arXiv:2304.12561v1 [cs.CV] 25 Apr 2023

TCR：基于注意力优化的短视频标题生成与封面选择

Yakun Yu[1⋆]、Jiuding Yang[1⋆]、Weidong Guo[2⋆]、Hui Liu[2], 、Yu Xu[2], 和 Di Niu[1]

[1] 加拿大埃德蒙顿阿尔伯塔大学{yakun2、
jiuding、dniu}@ualberta.ca[2] 中国深圳腾讯

摘要。 随着用户生成短视频的广泛普及，内容创作者越来越难以将其内容推广给潜在观众。自动为短视频生成具有吸引力的标题和封面，有助于吸引观众的注意力。现有的视频描述生成研究大多专注于生成对动作的事实性描述，这与旨在吸引观众注意力的视频标题并不相符。此外，基于多模态信息进行封面选择的研究也较为稀缺。这些问题促使人们需要专门的方法来支持短视频标题生成与封面选择（TG-CS）这一联合任务，同时也需要创建相应的数据集以支持相关研究。在本文中，我们首先收集并提出了一个名为短视频标题生成（SVTG）的真实世界数据集，其中包含具有吸引力的标题和封面的视频。随后，我们提出了一种用于 TG-CS 的基于注意力优化的标题生成与封面选择（TCR）方法。该优化过程会逐步选择高质量样本，以及每个样本中高度相关的帧和文本词元，以优化模型训练。大量实验表明，在生成标题方面，我们的 TCR方法优于各种现有的视频描述生成方法，并且能够为包含噪声的真实世界短视频选择更优的封面。

关键词： 标题生成 · 封面选择 · 多模态学习。

1 引言

视频标题和视频封面是用户使用应用程序（例如 TikTok）时吸引其注意力的两个关键要素。然而，许多短视频没有经过精心编辑的标题，只有简单的描述，或者没有除首帧以外的封面图像。每日上传的视频数量庞大，也进一步导致平台难以及时编辑视频封面。因此，显然有必要利用人工智能提高视频标题和封面图像选择的质量，尤其是为了提升用户参与度。

近年来，人们研究并开发了各种视频描述生成技术，包括视频字幕生成 [5,15], 基于视频的摘要生成 [6,7],

⋆ 贡献相同。

等等。这些技术主要用于客观描述视频中物体的动作和互动。例如，对于一段女子练瑜伽的视频，生成的字幕通常是“女子正在锻炼，而她的猫一直在打扰她”，这种描述虽然准确，却平淡无奇。然而，视频上传者更愿意用幽默的方式将这个场景命名为“我的猫把我当成了他的玩具”，以吸引观众的注意。因此，现有研究在生成引人入胜、贴近人类表达的视频标题方面仍有不足。然而，现实社交媒体平台上的开放短视频数据集匮乏，阻碍了贴近人类表达的视频标题生成技术的发展。此外，关于视频封面选择的研究也很少 [12], ，这同样是一个挑战，原因仍在于缺少视频封面选择数据集。

为了解决上述问题，我们收集了一个短视频标题生成（SVTG）数据集，其中包含 8,652 个短视频样本，每个样本都配有经人工编辑的吸引人标题，以及由原内容创作者选定的视觉封面。据我们所知，SVTG 是首个公开可用的数据集，旨在针对真实世界中的短视频联合进行视频标题生成与封面选择（TG-CS），尤其着眼于提升点击率，而不是简单描述场景。

此外，我们提出了一种新颖的结合精炼的视频标题生成与封面选择（TCR）方法。该方法基于短视频中的多模态信息，将标题封面生成器与基于交叉注意力的精炼学习相结合，用于 TG-CS 任务。具体而言，我们采用多模态标题-封面生成器来捕捉模态之间的时间依赖关系。为处理从现实生活短视频中提取的噪声模态数据，我们进一步提出了一种精炼学习策略，以细化模型训练。

我们的主要贡献总结如下：(1) 我们构建了一个新的短视频数据集 SVTG，可推动短视频标题生成研究超越对对象进行不加修饰的事实描述，并促进基于多模态信息的视频封面选择研究。我们的目标是生成引人注目的视频标题和封面，以吸引在线用户的注意力；(2) 我们提出了 TCR 方法，通过充分捕捉模态间的依赖关系并采用细化训练，同时生成吸引人的标题和封面；(3) 我们在 SVTG 数据集和公开的 How2 数据集上评估了 TCR。结果表明，TCR 在两个数据集上的标题生成任务以及 SVTG 上的封面选择任务中均表现有效。我们将开放数据集和代码，以推动未来针对真实社交媒体视频的标题生成与封面选择研究。

2 相关工作

2.1 文本生成

文本生成主要包括文本摘要和视频字幕生成。前者旨在为长文档生成简短摘要，同时保留文档的主要思想。后者侧重于使用精炼的文本描述视频的视觉内容。

表 1. SVTG数据集的统计信息。

	#视频数	#类别	#作者	#句子数	词汇表
训练集	8,052	37	3,478	155,826	100,361
验证集	200	25	172	3,802	6,770
Test	400	27	288	7,352	10,494
All	8,652	37	3,618	166,980	105,487

传统的文本摘要利用单一的文本模态来生成摘要 [10,16]。近期研究 [6,7] 借助其他模态（例如视觉模态）在摘要任务上取得了令人鼓舞的结果。例如，Liu 等 [9] 提出了基于循环神经网络和基于 Transformer 的编码器-解码器模型，其输入为视频及其真实标注/自动语音识别转录文本。编码器-解码器网络广泛用于视频字幕生成，其中编码器先将输入转换为高级上下文表示，然后解码器接收这些表示并生成作为字幕的词语。例如，Tan 等[15] 提出了一个基于循环神经网络的编码器-解码器模型，并结合三个视觉推理模块来生成视频描述。

上述方法在给定视频及其经过良好编辑的相应转录文本时，能够很好地生成视频场景的事实性描述，但当它们试图基于含噪的多模态信息生成超越事实性描述的标题时则表现不佳。相比之下，我们的方法能够处理含噪的模态数据，并捕捉模态之间的细微关系，因此既能生成事实性描述，也能生成吸引人的标题。

2.2 封面选择

封面选择旨在找到能够传达视频主旨的代表性帧。Song等人 [14] 提出了一种采用聚类方法的缩略图选择系统。Ren等人 [12] 根据预测分数选择最佳帧。他们仅利用视觉模态，而对于真实世界短视频而言，这远远不够，因为关键帧应由多种模态共同确定。尽管 Li等人 [7] 提出了 VMSMO，利用多模态信息（即新闻视频与新闻文章配对）训练封面选择模型，但该模型的性能高度依赖于经过人工精心编辑的新闻文章质量。然而，SVTG数据集不提供此类人工创建的文章。因此，将他们的方法用于我们的任务进行比较并不合适。此外，我们使用统一模型而非独立的封面选择模型来提取封面，这种做法更直接。

3 SVTG数据集

据我们所知，SVTG 是首个公开可用的中文数据集，专为短视频的联合标题生成与封面选择而设计。我们认为，如果标题具有吸引力，



图1。（a）SVTG数据示例与（b）其他基准数据集示例的比较。SVTG示例顶部为连续帧，中部为标题，底部为通过自动语音识别/光学字符识别提取的转录文本。其他示例展示了其真实标注（GT）和用户生成的描述。

就意味着它除了描述事实信息（例如物体间交互）之外，还表达了视频创作者通过一些情感词吸引观众的意愿。SVTG数据收集见补充材料 A。表 1 展示了我们数据集的统计信息。

为展示 SVTG 的独特性，我们将其与另外四个基准数据集进行了比较，如图1所示。与这些数据集相比，SVTG 具有三个独特特征。首先，SVTG 视频拥有更具吸引力的标题，这能够推动深度学习模型的研究，使其生成超越简单事实性描述的短视频标题。正如我们在图1(a)中观察到的，SVTG 视频通常包含强烈的情感词（例如“必须”和“营养丰富”），以吸引观众的注意力，而不仅仅是描述视频场景。相比之下，从图1(b)中可以看出，其他基准数据集中的字幕、摘要和标题通常强调物体动作（例如“站立”）以及物体间交互（例如“男子使用酒吧”）。这些描述往往朴实客观，不带任何情感色彩。如果我们按照其他数据集的标注方式对图1(a)中的视频进行标注，标题听起来会像“倒入牛奶并加入水果制作蛋糕”，对在线观众的吸引力也会降低。

其次，SVTG 的视频标题生成更具挑战性，也更贴近现实生活视频。如图1(b)所示，现有基准数据集将人工标注的视频描述、字幕或文章作为文本模态。然而，在现实世界应用中，获得这类人工标注的视频描述成本高昂，甚至不可行。在现实世界场景中，当内容创作者将短视频上传至平台时，平台需要根据视频本身选择封面并生成标题。SVTG 中的所有信息均来自视频本身，无需其他人工标注。因此，SVTG 更

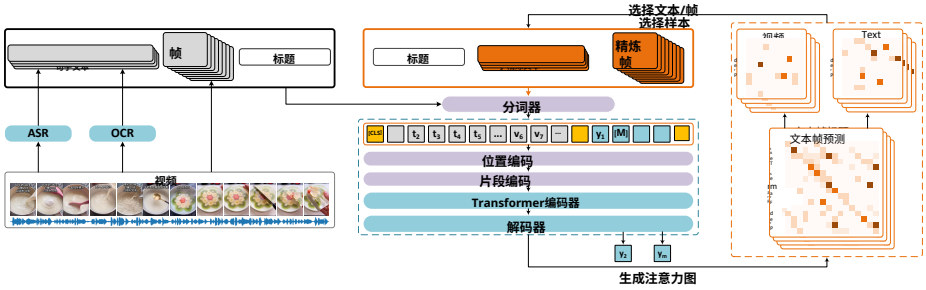

图 2. 所提出的 TCR 方法的整体架构，包括黑框中的数据预处理、蓝色虚线框中的标题封面生成器，以及橙色虚线框中的基于多模态交叉注意力的细化模块。“Title”表示真实标题。注意力图中的“Pred”表示预测标题。分词序列中的“[M]”表示掩码标记。

贴近现实世界应用。然而，与人工标注的摘要或描述相比，自动语音识别和光学字符识别检测到的文本可能包含大量噪声，使得 SVTG 比其他数据集更具挑战性。

第三，据我们所知，SVTG 是首个包含视频封面、用于促进短视频联合标题生成与封面选择的数据集。我们认为，标题生成和封面选择这两项任务相结合，对于吸引观众注意力至关重要，因此不应将二者分开考虑。此外，统一的数据集有助于鼓励模型学习这一联合任务中的共享关联。有关这些基准数据集的更多详情，请参见补充材料 B。

4 方法

我们提出了一种名为 TCR 的自注意力方法，用于在 SVTG 数据集上进行联合标题生成与封面选择。图 2 展示了 TCR 方法的结构。它由一个多模态标题-封面生成器和一个细化模块组成。生成器通过充分利用不同模态之间的依赖关系来生成标题并提取封面。细化模块通过筛选高质量样本以及每个样本中高度相关的词元和帧，帮助优化模型训练。下面我们将详细介绍 TCR 方法。

4.1 标题生成与封面选择

由于帧和文本输入都具有时间顺序，我们可以直接将二者结合起来，显式学习所有文本词元与帧之间的依赖关系。我们首先将训练数据预处理为以下格式：

– 文本表示：给定从视频中通过自动语音识别和光学字符识别提取的转录文本，将自动语音识别和光学字符识别文本合并后，分词为由以下词片组成的序列：

词片 [17],，即 $T = (t_1, t_2, ..., t_n)$，其中 n 是序列中的词元数量。– 标题表示：标题表示为由词元组成的序列 $Y = (y_1, y_2, ..., y_m)$，其中 m 表示 Y 中的词元数量。– 视频表示：给定长度为 L 的视频帧序列，将其输入预训练 ResNet-101 二维卷积神经网络 [3] 以获取视觉特征 $V = \{v_l\}_{l=1}^{L} \in \mathbb{R}^{L \times d_v}$，其中 d_v 是预训练 ResNet 模型的特征维度。

标题封面生成器由三个嵌入层、12层 Transformer 编码器以及顶部作为解码器的语言建模头组成。一方面，我们将帧视为 L 虚拟词元 $\{v_i\}_{i=1}^{L}$，并将其与文本词元 T 拼接，形成长度为 $L+n$ 的新序列。我们还在序列开头添加特殊词元 [CLS]，并在序列末尾添加另一个特殊词元 [SEP]。这些词元（包括 Y）由三个可训练的嵌入层进行嵌入：词元级嵌入、位置级嵌入和片段级嵌入。

另一方面，我们将帧特征 V 输入全连接（FC）层，将其投影到与文本词元嵌入相同的低维空间中，并将投影后的帧特征视为帧词元嵌入。为了建立文本词元与帧之间的联系，使它们能够相互关注，我们通过简单地用帧嵌入替换虚拟词元嵌入来结合各自的嵌入。因此，文本词元和视频帧的最终表示可以通过将其对应的三个嵌入相加而获得。随后，我们将这些表示作为 Transformer 编码器的输入。

假设输入为 $X_{VT} \in \mathbb{R}^{L' \times d_h}$，其中 d_h 是我们模型的隐藏维度，第个 Transformer 编码器的输出记为 $O_i \in \mathbb{R}^{L' \times d_h}$。每个编码器通过多个自注意力头聚合前一个编码器的输出，因此 O_i 的计算方式为：

$$O_i = \begin{cases} X_{VT}, & i = 0; \\ \text{Encoder}_i(O_{i-1}), & i > 0. \end{cases}$$

通过这种方式，我们可以获得不同编码器层的视觉语言融合表示。令 A_i 表示i号编码器中每个注意力头的注意力分数。

$$\begin{gathered} A_i = \text{Dropout}(\text{Softmax}(\frac{Q_i K_i^T}{\sqrt{d_h/12}} + \text{Attention Mask})), \\ Q_i,\ K_i = O_{i-1} W_{Q_i},\ O_{i-1} W_{K_i}. \end{gathered} \tag{1}$$

Q 和 K 是从各编码器输入经线性变换得到的查询和键。注意力掩码 $\in \mathbb{R}^{L' \times L'}$（元素 $\in \{0, -\infty\}$）用于确保待生成的目标词元只关注左侧信息，包括其自身、已生成的词元、帧和文本词元 [2]。语言模型头由两个稠密层组成，中间夹有一个层归一化层（LN），用于

计算用于解码的词汇分布。生成器训练完成后，我们选择根据公式 1 得到的注意力分数最高的帧作为视频封面。

4.2 基于注意力的模型优化

考虑到自动语音识别/光学字符识别转录文本和连续帧中可能存在冗余噪声，进行噪声过滤显然有助于提升生成标题的质量。为此，我们提出了一种基于注意力的优化模块，它不仅能够自动选择每个样本中多模态输入的关键部分，还能筛选出质量更高的样本。此外，该模块可以与上述生成器交互，找到不同输入词元的最佳组合，逐步优化生成器。

词元级优化。对于标题封面生成器，我们可以先为训练数据生成标题。随后，我们通过公式 1 获取生成标题与输入（即 ASR/OCR词元和帧词元）之间的交叉注意力分数。我们假设生成标题重点关注的词元对于构建性能良好的生成器至关重要。因此，我们将这些词元提取为优化后的句子/帧。对于句子优化，我们首先为每个生成词元定位注意力分数最高的句子。然后，自然会将生成标题频繁关注的排名前 u 的句子选为关键句子。

视频标题通常包含一些词语（例如篮球比赛视频中的“空中接力”），这些词语经常出现在 ASR/OCR文本中，却很少与视频帧相关。因此，与选择相关句子的方式相比，我们需要以更细的粒度选择相关帧。为此，我们首先根据每个帧与生成标题中各词元之间的注意力分数，为每个帧分配一个 $weight \in [0, L]$，然后计算每个帧的总权重。权重越高，表示对标题生成的贡献越大。最后，我们选出权重最高的前 v 帧，作为细化后的视频帧或关键帧。

样本级优化。 除了优化每个样本中的词元外，我们还通过在样本级别选择质量更高的数据来对数据进行去噪。如果生成器M 生成的标题与其真实标题相似，则该训练数据会被视为高质量数据。相似度通过 Rouge 分数 [8]计算。

显然，通过在词元级和样本级进行细化，我们可以多次优化标题封面生成器，直到获得最佳测试集结果。详细的细化算法见补充材料 C。

5 实验结果

5.1 基线模型

我们将所提出的方法与以下基线模型进行比较：

Lead-3 直接选取文本中的前三个句子作为标题 [10]。

HSG 构建一个包含句子节点和词节点的异构图，然后通过节点分类选择句子作为文档摘要 [16]。

MAST 是一个由编码器层、分层注意力层和解码器组成的模型，可基于多模态输入生成文本摘要 [6]。

神经机器翻译 是一种基于循环神经网络的模型 [1] ，旨在执行序列到序列任务，并针对视频描述任务进行了修改。

MFN通过多阶段融合模型，基于视频及其自动语音识别或真实转录文本生成摘要；该模型还包含一个用于移除冗余信息的遗忘门模块 [9]。

Bert2Bert 包含文本BERT编码器、视频Transformer编码器和BERT解码器，是近期视频/图像 +语言研究中常用的框架 [4]。

5.2 实现细节

我们按照表 1 中的划分标准，在 SVTG数据集上评估上述基线模型和我们的 TCR 模型。我们使用 PyTorch [11] 在一块 NVIDIA V100 GPU 上进行实验。为预处理视频，我们为每个视频采样 25 帧，并从这些采样帧中提取视觉特征。

我们的多模态生成器改编自一个文本预训练模型架构 [2]。词汇表大小为 21,128。我们使用 AdamW优化器，学习率为 $1e-5$。前 10% 的总训练步数采用线性预热调度，之后进行线性衰减。我们采用随机失活进行正则化。批量大小设为 16，最大输入序列长度设为 512。训练期间，我们仅对 20% 的标题词元进行掩码，掩码标记的最大数量设为 20。我们最多训练 20 个轮次，并根据预测标题词元和真实词元计算交叉熵损失。我们使用验证集选择评估指标平均分最高的模型。测试时，我们采用束搜索，束宽设为 5，并报告最终测试结果。细化时，我们选取排名前 3 的句子/帧，并进行一次迭代。

此外，我们在公开的 How2数据集[13] 上评估 TCR，以进一步验证其生成能力。How2数据集包含 79,114 个视频，以及相应的用户生成的描述和摘要。该数据集的训练集包含 73,993 个样本，验证集包含 2,965 个样本，测试集包含 2,156 个样本。由于 TCR 方法旨在处理带有噪声的多模态信息的真实世界短视频，为了进行公平比较，我们按照 Liu等人 [9] 的工作，使用自动语音识别转录文本，而非提供的描述。

5.3 结果

首先，我们报告基线模型和所提出的 TCR 在标题生成方面的表现。接着展示封面选择的评估结果。随后，我们从不同角度呈现消融测试。

表2。 SVTG 与基线方法的 ROUGE 分数比较。

模型	R-1	R-2	R-L
Lead-3	29.44	18.40	37.35
HSG	41.06	27.45	40.59
NMT	41.20	28.74	41.91
MFN-rnn	16.64	1.47	14.36
MAST	39.73	27.42	41.00
MFN-Transformer	19.24	3.99	18.40
Bert2Bert	24.84	5.75	22.58
去除文本模态的TCR	8.53	0.31	9.30
去除视觉模态的TCR	46.35	32.67	46.07
TCR	**47.62**	**33.98**	**47.08**

标题生成表现。按照以往研究的做法 [8]，我们采用标准的 Rouge F1 指标评估生成标题的质量 [6, 7, 9, 10,16]。ROUGE-1、ROUGE-2 和 ROUGE-L 分别表示一元词组、二元词组和最长公共子序列与真实标题的重叠程度。

我们首先比较了我们的方法与基线模型在SVTG上的结果，如表2所示。从表中可以观察到，在所有Rouge分数方面，TCR均优于所有基线模型。具体而言，TCR在ROUGE-1分数、ROUGE-2分数和ROUGE-L分数上分别比最佳基线模型高出6.42个R-1分数、5.24个R-2分数和5.17个R-L分数，表明在输入含噪的ASR/OCR文本和帧时，TCR具备更强的视频标题学习能力。尽管这些基线模型能够根据用户标注的文本描述生成摘要和字幕，但它们无法生成社交媒体语言风格的标题，也无法处理含噪的文本信息。相比之下，TCR通过充分利用所有文本/帧标记之间的依赖关系以及去噪训练，能够生成更准确的标题。

表3. How2 的 Rouge 分数比较。

模型	ROUGE-1	ROUGE-2	ROUGE-L
S2S	48.1	28.2	43.4
FT	51.1	31.0	45.8
HA-RNN	53.9	34.2	48.7
HA-Transformer	55.1	36.0	50.1
MFN-rnn	60.0	43.6	56.1
MFN-Transformer	59.3	42.1	55.0
TCR	**60.9**	**44.2**	**60.6**

为进一步考察 TCR 的生成能力，我们将其应用于一个公开数据集 How2，该数据集旨在为视频生成事实摘要。

表4。对选定帧和视频自带默认封面的评估得分。

图片	*Superior*	*Informativeness*
我们选定的帧 31.5%		15.5%
默认封面	25.5%	15.5%
无法判断	43.0%	69.0%

表2展示了在 How2 上的结果。为进行公平比较，我们采用了文献 [9] 中的数据集和基线模型。有关这些基线模型的详细信息，请参见补充材料D。从表2可以看出，TCR 在所有 Rouge 分数上均优于基线模型。值得注意的是，它比表现最佳的基线模型 MFN-rnn 高出 4.5 个 R-L分数。这一发现再次表明，TCR 不仅擅长在事实性描述范围内生成内容（例如 How2 中的摘要），也擅长在事实性描述范围之外生成内容（例如 SVTG 中的标题）。此外，通过比较表2和表2中 MFN 的表现，我们可以得出结论：MFN 无法泛化到 SVTG，因为 SVTG 有别于传统的视频摘要/字幕数据集。

封面选择性能。 我们通过人工判断，从以下方面将选定的封面图像与视频自带的默认封面进行比较：信息丰富度（哪张图片与视频标题所传达的主题相对应），以及优越性（哪张图片更适合作为能够吸引您的视频封面）。有关人工评估的更多详情，请参阅补充材料*E*。

结果如表4所示。从表中可以观察到，在大多数情况下，参与者无法判断哪一张更好，因为这两张图片几乎完全相同。此外，我们选出的封面在优越性方面比默认封面高出23.5%，而在信息丰富度方面与默认封面相当，这证明了我们的方法能够从一系列帧中自动选择封面图像。

消融实验。 所有消融实验均在SVTG数据集上进行。表2的最后三行展示了不同模态影响的消融结果。从表中可以看出，TCR优于去除视觉模态的TCR和去除文本模态的TCR，这表明，与单一模态相比，多模态信息更有利于标题生成。还可以很容易地观察到，文本模态占主导地位，而视觉模态对标题生成的贡献较小，因为自动语音识别和光学字符识别检测到的文本通常包含描述短视频内容亮点的词语。补充材料F进一步提供了对所选句子/帧数量以及细化迭代次数的消融测试。

5.4 定性分析

图3展示了示例，包括基线方法和我们提出的模型生成的标题，以及封面选择对比。结果显示

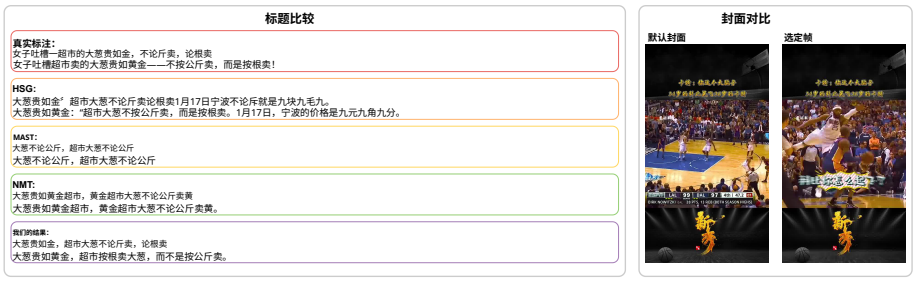

图3. （a）基线方法和TCR生成的标题示例；（b）选定帧及视频自带的默认封面。视频标题为“Kobe Bryant pump fakes Vince Carter”。

图3(a)表明，我们提出的模型能够生成比基线方法更准确的标题。此外，如图3(b)所示，选定帧比视频自带的默认封面更好，因为选定帧捕捉到了防守球员跳跃的瞬间，而默认封面没有捕捉到。为了更好地理解细化过程以及我们的模型所学到的内容，我们还在补充材料G中将注意力分数进行了可视化。

6 结论

本文首先收集了一个名为SVTG的真实世界数据集，专门用于支持短视频的联合标题生成与封面选择任务。据我们所知，SVTG是首个包含吸引人标题的短视频数据集，而其他基准数据集则包含未经修饰的事实性字幕或摘要。随后，我们提出了TCR方法，该方法由标题封面生成器和模型细化模块组成。标题封面生成器通过充分利用不同模态之间的依赖关系来生成标题并提取封面图像。细化模块进一步筛选高质量训练样本，以及每个样本中的相关文本/帧，从而逐步细化生成器。在SVTG数据集和公开的How2数据集上进行的实验表明，面对噪声数据时，所提出的TCR方法在标题生成方面效果显著。此外，封面选择评估结果表明，与默认封面相比，我们选定的封面更受青睐。这进一步证明了TCR方法在封面选择方面的有效性。

参考文献

1. Caglayan, O., Garc´ıa-Mart´ınez, M., Bardet, A., Aransa, W., Bougares, F., Barrault, L.：Nmtpy：用于高级神经机器翻译系统的灵活工具包。《布拉格数学语言学通报》**109**(1)，15（2017）2. Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., Hon, H.W.：用于自然语言理解与生成的统一语言模型预训练。《神经信息处理系统进展》**32** （2019）

3. He, K., Zhang, X., Ren, S., Sun, J.：用于图像识别的深度残差学习。载于：IEEE计算机视觉与模式识别会议论文集。第770–778页（2016）4. Hu, R., Singh, A.：Unit：采用统一Transformer的多模态多任务学习。载于：IEEE/CVF国际计算机视觉会议论文集。第1439–1449页（2021）5. Iashin, V., Rahtu, E.：多模态密集视频字幕生成。载于：IEEE/CVF计算机视觉与模式识别会议研讨会论文集。第958–959页（2020）6. Khullar, A., Arora, U.：MAST：采用三模态层次注意力的多模态抽象式摘要。载于：首届文本之外的自然语言处理国际研讨会论文集。第60–69页（2020）7. Li, M., Chen, X., Gao, S., Chan, Z., Zhao, D., Yan, R.：VMSMO：学习为基于视频的新闻文章生成多模态摘要。载于：2020年自然语言处理中的经验方法会议（EMNLP）论文集。第9360–9369页（2020）8. Lin, C.Y.：Rouge：用于摘要自动评估的软件包。载于：文本摘要领域拓展。第74–81页（2004）9. Liu, N., Sun, X., Yu, H., Zhang, W., Xu, G.：采用遗忘门的多阶段融合，用于开放域视频中的多模态摘要。载于：2020年自然语言处理中的经验方法会议（EMNLP）论文集。第1834–1845页（2020）10. Nallapati, R., Zhai, F., Zhou, B.：Summarunner：用于文档抽取式摘要的基于循环神经网络的序列模型。载于：AAAI人工智能会议论文集。第31卷（2017）11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 等：PyTorch：一种命令式风格的高性能深度学习库。《神经信息处理系统进展》**32**（2019）12. Ren, J., Shen, X., Lin, Z., Mech, R.：短视频中的最佳帧选择。载于：IEEE/CVF冬季计算机视觉应用会议论文集。第3212–3221页（2020）13. Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., Metze, F.：How2：用于多模态语言理解的大规模数据集。载于：NeurIPS（2018）14. Song, Y., Redi, M., Vallmitjana, J., Jaimes, A.：点击还是不点击：从视频中自动选择美观缩略图。载于：第25届ACM信息与知识管理国际会议论文集。第659–668页（2016）15. Tan, G., Liu, D., Wang, M., Zha, Z.J.：学习离散地组合用于视频字幕生成的推理模块网络。载于：第二十九届国际人工智能联合会议国际会议论文集。第745–752页（2021）16. Wang, D., Liu, P., Zheng, Y., Qiu, X., Huang, X.J.：用于抽取式文档摘要的异构图神经网络。载于：第58届计算语言学协会年会论文集。第6209–6219页（2020）17. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., 等：Google神经机器翻译系统：弥合人类翻译与机器翻译之间的差距。arXiv预印本 arXiv:1609.08144（2016）