基于视频大模型的通用视频高光剪辑

一、赛题背景

随着短视频、直播录制、活动记录、生活影像和自媒体内容生产快速增长，用户每天产生大量类型多样的视频素材。此类视频通常时长较长、精彩片段分布稀疏、镜头抖动明显、画面质量差异较大，人工筛选高光片段并完成裁切、构图和成片制作需要较高时间成本。传统视频剪辑主要依赖人工经验，难以在海量素材中稳定识别内容变化、事件节奏、镜头运动、场景氛围、声音线索和视觉美感等综合因素。

本赛题面向“从长视频中自动发现并生成高光短片”的真实需求，鼓励参赛者综合运用视频理解、时序定位、语义重点分析、视觉美学评价、多模态建模和智能剪辑等技术，构建可复现、可评测、可落地的高光片段自动识别与剪辑算法，提升通用视频内容生产效率和智能化水平。

二、赛题应用场景

赛题对应的实际应用场景为通用视频自动剪辑平台。用户上传一段原始视频后，系统自动识别其中具有传播价值或回看价值的精彩内容，按帧给出推荐构图区域，最终生成适合横屏或竖屏发布的裁剪视频或高光短片。该能力可服务于体育赛事、户外活动、演出活动、旅行记录、生活 Vlog、课堂讲座、会议培训、内容运营和媒体生产等多类场景。例如，个人用户可以快速从长视频中获得可分享的高光片段；活动组织方可以批量生成集锦素材、宣传短片和社交媒体内容；内容运营团队可以快速定位有信息量、有情绪张力或画面表现力的片段；平台服务方可将该能力作为增值功能提升用户体验，降低人工剪辑成本，提高内容生产效率和传播效率。

三、赛题任务

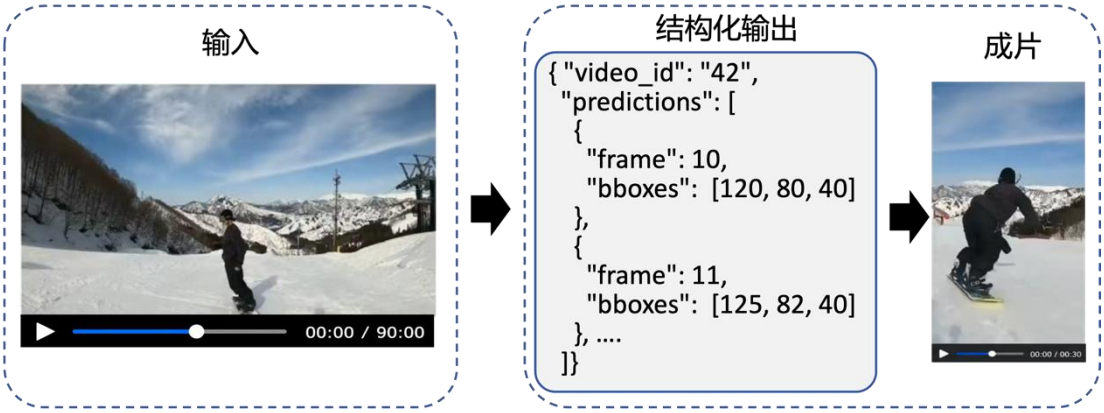
本赛题旨在利用视频理解、多模态内容分析与智能剪辑技术，实现通用视频场景下精彩内容的精准高效发现与自动裁剪，即对原始视频中的画面内容、事件变化、镜头运动、音频节奏和视觉质量等信息进行协同处理分析，依托内容语义理解与视觉美学评价，自动确定需要保留的高光帧序列，并逐帧输出推荐构图区域。

（一）任务描述

在通用视频高光剪辑任务中，视频时序内容作为核心分析对象，内容语义、事件节奏、镜头变化、音频线索和视觉质量等信息共同构成判断高光内容和构图区域的重要依据。给定一段原始视频，参赛者需要对视频帧序列、画面重点、镜头运动、声音节奏和画面美学质量进行协同分析和处理，从而筛选出应进入成片的高光帧，并为每个高光帧输出推荐构图框。该构图框可以围绕人物、物体、动作或文字等关键内容生成，也可以在无明确主体时给出整体画面或局部重点区域，实现从原始长视频到结构化裁剪结果的自动映射。赛题要求参赛者设计基于视频理解、多模态内容分析或视频

大模型的算法，对通用视频中的时序特征、空间重点信息和质量评价信息进行联合建模，最终完成高光帧选择与逐帧推荐构图输出，如图 1 所示。

图 1 基于视频大模型的通用视频高光剪辑的任务描述



（二）任务输入输出说明

1. 算法输入：原始视频文件、测试样本索引和指定输出格式 targetRatioWH；其中 targetRatioWH 用于指定该样本最终高光成片的目标画幅比例（16:9；9:16）。targetRatioWH 采用[target_w,target_h]形式，例如[16,9]表示横屏，[9,16]表示竖屏。测试阶段赛方提供 JSONL 索引文件，每行对应一个视频样本，需记录 video_id 及 targetRatioWH。

2. 算法输出：参赛者需按照索引文件读取视频，对进入成片的高光帧进行选择，并生成逐帧构图框预测结果。每个视频对应一组逐帧预测结果，统一写入 predictions 字段。predictions 为数组，每个元素包含 frame、bboxes 等字段，其中 frame 为原视频帧序号。字段名沿用 bboxes，但每个 frame 仅提交一个构图框，bboxes 为单个三元组[x,y,w]，坐标以原视频分辨率为基准。x、y 表示推荐构图框左上角坐标，w 表示构图框宽度。targetRatioWH=[$target\_w, target\_h$]，评测程序自动计算高度 h，即 h = w × target_h / target_w，不要求参赛者同时输出 w 和 h。若该帧存在明确主体，bboxes 可围绕关键主体生成；若无明确主体，bboxes 可表示整体画面或最具信息量、观赏性的局部区域。对于无明显高光内容的视频，可输出空 predictions 列表。

四、数据集及数据说明

（一）数据来源

赛方不提供训练集，参赛者可自行使用合规公开视频数据、公开数据集、自建数据和公开预训练模型进行算法训练与优化，并需在技术报告中说明数据来源、采集或清洗方式、使用范围和许可证情况。赛方仅提供样例数据、数据格式说明、提交样例、基础评测脚本以及初赛、复赛、半决赛阶段测试集。测试集不开放标注，仅可用于赛

事提交评测；参赛者不得将测试数据上传至公开平台、第三方模型服务、公开仓库或社交媒体，不得私自传播、共享、转存、人工标注、用于模型训练或反向构造规则。

（二）数据内容

测试数据以通用视频为主，覆盖横屏、竖屏、第一视角、第三人称跟拍、固定机位、屏幕录制等拍摄方式，场景可包括体育赛事、户外活动、演出记录、旅行生活、活动集锦、课堂讲座、会议培训和媒体素材等。

样例数据用于说明视频文件组织、JSONL 索引结构和提交格式，不代表正式测试集的完整分布。正式测试集将重点考查模型在不同内容类型、不同拍摄视角、不同画面质量和不同视频时长下的泛化能力。

（三）数据格式

以 JSONL 文件组织测试样本索引和参赛预测结果。索引文件中每行对应一个视频样本，包含 video_id、视频文件名和 targetRatioWH 等必要索引信息；其他视频属性可由程序读取原始视频获得。targetRatioWH 采用[target_w,target_h]形式。参赛者提交结果时，应提交预测结果 JSONL 文件，每行对应一个视频样本，并在该样本下补充 predictions 字段。predictions 为该视频的逐帧预测列表，每个元素包含 frame、bboxes 等字段，示例形式为{"frame":10,"bboxes":[120,80,40]}。其中 bboxes 为单个构图框三元组，第三个数值 w 表示构图框宽度，评测程序根据 targetRatioWH 自动计算高度 h，不要求参赛者输出 h。

字段约束：frame ≥ 0，且不得超出原视频总帧数；bboxes 需满足 x ≥ 0、y ≥ 0、w> 0。将[x,y,w]换算为[x,y,w,h]后，要求构图框不得超出原视频画面范围。视频总帧数和画面尺寸由评测程序读取原始视频获得，不要求参赛者在提交文件中额外填写。同一视频内 frame 不应重复，predictions 建议按 frame 升序排列。字段缺失、视频编号无法匹配、帧号越界、坐标越界、NaN 数值、空框或换算后构图框越界等异常情况可被评测程序判定为无效预测。

（四）数据特色

本赛题测试数据面向真实通用视频内容生产场景，突出内容类型、拍摄视角、视频时长和画面质量的多样性。数据中既包含人物、物体或动作重点明显、镜头跟随清晰的片段，也包含无明确主体、重点分散、主体尺度变化、快速运动、遮挡、镜头抖动、背景干扰、曝光变化和横竖屏比例差异等真实拍摄情况。此类数据设计有助于全面考查算法对高光帧选择、内容完整性、重点可辨识性、画面观赏性和推荐构图稳定性的综合判断能力，避免模型仅依赖单一内容类别、固定镜头或简单画面质量规则取得高分。

五、算法设计要求

（一）模型类型

鼓励参赛者围绕通用视频高光剪辑任务，基于视频理解大模型或多模态大模型进行算法设计，并结合 SFT、RLHF 等微调方法提升模型对精彩内容、高光帧选择、画面重点和构图质量的判断能力。参赛者可基于 Qwen、SAM 等开源模型进行轻量化适配，探索时序建模、语义重点定位、画面美学评估、长视频分段推理、推理加速与显存优化等方案，并保证训练、推理和后处理流程可复现。

（二）创新性

鼓励参赛者围绕高光剪辑任务提出创新性改进，例如高召回候选帧或候选序列生成、内容完整性建模、帧级冗余控制、语义重点或美学构图驱动的构图框生成、画面质量与内容精彩度联合评分、轻量化推理和长视频分段处理等。

（三）鲁棒性与泛化能力

算法应能适应不同内容类型、视频比例、拍摄视角、视频时长和画面质量。当视频存在模糊、抖动、遮挡、复杂背景、重点区域不明显、主体尺度变化或光照变化时，模型仍应稳定输出可解析、可评测的结果，避免因单个异常视频导致批量推理中断。

（四）工程实现要求

参赛方案应在规定评测环境和时间限制内完成测试视频推理，输出 JSONL 预测结果文件需满足赛方格式要求。提交代码应包含清晰的环境依赖、模型加载说明、推理入口、结果生成逻辑和必要的异常处理，确保评审团队能够复现线上提交结果。

六、性能指标要求

本赛题采用自动评测为主、人工复核为辅的评价方式。自动评测以视频平均连续 IoU 加权 F 分数作为最终指标：先对每个视频分别计算连续 IoU 加权F_i，再对所有视频取平均。评测程序同时保留每条有效预测的真实IoU值、按视频统计的平均IoU、每个视频的F_i和最终平均分，便于结果复核与误差分析。

（一）预测匹配规则

评测时以 video_id 和 frame 为基础匹配条件。评测程序先根据目标画幅比例将参赛者提交的[x,y,w]换算为标准构图框[x,y,w,h]，再将同一视频、同一 frame 上的有效预测 bboxes 与该 frame 的真实构图区域直接计算 IoU。空间 IoU 计算方式为：

$$IoU = \frac{area(B_{pred} \cap B_{gt})}{area(B_{pred} \cup B_{gt})}$$

其中B_{pred}为预测 bboxes，B_{gt}为真实构图区域。每个 frame 原则上只提交一个预测结果；若同一视频同一 frame 存在多个预测结果，仅按提交顺序保留第一个有效预测用于计算IoU，其余重复预测计为误报且不获得IoU得分。frame 不一致的预测不参与跨帧匹配，即使构图区域相似也不计为正确。

（二）视频平均连续 IoU 加权 F1 计算方式

对第i个视频，统计该视频内有效预测数量$N_{pred,i}$、真实标注数量$N_{gt,i}$以及所有有效同帧预测的IoU总和$S_{IoU,i}$。对有效同帧预测，其贡献值为该预测与对应真实构图区域的真实IoU；frame 错误、重复预测或格式无效的预测不获得IoU得分，并计入该视频的$N_{pred,i}$；未被有效同帧预测覆盖的真实标注计入该视频的$N_{gt,i}$但不增加$S_{IoU,i}$。

其中每个视频的精确率、召回率和F_i分数分别为：

$$P_i = \frac{S_{IoU,i}}{N_{pred,i}};\ R_i = \frac{S_{IoU,i}}{N_{gt,i}};\ F_i = \frac{2 \times P_i \times R_i}{P_i + R_i}$$

当某个视频的$N_{gt,i}$=0 且$N_{pred,i}$=0 时，说明该视频无高光标注且参赛者未输出高光预测，该视频F_i记为 1；当$N_{gt,i}$=0 但$N_{pred,i}$>0，或$N_{gt,i}$>0 但$N_{pred,i}$=0 时，该视频F_i记为 0；其余情况按上述公式计算。设测试集视频数量为V，最终成绩为所有视频F_i的平均值：

$$F_{video} = \frac{F_1 + F_2 + \cdots + F_V}{V}$$

排行榜百分制展示按下式换算：

$$Score = F_{video} \times 100$$

（三）无效结果判定

若预测结果存在 frame 越界、bboxes 坐标越界、字段缺失、数值为 NaN、视频编号无法匹配、JSONL 结果文件结构错误、推理超时或无法复现等情况，评测程序可将对应预测或对应提交判定为无效。

七、功能要求

参赛者提交的解决方案应支持批量读取测试视频并自动生成标准 JSONL 提交文件；支持对每个视频输出逐帧 predictions 列表；支持为高光帧输出推荐构图位置和单一尺度值，由评测程序依据目标画幅比例换算为 9:16、16:9 等不同比例的视频剪辑构图框；支持结果格式校验，对帧号越界、bbox 坐标越界、重复帧、空列表、NaN 数值等异常结果进行处理；支持日志记录；总决赛阶段需支持从原始视频输入到 JSONL 结果文件生成的端到端复现流程。

八、开发环境

（一）软件环境

操作系统建议为 Linux（Ubuntu 20.04/22.04）；编程语言建议为 Python 3.8 及以上；深度学习框架可使用 PyTorch、Jittor、TensorFlow 等主流框架；若采用视频理解大模型或多模态大模型路线，可按需使用 Transformers、ModelScope 等工具。提交时需提供 requirements.txt、environment.yml 或 Dockerfile。

（二）硬件环境

模型开发和训练阶段可使用本地工作站或云端 GPU 资源。训练所用数据由参赛团队自行准备并保证合规。模型如需特殊硬件、超大显存或额外商业服务，需提前说明，赛方可基于公平性要求限制使用。

九、成绩评价

初赛和复赛采用自动评测方式，排行榜依据视频平均连续IoU加权F分数排序。初赛成绩主要用于算法验证和有效提交判断，复赛成绩用于省赛排名和决赛晋级。

自动评测根据统一测试集和统一评测脚本计算。参赛者需保证提交的 JSONL 文件能够被评测程序正确解析，且不得修改测试集原始 video_id、视频文件名或读取路径等索引信息。对于帧号越界、坐标越界、字段缺失、数值包含 NaN、视频编号无法匹配、JSONL 结果文件结构错误等情况，评测程序可直接判定对应样本或对应提交无效。提交格式错误、无法运行或无法复现的成绩视为无效成绩。

线下决赛阶段，在半决赛客观评分基础上，由评审专家对技术报告、代码复现、演示效果、方案创新性和答辩表现进行综合评价。若复现结果与线上提交结果存在显著差异，赛方可按复现结果计分或取消成绩。

十、解题思路

知识点：重点围绕视频理解大模型的任务适配与微调展开，包括通用高光指令数据构建、长视频时序采样、高光帧选择、语义重点定位与逐帧构图框生成、多模态特征融合、SFT 监督微调、偏好对齐和强化学习优化。参赛者需理解如何把“精彩程度、内容完整性、重点可辨识性、画面稳定性、误报控制”等剪辑偏好转化为可训练、可评测的模型目标。

思路引导：可先基于开源视频大模型或多模态大模型构建初始方案，在公开数据集/标注数据集基础上，使用 SFT 让模型学会按赛事格式输出 frame 和 bboxes。具体实现上，可由视频大模型负责高光帧选择、内容语义理解和构图意图判断，再结合 SAM 或同类分割模型生成候选 mask，通过提示点、提示框或语义区域提示完成关键区域分割，并将 mask 外接矩形、主体区域或画面重点区域转换为目标比例下的左上角坐标和单一尺度值，形成最终 bboxes；对于无明确主体的帧，可利用视频大模型给出的语义重点或美学中心作为 SAM 提示，生成更稳定的局部构图候选。在此基础上，引入强化学习或偏好优化方法，将视频平均连续 IoU 加权 F 分数、真实 IoU、漏检惩罚、误报惩罚、SAM 候选框质量、构图稳定性和时序平滑性等设计为奖励信号，使模型更倾向于输出 frame 准确、构图框与真实重点区域高度重合且冗余较少的高光结果。强化学习优化可采用 DPO、GRPO 或同类方法，但应避免只追求单一分数而破坏成片连贯性。方案还需配合 SAM 推理加速、mask 到 bbox 后处理、跨帧框平滑、输出格式校验、显存与速度优化等工程模块，并完整记录数据来源、微调配置、SAM 模型版本与提示策略、奖励函数、推理参数和后处理规则，确保结果可复现、

可解释、可公平评测。

十一、赛题约束条件

（一）算法约束

允许使用公开预训练权重、通用视觉/视频基础模型和开源多模态模型，但不得包含赛事测试集或未授权私有数据。禁止调用商业闭源模型 API、在线人工服务或外部云端推理服务完成测试集推理，除非赛方另行明确允许。提交结果必须由参赛队伍提交的代码和模型自动生成，赛方有权要求现场或离线复现。

（二）数据使用约束

赛事提供的样例数据和测试数据仅限本次比赛使用，不得传播、转售或用于非赛事相关商业用途。未经赛方许可，不得将测试集上传至公开平台、第三方模型服务、公开仓库、社交媒体或其他外部系统，不得以任何形式共享、转存、人工标注、人工修正预测结果、用于模型训练或反向构造规则。参赛者可使用合规公开数据集或自建数据进行训练，但须在技术报告中完整披露数据来源、采集/清洗方式、许可证和使用范围。

参赛团队和组织方均应履行保密义务。违反数据保密、传播限制或公平竞赛要求的，赛方有权取消参赛资格及获奖资格，并保留追究相关责任的权利。

十二、参考资源

（一）文献资料

[1] Li Y, Cheng J, Jia S, et al. TempSamp-R1: Effective Temporal Sampling with Reinforcement Fine-Tuning for Video LLMs[C]//Advances in Neural Information Processing Systems, 2026, 38: 40692-40716.

[2] https://github.com/HVision-NKU/TempSamp-R1

[3] Qwen Team. Qwen3-vl technical report, 2025. arXiv:2511.21631.

[4] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. arXiv:2501.12948.

（二）在线资源

1. AIC 官网算法挑战赛道已发布赛题页面（https://www.aicomp.cn/tracks/tracks-1）。

2. ActivityNet、THUMOS、FineGym、Sports-1M、YouTube-8M 等公开视频理解数据集。

十三、提交要求

初赛、复赛：参赛队伍需提交测试集预测结果 JSONL 文件。

半决赛：参赛队伍需提交算法代码、预测结果 JSONL 文件、模型文件与环境、技术报告。具体提交内容如下：

（一）算法代码 参赛队伍应提交完整的算法代码，包括数据预处理、模型训练、预测推理等各个环节的代码。代码需使用符合规范的 Python 语言编写，具备清晰的注释和文档说明，以便评审人员理解和运行。

（二）预测结果文件、模型文件与环境 参赛队伍应提交测试的预测结果 JSONL 文件、训练好的模型文件（由于模型可能较大，请提供模型下载链接）和推理脚本，并提供模型的加载和使用说明，包括所需的运行环境、依赖库等信息。模型文件应能够在指定的测试环境中正常运行并输出预测结果 JSONL 文件。

（三）技术报告 参赛队伍应提交详细的技术报告，内容包括算法设计思路、模型架构图、实验设置（如训练参数、数据增强方法等）、性能分析（对主要指标和次要指标的详细分析）、算法的创新点及不足之处的分析和参考文献。技术报告格式采用 PDF，页数不限。内容格式可参考“附件：算法赛题技术方案大纲”

总决赛：提交内容及具体要求以组委会后续正式通知为准。

十四、奖金设置

为了鼓励参赛选手参赛积极性，本赛题根据总决赛成绩，对全国总决赛一等奖前 7 名参赛团队。奖金由赛题方负责，最终奖项数量、奖金金额和证书设置以组委会正式发布规则为准。

1. 冠军奖：第 1 名，奖金 50000 元/每团队
2. 亚军奖：第 2-3 名，奖金 30000 元/每团队
3. 季军奖：第 4-7 名，奖金 10000 元/每团队

十五、其他说明

公平性：严禁任何形式的作弊行为，包括但不限于测试数据泄露、上传测试集至公开平台或第三方服务、私自传播测试集、人工标注测试集、使用测试集训练模型、训练数据与测试数据重叠、人工修正测试结果、抄袭他人代码等。一经发现，立即取消参赛资格，并追究相关责任。

知识产权：参赛者提交的作品必须为原创，未在其他比赛中获奖或公开发表。比赛主办方有权对参赛作品进行展示、宣传等相关活动，但知识产权仍归参赛者所有。赛事提供的样例数据、测试数据、评测脚本和测试结果仅限本次比赛使用，未经许可不得用于其他用途。

十六、联系方式

赛题交流 QQ 群：389921008

邮 箱：yunhengli@mail.nankai.edu.cn

报名官网：www.aicomp.cn