中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-17

论文引用格式:Shi Qian, Hu Zhenyi, Liu Maofu, Hu Huijun, Zhang Xiaokang. An Event Knowledge-Enhanced Method for Multimodal Summarization of Natural Disaster News[J/OL]. Journal of Image and Graphics, XXXX:1-17. DOI: 10.11834/jig.260497.(石纤, 扈振义, 刘茂福, 胡慧君, 张效康. 事件知识增强的自然灾害多模态摘要生成方法[J/OL]. 中国图象图形学报, XXXX:1-17. DOI: 10.11834/jig.260497.)[DOI: 10.11834/jig.260497]

事件知识增强的自然灾害多模态摘要生成方法

石纤[1],扈振义[1],刘茂福[1],胡慧君[1],张效康[2]

1. 武汉科技大学计算机学院,湖北　武汉　430065; 2. 武汉大学人工智能学院,湖北　武汉　430072

摘　要: **目的**　自然灾害多模态摘要旨在为多篇图文新闻报道生成自然灾害事件概要,以辅助快速掌握灾情进展。然而,通用多模态大模型在处理自然灾害领域的长文本与多图像时,对灾害事件要素的利用不稳定,进而影响摘要的语义聚焦程度与事实一致性。**方法**　针对上述问题,提出一种事件知识增强的自然灾害多模态摘要生成方法EKE-MMS(Event Knowledge-Enhanced Multimodal Summarization for Natural Disasters)。该方法通过事件知识筛选灾害关键句,利用灾害类型知识模板引导局部摘要生成,再结合新闻发布时间和事件基本描述对局部摘要进行时序融合,最终生成事件级摘要。**结果**　在自建自然灾害多模态摘要数据集上的实验结果表明,针对BERTScore、ROUGE-L和CFSR,EKE-MMS方法最优分别提升了5.62%、3.27%和4.21%,且在人类偏好奖励模型的评测中表现更优。**结论**　EKE-MMS能够有效缓解复杂多源输入下灾害事件要素利用不稳定问题,提升摘要的语义质量和生成事实一致性。

关键词: 多模态摘要;自然灾害;事件知识;跨文档聚合;事实一致性

An Event Knowledge-Enhanced Method for Multimodal Summarization of Natural Disaster News

Shi Qian[1], Hu Zhenyi[1], Liu Maofu[1], Hu Huijun[1], Zhang Xiaokang[2]

1. School of Computer Science, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China; 2. School of Artificial Intelligence, Wuhan University, Wuhan, Hubei 430072, China

Abstract: **Objective** Multimodal summarization for natural disasters aims to generate concise and coherent event-level summaries from multiple news articles and their associated images, supporting rapid understanding of disaster development and emergency decision-making. Unlike general news summarization, this task requires accurate preservation of event elements such as occurrence time, location, intensity, casualties, affected population, rescue forces, relief supplies, infrastructure damage, and response progress. However, natural disaster reports usually contain long texts, heterogeneous images, duplicated descriptions, complementary facts, stage-specific updates, and occasional conflicts across documents. Under such complex multi-source inputs, general-purpose multimodal large language models may use disaster event elements unstably. They may omit salient quantitative or descriptive facts with high summarization value, or generate incor-

收稿日期:2026-08-07;**修回日期**:2026-10-31

基金项目:湖北省自然科学基金创新发展联合基金重点项目(2026AFC1121);国家自然科学基金面上项目(42371374);"十四五"湖北省优势特色学科(群)项目(2023D0302)

Supported by: Key Program of the Innovation and Development Joint Fund under the Natural Science Foundation of Hubei Province (2026AFC1121); General Program of the National Natural Science Foundation of China (42371374); "14th Five-Year Plan" Hubei Province Advantageous and Characteristic Disciplines (Groups) Project (2023D0302)

© 中国图象图形学报版权所有

rect values, times, locations, entities, and relations inconsistent with the source documents. These problems reduce factual completeness and factual consistency. To address them, this study proposes an Event Knowledge-Enhanced Multimodal Summarization method for Natural Disasters, termed EKE-MMS. **Method** EKE-MMS takes event knowledge as the central constraint and organizes content selection and summary generation around disaster occurrence, consequences, and response. The framework contains two modules: event knowledge-guided key-sentence extraction and event knowledge-constrained event-element generation. It also uses two preconstructed knowledge resources: key-sentence extraction rules and a disaster event knowledge template library. In the first module, each news article is segmented into sentences and assigned unique indices. The indexed text and extraction rules are jointly provided to a fixed language model. The rules guide the model to retain sentences describing the disaster, its consequences, and response measures, while filtering peripheral background, weakly related information, historical reviews, and redundant descriptions. When several sentences express the same fact, the model keeps the statement that is more specific, quantitative, and valuable for summarization. Instead of rewriting selected content, the model outputs only sentence indices, which are traced back to the original article in their original order. This index-based strategy preserves source wording and reduces factual deviations caused by secondary generation. The number of retained sentences is adaptively determined by valid facts, information density, and redundancy. The event knowledge-constrained event-element generation module comprises local image-text semantic calibration and global temporal fusion. During local calibration, the disaster type label selects the corresponding template from the knowledge library. The selected key sentences, associated images, and template are jointly input into a multimodal large language model to generate a local multimodal summary. This stage follows a text-dominant and image-assisted strategy: text provides the primary factual basis, whereas images supplement visually supported information such as disaster scenes, infrastructure damage, rescue activities, and environmental conditions. The template constrains event-element preservation and cross-modal evidence use, reducing visual misinterpretation and unsupported factual completion. Local summaries are first ordered by publication time and aligned using the fact occurrence time or statistical cutoff time as the primary temporal evidence. Together with the basic event description and a global fusion prompt, they are used to merge repeated facts, aggregate complementary information, and distinguish stage-specific updates and corrections from genuine conflicts before producing the event-level summary. Experiments were conducted with Qwen2.5-VL-7B-Instruct, LLaVA-v1.6-Mistral-7B, InternVL3-8B, and MiniCPM-V-4.5-8B. For each model, its original end-to-end setting served as the baseline. Performance was evaluated using ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, CFSR, and Skywork-Reward-V2-Llama-3.2-3B. **Result** EKE-MMS consistently improved all four multimodal generators over their original end-to-end settings. Across the evaluated models, the maximum gains in BERTScore, ROUGE-L, and CFSR reached 5.62, 3.27, and 4.21 percentage points, respectively. The method also achieved better Skywork preference scores, indicating improvements in lexical coverage, semantic similarity, overall summary quality, and factual reliability. With MiniCPM-V-4.5-8B, the complete EKE-MMS obtained 52.22, 35.38, 47.58, 79.90, 74.14, and 94.29 on ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, Skywork, and CFSR, respectively. Ablation studies further verified each component. Removing event knowledge-guided key-sentence extraction caused clear degradation because models had to process noisy and redundant reports. For Qwen2.5-VL-7B-Instruct, ROUGE-1 decreased from 52.37 to 48.55 and CFSR from 93.29 to 92.04. Removing image input reduced semantic and factual performance; on LLaVA-v1.6-Mistral-7B, BERTScore decreased from 78.50 to 74.25, Skywork from 67.13 to 59.21, and CFSR from 93.22 to 92.48. Removing disaster event knowledge templates also weakened results; on MiniCPM-V-4.5-8B, Skywork decreased from 74.14 to 70.59 and CFSR from 94.29 to 92.88. These findings show that key-sentence extraction, visual evidence, and disaster-type-specific knowledge constraints make complementary contributions. **Conclusion** EKE-MMS effectively alleviates unstable utilization of disaster event elements under complex multi-source inputs. Event knowledge-guided key-sentence extraction improves salient fact preservation, image evidence supplements disaster information, and disaster event knowledge templates regulate event-element organization and expression. Together with global temporal fusion, these components improve factual completeness, semantic quality, and factual consistency. The consistent gains across four multimodal large language models demonstrate that EKE-MMS is adaptable to different generators and effective for natural disaster multi-document multimodal summarization.

© 中国图象图形学报版权所有

Key words: multimodal summarization; natural disaster; event knowledge; cross-document aggregation; factual consistency

论文引用格式: Shi Qian, Hu Zhenyi, Liu Maofu, Hu Huijun, Zhang Xiaokang. An Event Knowledge-Enhanced Method for Multimodal Summarization of Natural Disaster News. Journal of Image and Graphics, XXXX: XXXX-XXXX. DOI: 10.11834/jig.260497.
(石纤，扈振义，刘茂福，胡慧君，张效康. 事件知识增强的自然灾害多模态摘要生成方法. 中国图象图形学报, XXXX: XXXX-XXXX. DOI: 10.11834/jig.260497.)[DOI:10.11834/jig.260497]

0 引 言

多模态摘要(Multimodal Summarization, MMS)旨在联合建模文本、图像、音频或视频等两种及以上模态的关键信息，并生成简洁、连贯且覆盖核心语义的摘要(Jangra 等，2023)。近年来，大语言模型与多模态生成技术的发展进一步推动了MMS研究(杜鹏飞 等，2021)。在自然灾害新闻与应急信息整理等时效性较强的场景中，MMS能够从多篇图文报道中提炼灾害发生、影响及处置进展，为灾情判断和应急决策提供信息支撑。然而，自然灾害新闻通常具有篇幅较长、来源多样、跨文档关系复杂、模态信息异构及事实精度要求高等特点。面对上述复杂多源输入，通用多模态大模型对灾害事件要素的利用不稳定，具体表现为遗漏具有较高摘要价值的关键事实和生成与源文档中事实不一致的事件要素，进而影响摘要的语义聚焦程度与事实一致性。

自然灾害多文档输入中通常同时包含重叠信息、互补信息、阶段性更新及局部冲突。已有多文档摘要研究指出，同一主题下的多篇文档具有复杂的跨文档关系(Zhang 等，2024)；相关跨模态事件理解研究也表明，多媒体新闻中的噪声、冗余及弱相关内容可能干扰事件信息识别，因此需要对输入内容进行筛选(Liu 等，2025)。在复杂多源输入的摘要生成过程中，背景与重复信息会降低显著事件要素的相对突出程度，使模型在摘要生成过程中遗漏具有较高摘要价值的具体要素。如图1(a)所示，原始报道详细给出了“帐篷570顶、床560张、棉被1950套、雨衣雨鞋130套、饮用水450件、各类食品3100余箱”等物资调度信息，而InternVL3-8B(Zhu 等，2025)仅将其概括为“调运帐篷、床、棉被、饮用水等物资”。该结果保留了物资调度这一处置事件，却遗漏了反映救援规模和处置进展的定量事实，体现了事件要素利用不稳定在内容筛选与保留层面的具体表现。

除显著事件要素保留不足外，事件要素利用不稳定的另一种表现是事件要素生成偏移，即摘要中的时间、数值偏离输入证据。自然灾害摘要对发生时间、地点、受灾人数、救援力量、物资数量和影响范围等要素的准确性要求较高，一旦出现数值替换、时间错误或事件主体混淆，将直接降低摘要的可信度。已有研究表明，语言模型在处理长上下文时，对关键信息的稳定利用仍存在局限(Liu 等，2024)。如图1(b)所示，原始输入中的“1524人受灾”被InternVL3-8B(Zhu 等，2025)错误生成为“1122名群众受灾”，属于典型的数值要素生成偏移。相关研究还表明，大视觉语言模型可能生成与图像内容不一致的对象、属性或场景信息(Li 等，2023)，生成式摘要模型也可能产生与源文档不一致的实体、数值及事实关系(Zhu 等，2021)。

上述现象共同表明，在复杂多源灾害报道中，事件要素利用不稳定主要源于内容筛选和跨文档事实组织缺少明确约束。背景、弱相关及重复信息可能削弱显著事件要素，而图像语义的不确定性以及不同报道间事实时间、信息状态的差异，可能进一步引发信息误用或事件要素生成偏移。仅依赖通用多模态大模型的端到端生成，难以同时处理上述问题，因此需要在内容筛选、图文信息使用和跨文档组织阶段引入面向灾害事件要素的知识与时序约束，以减少显著事件要素遗漏和事件要素生成偏移风险。

为此，本文提出一种事件知识增强的自然灾害多模态摘要生成方法(Event Knowledge-Enhanced Multimodal Summarization for Natural Disasters, EKE-MMS)。该方法以事件知识为核心约束。首先，事件知识引导的关键句抽取模块依据事件知识规则，自适应筛选灾害事件本身、灾害衍生后果和灾害处置措施等关键内容，降低外围背景、弱相关信息及重复叙述的干扰；其次，局部图文语义校准模块根据灾害类型从知识模板库中检索相应模板，采用“文本主

© 中国图象图形学报版权所有

导、图像辅助”的方式生成局部多模态摘要，以降低视觉信息误用和无依据的事实补全风险；最后，全局时序融合模块以局部摘要中的事实发生时间或统计截止时间为主要时序依据，以新闻发布时间为辅助顺序，并结合事件基本描述区分阶段性更新、事实修正和无法由时序演化解释的事实冲突，进而完成跨文档语义聚合与事件级摘要生成。

本文的主要工作如下：

(1)提出事件知识增强的自然灾害多模态摘要生成方法EKE-MMS。该方法通过事件知识引导模型筛选具有较高摘要价值的灾害事实，并利用灾害事件知识模板约束局部图文摘要生成，以缓解复杂多源输入下事件要素利用不稳定导致的显著事件要素保留和事实一致性不足问题。

(2)构建面向自然灾害新闻事件的多文档多模态摘要数据集，覆盖新闻文本、新闻配图和事件级参考摘要，为中文自然灾害事件级多模态摘要任务提供评测基础。

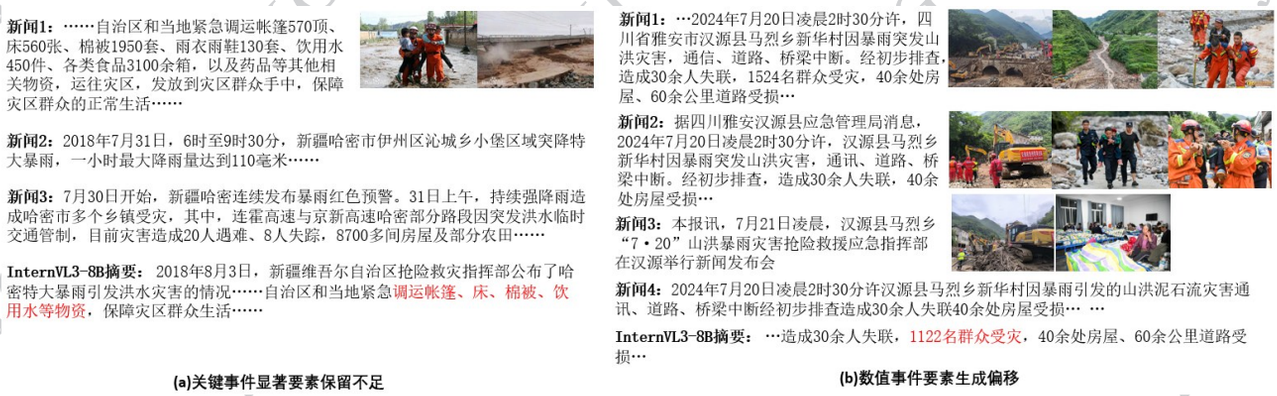


图1　通用多模态大模型在自然灾害摘要中的事件要素利用不稳定

Fig. 1　Unstable utilization of event elements by general-purpose multimodal large language models in natural disaster summarization

1　相关工作

1.1　多模态摘要生成

多模态摘要是自动摘要在多模态信息环境下的扩展，其研究已由早期的特征级融合逐步发展至跨模态语义对齐和多模态大模型驱动的生成式摘要阶段(Jangra 等，2023；Zhang，2025)。与文本摘要相比，多模态摘要还需利用图像等非文本信息补充场景、对象和状态语义。本文研究的自然灾害新闻摘要属于输入为多模态、输出为文本模态的多文档多模态摘要任务，即基于同一事件的多篇新闻文本及其配图生成事件级摘要。

早期多模态摘要研究主要关注文本与视觉特征的融合方式。Li等人(2020)提出多模态选择性编码网络，通过门控机制在编码阶段选择性引入视觉信息，以增强模型对文本关键内容的识别能力。随着视觉语言预训练的发展，跨模态表示学习进一步提升了模型对图文语义关系的建模能力(Radford 等，2021；Li 等，2023)。在此基础上，研究者开始关注视觉信息贡献不均和视觉冗余问题。Xiao 等人(2023)提出CFSum网络，通过图像预过滤模块剔除无关视觉信息，并在词级和短语级两个粒度上建模视觉信息对摘要生成的贡献。上述研究表明，视觉信息并非始终对摘要生成具有正向作用，如何筛选并利用与摘要任务相关的有效视觉线索，是多模态摘要生成中的关键问题。

除特征融合外，跨模态语义对齐与信息覆盖度也是多模态摘要的重要研究方向。He 等人(2023)提出Align-and-Attend模型，通过双对比损失增强样本内与样本间的跨模态对齐能力。Zhang 等人(2023)提出CISum框架，将图像转换为视觉描述，并通过跨模态交互学习提升生成摘要对多模态语义的覆盖程度。Yu 等人(2026)通过双时相特征增强与多关系差异耦合，加强变化区域及其语义差异的跨模态建模。近年来，多模态大模型的发展进一步推动了生成式多模态摘要研究。Patil 等人(2024)提出REFINESUMM框架，利用多模态大模型构建并自我精炼图文摘要数据；Tan 等人(2025)面向科学论文场景提出Uni-SciSum，以统一框架处理文本、图像、

© 中国图象图形学报版权所有

视频和音频等异构信息。总体而言，现有研究主要围绕模态融合、语义对齐与生成能力展开。尽管已有工作尝试面向科学论文等特定领域建模，但对领域事件场景中的显式事件知识约束以及关键事件要素完整保留的关注仍显不足。

1.2 领域事件多模态摘要

多模态摘要的研究场景已由通用图文场景扩展至医疗、科学文档和危机响应等垂直领域。与开放域多模态摘要任务相比，领域多模态摘要除需融合文本、图像等多源信息外，还需利用领域知识并施加事实约束。例如，临床文档与放射学报告摘要通常结合医学图像、检查文本及解剖学相关知识，以提升生成结果的医学专业性和事实可靠性（Ghosh 等，2024；Hu 等，2023）。这说明，领域知识能够为垂直场景下的摘要内容选择和语义约束提供重要依据。

在垂直领域任务中，针对具有动态演化特征的事件场景，领域事件多模态摘要还需围绕事件要素及其演化关系组织多源信息。危机响应与灾害信息处理的相关研究通常从新闻报道、社交媒体和应急报告中提取关键事实，并据此生成反映事件演化过程的摘要或时间线（McCreadie 和 Buntain，2023）。多模态新闻事件研究则表明，结合图文信息、上下文或外部知识有助于增强事件要素识别和事件关系建模能力（Lin 等，2024）。自然灾害新闻摘要属于典型的领域事件多模态摘要任务，但现有研究对复杂多源输入下关键事件显著要素保留及跨文档事实一致性的关注仍相对不足。

1.3 知识增强与事实一致性

事实一致性要求摘要中的实体、时间、地点、数量和事件关系等要素与源文档保持一致。生成式摘要模型容易产生与输入不符或无法由输入验证的内容（Maynez 等，2020），因此相关研究通过一致性评估与事实约束提高摘要忠实性（Kryscinski 等，2020）。在自然灾害场景中，时间、地点、受灾人数、影响范围和救援进展等事件要素具有明确现实指向，相关要素发生偏移将直接降低摘要可信度。

为缓解事实错误，Cao 等人（2018）利用开放信息抽取和依存句法分析获得事实描述，以提升摘要对源文档的忠实性；Zhu 等人（2021）将源文档中的事实关系表示为知识图谱，并通过图注意力机制融入摘要生成过程。相关综述也指出，结构化知识、检索知识和任务先验能够提升文本生成的信息充分性与可控性（Yu 等，2022）。在具体领域的多模态生成任务中，Fu 等人（2025）通过知识增强与特征对齐提升胸片报告生成效果，表明领域知识与跨模态特征建模能够辅助生成模型更充分地利用任务相关信息。在自然灾害领域，Liu 等人（2026）将知识图谱引入自然灾害事件抽取，通过结构化领域知识增强灾害事件及其要素的建模，为自然灾害信息的结构化处理提供了参考。在多模态场景中，事实一致性还表现为文本事实与视觉证据之间的不匹配。Li 等人（2023）指出，在多模态场景中，大视觉语言模型还可能生成缺乏视觉证据支持的对象或属性，产生视觉幻觉。

总体而言，结构化事件归纳能够显式表示时间、地点和数值等事件要素，便于事实检索与校验，但受预定义模式限制，难以充分覆盖不同灾害类型中的开放性文本及视觉信息，也可能损失事实限定和阶段关系。为兼顾事实约束与表达灵活性，EKE-MMS 不将固定事件槽位作为唯一中间表示，而以事件知识规则和灾害类型知识模板约束关键内容筛选与局部摘要生成，通过编号映射保持关键事实与原文的对应关系，并利用多模态大模型完成开放性图文语义理解和自然语言组织。

2 本文方法

2.1 整体框架

EKE-MMS 以事件知识为核心约束，面向同一自然灾害事件的多篇新闻文本及其关联图像生成事件级摘要。针对复杂多源输入下模型对灾害事件要素利用不稳定所导致的关键事件显著要素保留不足和事件要素生成偏移问题，该方法围绕灾害发生、灾害后果和灾害处置等事件要素组织内容筛选与摘要生成过程，以提升摘要的事实完整性和事实一致性。整体架构如图 2 所示。

在形式化表示上，单个灾害事件 E 可表示为：

$$E = (D, e_{\text{type}}, \boldsymbol{D}_E) \tag{1}$$

式中，D 表示事件的结构化描述，e_{type} 表示灾害类型标签，$\boldsymbol{D}_E = \{(T_i, \boldsymbol{I}_i, t_i)_{i=1}^{N}\}$ 表示该事件下的新闻集合；N 表示新闻数量，i 表示新闻序号，T_i 表示第 i 篇新闻的文本；$\boldsymbol{I}_i = \{\boldsymbol{I}_{i1}, \cdots, \boldsymbol{I}_{im}\}$ 表示其关联的图片集合，m 表示该篇新闻关联的图片数量；t_i 表示新闻

© 中国图象图形学报版权所有

发布时间。目标是基于事件E生成最终的事件级摘要S_E:

$$S_E = f_\theta(E) \#(2)$$

式中,f_θ表示在事件知识增强机制下的摘要生成过程,$\boldsymbol{\theta}$表示该过程中生成模型的参数集合。

EKE-MMS采用事件知识软约束下的多阶段生成方式,由事件知识引导的关键句抽取和事件知识约束的事件要素生成两个模块组成,并调用预先构建的关键句抽取知识规则$\boldsymbol{K}_{\text{ext}}$和灾害事件知识模板库$\boldsymbol{P}_{\text{temp}}$。与将多源信息完全归纳为固定事件槽位不同,该方法利用上述知识约束内容选择与组织,同时保留自然语言对开放性灾害事实及其上下文关系的表达能力。其中,前者用于筛选具有较高摘要价值的灾害事实,降低外围背景、弱相关信息和重复叙述的干扰;后者根据事件类别选择相应模板,用于约束局部图文语义校准中的事件要素保留和图文信息使用。

事件知识约束的事件要素生成模块包括局部图文语义校准和全局时序融合。局部图文语义校准采用"文本主导、图像辅助"的方式生成局部多模态摘要,以自然语言保留核心事件要素、开放性图文信息及其语义关系;全局时序融合综合局部摘要中的事实发生时间与信息状态、新闻发布时间、事件基本描述和全局融合提示词,在事实时间优先、发布时间辅助的原则下,区分并处理阶段性更新、事实修正和真正冲突,最终生成事件级摘要。因此,事件知识主要负责约束内容筛选与组织,多模态大模型则完成开放性图文语义理解和自然语言表达。

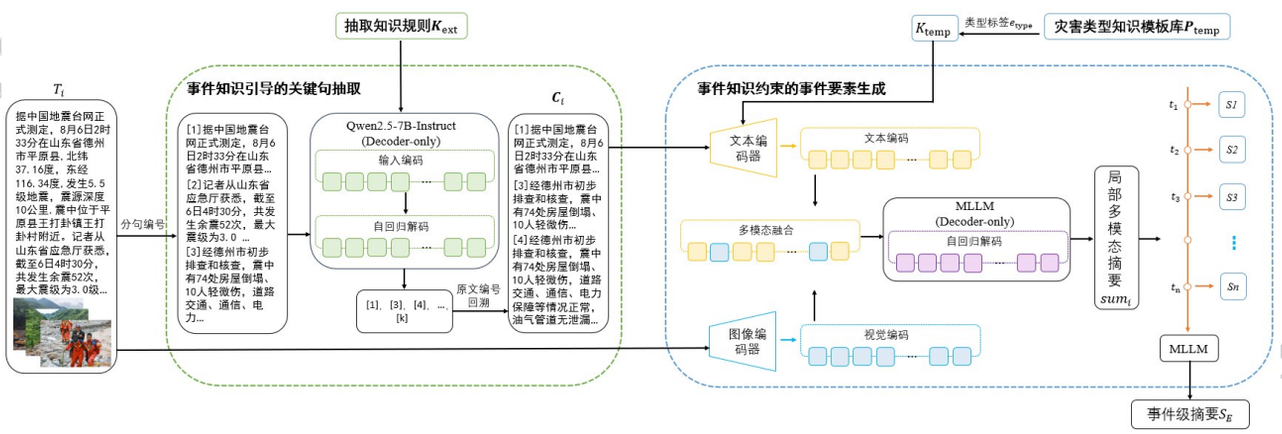


图2 EKE-MMS方法框架

Fig. 2 Framework of the EKE-MMS Method

2.2 事件知识与数据集构建

为支持EKE-MMS中的知识约束及实验评价,本文分别构建事件知识和自然灾害多文档多模态摘要数据集。事件知识用于关键内容筛选与局部图文语义校准,数据集则为各方法提供统一的测试与评价基础。

2.2.1 事件知识构建

自然灾害新闻通常围绕相对稳定的事件要素展开,事件级摘要需要概括灾害发生、灾害影响及处置过程。为提升模型对灾害事件要素的稳定利用能力,本文预先构建两类事件知识:用于关键内容筛选的关键句抽取知识规则$\boldsymbol{K}_{\text{ext}}$,以及用于局部图文语义校准的灾害事件知识模板库$\boldsymbol{P}_{\text{temp}}$。两类知识在人员伤亡、灾害影响和处置措施等内容上存在一定交叉,但其功能定位不同。

关键句抽取知识规则$\boldsymbol{K}_{\text{ext}}$采用由灾害事件本身、灾害衍生后果和灾害处置措施构成的三维事件结构。其中,灾害事件本身包括发生时间、发生地点、灾害类型、强度指标、影响范围和诱发原因;灾害衍生后果包括人员伤亡、财产损失、基础设施破坏及民众生活影响;灾害处置措施包括应急响应、救援力量、物资调配、医疗救助和灾后重建。该规则适用于不同灾害类型,用于限定关键句筛选范围,提高关键事件显著要素在复杂输入中的信息密度。

针对不同灾害类型在事实属性上的差异,本文进一步构建灾害事件知识模板库$\boldsymbol{P}_{\text{temp}}$。该模板库覆盖地震、洪水、火灾、台风、暴雨、泥石流、干旱、寒潮和冰雹等17类自然灾害,其分类设计参考EM-DAT

© 中国图象图形学报版权所有

国际灾害数据库及我国应急管理部门的相关分类标准。每类模板由通用事件要素、类别专属事件要素和事实使用约束组成。其中，核心事件要素包括人员伤亡、财产损失、基础设施破坏、民众生活与生态影响，以及应急响应、救援行动、物资调配和灾后安置重建等信息；专属事件要素用于描述不同灾害类型的特有属性。例如，地震事件包含震级、震中经纬度和震源深度等要素，台风事件包含台风名称、生成与登陆时间、登陆地点、中心位置、移动方向与速度、最大风力、最低气压和风圈半径等要素。图3为事件知识与提示模板示意图。

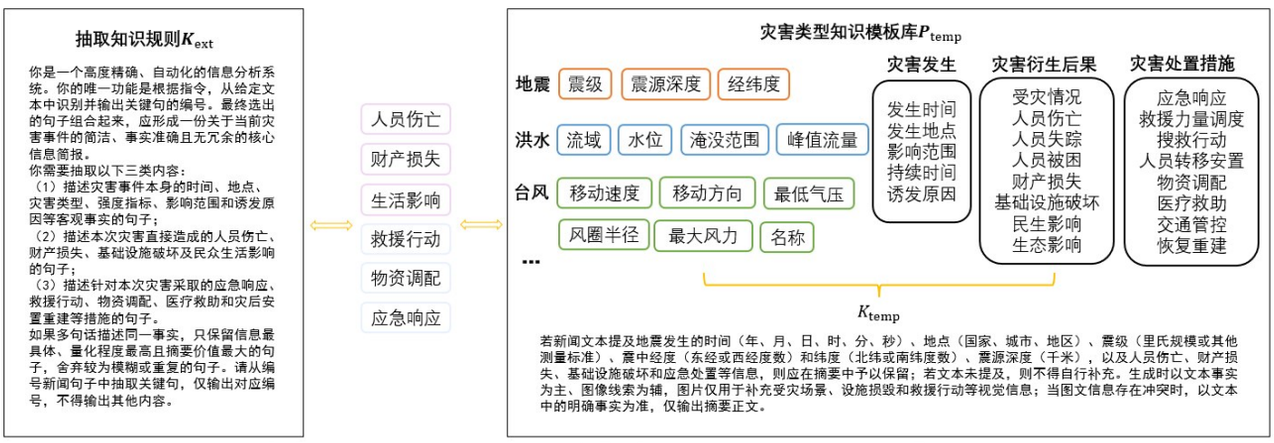


图3　事件知识与提示模板示意

Fig. 3　Event Knowledge and Prompt Template

在摘要生成过程中，系统根据当前事件的灾害类型标签从$\boldsymbol{P}_{temp}$中选择对应的K_{temp}，并将其用于局部图文语义校准。$\boldsymbol{K}_{ext}$用于提高显著事件要素的识别与保留能力，K_{temp}用于约束局部跨模态生成过程中的事件要素表达。

2.2.2　自然灾害多文档多模态摘要数据集构建

为支持中文自然灾害事件级多文档多模态摘要方法的评价，本文构建了一个面向自然灾害新闻场景的数据集。该数据集以自然灾害事件为基本组织单元，每个事件由多篇新闻文本、多张关联图片及一个事件级参考摘要构成。数据构建过程主要包括事件筛选、新闻采集、图文清洗和参考摘要整理四个步骤。

首先，在事件筛选阶段，本文以百度百科和国家减灾网中的自然灾害事件为候选来源，根据事件名称、灾害类型、发生时间和发生地点确定事件边界，剔除信息缺失、事件边界不清或重复记录的样本。其次，在新闻采集阶段，本文以筛选后的灾害事件名称为检索入口，从新华网、人民网、央视新闻等主流媒体采集相关新闻报道，并保留新闻标题、正文、发布时间和原文配图等信息。每篇新闻及其配图保持原始对应关系。

在数据清洗阶段，本文分别对新闻文本和图像数据进行处理。对于新闻文本，主要依据来源链接、标题规范化结果和正文相似度进行去重，剔除重复报道、正文缺失或与目标事件关联较弱的样本；对于图像数据，剔除加载失败、重复、清晰度较低或与事件内容明显无关的图片。经过清洗后，每个事件样本保留与事件主题相关的多篇新闻及其对应图片，从而形成事件级多模态输入。参考摘要来自对应灾害事件的百度百科词条简介部分。

最终，本文构建的数据集共包含1091个自然灾害事件、4560篇新闻文本和4509幅图片，平均每个事件约包含4篇新闻和4张图片。数据覆盖17类自然灾害，包括冰雹、台风、地震、寒潮、山洪、干旱、暴雨、暴风雪和泥石流等。由于本文不涉及监督训练或模型参数微调，数据集中的1091个事件均用于测试评估；所有对比方法和消融方法均在相同事件集合上进行生成与评价。

该数据集具有以下特点。第一，数据以灾害事件为基本单元，每个事件包含多篇新闻报道和多张相关图片，能够体现灾害新闻场景中的跨文档、多模态信息组织特征。第二，数据保留新闻发布时间，可用于建模灾害事件从发生、影响扩散到救援处置的

© 中国图象图形学报版权所有

时序演化过程。第三，灾害事件摘要对时间、地点、受灾范围、伤亡人数、救援力量和物资调度等事实信息具有较高精确性要求，因此该数据集能够较好地检验模型在长文本理解、图文对齐和事实一致性方面的表现。

2.3 事件知识引导的关键句抽取

自然灾害新闻通常包含外围背景、一般性知识、重复叙述以及与当前灾害关联较弱的延展信息。直接输入新闻原文可能降低伤亡情况、物资数量和救援进展等关键事件显著要素的信息密度，增加关键事实遗漏风险。为此，EKE-MMS在局部多模态摘要生成之前设置事件知识引导的关键句抽取模块，利用关键句抽取知识规则$\boldsymbol{K}_{\text{ext}}$，从单篇新闻文本$T_i$中筛选与当前灾害直接相关且具有摘要价值的关键内容集合$\boldsymbol{C}_i$：

$$\boldsymbol{C}_i = M_{noise}(T_i, \boldsymbol{K}_{\text{ext}}) \#(3)$$

式中，$M_{noise}(\cdot)$表示事件知识引导的关键句抽取过程。

与采用固定压缩率或预设保留句数的方法不同，本文采用内容自适应的关键句抽取策略。由于不同新闻在篇幅、事实密度和冗余程度方面存在差异，统一压缩率容易造成事实丰富文本的过度压缩，或背景冗长文本的抽取不足。因此，该模块不限定关键句数量，而是在$\boldsymbol{K}_{\text{ext}}$的约束下，根据有效事件事实的实际分布自适应确定保留规模。对于事实密度较高且冗余内容较少的新闻，模型保留相对充分的事件要素；对于背景铺垫、弱相关内容或重复信息较多的新闻，模型则形成更加紧凑的关键内容集合。对于重复描述同一事实的句子，规则要求保留信息最具体、量化程度最高且摘要价值最大的表述。

具体而言，系统首先对T_i进行分句，并为各句子分配唯一编号。模型在提示语约束下仅输出满足筛选条件的句子编号，而不直接生成或改写新闻内容。随后，系统根据输出编号回溯原文中的对应句子，并按照原始顺序组成关键内容集合$\boldsymbol{C}_i$。由于抽取结果直接来源于原始报道，该策略能够减少模型二次改写可能引入的事实偏移和无依据生成，并增强关键事实的可追溯性。

经过该模块处理后，外围背景、弱相关内容、不重要信息和重复叙述得到压缩，具有较高摘要价值的灾害事件要素在输入中的信息密度得到提升，从而降低关键事实遗漏风险，并为后续局部图文语义校准提供更加聚焦、紧凑且可靠的文本基础。

2.4 事件知识约束的事件要素生成

经过事件知识引导的关键句抽取后，关键内容集合$\boldsymbol{C}_i$已保留单篇新闻中与当前灾害相关且具有较高摘要价值的主要事实。然而，事件要素的稳定利用仍面临两个层面的挑战：在单篇新闻内部，文本事实与图像线索之间可能存在语义差异，模型若不能合理区分二者的证据作用，容易遗漏关键事实或产生无依据补全；在多篇新闻之间，发布时间、信息来源和报道重点的差异可能导致事件要素重复、阶段性更新或相互冲突。为此，本文通过局部图文语义校准和全局时序融合两个阶段，对跨模态证据利用和跨文档事实组织过程进行约束。

2.4.1 局部图文语义校准

局部图文语义校准旨在事件知识约束下融合单篇新闻的文本事实与视觉线索，生成事实可靠的局部多模态摘要。自然灾害新闻除包含时间、地点和灾情数值等边界明确的事件要素外，还包含事实限定、设施损毁、民生影响和救援行动等开放性信息，固定事件槽位可能难以充分保留相关内容及其语义关系。因此，该阶段采用自由文本生成局部摘要作为中间表示，并利用灾害事件知识模板对其内容范围和事实使用进行约束。对于同一灾害事件，系统根据事件级灾害类型标签e_{type}，从灾害事件知识模板库$\boldsymbol{P}_{\text{temp}}$中检索对应模板：

$$K_{\text{temp}} = Retrieve\left(\boldsymbol{P}_{\text{temp}}, e_{\text{type}}\right) \#(4)$$

同一事件下的多篇新闻共享该知识模板。随后，将关键内容集合$\boldsymbol{C}_i$、新闻图片集合$\boldsymbol{I}_i$和灾害事件知识模板K_{temp}共同输入多模态大模型，生成局部多模态摘要sum_i：

$$sum_i = M_{align}(\boldsymbol{C}_i, \boldsymbol{I}_i, K_{\text{temp}}) \#(5)$$

式中，$M_{align}(\cdot)$表示局部图文语义校准过程。K_{temp}作为软结构约束，引导模型关注与当前灾害类别相匹配的核心事件要素和专属事件要素，规范事件要素的组织与表达，同时允许保留模板未能完全枚举的开放性灾害信息。

在图文融合过程中，本文采用“文本主导、图像辅助”的校准策略。模型识别图片中的受灾场景、设施损毁、救援行动、人员转移和环境状态等视觉线索，并将其作为文本事实的补充。对于时间、地点、人员伤亡、灾情数值、救援力量和物资数量等需要精确表达的事件要素，以文本中的明确事实为主要依

© 中国图象图形学报版权所有

据；图像主要用于补充受灾现场、损毁状态和救援行为等场景信息，不用于推断视觉内容无法直接确认的时间、地点和数值事实。当图文信息不一致或视觉证据不足时，模型遵循文本事实优先和不得无依据补全的约束，不引入缺乏输入证据支持的信息。

通过上述过程，局部图文语义校准在保留开放性图文语义的同时，对摘要的内容范围和事实使用进行约束，有助于保留显著事件要素，减少视觉信息误用和事件要素生成偏移，提高局部摘要的事实可靠性。

2.4.2　全局时序融合

局部图文语义校准以单篇新闻为处理单位，尚未解决多篇报道之间的事实重复、阶段性更新和表述冲突问题。因此，在获得各篇新闻的局部多模态摘要后，本文进一步结合新闻发布时间 t_i 和事件基本描述 D，执行全局时序融合：

$$S_E = M_{fuse}(\{(sum_i, t_i)\}_{i=1}^{N}, D) \#(6)$$

式中，$M_{fuse}(\cdot)$ 表示全局时序融合过程，S_E 表示最终生成的事件级摘要。局部摘要 sum_i 已保留事实发生时间、统计截止时间及“截止”、“初步统计”和“最新通报”等信息状态；新闻发布时间 t_i 用于组织报道顺序，事件基本描述 D 作为全局事实组织和不一致信息校准的辅助依据。

首先按新闻发布时间升序排列局部摘要，并对同类事件要素进行时序对齐。新闻发布时间仅用于确定报道顺序；当其与事实发生时间或统计截止时间不一致时，以后者为主要判断依据，较晚报道中的早期事实不视为最新进展。事实时间缺失时，结合新闻发布时间和信息状态辅助判断；依据不足时，不将差异直接判定为阶段性更新。

模型依据事实对象、事实时间、统计口径和信息状态，将跨文档差异划分为阶段性更新、事实修正和真正冲突：若同一事件要素在不同事实时间下发生变化，则判定为阶段性更新，突出较新的明确状态并按需保留发展过程；若相同事实对象、事实时间和统计口径下的后续报道包含明确的修正或核实标记，则判定为事实修正，并采用修正后的信息；若不同报道针对相同事实对象、事实时间和统计口径给出互斥取值，且无更新或修正依据，则判定为真正冲突。对于真正冲突，若事件基本描述 D 包含与当前事实时间和统计口径一致的明确依据，则采用与其一致的事实；否则，优先保留信息状态、事实时间和统计口径更明确的表述。若仍无法确定唯一事实，则省略冲突取值，仅保留多篇报道共同支持的内容。对于重复且一致的事实，模型进行合并表达。

该机制用于区分和处理阶段性更新、事实修正和真正冲突，减少过时、重复及互斥信息对事件级摘要的影响，有助于提高摘要的时序连贯性与事实一致性。

3　实　验

3.1　实验设置

事件知识引导的关键句抽取模块中，本文采用纯语言大模型 Qwen2.5-7B-Instruct（Yang 等，2025）执行关键句抽取任务。为控制不同实验中的生成参数差异，所有生成模型均采用统一的推理配置：温度设置为 0，使用贪婪解码策略，最大生成长度设置为 500 token。

本文从基于词面覆盖的 ROUGE（Lin，2004）指标、基于语义表示的 BERTScore（Zhang 等，2019）指标、人类偏好奖励模型以及事实一致性四个方面对摘要结果进行综合评价。ROUGE-N 定义为：

$$V_{\mathrm{ROUGE-N}} = \frac{\sum_{S \in Reference} \sum_{gram_n \in S} Count_{\mathrm{match}}(gram_n)}{\sum_{S \in Reference} \sum_{gram_n \in S} Count(gram_n)} \#(7)$$

式中，$Reference$ 表示参考摘要集合；$gram_n$ 是 n 元词组。$Count_{\mathrm{match}}(gram_n)$ 为生成摘要与参考摘要中相同 n-gram 的匹配计数。ROUGE-L 计算公式如下：

$$F_{\mathrm{LCS}} = \frac{(1+\beta^2) R_{\mathrm{LCS}} P_{\mathrm{LCS}}}{R_{\mathrm{LCS}} + \beta^2 P_{\mathrm{LCS}}} \#(8)$$

式中，$R_{\mathrm{LCS}} = (LCS(X,Y))/m$；$P_{\mathrm{LCS}} = (LCS(X,Y))/n$；$LCS(\cdot)$ 是指两个字符串中长度最长的公共子序列，用于度量预测文本和参考答案之间的相似度；X 指预测文本；Y 指参考文本；β 是用于控制召回率（R_{LCS}）和精确度（P_{LCS}）的权重。BERTScore 基于向量语义匹配，记 C 与 R 分别为生成摘要与参考摘要的词向量集合，则精确率 P、召回率 R 与调和平均 $F1$ 为：$P = \frac{1}{|C|}\sum_{x \in C} \max_{y \in R} \cos(x,y)$，$R = \frac{1}{|R|}\sum_{y \in C} \max_{x \in R} \cos(x,y)$，$F_1 = \frac{2PR}{P+R}$，$\cos(x,y)$ 表示余弦相似度。奖励模型评分采用 Skywork-Reward-V2-Llama-3.2-3B（Liu 等，

© 中国图象图形学报版权所有

2025)对生成摘要给出原始得分raw_score,并通过Sigmoid归一化为[0,1]区间的综合质量得分:

$$sigmoid_score = \frac{1}{1 + e^{-raw_score}} \#(9)$$

*sigmoid_score*越接近1,表示摘要整体质量越高。事实一致性评估将原始新闻文本及结构化描述作为事件级证据池$P_E = \mathrm{D} \cup \{T_i\}_{i=1}^{N}$,采用Qwen2.5-7B-Instruct对待评估摘要进行关键事实单元抽取,再判断每个事实单元是否被证据池支持。CFSR(Claim Fact Support Rate,关键事实支持率)表示生成摘要中的关键事实单元能够被原始新闻证据支持的比例,作为事实一致性指标。

3.2 实验结果

3.2.1 对比实验

为验证EKE-MMS在不同多模态生成器上的适用性,本文选取Qwen2.5-VL-7B-Instruct(Bai等,2025)、LLaVA-v1.6-Mistral-7B(Liu等,2023)、InternVL3-8B(Zhu等,2025)和MiniCPM-V-4.5-8B(Yu等,2025)进行实验,分别比较各模型直接生成与引入EKE-MMS后的性能变化。

为比较不同摘要生成方法的效果,本文参考已有研究中的多阶段摘要(Zhang等,2022)、抽取后生成(Chen和Bansal,2018)和外部引导摘要(Dou等,2021)策略,在统一的MiniCPM-V-4.5-8B生成器上构建Multi-stage Summarization、Extract-then-Generate和Template-Guided Generation三类单一策略控制基线,三类基线用于考察单一通用策略能否取得与完整方法相近的效果。

Multi-stage Summarization:先生成单篇新闻的局部摘要,再将多个局部摘要融合为事件级摘要。

Extract-then-Generate:先从新闻中抽取关键句,再直接生成事件级摘要。

Template-Guided Generation:将原始新闻、图片和灾害事件知识模板共同输入模型以约束摘要生成。

采用统一生成器能够控制基座模型能力差异,使实验结果主要反映不同摘要生成流程的影响。此外,本文引入最大边际相关性算法MMR(Maximal Marginal Relevance)作为传统抽取式基准,该方法通过平衡句子相关性与冗余性完成多文档信息抽取。

表1展示了不同方法在自然灾害多文档多模态摘要数据集上的实验结果。总体而言,EKE-MMS在词面覆盖、语义相似性、偏好评价和事实一致性等维度上取得了较优表现,表明事件知识引导的关键句抽取、局部图文语义校准和全局时序融合有助于模型稳定识别、保留和组织灾害事件要素,从而改善摘要的语义质量和事实一致性。

传统抽取式方法MMR的R-1、R-2和R-L分别为36.10、19.93和30.48,表明该方法能够保留部分与参考摘要重合的词汇和短语,但其整体词面覆盖相对有限。在Skywork偏好评价中,MMR较采用MiniCPM-V-4.5-8B作为生成器的EKE-MMS低27.06个百分点。MMR主要依据句子相关性和冗余性选择原文内容,无法利用新闻图片、灾害事件知识模板和发布时间信息,因此在跨模态信息补充、事件演化组织和多源事实融合方面存在局限。

在统一生成器的控制对比中,Multi-stage Summarization、Extract-then-Generate和Template-Guided Generation相较于直接生成均取得了不同程度的性能提升,说明局部摘要生成、关键内容抽取和模板约束均有助于改善摘要质量。其中,Extract-then-Generate的R-1、BS和CFSR分别达到51.89、79.00和92.87,在三类控制基线中表现较优,表明预先筛选关键内容有助于提高输入文本的信息密度。然而,该方法未利用图像证据进行局部语义校准,也未依据发布时间组织多篇报道中的阶段性信息。Multi-stage Summarization能够降低一次性处理多篇原始新闻的输入压力,但其局部摘要生成阶段缺少关键句抽取知识规则和灾害事件知识模板的约束;Template-Guided Generation能够约束事件要素表达,但未在生成前过滤原始长文本中的外围背景、弱相关内容和重复信息。

相比之下,EKE-MMS依次执行事件知识引导的关键句抽取、局部图文语义校准和全局时序融合,分别从输入内容筛选、跨模态证据利用和跨文档事实组织三个层面提升模型对灾害事件要素的稳定利用能力。在不同生成器上,引入EKE-MMS后,四种多模态大模型的综合性能均有所改善。其中,LLaVA-v1.6-Mistral-7B的BS提高5.62个百分点,Skywork得分提高16.92个百分点,CFSR提高接近3个百分点,其R-1由43.78提高至48.85,提升5.07个百分点。上述结果表明,EKE-MMS对摘要的语义质量、偏好评价和事实一致性具有积极作用。

在ROUGE指标方面,Qwen2.5-VL-7B-Instruct

© 中国图象图形学报版权所有

的R-1、R-2和R-L分别提高1.19、1.59和1.07个百分点，MiniCPM-V-4.5-8B的三项ROUGE指标提升均超过2个百分点，说明EKE-MMS在多数生成器上能够提高生成摘要与参考摘要之间的词面覆盖程度。

InternVL3-8B引入EKE-MMS后，不同类型评价指标呈现出较为明显的变化差异。其中，R-1、R-2和R-L降幅分别为4.00、4.86和3.03个百分点；与此相对，BS由73.80提高至78.23，Skywork由51.89提高至66.02，CFSR由88.42提高至92.63。该结果表明，EKE-MMS虽然未提高InternVL3-8B与参考摘要之间的表层词汇重合度，但在语义相似度、整体质量和事实一致性方面均取得了较为明显的改善。为进一步分析ROUGE指标下降的原因，本文对摘要的生成长度进行了统计。InternVL3-8B直接生成摘要的平均长度为332.12字符，总体标准差为168.93引入EKE-MMS后，平均长度降至262.01字符，总体标准差降至134.21。与直接生成相比，EKE-MMS生成的摘要平均缩短70.10字符，降幅为21.11%，且输出长度分布更加集中。参考摘要的平均长度为274.68字符，直接生成结果比参考摘要平均长57.44字符，而EKE-MMS生成结果与参考摘要平均长度仅相差12.66字符。该统计结果说明，EKE-MMS有效压缩了背景铺垫、重复描述及弱相关内容，使模型生成的摘要更加紧凑。由于ROUGE主要依据词汇及短语重合进行评价，直接生成的较长摘要能够覆盖更多参考摘要中的表层表达；EKE-MMS输出长度明显缩短，减少部分词汇匹配机会，从而导致ROUGE下降。结合BS、Skywork和CFSR的同步提升可以看出，该指标变化并非摘要质量的整体下降，而是表现为表层词汇覆盖减少的同时，语义质量、人类偏好和事实可靠性得到改善。

EKE-MMS相较于传统抽取式方法、单一摘要生成范式和多模态大模型的直接生成方式，整体上取得了更优的实验结果。事件知识引导的关键句抽取有助于提高显著事件要素在输入中的信息密度，局部图文语义校准能够补充视觉线索并规范事件要素表达，全局时序融合则用于组织多篇报道中的重复、更新和冲突信息，三者共同改善自然灾害事件级摘要的语义质量和事实一致性。

表1　不同模型方法实验结果对比

Table 1　Comparison of Experimental Results for Different Models

/%

方法	R-1	R-2	R-L	BS	Skywork	CFSR
MMR	36.1	19.93	30.48	74.58	47.08	98.20
Multi-stage Summarization	51.02	35.01	46.55	78.92	69.28	91.66
Extract-then-Generate	51.89	34.44	45.29	79.00	72.84	92.87
Template-Guided Generation	50.36	33.10	44.52	78.52	68.79	92.03
Qwen2.5-VL-7B-Instruct	51.18	32.72	47.00	77.60	65.74	91.78
EKE-MMS (Qwen2.5-VL-7B-Instruct)	**52.37**	**34.31**	**48.07**	**79.39**	**73.62**	**93.29**
LLaVA-v1.6-Mistral-7B	43.78	31.24	41.12	72.88	50.21	90.26
EKE-MMS (LLaVA-v1.6-Mistral-7B)	**48.85**	**32.27**	**44.07**	**78.50**	**67.13**	**93.22**
InternVL3-8B	46.43	28.85	42.36	73.80	51.89	88.42
EKE-MMS (InternVL3-8B)	42.43	23.99	39.33	**78.23**	**66.02**	**92.63**
MiniCPM-V-4.5-8B	50.28	33.05	44.31	78.49	66.66	91.41
EKE-MMS (MiniCPM-V-4.5-8B)	**52.22**	**35.38**	**47.58**	**79.90**	**74.14**	**94.29**

3.2.2　消融实验

为验证EKE-MMS中各模块的有效性，本文在固定生成参数设置下进行消融实验。结果如表2所示。具体设置如下：

EKE-MMS-w/o Noise：在完整方法基础上移除事件知识引导的关键句抽取模块，其他不变。

© 中国图象图形学报版权所有

EKE-MMS-w/o Image：在完整方法基础上移除局部图文语义校准阶段中的图像模态输入，其他不变。

EKE-MMS-w/o Template：在完整方法基础上移除灾害事件知识模板约束，其他不变。

总体来看，完整EKE-MMS在四种基座模型上均取得了较优的综合结果，说明事件知识引导的关键句抽取、图像模态输入和灾害事件知识模板均对摘要生成具有积极作用。以MiniCPM-V-4.5-8B为例，完整方法在R-1、R-2、R-L、BS、Skywork和CFSR上分别取得52.22、35.38、47.58、79.90、74.14和94.29，均高于对应消融配置，表明三项关键设计因素均对完整方法的综合性能作出贡献。

移除事件知识引导的关键句抽取后，模型需要直接处理包含外围背景、弱相关内容和重复叙述的原始新闻，多项指标随之下降。在Qwen2.5-VL-7B-Instruct上，R-1由52.37下降至48.55，R-2由34.31下降至30.23，R-L由48.07下降至45.72，CFSR由93.29下降至92.04；在MiniCPM-V-4.5-8B上，R-L由47.58下降至45.29，Skywork由74.14下降至67.29，CFSR由94.29下降至92.52。上述结果表明，依据关键句抽取知识规则筛选新闻内容，有助于降低噪声和冗余信息的干扰，提高显著事件要素在输入中的信息密度，进而降低关键事实遗漏风险。

移除局部图文语义校准阶段的图像输入后，多数模型的BS、Skywork和CFSR出现下降。以LLaVA-v1.6-Mistral-7B为例，BS由78.50下降至74.25，Skywork由67.13下降至59.21，CFSR由93.22下降至92.48；在MiniCPM-V-4.5-8B上，R-L由47.58下降至44.74，Skywork由74.14下降至67.41，CFSR由94.29下降至93.21。该结果说明，图像能够为受灾场景、设施损毁、救援行动和环境状态等信息提供补充线索，有助于提升模型对跨模态事件要素的利用能力以及局部摘要的语义完整性。需要说明的是，该消融配置仅移除图像输入，仍保留文本关键内容和灾害事件知识模板。

移除灾害事件知识模板后，模型无法获得与当前灾害类别对应的核心事件要素和专属事件要素，多项指标均出现不同程度下降。在Qwen2.5-VL-7B-Instruct上，Skywork由73.62下降至67.22，CFSR由93.29下降至92.55；在LLaVA-v1.6-Mistral-7B上，R-1由48.85下降至45.26，Skywork由67.13下降至60.10，CFSR由93.22下降至92.19；在MiniCPM-V-4.5-8B上，Skywork由74.14下降至70.59，CFSR由94.29下降至92.88。上述结果表明，灾害事件知识模板能够为模型提供面向灾害类别的事件要素约束，有助于规范摘要内容组织并减少事件要素表达偏移。

综上，三组消融实验分别验证了EKE-MMS各组成部分的作用：事件知识引导的关键句抽取从输入筛选层面提高显著事件要素的保留能力；图像输入从跨模态证据利用层面补充灾害现场信息；灾害事件知识模板从生成约束层面规范事件要素的组织与表达。三者共同提升模型对灾害事件要素的稳定利用能力，使完整EKE-MMS在语义质量和事实一致性方面取得更稳定的综合表现。

3.2.3　执行效率与案例分析

考虑到自然灾害信息汇总对时效性的要求，本文比较了多模态大模型直接生成事件级摘要与采用EKE-MMS生成事件级摘要的端到端推理耗时。在本文实验设置下，由于EKE-MMS增加了事件知识引导的关键句抽取、事件知识约束的事件要素生成过程，其推理耗时高于直接生成，但每个事件的平均额外耗时低于1分钟。

本文研究面向多篇新闻文本和多张关联图片的事件级信息汇总，主要服务于灾害信息整理与辅助决策，而非秒级预警触发。因此，上述额外耗时在本文任务场景下具有一定可接受性。结合前述实验结果，EKE-MMS在增加一定推理成本的同时，改善了摘要的内容覆盖、语义质量和事实一致性，体现了生成质量与执行效率之间的合理权衡。

为进一步分析EKE-MMS的实际生成效果，本文从语料库中选取四个代表性案例进行定性比较，重点考察显著事件要素保留、视觉线索利用以及时间和数值要素的事实一致性，结果如图4所示。

如图4(a)所示，在地震救援案例中，基线模型仅以“已调集救援力量”等概括性内容描述处置进展，未充分保留原文中的救援人员、车辆和装备信息。相比之下，EKE-MMS生成的摘要保留了“60名指战员、30辆消防救援车、5300件套救援器材装备”等具体事实。该案例表明，事件知识引导的关键句抽取能够提高具有较高摘要价值的处置要素在输入中的信息密度，有助于减少显著事实遗漏。

如图4(b)所示，在包含灾后现场图片的地震案

© 中国图象图形学报版权所有

表2　消融实验结果

Table 2　Ablation Experiment Results

/%

模型	方法	R-1	R-2	R-L	BS	Skywork	CFSR
Qwen2.5-VL-7B-Instruct	EKE-MMS	52.37	34.31	48.07	79.39	73.62	93.29
	EKE-MMS-w/o Noise	48.55	30.23	45.72	78.23	71.94	92.04
	EKE-MMS-w/o Image	48.51	30.55	44.33	78.68	68.92	92.69
	EKE-MMS-w/o Template	49.19	30.7	44.56	78.55	67.22	92.55
LLaVA-v1.6-Mistral-7B	EKE-MMS	48.85	32.27	44.07	78.50	67.13	93.22
	EKE-MMS-w/o Noise	46.12	31.88	42.20	75.11	60.89	91.30
	EKE-MMS-w/o Image	45.61	31.87	42.22	74.25	59.21	92.48
	EKE-MMS-w/o Template	45.26	31.55	42.65	75.02	60.10	92.19
InternVL3-8B	EKE-MMS	42.43	23.99	39.33	78.23	66.02	92.63
	EKE-MMS-w/o Noise	41.10	22.4	39.47	74.56	51.89	89.27
	EKE-MMS-w/o Image	42.41	23.42	39.35	78.57	59.45	91.02
	EKE-MMS-w/o Template	43.11	24.7	39.9	78.28	58.23	91.06
MiniCPM-V-4.5-8B	EKE-MMS	52.22	35.38	47.58	79.90	74.14	94.29
	EKE-MMS-w/o Noise	50.97	34.28	45.29	78.55	67.29	92.52
	EKE-MMS-w/o Image	50.68	33.99	44.74	79.14	67.41	93.21
	EKE-MMS-w/o Template	51.77	33.66	45.37	79.43	70.59	92.88

例中，基线模型主要概括文本中的行政响应信息，对视觉内容的利用较为有限。EKE-MMS在以文本事实为主要依据的基础上，补充了图片中呈现的“货架物品掉落”、“地面积水”和“砖墙倒塌”等现场线索。该案例说明，局部图文语义校准能够合理利用与灾害事件相关的视觉证据，补充文本对受灾场景和设施损毁状态的描述，从而提高局部摘要的信息完整性。

图4(c)和图4(d)进一步展示了事件要素生成偏移现象。在图4(c)中，基线模型将原文中的“2015年”生成为“2011年”，造成时间要素不一致；在图4(d)中，基线模型将“1524人受灾”生成为“1122人受灾”，造成数值要素偏移。相比之下，EKE-MMS在上述案例中保留了与原始报道一致的时间和数值信息，表明事件知识约束下的生成流程有助于减少时间、数值等事件要素的表达偏移。

综合来看，EKE-MMS能够从输入筛选、跨模态证据利用和事实组织三个层面提高模型对灾害事件要素的稳定利用能力。其中，事件知识引导的关键句抽取有助于保留具有较高摘要价值的显著事实，局部图文语义校准能够补充受灾现场的视觉线索，事件知识约束的事件要素生成则有助于减少无依据补全及时间、数值等要素偏移，从而改善事件级摘要的语义连贯性和事实一致性。

新闻1：据青海省地震局6月17日消息，据中国地震台网正式测定，2021年6月16日16时48分（北京时间），在青海省海西州茫崖市发生5.8级地震，震源深度10公里……



新闻2：据中国地震台网测定，北京时间2021年6月16日16时48分在青海海西州茫崖市，北纬38.14度，东经93.81度，发生5.8级地震，震源深度10千米……

新闻3：……一支重型和一支轻型地震救援队160 名指战员、30辆消防救援车、5300 件套救援器材装备集结完毕，随时做好救援出动准备.…及药品等其他相关物资……

基线输出：……距冷湖镇约77公里，青海省消防救援总队已调集救援力量，随时准备救援……

EKE-MMS输出：……距冷湖镇约77公里，青海省消防救援总队已调集救援队160名指战员、30辆消防救援车、5300件套救援器材装备，随时准备救援……

（a）　案例一灾害处置关键事实保留对比

© 中国图象图形学报版权所有

新闻1：央视网消息：据中国地震台网正式测定，今天（8月6日）凌晨2时33分，在山东省德州市平原县发生5.5级地震，震源深度10千米，山东全省普遍有震感。本次地震震中位于北纬37.16度，东经116.34度……

新闻2：记者从应急管理部获悉，8月6日2时33分，山东德州市平原县，北纬37.16度，东经116.34度，发生5.5级地震，震源深度10公里……为保证列车安全，国铁济南局启动应急响应……



新闻3：新华社济南8月6日电，据中国地震台网正式测定，8月6日2时33分在山东省德州市平原县发生5.5级地震，震源深度10公里……



基线输出：2025年8月6日凌晨，山东德州市平原县发生5.5级地震，震源深度10公里……山东消防救援队伍已集结到位，做好应急准备。平原县因地震对燃气管网造成破坏，已紧急停气……

EKE-MMS输出：2025年8月6日凌晨，山东德州市平原县发生5.5级地震，震源深度10公里，震中位于北纬37.16度，东经116.34度……地震导致货架上的物品掉落，地面有积水，砖墙倒塌，部分墙体和屋顶受损，地震发生后，山东德州消防救援支队已前往震中救援……

（b） 案例二 灾后现场视觉线索利用对比

新闻1：辽宁职工报《北方周末》副主编、绥中门户网站站长刘洪刚称6月1日16时许，葫芦岛市区突降冰雹，较多车辆和户外设施受损…

新闻2：葫芦岛昨降大冰雹四喜丸子那么大葫芦岛本报讯 昨日16时许，葫芦岛市多地突降冰雹…

新闻3：据中国网络咨询台等媒体报道，昨日15时30分左右，辽宁葫芦岛港附近，突降大雨并伴随着冰雹，这让很多走在路上的行人措手不及赶紧找地方躲避…

新闻4：……2015年6月1日16时许葫芦岛市区冰雹昨日起至9月20日沈阳将进入汛期，今年首次将新民、辽中、康平、法库城市防汛工作纳入全市城市防汛范畴……



基线输出：…2011年6月1日16时许，辽宁葫芦岛市区遭遇冰雹袭击…

EKE-MMS输出：…2015年6月1日16时许，辽宁葫芦岛市区遭遇冰雹袭击…

（c） 案例三 时间要素事实一致性对比

新闻1：…受暴雨影响，马烈河突发大水，导致新华村1524人受灾。初步排查，有30余人失联，40余处房屋、60余公里道路受损…



新闻2：本报讯，7月21日凌晨，汉源县马烈乡“7·20”山洪暴雨灾害抢险救援应急指挥部在汉源举行新闻发布会



新闻3：据四川雅安汉源县应急管理局消息，2024年7月20日凌晨2时30分许……经初步排查，造成30余人失联，40余处房屋受损…



基线输出：…造成30余人失联，1122名群众受灾，40余处房屋、60余公里道路受损…

EKE-MMS输出：…造成30余人失联，1524人受灾，40余处房屋、60余公里道路受损…

（d） 案例四 数值要素事实一致性对比

((a) Key facts retention and comparison in disaster handling; (b) Utilization of Visual Cues at Post-Disaster Sites through Contrast; (c) Comparison of Time Element Consistency; (d) Comparison of Numerical Element Fact Consistency)

图4 案例分析图

Fig. 4 Case Analysis Diagram

4 结 论

针对复杂多源输入下灾害事件要素利用不稳定所导致的语义聚焦程度与事实一致性不足问题，本文构建了自然灾害事件级多模态摘要数据集，并提出事件知识增强的自然灾害多模态摘要生成方法。通过事件知识引导关键内容筛选和局部图文摘要生成，再通过时序融合组织不同报道中的阶段更新与事实差异。实验表明，相较于传统抽取式方法和单一策略控制基线取得了较优的综合表现，应用于不同多模态生成器时，摘要的语义质量和事实支持率总体得到改善。

但当前时序融合依赖报道中可辨识的事实时间、统计口径和修正提示。相关信息缺失时，阶段更新、事实修正与事实冲突仍难以可靠区分。此外，持续到达的新闻流中的摘要更新能力尚未得到验证。后续将研究跨时段事实追踪与来源证据校验，并优化多轮生成效率。

© 中国图象图形学报版权所有

参考文献(References)

Bai S, Chen K Q, Liu X J, Wang J L, Ge W B, Song S B, et al. 2025. Qwen2.5-VL technical report[EB/OL]. [2025-09-01]. https://arxiv.org/pdf/2502.13923.pdf

Cao Z Q, Wei F, Li W J and Li S J. 2018. Faithful to the original: Fact aware neural abstractive summarization//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, Louisiana, USA: AAAI Press: 4784-4791 [DOI: 10.1609/aaai.v32i1.11912]

Chen Y C and Bansal M. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics: 675-686 [DOI: 10.18653/v1/P18-1063]

Dou Z Y, Liu P F, Hayashi H, Jiang Z B and Neubig G. 2021. GSum: a general framework for guided neural abstractive summarization//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics: 4830-4842 [DOI: 10.18653/v1/2021.naacl-main.384]

Du P F, Li X Y and Gao Y L. 2021. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 32(2): 327-348 (杜鹏飞, 李小勇, 高雅丽. 2021. 多模态视觉语言表征学习研究综述. 软件学报, 32(2): 327-348) [DOI: 10.13328/j.cnki.jos.006125]

Fu J, Liu L, Fu X D, Liu L J and Peng W. 2025. Knowledge enhancement and feature alignment for chest radiology images report generation. Journal of Image and Graphics, 30(5):1528-1542 (符杰, 刘骊, 付晓东, 刘利军, 彭玮. 2025. 结合知识增强和特征对齐的胸片图像报告生成. 中国图象图形学报, 30(5):1528-1542) [DOI:10.11834/jig.240536]

Ghosh A, Tomar M S, Tiwari A, Saha S, Salve J A and Sinha S. 2024. From sights to insights: Towards summarization of multimodal clinical documents//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics: 13117-13129 [DOI: 10.18653/v1/2024.acl-long.708]

He B, Wang J, Qiu J L, Bui T, Shrivastava A and Wang Z. 2023. Align and attend: Multimodal summarization with dual contrastive losses//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14867-14878 [DOI: 10.1109/CVPR52729.2023.01428]

Hu J P, Chen Z H, Liu Y, Wan X and Chang T H. 2023. Improving radiology summarization with radiograph and anatomy prompts//Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics: 12066-12080 [DOI: 10.18653/v1/2023.findings-acl.764]

Jangra A, Mukherjee S, Jatowt A, Saha S and Hasanuzzaman M. 2023. A survey on multi-modal summarization. ACM Computing Surveys, 55(13s): 1-36 [DOI: 10.1145/3584700]

Kryscinski W, McCann B, Xiong C and Socher R. 2020. Evaluating the factual consistency of abstractive text summarization//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online: Association for Computational Linguistics: 9332-9346 [DOI: 10.18653/v1/2020.emnlp-main.750]

Li H R, Zhu J N, Zhang J J, He X D and Zong C Q. 2020. Multimodal sentence summarization via multimodal selective encoding//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics: 5655-5667 [DOI: 10.18653/v1/2020.coling-main.496]

Li J N, Li D X, Savarese S and Hoi S C H. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA: PMLR: 19730-19742

Li Y F, Du Y F, Zhou K, Wang J P, Zhao X and Wen J R. 2023. Evaluating object hallucination in large vision-language models//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics: 292-305 [DOI: 10.18653/v1/2023.emnlp-main.20]

Lin C Y. 2004. ROUGE: A package for automatic evaluation of summaries//Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics: 74-81

Lin Z H, Xie J Y and Li Q. 2024. Multi-modal news event detection with external knowledge. Information Processing & Management, 61(3): 103697 [DOI: 10.1016/j.ipm.2024.103697]

Liu C Y, Zeng L, Xiao Y, He J, Liu J, Wang C, et al. 2025. Skywork-Reward-V2: Scaling preference data curation via human-AI synergy [EB/OL]. [2025-09-01]. https://arxiv.org/pdf/2507.01352.pdf

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023. Visual instruction tuning [EB/OL]. [2025-09-01]. https://arxiv.org/pdf/2304.08485.pdf

Liu M F, Zhou B Y, Hu H J, Qiu C and Zhang X K. 2025. Cross-modal event extraction via visual event grounding and semantic relation filling. Information Processing & Management, 62(3): 104027 [DOI: 10.1016/j.ipm.2024.104027]

Liu M F, Zu Q, Shi Z C, Zhang X K, Li Y S and Zhang L F. 2026. A knowledge graph enhanced method for natural disaster event extraction[J/OL]. Geomatics and Information Science of Wuhan University: 1-21 (刘茂福, 祖琦, 史志成, 张效康, 李彦胜, 张乐飞. 2026. 知识图谱增强的自然灾害事件抽取方法[J/OL]. 武汉大学学报(信息科学版): 1-21 [2026-07-26]. https://doi.org/10.

© 中国图象图形学报版权所有

13203/j.whugis20250232

Liu N F, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, et al. 2024. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12: 157-173 [DOI: 10.1162/tacl_a_00638]

Maynez J, Narayan S, Bohnet B and McDonald R. 2020. On faithfulness and factuality in abstractive summarization//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics: 1906-1919 [DOI: 10.18653/v1/2020.acl-main.173]

McCreadie R and Buntain C. 2023. CrisisFACTS: Building and evaluating crisis timelines//Proceedings of the 20th International Conference on Information Systems for Crisis Response and Management. Omaha, NE, USA: ISCRAM Association: 320-339 [DOI: 10.59297/JVQZ9405]

Patil V, Ribeiro L F R, Liu M, Bansal M and Dreyer M. 2024. REFINESUMM: Self-refining MLLM for generating a multimodal summarization dataset//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics: 13773-13786 [DOI: 10.18653/v1/2024.acl-long.743]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 8748-8763

Tan Z S, Zhong X Y, Ji J Y, Jiang W and Chiu B. 2025. Enhancing large language models for scientific multimodal summarization with multimodal output//Proceedings of the 31st International Conference on Computational Linguistics: Industry Track. Abu Dhabi, UAE: Association for Computational Linguistics: 263-275

Xiao M, Zhu J N, Lin H T, Zhou Y and Zong C Q. 2023. CFSum Coarse-to-Fine Contribution Network for Multimodal Summarization//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics: 8538-8553 [DOI: 10.18653/v1/2023.acl-long.476]

Yang A, Yang B S, Zhang B C, Hui B Y, Zheng B, Yu B W, et al. 2025. Qwen2.5 technical report[EB/OL]. [2025-09-01]. https://arxiv.org/pdf/2412.15115.pdf

Yu J X, Fu Y J and Liu J. 2026. Bi-temporal feature enhancement-guided multi-relational difference coupling for remote sensing change captioning. Journal of Image and Graphics, 31(9): 3293-3307 (于洁潇, 付雨杰, 刘婧. 2026. 双时相特征增强引导的多关系差异耦合遥感变化字幕生成. 中国图象图形学报, 31(9): 3293-3307) [DOI:10.11834/jig.260171]

Yu T Y, Wang Z F, Wang C Y, Huang F W, Ma W S, He Z H, et al. 2025. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe[EB/OL]. [2025-10-01]. https://arxiv.org/pdf/2509.18154.pdf

Yu W H, Zhu C G, Li Z T, Hu Z T, Wang Q Y, Ji H, et al. 2022. A survey of knowledge-enhanced text generation. ACM Computing Surveys, 54(11s): 1-38 [DOI: 10.1145/3512467]

Zhang H P. 2025. Bridging multimodal and video summarization: A unified survey//Proceedings of The 5th New Frontiers in Summarization Workshop. Hybrid: Association for Computational Linguistics: 157-171 [DOI: 10.18653/v1/2025.newsum-main.11]

Zhang L T, Zhang X M, Guo Z M and Liu Z P. 2023. CISum: Learning cross-modality interaction to enhance multimodal semantic coverage for multimodal summarization//Proceedings of the 2023 SIAM International Conference on Data Mining. Minneapolis, Minnesota, USA: SIAM: 370-378 [DOI: 10.1137/1.9781611977653.ch42]

Zhang M, Lu J Y, Yang J H, Zhou J, Wan M and Zhang X J. 2024. From coarse to fine: Enhancing multi-document summarization with multi-granularity relationship-based extractor. Information Processing & Management, 61(3): 103696 [DOI: 10.1016/j.ipm.2024.103696]

Zhang T Y, Kishore V, Wu F, Weinberger K Q and Artzi Y. 2019. BERTScore: Evaluating text generation with BERT [EB/OL]. [2025-09-01]. https://arxiv.org/pdf/1904.09675.pdf

Zhang Y S, Ni A S, Mao Z M, Wu C H, Zhu C G, Deb B, et al. 2022. Summ^N: a multi-stage summarization framework for long input dialogues and documents//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics: 1592-1604 [DOI: 10.18653/v1/2022.acl-long.112]

Zhu C G, Hinthorn W, Xu R C, Zeng Q K, Zeng M, Huang X D, et al. 2021. Enhancing factual consistency of abstractive summarization//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics: 718-733 [DOI: 10.18653/v1/2021.naacl-main.58]

Zhu J G, Wang W Y, Chen Z, Liu Z Y, Ye S L, Gu L X, et al. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models[EB/OL]. [2025-09-01]. https://arxiv.org/pdf/2504.10479.pdf

作者简介

石纤,女,硕士研究生,主要研究方向为多模态摘要。E-mail: 2718841615@qq.com

刘茂福,通信作者,男,教授,主要研究方向自然语言处理。E-mail: liumaofu@wust.edu.cn

扈振义,男,硕士研究生,主要研究方向为自然语言处理。E-mail: zhenyihu@wust.edu.cn

胡慧君,女,副教授,主要研究方向为智能信息处理。E-mail: huhuijun@wust.edu.cn

张效康,男,教授,主要研究方向为遥感影像智能解译。E-

© 中国图象图形学报版权所有

mail：natezhangxk@gmail.com

© 中国图象图形学报版权所有