TripleSumm：面向视频摘要的自适应三模态融合

Sumin Kim* **Hyemin Jeong*** **MinguKang*** **YejinKim Yoori Oh†** **JoonseokLee†**首尔国立大学
{sumink, hyeminjeong, gms5560, a2000yejin, yoori0203, joonseok}@snu.ac.kr

arXiv:2603.01169v1 [cs.CV] 1 Mar 2026

摘要

视频内容的指数级增长使得有效的视频摘要成为必要，以便从长视频中高效提取关键信息。然而，现有方法难以充分理解复杂视频，主要是因为它们采用静态或模态无关的融合策略。这些方法未能考虑视频数据中固有的、随帧变化的模态显著性动态差异。为克服这些局限，我们提出**TripleSumm**，一种新颖的架构，可在帧级自适应地加权并融合视觉、文本和音频模态的贡献。此外，多模态视频摘要研究的一个重要瓶颈是缺乏全面的基准。针对这一瓶颈，我们引入**MoSu**（最常重播的多模态视频摘要），这是首个提供全部三种模态的大规模基准。大量实验表明，TripleSumm取得了最先进的性能，在包括MoSu在内的四个基准上以显著优势超越现有方法。我们的代码和数据集可在以下地址获取：https://github.com/smkim37/TripleSumm。

1 引言

随着智能移动设备与数据通信的最新进展，视频内容在 YouTube、TikTok 等各类平台上呈爆发式增长。与此同时，近期趋势表明人们的偏好已转向短视频内容，从而增加了通过摘要摘取长视频主要情节的需求。视频摘要是提取能够充分代表原始视频内容的关键片段的任务，旨在满足这一需求。

现有的视频摘要研究（Apostolidis 等人，2021b；Son 等人，2024；Kim 等人，2025a）主要致力于寻找一种将帧级视觉特征映射为重要性分数以作为摘要的模型架构。然而，人类对视频的理解本质上是一个多模态过程，它整合了视觉之外的多种线索。大多数现有架构仅关注视觉模态，因此忽略了其他模态中存在的互补信息。图1以音乐试镜为例，说明了用于理解内容的主要模态会随视频进程而变化。在（a）处，文本模态（语音）对于把握评委的评价最具信息量；而在（b）处，视听线索对于欣赏机器人的表演更为重要。在（c）处，三种模态共同发挥作用。观察到即使在同一视频中，特定模态的重要性也会显著变化，这促使我们以自适应方式利用多种模态，动态地为信息量最大的模态赋予更高权重，从而实现更有效的视频摘要。

认识到这一点后，近期研究开始利用多模态信号进行视频摘要（He 等人，2023年；Li 等人，2023年；Hua 等人，2025年；Guo 等人，2025年）。然而，在视频摘要任务中，如何以最合适的方式融合多模态信号，仍未得到充分探索。摘要的核心目标不仅仅是预测真实标签，更是区分重要性较高与较低的视频帧。为此，需要采用判别式方法，综合考虑两个关键因素，恰当地衡量特征显著性：模态内时间依赖，

*贡献相同
†通讯作者

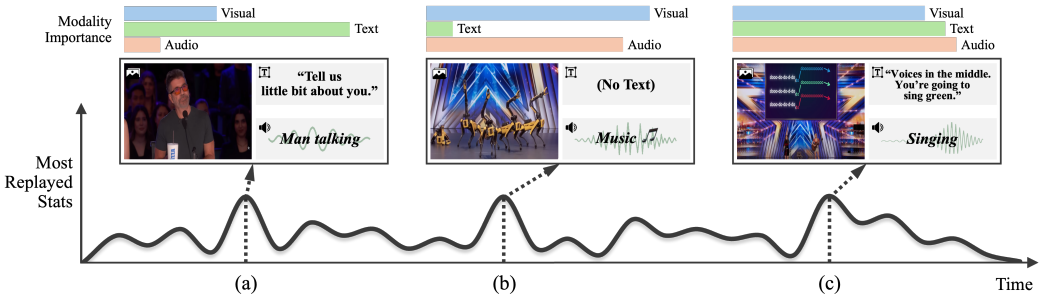


图 1：**模态动态重要性的示意图**。在 (a) 中，文本最显著；而在 (b) 中，视觉-音频占主导地位。在 (c) 中，三种模态均有显著贡献。这凸显了采用自适应模型的必要性，该模型能够逐帧动态调整各模态的显著性权重。

即比较相邻时间步上的同一模态特征；以及跨模态连贯性，即比较同一时间点上的不同模态。尽管使用了丰富的三模态输入，大多数现有架构仍采用简单或静态的融合机制（例如，标准自注意力或交叉注意力），或优先考虑单一模态，因而无法动态聚焦于每帧中信息量最大的线索。当非视觉线索具有很强的描述性时，这会导致性能受损。

本文提出了**TripleSumm**，这是一种新颖的视频摘要模型，能够灵活、自适应地利用多模态特征，并在模态缺失时保持稳健，涵盖视频的三个关键方面：视觉、文本和音频。具体而言，TripleSumm 采用以下两个新颖组件。首先，*多尺度时序模块* 采用具有不同窗口大小的分层滑动窗口结构，有效检测细微的时序变化，同时保留视频的整体叙事。其次，*跨模态融合模块* 引入融合令牌，明确学习在每个时刻应优先考虑哪种模态。

除了模型架构之外，训练数据也是多模态视频摘要的另一项瓶颈。尽管训练数据至关重要，目前仍没有公开可用的视频数据，能够以足够大的规模提供视觉-文本-音频特征和重要性分数标注。现有数据集的规模往往有限（Gygli 等人，2014年；Song 等人，2015）或模态种类有限（Narasimhan 等人，2022年；Sul 等人，2023年；Argaw 等人，2024）。为此，我们推出了一个大规模视频摘要数据集，提供全部三种关键模态的特征，即**Mo**stReplayedMultimodal Video **Su**mmarization（**MoSu**）。该数据集由 52,678 个真实场景视频组成，并汇总了每个视频至少 50,000 名观众的观看行为，可作为高度可靠的三模态视频摘要训练与评估集。

通过全面的实验，我们验证了 TripleSumm 在多个基准上大幅优于现有模型。值得注意的是，即使缺少一种或多种模态，我们的方法也能稳健地生成合理的视频摘要，并根据内容动态依赖各个模态。我们的定性分析清楚地表明，所提出的方法能够以逐帧方式自适应地融合多模态信息。

我们的主要贡献可概括如下：

- 我们提出了 **TripleSumm**架构，能够在帧级自适应地融合视觉、文本和音频模态。借助其时序与模态模块，该架构能够动态调整各模态的重要性，并有效捕捉视频的微观与宏观层面信息。
- 我们推出了 **MoSu**，这是首个为每段视频提供三模态特征的大规模视频摘要数据集，为多模态视频摘要奠定了可靠的基础。
- 我们证明，TripleSumm 在包括 MoSu 在内的四个主要视频摘要基准上达到了最先进水平，同时保持了较高的参数效率。

2 相关工作

视频摘要。 视频摘要的早期研究主要侧重于利用循环神经网络或长短期记忆网络对视觉特征中的时间依赖关系进行建模（Zhang等人，2016年；Zhao 等人，2017；2018年；2021b）。较新的模型采用自注意力机制来捕捉全局长程依赖关系（Fajtl等人，2018年；Jung et al.,

2019；Zhu 等人，2021年；Wang 等人，2020年；Apostolidis 等人，2021b；Jiang 和 Mu，2022年；Terbouche 等人，2023年；Son 等人，2024）。此外，一些研究采用基于图的模型（Park et al.，2020；Zhao et al.，2021a；Zhu et al.，2022；Zhang等人，2024）、生成对抗网络（Mahasseni et al.，2017；Yuan et al.，2019；Apostolidis et al.，2019；2021a）或扩散模型（Yu et al.，2024；Shang et al.，2025年；Kim et al.，2025a）。

尽管架构取得了这些进展，但仅依赖视觉特征会限制对视频的全面理解。随着研究转向多模态方法，转录文本或图像字幕等文本数据已被纳入其中（Narasimhan 等人，2021年；Huang 等人，2021年；Narasimhan 等人，2022年；Ghauri 等人，2021年；Li 等人，2023年；Argaw 等人，2024年；Qiu 等人，2024年），近期的 CFSum（Guo 等人，2025年）则率先利用视觉、文本和音频信号进行摘要。然而，这些方法中的大多数仍采用简单或静态机制（例如标准自注意力或固定的交叉注意力模块）来融合多模态特征，缺乏充分的自适应能力。因此，它们要么一视同仁地处理各种模态，要么优先考虑视觉信息，仅将文本视为补充特征（Narasimhan 等人，2021年；Li 等人，2023年）。由于这种固有偏见以及缺乏动态加权，这些模型无法充分概括非视觉线索（文本或音频）主导内容的视频。研究人员还利用大语言模型捕捉多模态上下文，发挥其强大的推理能力（Lin 等人，2024年；Hua 等人，2025年；Lee 等人，2025年；Kang 等人，2025年）。与此同时，音频仍是视频摘要中利用不足的模态，近期只有少数研究例外（Badamdorj 等人，2021年；Liu 等人，2022年；Zhao 等人，2022年；Xie 等人，2022年）。我们的工作旨在平衡利用视觉、文本和音频这三种主要模态，从而实现全面的视频摘要。

视频摘要数据集。 SumMe（Gygli 等人，2014年）和 TVSum（Song 等人，2015）得到了广泛使用，但存在两个关键局限：规模极小（分别仅包含 25 个和 50 个视频），且完全基于视觉线索进行单模态标注，因此不适用于多模态研究。大规模基准 Mr. HiSum（Sul 等人，2023年）解决了规模问题，但其有意排除音乐等音频丰富类别，使模态差距仍未得到解决。较新的数据集开始关注纳入文本信息。然而，它们往往引入新的挑战，例如对教学视频存在领域偏差（例如，（Argaw 等人，2024））或直播流存在领域偏差（例如，（He 等人，2023年）），或者真实摘要过于稀疏（例如，（Qiu 等人，2024））。为满足对大规模、三模态且多样化数据集的需求，我们提出了 MoSu 数据集，详见第 4节。

3 TripleSumm：所提出的方法

我们提出了视频摘要模型 **TripleSumm**，如图 2 所示。我们首先介绍多模态特征提取（第 3.1 节），随后说明如何利用两个核心组件融合时序信息和跨模态信息： 多尺度时序模块 和 跨模态融合模块 （第 3.2 节）。最后，我们介绍推理过程，说明如何预测帧级重要性分数，并据此生成最终的摘要视频（第 3.3 节）。

3.1 输入表示

原始视频由多个采样率不同的模态流组成。我们对这些模态流进行预处理，生成一组同步特征序列，并将每个序列重采样至 N 个时间步。（详情见第 5.1 节。）在每个时间步 $i = 1, \ldots, N$，我们都有多模态信号。在三模态设置下，我们假设包含视觉 $\mathcal{V} = \{\mathbf{V}_1, \ldots, \mathbf{V}_N\}$、文本 $\mathcal{T} = \{\mathbf{T}_1, \ldots, \mathbf{T}_N\}$ 和音频 $\mathcal{A} = \{\mathbf{A}_1, \ldots, \mathbf{A}_N\}$，其中 $\mathbf{V}_i$、$\mathbf{T}_i$ 和 $\mathbf{A}_i$ 分别是模态特定原始数据；例如，$\mathbf{V}_i$ 是一幅二维RGB图像。我们将基于这种三模态设置介绍所提出的方法，但该方法本身也可以扩展到任意一组多模态特征。

对于每个模态序列中的每个项目，我们使用各模态专属的预训练编码器提取特征表示，记为 $\mathbf{X}^{\{v,t,a\}}$:

$$\mathbf{X}^v = \phi(\mathbf{V}) \in \mathbb{R}^{N \times D_v}, \quad \mathbf{X}^t = \psi(\mathbf{T}) \in \mathbb{R}^{N \times D_t}, \quad \mathbf{X}^a = \omega(\mathbf{A}) \in \mathbb{R}^{N \times D_a}, \tag{1}$$

其中 ϕ 是图像编码器（例如，（Szegedy 等人，2015；Radford 等人，2021年）），ψ 是文本编码器（例如，（Liu 等人，2019）），而 ω 是音频编码器（例如，（Gong 等人，2021年））。每个预训练

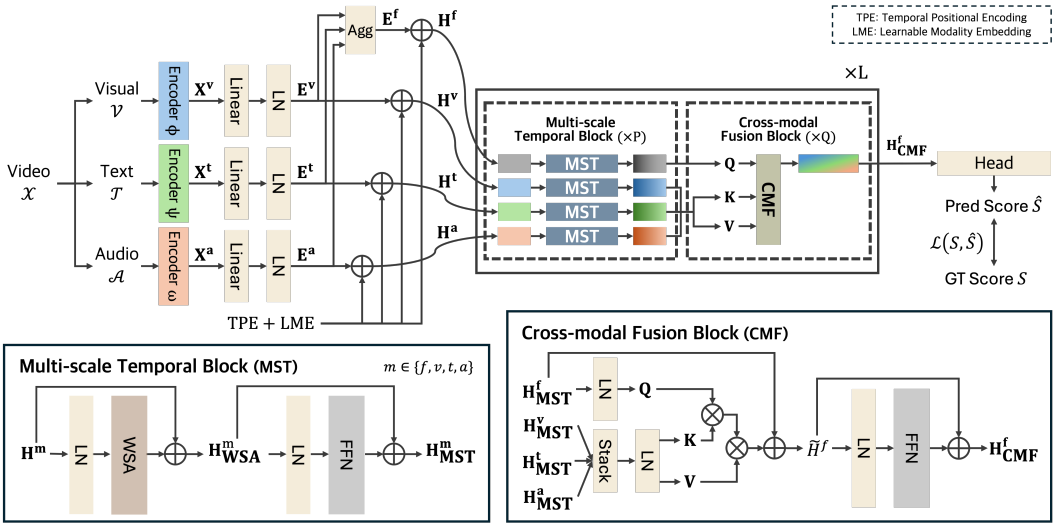


图 2：**TripleSumm 的整体架构**。首先对视觉、文本和音频特征进行编码和线性投影，然后将其聚合为融合令牌，再通过**多尺度时序模块**（MST，左下）进行精炼，并在**跨模态融合模块**（CMF，右下）中融合。融合后的表示经由预测头处理，生成帧级重要性分数。

编码器都会生成一个与自身大小相同的嵌入，分别记为 D_v、 D_t和 D_a ，对应视觉、文本和音频。

这些模态特定特征位于不同的潜空间中。为了在后续模块中有效地融合它们，我们将其投影到大小为 D的共同嵌入空间。为此，我们对每种模态特定特征$\mathbf{X}^{\{v,t,a\}}$ 应用线性投影（Linear）和层归一化（LN）（Ba 等人，2016年），以生成各模态嵌入$\mathbf{E}^{\{v,t,a\}} \in \mathbb{R}^{N\times D}$:

$$\mathbf{E}^{\{v,t,a\}} = \mathrm{LN}(\mathrm{Linear}(\mathbf{X}^{\{v,t,a\}})). \tag{2}$$

我们将时间步 i处的嵌入（即$\mathbf{E}^{\{v,t,a\}}$的第 i行）记为$\mathbf{e}_i^{\{v,t,a\}}$ 。在这些各模态嵌入的基础上，我们引入另一种跨模态嵌入，记为$\mathbf{E}^f = \{\mathbf{e}_1^f, \ldots\ldots, \mathbf{e}_N^f\} \in \mathbb{R}^{N\times D}$，其中$\mathbf{e}_i^f$ = 为 $\mathrm{Agg}(\mathbf{e}_i^v,\ \mathbf{e}_i^t,\ \mathbf{e}_i^a)$。聚合函数（Agg）可以是确定性函数（例如，平均），也可以是可学习模型（例如，多层感知机）。引入该嵌入的主要动机是避免传统跨模态融合方法可能带来的偏差，例如，将一种模态不对称地用作查询来关注其他模态。我们的融合嵌入$\mathbf{E}^f$ 充当整合所有模态的锚点，促进各模态平等参与。

最后，我们通过加入时间位置编码（TPE）来构建最终的令牌嵌入$\mathbf{tpe}_i \in \mathbb{R}^D$ (Vaswani等人，2017)以及可学习模态嵌入（LME）$\mathbf{lme}^{\{f,v,t,a\}} \in \mathbb{R}^D$，以区分每个时间步以及模态的来源:

$$\mathbf{h}_i^{\{f,v,t,a\}} = \mathbf{e}_i^{\{f,v,t,a\}} + \mathbf{tpe}_i + \mathbf{lme}^{\{f,v,t,a\}}, \quad \text{for } i = 1, \ldots, N. \tag{3}$$

以矩阵形式，在所有时间步上堆叠得到的最终输入矩阵记为$\mathbf{H}^{\{f,v,t,a\}} \in \mathbb{R}^{N\times D}$。

3.2 时序与跨模态细化

我们所提出架构的核心是一种分层的“细化与融合”策略，旨在从输入序列（$\mathbf{H}^{\{f,v,t,a\}}$）中学习整合表征。这通过交错两个关键组件来实现：多尺度时序模块（MST）用于各模态内部的时序细化，跨模态融合模块（CMF）用于模态间的信息交换。我们堆叠 L个交错层，其中每一层由 P个时序模块和 Q 个跨模态模块组成。

多尺度时序模块。 对于输入序列 $\mathbf{H}^{\{f,v,t,a\}}$，MST 通过采用窗口自注意力（WSA）（Beltagy 等人，2020；Liu 等人，2021年），限制查询令牌$\mathbf{q}_i \in \mathbb{R}^D$ 在时间步 i 的自注意力范围，使其仅关注以 w 为中心、大小为 i的局部窗口内的键和值。这些局部键和值记为

by $\mathbf{K}_i \in \mathbb{R}^{w\times d_k}$ 和 $\mathbf{V}_i \in \mathbb{R}^{w\times d_k}$，通过收集窗口内的键向量 $\mathbf{k}_j$ 和值向量$\mathbf{v}_j$ 构建，其中 $j = |i-j| \le w$，$w \in \{n \in \mathbb{N} \mid 1 \le n \le N$ 为 n 且为奇数$\}$，d_k 为键向量的维度。形式上，MST 的整体过程如下：

$$\begin{aligned}\mathbf{h}^m_{\text{WSA}} &= \text{Attn}(\text{LN}(\mathbf{h}^m_i), \text{LN}(\mathbf{h}^m_{\{j:|i-j|\le w\}}), \text{LN}(\mathbf{h}^m_{\{j:|i-j|\le w\}})) + \mathbf{h}^m, \\ \mathbf{h}^m_{\text{MST}} &= \text{FFN}(\text{LN}(\mathbf{h}^m_{\text{WSA}})) + \mathbf{h}^m_{\text{WSA}},\end{aligned} \tag{4}$$

其中 $m \in \{f, v, t, a\}$，$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\mathbf{QK}^\top/\sqrt{d_k}\right)\mathbf{V}$ 是标准注意力机制（Vaswani 等人，2017），FFN 是前馈网络。“多尺度”特性通过在每一层调整窗口大小 w 来实现，使模型能够捕捉从局部到全局尺度的语义信息。具体而言，初始层采用较小的窗口大小，以捕捉相邻帧之间细粒度的局部依赖。随后，较深层逐步扩大窗口，以捕捉跨帧的长程依赖。WSA 以共享参数应用于每种模态，使模型能够捕捉通用的时间模式，同时提高参数使用效率。关于参数共享的消融实验详见附录 C.5。此外，WSA 不仅对模型捕捉输入视频中的多粒度上下文至关重要，还有助于提高计算效率，将标准自注意力的复杂度从 $O(N^2)$ 降低至每层 $O(w \cdot N)$。

跨模态融合模块。 而 MST 专注于优化各模态内部的时间模式，CMF 则旨在对不同模态之间的交互进行建模，并在每个时间步独立执行。为了使模型能够在每个时间步选出信息量最大的模态，同时避免偏向任何特定模态，该模块采用交叉注意力机制，在每个时间步 i 将融合词元$\mathbf{h}^f_i$ 作为唯一查询，并将相应的各模态专属词元$\mathbf{h}^{\{v,t,a\}}_i$ 作为键和值。随后，查询令牌会关注这组三个上下文令牌，从而能够在特定时刻对最相关模态的信息进行加权和聚合。

For因此，给定前一个时间模块在时间步 i输出的特征$\mathbf{h}^{\{f,v,t,a\}}$ MST i，它会执行 s

$$\tilde{\mathbf{h}}^f_i = \text{Attn}(\text{LN}(\mathbf{h}^f_{\text{MST},i}),\ \text{LN}(\mathbf{h}^{\{v,t,a\}}_{\text{MST},i}),\ \text{LN}(\mathbf{h}^{\{v,t,a\}}_{\text{MST},i})) + \mathbf{h}^f_{\text{MST},i}, \tag{5}$$

$$\mathbf{h}^f_{\text{CMF},i} = \text{FFN}(\text{LN}(\tilde{\mathbf{h}}^f_i)) + \tilde{\mathbf{h}}^f_i, \tag{6}$$

其中，$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\mathbf{QK}^\top/\sqrt{d_k}\right)\mathbf{V}$**</style}，为键向量的维度。层归一化 LN() 在每个模态内分别进行。更新后的融合嵌入$\mathbf{H}^f$** CMF选择性地融合了从所有三种模态中提取的上下文信息。

TripleSumm 的一个关键设计原则是将时间融合与跨模态融合分开。时间上下文仅由 MST 建模，使其能够专注于在每个对应的时间步整合模态信息。另一方面，跨模态融合完全由 CMF 在每个时间步独立执行。这种设计不仅有助于模型显式学习视频中的两种正交模式（时间模式和多模态模式），还可通过暂时合并批次维度与时间维度，实现高效的并行处理，从而并行处理所有词元。

3.3 模型训练和推理

最后，预测头通过线性投影将经过优化的融合特征 $\mathbf{H}^f$ CMF 映射为每个视频帧的重要性分数$\hat{}$ $S \in [0, 1]$，该分数表示该视频帧被纳入最终视频摘要的概率。有关预测头的更多细节，请参阅附录 A.1。

$\hat{}$ 模型通过最小化预测分数向量 $S = \{\hat{s}_1, \ldots, \hat{s}_N\}$ 与真实分数 $S = \{s_1, \ldots, s_N\}$之间的平方 L2 损失进行训练。形式上，我们的损失定义为

$$\mathcal{L}(S, \hat{S}) = \|S - \hat{S}\|_2^2 = \|S - \text{Linear}(\mathbf{H}^f_{\text{CMF}})\|_2^2. \tag{7}$$

按照视频摘要领域的惯例（Otani 等人，2019；Son 等人，2024；Kim 等人，2025a），最终摘要通过在固定长度预算下，选择一组时间连贯的镜头来获得，使预测的帧级重要性分数最大化。详细的分割与选择流程见附录 A.2。

4 MoSu 数据集

视频摘要长期以来一直受困于数据质量和规模不足。例如，过去十年间使用最广泛的公开数据集是 SumMe（Gygli 等人，2014年）和 TVSum（Song 等人，2015），两者分别仅包含 25 个和 50 个视频。显然，这些数据集规模极小，导致严重的过拟合问题；但由于标注成本高昂，直到最近才有其他数据集出现。Mr. HiSum（Sul 等人，2023年）利用 YouTube 上的“最常重播”统计数据作为逐帧重要性的可靠代理指标，反映观众的集体参与度（详情见附录 B.2）。然而，尽管具有这一优势，Mr. HiSum 仍不适用于三模态视频摘要，因为它只提供视觉特征，不包含配套的文本或音频模态；此外，它还移除了以音频为中心的视频类别，进一步限制了其在多模态融合中的适用性。

为了可靠地训练和评估三模态视频摘要模型，我们推出了一个名为 **MoSu(Mo**st Replayed Multimodal Video **Su**mmarization) 的全新大型数据集，该数据集从 YouTube-8M（Abu-El-Haija 等人，2016年）整理而成，包含来自知识图谱实体的 3,862 个类别及其多标签注释。对于注释，我们采用与 Mr. HiSum（Sul 等人，2023年）类似的流程，收集“最常重播”统计数据。我们通过筛选符合以下条件的视频来构建该数据集：1）提供英文转录和音轨（原始提供，或由 YouTube 的字幕翻译功能自动生成），以满足三模态条件；2）观看次数超过 50,000，以获得 Most Replayed 统计数据；3）时长至少为 120 秒，以确保内容足够长，从而能够进行有意义的摘要。

数据集统计信息。MoSu 包含 52,678 个视频，总时长近 4,000 小时，涵盖 3,406 个类别，主题范围广泛。MoSu 中视频的平均时长为 272.3 秒，介于 120 至 501 秒之间。与表 1 中的其他视频摘要数据集相比，MoSu 是首个提供全部三种模态（视觉、文本和音频）的大规模数据集，而此前的数据集仅包含视觉模态（例如 Mr. HiSum）或双模态（例如 MMSum）。

数据集	模态			视频时长		类别
	视觉	文本	音频			
SumMe	✓			25	1.1	–
TVSum	✓			50	3.5	10
Mr.HiSum	✓			31,892	1,788.0	3,509
TL;DW?	✓	✓		12,160	628.5	185
BLiSS	✓	✓		13,303	1,109.0	–
LfVS-P	✓	✓		250K	55,416.6	6,700
MMSum	✓	✓		5,100	1,229.9	170
MoSu	✓	✓	✓	52,678	3,983.7	3,406

表 1：**各类视频摘要数据集的统计信息**。总时长以小时为单位。

主题类别。鉴于这一发现视频摘要的复杂度会因视频主题而显著不同（Sul 等人，2023年），我们也旨在分析 MoSu 中不同主题类别下的视频摘要质量。由于从 YouTube-8M 继承而来的原始 3,406 个类别划分过细，我们将其聚类为 10 个不同的主题组。具体而言，我们首先对维基百科中 3,862 条实体描述的 SBERT 嵌入（Reimers 与 Gurevych，2019）进行 K 均值聚类，以构建主题空间。我们还使用拼接后的实体标签将每个视频映射到同一嵌入空间，并选择最近的聚类质心作为其伪类别。所得主题组的详情见附录 B.3；各聚类的摘要性能见附录 G.2，其中展示了这些主题组如何影响摘要的难易程度。

5 实验

5.1 实验设置

数据集。我们的主要评估在第4节介绍的 MoSu 数据集上进行。为进一步验证所提出模型的泛化能力，我们还在三个广泛使用的外部数据集上评估其性能：大规模的 Mr. HiSum（Sul 等人，2023年），以及两个由人工标注的 SumMe（Gygli 等人，2014年）和 TVSum（Song 等人，2015）。对于 Mr. HiSum 和 MoSu，我们遵循原始数据划分。对于 SumMe 和 TVSum，我们在 5 折交叉验证框架下采用两种不同协议进行评估：不包含测试集的传统训练/验证（TV）划分（Li 等人，2023年；Son 等人，2024；Lee 等人，2025年），以及训练/验证/测试（TVT）划分（Kim 等人，2025a），以纠正 TV 设置中固有的过拟合问题。有关协议的详细信息见附录B.5。

方法	模态 V	T	A	τ ↑	ρ ↑	mAP50 ↑	mAP15 ↑	参数量 ↓	GFLOPs ↓
VASNet (Fajtl等人，2018年)	✓			0.151	0.219	64.49	31.05	8.13M	1.99G
PGL-SUM (Apostolidis 等人，2021b)	✓			0.151	0.218	64.97	30.63	5.31M	1.21G
CSTA (Son 等人，2024)	✓			0.291	0.398	71.77	40.65	10.56M	11.37G
A2Summ (He 等人，2023年)	✓	✓		0.181	0.257	66.48	35.70	2.48M	1.35G
SSPVS （Li 等人，2023年）	✓	✓		0.190	0.271	66.10	32.65	112.81M	43.64G
Joint-VA （Badamdorj 等人，2021年）	✓		✓	0.190	0.272	65.68	32.25	4.21M	1.63G
UMT (Liu 等人,2022年)	✓		✓	0.239	0.334	68.83	36.73	4.66M	1.39G
CFSum (Guo 等人,2025年)	✓	✓	✓	0.277	0.374	70.97	38.20	19.83M	8.52G
TripleSumm（本文方法）	✓	✓	✓	**0.351**	**0.472**	**74.72**	**44.42**	**1.37M**	**0.97G**

Table 2: **在 MoSu 上的比较**。 最佳和次佳结果分别以**粗体** 和下划线标示 .

（a）Mr. HiSum

方法	TVT τ	TVT ρ	TVT mAP50	TVT mAP15
VASNet	0.069	0.102	58.69	25.28
PGL-SUM	0.097	0.141	61.60	27.45
CSTA	0.128	0.185	63.38	30.42
A2Summ	0.121	0.172	63.20	32.34
SSPVS	0。078	0。113	59.48	26.35
Joint-VA	0.161	0.231	65.88	35.23
UMT	0.178	0.253	66.81	35.65
我们的 (视觉)	0.187	(	67.16	35.57
本文 （完整）	**0.258**	(	**70.72**	**40.88**

(b) SumMe

方法	TVT τ	TVT ρ	TV τ	TV ρ
VASNet	0.089	0.099	0.160	0.170
PGL-SUM	0.104	0.116	0.192	0.213
CSTA	0.133	0.148	0.246	0.274
A2Summ	0.088	0.096	0.108	0.129
SSPVS	0.142	0.157	0.192	0.257
Joint-VA	0.117	0.129	0.230	0.256
UMT	0.148	0.165	0.241	0.268
本文 （完整）	0.162			
本文 （MoSu）	**0.172**			

(c) TVSum

方法	TVT τ	TVT ρ	TV τ	TV ρ
VASNet	0.153	0.205	0.160	0.170
PGL-SUM	0.141	0.186	0.157	0.206
CSTA	0.168	0.221	0.194	0.255
A2Summ	0.157	0.206	0.137	0.165
SSPVS	0.171	0.226	0.181	0.238
Joint-VA	0.142	0.188	0.166	0.220
UMT	0.144	0.189	0.179	0.235

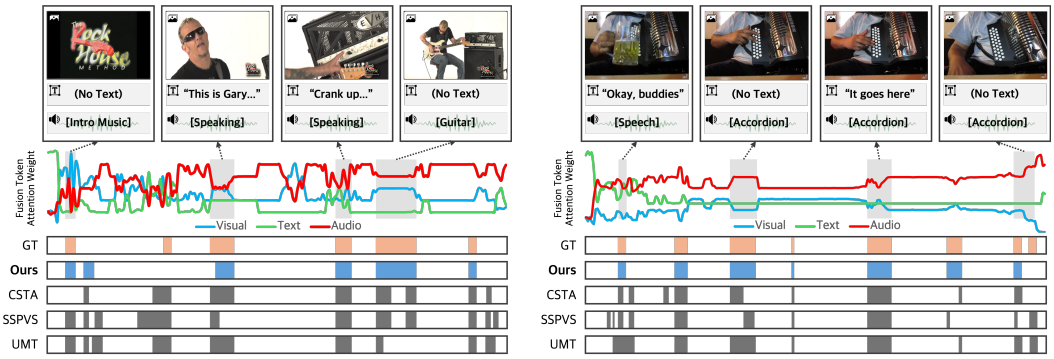
表 3： **（a）Mr. HiSum、（b）SumMe 和（c）TVSum 上的性能比较**。 Ours(Full)表示在各自数据集上训练的模型。Ours(Visual) 仅使用视觉特征进行训练，Ours(MoSu) 则在 MoSu 上预训练，并在各个目标数据集上微调。

数据预处理。 如第3.1节所述，我们的模型需要长度相同且在时间上对齐的特征序列。我们对 MoSu 中的每种模态进行如下预处理。对于视觉模态，我们以 1 fps 的速率采样视频帧，并使用预训练的 CLIP（Radford 等人，2021年）对其进行编码。对于文本模态，我们从 YouTube 提取带时间戳的转录文本，并通过获取句子级别的 [CLS] 标记来获得 RoBERTa（Liu 等人，2019）特征；随后将该特征广播至其持续时间所覆盖的所有帧；没有转录文本的帧则填充默认嵌入向量。对于音频，我们使用中心窗口方法以 1 秒为间隔提取特征：对于每一秒t，从时间区间 $[t-5, t+5]$ 中裁剪出一段 10 秒的片段，并使用预训练的音频频谱Transformer（Gong 等人，2021年）将其编码为特征向量。

在 Mr. HiSum、SumMe 和 TVSum 上，为了进行公平比较，我们采用它们官方提供的视觉特征。由于这些数据集不提供文本或音频流，我们使用图像描述模型 Qwen2.5-VL-7B-Instruct（Bai 等人，2025年）生成视频帧级文本描述，并直接从视频中提取原始音频。对于每种模态，我们随后遵循各数据集原先定义的预处理流程。详细的预处理步骤见附录 B.4。

实现细节。 我们使用 $L = 2$ 个交错层，以及 $P = 2$ 个多尺度时序块（因此总共有4个时序块）以及 $Q = 2$ 个跨模态融合块。为了便于从局部到全局捕获特征，我们逐步增大注意力窗口大小（w），从5直至整个序列长度 N。更多细节见附录 A.1。

评估指标。遵循Otani 等人(2019)，我们主要报告基于排名的相关性指标，即 Kendall 的 τ (Kendall, 1945年）和 Spearman 的 ρ (Zwillinger 和 Kokoska，1999)，用于评估视频帧级重要性预测，与近期研究的做法一致（Son 等人，2024；Terbouche 等人，2023年；Kim 等人，2025a）。在 MoSu 和 Mr. HiSum 数据集上，鉴于使用 Most Replayed 统计数据作为真实值，我们进一步采用平均精度均值（mAP）评估精彩片段检测性能，遵循Sul 等人(2023年）的方法。按照这一流程，每段视频被划分为 5 秒片段，并根据预测的重要性进行排序。随后，将排名前 50%（mAP50）和前 15%（mAP15）的片段作为预测精彩片段，与真实值进行评估。



（a）吉他演示：随着内容变化，注意力在视觉、文本和音频线索之间转移。

(b) 视觉-文本信息较少的手风琴演奏视频，其中音频线索占主导。

图 3：**MoSu 上的定性示例**。中间的图展示了 Fusion Token 的注意力权重，说明我们的模型如何动态估计各模态的显著性，从而即使部分模态缺失，也能保持较高的摘要准确率。

5.2 定量评估

在 MoSu 上的评估。如表2所示，TripleSumm 在 MoSu 上取得了显著的最先进性能，在所有指标上均大幅领先于所有单模态和多模态基线。除了准确率提升外，TripleSumm 的效率也很高，参数量仅为 1.37M，显著少于 CSTA（10.56M）和 UMT（4.66M）等强基线。结果表明，TripleSumm 在保持架构显著轻量且高效的同时，实现了卓越的预测性能。

在其他数据集上的评估。表3(a) 显示，我们的方法在 Mr. HiSum 的大多数指标上都达到了最先进水平。值得注意的是，即使仅使用视觉特征，我们的方法也优于所有基线 Ours (Visual)，这验证了所提出的多尺度时序模块在单一模态中捕捉显著时间特征的有效性。此外，我们的完整模型显著优于仅使用视觉信息的版本，凸显了利用多模态信息的优势。
表 3(b–c) 表明，在 SumMe 和 TVSum 上进行训练和评估时，我们的完整模型大幅优于所有基线；这两个数据集均为人工标注数据集。这一点尤其值得注意，因为真实标注完全基于视觉信息。模型在使用全部三种模态时取得的性能表明，非视觉模态为识别关键视觉时刻提供了重要的上下文线索，这意味着我们的三模态方法能够学习模态之间的协同关系。最后，在 MoSu 上预训练并在目标数据集上微调后，我们的模型表现更佳（最后一行，Ours(MoSu)）。这一结果表明，从大规模 MoSu 数据集中学习到的丰富表征能够有效迁移。

5.3 定性演示

我们在图中展示了使用 TripleSumm 生成视频摘要的几个定性示例。3。如中间图表所示，该图展示了*Fusion Token*的注意力权重。随着内容变化，TripleSumm 会逐帧调整各模态的显著性。通过这些权重，我们可以看出，模型恰当地学习并反映了各模态的相对重要性，与人工直觉高度一致，而非出现混乱波动或收敛到静态均值。图3a中的吉他演示视频主要依赖音频信号，同时也能在片头标志（第一个场景）时正确转向视觉模态，在旁白（第二个场景）时转向文本。图3b 展示了一段手风琴演奏，画面变化很少且没有文本。尽管存在缺失模态，模型仍能通过恰当地关注音频做出有力预测。这一能力源于我们的设计：使用中性融合令牌而非特定模态的令牌作为查询，使模型能够不受偏见影响地权衡各模态。整体而言，我们观察到，与基线相比，TripleSumm 通过在每个时刻恰当地权衡各模态，生成了更忠实于真实摘要的摘要。更多定性结果见附录H。

(a) 输入模态

V	T	A	τ	ρ	mAP50	mAP15
✓			0.298	0.404	72.20	40.24
	✓		0.267	0.364	70.01	37.37
		✓	0.271	0.370	69.96	37.88
	✓	✓	0.307	0.416	72.41	41.20
✓	✓		0.328	0.442	73.77	42.39
✓		✓	0.331	0.448	73.85	42.64
✓	✓	✓	**0.351**	**0.472**	**74.72**	**44.42**

(b) 注意力窗口大小

窗口大小 (w)	τ	ρ	mAP50	mAP15
常量（局部）	0.297	0.409	71.61	39.73
常量（宽）	0.304	0.416	72.54	40.26
常量（全局）	0.317	0.431	73.46	41.62
宽到局部	0。325	0。441	73.20	42.13
局部到宽	0.335	0.455	73.83	43.41
全局到局部	0.345	0.461	74.36	43.61
局部到全局	**0.351**	**0.472**	**74.72**	**44.42**

(c) MST 与 CMF 块

MST	CMF	τ	ρ	mAP50	mAP15
	✓	0.250	0.352	70.01	37.35
✓		0.337	0.452	73.90	43.08
✓	✓	**0.351**	**0.472**	**74.72**	**44.42**

(d) 模态融合

方法	τ	ρ	mAP50	mAP15
静态	0.338	0.458	74.08	43.31
全局	0.344	0.461	74.26	43.59
动态	**0.351**	**0.472**	**74.72**	**44.42**

表4：**关于输入模态、注意力窗口大小、MST 和 CMF模块以及模态融合方法的消融研究。**

5.4 消融研究

我们进一步开展了若干 关于 MoSu 的一般消融研究，除非另有说明方面。

输入模态消融。 我们评估了本方法在所有可能输入模态组合下的表现，见表4(a)。在单模态设置中，视觉模态（V）表现最强，而音频（A）略优于文本（T）。这凸显了音频流的作用：它持续提供丰富的上下文线索，而转录文本在“真实场景”视频中往往较为稀疏，导致许多片段缺乏语义信息。每种双模态组合均优于最强的单一模态，体现出显著的协同效应。在此协同基础上，完整的三模态配置取得了最佳性能，证实三种模态均为摘要提供了独特而有价值的信号。

窗口大小消融。 我们比较了多尺度时间块的若干分层窗口划分方案。我们将 $w \in \{3, 5, 7\}$ 记为局部窗口，并将其3至9倍大小记为宽窗口。全局窗口表示 $w = N$，其中 N 为视频的总帧数。基于这些窗口大小，我们试验了多种调度策略：使用恒定的窗口大小，以及逐步扩大或缩小。例如，局部到全局是指在第一层采用局部窗口大小，并在后续层中逐步采用更大的尺寸，直至最后一层覆盖全部帧；例如， $[3, 9, 27, N]$ 或 $[5, 15, 45, N]$。

我们首先从表 4(b) 中观察到，采用固定窗口大小会导致性能不佳。这里，常量（全局）等同于标准自注意力，即在所有层中使用所有帧，但其性能不及以下混合策略。在层次化策略中，我们观察到两种模式。首先，较窄到较宽的策略（局部到宽窗口/全局）往往优于相反的策略（宽窗口/全局到局部），这表明，先通过自底向上方法学习细粒度细节，再据此建立更广泛的上下文，更适合视频摘要。此外，对于自底向上和自顶向下策略，加入一个全局窗口层都能略微提升性能，但代价是增加计算开销。我们基于多种 w 取值开展了更细粒度的实验，并据此得出这些结论，详见附录 C.2。

关于 MST 和 CMF 模块的消融研究。我们通过消融研究验证了方法中两个组件的作用：多尺度时序（MST）模块和跨模态融合（CMF）模块。表 4(c) 中的结果清楚地表明，两个模块对于实现模型的完整性能都不可或缺；移除其中任意一个都会导致性能显著下降。比较这两个组件的影响后，我们发现移除 MST 会造成更严重的性能退化。这凸显了 MST 模块对于准确捕捉长程时序依赖和多尺度上下文至关重要，而这些正是实现精确帧级重要性预测所必需的。总之，CMF 模块通过动态融合处理后的特征，确实带来了最终的多模态增益；而 MST 模块则构成了基础骨干，对于利用视频复杂的长程时序结构不可或缺。这证实了它们在所提出架构中各自发挥的作用。

关于模态融合方法的消融研究。我们论文的一个主要假设是，每种模态的显著性会随视频帧而变化。为验证该假设是否成立，我们对这一设计进行了消融研究。融合三模态词元最简单的方法，是以相同权重对它们取平均（静态）。稍微高级一些的变体，则是通过对交叉模态注意力分数进行时间平均，为每种模态学习一个标量权重（全局）。随后，这些全局权重会在整个视频中统一应用，使模型能够适应不同模态，但无法适应不同视频帧。最后，动态融合会在每个视频帧独立计算交叉注意力分数，并据此确定各模态特征在每帧中的相对贡献。

如表 4(d) 所示，随着模态和时间适应性的加入，性能逐渐提升（动态 $>$ 全局 $>$ 静态）。这一结果证实，模型有效利用了帧级灵活性，且未坍缩为平凡解，验证了我们的 动态 融合机制稳定且对实现有效的多模态摘要至关重要。

5.5 长视频的零样本性能

为更严格地评估模型的可扩展性和泛化能力，我们在 50 段训练期间未见过、明显更长的视频上开展了另一项零样本实验。这些测试视频平均时长为 70.4 分钟，涵盖广泛的主题。关于该测试集的更多详情见附录 B.6。

方法	τ	ρ	mAP50	mAP15
随机	0.000	0.000	50.59	16.18
VASNet	0.024	0.036	54.15	18.95
PGL-SUM	0.024	0.035	54.29	17.55
CSTA	0.083	0.123	58.09	22.26
A2Summ	0.042	0.062	54.04	18.66
SSPVS	0.033	0.048	53.60	18.45
Joint-VA	0.052	0.077	53.99	19.90
UMT	0.066	0.097	56.05	23.10
CFSum	0.061	0.089	56.32	20.59
TripleSumm			**59.70**	**23.27**

表 5：**长视频上的零样本性能**。所有模型均在 MoSu 上训练，并直接在长视频数据集上测试。

所有模型均在 MoSu 上进行训练，表 5报告了它们在该长视频集上的零样本推理性能。结果清楚地证明了 TripleSumm 卓越的泛化能力，尤其是在要求严苛的长篇场景中。对于这些超长视频，我们的模型在基于排名的相关性指标上显著优于所有基线。具体而言，我们的 TripleSumm 在 Kendall 的 τ (0.128) 和 Spearman 的 ρ(0.189) 上取得了最高分。这证实了我们的自适应融合架构能够更加有效地泛化到复杂的长篇内容，在面对全新的领域和语义结构时，仍能保持最准确的视频帧级重要性预测。

据我们所知，这是首个针对数小时长视频的摘要基准。考虑到对于这些叙事结构更复杂的长视频而言，视频摘要的挑战更大，但也更有意义，我们相信，这项实验也为推动视频摘要在真实场景中的应用作出了重要贡献。

6 结论

我们主要探索了将视觉、文本和音频三种模态深度融合，以完成视频摘要任务。我们提出的方法 **TripleSumm** 会动态评估各模态在不同时刻的显著性，并自适应地加以利用，从而生成质量更高的摘要视频。为促进后续研究，我们还推出了新的大规模三模态数据集 **MoSu**。实验表明，TripleSumm 在多个基准上均取得了极具竞争力的结果。这些发现显示，整合全部三种模态有望更全面地理解视频内容。我们希望，所提出的 TripleSumm 模型和 MoSu 数据集能够推动多模态视频摘要领域的未来发展。

尽管为了在现有基准上进行公平比较，我们遵循标准的三步协议（帧重要性评分、时间分割和片段选择），但我们认为，开发一种完全端到端可训练的模型将是一个很有前景的未来方向。探索能够直接学习选择连贯摘要片段的方法，而非仅学习帧级分数，将为推动该领域的发展提供宝贵机会。

致谢

本研究还获得了首尔大学的 SOFT Foundry Institute、三星电子、Youl-chon Foundation、韩国国家研究基金会 (NRF) 的资助（RS-2021-NR05515、RS- 2024-00336576、RS-2023-0022663、RS-2025-25399604、RS-2025-25421642），以及韩国政府资助的信息通信技术规划评估院 (IITP) 的资助（RS-2022- II220264、RS-2024-00353131）。

大语言模型的使用

我们承认，在本文撰写过程中使用了大语言模型，以提升可读性、纠正语法并优化句子结构。所有智力贡献，包括核心思想、方法和分析，均由作者独立完成。

参考文献

Sami Abu-El-Haija、Nisarg Kothari、Joonseok Lee、Paul Natsev、George Toderici、Balakrishnan Varadarajan 和 Sudheendra Vijayanarasimhan。YouTube-8M：大规模视频分类基准。*arXiv:1609.08675*，2016年。

Evlampios Apostolidis、Eleni Adamantidou、Alexandros I Metsai、Vasileios Mezaris 和 Ioannis Patras。通过注意力驱动的对抗学习实现无监督视频摘要。载于国际多媒体建模会议 *(MMM)* 论文集，2019年。

Evlampios Apostolidis、Eleni Adamantidou、Alexandros I. Metsai、Vasileios Mezaris 和 Ioannis Patras。AC-SUM-GAN：结合演员-评论家与生成对抗网络，实现无监督视频摘要。*IEEE*视频技术电路与系统汇刊 *(TCSVT)*，31(8):3278–3292，2021a。

Evlampios Apostolidis、Georgios Balaouras、Vasileios Mezaris 和 Ioannis Patras。结合全局与局部注意力及位置编码的视频摘要。载于*IEEE*多媒体国际研讨会 *(ISM)* 论文集，2021b。

Dawit Mureja Argaw、Seunghyun Yoon、Fabian Caba Heilbron、Hanieh Deilamsalehy、Trung Bui、Zhaowen Wang、Franck Dernoncourt 和 Joon Son Chung。《利用大语言模型扩展视频摘要预训练》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2024。

Jimmy Lei Ba、Jamie Ryan Kiros 和 Geoffrey E Hinton。层归一化。*arXiv:1607.06450*，2016年。

Taivanbat Badamdorj、Mrigank Rochan、Yang Wang 和 Li Cheng。《联合视觉与音频学习用于视频高光检测》。载于*IEEE/CVF*国际计算机视觉会议论文集（*ICCV*），2021年。

Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。《Qwen2.5-VL技术报告》。*arXiv:2502.13923*，2025年。

Iz Beltagy、Matthew E Peters 和 Arman Cohan。Longformer：长文档转换器 r. *arXiv:2004.05150*，2020。

Jiri Fajtl、Hajar Sadeghi Sokeh、Vasileios Argyriou、Dorothy Monekosso 和 Paolo Remagnino。利用注意力机制对视频进行摘要。发表于亚洲计算机视觉会议（*ACCV*）论文集，2018年。

Junaid Ahmed Ghauri、Sherzod Hakimov 和 Ralph Ewerth。通过结合多组特征集与并行注意力进行监督式视频摘要。*arXiv:2104.11530*，2021年。

Yuan Gong、Yu-An Chung 和 James Glass。AST：音频频谱Transformer。*arXiv:2104.01778*，2021年。

Maosheng Guo、Yu Zhang 和 Ting Liu。高斯变换器：一种用于自然语言推理的轻量级方法。发表于 *AAAI*人工智能会议（*AAAI*）论文集，2019。

Yaowei Guo、JiazhengXing、Xiaojun Hou、Shuo Xin、Juntao Jiang、Demetri Terzopoulos、Chenfanfu Jiang 和 Yong Liu。Cfsum：一种基于 Transformer 的多模态视频摘要框架，采用粗细融合。在*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），2025年。

Michael Gygli、Helmut Grabner、Hayko Riemenschneider 和 Luc Van Gool。根据用户视频创建摘要。在欧洲计算机视觉会议论文集（*ECCV*），2014年。

Bo He、Jun Wang、Jielin Qiu、Trung Bui、Abhinav Shrivastava 和 Zhaowen Wang。对齐与关注：采用双重对比损失的多模态摘要。在*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2023年。

Hang Hua、Yunlong Tang、Chenliang Xu 和 Jiebo Luo。V2xum-llm：通过时间提示指令微调实现跨模态视频摘要。载于*AAAI*人工智能会议论文集（*AAAI*），2025年。

Jia-Hong Huang、Luka Murn、Marta Mrak 和 Marcel Worring。Gpt2mvs：用于多模态视频摘要的生成式预训练变换器-2。载于*ACM*国际多媒体检索会议论文集（*ICMR*），2021年。

Hao Jiang 和 Yadong Mu。通过跨任务样本迁移实现联合视频摘要与时刻定位。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2022年。

Yunjae Jung、Donghyeon Cho、Dahun Kim、Sanghyun Woo 和 In So Kweon。无监督视频摘要的判别式特征学习。载于*AAAI*人工智能会议论文集（*AAAI*），2019年。

Jeonghun Kang、Soonmok Kwon、Joonseok Lee 和 Byung-Hak Kim。DIAMOND：一种用于上下文感知的棒球精彩片段摘要的 LLM 驱动智能体。载于计算语言学协会年会论文集（*ACL*），智能体语言模型研究研讨会，2025年。

Maurice G Kendall。排名问题中并列名次的处理。*Biometrika*，33(3):239–251，1945年。

Kwanseok Kim、Jaehoon Hahm、Sumin Kim、Jinhwan Sul、Byunghak Kim 和 Joonseok Lee。Summdiff：基于扩散的视频摘要生成式建模。载于*IEEE/CVF*国际计算机视觉会议论文集（*ICCV*），2025a。

Minsun Kim、Dawon Lee 和 Junyong Noh。使用播放次数最多的数据生成用户指定时长的精彩片段视频。载于*ACM CHI*计算机系统中的人为因素会议论文集（*CHI*），2025b。

Min Jung Lee、Dayoung Gong 和 Minsu Cho。基于大型语言模型的视频摘要。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2025年。

Haopeng Li、Qiuhong Ke、Mingming Gong 和 Tom Drummond。基于多模态自监督学习的渐进式视频摘要。载于*IEEE/CVF*冬季计算机视觉应用会议（*WACV*）论文集，2023年。

Jingyang Lin、Hang Hua、Ming Chen、Yikang Li、Jenhao Hsiao、Chiuman Ho 和 Jiebo Luo。Videoxum：视频的跨模态视觉与文本摘要。*IEEE*多媒体汇刊（*TMM*），26：5548–5560，2024。

Ye Liu、Siyuan Li、Yang Wu、Chang-Wen Chen、Ying Shan 和 Xiaohu Qie。UMT：用于联合视频片段检索和精彩片段检测的统一多模态 Transformer。在*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2022年。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。RoBERTa：一种稳健优化的 BERT 预训练方法。*arXiv:1907.11692*，2019。

Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。Swin Transformer：采用移位窗口的分层视觉 Transformer。发表于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，2021年。

Behrooz Mahasseni、Michael Lam 和 Sinisa Todorovic。基于对抗式 LSTM 网络的无监督视频摘要。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2017。

Medhini Narasimhan、Anna Rohrbach 和 Trevor Darrell。CLIP-It！语言引导的视频摘要。发表于神经信息处理系统进展（*NeurIPS*），2021年。

Medhini Narasimhan、Arsha Nagrani、Chen Sun、Michael Rubinstein、Trevor Darrell、Anna Rohrbach 和 Cordelia Schmid。TL;DW？结合任务相关性与跨模态显著性的教学视频摘要。发表于欧洲计算机视觉会议（*ECCV*）论文集，2022年。

Mayu Otani、Yuta Nakashima、Esa Rahtu 和 Janne Heikkila。《重新思考视频摘要的评估》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2019。

Jungin Park、Jiyoung Lee、Ig-Jae Kim 和 Kwanghoon Sohn。《SumGraph：通过递归图建模实现视频摘要》。载于欧洲计算机视觉会议论文集（*ECCV*），2020。

Danila Potapov、Matthijs Douze、Zaid Harchaoui 和 Cordelia Schmid。《类别特定的视频摘要》。载于欧洲计算机视觉会议论文集（*ECCV*），2014年。

Jielin Qiu、Jiacheng Zhu、William Han、Aditesh Kumar、Karthik Mittal、Claire Jin、Zhengyuan Yang、Linjie Li、Jianfeng Wang、Ding Zhao 等。《MMSum：用于视频多模态摘要和缩略图生成的数据集》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2024。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议（*ICML*）论文集，2021。

Nils Reimers 和 Iryna Gurevych。句子BERT：使用孪生 BERT 网络生成句子嵌入。*arXiv:1908.10084*，2019。

Zirui Shang、Yubo Zhu、Hongxi Li、Shuo Yang 和 Xinxiao Wu。使用去噪扩散概率模型进行视频摘要。载于*AAAI*人工智能会议（*AAAI*）论文集，2025年。

Noam Shazeer。GLU变体改进 Transformer。*arXiv:2002.05202*，2020。

Jaewon Son、Jaehun Park 和 Kwangsu Kim。CSTA：用于视频摘要的基于CNN的时空注意力。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2024。

Yale Song、Jordi Vallmitjana、Amanda Stent 和 Alejandro Jaimes。TVSum：使用标题总结网络视频。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2015 年。

Jinhwan Sul、Jihoon Han 和 Joonseok Lee。Mr. hisum：用于视频高光检测和视频摘要的大规模数据集。发表于神经信息处理系统进展（*NeurIPS*），2023年。

Christian Szegedy、Wei Liu、Yangqing Jia、Pierre Sermanet、Scott Reed、Dragomir Anguelov、Dumitru Erhan、Vincent Vanhoucke 和 Andrew Rabinovich。通过卷积走向更深层。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2015年。Hacene Terbouche、Maryan Morel、Mariano Rodriguez 和 Alice Othmani。用于视频摘要的多标注注意力模型。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2023年。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需的一切。发表于神经信息处理系统进展（*NIPS*），2017年。Junyan Wang、Yang Bai、Yang Long、Bingzhang Hu、Zhenhua Chai、Yu Guan 和 Xiaolin Wei。双重查询：用于视频摘要的双重混合注意力元学习。发表于*ACM*多媒体国际会议论文集（*MM*），2020年。Jiehang Xie、Xuanbai Chen、Shao-Ping Lu 和 Yulu Yang。一种知识增强的多模态视频摘要框架。发表于*ACM*多媒体国际会议论文集（*MM*），2022年。Minghao Xu、Hang Wang、Bingbing Ni、Riheng Zhu、Zhenbang Sun 和 Changhu Wang。基于集合学习的跨类别视频高光检测。发表于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，2021年。Qinghao Yu、Hui Yu、Ying Sun、Derui Ding 和 Muwei Jian。基于特征融合扩散模型的无监督视频摘要。*IEEE*计算社会系统汇刊（*TCSS*），11(5)：6010–6021，2024年。Li Yuan、Francis EH Tay、Ping Li、Li Zhou 和 Jiashi Feng。Cycle-sum：用于无监督视频摘要的循环一致对抗式LSTM网络。发表于*AAAI*人工智能会议论文集（*AAAI*），2019年。Ke Zhang、Wei-Lun Chao、Fei Sha 和 Kristen Grauman。利用长短期记忆网络进行视频摘要。发表于欧洲计算机视觉会议（*ECCV*）论文集，2016年。Yunzuo Zhang、Yameng Liu、Weili Kang 和 Ran Tao。VSS-Net：用于视频摘要的视觉语义自挖掘网络。*IEEE*视频技术电路与系统汇刊（*TCSVT*），34(4)：2775–2788，2024年。Bin Zhao、Xuelong Li 和 Xiaoqiang Lu。用于视频摘要的分层循环神经网络。发表于*ACM*多媒体国际会议论文集（*MM*），2017年。Bin Zhao、Xuelong Li 和 Xiaoqiang Lu。HSA-RNN：用于视频摘要的层次结构自适应RNN。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2018年。Bin Zhao、Haopeng Li、Xiaoqiang Lu 和 Xuelong Li。用于视频摘要的重构式序列图网络。*IEEE*模式分析与机器智能汇刊（*TPAMI*），44(5)：2793–2801，2021a。Bin Zhao、Xuelong Li 和 Xiaoqiang Lu。TTH-RNN：用于视频摘要的张量链分层循环神经网络。*IEEE*工业电子汇刊（*TIE*），68(4)：3629–3637，2021b。Bin Zhao、Maoguo Gong 和 Xuelong Li。用于视频摘要的分层多模态Transformer。*Neurocomputing*，468：360–369，2022年。Wencheng Zhu、Jiwen Lu、Jiahao Li 和 Jie Zhou。DSNet：一种灵活的视频检测到摘要网络。*IEEE*图像处理汇刊（*TIP*），30：948–962，2021年。Wencheng Zhu、Yucheng Han、Jiwen Lu 和 Jie Zhou。视频摘要中的时空图关系推理。*IEEE*图像处理汇刊（*TIP*），31：3017–3031，2022年。

Daniel Zwillinger 和 Stephen Kokoska。*CRC* 标准概率与统计表及公式。CRC Press，1999。

附录

A 实现细节

A.1 实验设置

模型的超参数详见表I。架构由 2 个交错层组成，每层包含 2 个多尺度时序模块和 2 个跨模态融合块。整个架构共包含 4 个多尺度时序模块。为捕捉由局部到全局尺度的特征，注意力窗口大小 w 在这些模块中逐步增大，从局部自注意力（窗口大小为 5）增至 15、45 和 N，其中 N 为输入序列长度。前馈网络（FFN）层采用 Swish 门控线性单元（SwiGLU）（Shazeer，2020）。最终的分数预测头依次由线性投影、GeLU 激活、层归一化、另一线性投影以及最终的 Sigmoid 激活组成。该预测头的隐藏维度设为 192。架构参数包括维度为 128 的嵌入维度（D）。模型使用 AdamW 优化器训练了 100 轮。学习率初始设为 1×10^{-4}，并通过余弦调度器进行调整。所有实验均在单台 NVIDIA RTX A100 上完成。

类别	超参数	值
模型架构	嵌入维度（D）	128
	多尺度时序模块数量（P）	2
	跨模态融合模块数量（Q）	2
	交错模块数量（L）	2
	注意力头数量	4
	预测头的隐藏维度	192
训练细节	轮次	100
	批量大小	64
	Dropout率	0.1
	初始学习率	1×10^{-4}

表 I：**所提出模型的超参数**。

A.2 摘要生成流程

ˆ为根据预测的逐帧重要性分数 S构建最终摘要，我们遵循以往工作采用的标准流程（Son 等人，2024；Kim 等人，2025a）。首先，使用核时间分割（KTS）将每个输入视频划分为时间连贯的镜头（Potapov等人，2014年；Zhang等人，2016年）。随后，将每个镜头内预测的帧级分数取平均，作为该镜头的重要性分数。接下来，我们选择一组最优镜头，使镜头分数总和最大，同时满足预先设定的长度预算（例如，原视频时长的 15%）；这一选择过程被建模为 0/1背包问题，其中镜头长度为权重，镜头分数为价值。最后，按照原有时间顺序拼接所选镜头，生成最终摘要视频。

B 数据集详情

B.1 THEMOSU 数据集的详细统计信息

我们在第 4 节介绍了 MoSu 数据集，现进一步说明其特征。首先，图 I 展示了该数据集中视频时长的分布，范围为 120 至 500 秒。表 II 给出了视频时长、文本词元和音轨的统计数据。由于我们筛除了所有没有有效音轨的视频（例如，创作者将音频静音或移除音频时就可能出现这种情况），因此该数据集中的所有视频都包含音频信息，适用于三模态融合。

统计类别	数值
时长统计	
平均时长	272.25 秒
标准差	102.43 秒
最短时长	120.00 秒
最长时长	501.00 秒
文本统计	
词元总数	32.6M
每个视频的平均词元数	619.1
转录密度	61.84%
音频统计	
音频可用性	100%

表 II：**MoSu 数据集的详细统计信息**。转录密度是指具有有效文本的视频时长的平均占比。

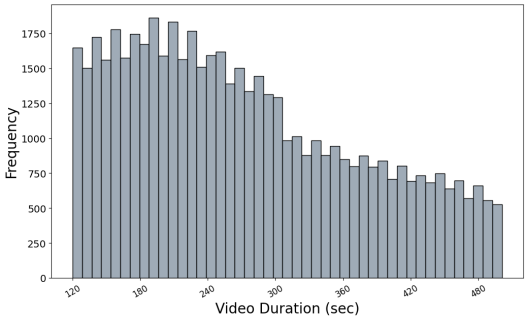


图 I：**MoSu 数据集中的视频时长**。

B.2 真实值预处理

YouTube 的“最常重播”统计数据被用作识别视频关键时刻的真实值。它是从数万次观看中收集的重播次数归一化频率，以 100 个分数构成的序列形式提供，每个分数对应一个均匀划分的片段。遵循Sul 等人(2023年)；Kim 等人(2025b) 的先例，我们仅使用观看次数超过 50,000 次的视频，以确保统计可靠性。

处理 YouTube 最常重播偏差。这些统计数据中观察到一种常见的异常现象：视频开始后的最初几秒内，重播分数会出现异常峰值，而且通常发生在主要内容开始之前。这种现象被称为 YouTube 最常重播偏差，可能是由于观众在视频开头注意力更加集中，或受到初始缓冲调整的影响。若不加以处理，这种偏差可能会误导模型，使其将高重要性与时间位置而非语义内容相关联。为减轻这一影响，并确保模型学习基于内容的重要性，我们将每个视频前五秒对应的真实分数置零。

B.3 主题类别聚类

表 III 展示了每个已识别聚类中的视频数量，以及出现频率最高的五个实体标签。观察发现，排名靠前的实体与分配的聚类主题之间具有高度的语义一致性，这验证了聚类方法的有效性。我们对每个聚类按 8:1:1 的比例进行分层划分，将数据集划分为训练集、验证集和测试集。这样可确保各聚类在所有数据子集中均保持相同比例。

B.4 数据预处理详情

为确保可复现性，我们提供了特征提取流程的详细设置。该流程以 1 秒为间隔，将所有视频处理为时间对齐的特征序列。

视觉特征。 对于 MoSu 数据集，我们每秒采样一帧，并使用官方预训练 CLIP 模型[1] 为每个视频帧提取一个 768 维特征向量。对于外部基准数据集，为确保公平比较，我们直接采用其官方提供的视觉特征：Mr. HiSum 使用 Inception-v3 特征，SumMe 和 TVSum 使用 GoogLeNet 特征。

文本特征。 文本特征来源于两个渠道：MoSu 数据集的 YouTube 转录文本，以及外部基准数据集的生成字幕。对于 MoSu 数据集，使用了 YouTube 提供的带时间戳转录文本。英文转录文本直接使用，非英文转录文本则通过 YouTube 的自动翻译 API 进行翻译。没有任何转录文本的视频被排除。对于外部基准数据集，则使用视觉语言

[1] openai/clip-vit-large-patch14

聚类	视频	出现频率最高的五类实体
电子游戏	11,578	比赛、电子游戏、动作冒险游戏、使命召唤、策略电子游戏
乐器	7,647	吉他、弦乐器、音乐家、原声吉他、电吉他
时尚与材料	5,588	化妆品、头发、发型、时尚、艺术
汽车	5,207	车辆、轿车、发动机、驾驶、跑车
电子设备	4,291	小工具、手机、智能手机、三星Galaxy、iPhone
美食与烹饪	3,919	食物、食谱、烹饪、甜点、菜肴
体育	3,784	比赛、重量训练、篮球、健身房、篮球动作
动画与漫画	3,752	卡通、喜剧（戏剧）、动画、艺术、喜剧演员
车辆与交通	3,731	车辆、摩托车、骑行、自行车、摩托车运动
动物	3,181	动物、宠物、狗、钓鱼、鱼

表 III：簇大小及其出现频率最高的五类实体。

模型[2]。随后，所有获取的文本均经过统一编码处理。首先，移除了非语言注释（例如，[音乐], [掌声]）。随后，使用预训练 RoBERTa 模型的 [CLS] 词元嵌入[3]将每个句子编码为 768 维特征向量。然后，将该向量广播至其持续时间所覆盖的所有 1 秒时间戳。对于没有关联文本的时间戳，则分配一个默认向量，该向量通过将一个<填充>词元输入编码器预先计算得到。

音频特征。 为了处理音频数据，我们为音频流的每一秒提取一个特征向量。具体做法是采用中心窗口方法，以捕获充分的语义上下文，并匹配模型的标准输入大小。具体而言，对于每一秒 t，我们从对称区间 $[t-5, t+5]$中提取一个 10 秒片段。这种双向加窗确保时间点 t 的表征能够结合前后发生的事件。随后，我们利用预训练音频频谱Transformer模型[4]的 [CLS] 令牌的输出嵌入，将这些片段编码为固定大小的特征向量。最终，整个音频流被转换为一系列上下文化表示。

B.5 SumMe 和 TVSum 中的数据划分

为确保内容完整并提高可复现性，我们详细介绍主论文中使用的两种评估方案。两种方案均采用 5 折交叉验证框架，但在模型开发与评估时的数据划分方式上存在显著差异。

传统训练/验证（TV）划分。 遵循早期视频摘要研究中的标准做法，将数据集随机划分为五个大小相等的折。每轮迭代中，使用其中四折进行训练，剩余一折用于评估，最终结果取五折的平均值。然而，该方案没有单独划分专用测试集。因此，超参数调优和早停决策实际上都是依据留出的评估折上的性能作出的，这使模型极易过拟合。

训练/验证/测试（TVT）划分。 为了解决传统 TV 设置固有的过拟合问题，我们还采用 Kim 等人 (2025a)所采用的方法，在五折交叉验证框架中使用更严格的 TVT 划分来评估我们的方法。在此设置下，五个折被明确划分为训练集、验证集和测试集（通常使用三折进行训练、一折进行验证、一折进行测试，约为 6:2:2 的比例）。验证折仅用于超参数调优和早停，确保在模型开发阶段测试折完全不可见。该协议能够对模型真正的泛化能力进行更严格、更可靠的评估。

[2] Qwen/Qwen2-VL-7B-Instruct [3] FacebookAI/roberta-base [4] MIT/ast-finetuned-audioset-10-10-0.4593

B.6 长视频测试集

为了在更贴近现实且更具挑战性的设置下评估视频摘要模型，我们收集了一个专门的评估集，其视频时长显著超过其他任何公开视频摘要数据集。传统的视频摘要基准，例如 SumMe（Gygli 等人，2014）和 TVSum（Song 等人，2015），所含视频的平均时长为 2–5 分钟。近期的数据集，如 LfVS-P（Argaw 等人，2024）和 MMSum（Qiu 等人，2024），则将平均时长延长至 13–14 分钟，如表 I所示。然而，这些数据集中的视频时长仍远不及现实世界中的长视频，例如数小时长的电影或纪录片。

我们收集了 50 个多样化的长视频，旨在通过 MoSu 训练集未涵盖的领域来挑战模型。MoSu 涵盖了广泛的类别（例如，电子游戏、时尚、动物），而该数据集则包含复杂的叙事结构和专业内容。例如，其中包括长篇电影和纪录片、专业技术教程、完整的体育比赛以及多语言脱口秀。这些视频的真实摘要采用与 MoSu 相同的“最常重播”分数获取。该新测试集的关键统计数据汇总于表 IV中。值得注意的是，平均视频时长为 70.4 分钟，是 MoSu 训练视频的 10 至 20 倍。

统计数据	**时长（秒）**
平均时长	4224.0
标准差	1263.6
最短时长	2413.0
最长时长	7207.0

表 IV：**长视频数据集的统计数据**。

C 架构消融研究

C.1 模型层

层数			τ ↑	ρ ↑	**mAP50** ↑	**mAP15** ↑	**参数量** ↓
整体 (L)	**MST (P)**	**CMF（Q）**					
8	8	8	0.326	0.442	73.41	42.2	17.15M
4	4	4	0.346	0.465	74.39	43.34	4.53M
2	**2**	**2**	**0.351**	**0.472**	**74.72**	**44.42**	1.37M
2	2	1	0.330	0.451	73.67	42.96	1.11M
2	1	2	0.339	0.461	74.41	43.33	1.11M
2	1	1	0.340	0.461	74.38	43.40	0.85M
1	1	1	0.314	0.430	73.4	41.61	**0.58M**

表 V：**关于模型层数的消融研究**。整体表示总层数，而 MST 和 CMF 分别指多尺度时序模块和跨模态融合模块。

为探究网络深度的影响，我们同时改变总层数以及多尺度时序（MST）模块与跨模态融合（CMF）模块之间的层分配，结果见表 V。增加 CMF 层数并未带来额外收益，反而导致所有评估指标上的整体性能略有下降；而降低 MST 深度则会持续降低性能。这些结果表明，时间建模需要中等深度的 MST 才能获得有效表征，而单个融合层足以实现跨模态整合。

C.2 窗口大小

为分析时序感受野对摘要性能的影响，我们针对基于窗口的自注意力机制测试了多种窗口大小配置，如下所示

窗口大小 (w)	$\tau \uparrow$	$\rho \uparrow$	mAP50 ↑	mAP15 ↑
基线				
标准自注意力	0.317	0.431	73.46	41.62
固定窗口				
$5, 5, 5, 5$	0.297	0.409	71.61	39.73
$15, 15, 15, 15$	0.318	0.430	72.64	41.15
$45, 45, 45, 45$	0.322	0.435	72.95	42.10
全局到局部				
$N, 27, 9, 3$	0.340	0.462	74.22	43.09
$N, 45, 15, 5$	0.345	0.461	74.36	43.61
$N, 63, 21, 7$	0.347	0.468	74.52	43.54
局部到全局				
$3, 9, 27, N$	0.346	0.466	74.58	43.91
$\mathbf{5, 15, 45, N}$	**0.351**	**0.472**	**74.72**	**44.42**
$7, 21, 63, N$	0.348	0.468	74.67	43.92

表 VI：**不同窗口大小配置的消融研究**。 N 表示完整序列长度（全局注意力）。

见表 VI。具体而言，我们比较了固定窗口大小和分层配置（全局到局部和局部到全局）。结果表明，局部到全局策略，具体而言是 $[5, 15, 45, N]$，在所有指标上都取得了最佳性能。该配置逐步扩大从较低层到较高层的感受野，使模型能够在早期阶段捕捉细粒度局部时序动态，并在后期阶段整合全局上下文。相反，固定的小窗口无法捕捉长程依赖，而从全局注意力开始（全局到局部）似乎不太能保留关键的局部细节。这些发现证实，逐层扩大的窗口结构最适合建模视频内容的多尺度特性。

C.3 可学习窗口

窗口大小（w） $\tau \uparrow$		$\rho \uparrow$	mAP50 ↑	mAP15 ↑
标准自注意力	0.317	0.431	73.46	41.62
+ 可学习	0.318 (+0.001)	0.429 (-0.002)	73.65 (+0.19)	41.44 (-0.18)
全局到局部	0.345	0.461	74.36	43.61
+ 可学习	0.344 (-0.001)	0.465 (+0.004)	74.49 (+0.13)	43.77 (+0.16)
局部到全局	0.351	**0.472**	74.72	**44.42**
+ 可学习	**0.353 (+0.002)**	0.469 (-0.003)	**74.83 (+0.11)**	44.32 (-0.10)

表 VII：**可学习高斯调制对不同注意力策略的影响**。

理想情况下，最好能根据数据自适应地选择一组窗口大小。然而，根据我们最初的实验，从头开始学习这些窗口大小极不稳定，这可能是由于自由度过高。为此，我们采用基于高斯调制自注意力机制的可学习窗口配置，其概念与（Guo 等人，2019）类似，并使用固定窗口大小的预训练模型进行初始化。核心思想是将高斯分布的参数（包括标准差 σ）设为可学习值，使模型能够自适应地调整每一层的感受野宽度（或尺度）。具体而言，通过使用高斯掩码 $\mathbf{G}_{i,\,j}$调制注意力分数，参数 σ 直接控制注意力权重随时间距离 $|i-j|$增加而衰减的速率。从数学上说，较小的 σ 会使指数项对较远的视频帧迅速趋近于零，从而有效地增强局部性；而较大的 σ 则会产生更平坦的分布，使模型能够捕获全局上下文。

如表 VII所示，这种自适应机制并未带来显著差异。在包括标准自注意力、全局到局部和我们的局部到全局策略在内的所有配置中，其效果都微乎其微，部分指标甚至有所下降。这表明，我们手动设计的

局部到全局结构已能高效捕捉所需的多尺度时间依赖。相比我们经过精心调优的固定窗口，学习感受野方差所增加的复杂性并未带来切实收益。

C.4 嵌入维度

D	$\tau \uparrow$	$\rho \uparrow$	mAP50 ↑	mAP15 ↑	参数量 ↓	GFLOPs ↓
768	0.234	0.324	68.25	34.50	39.71M	31.77G
512	0.313	0.424	72.40	40.84	18.07M	14.25G
384	0.344	0.464	74.44	43.18	10.42M	8.10G
256	0.347	0.467	74.54	44.1	4.85M	3.67G
192	0.350	0.471	74.59	43.84	2.85M	2.10G
128	**0.351**	**0.472**	**74.72**	**44.42**	1.37M	0.97G
96	0.348	0.467	74.56	43.8	**0.84M**	**0.57G**

表 VIII： **不同嵌入维度的消融研究**。 D 表示嵌入维度。

我们进一步研究了嵌入维度 D 对模型效率和摘要准确率的影响。如表 VIII所示，我们观察到模型在$D = 128$ 至 256左右时性能最佳。性能在增至 $D = 384$时略有下降，而更大的 D 则会导致模型严重过拟合。即使嵌入维度极低，仅为 $D = 96$，模型仍保持着有竞争力的性能，展现了特征表示的鲁棒性和紧凑性。基于这些结果，我们选择 $D = 128$ 作为最优设置，在轻量级架构与高质量摘要之间取得良好平衡。

C.5 共享参数

设置	$\tau \uparrow$	$\rho \uparrow$	mAP50 ↑	mAP15 ↑	参数量 ↓
不使用共享参数	0.343	0.459	74.24	43.24	2.95M
使用共享参数	**0.351**	**0.472**	**74.72**	**44.42**	**1.37M**

表 IX： **共享参数对性能的影响**。

为评估多尺度时序模块中参数共享的影响，我们进行了消融研究（表 IX）。无共享参数配置为每种模态分配一个独立的时序模块，而共享参数配置则采用一个由视觉、文本和音频流共用的模块。

参数共享将可学习权重的数量从 2.95 M 减少到 1.37 M，降幅近三分之二，同时在所有评估指标上均取得了更高的分数。这一结果表明，参数共享使时序模块能够捕捉不同模态共有的多尺度模式。该共享模块同时在所有模态上训练，因此接触到的训练序列约为每种模态使用独立模块时的四倍。数据接触量的增加，加上多尺度时序结构本身对长程和短程动态的建模能力，清楚地解释了为何共享参数设计尽管参数数量大幅减少，仍能取得更优的性能。

C.6 融合令牌

我们探索了将三模态特征（$\mathbf{e}_i^v$，$\mathbf{e}_i^t$，$\mathbf{e}_i^a$）聚合到融合令牌 $\mathbf{e}_i^f$ 中的三种方式：1）以视觉为中心的基线，仅使用视觉特征（$\mathbf{e}_i^f = \mathbf{e}_i^v$；无融合）；2）将融合令牌本身视为可学习参数（$\mathbf{e}_i^f \in \mathbb{R}^D$；可学习）；3）对视觉、文本和音频特征进行简单平均（$\mathbf{e}_i^f = (\mathbf{e}_i^v + \mathbf{e}_i^t + \mathbf{e}_i^a)/3$；平均）。

令人惊讶的是，表 4(d) 表明，简单的 平均初始化取得了最高性能。该无融合 基线主要依赖视觉信息，因此在

语音或环境声音等非视觉线索占主导地位，或某些模态缺失时，它可能会受到影响。平均对所有模态赋予相同权重，虽然简单，但能有效平衡所有模态，将其用作查询。

方法	τ ↑	ρ ↑	mAP50 ↑	mAP15 ↑
无融合	0.337	0.451	73.96	43.48
可学习	0.341	0.462	74.41	43.71
平均	**0.351**	**0.472**	**74.72**	**44.42**

表 X：**关于融合令牌聚合的消融研究**。

D 计算成本分析

方法	τ ↑	ρ ↑	参数量 ↓	GFLOPs ↓	推理时间 ↓
VASNet	0.151	0.219	8.31M	1.99G	7.36ms
PGL-SUM	0.151	0.218	5.31M	1.21G	13.55ms
CSTA	0.291	0.398	10.56M	11.37G	9.48ms
A2Summ	0.181	0.257	2.48M	1.35G	44.29ms
SSPVS	0.190	0.271	112.81M	43.64G	14.27ms
Joint-VA	0.190	0.272	4.21M	1.63G	**2.58ms**
UMT	0.239	0.334	4.66M	1.39G	4.94毫秒
CFSum	0.277	0.374	19.83M	8.52G	3.87ms
TripleSumm	**0.351**	**0.472**	**1.37M**	**0.97G**	2.81ms

表 XI：**计算成本和性能比较**。我们报告秩相关得分（τ、 ρ）、模型大小（参数量）、推理成本（GFLOPs）和推理时间，以评估效率与准确率之间的权衡。

我们比较 TripleSumm 与其他方法在计算复杂度和推理速度方面的表现，以验证其在实际应用中的适用性。表XI报告了摘要准确率（τ 和 ρ）、竞争模型的可学习参数量（参数量）、GFLOPs 和推理时间。TripleSumm 在效率方面达到了新的最先进水平，仅需 1.37M 个可学习参数和 0.97 GFLOPs，显著低于所有其他基线。尽管模型轻量，我们的模型在相关性指标（$\tau = 0.361$， $\rho = 0.484$）上仍大幅优于所有基线。此外，TripleSumm 的推理时间仅为 2.81 ms，速度与最快的方法（Joint-VA）相当，同时提供了更优的摘要质量。此分析证实，TripleSumm 成功打破了性能与效率之间的权衡，为视频摘要提供了高度可扩展的解决方案。

E 自适应学习动态的补充分析

E.1 对学习到的模态显著性的消融研究

为验证我们的模型是否能在每个视频帧上有效地为真正重要的模态分配注意力权重，我们在 MoSu 测试集上进行了另一项消融研究。在推理时，我们使用 *Fusion Token* 在每个视频帧上分配给三种模态（视觉、文本和音频）的注意力权重。随后，我们根据权重对它们进行排序：每个时间步分别对应最高权重（排名 1）、次高权重（排名 2）和第三高权重（排名 3）。根据这些排名，我们仅选取部分模态（例如，“仅排名 1”或“排名 $1+2$”）。
如表 XII所示，完整的 TripleSumm 模型始终利用所有模态，在全部指标上均取得最佳性能，证实三种模态的加权贡献是最优策略。正如预期，当模型被迫忽略更多高权重模态时，性能逐渐下降，在仅限于最不重要的模态（“仅 Rank 3”）时表现最差。这表明注意力机制成功抑制了噪声，并优先处理信息性线索。显著的性能提升从

“仅 Rank 1”到“Rank 1 + 2”验证了第二重要的模态提供了关键的互补信息。总之，该消融研究证实，Fusion Token 的注意力机制成功学习了各模态的帧依赖显著性，从而使 TripleSumm 取得更优性能。

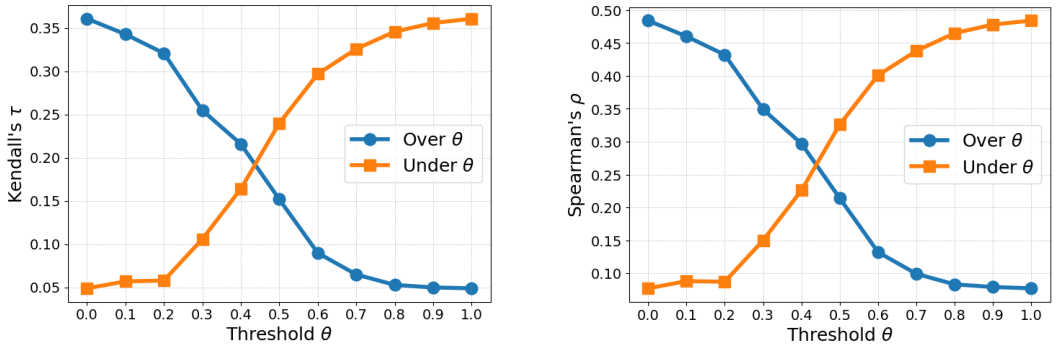
方法	$\tau \uparrow$	$\rho \uparrow$	mAP50 ↑	mAP15 ↑
仅第 3 名	0.058	0.087	59.44	25.20
仅第 2 名	0.136	0.194	63.34	29.99
排名 2 + 3	0.158	0.220	64.06	30.81
仅排名 1	0.196	0.273	66.09	33.28
排名 1 + 2	0.273	0.373	70.28	38.69
TripleSumm（我们的方法）	**0.351**	**0.472**	**74.72**	**44.42**

表 XII：**基于估计显著性移除模态特征的消融实验。**

E.2 注意力权重分布分析

为了考察 *Fusion Token*学到的注意力权重分布，我们进行了另一项与上述类似的消融研究（附录 E.1），但采用特定阈值 θ 而非相对重要性。本实验旨在评估学到的权重是高度集中于少数主导模态，还是更广泛地分布在输入特征中。

如图 II所示，分析采用了两种过滤策略。“Over θ”仅保留权重$\geq \theta$ 并将其他权重设为零），而“Under θ”仅保留权重 $\leq \theta$ 并将其他权重设为零）。性能指标 Kendall’s τ 和 Spearman’s ρ随着阈值 θ 从0.0 变化至 1.0，呈现出显著平滑的线性反比关系。这种线性关系表明，注意力权重并未集中于少数固定的特征子集。若权重是稀疏的，我们预期性能会出现剧烈的非线性变化。相反，“Over θ”性能逐渐下降，而“Under θ”性能逐渐提升（随着θ 增大并纳入更多细微权重），这表明 Fusion Token 广泛而细致地分配其注意力权重。这种广泛分布证实，模型从多种特征和模态中提取了互补信息，也验证了 TripleSumm 能够有效利用分布式显著性，而非依赖单一主导线索。



(a) Kendall 的 τ（阈值为 θ）　　(b) Spearman 的 ρ（阈值为 θ）

图 II：**通过对模态注意力权重设置阈值进行性能评估。** 该图展示了对 Fusion Token 学到的模态注意力权重应用阈值θ 时，模型的性能；性能分别以 (a) Kendall 的 τ和 (b) Spearman 的 ρ衡量。**高于** θ 的线表示仅保留权重 $\geq \theta$ 时的性能，而**低于** θ 的线表示仅保留权重 $\leq \theta$ 时的性能。

F 外部数据集的完整结果

F.1 TVSum 和 SumMe

方法	SumMe		TVSum	
	τ ↑	ρ ↑	τ ↑	ρ ↑
随机	0.000	0.000	0.000	0.000
人工	0.205	0.213	0.177	0.204
SUM-GAN(Mahasseni 等人，2017)	0.049	0.066	0.024	0.031
VASNet(Fajtl et al., 2018)	0.160	0.170	0.160	0.170
DSNet-AF(Zhu 等人，2021年)	0.037	0.046	0.113	0.138
DSNet-AB(Zhu 等人，2021年)	0.051	0.059	0.108	0.129
AC-SUM-GAN(Apostolidis et 等，2021a)	0.102	0.088	0.031	0.041
DMASum(Wang 等人, 2020)	0.063	0.089	0.203	0.267
CLIP-It(Narasimhan 等人,2021年)	-	-	0.108	0.147
PGL-SUM(Apostolidis et al., 2021b)	0.192	0.213	0.157	0.206
UMT(Liu et al., 2022)	0.241	0.268	0.179	0.235
iPTNet(Jiang 与 Mu，2022年)	0.101	0.119	0.134	0.163
A2Summ(He et al., 2023)	0.108	0.129	0.137	0.165
AAAM(Terbouche 等人, 2023年)	-	-	0.169	0.223
MAAM(Terbouche 等人, 2023年)	-	-	0.179	0.236
VSS-Net(Zhang 等人,2024)	-	-	0.190	0.249
Joint-VA(Badamdorj et 等，2021年)	0.230	0.256	0.166	0.220
SSPVS(Li 等人,2023年)	0.192	0.257	0.181	0.238
CSTA(Son 等，2024)	0.246	0.274	0.194	0.255
LLMVS(Lee 等，2025年)	0.253	0.282	0.211	0.275
SummDiff(Kim 等，2025a)	0.256	0.285	0.195	0.255
本文方法(Full)	0.265	0.296	0.211	0.275
本文方法(MoSu)	**0.282**	**0.314**	**0.217**	**0.282**

Tab第十三： **在 SumMe 和 TVSum 数据集上的性能比较**。结果报告 d
und采用传统的训练/验证（TV）设置进行 5 折交叉验证。

表 XIII展示了在 5 折交叉验证（5FCV）框架下，采用传统训练/验证（TV）划分时，SumMe 和 TVSum 数据集上的性能比较。提供这些结果主要是为了确保与先前研究的可比性，因为以往文献主要采用这种评估设置。

传统的 TV 划分便于与现有基线进行比较，也展示了我们的方法在不同随机划分下的稳健性；而 TVT 划分通过消除超参数优化过程中可能出现的测试集污染，更可靠地衡量了真正的泛化性能。在这两种评估设置下，TripleSumm 均始终优于以往方法，尤其是在使用我们的 MoSu 数据集进行预训练时，提升尤为显著。在不同评估协议下始终如一的性能提升，证明了我们多模态融合设计的有效性，也验证了大规模预训练数据的益处。

F.2 MR. HISUM 数据集

方法	τ ↑	ρ ↑	mAP50 ↑	mAP15 ↑
SUM-GAN (Mahasseni 等,2017)	0.067	0.095	56.62	23.56
VASNet (Fajtl等人,2018年)	0.069	0.102	58.69	25.28
AC-SUM-GAN (Apostolidis 等人, 2021a)	0.012	0.018	55.35	21.88
SL 模块 (Xu 等人,2021年)	0.060	0.088	58.63	24.95
PGL-SUM (Apostolidis 等人, 2021b)	0.097	0.141	61.60	27.45
Joint-VA (Badamdorj 等人, 2021年)	0.161	0.231	65.88	35.23
iPTNet (Jiang 和 Mu, 2022年)	0.020	0.029	55.53	22.74
UMT (Liu 等人, 2022年)	0.178	0.253	66.81	35.65
A2Summ (He 等人,2023年)	0.121	0.172	63.20	32.34
SSPVS (Li 等人, 2023年)	0.078	0.113	59.48	26.35
CSTA (Son 等人, 2024)	0.128	0.185	63.38	30.42
我们的 (视觉)	0.187	0.258	67.16	35.57
我们的 (完整)	**0.258**	**0.352**	**70.72**	**40.88**

表 XIV： **在 Mr. HiSum 数据集上的比较。**

为提供完整的基准，我们在表 XIV 中报告了 Mr.HiSum 数据集上的详细结果。TripleSumm 在所有相关性指标上均取得最高分，明显领先于以往最先进的方法。

G 数据集消融研究

G.1 数据规模扩展

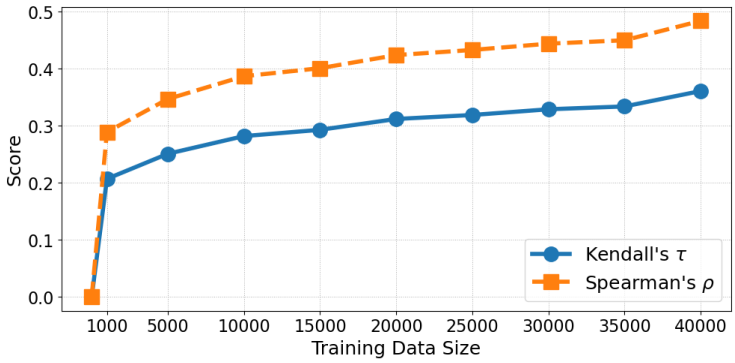


图 III： **评估 MoSu 数据集规模对 TripleSumm 性能的影响。**

为考察训练数据量对性能的影响，我们使用规模逐步增大的 MoSu 子集训练 TripleSumm。如图 III所示，Kendall 的 τ 和 Spearman 的 ρ 均与训练样本数量呈现明显的正相关。值得注意的是，即使使用整个 MoSu 数据集，性能曲线仍未达到饱和，这表明增加数据有望进一步提升性能。这一发现凸显了 SumMe 和 TVSum 等传统基准数据集的数据匮乏问题，并强调了 MoSu 等大规模数据集对于推动视频摘要领域发展的必要性。

G.2 类别分布

为评估标签平衡情况，我们分析了 MoSu 数据集中语义簇的分布（附录 B.3）。该数据集由十个规模和主题各异的不同聚类组成；不过，样本分布足够均衡，不会让任何单一类别在训练过程中占据主导地位。

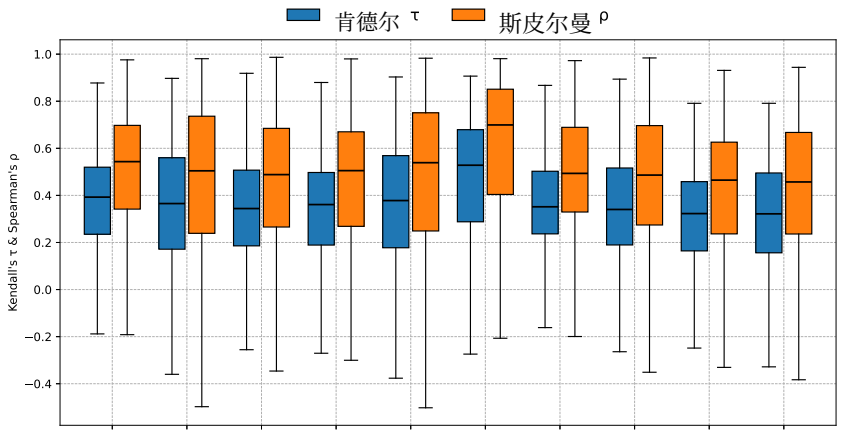
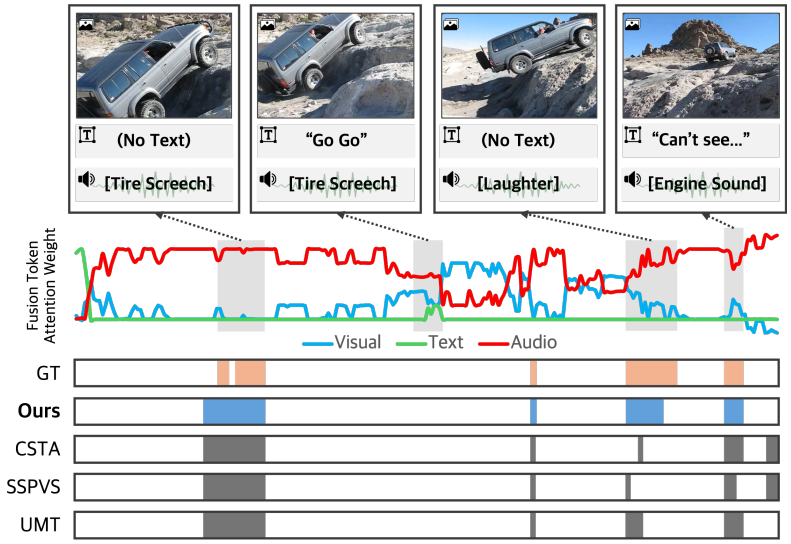


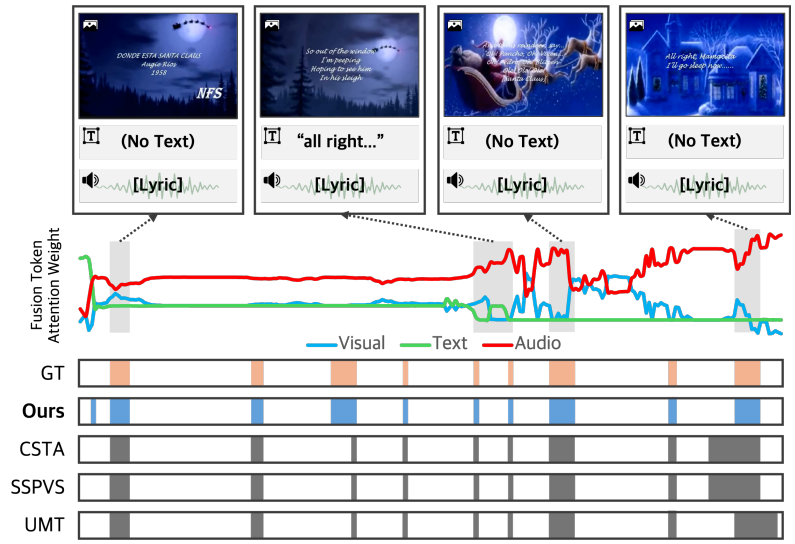
图 IV：**MoSu 数据集上的逐聚类性能**。 箱线图展示了每个语义簇的肯德尔 τ（蓝色）和斯皮尔曼 ρ（橙色）分布。

按簇进行的性能评估（以相关性指标衡量）进一步证实了这种平衡（图 IV）。各簇的中位数与整体均值相近，且与样本量没有明显相关性。例如，样本较少的簇往往也能取得与样本较多的簇相当的性能，而一些样本充足的簇则表现出更大的方差或略低的中位数。值得注意的是，“体育”和“动画与漫画”簇的中位数更高、分布更集中，表明模型对这些类别的预测更稳定。这些结果表明，模型准确率更多受到各簇内容难度（例如视觉复杂度、音频变化）的影响，而非数据量。

H 更多定性结果



(a) 车辆越过障碍物，轮胎尖啸声这一音频线索对于理解该动作至关重要。



(b) 一段以音轨为核心的音乐视频，其中视觉和文本线索的重要性较低。

图 V：**MoSu 上的更多定性示例**。中间的图展示了 Fusion Token 的注意力权重，说明我们的模型如何动态估计每种模态的显著性，从而在部分模态缺失时仍能保持较高的摘要准确率。