

# SHAPE引擎四层架构

(Semantic Hierarchical Attention Profiling Engine)

## 第一层：数据预处理层 (Data Preprocessing Layer)

输入：课堂录像  
(Video + Audio)

音视频同步  
(Cross-Correlation)

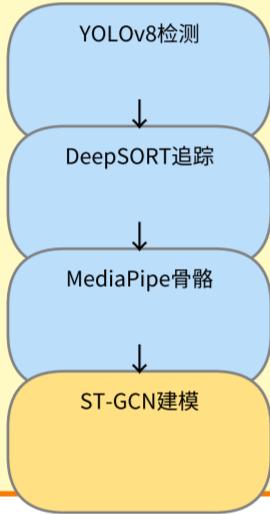
创新点1：语义驱动分段

(完整率 76.6%→95.3%)

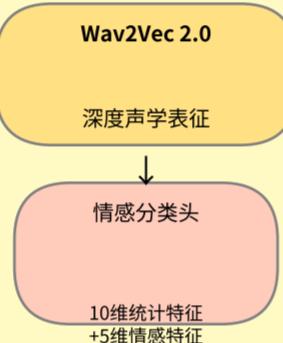
输出：N≈175个  
语义单元/课

## 第二层：特征提取层 (Feature Extraction Layer)

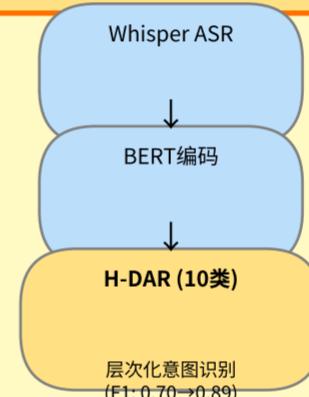
视觉模态 (20维)



音频模态 (15维)



创新点2：文本模态 (35维)



vs MFCC: +6.4%  
vs 单帧: +17.7%  
vs 关键词: +12.6%

创新点3：SHAPE跨模态注意力融合 (Fusion & Classification Layer)

模块1：  
特征投影



模块2：  
跨模态注意力



模块3：  
BiLSTM



模块4：  
注意力池化



模块5：  
风格分类器

→512维

6个α权重

时序建模

β权重

7类输出

性能提升：

vs Early Fusion: +6.2% | vs Late Fusion: +3.8%

vs 最佳单模态: +13.1% (78.3%→91.4%)

## 第四层：画像生成层 (Profiling & Application Layer)

风格分类结果

主导风格+置信度  
Top-2覆盖率98.1%

模态贡献度分析

基于α权重  
情感型:音频0.62  
互动型:视觉0.50

曲型片段提取

基于β权重  
Top-K关键时刻  
可解释性分析

最终性能：准确率93.5% | F1=0.91 | Kappa=0.89

注：三项创新形成完整改进链，语义分段→H-DAR→SHAPE，最终准确率93.5%

