课堂多模态视频“异常/违规”行为分析研究路线（面向45分钟整课）

> 目标：在45分钟课堂视频中，自动发现并定位学生与教师的异常/违规行为（如：学生睡觉、玩手机；教师走神、粗鄙言论、体罚等），并给出可解释证据（时间戳、画面、语音文本与理由）。

---

1. 任务拆解与定义

- 时空定位（Where & When）
  - *短时原子行为*：如“拿出/使用手机”“打瞌睡”“拍打学生”等。
  - *复合行为/情境*：如“讲课过程中持续走神”“与学生发生肢体冲突”。
  - 输出：`[start, end]` 片段 + 关键帧/关键clip + 主体ID（教师/学生）。
- 语义判定（What & Why）
  - 基于视觉（人/物体/骨架/动作）+ 音频（ASR 文本、情绪/事件）+ 上下文（课程阶段、先后因果）综合推理，判断是否违规并给出证据链。
- 合规与可解释性
  - 所有结论需附证据包（关键帧/短clip、ASR片段、时间戳、规则条款/提示词）。
  - 记录模型置信度与不确定性，支持人工复核。

---

2. 数据与伦理合规

- 采集：获得师生及监护人书面同意；划定拍摄区域，避免黑板外、教室外隐私区域。
- 脱敏：默认对人脸做可逆/不可逆脱敏（马赛克/遮罩）；语音脱敏关键词（姓名、电话等）。
- 留存策略：区分“原始视频”“特征/摘要”“审计日志”，尽量只长期保留特征与事件摘要。
- 偏见与误报：定期抽检不同年级、学科、班型；误报触发人工复核而非自动处罚。

---

3. 总体系统架构

1. 离线预处理（GPU-0）
    - 解码 → 统一帧率（如 25fps）→ 语音分离/降噪 → 语音转写（ASR）→ 说话人分离（教师/学生）。
2. 基础感知（GPU-1）
    - 检测与跟踪：人/手机/教鞭/桌椅等（目标检测 + 多目标跟踪）。
    - 姿态/骨架：人体关键点/手势。
    - 短时动作识别：睡觉（低头+闭眼+静止）、玩手机（手-手机-目光关联）、推搡/击打等。
3. 时序定位与候选生成（GPU-2）
    - TAD/STAD（时间动作检测/时空动作检测）生成候选片段；
    - TSG/TVG（文本-视频时序定位）基于规则模板/自然语言检索“可能违规”的片段。
4. 多模态大模型推理（GPU-3）
    - 采用长视频友好的 MLLM（带记忆/分层/流式机制），对候选片段进行二次核验与解释；
    - 融合 ASR 文本（粗鄙/辱骂/体罚指令等）与视觉证据，输出判定+理由+证据。
5. 审计与人机协同
    - 事件工单（时间轴预览、关键证据对齐）；一键标注“误报/真阳性”，闭环持续学习。

---

4. 关键模块与可选模型

> 下列皆为开源/可本地部署优先，适配 4×A100（80G）或 40G。

4.1 音频/语音

- ASR：Whisper、Paraformer（普通话、低延迟），必要时流式改造；
- 说话人分离：pyannote.audio（教师/学生/多人争执）；
- 音频事件/情绪：SenseVoice（情绪、笑声、噪声、嘈杂环境等）。

4.2 视觉基础

- 目标检测：YOLOv8/10（人、手机、棍棒等小目标）；
- 多目标跟踪：ByteTrack（稳定ID，便于行为归属）；
- 人体姿态：MMPose（全身/手部关键点，支撑“体罚/推搡/拍打”等动作判断）。

4.3 行为与时序定位

- 短时动作识别（HAR）：VideoMAE / InternVideo2（clip级识别）
- 时间动作检测（TAD）/时空动作检测（STAD）：ActionFormer/DETR式查询/边界回归等实现。
- 文本引导时序定位（TSG/TVG）：针对“是否出现粗鄙言论”“是否发生肢体接触”等自然语言查询。

4.4 多模态大模型（长视频友好）

- 记忆增强：MovieChat、MA-LMM（稀疏记忆/外部记忆）；
- 流式/层级：TimeSearch、ReVisionLLM、VideoTree；
- 高效Token策略：Video Token Merging、PruneVid/LLM-VTP、Adaptive Keyframe/Clip Selection。

---

5. 长视频处理策略（避免“抽帧丢信息”）

- 候选优先：用 TAD/STAD + TSG 生成候选时间段（高召回），避免对整段视频一刀切抽帧。
- 查询自适应采样：按具体问题（如“是否体罚”）进行关键帧/关键clip自适应选择（覆盖+相关性权衡）。
- 层级记忆：将 45 分钟切分为“章节 → 事件 → 关键clip”，LLM在不同粒度读写记忆槽；
- Token压缩：视觉token合并/剪枝（静态/重复内容合并），保留手-物体-接触等关键信息；
- 流式增量：边看边写“时间线摘要”，支持事后回溯与多轮提问。

---

6. 规则建模与违规判定

- 策略库：学校/地区规范 → 结构化规则条款（如“辱骂/体罚/侮辱性词汇/器具使用/长时间离

岗”等）。

- **融合判定**：视觉置信度 × ASR粗鄙/仇恨/辱骂检测 × 触碰/推搡的骨架证据 × 时序一致性。
- **解释模板**：输出“**何时**（时间戳）、**谁**（轨迹ID/角色）、**做了什么**（动作/语音）、**为何违规**（匹配条款）”。

---

7. 训练与标注

- **最小可行标签集**：
  - 视觉：`使用手机/睡觉/击打/推搡/持续离岗/眼神偏离黑板/接触学生` 等；
  - 语音：`脏话/辱骂/威胁/讽刺/大喊/嘲笑` 等（含语调/情绪要素）；
  - 关系：`手-手机-目光`，`教师-学生-接触`。
- **半自动标注**：基于初版检测/TSG生成伪标签 → 人工纠正 → 迭代。
- **困难样本挖掘**：遮挡、背影、拥挤、噪声、方言、口罩。

---

8. 评测方案与指标

- **检测/定位**：mAP@tIoU（TAD/STAD）、段级F1/Rec@K、偏移误差（秒）。
- **语音文本**：ASR WER/CER、粗鄙/仇恨检测F1、说话人分离DER。
- **综合事件**：端到端事件级精确率/召回率/F1；误报复核通过率。
- **解释质量**：人工打分（相关性、完整性、可读性）。

---

9. 计算与部署建议（4×A100）

- **并行划分**：解码/ASR、检测跟踪、TAD/TSG、MLLM 推理分别绑定 GPU；
- **特征缓存**：预提取 clip 级特征（如 2s/4s），共享给 TSG/LLM，避免重复编码；
- **混合精度**：FP16/FP8；大模型采用 vLLM/流水线并行；
- **吞吐实践**：45 分钟课程可在“离线批处理”模式下完成全量分析；在线监测模式以候选触发 LLM。

---

10. 预期效果（目标值，供里程碑考核）

- 端到端事件级F1 ≥ 0.80（主类目：玩手机/睡觉/粗鄙言论/肢体冲突）。
- 关键事件时间定位误差 ≤ 2–3 秒；
- 人机协同复核中，可解释性评分 ≥ 4/5。

---

11. 推荐开源组件（示例）

- 音频：Whisper / Paraformer（FunASR）/ SenseVoice；pyannote.audio（说话人分离）。
- 视觉：YOLOv8/10、ByteTrack、MMPose；VideoMAE / InternVideo2。
- 长视频与多模态：MovieChat、MA-LMM、TimeSearch、ReVisionLLM、VideoTree、Token Merging/PruneVid/Keyframe-Selection。

---

12. 里程碑

1. 第1月：数据与合规方案、原型流水线（ASR+检测+跟踪+基础TSG）。
2. 第2月：加入TAD/STAD与记忆式MLLM，完成事件级评测基线。
3. 第3月：引入自适应关键clip选择与token压缩，提升长视频准确性与效率。
4. 第4月：组织人工复核闭环与偏见审计，形成可交付报告模板与SDK。