IRIXEN-Embedding-V2 技术白皮书

> 本文档有中文与英文两个版本，可用页头的语言开关切换；可下载的 PDF 为中文版。文档编号 IRIXEN-WP-2026-001，版本 V1.0，编制单位 IRIXEN Team。

摘要

IRIXEN-Embedding-V2 是一个参数高效的多任务文本嵌入模型。它在一个**冻结的 Qwen3-Embedding-4B**（4.02B 参数）之上，只训练 **68.8M** 可训练参数——占全模型 **1.68%**：一部分是覆盖 q/k/v/o 四个投影的低秩适配（LoRA，r=64、α=128），另一部分是一组把表示映射到不同几何空间的**乘积流形几何头**：球面 S^{1024}、双曲头 L^1（H^{512}，c=0.5）、双曲头 L^2（H^{512}，c=2.0）与欧氏头 R^{1024}。

在六项基准上模型 5 胜 1 负：MIRACL 检索 **95.10%**（较基座 +3.65）、四种未参与训练的语种上跨语言检索 **93.90%**（+5.35）、T2 中文重排 **82.87%**（+11.07）、iFLYTEK 119 类分类 **64.05%**（+15.70）、CLS 零样本聚类 **76.94%**（+1.86）；未参与训练的 STS 任务为 86.61%，低于基座 7.09 个百分点。

结论前置：**增益来自表示空间的几何结构，而不是参数堆叠**。全部训练只用了 1,000 步基座训练加 800 步分类头精训，单张 MI300X 约 90 分钟，bf16 推理显存约 10 GB。STS 的缺口我们不作粉饰，并在第 6.1 节给出可直接落地的路由方案。

1. 概述

1.1 模型定位

IRIXEN-Embedding-V2 是一个参数高效的多任务文本嵌入模型。模型以冻结的 Qwen3-Embedding-4B 为基座，通过低秩适配（LoRA）与乘积流形几何头（Product Manifold Geometry Heads）的组合，实现单模型对检索、重排、分类、聚类、相似度等多类语义任务的统一表示学习。

模型全量参数规模 4.02B（基座，冻结），新增可训练参数 68.8M，仅占全模型 1.68%；bf16 推理显存需求约 10 GB；在单卡 AMD MI300X 上完成全部训练约需 90 分钟。

1.2 核心命题

不同的语义任务需要不同的几何空间。将检索、分类、聚类、相似度等任务强行映射至单一欧氏空间，是当前文本嵌入模型的隐性瓶颈。IRIXEN-Embedding-V2 通过乘积流形架构，使每一类任务落在与其内在结构最匹配的几何空间上，从**模型结构层面**而非参数调优层面提升表示质量。

1.3 关键指标摘要

表 **1-1** · 关键指标摘要

指标项	数值	备注
基座模型	Qwen3-Embedding-4B（4.02B）	训练与推理阶段均冻结
可训练参数	68.8M	占全模型 1.68%
推理显存需求	约 10 GB	bf16 精度
训练硬件	AMD MI300X 192GB × 1	单卡完成全部训练
总训练时长	约 90 分钟	含分类头二次精训
MIRACL 检索（10 语言）	95.10%	较基座 +3.65 个百分点
MIRACL 跨语言（4 未训语言）	93.90%	较基座 +5.35 个百分点
iFLYTEK 分类（119 类）	64.05%	较基座 +15.70 个百分点

1.4 术语与缩略语

表 **1-2** · 术语与缩略语

术语	含义
文本嵌入（Embedding）	将文本映射为固定维度稠密向量，用于检索、分类等下游计算
LoRA	低秩适配（Low-Rank Adaptation），仅训练少量增量参数的参数高效微调方法
乘积流形	由多个不同几何性质的空间（球面、双曲、欧氏）直积构成的表示空间
双曲空间 H	具有负曲率的几何空间，体积随半径指数增长，天然适配树状层次结构
Lorentz 模型	双曲空间的一种实现形式，以曲率参数 c 控制空间弯曲程度
多头池化	同时采用注意力池化、末位词元池化、均值池化并经可学习门控融合的表示聚合方式
重排（Rerank）	对初步召回结果按相关性重新排序的任务
MIRACL	多语言检索基准数据集
STS	语义文本相似度（Semantic Textual Similarity）任务，以 Spearman 相关系数衡量
bf16 / fp32	bfloat16 与 float32 数值精度
ROCm / SDPA	AMD GPU 计算平台 / 缩放点积注意力实现

2. 总体技术架构

2.1 架构分层

IRIXEN-Embedding-V2 自下而上由五层构成：

冻结基座层：Qwen3-Embedding-4B，4.02B 参数，32 层 Transformer，隐藏维度 2560；

LoRA 适配层：秩 r=64、缩放系数 α=128，作用于 q/k/v/o 四个投影矩阵，可训练参数 47.19M；

多头池化融合层：注意力池化、末位词元池化、均值池化三路并行，经 Softmax 可学习门控加权融合；

乘积流形几何头层：球面头 S^{1024}、双曲头 L^1（H^{512}，c=0.5）、双曲头 L^2（H^{512}，c=2.0）、欧氏头 R^{1024}；

任务头层：分类器 RegClassifier，输入为双曲头 L^1 与 L^2 特征拼接的 1026 维向量，输出 119 类。

2.2 数据流示意

模型端到端的数据流如下图所示。基座与 LoRA 只负责把 token 序列编码为上下文表示，任务差异全部由池化融合与几何头承担。

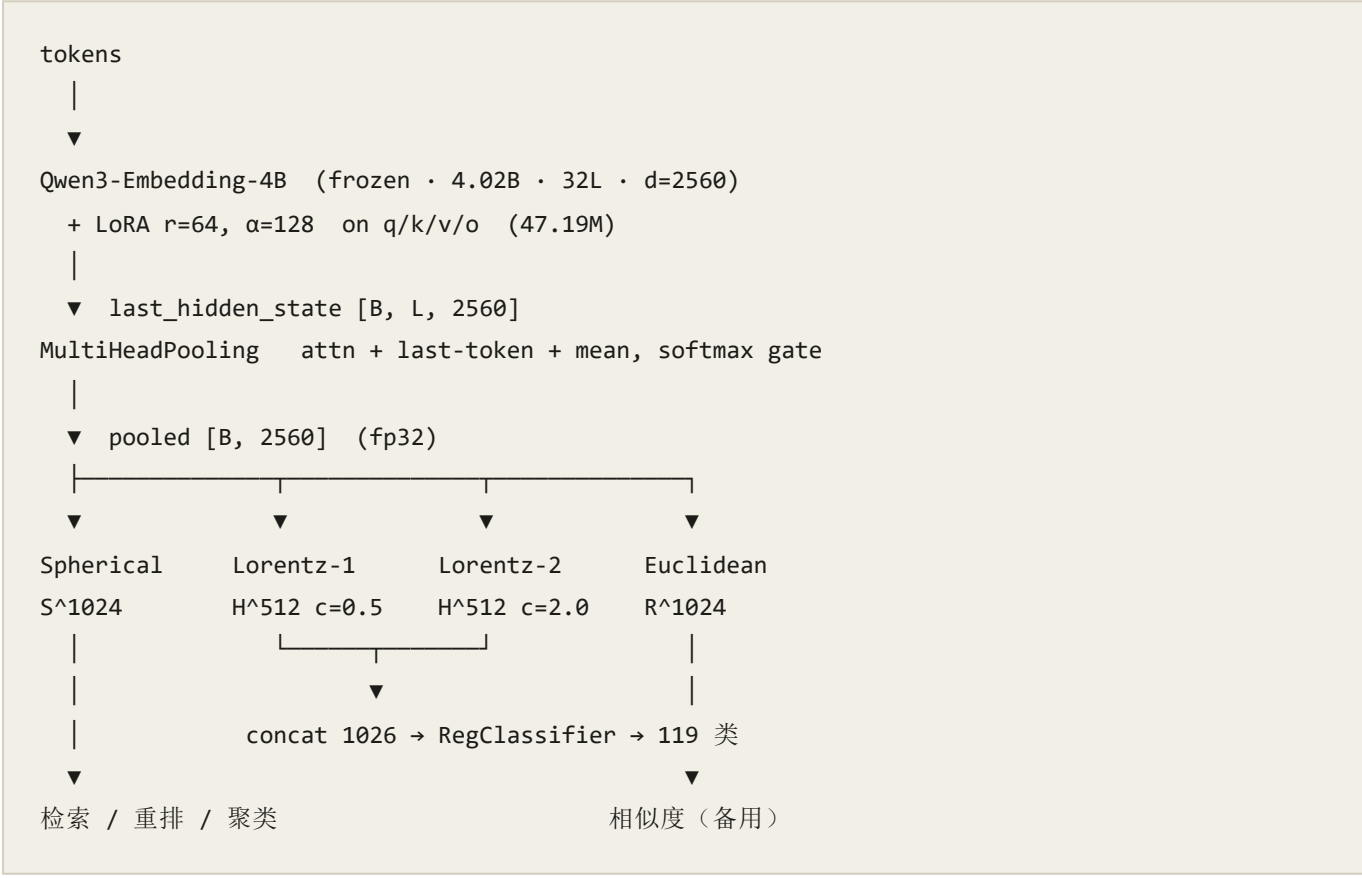


图 2-1 · 端到端数据流：共享编码 + 分任务几何头

2.3 模型参数构成

表 2-1 · 模型参数构成

组件	参数量	说明
Qwen3-Embedding-4B	4,021,762,816	冻结，不参与训练
LoRA（r=64, α=128）	47,185,920	作用于 q/k/v/o 投影
MultiHeadPooling	2,563	注意力查询向量与融合门控
SphericalHead	7,342,080	2560→2048→1024，两层 MLP
LorentzHead-1	3,146,240	2560→1024→512，曲率 c=0.5
LorentzHead-2	3,146,240	2560→1024→512，曲率 c=2.0
EuclideanHead	7,342,080	2560→2048→1024，两层 MLP
RegClassifier	589,000	LayerNorm + Dropout(0.3) + 1026→512→119
可训练参数合计	**68.8M**	占全模型 **1.68%**

3. 核心技术

3.1 乘积流形架构

传统嵌入模型将语义映射至单一欧氏空间 R^d。IRIXEN-Embedding-V2 采用乘积流形作为统一表示空间：

$$M = S^{1024} \times H^{512}(c=0.5) \times H^{512}(c=2.0) \times R^{1024}$$

其中球面分量承担检索与重排，两个不同曲率的双曲分量分别承担大类与细类层次结构建模，欧氏分量承担连续相似度计算。各流形分量的几何属性与适配任务如下：

表 **3-1** · 乘积流形各分量几何属性与任务适配

流形分量	维度	几何性质	适配任务	数学原理
Spherical	S^1024	单位球面	检索 / 重排	方向判别天然匹配余弦距离
Lorentz-1	H^512, c=0.5	双曲面	大类层次	低曲率近似平坦空间，适合粗分
Lorentz-2	H^512, c=2.0	双曲面	细类层次	高曲率强弯曲，适合深层细分
Euclidean	R^1024	平坦空间	连续相似度	无曲率偏差

关于双曲率并行的必要性：双曲空间的体积随半径指数增长，天然承载树状层次结构。浅层大类（如「科技 / 财经」）适合低曲率空间，深层细分类别（如「AI 芯片 / 新能源车」）适合高曲率空间。以两个不同曲率的双曲头并行建模层次结构，是 IRIXEN-Embedding-V2 的结构性创新。

3.2 多头池化融合

模型不依赖单一的末位词元池化或均值池化，而是采用三路池化并经可学习门控自适应融合：

```
pooled = g₁ · AttentionPool(h) + g₂ · LastToken(h) + g₃ · MeanPool(h)
其中 g = softmax(learnable_gate)，随训练自适应更新
```

该设计解决了不同任务对池化策略偏好不一致的问题：检索任务偏好末位词元池化，分类任务偏好注意力池化，STS 类任务偏好均值池化。门控机制使同一模型可按任务自动收敛至合适的池化组合。

3.3 分类头两阶段精训

针对分类任务（iFLYTEK 119 类），模型采用两阶段训练策略：

阶段一：联合训练基座 LoRA、双曲几何头与分类头；

阶段二：冻结基座 LoRA 与双曲头，仅重新训练分类头，并引入 LayerNorm、Dropout(0.3) 与 Label Smoothing(0.1) 正则化配置。

阶段二的作用显著：过拟合差距（train-val gap）由 +7.9% 收敛至 +4.6%，测试集准确率提升 2.55 个百分点。

4. 训练方案

4.1 训练数据

表 4-1 · 训练与验证数据构成

数据集	规模	语言覆盖	用途
MIRACL 训练集	28,630 三元组	zh / en / es / fr / de / ru / ja / ko / ar / it（10 语言）	检索任务训练
MIRACL 验证集	11,452 三元组	hi / th / vi / tr（4 种完全未训语言）	跨语言泛化验证
T2 重排数据	17,047 三元组	中文	重排任务训练
iFLYTEK	24,649 样本	中文，119 类	分类任务训练

4.2 损失函数

检索任务采用带边距的三元组损失（margin triplet loss）：

```
L = relu(0.05 + ⟨q, n⟩ − ⟨q, p⟩)
```

其中 q 为查询向量，p 为正例向量，n 为负例向量。分类任务采用带标签平滑的交叉熵损失：

```
L = CrossEntropy(Classifier(LorentzFeature), y)
label_smoothing = 0.1
```

4.3 训练配置

表 4-2 · 训练配置

配置项	取值
基座状态	Qwen3-Embedding-4B（4.02B 参数），全程冻结
LoRA 配置	r=64, α=128，覆盖 q/k/v/o 四个投影
池化方式	MultiHeadPooling（attention + last-token + mean，可学习门控）
基座训练步数	1,000 步
分类头精训步数	800 步
优化器	AdamW
学习率调度	1.5e-4 → 8e-5 → 5e-5（三段式）
任务采样比例	检索 : 分类 = 0.5 : 0.5
数值精度	基座 bf16，几何头与分类头 fp32
梯度裁剪	1.0
训练硬件	AMD MI300X 192GB，单卡
总训练时长	约 90 分钟

5. 性能评估

5.1 评估口径

以下对比实验在同一份干净数据、同一 tokenizer、同一采样方法条件下进行。基线模型 Qwen3-Embedding-4B 采用官方最优指令，IRIXEN-Embedding-V2 采用对应流形输出，确保对比结果只反映模型结构差异，而不是提示词或数据差异。

5.2 对比结果

表 5-1 · 与基线模型性能对比（5 胜 1 负）

任务	IRIXEN-EMBEDDING-V2	QWEN3-EMB-4B（基线）	增益
MIRACL 检索（10 语言）	95.10%	91.45%	+3.65
MIRACL 跨语言（4 未训语言）	93.90%	88.55%	+5.35
T2 中文重排	82.87%	71.80%	+11.07
iFLYTEK 分类（119 类）	64.05%	48.35%	+15.70
CLS 零样本聚类	76.94%	75.08%	+1.86
STS Spearman（未训练）	86.61%	93.70%	−7.09

5.3 结论分析

分类任务领先 15.70 个百分点。双曲几何对层次结构的建模优势源于空间体积随半径指数增长的数学性质，属结构性优势而非调参结果。

跨语言泛化达 93.90%。训练仅使用 10 种语言，验证使用 4 种训练中完全未出现的语言（印地语、泰语、越南语、土耳其语），表明模型学到的方向空间是语言无关的语义特征表示。

STS 任务未训练，本报告如实披露。该任务上模型低于基线 7.09 个百分点；生产环境建议将该类请求路由至基座 Qwen3-Embedding-4B 并附加 STS 指令（可达 93.70%）。

6. 部署与使用建议

6.1 任务路由建议

基于第 5 章的评估结果，生产环境建议采用如下多模型路由策略。同一份权重即可覆盖五类请求中的四类，只有 STS 需要回退到基座指令模型。

表 6-1 · 任务路由建议

任务类型	推荐模型与输出头	参考指标
检索	IRIXEN-Embedding-V2（Spherical 头）	95.10%（跨语言 93.90%）
重排	IRIXEN-Embedding-V2（Spherical 头）	82.87%
分类	IRIXEN-Embedding-V2（分类器）	64.05%
聚类	IRIXEN-Embedding-V2（Lorentz 头）	76.94%
相似度（STS）	Qwen3-Embedding-4B 基座 + STS 指令	93.70%

6.2 运行环境要求

表 6-2 · 运行环境要求

项目	要求
推理显存	约 10 GB（bf16 精度）
推理批大小	建议 32
首次加载时间	约 30 秒
注意力实现	推荐 sdpa（ROCm 环境兼容性最佳）

6.3 推理接口

模型提供标准 Python 推理接口。检索与分类共用同一份权重，只在输出头上分流：

```
from inference import IRIXEN2

m = IRIXEN2(base_path="...", model_dir="./IRIXEN-Embedding-V2")

# 检索：编码查询与文档，向量点积即得相关性得分
eq = m.encode(["查询文本"], task="retrieval")
ed = m.encode(["文档文本"], task="retrieval")
sim = eq @ ed.T

# 分类：直接输出类别标签
labels = m.classify(["应用描述文本"])
```

图 6-1 · 检索与分类的调用方式

6.4 兼容性注意事项

在 AMD ROCm 环境下，应使用 `attn_implementation="sdpa"`；FlashAttention 2 在 ROCm 上存在推理卡死风险（已在开发过程中实测确认）。

`flash-attn` 2.8.3 为可选依赖，非必需组件。

7. 局限性说明

为便于使用方做出合理选型判断，现将模型已知局限性如实披露如下：

STS 未训练。模型不包含 STS 任务训练，该任务性能低于基线；生产环境建议按第 6.1 节路由至基座模型处理。

分类头语言范围限定。iFLYTEK 分类器仅针对中文主题分类，跨语种分类场景需重新训练分类头。

训练尚未饱和。当前为基座 1,000 步加分类头 800 步的训练规模，仍存提升空间；按现有趋势继续训练，iFLYTEK 分类准确率预计可达 65% 以上。

ROCm 兼容性约束。注意力实现需按第 6.4 节选择，FlashAttention 2 在 ROCm 上不可用。

8. 技术栈与依赖

表 **8-1** · 技术栈与依赖

组件	版本 / 说明
PyTorch	2.x + ROCm 7.2
transformers	5.15
peft	0.20
flash-attn	2.8.3（可选，推荐 sdpa 替代）
训练硬件	AMD MI300X 192GB

9. 引用方式

引用本模型时请使用如下格式：

```
IRIXEN-Embedding-V2 (2026). Multi-Head Pooling + Dual-Curvature Product Manifold
Embedding. IRIXEN Team.
```

附录 · 文档信息

表 **A-1** · 文档信息

项目	内容
文档编号	IRIXEN-WP-2026-001
版本号	V1.0
编制单位	IRIXEN Team
编制日期	2026 年 10 月 11 日
适用范围	模型评估、选型与生产部署参考

本文所有指标均可在同一份干净数据、同一 tokenizer 与同一采样方法下复现；如需核对具体评测脚本与采样清单，可通过站内工单联系。