1. 为什么要制作数据集

(针对上周师兄提出的疑问，补充立项依据)

目前通用的多模态模型（如 Qwen-VL 原生权重）虽然在“通用描述”上表现优秀，但在处理我们特定的**异常检测/细粒度动作分析**任务时存在明显的短板。我制作 `xd_dataset` 的核心目的有三点：

1. **对齐任务范式**：
    - 原生模型的训练数据多为静态图文对（Captioning），缺乏对“时间维度”的敏感性。
    - 本数据集将连续视频帧与时间戳事件（Annotations）结合，旨在**使模型建立【视觉特征 -> 时间定位】的映射能力**，而不仅仅是识别画面里有什么。
2. **增强领域适应性 (Domain Adaptation)**：
    - 针对特定场景（如监控视角、特定动作交互），通用数据集缺乏此类长尾分布的数据。
    - 通过此数据集微调，可以纠正模型在特定领域的“注意力漂移”问题（即防止模型只关注背景而忽略关键动作）。
3. **为“可控生成”做数据铺垫**：
    - 我们后续的研究计划涉及干预模型的注意力（Attention Steering）。高质量的标注数据是训练“Attention Loss”或评估 Attention 准确性的基石。

2. Qwen3-VL Attention

2.1 实验设置

- **模型**：Qwen3-VL-2B-Instruct
- **输入**：视频帧/图片 + Prompt "Describe the image." / "A person high-fiving a dog"
- **方法**：提取 Prompt 最后一个 Token 对视觉 Token 的 Attention Map，进行 Layer-wise 和 Head-wise 的可视化。

2.2 关键发现 I：注意力头的“专业化分工”

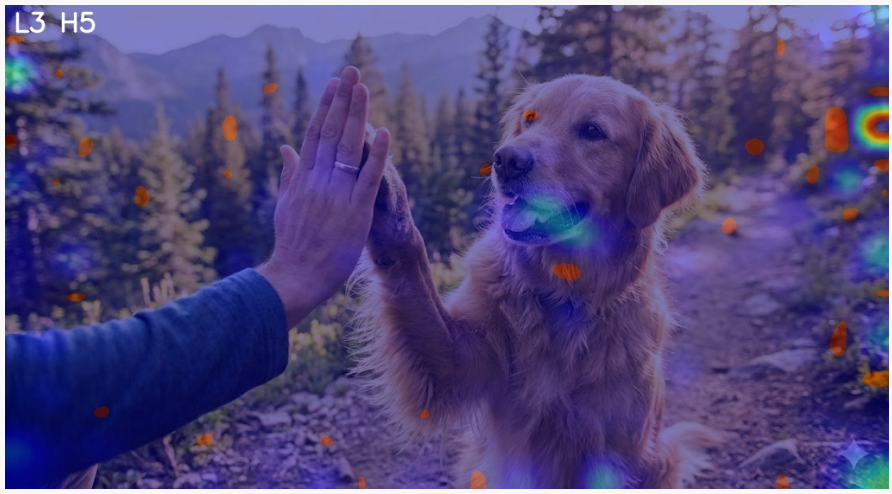
我发现并非所有 Layer 都在关注图像内容 无法识别哪些“注意力头“更加重要。

(对比图：一张杂乱的 Layer 0 图片 vs 一张清晰聚焦物体的 Layer 12/20 图片)

(图注：Layer 20 的注意力明显更集中于语义主体，而 Layer 0 则较为发散)

2.3 关键发现 II：输入依赖性 (Input Dependence)

这是一个非常有趣的现象。我们观察了同一个 Attention Head (`L3 H5`) 在不同输入下的表现：

输入内容	表现	示例图
Case A: 远景/干净背景	Head 精精准定位了手与爪子的接触点	
Case B: 近景/复杂背景	Head 被背景纹理干扰，注意力弥散，丢失焦点	

结论：同一个 Attention Head 的功能是固定的（探测特征），但其**激活状态高度依赖于输入分布。**这意味着我们不能简单地寻找“万能头”，而需要寻找“鲁棒头”。

---

3. 遇到的挑战与思考

现象：
在测试中，我发现通过 Attention Map 并没有观察到模型对“手掌”部分有极高的关注度（热力值偏低），但模型生成的文本却准确描述了“High-fiving”。

原因分析：

1. **先验知识 (Prior Knowledge)**：LLM 部分极其强大，依靠 Prompt 里的文本共现概率（"Person" + "Dog" -> "High-five"）进行了“脑补”，而非完全依赖视觉。
2. **注意力稀疏性**：模型可能只需要关注几个关键像素点（Sparse Attention）就能完成识别，归一化后的热力图容易掩盖这些微弱但关键的信号。

---

4. 下一步计划

基于上述发现，下周计划：

1. **探索更鲁棒的注意力筛选机制**：
    - 目前的分析主要基于定性观察和简单的能量占比。计划尝试探索更多维度的量化指标，在更多样化的数据集上验证注意力头的稳定性，尝试找出更通用的规律。
2. **进一步验证注意力与生成的关联**：
    - 尝试设计实验来验证“高关注度”是否真的不仅带来“可解释性”，还能带来“更好的生成结果”。探究是否可以通过调整注意力机制来优化模型在特定场景下的表现。
3. **架构融合探索 (Architecture Integration - V-JEPA)**：
    - **引入背景 (What is V-JEPA?)**：参考 Meta FAIR 提出的 **V-JEPA (Video Joint Embedding Predictive Architecture)**。它的核心理念是 不重建像素，而是预测潜变量特征 (Latent Representations)。
    - **融合方向**：探索利用 V-JEPA 提取的鲁棒时空特征作为“视觉锚点”，用来矫正 Qwen3-VL 的注意力分布，或者作为辅助监督信号，解决生成式模型过度依赖文本先验而忽视视觉物理逻辑的问题。