# 语义引导的无监督视频摘要

1st HaizhouLiu上海理工大学控制科学与工程系中国上海 200093
233370837@st.usst.edu.cn

2nd HaodongJin上海理工大学控制科学与工程系中国上海 200093
231260086@st.usst.edu.cn

3rd Yiming Wang格拉斯哥大学心理学与神经科学学院英国格拉斯哥 G12 8QQ
yiming.wang@glasgow.ac.uk

4th HuiYu格拉斯哥大学心理学与神经科学学院英国格拉斯哥 G12 8QQ hui.yu@glasgow.ac.uk

arXiv:2601.14773v1 [cs.AI] 21 Jan 2026

摘要—**视频摘要是社会理解的一项关键技术，能够高效浏览海量多媒体内容，并从社交平台中提取关键信息。现有的大多数无监督摘要方法依赖生成对抗网络（GAN），通过对抗训练改进关键帧选择，并生成连贯的视频摘要。然而，这类方法主要利用单模态特征，忽略了语义信息在关键帧选择中的引导作用，而且往往存在训练不稳定的问题。为解决这些局限，我们提出了一种新颖的语义引导的无监督视频摘要方法。具体而言，我们设计了一种新颖的帧级语义对齐注意力机制，并将其集成到关键帧选择器中，以引导对抗框架内基于 Transformer 的生成器更好地重建视频。此外，我们采用增量训练策略，逐步更新模型组件，从而有效缓解生成对抗网络训练的不稳定性。实验结果表明，我们的方法在多个基准数据集上取得了优异的性能。**

索引词—**语义引导、社会智能、视频摘要。**

## I. 引言

随着摄像机和录制设备的迅速普及，每天都有海量视频数据被采集，用以记录各类社会事件。然而，视频内容数量庞大，也带来了一个重大挑战：如何高效地搜索、分类和浏览感兴趣的特定信息。为了解决这一问题，视频摘要 [1], [2], [3]已成为一种有效技术，使用户能够快速浏览大规模视频集合并提取关键信息。它已广泛应用于各种社会与视觉场景，例如事件精彩片段检测 [4], [5], 交通预测 [6], [7], 灾害监测 [8], [9], 和安全监控 [10], [11]。

通讯作者：H. Yu，电子邮箱：hui.yu@glasgow.ac.u.k. 本研究由 UKRI (EP/Z000025/1) 和欧洲地平线计划下的 MSCA 项目资助，用于支持 ACMOD 项目（资助编号：101130271）。

随着深度学习的发展，视频摘要方法已形成三类：监督方法 [12]、弱监督方法 [13], 和无监督方法 [14]。监督方法依赖大规模配对标注，引导模型识别显著内容。然而，人工标注过程成本高、耗时长，而且本质上具有主观性。弱监督方法通过引入先验知识或部分标签来提升泛化能力，从而降低标注成本，但仍会受到标签偏差的影响，并且难以捕捉视频语义的复杂性和多样性。相比之下，无监督方法因无需人工标签而日益受到关注。它们通过启发式策略生成摘要，其中对抗学习 [15]和强化学习 [16] 是应用最广泛的方法。

强化学习方法通过设计奖励函数提升摘要的代表性和多样性，但其效果在很大程度上取决于奖励设计，而且往往难以捕捉长程帧依赖关系。另一方面，生成对抗网络（GAN）采用对抗学习来驱动无监督视频摘要的序列生成模型。现有基于生成对抗网络的方法通常采用循环神经网络（例如长短期记忆网络） [17], [18], [19] 作为生成器，以捕捉多尺度时间依赖关系。尽管循环网络能够建模帧间的上下文关系，但在处理长期依赖关系和并行训练方面效率低下。此外，大多数现有方法仅关注单模态特征重建，忽视了视频中固有的视觉-语义相关性，因此无法有效地建模多模态交互。

对于一个完整的视频，其底层内容是唯一确定且保持不变的。因此，同一视频的不同模态（例如视觉信息和语义信息）天然传达相同的主题内容。换言之，

视频内的不同模态之间天然存在很强的关联。具体而言，同一帧的视觉特征与文本特征应当呈现确定性的映射，因此，原始视频与其跨模态摘要之间的数据分布也应相应地保持一致。为捕捉语义信息，我们利用具备零样本分类能力的 CLIP 编码器 [20],从视频帧中提取潜在语义特征。通过计算视觉特征序列与语义特征序列之间的余弦相似度，我们建立跨模态对齐，并对二者的对应关系进行量化。随后，将这一相似度分数视为帧选择器分配给每一帧的重要性分数，为摘要生成提供指导。受此思路启发，我们利用视频帧序列中的语义信号作为线索，驱动对抗训练，使模型能够生成最大程度表征原始视频内容的摘要。

因此，我们提出了一种语义引导的无监督视频摘要方法。如图 1 所示，我们首先分别采用卷积神经网络和 CLIP 编码器提取视觉特征和文本特征。随后，集成了帧级语义对齐注意力的帧选择器计算关键帧分数，以引导基于 Transformer 的生成器重构视频摘要。训练过程持续进行，直至判别器无法再区分真实视频与生成的摘要，从而实现语义引导的对抗学习。此外，我们采用增量训练策略逐步更新对抗组件，有效缓解生成对抗网络训练的不稳定性。

本工作的主要贡献总结如下:

1）我们设计了一种由帧级语义对齐注意力驱动的新型对抗学习框架，并引入增量训练策略，以有效稳定生成对抗网络训练。2）我们提出了一种新的注意力机制，通过跨模态余弦相似度量化关键帧的重要性，从而实现跨模态交互，并联合建模内在的视觉-语义相关性，进而提升摘要质量。3）我们设计了一种采用自编码架构的基于 Transformer 的生成器，用于捕获文本和视觉依赖关系，并实现多模态特征的自适应学习，以弥补单模态建模的局限性并生成高质量摘要。

## II. 相关工作

### A. 无监督视频摘要

无监督视频摘要方法 [21] 通常采用启发式策略进行视频重建，大致可分为基于强化学习（RL）和基于生成对抗网络（GAN）的方法。基于强化学习的方法依赖精心设计的奖励函数来指导关键帧选择。然而，这类方法的有效性在很大程度上取决于奖励设计，而奖励设计往往难以制定，且因任务而异。相比之下，基于生成对抗网络的方法通过对抗训练范式实现无监督学习。训练摘要生成器生成紧凑的摘要，同时由判别器尝试区分这些摘要与原始视频。在训练过程中，生成器逐渐改进，最终生成能够骗过判别器的高质量视频摘要。Mahasseni 等人 [22] 首次将对抗范式引入无监督视频摘要，通过最小化原始视频与生成摘要之间的分布差异，为这一研究方向奠定了基础。在此基础上，Apostolidis 等人 [23]提出在输入与摘要生成器之间加入线性压缩层以降低特征维度，并采用逐步标签学习策略来提升对抗训练效率。之后，他们进一步扩展了这一方法，提出 SUM-GAN-AAE 模型 [24], ，将注意力机制引入变分自编码器和自编码器层，显著增强了模型捕捉上下文依赖关系的能力。尽管基于生成对抗网络的无监督视频摘要 [25], [26] 已取得显著进展，但这类方法仍受生成对抗网络训练固有不稳定性及相关挑战的制约，因此摘要生成的性能仍有很大的提升空间。

## III. 提出的方法

### A. 技术框架概述

我们将视频摘要表述为一个结构化预测问题：给定一段连续视频帧序列作为输入，目标是输出一组能够有效表征原始视频的关键帧。如图 1 所示，首先使用卷积神经网络和 CLIP 编码器处理包含 $T$ 帧的视频序列，提取每一帧的视觉特征 $\{X_t\}_{t=1}^{T}$ 和语义特征 $\{T_t\}_{t=1}^{T}$ 。随后，将帧序列的多模态特征输入帧选择器，生成帧重要性分数序列 $\{S_t\}_{t=1}^{T}$。每个分数 $S_t$ 的取值范围为 $[0,1]$, ，数值越大表示帧 $t$ 在视频中的重要性越高。基于这些重要性分数，对每一帧的视觉特征和语义特征重新加权，得到 $W_t = S_t \cdot X_t$ 和 $V_t = S_t \cdot T_t$，然后将它们输入基于 Transformer 的生成器，生成重建后的摘要特征 $\{X_t\}$ 和 $\{T_t\}$。最后，将重建后的多模态摘要特征（$X_t$， $T_t$）与原始特征（$X_t$， $T_t$）共同输入判别器，使对抗训练框架确保生成的摘要与原始视频的语义和视觉分布高度相似。

### B. 语义对齐注意力帧选择器

帧选择器为每一帧分配重要性分数，以衡量其在整个视频中的代表性。根据这些分数，对帧的视觉特征和语义特征进行重新加权，从而形成新的序列特征分布。帧评分越准确，

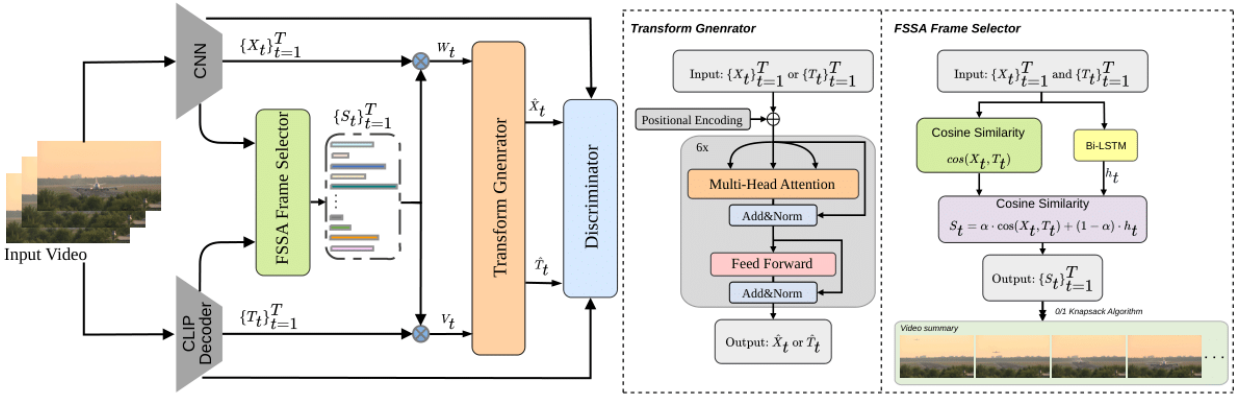


图 1。 所提出的无监督视频摘要框架的总体流程 **架构概览**。符号： ⊕表示拼接， ⊗ 表示逐元素乘法。 s

序列就越能保留最能代表原始视频内容的帧，从而帮助生成器重建质量更高的摘要。然而，大多数现有方法采用双向长短期记忆网络（双向长短期记忆网络） [27]来捕捉帧之间的时间依赖关系，并据此分配分数。这类方法通常仅依赖对视频序列视觉特征的建模，而忽略了语义信息对帧评分的贡献。

为解决这一局限性，我们提出了一种帧级语义对齐注意力（FSSA）机制。该机制将视频帧的视觉特征与语义特征之间的余弦相似度融入帧分数序列，使语义信息能够在评分过程中发挥引导作用，最终使生成的视频摘要能够更好地捕捉视频的核心内容。

形式化地，帧分数定义为

$$S_t = \alpha \cdot \cos(X_t, T_t) + (1-\alpha) \cdot h_t \tag{1}$$

其中， $S_t$ 表示帧 $t$的重要性分数，$\cos(X_t, T_t)$表示视觉特征 $X_t$ 与语义特征 $T_t$之间的跨模态余弦相似度，而 $h_t$ 是双向长短期记忆网络生成的分数。较高的 $S_t$ 表示当前帧抽象出的深层特征更能代表整体视频内容。

为充分发挥跨模态注意力在帧选择中的优势，我们引入一个可学习参数 $\alpha \in [0,1]$。训练过程中， $\alpha$ 会动态调整跨模态相似度分数与双向长短期记忆网络分数之间的相对权重，使模型能够针对不同的视频语境自适应地确定最优评分方案。

### C. 无监督学习策略

对于包含 $T$ 帧的视频，我们将序列的原始视觉和语义特征向量记为$\{X_t, T_t\}_{t=1}^T$，将重建特征记为 $\{\hat{X}_t, \hat{T}_t\}_{t=1}^T$。

训练过程分为三个阶段。首先，执行一次前向传播，计算 ( $X_t$， $T_t$) 与 ( $\hat{X}_t$， $\hat{T}_t$) 之间的重建损失，并相应地更新判别器。其次，在参数部分更新后，再次执行前向传播，计算 $T_t$ 与 $T_t$之间的 $\hat{}$ 差异，然后更新帧选择器。最后，进行第三次前向传播，将 $\hat{}$输出 $X_t$ 与 $X_t$进一步比较，并优化帧选择器和生成器，以最小化重建损失。

整体无监督训练目标结合了重建损失和稀疏损失。具体而言，重建损失定义为

$$L_{\text{rec}} = \mu \cdot \|X_t - \hat{X}_t\|_2^2 + \nu \cdot \|T_t - \hat{T}_t\|_2^2 + \text{SSIM}(X_t, \hat{X}_t) \tag{2}$$

$\hat{}\,\hat{}$ 其中， $X_t$ 和 $T_t$ 表示原始视频特征， $X_t$ 和 $T_t$表示重建特征， $\mu$ 和 $\nu$ 为超参数，SSIM 表示结构相似性损失 [28]。

为确保紧凑性并避免平凡解，我们进一步对帧重要性分数引入稀疏约束：

$$L_{\text{sparsity}} = \left| \frac{1}{N} \sum_{j=1}^{N} \left( S_t^j - \lambda \right) \right| \tag{3}$$

其中， $S_t^j$ 表示序列中第 $j$ 帧的重要性分数， $\lambda$ 为稀疏正则化超参数， $N$ 为帧总数。

总体训练损失结合了重建约束和稀疏约束，定义如下：

$$L_{\text{total}} = L_{\text{rec}} + L_{\text{sparsity}} \tag{4}$$

该损失共同确保特征重建的准确性，同时促进生成紧凑且信息丰富的视频摘要。

## IV. 实验

### A. 数据集与实验设置

我们在两个广泛使用的基准数据集 SumMe 和 TVSum 上评估了所提出的语义引导的无监督视频摘要模型。

SumMe [29] 包含 25 个视频，时长从 1 到 6 分钟不等，涵盖多视角社交事件。每个

视频由 15–18 位用户进行标注，他们通过选择关键镜头生成摘要；帧级重要性分数则以单一参考真实标注的形式提供。

TVSum [30] 由 50 个视频组成，每段视频时长为 1 至 11 分钟。每段视频由 20 位用户进行标注，他们按照 1–5 的等级为帧级重要性分数打分。摘要长度限制为视频总时长的 5%–15%。

使用 F1分数评估性能，其计算方式为精确率（$P$）和召回率（$R$）的调和平均值：

$$F1 = \frac{2PR}{P+R} \tag{5}$$

其中，自动生成的摘要（$A$）与用户生成的摘要（$G$）之间的重叠部分按帧级进行衡量。对于 TVSum，报告所有用户的平均 F1分数；对于 SumMe，则采用标注者中的最高 F1 分数。为确保满足摘要长度限制，我们采用背包算法，将摘要限制在视频时长的 15% 以内。

在实验设置方面，所有视频均以 2 fps 的帧率降采样。模型使用 PyTorch 实现，并在 NVIDIA GeForce RTX 3090 GPU 上训练。我们为判别器使用学习率 $\lambda = 0.001$，为其他模块使用学习率 $\lambda' = 0.002$，并采用 Adam优化器。每个数据集均随机划分为 80% 的训练集和 20% 的测试集，报告的结果为五次独立运行的平均值。

### *B.* 定量比较

为确保评估的公平性，我们采用两种评估框架来量化算法生成的摘要质量：一种框架利用每段视频的多位标注者摘要，另一种则采用单个参考摘要。如表I 和表II所示，我们的方法在 SumMe 和 TVSum 数据集上均取得了最佳表现。此外，我们发现，在单一参考摘要评估设置下获得的 F1分数始终高于多用户标注设置下的分数。一个可能的原因是，表I中的评估汇总了多位用户的标注，因此提供了更客观的基准。相比之下，表II中的评估依赖单个参考摘要，可能会受到标注者主观性的影响。这些结果凸显了采用多种评估指标的必要性，因为它们能够反映摘要性能的不同方面，从而更全面地评估模型的有效性。

总体而言，我们的方法在两个基准数据集上均取得了最高的平均值 F1分数，比排名第二的结果高出 3.2 个百分点。这种持续的提升证明了我们的方法能够有效捕捉视频中的关键信息。此外，结果显示，TVSum 数据集上的 F1分数始终高于 SumMe。一个可能的解释是，TVSum 包含更多时长更长、内容更多样的视频。这些特征提供了更丰富的语义信息，使我们的模型能够学习到更优且泛化能力更强的特征。

表 I 使用多个用户标注的摘要方法与其他无监督方法进行比较。（“±”表示性能优于或劣于我们的方法。）

| 方法 | SumMe | TVSum | 平均值 |
|---|---|---|---|
| 镶嵌 | 41.4(-) | 64.1(-) | 52.8(-) |
| DR-DSN | 41.4(-) | 57.6(-) | 49.5(-) |
| DSR-RL-LSTM | 43.8(-) | 61.4(-) | 52.6(-) |
| DSR-RL-GRU | 50.3(-) | 60.2(-) | 55.3(-) |
| 无配对视觉状态网络 | 47.5(-) | 55.6(-) | 51.6(-) |
| SUM-Ind LU | 51.9(-) | 61.5(-) | 56.7(-) |
| AC-SUM-GAN | 50.8(-) | 60.6(-) | 55.7(-) |
| CAAN | 50.8(-) | 59.6(-) | 55.2(-) |
| CSNet | 51.3(-) | 58.8(-) | 55.1(-) |
| Cycle-SUM | 41.9(-) | 57.6(-) | 49.8(-) |
| RS-SUM | 52.0(-) | 61.1(-) | 56.6(-) |
| SUM-GDA | 50.0(-) | 59.6(-) | 54.8(-) |
| SUM-GAN-sl | 47.3(-) | 58.0(-) | 52.7(-) |
| SUM-GAN-AAE | 48.9(-) | 58.3(-) | 53.6(-) |
| SUM-GAN-GEA | 53.4(-) | 61.3(-) | 57.4(-) |
| DMFF | 51.8(-) | 61.0(-) | 56.4(-) |
| PRLVS | 46.3(-) | 63.0(-) | 54.7(-) |
| SSPVS | 48.7(-) | 60.3(-) | 54.5(-) |
| M3Sum(SP) | 43.6(-) | 56.9(-) | 50.2(-) |
| AMFM | 51.8(-) | 61.0(-) | 56.4(-) |
| Ours | **55.9** | **65.3** | **60.6** |

表 II 使用单一真实摘要与其他视频摘要方法的比较（星号表示无监督方法，“±”表示性能优于或劣于我们的方法）。

| 方法 | SumMe | TVSum |
|---|---|---|
| SASUM | 45.3(-) | 58.2(-) |
| DTR - 生成对抗网络 | 44.6(-) | 59.1(-) |
| A - AVS | 43.9(-) | 59.4(-) |
| M - AVS | 44.4(-) | 61.0(-) |
| AALVS | 46.2(-) | 63.6(-) |
| *DMFF | 62.7(-) | 65.3(-) |
| *Cycle - 求和 | 41.9(-) | 57.6(-) |
| *SUM - GAN - sl | 46.8(-) | 65.3(-) |
| *SUM - GAN - AAE | 56.9(-) | 63.9(-) |
| *SUM - GAN - GEA | 65.9(-) | 68.2(-) |
| *AC - SUM - 生成对抗网络 | 60.7(-) | 64.8(-) |
| *SUM - 生成对抗网络 | 38.7(-) | 50.8(-) |
| *SUM - 生成对抗网络$_{dpp}$ | 39.1(-) | 51.7(-) |
| *SUM - 生成对抗网络$_{sup}$ | 41.7(-) | 56.3(-) |
| *Cycle - 求和$_{dpp}$ | 42.3(-) | 57.9(-) |
| *Cycle - 求和$_{sup}$ | 44.8(-) | 58.1(-) |
| *我们的方法 | **63.7** | **70.1** |

### *C.* 定性比较

图2 展示了我们的模型在两个基准数据集上生成的摘要的视觉示例。在 SumMe 数据集中，第一个示例描绘了一架飞机着陆；在 TVSum 数据集中，第 15 个示例展示了使用一种产品治疗狗的耳朵。这些具有代表性的案例凸显了不同数据集间视频内容的多样性，为评估摘要质量提供了有意义的依据。图的 (a) 部分展示了由人类受试者标注的真实摘要，以及关键帧得分分布直方图。(b) 部分展示了我们的模型生成的摘要及其对应的预测得分分布。二者对比可见，

模型选取的大多数关键帧都与真实分布中观察到的峰值高度吻合。这种吻合表明，我们的方法不仅能够识别最显著的时刻，还能捕捉事件的语义连续性。因此，生成的摘要更忠实地反映了人类主观感知，表明该模型兼顾了信息丰富度与连贯性。

### *D.* 消融研究

表III报告了不同模型变体在 SumMe 和 TVSum 数据集上的表现。结果清楚地表明，所提出的方法显著优于基于生成对抗网络（GAN）的传统无监督视频摘要方法。通过比较可以发现，FSSA 帧选择器有效促进了跨模态信息交互。它利用语义-视觉对齐所导出的帧级注意力分布，引导重建出更具代表性、质量更高的摘要。此外，基于 Transformer 的生成器进一步增强了跨模态感知能力，并实现了细粒度摘要重建，这一点在表 III所报告的性能提升中得到了持续体现。Exp5 的结果不太理想，其表现与采用单一策略时相当。不过，这一结果有力地证明了语义引导在我们的方法中的重要性。具体而言，具有语义对齐注意力的帧选择器是我们无监督框架的核心，对有效的视频重建起着决定性作用。

表 III 消融研究。

| 实验 | 多模态 | FSCA | Transform Gnenrator | SumMe | TVSum |
|---|---|---|---|---|---|
| SUM - GAN - AAE | | | | 48.9 | 58.3 |
| Exp1 | ✓ | | | 49.3 | 59.6 |
| Exp2 | | ✓ | | 52.7 | 62.4 |
| Exp3 | | | ✓ | 51.5 | 60.3 |
| Exp4 | ✓ | ✓ | | 52.4 | 63.6 |
| Exp5 | ✓ | | ✓ | 51.9 | 59.1 |
| Exp6 | | ✓ | ✓ | 53.9 | 64.3 |
| Ours | ✓ | ✓ | ✓ | 55.9 | 65.3 |

## 五、结论与未来工作

本文提出了一种新颖的无监督视频摘要框架，称为语义引导的无监督视频摘要。我们方法的核心思想是，计算同一视频序列中帧级表示之间的跨模态余弦相似度，并将该相似度分布作为语义线索来引导摘要生成。整个框架采用增量优化策略进行训练，从而确保对抗学习稳定且有效。在基准数据集上开展的大量实验表明，与现有方法相比，所提出的方法效果显著且具有优越性。

在未来的工作中，我们计划将音频特征纳入框架，以进一步提升摘要质量。通过探索更丰富的多模态特征表示，我们希望将语义引导的视频摘要应用于更多样化的真实场景。

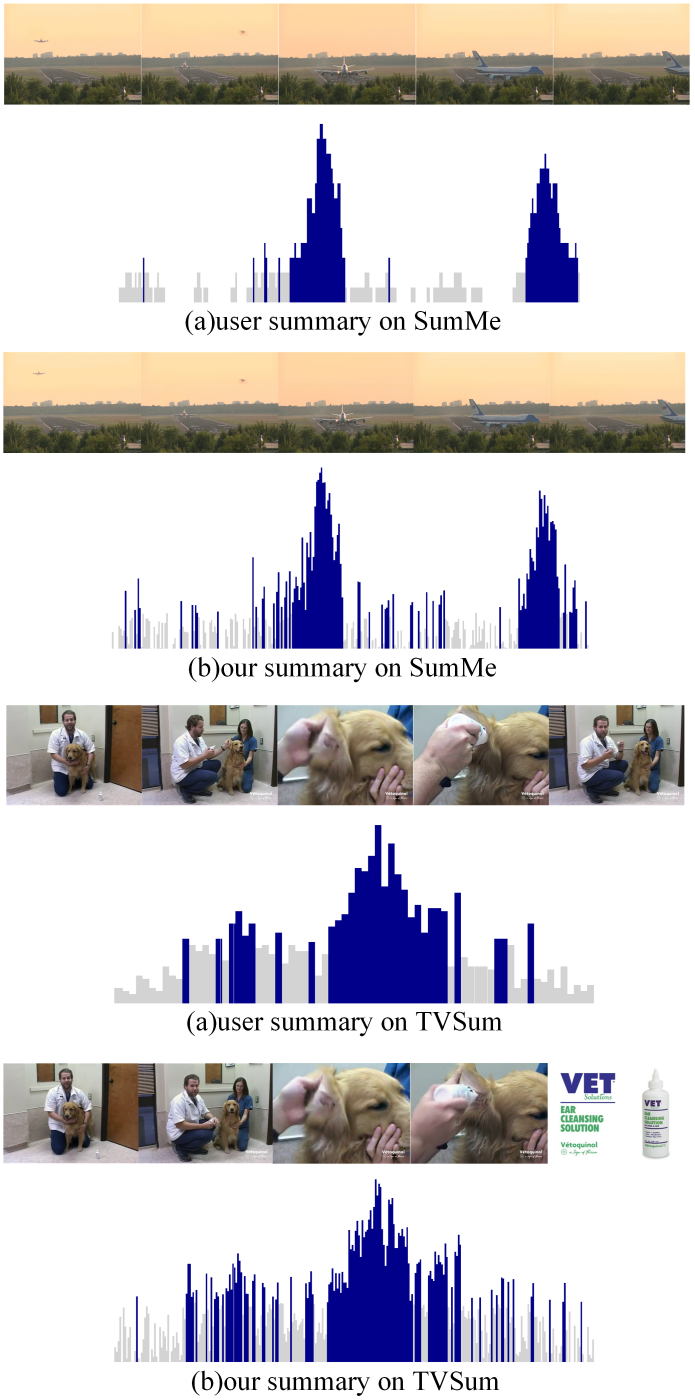


图2. 所提方法生成的部分摘要可视化结果。灰色表示真实标注，蓝色表示模型预测的摘要结果，柱子的高度表示帧的重要性。左侧的视频摘要来自 SumMe 数据集中的第一个视频，右侧的则来自 TVSum 数据集中的第15个视频。

## 参考文献

[1] H. Li、Y. Zhu、Z. Shang、Z. Wang 和 X. Wu，《视频摘要综述：挑战与进展》，*IEEE*汇刊

电路与系统视频技术，第 1–1 页，2025年。[2] Y. Tang、J. Bi、S. Xu、L. Song、S. Liang、T. Wang、D. Zhang、J. An、J. Lin、R. Zhu、A. Vosoughi、C. Huang、Z. Zhang、P. Liu、M. Feng、F. Zheng、J. Zhang、P. Luo、J. Luo 和 C. Xu，《大语言模型视频理解：综述》，*IEEE*《电路与系统视频技术汇刊》，第 1–1 页，2025年。[3] C. Ye、W. Chen、B. Hu、L. Zhang、Y. Zhang 和 Z. Mao，《通过探索对应字幕之间的连贯性改进视频摘要》，*IEEE*《图像处理汇刊》，第 34 卷，第 5369–5384 页，2025年。[4] T. Li、Z. Sun 和 X. Xiao，《利用表征激活序列学习进行无监督模态可迁移视频高光检测》，*IEEE*《图像处理汇刊》，2024年。[5] R. Wang、J. Feng、F. Zhang、X. Luo 和 Y. Luo，《用于联合视频片段检索与高光检测的模态感知异质图》，*IEEE*《电路与系统视频技术汇刊，2024年。[6] Y. Lu、H. Ma、E. Smart 和 H. Yu，《增强自动驾驶决策：面向复杂高速公路场景的混合深度强化学习与运动学自动驾驶框架》，*IEEE*《智能交通系统汇刊》，第 26 卷，第 3 期，第 3198–3209 页，2025年。[7] L. Fan、Y. Wang、H. Zhang、C. Zeng、Y. Li、C. Gou 和 H. Yu，《基于基础模型的复杂道路多模态感知与决策系统》，*IEEE*《系统、人类与控制论：系统汇刊》，第 54 卷，第 11 期，第 6561–6569 页，2024年。[8] K. Shi、X. Peng、H. Lu、Y. Zhu 和 Z. Niu，《社会传感器在自然灾害应急管理中的应用：综述》，*IEEE*《计算社会系统汇刊》，第 10 卷，第 6 期，第 3143–3158 页，2022年。[9] R. Qin、F.-Y. Wang、X. Zheng、Q. Ni、J. Li、X. Xue 和 B. Hu，《面向计算社会系统的 Sora：从反事实实验到平行智能下的人工事实实验》，*IEEE*《计算社会系统汇刊，第 11 卷，1550，2024 2，1531 无. 第 – 页 .[10] A. Goyal、M. Mandal、V. Hassija、M. Aloqaily 和 V. Chamola，《Captionomaly：用于社会监控系统异常描述的深度学习工具箱》，*IEEE*《计算社会系统汇刊》，第 11 卷，第 1 期，第 207–215 页，2023年。[11] D. Li、X. Nie、R. Gong、X. Lin 和 H. Yu，《基于上下文学习的监控视频多分支生成对抗网络异常事件检测》，*IEEE*《电路与系统视频技术汇刊》，第 34 卷，第 5 期，第 3439–3450 页，2024年。[12] W. Zhu、Y. Han、J. Lu 和 J. Zhou，《视频摘要中的时空图关系推理》，*IEEE*《图像处理汇刊》，第 31 卷，第 3017–3031 页，2022年。[13] Z. Hu、A. Sun、Z. Wang、J. Li、Z. Song、R. Hong 和 M. Wang，《自适应双重视频摘要：从动态关键帧到字幕》，*IEEE*多媒体汇刊，第 1–12 页，2025年。[14] X. Cheng、L. Yang 和 R. Li，《基于时空语义图和增强注意力机制的无监督视频摘要》，*IEEE*《计算社会系统汇刊》，第 1–14 页，2025年。[15] Q. Yu、H. Yu、Y. Sun、D. Ding 和 M. 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》，第 11 卷，第 5 期，第 6010–6021 页，2024年。[16] T. Liu、Q. Meng、J.-J. Huang、A. Vlontzos、D. Rueckert 和 B. Kainz，《通过强化学习与 3D 时空 U-Net 进行视频摘要》，*IEEE*《图像处理汇刊》，第 31 卷，第 1573–1586 页，2022年。[17] Y. Zhang、X. Liang、D. Zhang、M. Tan 和 E. P. Xing，《利用在线运动自编码器进行无监督对象级视频摘要》，《模式识别快报》，第 130 卷，第 376–385 页，2020年。[18] T. Hussain、K. Muhammad、A. Ullah、Z. Cao、S. W. Baik 和 V. H. C. de Albuquerque，《利用卷积神经网络和双向长短期记忆网络进行云辅助多视图视频摘要》，*IEEE*《工业信息学汇刊，第 16 卷，第 1 期，第 77–86 页，2020年。[19] Y. Xu、Z. Wu、L. Li、S. Li、W. Li、M. Li、Y. Rao 和 S. Deng，《混合孪生掩码自编码器：无监督视频摘要器》，*IEEE*《电路与系统视频技术汇刊》，第 1 1 页，2025 –。[20] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等，《从自然语言监督中学习可迁移视觉模型》，载于国际机器学习会议。PmLR，2021年，第 8748–8763 页。

[21] S. Zang、H. Jin、Q. Yu、S. Zhang 和 H. Yu，《使用 U 形非局部网络进行视频摘要》，*International Journal of Network Dynamics and Intelligence*，第 100 013–100 013 页，2024年。[22] B. Mahasseni、M. Lam 和 S. Todorovic，《利用对抗式长短期记忆网络进行无监督视频摘要》，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，第 202–211 页。[23] E. Apostolidis、A. I. Metsai、E. Adamantidou、V. Mezaris 和 I. Patras，《一种用于改进无监督视频摘要中对抗训练的逐步式基于标签的方法》，载于第*1*届人工智能赋能智能电视内容制作、访问与交付国际研讨会论文集，2019年，第 17–25 页。[24] E. Apostolidis、E. Adamantidou、A. I. Metsai、V. Mezaris 和 I. Patras，《通过注意力驱动的对抗学习进行无监督视频摘要》，载于多媒体建模：第*26*届国际会议 *MMM 2020*，韩国大田，*2020*年*1*月*5*日至*8*日，论文集，第一部分 *26*。施普林格，2020年，第 492–504 页。[25] Y. Xu、X. Li、L. Pan、W. Sang、P. Wei 和 L. Zhu，《具有上下文潜在序列学习的自监督对抗式视频摘要器》，*IEEE*《电路与系统视频技术汇刊》，第 33 卷，第 8 期，第 4122–4136 页，2023年。[26] E. Apostolidis、E. Adamantidou、A. I. Metsai、V. Mezaris 和 I. Patras，《AC-SUM-GAN：将演员-评论家与生成对抗网络相结合以实现无监督视频摘要》，*IEEE*《电路与系统视频技术汇刊》，第 31 卷，第 8 期，第 3278–3292 页，2021年。[27] L. Yuan、F. E. H. Tay、P. Li 和 J. Feng，《利用循环一致性对抗式长短期记忆网络进行无监督视频摘要》，*IEEE*多媒体汇刊，第 22 卷，第 10 期，第 2711–2722 页，2020年。[28] T. Zhao、J. Wang、Z. Wang 和 C. W. Chen，《基于结构相似性的粗粒度可伸缩视频编码》，*IEEE*广播汇刊，第 61 卷，第 2 期，第 210–221 页，2015年。[29] M. Gygli、H. Grabner、H. Riemenschneider 和 L. Van Gool，《从用户视频创建摘要》，载于计算机视觉—*ECCV 2014*：第*13*届欧洲会议，瑞士苏黎世，*2014*年*9*月*6*日至*12*日，论文集，第七部分 *13*。施普林格，2014年，第 505–520 页。[30] Y. Song、J. Vallmitjana、A. Stent 和 A. Jaimes，《TVSum：使用标题概述网络视频》，载于*IEEE*计算机视觉与模式识别会议论文集，2015年，第 5179–5187 页。