专用Visual Encoder预训练计划

1. 项目目标与核心愿景

本项目旨在为长鑫存储构建一个自主可控、面向芯片制造领域深度优化的Visual Encoder（视觉编码器）。其核心愿景是：以学术界前沿的AIMV2多模态自回归预训练架构为技术蓝图，构建一个不仅具备顶尖通用文档解析能力，更在芯片制造核心文档（如JEDEC标准、EFA/PFA报告）的理解上实现显著超越的基础视觉模型。

该模型将作为公司下一代多模态大模型（MLLM）的核心“眼睛”，为工艺缺陷诊断、测试报告分析、知识库构建等核心业务场景提供强大的视觉理解基石。

核心目标量化如下：

1. 通用能力对标：在公开基准测试（如OmniDocBench、dots.ocr-bench）上，模型的核心文档解析指标复现或达到原始dots.ocr模型的水平，证明我们具备构建SOTA级通用视觉编码器的能力。
2. 领域能力超越：在内部构建的JEDEC标准文档解析评测集上，模型在文本识别准确率、版面元素（图表、公式、表格）恢复完整性、关键信息提取精度等综合指标上，相较于直接使用原始dots.ocr模型，实现5%以上的绝对性能提升。

2. 技术路线：AIMV2架构的领域化演进

为实现上述目标，我们将放弃传统的“预训练模型微调”路径，转而采用更具前瞻性和潜力的 **“从零开始、领域注入”** 的预训练策略。技术路线基于AIMV2论文，具体演进如下：

阶段一：架构借鉴与初始化 (2025年12月)

- **核心理念**：采用AIMV2的**多模态自回归预训练**框架。该框架通过一个统一的解码器，自回归地预测图像块和文本标记，能从每一个图像块和文本词中获得密集的训练信号，已被证明能学到比对比学习（如CLIP）或纯描述学习更强的视觉表征。
- **我们的调整**：
  - **视觉编码器**：采用与AIMV2一致的ViT架构，但初始化参数采用在大量自然图像上预训练好的权重（如SigLIP），而非完全随机初始化，以加速收敛。

    - **训练目标**：沿用AIMV2的混合损失函数 `L_text + α * L_img`，其中图像部分使用L2像素回归损失。论文指出，这种联合目标在文本密集型VQA任务上显著优于纯描述或纯对比目标。
- **与dots.ocr的区别**：dots.ocr是一个**端到端的文档解析VLM**，其视觉编码器是固定的、黑盒的。我们则是**构建并训练一个专有的、透明的视觉编码器**，其能力可通过后续不同任务的连接器（如计划中的2层MLP）灵活释放。

阶段二：数据策略——领域深度渗透 (2026年1月 - 2026年3月)

这是实现领域超越的关键。我们将构建一个阶梯式的数据供应链：

1. **通用数据层（基座能力）**：使用AIMV2论文中提到的公开大规模图文对数据集（如DFN-2B, COYO），确保模型获得广泛的视觉概念和语言关联基础。
2. **领域数据层（核心能力）**：
    - **来源**：系统化收集并预处理公司内部的**JEDEC标准PDF、EFA/PFA报告（脱敏）、芯片设计文档、工艺手册**等。
    - **处理工具**：首先利用**dots.ocr模型**（作为高精度OCR工具）对这些领域文档进行初步解析，自动化生成“文档图像-结构化文本”的高质量配对数据。此步骤同时可评估dots.ocr在领域数据上的基线性能。
    - **数据增强**：针对芯片文档特点，模拟生成不同分辨率、噪点、压缩比的图像变体，以及部分遮挡的图表，提升模型鲁棒性。
3. **训练数据配比**：在预训练中，逐步提高领域数据的采样权重，让模型在通用知识的基础上，深度吸收芯片制造领域的视觉模式和专业术语。

阶段三：渐进式训练与分辨率适应 (2026年3月 - 2026年4月)

遵循AIMV2的成功经验，采用两阶段训练：

1. **基础分辨率预训练**：在224px或336px分辨率上，使用混合数据完成第一阶段预训练，让模型掌握基本语义。
2. **高分辨率与原生分辨率适应**：
    - **高分辨率微调**：对预训练模型进行448px等高分辨率微调，以捕获芯片图像和文档中更精细的细节（如微小的缺陷特征、电路纹理）。
    - **原生分辨率训练（关键创新）**：采用AIMV2论文中“**Native Resolution Fine-tuning**”策略，训练模型处理任意分辨率和长宽比的原始图像。这对于处理尺寸、格式不一的工业文档和SEM图像至关重要，能避免因统一缩放造成的信息损失。

3. 实施里程碑与关键产出

时间段	阶段	主要任务	关键产出与里程碑
2025年12月	筹备与基线建立	1. 环境搭建，复现dots.ocr推理流程。 2. **构建JEDEC评测集**：人工标注100-200份代表性文档，作为领域能力的黄金测试集。 3. 在评测集上运行原始dots.ocr，建立**性能基线**。	1. 可运行的dots.ocr环境。 2. **《JEDEC文档解析评测集V1.0》及基线性能报告**。
2026年1月	架构准备与数据工程	1. 基于Megatron-LM等框架，搭建AIMV2训练代码。 2. 启动领域数据自动化处理流水线，使用dots.ocr批量处理文档，生成训练对。	1. 核心训练代码库。 2. 首批高质量领域图文训练数据集（>10万对）。
2026年2-3月	核心预训练	1. 启动混合数据（通用+领域）的**第一阶段预训练**。 2. 在通用基准（OmniDocBench子集）和JEDEC评测集上进行**中期验证**。	1. **专用Visual Encoder v0.5模型**。 2. 中期评估报告，验证通用能力达标的可行性。
2026年4月	分辨率适应与精调	1. 进行**高分辨率及原生分辨率适应**训练。 2. 使用更精细的领域数据进行轻量级指令微调，强化版面分析与信息提取。	1. **专用Visual Encoder v1.0最终模型**。 2. 在JEDEC评测集上实现**相对dots.ocr基线提升>5%** 的验证报告。
2026年5月	集成与交付		1. **《专用Visual Encoder技术白皮书》** 与完整模型权重。

		1. 将训练好的Visual Encoder与DeepSeek-V3.2 LLM通过2层MLP连接，构建MLLM原型。 2. 在EMMI hotspot分析等场景进行端到端POC验证。	2. POC验证报告，展示其在MLLM pipeline中的价值。

4. 成功标准与风险评估

成功标准：

1. **定量**：JEDEC评测集综合得分提升≥5%；在OmniDocBench文本识别等核心指标上与dots.ocr相当。
2. **定性**：模型对芯片文档中的复杂图表、特殊符号、表格的逻辑结构理解明显优于通用模型。
3. **工程**：模型能稳定输出高质量的图像特征，支持后续MLLM训练；数据流水线可复用。

主要风险与应对：

- **风险1：领域数据质量与数量不足**。
  - **应对**：前期重点构建高质量、小规模的黄金评测集；数据流水线采用“dots.ocr初筛 + 关键样本人工校验”模式，保证数据质量优先于数量。
- **风险2：AIMV2架构训练成本高昂**。
  - **应对**：申请专用算力（如48-64卡H100集群）；采用混合精度训练、梯度检查点等技术优化；与论文不同的是，我们使用预初始化编码器，可缩短训练周期。
- **风险3：领域性能提升未达预期**。
  - **应对**：在中期验证后，及时调整数据混合比例、训练目标权重（α参数），或引入针对文档理解的辅助损失函数。

5. 资源需求

- **算力**：稳定且充足的GPU集群（**建议48-64卡H100 80G或等效算力**），用于为期2-3个月的大规模预训练。

- **数据**：**公司内部JEDEC、EFA/PFA等文档的脱敏使用权限**，以及构建评测集所需的领域专家工时支持。
- **协作**：与IT/数据平台部门合作，建立安全的数据处理环境；与缺陷分析、工艺部门专家合作，进行数据标注和效果评估。

本计划将使我们不仅获得一个强大的专用视觉编码器，更将建立起一套从领域数据生成到视觉基础模型训练的内部闭环能力，为长鑫存储在AI时代的核心竞争力构筑坚实壁垒。