西安交通大學
XI'AN JIAOTONG UNIVERSITY

软件学院
西安交通大学

基于层级语义建模与多智能体协同推理的免训练视频摘要方法

Training-free Video Summarization via Hierarchical Semantic Modeling and Multi-agent Collaborative Reasoning

汇报人：刘明圻

指导教师：徐亦飞

2026 年 4 月

1 研究背景与动机

研究背景

- 视频数据爆炸式增长，信息过载问题日益突出
- 视频摘要：自动保留最关键的帧或片段，帮助用户快速把握视频内容
- 传统方法普遍依赖标注与任务特定训练，跨场景泛化能力有限

研究趋势

多模态大语言模型（MLLMs）的快速发展，为免训练（**Training-free**）视频摘要提供了新范式：无需额外训练，仅依靠语言推理即可完成语义理解与关键片段筛选

现有方法的不足与本文动机

西安交通大学 XI'AN JIAOTONG UNIVERSITY

现有方法局限

1. 高层语义理解不足，难以把握视频主题
2. 单一评分机制，难以覆盖多维评价目标
3. 时序上下文利用有限，长视频一致性不足

本文两条技术路线

1. 层级语义建模
   实现帧 → 场景 → 视频的逐级理解
2. 多智能体协同推理
   由 Planner + 多 Expert 联合完成多维评分

论文结构

1. 第一章：绪论——研究背景、研究现状与研究内容
2. 第二章：相关理论基础与技术原理
3. 第三章：基于层级语义建模的免训练视频摘要方法
4. 第四章：基于多智能体协同推理的视频摘要方法
5. 第五章：结论与展望

2 相关理论基础

视频摘要任务概述

- 任务定义：给定视频 $\mathcal{V} = \{f_1, \ldots, f_T\}$，自动选择关键帧或关键片段生成摘要
- 常用数据集：
  - **SumMe**：包含 25 段用户生成视频，内容涵盖旅行、烹饪、极限运动和日常活动等开放域场景。每段视频由 15–18 位标注者分别给出代表性摘要片段。
  - **TVSum**：包含 50 段 YouTube 视频，覆盖新闻、纪录片、教程、Vlog 等 10 个主题类别。每段视频提供 20 位标注者逐帧给出的重要性评分。
- 评价指标：F_1 分数，以及 Spearman ρ、Kendall τ 秩相关系数

关键技术基础

西安交通大學 XI'AN JIAOTONG UNIVERSITY

大语言模型

- Transformer 自注意力架构
- 提示工程 (Prompt Engineering)
- 多模态语义对齐

智能体与协作

- 基于 LLM 的 Agent 决策
- 通过角色化提示实现职责分工
- 多智能体协同机制

3 层级语义建模的免训练视频摘要方法

方法总览（第三章）

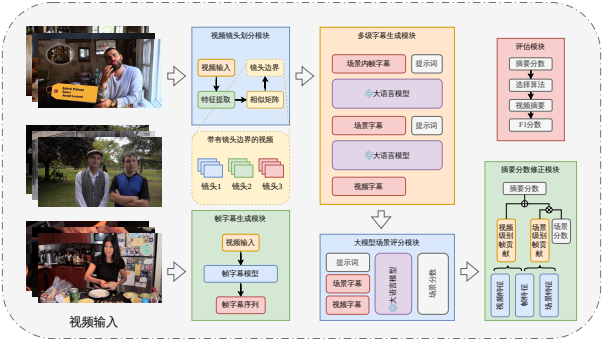


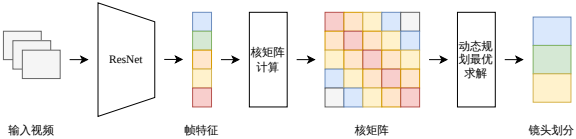
图: 基于大语言模型的免训练视频摘要框架结构图

阶段一：视频场景划分

- 利用 KTS 算法对视频进行结构化分段
- 优化目标：

$$\min_{\mathcal{B}} J(\mathcal{B}) = \sum_{k=1}^{K} \mathrm{Var}(s_k) + \lambda K$$

- 输出语义相近的场景单元 $\{s_1, s_2, \ldots, s_K\}$



阶段二：多级语义描述生成

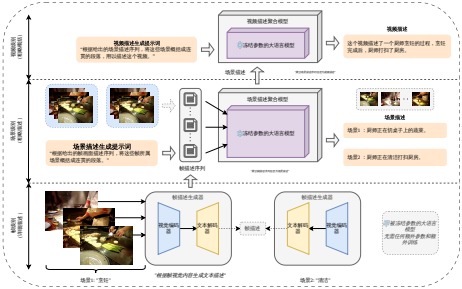
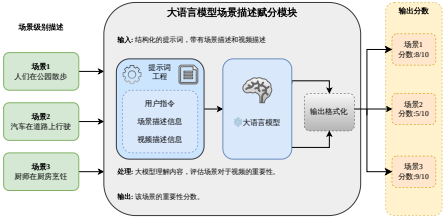


图: 多级视频语义描述生成流程

- 帧级：逐帧生成视觉描述
- 场景级：聚合同一场景的文本信息
- 视频级：生成全局主题摘要

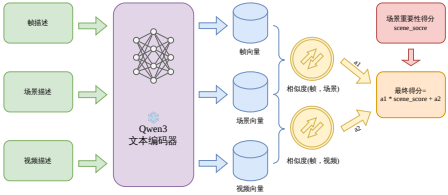
阶段三：全局主题约束的场景语义评分



- 以视频级摘要 Cap_v 作为全局参照
- 评估每个场景对整体主题的贡献程度

$$S_{\mathrm{LLM}}(k) = \mathrm{Norm}\Big(\mathrm{LLM}(\Pi(Cap_v, Cap_s(k)))\Big)$$

阶段四：帧级语义分数增强



- 场景内部：衡量帧与场景的一致性
- 全局范围：衡量帧与视频主题的一致性
- 二者融合得到最终帧级分数

$$F(p_j) = \alpha\, S_{\mathrm{LLM}}\big(\sigma(j)\big)\, S_{FS}(p_j) + S_{FV}(p_j)$$

第三章实验结果——F_1 分数对比

方法	类别	SumMe	TVSum
dppLSTM (2016)	监督	38.6	54.7
ACGAN (2019)	监督	47.2	59.4
MPFN (2024)	监督	51.9	62.4
CM-AVS (2026)	监督	52.4	**67.5**
DR-DSN (2018)	无监督	41.4	57.6
OM-AE (2019)	无监督	37.7	51.5
SUM-SR (2024)	无监督	51.3	60.2
SegSum (2025)	无监督	**54.0**	62.0
M3Sum-SP (2024)	免训练	43.6	56.9
M3Sum-PCoT (2024)	免训练	41.9	57.6
本章方法	免训练	**48.1**	**61.1**

表: 第三章方法与代表性方法的 F_1 对比（%）

- 两个数据集上均明显优于现有免训练基线
- TVSum 上达到 61.1%，已接近主流无监督方法

第三章实验结果——秩相关性对比

方法	SumMe		TVSum	
	ρ	τ	ρ	τ
dppLSTM (2016)	0.049	0.040	0.055	0.042
VASNet (2018)	0.170	0.160	0.170	0.160
DSNet-AB (2020)	0.059	0.051	0.129	0.108
DSNet-AF (2020)	0.046	0.037	0.138	0.113
DMASum (2020)	0.089	0.063	0.267	0.203
PGL-SUM (2021)	–	–	0.157	0.206
MSVA (2021)	0.230	0.200	0.210	0.190
iPTNet (2022)	0.119	0.101	0.163	0.134
CSTA (2024)	0.274	0.246	0.255	0.194
LLMVS (2025)	**0.282**	**0.253**	0.275	**0.211**
本章方法	**0.267**	**0.218**	**0.280**	**0.191**

表: 第三章方法与代表性方法的秩相关性对比

- TVSum 上 $\rho = 0.280$ 为最高值
- SumMe 上整体仍优于多数传统监督方法

第三章实验结果——消融实验

消融方案	SumMe	Δ	TVSum	Δ
w/o Caption & Scene Scoring	45.8	−2.3	58.9	−2.2
w/o S_{FS}（帧–场景相关性）	46.7	−1.4	60.6	−0.5
w/o S_{FV}（帧–视频相关性）	42.2	−5.9	60.3	−0.8
w/o Frame Refinement	42.2	−5.9	52.8	−8.3
完整方法	**48.1**	—	**61.1**	—

表: 第三章消融实验（Max/Avg F_1 %）

- 帧级增强模块对两个数据集的性能贡献最大
- 帧–视频全局相关性 S_{FV} 在 SumMe 上尤为关键

定性分析——SumMe Playing Ball 视频

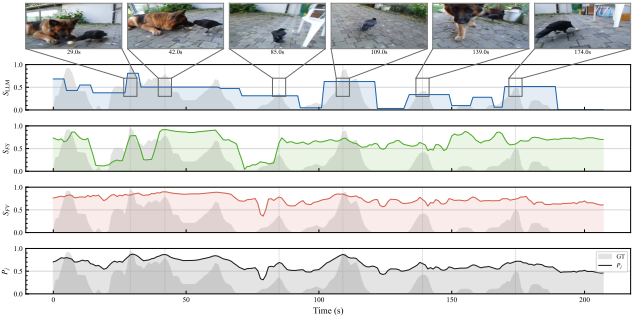


图: SumMe：Playing Ball 帧级评分曲线与关键帧可视化

- 场景分数能清楚地区分核心交互与静态过渡
- 局部与全局相关性共同补偿关键帧
- 最终高分区域与人工标注基本一致

定性分析——TVSum WxtbjNsCQ8A 视频

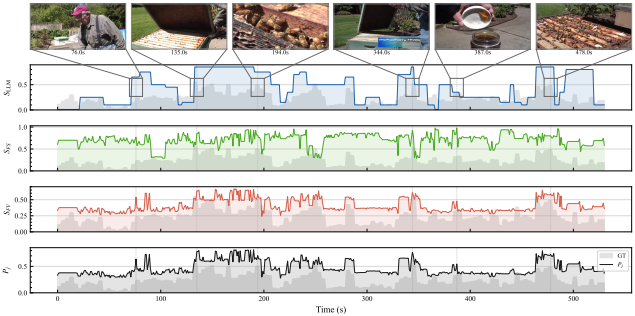


图: TVSum：养蜂人劳作视频帧级评分曲线与关键帧可视化

- 主题越清晰，场景分数的递进趋势越明显
- 核心劳动片段保持持续高分
- 融合评分在关键阶段形成连续高响应

参数敏感性分析——α 对 F_1 的影响

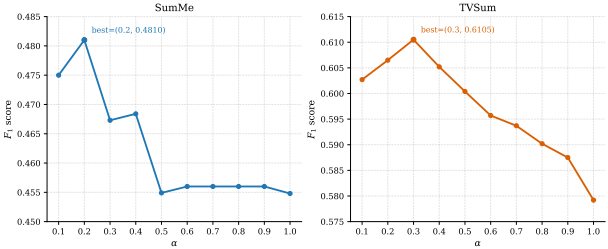


图: 平衡系数 α 对 SumMe / TVSum F_1 分数的影响

- SumMe 最优：$\alpha = 0.2$ ($F_1 = 48.1\%$)
- TVSum 最优：$\alpha = 0.3$ ($F_1 = 61.1\%$)
- 较小的 α 更利于平衡场景先验与全局主题匹配

4 多智能体协同推理的视频摘要方法

方法总览（第四章）

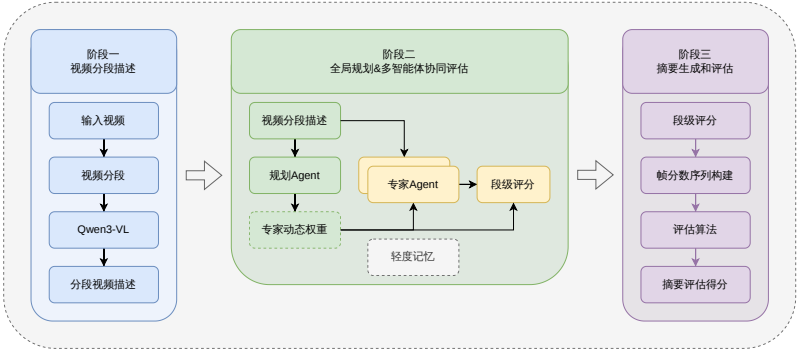
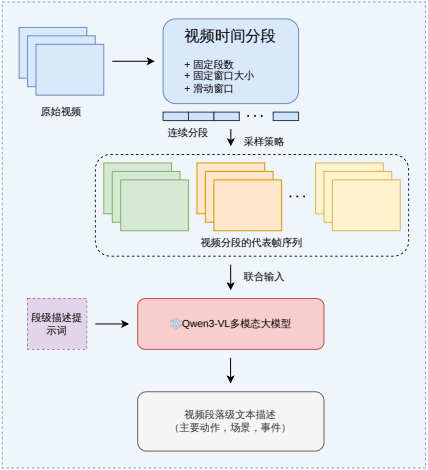


图: 基于大语言模型智能体协同的视频摘要方法总览

阶段一：视频结构化处理与段级描述生成



1. 均匀分段：将视频划分为 N 个连续片段
2. 片段内采样：每段均匀选取代表帧
3. 段级描述：由 VLM 生成片段级文本描述

$$Cap_s(k) = \mathrm{VLM}_{\mathrm{seg}}\left(I_k; \theta_{\mathrm{VLM}}, \pi_{\mathrm{seg}}\right)$$

阶段二：全局规划与多专家协同评分

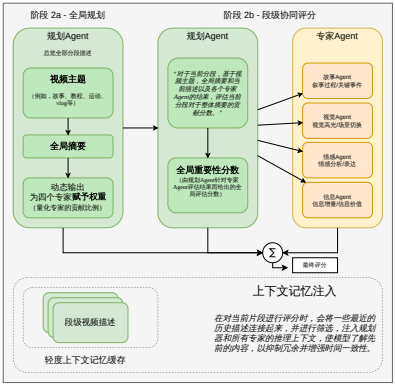


图: 全局规划与多专家协同评分流程

Planner Agent 与 Expert Agents

Planner Agent 全局规划

- 推断视频主题 z
- 生成全局摘要 Cap_v
- 分配专家权重 **w**
- 融合各 Expert 证据给出主分数

四类 Expert Agent

- **Story**：叙事推进
- **Visual**：视觉显著性
- **Emotion**：情绪表达
- **Information**：信息增量

分数融合：

$$S_{\text{final}}(k) = S_{\text{plan}}(k) + \sum_{m \in \{s,v,e,i\}} w_m \, S_m(k)$$

阶段三：记忆增强与帧级摘要生成

西安交通大学 XI'AN JIAOTONG UNIVERSITY

1. 上下文记忆注入——将前序片段描述拼接注入当前评分提示

$$\mathcal{M}_k^{(h)} = \{Cap_s(\max(1, k-h)), \ldots, Cap_s(k-1)\}$$

2. 帧级分数映射——段级分数回填至帧轴

$$F(t) = S_{\text{final}}\big(k(t)\big)$$

3. 背包优化——0/1 动态规划选择最优片段集合

$$\max_{\mathbf{x}} \sum_j x_j u_j \quad \text{s.t.} \quad \sum_j x_j l_j \le L, \; x_j \in \{0, 1\}$$

第四章实验结果——F_1 分数对比

方法	类别	SumMe	TVSum
dppLSTM (2016)	监督	38.6	54.7
ACGAN (2019)	监督	47.2	59.4
MPFN (2024)	监督	51.9	62.4
CM-AVS (2026)	监督	52.4	**67.5**
DR-DSN (2018)	无监督	41.4	57.6
OM-AE (2019)	无监督	37.7	51.5
SUM-SR (2024)	无监督	51.3	60.2
SegSum (2025)	无监督	**54.0**	62.0
M3Sum-SP (2024)	免训练	43.6	56.9
M3Sum-PCoT (2024)	免训练	41.9	57.6
第三章方法	免训练	48.1	61.1
本章方法	免训练	**53.6**	**55.6**

表: 第四章方法与代表性方法的 F_1 对比（%）

- 在 SumMe 上超过全部监督基线
- 相比第三章方法 SumMe 提升 5.5 个百分点
- 说明多智能体协同推理能够进一步提升摘要质量

第四章实验结果——秩相关性对比

方法	SumMe		TVSum	
	ρ	τ	ρ	τ
dppLSTM (2016)	0.049	0.040	0.055	0.042
VASNet (2018)	0.170	0.160	0.170	0.160
DSNet-AB (2020)	0.059	0.051	0.129	0.108
DSNet-AF (2020)	0.046	0.037	0.138	0.113
DMASum (2020)	0.089	0.063	0.267	0.203
PGL-SUM (2021)	–	–	0.157	0.206
MSVA (2021)	0.230	0.200	0.210	0.190
iPTNet (2022)	0.119	0.101	0.163	0.134
CSTA (2024)	0.274	0.246	0.255	0.194
LLMVS (2025)	**0.282**	**0.253**	0.275	0.211
第三章方法	0.267	0.218	0.280	0.191
本章方法	**0.282**	0.214	**0.323**	**0.224**

表: 第四章方法与代表性方法的秩相关性对比

- TVSum ρ=0.323 为全表最高
- 说明排序结果与人工标注最一致
- 在完全免训练条件下，表现已达到甚至超过 LLMVS

定性分析——SumMe Car Over Camera 视频

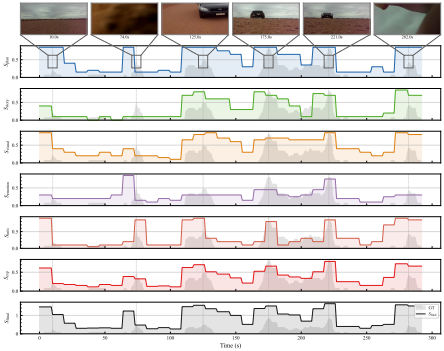


图: SumMe：Car Over Camera 多智能体评分曲线与关键帧可视化

- 前段空镜被压低，说明系统能识别低信息片段
- 车辆出现与扬尘片段形成全视频主峰值
- 末段人物出镜被识别为叙事收束节点

定性分析——TVSum qqR6AEXwxoQ 视频

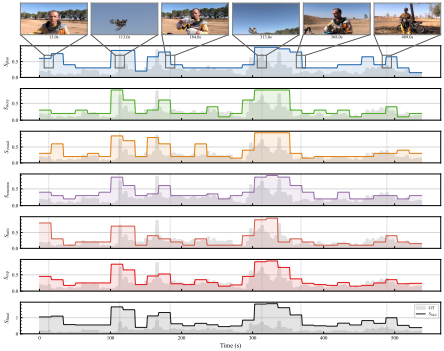


图: TVSum：越野摩托视频多智能体评分曲线与关键帧可视化

- 口述讲解片段整体维持低分
- 跳跃与特技片段获得最高融合评分
- 系统能够区分核心动作与重复运动画面

第四章实验结果——消融实验

消融方案	SumMe	Δ	TVSum	Δ
w/o Planner Agent	51.74	−1.86	47.27	−8.33
w/o Expert Agents	50.43	−3.17	49.28	−6.32
w/o Story Expert	47.58	−6.02	48.71	−6.89
w/o Visual Expert	49.82	−3.78	50.80	−4.80
w/o Emotion Expert	49.66	−3.94	49.32	−6.28
w/o Information Expert	51.99	−1.61	49.77	−5.83
完整方法	**53.6**	—	**55.6**	—

表: 智能体模块消融实验（F_1 %）

- 去除 Story Expert 导致最大性能退化，说明叙事推进是核心维度
- Planner Agent 在 TVSum 上作用尤为突出（Δ= − 8.33）

第四章参数分析——分段数量 N 的影响

西安交通大学 XI'AN JIAOTONG UNIVERSITY

分段数 N	SumMe Max F_1	TVSum Avg F_1
8	51.89	51.41
16	50.66	52.25
32	**53.6**	**55.6**

表：固定 DeepSeek 后端下不同分段粒度的性能对比（%）

- N 增大意味着单个推理单元更短、段级描述更细
- $N = 8$ 与 $N = 16$ 表现接近，粗粒度设置优势不明显
- 当 $N = 32$ 时，SumMe 与 TVSum 均达到最优性能
- 更细粒度的候选片段有利于 Planner 与 Expert 做出更精确判断

5 结论与展望

主要贡献

1. 层级语义建模的免训练视频摘要方法
   - 构建帧 → 场景 → 视频三级语义表示链路
   - 引入全局主题约束与帧级语义增强，生成主题一致的摘要
2. 多智能体协同推理的视频摘要框架
   - 设计 Planner 全局规划 + 四类 Expert 多维协同评分机制
   - 引入上下文记忆增强，提高摘要时序完整性
3. 两条技术路线在 SumMe / TVSum 上均取得有竞争力的结果，且无需额外训练与参数更新

未来工作方向

1. 提升长视频建模能力：结合内容感知的自适应分段与动态检索式记忆机制，使推理单元更贴近真实事件边界，并增强长程语义一致性
2. 面向用户需求的个性化摘要：引入用户意图感知与动态专家权重调整机制，使摘要结果能够更好适配不同用户偏好与查询目标

西安交通大學
XI'AN JIAOTONG UNIVERSITY

谢 谢

基于层级语义建模与多智能体
协同推理的免训练视频摘要方法

刘明圻

2026 年 4 月