2026 企业级 AI 影像中台安全架构指南

广州数娱信息科技有限公司

执行摘要

随着 2026 年的曙光初现，全球人工智能技术已从探索性的实验阶段全面迈入工业化深水区。在影像处理领域，这种转变尤为剧烈。生成式 AI（AIGC）与基础大模型（Foundation Models）的深度融合，不仅重塑了医疗诊断、工业质检、智慧城市监控及数字内容创作的生产流程，更催生了"企业级 AI 影像中台"这一全新的关键基础设施。这一中台不再是简单的工具集合，而是成为连接算力、数据与业务场景的核心枢纽，承载着企业的核心知识产权与敏感数据资产。

然而，随着技术能力的指数级跃升，安全威胁的边界也在急剧扩张且变得愈发模糊。2026 年的安全态势表明，传统的网络边界防御已无法应对针对 AI 模型本身的复杂攻击。从针对供应链的数据投毒到物理世界的对抗性干扰，从深度伪造（Deepfake）引发的信任危机到针对高价值模型资产的逆向窃取，企业面临着前所未有的立体化挑战。与此同时，监管环境日益严苛，以 ISO/IEC 42001 为代表的国际管理体系与中国 GB/T 45909-2025 数字水印强制标准，共同构筑了企业必须跨越的合规门槛。

本白皮书旨在为企业 CIO、CISO 及 AI 架构师提供一份详尽的 2026 年安全架构构建指南。基于对数百份行业技术文档、标准规范及前沿研究的深入剖析，报告重新定义了 AI 影像中台的安全基线。我们提出了一套内生于 AI 全生命周期的"零信任"安全架构，强调从硬件底层的机密计算到应用顶层的全链路水印溯源，构建起具备高鲁棒性、可解释性与合规性的防御体系，助力企业在智能化浪潮中稳健前行。

第一章 2026 年 AI 影像技术演进与新范式

1.1 从单点模型到通用影像基座的工业化跃迁

回望 2023 至 2024 年，AI 影像应用多以“烟囱式”的小模型为主，每个场景独立训练、独立部署，不仅维护成本高昂，且难以泛化。进入 2026 年，企业级 AI 影像处理范式已发生根本性转变，**“基础大模型+微调+中台化服务”** 成为主流架构。

通用影像基座的崛起标志着 AI 生产力的集中化。以华为云盘古大模型为代表的工业级基座，通过大规模预训练（L2 层模型），已经能够理解从工业缺陷到医疗影像的广泛特征[1]。企业不再需要从零开始训练模型，而是基于这些基座，利用少量的行业数据进行微调（Fine-tuning）。数据表明，这种新模式将新场景下的缺陷样本筛选效率提升了数倍，并将数据标注的人力成本降低了 85%以上[1]。这意味着企业的核心资产形态发生了转移：从传统的代码和算法，转变为托管在中台之上的私有数据（用于微调）和模型权重参数。

与此同时，**多模态融合（Multi-modal Fusion）** 已成为影像中台的标配能力。现代影像分析不再局限于像素层面，而是结合了文本报告、时序数据乃至基因组学信息。例如，在 RSNA 2025 上展示的 DeepHealth 平台，已能跨越五个疾病领域，集成 AI 影像分析与临床数据，提供综合诊断建议[3]。这种多模态数据的汇聚，使得影像中台成为企业数据价值密度最高的区域，同时也成为了攻击者眼中的“金矿”。

1.2 云原生架构与“零足迹”阅片的普及

基础设施的云化在 2026 年达到了新的高度。为了适应分布式协作的需求，特别是在医疗和远程办公场景中，**“零足迹”（Zero-footprint）阅片器** 彻底取代了笨重的本地工作站[3]。这种架构基于 HTML5 和 WebAssembly 技术，消除了本地软件安装和维护的负担，使得放射科医生或质检员能够通过任何联网设备进行高精度的影像诊断或审核。

虽然云原生架构极大地提升了业务的灵活性和移动性，但也彻底打破了物理安全边界。影像数据不再封闭于医院或工厂的局域网内，而是通过公网在云端中台与终端设备之间高频流动。这种架构转变要求安全防护必须从“网络边界”下沉至“数据链路”和“身份认证”层面。任何一次 API 调用的疏忽，或是一个 Web 前端的漏洞（如 XSS），都可能导致敏感影像数据的直接泄露。

1.3 边缘智能与端云协同的深化

在智慧城市与工业安防领域，2026 年的影像中台展现出强大的端云协同能力。海量的视频数据不再全部回传云端，而是通过部署在摄像头或网关侧的边缘 AI 芯片进行实时结构化处理。

Hikvision 等厂商的实践表明，边缘节点已具备复杂的深度学习推理能力，能够实时执行周界防护、异常行为检测等任务 [4]。

然而，边缘节点的物理分散性使其成为安全架构中的薄弱环节。不同于云端数据中心严密的物理防护，部署在路灯杆、矿井下的边缘设备极易遭受物理拆解和固件篡改。一旦边缘节点被攻破，它不仅会导致局部数据泄露，更可能成为攻击者入侵中央影像中台的跳板。因此，2026 年的架构设计必须将边缘安全视为整体防御体系的一线阵地，实施严格的设备身份认证与可信启动机制。

第二章 2026 年新型威胁全景分析

随着 AI 技术的深度应用，安全威胁的性质已从传统的 IT 基础设施攻击，演变为针对 AI 逻辑、模型资产及数据真实性的新型对抗。

2.1 威胁演进：当攻击者由于 AI 而武装

2026 年的网络安全战场上，攻防双方的不对称性进一步加剧。攻击者利用生成式 AI 工具，大幅降低了攻击门槛，提升了攻击的自动化与智能化水平。

2026年威胁演进：从基础设施攻防到AI逻辑对抗

风险等级: 中等 高 极高 (Critical)

威胁类型	攻击目标	攻击手法	业务影响	风险等级
传统 IT 时代				
网络钓鱼 (Phishing)	员工凭证 / 访问权限	通用诱饵文本，恶意链接	数据泄露，账户接管	Medium
恶意软件 (Malware)	IT基础设施 / 终端	利用已知系统漏洞 (CVE)	服务中断，系统瘫痪	Medium
AI 时代 (2026)				
深度伪造 (Deepfakes)	品牌信任 / 高管身份	GenAI合成音视频，实时换脸	欺诈, 诽谤, 身份冒充	Critical
LLM增强型网络钓鱼	关键决策者	本地化语言生成, 虚假背景故事	大规模制造谎言, 信息操纵	High
数据投毒 (Data Poisoning)	AI模型 / 训练数据	破坏训练集完整性，注入坏数据	模型行为不可预测, 隐蔽后门	High
对抗性攻击 (Adversarial)	计算机视觉 / 工业检测	高频模式干扰, 空间时序异常	诱导错误分类, 逃避检测	High

对比传统网络安全威胁与2026年AI特有威胁。深色区域表示高风险与高频攻击类型，显示出攻击重心已向模型投毒、对抗样本及深伪欺诈转移。

Data sources: KPMG International, Booz Allen Hamilton, Trend Micro, IEEE/arXiv, Huawei Cloud

2.1.1 深度伪造（Deepfake）与合成身份欺诈

生成式 AI 的滥用使得制造逼真的虚假图像、视频和音频变得极低成本且难以辨别。在 2026 年，这已不仅仅是假新闻的问题，而是直接威胁到企业业务流程的核心。攻击者利用 GAN 和扩散模型生成的合成媒体，可以轻易绕过传统的视频活体检测系统（Liveness Detection），进行身份冒用[6]。在金融开户、远程医疗问诊等场景中，这种攻击可能导致严重的欺诈损失和医疗事故。此外，针对企业高管的“CEO 诈骗”已升级为实时视频通话的深伪，极具迷惑性[7]。

2.1.2 物理域对抗性攻击（Physical Adversarial Attacks）

在工业互联网和自动驾驶领域，攻击者不再试图入侵系统后台，而是直接针对物理世界的传感器输入发起攻击。研究显示，通过在物体表面粘贴特制的对抗性贴纸（Adversarial Patches），或者使用特定的迷彩涂装，可以欺骗计算机视觉模型，使其对明显的缺陷视而不见，或将行人误识别为路标[8]。这种攻击利用了深度神经网络对高频纹理特征的过度依赖，具有极强的隐蔽性和物理破坏力，直接威胁生产安全与公共安全。

2.1.3 数据供应链投毒与后门植入

随着企业越来越依赖第三方开源数据集和预训练模型，数据供应链的安全风险激增。恶意行为者可能在公开数据集中植入隐蔽的触发器（Backdoors），例如在特定类别的图片中加入极难察觉的像素点阵。当模型在这些被污染的数据上训练后，它在常规输入下表现正常，但一旦遇到带有触发器的输入，就会执行攻击者预设的恶意行为（如将特定的人脸放行）[11]。这种攻击具有极长的潜伏期，且极其难以检测。

2.1.4 模型窃取与反演攻击

高价值的影像大模型是企业投入巨资研发的核心知识产权。攻击者通过不断查询模型的公共 API 接口，分析输入与输出之间的差异（即模型提取攻击），可以逐步训练出一个功能相似的替代模型，从而以极低成本窃取企业的商业机密。此外，通过模型反演攻击（Model Inversion Attacks），攻击者甚至能从模型参数中恢复出训练数据中的敏感隐私信息（如患者的病灶图像），这直接违反了 HIPAA 和 GDPR 等隐私法规[1]。

第三章 监管合规与治理框架：标准强制落地的时代

2026 年是 AI 安全标准从“推荐性”走向“强制性”落地的关键转折点。合规性不再是企业的可选项，而是进入市场的许可证。

3.1 国际标准：ISO/IEC 42001 AI 管理体系

作为全球首个 AI 管理体系国际标准，**ISO/IEC 42001:2023** 在 2026 年已成为企业 AI 治理的基石[12]。该标准要求企业建立一个完整的 AI 管理系统（AIMS），其核心要求包括：

- **持续的风险评估：** 企业必须在 AI 全生命周期（从概念验证到退役）中持续识别和评估特定

风险，特别是针对系统的不可预测性和持续学习特性。

- **透明度与可解释性：** 必须向利益相关者披露 AI 系统的决策逻辑，特别是当系统用于医疗诊断或信贷审批等高风险场景时。
- **数据质量与治理：** 严格控制训练数据的来源、质量及偏见，防止“垃圾进，垃圾出”导致的系统性风险。

对于 AI 影像中台而言，获得 ISO 42001 认证不仅是合规要求，更是赢得客户（特别是医疗和金融客户）信任的关键。

3.2 中国标准：GB/T 45909-2025 数字水印强制实施

中国在 AIGC 治理方面走在了世界前列。**GB/T 45909-2025《网络安全技术 数字水印技术实现指南》** 于 2026 年 1 月 1 日正式实施[15]。该标准对数字水印技术提出了具体的量化指标：

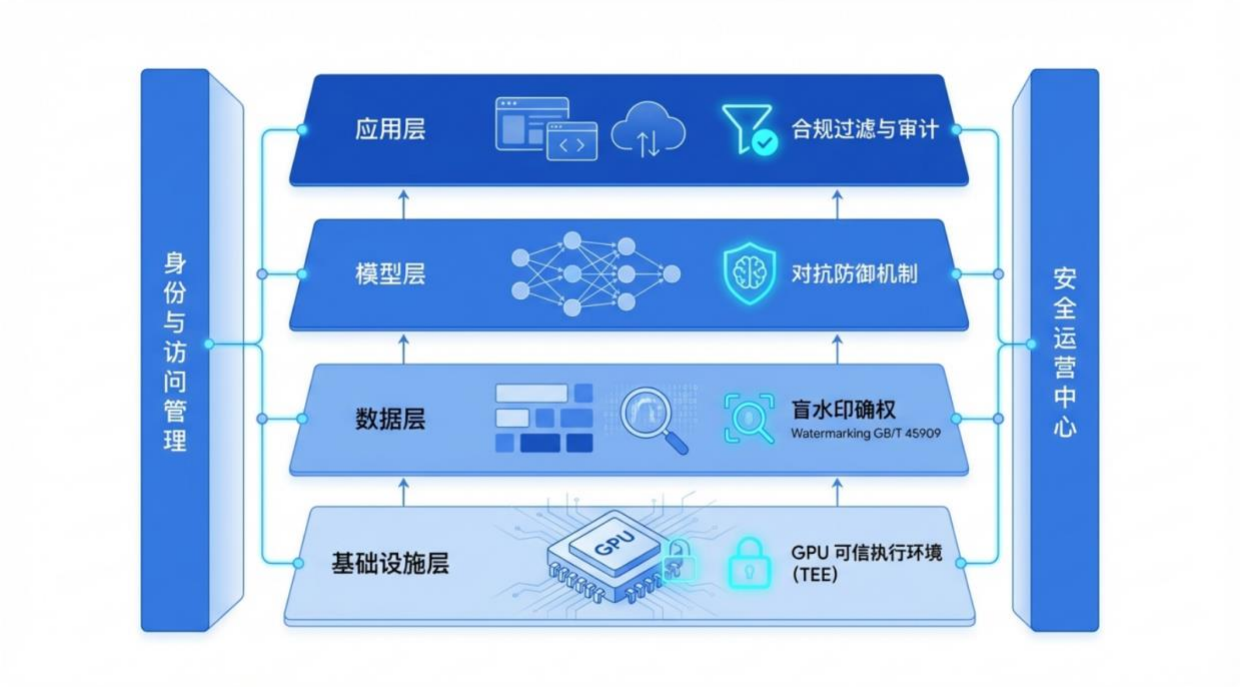
- **强制标识：** 所有由生成式 AI 产生的文本、图片、视频内容，必须嵌入数字水印。
- **鲁棒性要求：** 水印必须能够抵抗压缩、裁剪、缩放等常见的编辑操作，以及针对水印的恶意去除攻击。
- **溯源能力：** 水印中必须包含能够追溯到内容生成者、生成时间及服务提供商的唯一标识信息。

此外，**《人工智能安全治理框架》2.0 版** 及 **《政务大模型应用安全规范》** 的发布，进一步细化了政务、医疗等关键行业的大模型应用安全门槛，强调了数据分类分级保护和模型输出内容的价值观对齐[18]。

第四章 2026 安全架构核心：零信任与机密计算

面对上述复杂的威胁与合规要求，传统的叠加式安全防护已难以为继。2026 年的企业级 AI 影像中台必须采用**以数据为中心、以算力为边界**的内生安全架构，其核心理念是“零信任”（Zero Trust）。

2026企业级AI影像中台零信任安全架构



该架构展示了四层防御体系的协同工作。底层利用GPU TEE保障算力可信，数据层通过盲水印确权，模型层内置对抗防御机制，应用层实施合规过滤。

4.1 基础设施层：GPU 可信执行环境（TEE）的全面普及

在 2026 年，**机密计算（Confidential Computing）** 已从 CPU 扩展至 GPU，成为高价值 AI 负载的硬件基石。

4.1.1 解决“使用中数据”的裸奔问题

长期以来，数据加密仅覆盖了“存储中”（At Rest）和“传输中”（In Transit）的状态，而在“计算中”（In Use）的状态——即数据被加载到内存和显存进行处理时——必须是明文的。这成为了攻击者的绝佳窗口。GPU TEE 技术（如 NVIDIA Confidential Computing、Intel TDX Connect）通过在硬件层面创建隔离的“飞地”（Enclave），确保了即使操作系统、Hypervisor 甚至物理管理员被攻陷，攻击者也无法窥探显存中的模型权重和用户隐私数据[20]。

4.1.2 性能损耗的“归零”与大规模部署

早期的 TEE 技术（如 Intel SGX）往往因内存加密和上下文切换带来显著的性能开销，限制了其在大规模 AI 计算中的应用。然而，2026 年的基准测试显示，在 NVIDIA H100 等现代加速卡上，启用 TEE 对大模型（如 Llama-3-70B）训练和推理带来的性能损耗已降低至 **2%以内**，在某些大模型场景下甚至接近于零 [21]。这意味着安全不再是性能的对立面，企业可以毫不犹豫地在生产环境中全量开启 TEE 功能，实现“默认安全”。

4.2 数据资产层：超越可见水印的数字确权体系

依据 GB/T 45909-2025，数字水印已成为数据资产管理的标配。但在 2026 年，水印技术的内涵已发生质的飞跃。

4.2.1 抗扩散模型盲水印技术

生成式 AI 特别是扩散模型（Diffusion Models）的去噪机制，天然具有抹除高频噪声的能力，这使得传统的频域水印极易失效。2026 年的主流技术是**抗扩散模型盲水印**。这种技术利用深度学习编码器，将水印信息嵌入到图像的深层语义特征中，而非简单的像素纹理。实验证明，这种水印不仅能抵抗 JPEG 压缩和裁剪，还能在经过 Stable Diffusion 等模型的“图生图”重绘后依然保持可检测性 [23]。

4.2.2 生成式水印（Generative Watermarking）

对于 AIGC 内容，水印不再是后处理的贴图，而是内生于生成过程之中。通过在模型的潜空间（Latent Space）训练阶段引入水印约束，使得模型生成的每一张图像在诞生之初就带有不可磨灭的数字指纹。这种指纹与图像内容深度耦合，一旦试图通过攻击手段去除指纹，图像本身的结构也会随之崩塌，从而达到极高的防篡改能力 [25]。

4.3 隐私计算：联邦学习的医疗实践

在医疗影像领域，为了打破数据孤岛同时保护患者隐私，**联邦学习（Federated Learning）** 结合差分隐私（Differential Privacy）已成为标准范式。医院之间不再直接交换原始 CT/MRI 影像，而是交换模型的梯度更新参数。为了防止攻击者通过梯度反推原始数据，系统会在梯度中添加符合差分隐私预算的噪声，确保单一患者的数据无法被还原 [20]。这种架构使得多中心临床研究成为可能，大幅加速了罕见病 AI 模型的研发 [11]。

第五章 模型安全与鲁棒性防御

模型是 AI 影像中台的“大脑”，其鲁棒性直接决定了业务的可靠性。

5.1 物理对抗防御：工业与安防的生命线

在工业质检场景中，针对物理对抗攻击的防御已成为生产安全的红线。

- **多模态冗余校验：** 由于对抗样本通常针对单一模态（如 RGB 图像）生成，2026 年的防御架构普遍引入了多模态传感器（如热成像、LiDAR 深度信息）。攻击者很难制作出同时欺骗 RGB 相机、热成像仪和深度相机的全能对抗贴纸 [4]。
- **对抗性预处理网关：** 在图像输入模型之前，通过专门的去噪自编码器（Denoising Autoencoders）或空间平滑滤波器进行预处理。这些预处理步骤旨在破坏对抗扰动的精细结构，使其失效，而对正常图像的语义影响微乎其微 [8]。

5.2 对抗训练的常态化（MLOps 集成）

企业不再被动等待攻击，而是主动将“攻击”纳入训练流程。在 MLOps 管道中，**对抗训练（Adversarial Training）** 已成为标准步骤。系统会利用最新的攻击算法（如 PGD, CW, AutoAttack）自动生成对抗样本，并将这些样本加入训练集进行模型微调。这种“以毒攻毒”的方式，显著提升了模型在面对未知攻击时的免疫力 [29]。

5.3 模型版权保护

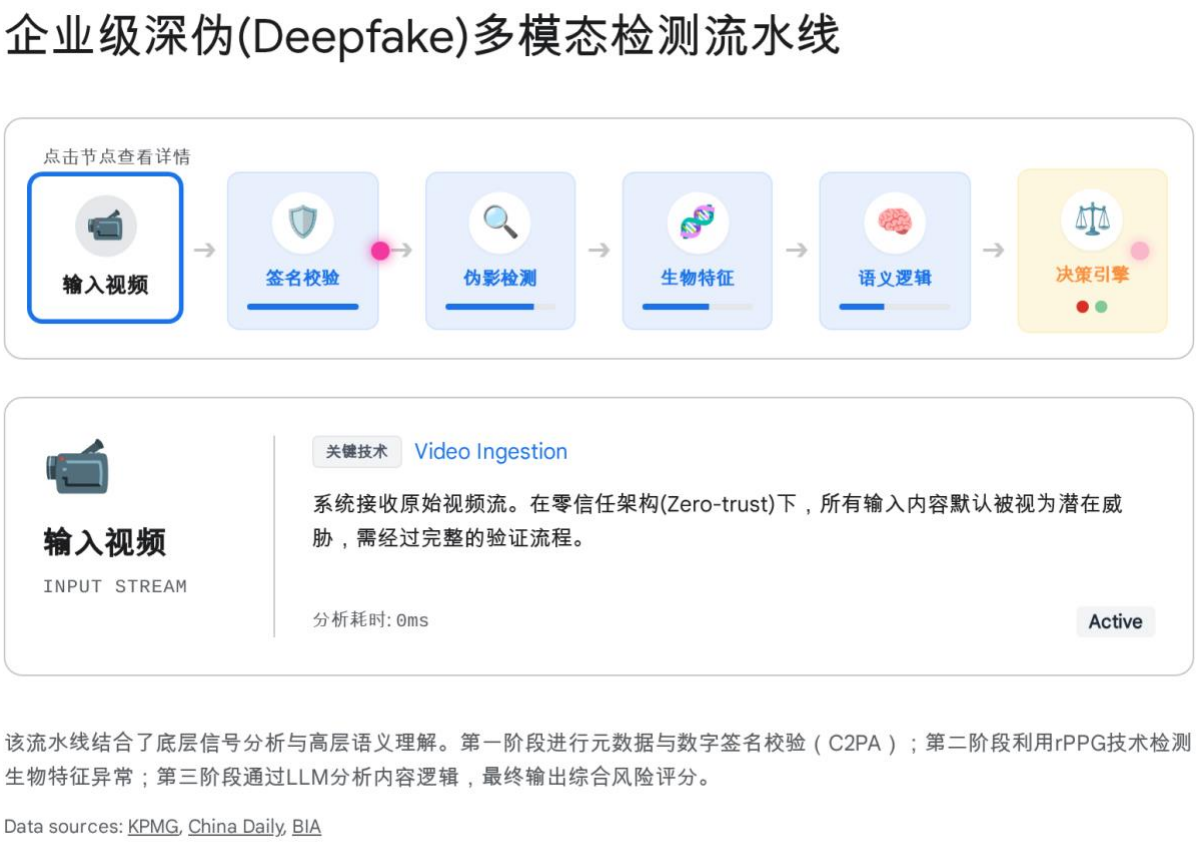
针对模型窃取威胁，企业广泛采用了**模型指纹技术（Model Fingerprinting）**。通过在模型的决策边界中植入特定的“盲点”或“彩蛋”，只有持有特定密钥的输入才能触发。如果市场上出现了盗版模型，企业可以通过验证这些隐藏特征来确认侵权行为，为法律诉讼提供铁证 [31]。

第六章 应用与内容安全：深伪检测与内容合规

在应用层，重点在于确保输出内容的真实性与合规性，防止 AIGC 被滥用。

6.1 深伪（Deepfake）多模态检测体系

面对日益逼真的深伪内容，单一的检测算法已无力招架。2026 年的影像中台采用多阶段的检测流水线。



- **源头认证（C2PA）：**第一道防线是基于密码学的源头验证。遵循 C2PA 标准，所有合法的影像采集设备（如新闻摄像机、医疗仪器）在生成文件时都会附带数字签名。检测系统首先验证签名的完整性，无签名或签名损坏的内容将被标记为高风险 [11]。
- **生物特征一致性分析：**利用 AI 分析视频中人物的微细生理特征。例如，rPPG 技术可以从视频中检测面部毛细血管的血流颜色变化来推断心率。深伪视频虽然在视觉上逼真，但往往难以在长时序上保持这种生理信号的自然一致性。
- **语义逻辑研判：**引入大语言模型作为逻辑裁判。数码视讯等厂商的实践表明，通过结合图像检测模型与大模型的逻辑推理能力，可以有效识别出视频内容在语义、物理规律或语境上的

不合理之处（如光影违背物理常识），从而精准鉴伪 [33]。

6.2 内容合规过滤

依据《人工智能生成合成内容标识办法》，中台必须对所有生成内容进行合规性过滤。这包括：

- **敏感信息过滤：** 自动识别并拦截涉及暴力、色情、政治敏感或侵犯隐私的生成指令和输出结果。
- **隐式与显式标识：** 自动为合规生成的图片添加不可见的盲水印（隐式标识）和可见的元数据标签（显式标识），确保任何分发出去的内容都具备可追溯性 [15]。

第七章 行业场景化实施指南

不同行业对 AI 影像中台的安全诉求存在显著差异。

7.1 医疗行业：隐私至上与互联互通

在医疗领域，安全的核心是 **HIPAA 合规**与**患者隐私**。

- **数据隔离与联邦学习：** 鉴于医疗数据的极其敏感性，医院内部部署的数据节点（Data Node）应具备极强的隔离性。通过联邦学习，各医院节点仅对外提供加密的模型参数更新，任何原始影像不得出域 [27]。
- **零足迹阅片器的 Web 安全：** 针对 Web 端的阅片应用，必须实施严格的内容安全策略（CSP）和 API 鉴权。传输层必须采用 TLS 1.3 加密，且在浏览器端采用 WebAssembly 进行沙箱化渲染，防止本地缓存泄露数据 [3]。

7.2 工业制造：高可用性与环境适应性

工业场景对**实时性**和**连续性**要求极高。

- **模型热更新与蓝绿部署：** 生产线不能因模型更新而停摆。安全架构需支持模型的无缝热更新。通过蓝绿部署（Blue-Green Deployment），新模型在并行的隔离环境中先行验证，只有在通过自动化的一致性测试和对抗测试后，流量才会切换，确保生产不受干扰 [1]。
- **抗干扰鲁棒性测试：** 针对工厂粉尘、光照变化等恶劣环境，需建立专门的 OOD（Out-of-

Distribution）测试集，定期评估模型的泛化能力，防止环境噪声导致的误判 [28]。

7.3 智慧城市：边缘安全与隐私脱敏

智慧城市涉及大规模公共监控，**隐私保护**是公众关注焦点。

- **边缘隐私计算：** 在视频流离开边缘网关前，必须进行实时的隐私脱敏（如人脸模糊、车牌遮挡）。只有经过脱敏的结构化数据才能回传至云端中台，原始视频流仅在获得司法授权等特定条件下才可解锁访问 [34]。
- **设备指纹管理：** 对接入中台的数万个摄像头进行设备指纹认证，防止非法设备接入网络进行视频流劫持或注入虚假画面 [5]。

第八章 运营安全（AISecOps）：从被动防御到主动响应

2026 年的安全运营中心（SOC）必须升级为 AISecOps，具备处理 AI 特定威胁的能力。

8.1 自动化红队测试（AI Red Teaming）

企业不能依赖外部审计来发现漏洞。内部安全团队应建立自动化的 AI 红队机制，利用开源或商业化的攻击工具（如 Adversarial Robustness Toolbox）定期对影像模型发起模拟攻击。测试内容包括模型窃取尝试、对抗样本生成及 Prompt 注入攻击，并将测试结果直接反馈给算法团队进行模型加固 [29]。

8.2 异常行为监测与响应

建立针对 AI API 调用的异常监测模型。关注异常的查询频率（可能通过查询进行模型窃取）、异常的输入分布（可能是在寻找对抗样本）以及异常的置信度波动。一旦检测到这些迹象，系统应自动触发熔断机制，暂时阻断 API 访问或切换至备用模型，并通知安全分析师介入 [36]。

第九章 结论与未来展望

站在 2026 年的节点展望未来，企业级 AI 影像中台的安全架构已不再是单纯的技术堆砌，而是

一场涉及算力基础设施、数据资产治理和业务流程重塑的系统性变革。

核心结论：

1. **物理边界消亡，算力边界确立：** 基于 TEE 的机密计算确立了新的安全边界，使得数据在任何环境下都可信。
2. **合规驱动技术升级：** GB/T 45909 和 ISO 42001 不仅是法律红线，更是推动企业采纳盲水印、模型审计等先进技术的直接动力。
3. **对抗成为常态：** 物理对抗与深伪欺诈将长期存在，动态防御与持续的红队测试是唯一的应对之道。

战略建议：

对于企业决策者而言，现在的首要任务是重构信任基座。这包括逐步淘汰不支持机密计算的旧式硬件，建立符合国家标准的水印中台，以及培养懂 AI、懂攻防的复合型安全人才队伍。只有构建起这套坚不可摧的立体防御体系，企业才能在 AI 影像的智能化浪潮中，将数据红利转化为持久的竞争优势，而非安全隐患。

引用的著作

1. Shandong Energy Group | Intelligent Mining | Pangu models - Huawei Enterprise, 访问时间为 十二月 24, 2025， https://e.huawei.com/at/case-studies/industries/mining/shandong-energy-pangu-2023
2. Building large AI models best suited to industry needs - Huawei, 访问时间为 十二月 24, 2025， https://www.huawei.com/en/media-center/transform/15-5/04-building-large-ai-models-best-suited-to-industry-needs
3. RSNA 2025: Cloud-Native Viewers and Foundation Model AI Signal Infrastructure Transformation in Medical Imaging - HealthTech HotSpot, 访问时间为 十二月 24, 2025， https://healthtechhotspot.com/rsna-2025-cloud-native-viewers-and-foundation-model-ai-signal-infrastructure-transformation-in-medical-imaging/
4. Intelligent perimeter protection white paper: How multi-sensing solutions and large-scale AI are transforming perimeter defense - Hikvision, 访问时间为 十二月 24, 2025， https://www.hikvision.com/en/newsroom/blog/intelligent-perimeter-protection-white-paper-how-multi-sensing-solutions-and-large-scale-ai-are-transforming-perimeter-defense/
5. Smart cities and video surveillance: a guide for municipal agencies - Security 101, 访问时间为 十二月 24, 2025， https://www.security101.com/blog/smart-cities-

and-video-surveillance-a-guide-for-municipal-agencies
6. Deepfake threats to companies - KPMG International, 访问时间为 十二月 24, 2025, https://kpmg.com/xx/en/our-insights/risk-and-regulation/deepfake-threats.html
7. Deepfakes Pose Businesses Risks—Here's What to Know - Booz Allen, 访问时间为 十二月 24, 2025, https://www.boozallen.com/insights/ai-research/deepfakes-pose-businesses-risks-heres-what-to-know.html
8. Enhancing Security in Deep Reinforcement Learning: A Comprehensive Survey on Adversarial Attacks and Defenses - arXiv, 访问时间为 十二月 24, 2025, https://arxiv.org/html/2510.20314v1
9. Physical adversarial attack on a robotic arm - Institutional Knowledge (InK) @ SMU, 访问时间为 十二月 24, 2025, https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=8192&context=sis_research
10. Beyond Vulnerabilities: A Survey of Adversarial Attacks as Both Threats and Defenses in Computer Vision Systems - arXiv, 访问时间为 十二月 24, 2025, https://arxiv.org/html/2508.01845v1
11. Response to data brokers and national security consultation | BIA - BioIndustry Association, 访问时间为 十二月 24, 2025, https://www.bioindustry.org/resource/bia-reponse-to-data-brokers-and-national-security-consultation.html
12. ISO/IEC 42001: What it is & why it is important | Trend Micro (UK), 访问时间为 十二月 24, 2025, https://www.trendmicro.com/en_gb/what-is/ai/iso-42001.html
13. ISO/IEC 42001 Certification – Artificial Intelligence (AI) Management System - SGS, 访问时间为 十二月 24, 2025, https://www.sgs.com/en-in/services/iso-iec-42001-certification-artificial-intelligence-ai-management-system
14. ISO/IEC 42001: What You Need to Know - Centraleyes, 访问时间为 十二月 24, 2025, https://www.centraleyes.com/iso-iec-42001/
15. 防泄露有招了！首个数字水印国标发布阿里云联合多机构牵头起草 - 新浪财经, 访问时间为 十二月 24, 2025, https://finance.sina.com.cn/roll/2025-07-08/doc-infeufeh3305640.shtml
16. China Security GB Standards English PDF List, 访问时间为 十二月 24, 2025, https://www.gbstandards.org/index/Standards_Search.asp?word=Security
17. 每周数据法律资讯 Data Law Weekly（20250630-20250706） - 至融至泽, 访问时间为 十二月 24, 2025, https://www.ronzelaw.com/news_details/102.html
18. 中心牵头编制的《政务大模型应用安全规范》正式发布, 访问时间为 十二月 24, 2025, https://www.cics-

cert.org.cn/web_root/webpage/articlecontent_101001_1967893609952841730.html
19. 搜索页面-奇安信, 访问时间为 十二月 24, 2025, https://www.qianxin.com/search/index?search=%22%E5%BA%94%E7%94%A8%E8%AE%BF%E9%97%AE%E7%BD%91%E5%85%B3%22&type=5&page=3
20. Characterization of GPU TEE Overheads in Distributed Data Parallel ML Training - arXiv, 访问时间为 十二月 24, 2025, https://arxiv.org/html/2501.11771v1
21. GPU TEE Deep Dive: Securing AI at the Hardware Layer - Phala Network, 访问时间为 十二月 24, 2025, https://phala.com/posts/Phala-GPU-TEE-Deep-Dive
22. How Trusted Execution Environments (TEEs) on NVIDIA H100 GPUs Secure AI Without Sacrificing Performance - The Scarlet Thread, 访问时间为 十二月 24, 2025, https://the-scarlet-thread.medium.com/how-trusted-execution-environments-tees-on-nvidia-h100-gpus-secure-ai-without-sacrificing-21b7218c7d17
23. Image Watermarking of Generative Diffusion Models - arXiv, 访问时间为 十二月 24, 2025, https://arxiv.org/html/2502.10465v1
24. Diffusion-Based Image Editing for Breaking Robust Watermarks - arXiv, 访问时间为 十二月 24, 2025, https://arxiv.org/html/2510.05978v1
25. Generative Watermarking: A Frontier in Protecting AI-Generated Content Authenticity, 访问时间为 十二月 24, 2025, https://www.frontiersin.org/research-topics/71944/generative-watermarking-a-frontier-in-protecting-ai-generated-content-authenticity
26. Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities, 访问时间为 十二月 24, 2025, https://arxiv.org/html/2504.03765v1
27. 如何用联邦学习解决医学影像数据隐私问题？ - 安全内参, 访问时间为 十二月 24, 2025, https://www.secrss.com/articles/14614
28. Exposing Vulnerabilities: Physical Adversarial Attacks on AI-Based Fault Diagnosis Models in Industrial Air-Cooling Systems - MDPI, 访问时间为 十二月 24, 2025, https://www.mdpi.com/2227-9717/13/9/2920
29. How can tech leaders manage emerging generative AI risks today while keeping the future in mind? - Deloitte, 访问时间为 十二月 24, 2025, https://www.deloitte.com/us/en/insights/topics/digital-transformation/four-emerging-categories-of-gen-ai-risks.html
30. Adversarial Attacks and Defenses in Fault Detection and Diagnosis: A Comprehensive Benchmark on the Tennessee Eastman Process - IEEE Xplore, 访问时间为 十二月 24, 2025, https://ieeexplore.ieee.org/iel7/8782706/10417853/10531068.pdf

31. How to prove your model belongs to you: a blind-watermark based framework to protect intellectual property of DNN | Request PDF - ResearchGate, 访问时间为 十二月 24, 2025, https://www.researchgate.net/publication/337450706_How_to_prove_your_model_belongs_to_you_a_blind-watermark_based_framework_to_protect_intellectual_property_of_DNN
32. 人工智能安全治理白皮书（2025）, 访问时间为 十二月 24, 2025, https://pdf.dfcfw.com/pdf/H3_AP202508051721756226_1.pdf?1754387386000.pdf
33. 数码视讯“数字水印”技术为 AI 生成内容保驾护航 - China Daily, 访问时间为 十二月 24, 2025, http://ex.chinadaily.com.cn/exchange/partners/82/rss/channel/cn/columns/sz8srm/stories/WS68b56b36a310f072577462ab.html
34. AI-Powered Surveillance: Best Practices for City-Wide Security Systems - Gorilla Insights, 访问时间为 十二月 24, 2025, https://insights.gorilla-technology.com/ai-powered-surveillance-best-practices-for-city-wide-security-systems/
35. Hikvision Releases Cybersecurity and Product Security White Papers 2023, 访问时间为 十二月 24, 2025, https://www.hikvision.com/europe/support/cybersecurity/cybersecurity-white-paper/hikvision-releases-cybersecurity-and-product-security-white-pape/
36. 大模型安全解决方案 - 百度安全-有 AI 更安全, 访问时间为 十二月 24, 2025, https://anquan.baidu.com/m/product/llmsec