IRIXEN-NanoVL-V1 技术白皮书

> 本文档以中文发布；模型权重、配置与推理代码的字段说明与本文一致。

摘要

IRIXEN-NanoVL-V1 是一个最低体量的视觉语言模型。它的语言主干 **98.82M 可训练参数**完全从零训练，视觉端复用一个**冻结的 SigLIP ViT-B/16**（93M），两者通过一个 1.57M 参数的两层 MLP 投影器连接。全模型总参数 **191.75M**，bf16 权重 **197 MB**，单张消费级显卡即可推理，实测吞吐 **100+ tok/s**。

在 400,000 条 ALLaVA-4V 图文样本上训练 12,000 步（192,000 样本-步）后，模型在三条测试样本上的 teacher-forcing 平均 top-1 为 **40.3%**、top-5 为 **63.5%**，相对 1/16000 的随机基线提升约 **6700 倍**。它能识别图像主题、色彩与粗略构图并生成语法通顺的中英文描述，但无法识别细粒度主体、图中的文字，也不擅长精确计数。

我们把结论写在最前面：**在这组约束下，40.3% 就是架构的天花板**，而不是训练没调好。这份白皮书同时给出三条被数据支持的解释，以及继续往上走需要付出的代价。

1. 设计约束与目标

项目的起点不是"做一个好用的视觉语言模型"，而是一个更窄的问题：**在单卡、可商用、可复现的约束下，一个视觉语言模型最小能做到多小，同时仍然"能看图、能说人话"？**

- **可训练参数 ≤ 100M**：训练与推理都要能落在消费级显卡上，不能用 7B 级基座。
- **语言部分从零训练**：不使用 LLaMA、Qwen 等现成基座，RoPE、RMSNorm、SwiGLU、tied embedding 全部自行实现。
- **全链路 Apache 2.0**：代码、权重、分词器、训练数据都能商用，分发内容中不得夹带许可不兼容的权重。
- **可复现**：训练时长、显存占用、数据规模、评测口径全部公开，结果可被第三方复核。

这四条约束共同决定了后面的每一个技术选择，包括为什么视觉端必须冻结、为什么输入分辨率只能停在 256×256。

2. 整体架构

一张 256×256 的 RGB 图像与一句问题文本进入模型后，经过三条链路：视觉编码、跨模态投影、语言解码。

```
图像 (256×256) → SigLIP ViT-B/16 [冻结] → 256 × 768 视觉 token
                                                  |
                                          投影器 (2 层 MLP + GELU)
                                                  |
问题文本 → tokenize → input_ids ──替换 <|image_pad|> 位置──→ 语言主干 (12L × 768H × 8 heads)
                                                  |
                                          逐 token 贪心解码 → 回答
```

2.1 视觉编码器（冻结，93M）

使用 SigLIP ViT-B/16 在 256×256 分辨率下的变体：`Conv2d(3, 768, kernel=16, stride=16)` 做 patch embedding，16 层 Transformer Encoder，输出 `[B, 256, 768]`，即 256 个视觉 token。整个编码器 `requires_grad=False`，训练全程不更新。

冻结的理由是数据量：从零训练一个 ViT 需要上亿级图文对，远超"最低体量"的目标。复用 SigLIP 是在单卡条件下唯一能跑通"从零训语言"的现实路径，代价是视觉端的表征能力被锁定在 SigLIP 的预训练分布里。

2.2 跨模态投影器（可训练，1.57M）

两层 MLP：`Linear(768 → 1024) → GELU → Linear(1024 → 768)`。它的作用是把视觉特征"翻译"到语言主干的嵌入空间，是唯一显式对齐两个模态的模块。1.57M 的容量上限，也是后文瓶颈分析中的第一条。

2.3 语言主干（可训练，97.25M）

12 层、隐藏维度 768、8 个注意力头的解码器：

- 每层结构：`RMSNorm → Attention(SDPA, RoPE) → +残差 → RMSNorm → SwiGLU(768→2048→768) → +残差`。
- 位置编码：RoPE，θ=10000，最大长度 1536。
- 词表：BPE 16,000；`embed_tokens` 与 `lm_head` 共享权重（tied embedding），省下一份 16000×768 = 12.29M 参数。
- 输出 `[B, T, 16000]` logits，逐 token 贪心解码，叠加 repetition penalty 与 no_repeat_ngram 抑制退化。

2.4 参数量构成

组件	参数量	训练状态
SigLIP ViT-B/16 视觉编码器	93M	冻结
2 层 MLP 投影器	1.57M	可训练
12 层语言主干（含 tied embedding）	97.25M	可训练
合计	**191.75M**	**98.82M** 可训练

3. 训练数据与流程

3.1 数据

训练集取自 **ALLaVA-4V**（Apache 2.0）子集：从 290 万条中按 `seed=42` 采样 400,000 条，中英比例约 50:50。图像统一缩放至 256×256，以 JPEG 字节形式存放在 parquet 中；对话格式为 `<image>` + 问题 → 描述 / 回答。

3.2 训练配置

项	值
优化器	AdamW（β_1=0.9，β_2=0.95，weight decay=0.1）
学习率	投影器 1e-3 / 语言主干 3e-4，cosine 衰减至 10%
有效批大小	4（batch）× 4（梯度累积）= 16
训练步数	12,000 步 = 192,000 样本-步
精度与平台	bf16，AMD ROCm 7.2（gfx1100）验证通过
硬件与耗时	单卡 51.5 GB 显存，约 36 分钟

3.3 序列构造与标签对齐

样本按因果语言模型的标准格式拼接，图像占位符 `<|image_pad|>` 连续出现 256 次，其位置的 embedding 被投影器输出替换：

```
user 段: [<|im_start|>, \n] + 问题 token + [<|image_pad|>] × 256 + [<|im_end|>, \n]
asst 段: [<|im_start|>, \n] + 回答 token + [<|im_end|>]

labels[t] = input_ids[t+1]   # 仅 asst 段有效，其余置 -100
```

这里有一个容易被忽略的实现细节：**labels 必须相对 input_ids 右移一位**。早期版本未做右移，模型学到的是"复制输入"，训练 loss 会假性降到 0.3 左右，而实际生成能力为零。这是本项目最值得记录的一次失败。

4. 评测

4.1 量化指标

评测使用 teacher forcing，统计每个位置预测正确率，测试集为 3 条人工挑选的样本，覆盖长描述、负空间提问与物体描述三类问题。

样本	语言	TOP-1	TOP-5
0（长描述 + 负空间问题）	英文	51.5%	72.4%
1（长描述 + 负空间问题）	中文	28.0%	50.5%
2（物体描述）	英文	41.4%	67.7%
平均	—	**40.3%**	**63.5%**

对照基线为随机猜测：词表 16,000，随机命中率约 **0.006%**。模型在其上提升约 6700 倍——这个倍数主要说明"确实学到了东西"，而不代表绝对能力强。

4.2 生成样例

样本 **0**，问：负空间与色彩对比如何影响这张图？

"The use of a bright, vibrant color contrast in the image creates an atmosphere that is both dynamic and intense. The dark background provides a stark contrast to the darker tones of the scene..."
语法通顺、主题相关（对比、背景、构图），但没有抓住"负空间"这个具体概念。

样本 **1**，同一问题用中文提问：

"图片上的光线和色彩对比，以及人物的明亮、柔和的色调，与背景形成鲜明对比。这种颜色选择通过营造一种动态而生动的氛围..."
中文语法基本正确，但仍未识别出画面主体是"身着盔甲的人物"。

5. 能力边界

5.1 能胜任

任务	说明
边缘设备 VQA	消费级显卡 100+ tok/s，CPU 亦可运行
图像主题识别	"图里有什么""什么颜色""几个人"这类粗粒度问题
教育演示与快速原型	197 MB 权重，加载约 60 秒
更大 VLM 的对照基线	作为"最小可用版本"参与消融与对比实验
Apache 2.0 商用	代码、语言主干、分词器全部可商用

5.2 不擅长

任务	原因
OCR / 读图中文字	256×256 输入下小于 16×16 像素的文字不可识别，模型会编造不存在的词（如 "DEATS"、"PREADA BOOD"）
细粒度主体识别	把"身着盔甲的人物"识别成"鸟/抽象动物"；64 或 256 个视觉 token 都不足以描清人物轮廓
中文流畅度	中文 top-1 28.0% 对比英文 51.5%；BPE 对中文碎片化更严重，中英混合数据也让语言意图更模糊
高精度计数	"图里有几个苹果"这类问题准确率极低
多轮对话记忆	单轮 VQA 设计，每轮独立推理，无历史上下文

6. 版本演进与瓶颈分析

指标	V1.0	V1.1	V1.2（最终）
可训练参数	63.13M	98.82M	**98.82M**
语言层数	8L	12L	**12L**
分词器	6,400	16,000	**16,000**
视觉端	B/32	B/32	**B/16**
视觉 token	64	64	**256**
训练步数	3,000	15,000	**12,000**
样本-步	48K	240K	**192K**
英文 top-1	41.8%	52.2%	**51.5%**
中文 top-1	25.4%	25.8%	**28.0%**
平均 top-1	34.4%	39.1%	**40.3%**

最值得注意的一行是视觉端：从 B/32（64 token）换到 B/16（256 token），**视觉 token 增加 4 倍，平均 top-1 只提升 1.2 个百分点**。这不是训练没跑好，而是三个瓶颈叠加的结果：

投影器容量太小：1.57M 参数的两层 MLP 要在 768→1024→768 之间完成 256 个视觉 token 的语义压缩，表达能力明显不足。

训练量反而下降：V1.1 跑了 240K 样本-步，V1.2 因序列变长只跑了 192K 样本-步，增益被抵消。

分辨率上限：ALLaVA 图像已被缩放到 256×256，B/16 每个 patch 覆盖 16×16 像素，已接近原图能提供的细节极限。

因此我们给出明确判断：**40.3% 平均 top-1 是"100M 可训练参数 + 256×256 输入"这组约束下的架构天花板**。继续提升只有三条路，且每条都要放弃某项约束——

视觉端换 SO400M：违反最低体量约束，权重从 197 MB 增至 GB 级。

语言主干扩到 1.5B+：失去边缘部署价值。

投影器扩到 50M+ 或加深层数：收益边际递减，且训练数据量不足以支撑。

7. 发布包与推理

发布包 `IRIXEN-NanoVL-V1` 共 7 个文件、198.9 MB，**不包含 SigLIP 权重**：

文件	大小	说明
`model.safetensors`	197.6 MB	98.82M 参数，bf16，tied embedding 只存一份
`config.json`	2.4 KB	结构参数与 vision_repo 元信息
`tokenizer.json`	1.21 MB	16,000 BPE，特殊 token ID 固定
`inference.py`	6.5 KB	自包含推理脚本（内嵌完整模型定义）
`README.md`	3.2 KB	模型卡：指标、局限与用法
`LICENSE`	0.5 KB	Apache 2.0
`requirements.txt`	107 B	torch / transformers / tokenizers 等依赖

`inference.py` 的设计目标是"拿到就能跑"：不依赖训练代码，用户不需要了解 RoPE 或 SwiGLU 的实现细节。

```
from inference import IRIXENNanoVLInference

m = IRIXENNanoVLInference("./IRIXEN-NanoVL-V1", "./siglip-base-p16-256")
print(m.generate(image, "Describe this image."))
```

命令行方式：`python inference.py <release_dir> <vision_dir> test.jpg`。

8. 许可与合规

层面	许可	说明
代码	Apache 2.0	模型结构、训练代码、推理代码
语言主干权重	Apache 2.0	97.25M，从零训练
投影器权重	Apache 2.0	1.57M
分词器	Apache 2.0	基于训练数据训练的 BPE
训练数据	Apache 2.0	ALLaVA-4V 本身为 Apache 2.0
视觉编码器	单独获取	SigLIP 由 Google 发布，用户需自行下载并遵守其原始许可；**发布包不含 SigLIP 权重**

这是刻意的合规设计：分发内容只有语言部分（197 MB），SigLIP 由用户在运行时自行获取。这让 IRIXEN-NanoVL-V1 的所有分发文件都处于纯粹的 Apache 2.0 之下。

9. 发布状态

产物	地址	状态
模型仓库	`modelscope.cn/models/lanche816/IRIXEN-NanoVL-V1`	已上传，197.65 MB
创空间	`modelscope.cn/studios/lanche816/IRIXEN-NanoVL-V1`	Streamlit 交互界面，可直接上传图片对话
权重	存于模型仓库	创空间通过 `snapshot_download` 拉取，无需重复推送

创空间启动流程：安装 6 个依赖（1–2 分钟）→ `load_model()` 触发 `snapshot_download` 拉取 197 MB 权重（30–60 秒）→ 显示"模型加载完成"后即可上传图片对话。

10. 结论

做到了什么：在 100M 可训练参数、单卡、40 万样本的约束下，从零训练出一个能看图、能说人话的视觉语言模型。英文 top-1 51.5% 足以识别主体与构图，中文 28.0% 也能生成语法通顺的描述；Apache 2.0 可商用，6.5 KB 的 `inference.py` 即可运行，教育与边缘场景足够。

没做到什么：它不是 LLaVA-7B 这一级别的替代品。把盔甲人物看成鸟、编造图中不存在的文字、中文提问用英文回答——这些不是训练 bug，而是 100M 参数、256×256 输入与 64/256 视觉 token 叠加后的物理上限。

定位：IRIXEN-NanoVL-V1 是"100M 级可训练参数下的视觉语言模型天花板"，不是一个优秀的 VLM，但在它的约束里，是一份诚实、可复现、可商用的交付。

附录：复现信息

训练数据：ALLaVA-4V（Apache 2.0），40 万条子集，采样种子 `seed=42`。

训练环境：AMD ROCm 7.2 / gfx1100，bf16，单卡 51.5 GB 显存。

评测口径：teacher forcing 逐位置 top-1 / top-5，3 条测试样本，随机基线 1/16000。

版本：V1.2（final），对应权重 `model.safetensors` 197.6 MB。

IRIXEN · 2026