迭代式地逼近最优结果,指导模型在高维参数空间快速定位最优解,有望在未知材料、新型工艺探索场景中发挥重要作用。

数字孪生与具身智能。数字孪生技术为制造装备提供实时高精度的数字模型,未来的大模型将直接嵌入生产设备,实现实时感知、预测与决策。通过具身智能技术,大模型不仅能够预测设备状态,还能实时调整设备操作,如动态补偿刻蚀条件或光刻曝光参数,从而提升产线控制的实时性与精度,进一步推动智能制造的落地应用。

开源生态、数据与标准。开源模型与数据生态的缺乏极大限制了集成电路大模型的快速发展。未来需建立面向集成电路领域的开源模型、基准数据集与标准化接口,如公开电路、版图、工艺数据集等。通过构建健康的社区生态与明确的标准体系,促进产业与学术界的协作创新,推动大模型在产业内更广泛、更深入的应用。

总体来说,集成电路大模型正朝着机理内嵌、统一表征、高可信输出、具身交互、大小模型协同、轻量高效、安全可控等方向发展。可以预见,在不远的将来,我们将目睹IC产业因为大模型技术的成熟而发生深刻变革:设计师和工艺师的工作方式被重新定义,EDA工具形态演化出智能助手,晶圆厂运营更加智能自主。集成电路大模型将成为连接设计与制造、虚拟与现实的核心纽带,驱动半导体行业迈上新的台阶。

6 结论

集成电路大模型融合人工智能与半导体设计制造,成为推动产业创新的重要前沿方向。然而,集成电路领域独特的复杂性对模型的泛化性、多模态数据融合、小样本泛化能力以及实时性等提出了巨大挑战,简单地将通用大模型移植到集成电路领域无法解决这些问题。因此,发展集成电路大模型必须建立全新架构与技术体系。

本文提出了“五层体系架构”,涵盖基础设施、算法底座、基座模型、交互接口与典型应用,从底层数据到应用落地构成完整生态闭环;同时提出了“五阶段构建范式”,明确从数据构建、预训练到应用部署的实施路径。本文进一步针对光刻仿真与OPC优化、TCAD仿真、良率建模与分析等关键任务,介绍其相关技术进展,充分展示了大模型的巨大潜力与实际价值。展望未来,随着统一仿真基座、黑盒优化、数字孪生与开源生态等方向的深入探索,集成电路大模型有望成为芯片产业发展的核心驱动力,推动产业进入更加智能、高效、协同和跨越式发展的新时代。

参考文献

- 1 Zhang Y. The 12-inch CMOS process integration platform is built to support the industrialization of research achievements from universities and research institutes: an interview with Wu Hanming, academician of the Chinese Academy of Engineering, and professor at the College of Microelectronics (Institute of Micro and Nano Electronics), Zhejiang University. *Micro/Nano Electron Intell Manufactur*, 2022, 4: 2–4 [张莹. 12英寸CMOS成套工艺公共研发平台建成,助力高校研究所成果实现产业化—访中国工程院院士,浙江大学微电子学院(微纳电子学院)吴汉明. *微纳电子与智能制造*, 2022, 4: 2–4]

- 2 Zhao W X, Zhou K, Li J, et al. A survey of large language models. ArXiv:2303.18223
- 3 Jiang Y, Lu X, Jin Q, et al. FabGPT: an efficient large multimodal model for complex wafer defect knowledge queries. In: Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design, 2024. 1–8
- 4 Jin Q, Liu Y, Jiang Y, et al. Unitho: a unified multi-task framework for computational lithography. In: Proceedings of the 2025 IEEE/ACM International Conference On Computer Aided Design (ICCAD), 2025. 1–9
- 5 Jiang Y, Jin Q, Lu X, et al. FabThink: a wafer analysis multimodal LLM via chain-of-thought-driven retrieval augmentation. In: Proceedings of the 44th IEEE/ACM International Conference on Computer-Aided Design, 2025. 1–8
- 6 Shao Z, Wang P, Zhu Q, et al. Deepseekmath: pushing the limits of mathematical reasoning in open language models. ArXiv: 2402.03300
- 7 Pu Y, He Z, Qiu T, et al. Customized retrieval augmented generation and benchmarking for EDA tool documentation QA. In: Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design, 2024. 1–9
- 8 Yu N, Chen H, Xu Q, et al. Wafer map defect recognition with few shot learning based on hybrid self-attention mechanism and prototype network. In: Proceedings of the 2021 China Automation Congress (CAC), 2021. 4128–4134
- 9 Geng H, Yang F, Zeng X, et al. When wafer failure pattern classification meets few-shot learning and self-supervised learning. In: Proceedings of the 2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD), 2021. 1–8
- 10 Jin Q, Jiang Y, Lu X, et al. SEM-CLIP: precise few-shot learning for nanoscale defect detection in scanning electron microscope image. In: Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design, 2024. 1–8
- 11 Kong Y, Ni D. A one-shot learning approach for similarity retrieval of wafer bin maps with unknown failure pattern. IEEE Trans Semicond Manufact, 2022, 35: 40–49
- 12 Chen Y, Wu Y, Wang J, et al. LLM-HD: layout language model for hotspot detection with GDS semantic encoding. In: Proceedings of the 61st ACM/IEEE Design Automation Conference, 2024. 1–6
- 13 Lu A, Wong H Y, et al. TCAD structure input file generation using large language model. In: Proceedings of the 2024 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2024. 1–4
- 14 Wu H, He Z, Zhang X, et al. ChatEDA: a large language model powered autonomous agent for EDA. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2024, 43: 3184–3197
- 15 Lu Y, Liu S, Zhang Q, et al. RTLML: an open-source benchmark for design RTL generation with large language model. In: Proceedings of the 2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC), 2024. 722–727
- 16 Huang S, Li J, Yu Z, et al. LLSM: LLM-enhanced logic synthesis model with EDA-guided CoT prompting, hybrid embedding and AIG-tailored acceleration. In: Proceedings of the 30th Asia and South Pacific Design Automation Conference, 2025. 974–980
- 17 Gao J, Cao W, Yang J, et al. Analoggenie: a generative engine for automatic discovery of analog circuit topologies. ArXiv: 2503.00205
- 18 Liu J X, Meng X F. Survey on privacy-preserving machine learning. J Comput Res Dev, 2020, 57: 346–362 [刘俊旭, 孟小峰. 机器学习的隐私保护研究综述. 计算机研究与发展, 2020, 57: 346–362]
- 19 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Commun ACM, 2020, 63: 139–144
- 20 Qiu X, Liu Y, Jin Q, et al. A diffusion-based approach to wafer defect image generation in semiconductor manufacturing. In: Proceedings of the 2025 Conference of Science and Technology of Integrated Circuits (CSTIC), 2025. 1–3
- 21 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017
- 22 Scarselli F, Gori M, Ah Chung Tsoi M, et al. The graph neural network model. IEEE Trans Neural Netw, 2008, 20: 61–80
- 23 Liu F, Zhang T, Dai W, et al. Few-shot adaptation of multi-modal foundation models: a survey. Artif Intell Rev, 2024, 57: 268
- 24 Lin Y, Li M, Watanabe Y, et al. Data efficient lithography modeling with transfer learning and active data selection. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2018, 38: 1900–1913
- 25 Yin Y, Chen R, Xu B, et al. Transfer learning for minimum operating voltage prediction in advanced technology nodes: leveraging legacy data and silicon odometer sensing. ArXiv:2509.00035
- 26 Zhou B, Jieming P, Sivan M, et al. Quantile online learning for semiconductor failure analysis. In: Proceedings of ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023. 1–5
- 27 Tran K T, Dao D, Nguyen M D, et al. Multi-agent collaboration mechanisms: a survey of LLMs. ArXiv:2501.06322
- 28 Zhou H, Wan X, Sun R, et al. Multi-agent design: optimizing agents with better prompts and topologies. ArXiv:2502.02533
- 29 Chen L, Wu Y, Wen C, et al. An agile framework for efficient LLM accelerator development and model inference. In: Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design, 2024. 1–9
- 30 Li J, Xu J, Huang S, et al. Large language model inference acceleration: a comprehensive hardware perspective. ArXiv:2410.04466
- 31 Zeng C, Liu S, Yang S, et al. GQSA: group quantization and sparsity for accelerating large language model inference. ArXiv:2412.17560
- 32 Lin J, Tang J, Tang H, et al. AWQ: activation-aware weight quantization for on-device LLM compression and acceleration. In:

Proceedings of MLSys, 2024. 87–100

- 33 Liang Y, Shi H, Shao H, et al. AccLLM: accelerating long-context LLM inference via algorithm-hardware co-design. ArXiv:2505.03745
- 34 Liu M, Ene T D, Kirby R, et al. ChipNeMo: domain-adapted LLMs for chip design. ArXiv:2311.00176
- 35 Nguyen C, Nguyen W, Suzuki A, et al. SemiKong: curating, training, and evaluating a semiconductor industry-specific large language model. ArXiv:2411.13802
- 36 Wang Z, He H, Wu T, et al. Lmlitho: A large vision model-driven lithography simulation framework. In: Proceedings of the 2025 IEEE/ACM International Conference On Computer Aided Design (ICCAD), 2025. 1–9
- 37 Viswanathan R, Azpiroz J, Selvam P. Process optimization through model based SRAF printing prediction. In: Proceedings of SPIE, 2012
- 38 Geng H, Zhong W, Yang H, et al. SRAF insertion via supervised dictionary learning. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2020, 39: 2849–2859
- 39 Li T, Tang Z, Wu T, et al. LLM-SRAF: sub-resolution assist feature generation using large language model. In: Proceedings of the 2025 Design, Automation & Test in Europe Conference (DATE), 2025. 1–7
- 40 Liu Q, Zhang L. Optimize the OPC control recipe with cost function. Photomask Technology. In: Proceedings of SPIE, 2010
- 41 Chen G, Yang H, Yu B, et al. Intelligent OPC engineer assistant for semiconductor manufacturing. AAAI, 2025, 39: 23144–23151
- 42 Wen W Y, Li J C, Lin S Y, et al. A fuzzy-matching model with grid reduction for lithography hotspot detection. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2014, 33: 1671–1680
- 43 Yang H, Su J, Zou Y, et al. Layout hotspot detection with feature tensor generation and deep biased learning. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2019, 38: 1175–1187
- 44 Torres J. ICCAD-2012 CAD contest in fuzzy pattern matching for physical verification and benchmark suite. In: Proceedings of the 2012 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2012. 349–350
- 45 Geng H, Yang H, Zhang L, et al. Hotspot detection via attention-based deep layout metric learning. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2022, 41: 2685–2698
- 46 Jeong H J, Choi J Y, Cho H M, et al. MOBO-driven advanced sub-3-nm device optimization for enhanced PDP performance. IEEE Trans Electron Devices, 2024, 71: 2881–2887
- 47 Liu S, Gao C, Li Y. Large language model agent for hyper-parameter optimization. ArXiv:2402.01881
- 48 Jeong C, Myung S, Huh I, et al. Bridging TCAD and AI: its application to semiconductor design. IEEE Trans Electron Devices, 2021, 68: 5364–5371
- 49 Zhao R, Chen H, Wang W, et al. Retrieving multimodal information for augmented generation: a survey. ArXiv:2303.10868
- 50 Jiang Y, Jin Q, Lu X, et al. Fabsage: a large multimodal model for IC defect detection, analysis, and knowledge querying. In: Proceedings of the 2024 IEEE LLM Aided Design Workshop (LAD), 2024. 1–5
- 51 Akbar M, Jazlan A, Ibrahim A, et al. Deep learning-based yield prediction for the die bonding semiconductor manufacturing process. In: Proceedings of the International Conference on Electrical, Control & Computer Engineering, 2023. 75–82
- 52 Ni Y. Analysis of the impact of artificial intelligence technology on integrated circuit manufacturing. Appl IC, 2024, 41: 41–43 [倪韵. 人工智能技术对集成电路制造的影响分析. 集成电路应用, 2024, 41: 41–43]
- 53 Thakur S, Ahmad B, Pearce H, et al. VeriGen: a large language model for Verilog code generation. ACM Trans Des Autom Electron Syst, 2023, 29: 1–31
- 54 Liu S, Wang W, Lu Y, et al. RTLCode: fully open-source and efficient LLM-assisted RTL code generation technique. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2025, 44: 1448–1461
- 55 Li X, Wang M, Han Y, et al. ChipGPT: how far are we from natural language hardware design. ArXiv:2305.14019
- 56 Blocklove J, Garg S, Karri R, et al. Chip-Chat: challenges and opportunities in conversational hardware design. ArXiv:2305.13243
- 57 Fu Y, Zhang Y, Yu Z, et al. GPT4AIGChip: towards next-generation AI accelerator design automation via large language models. ArXiv:2309.10730
- 58 Chen H C, Wu L, Gao M, et al. DocEDA: automated extraction and design of analog circuits from documents with large language model. ArXiv:2412.05301
- 59 Wang R, Huang R, Lin Y, et al. Layoutcopilot: an LLM-powered multi-agent collaborative framework for interactive analog layout design. ArXiv:2406.18873

Large models for integrated circuits: driving leapfrogging advances in IC manufacturing technologies

Qi SUN¹, Jinyuan DENG¹, Qian JIN¹, Yu LI¹, Zheyu YAN¹, Dawei GAO^{1,2,4}, Shisheng XIONG³, Zhenhua WU¹, Cheng ZHUO^{1*}, Ningsheng XU³ & Hanming WU¹

1. *College of Integrated Circuits, Zhejiang University, Hangzhou 311200, China*

2. *Zhejiang ICsprout Semiconductor Co., Ltd., Hangzhou 311200, China*

3. *Institute of Nanoscience and Nanotechnology, Fudan University, Shanghai 200433, China*

4. *Zhejiang Technology Innovation Center of CMOS IC Manufacturing Process and Design, Hangzhou 311200, China*

* Corresponding author. E-mail: czhuo@zju.edu.cn

Abstract With the continuous evolution of advanced process nodes, the complexity of integrated circuit (IC) manufacturing has increased significantly, raising the demand for cross-layer, multimodal, and intelligent manufacturing capabilities. Traditional approaches, relying primarily on expert experience, physical mechanisms, and mathematical models, show clear limitations in handling process iterations, workflow optimization, and cross-scenario technology transfer. This underscores the urgent need for intelligent technological systems with stronger generalization ability and industrial adaptability. In recent years, foundation models have demonstrated exceptional knowledge abstraction, semantic understanding, and generalization capabilities across multiple domains, offering new opportunities to advance intelligent IC manufacturing. However, IC manufacturing is typically characterized by heterogeneous, multimodal data, small sample availability, complex process couplings, and inherent trade-offs between efficiency and accuracy, making it difficult to apply general-purpose foundation models directly. Therefore, it is imperative to organically integrate multimodal data, physical mechanisms, and domain knowledge into model design to build dedicated foundation models tailored for manufacturing processes. This paper focuses on the core difficulties and key challenges of integrated circuit manufacturing, introduces the current development status of IC foundation models, and outlines future development pathways.

Keywords integrated circuit, design-technology co-optimization, intelligent manufacturing, artificial intelligence, large language model