西安交通大學
XI'AN JIAOTONG UNIVERSITY

计算机科学与技术学院
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

姓名：刘明圻

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：自动识别并保留最具代表性的帧或片段，支持用户快速理解视频主题
- 传统方法（监督/无监督/弱监督）依赖标注数据与任务特定训练，跨场景迁移能力有限

研究趋势

多模态大语言模型（MLLMs）的发展为免训练 (**Training-free**) 视频摘要提供了新范式——无需额外参数更新，利用语言推理能力完成语义分析与关键片段筛选

现有方法的不足与本文动机

西安交通大學 XI'AN JIAOTONG UNIVERSITY

现有方法局限

1. 高层语义建模不足——难以刻画视频主题
2. 单一评分器——无法兼顾多维评价目标
3. 时序上下文利用有限——长视频一致性差

本文两条技术路线

1. 层级语义建模
   帧 → 场景 → 视频的多级语义理解
2. 多智能体协同推理
   Planner + 多 Expert 多维度协同评分

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 $\mathcal{V} = \{f_1, \ldots, f_T\}$，选择关键帧/片段集合构成摘要
- 常用数据集：
  - SumMe（25 段用户视频，评测 Max F_1）
  - TVSum（50 段 YouTube 视频，评测 Avg F_1）
- 评价指标：F_1 分数、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
- 角色化提示实现职责分离
- 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）

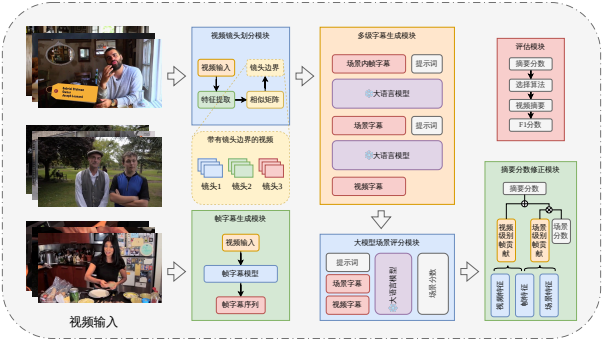


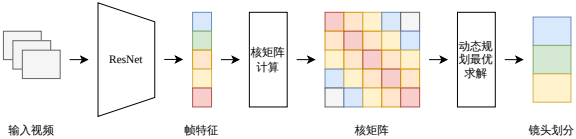
图: 基于大语言模型的免训练视频摘要框架结构图

阶段一：视频场景划分

- 采用 **KTS** 算法对视频进行结构化分段
- 优化目标：

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

- 将连续视频流组织为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

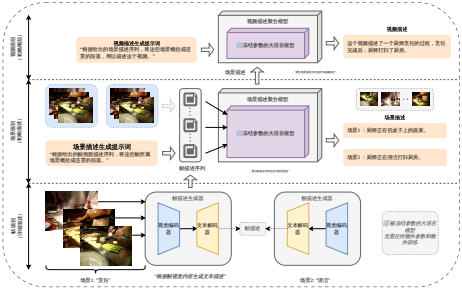
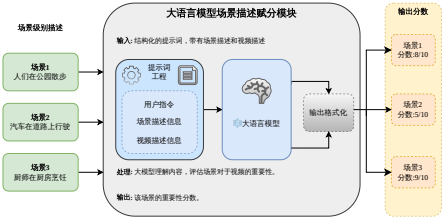


图: 多级视频语义描述生成流程

- 帧级描述：$Cap_f(t) = \mathrm{VLM}(I_t; \theta_{\mathrm{VLM}})$
- 场景级聚合：$Cap_s(k) = \mathrm{LLM}_{\mathrm{scene}}(\mathcal{T}_k; \theta_{\mathrm{LLM}})$
- 视频级摘要：$Cap_v = \mathrm{LLM}_{\mathrm{video}}(\{Cap_s(1), \ldots, Cap_s(K)\}; \theta_{\mathrm{LLM}})$

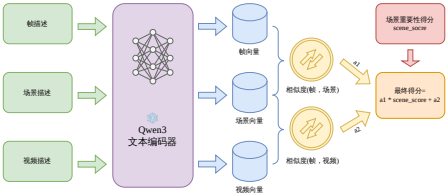
阶段三：全局主题约束的场景语义评分



- 联合输入场景描述 $Cap_s(k)$ 与视频级摘要 Cap_v
- LLM 评估该场景对整体主题的贡献

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

阶段四：帧级语义分数增强



- 帧–场景相关性：$S_{FS}(p_j) = \mathbf{z}_{f,j}^{\top} \mathbf{z}_{s,j}$
- 帧–视频相关性：$S_{FV}(p_j) = \mathbf{z}_{f,j}^{\top} \mathbf{z}_{v}$
- 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（%）

第三章实验结果——秩相关性对比

西安交通大學 XI'AN JIAOTONG UNIVERSITY

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

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贡献最大
- 帧–视频全局相关性 S_{FV} 在 SumMe 上至关重要

4 多智能体协同推理的视频摘要方法

方法总览（第四章）

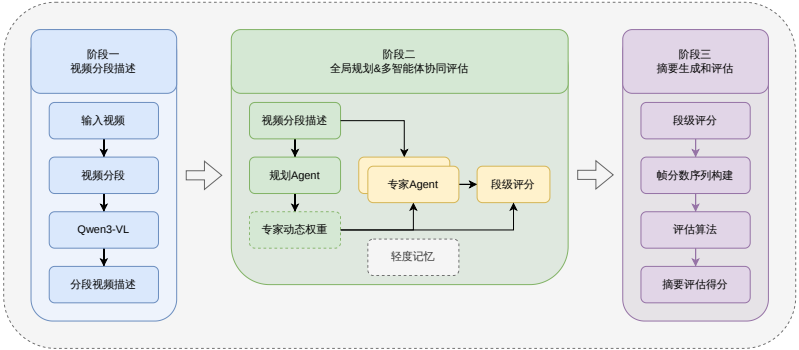
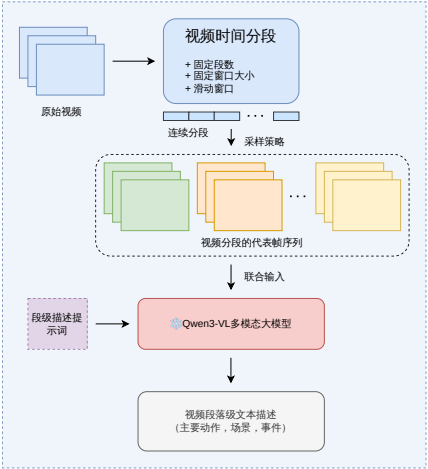


图: 基于大语言模型智能体协同的视频摘要方法总览

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：VLM 联合多帧输入生成段级文本

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

阶段二：全局规划与多专家协同评分

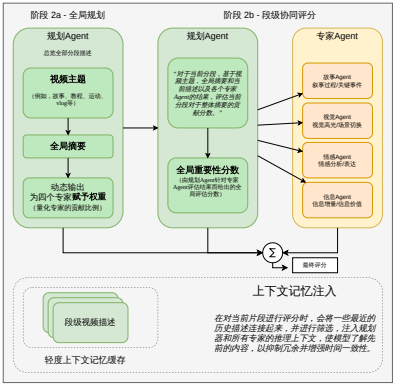


图: 全局规划与多专家协同评分流程

Planner Agent 与 Expert Agents

西安交通大学 XI'AN JIAOTONG UNIVERSITY

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 $\mathbf{w}$
- 参考各 Expert 证据给出综合主分数

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（%）

- SumMe 上超越全部监督基线，达到免训练方法最优
- 相比第三章方法 SumMe 提升 5.5 个百分点

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 取得最高 Spearman 系数
- 完全免训练条件下取得与 LLMVS 可比甚至更优的排序表现

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 %）

- Story Expert 消融导致最大性能退化
- Planner Agent 在 TVSum 上作用尤为突出（Δ= − 8.33）

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
   - 帧 → 场景 → 视频三级语义描述 + 全局主题约束场景评分
   - 帧级语义相关性增强，实现主题一致的摘要生成
2. 多智能体协同推理的视频摘要框架
   - Planner 全局规划 + 四类 Expert 多维度评分
   - 上下文记忆增强 + 背包优化摘要构造
3. 两条路线在 SumMe / TVSum 数据集上取得有竞争力的性能，无需额外训练与参数更新

实验结果总结

西安交通大學 XI'AN JIAOTONG UNIVERSITY

第三章方法

- SumMe: 48.1% F_1
- TVSum: 61.1% F_1
- 超越同类免训练方法 6–10 个百分点

第四章方法

- SumMe: 53.6% F_1
- TVSum: 55.6% F_1
- SumMe 上超越全部监督基线
- TVSum ρ=0.323 秩相关性最优

未来工作方向

1. 探索更自适应的视频分段策略，提升分段与事件边界一致性
2. 研究更稳定的记忆组织与检索机制，增强长视频全局一致性
3. 优化帧级映射机制，使段级推理结果更精确地对齐到帧级分数
4. 引入更强的多模态描述模型与更细分的专家角色
5. 提升系统运行效率与多模型协同扩展能力

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