Movie101：电影理解新基准

岳子豪，张琦，胡安文，张良，王子恒，金琴*
中国人民大学，信息学院
{yzihao, zhangqi1996, anwenhu, zhangliang00, zihengwang, qjin}@ruc.edu.cn

摘要

为了帮助视障人士欣赏电影，自动电影解说系统需要在没有演员台词的情况下，生成准确、连贯且富含角色信息的剧情解说。现有的研究将这一挑战作为一个普通的视频描述任务，通过一些简化手段，例如去除角色名、使用基于 n-grams 的指标来评估解说质量，这使得自动系统很难满足真实应用场景的需求。为此，我们构建了一个大规模的中文电影基准，名为 **Movie101**。该基准中的电影片段解说（Movie Clip Narrating, MCN）任务要求模型在没有演员说话的地方为完整的电影片段生成包含角色信息的解说段落，更贴近实际应用场景。同时，我们还提供了外部知识（如角色信息、电影类型等），以便模型能够更好地理解电影。此外，我们还提出了一种新的评测指标 MNScore 用于评估生成解说质量，它与人工评测更一致。我们的基准还支持时序解说定位（Temporal Narration Grounding, TNG）任务，以研究在给定文本解说的情况下，在整部电影中定位目标片段。对于这两个任务，我们提出的方法充分利用了外部知识，优于精心设计的基线。数据集和代码发布于https://github.com/yuezih/Movie101。

1 介绍

报告显示，截至 2020 年，全球约有 2.85 亿视障人士 (He et al., 2020)。尽管有相关规定出台以保证电影和电视节目的无障碍访问，但背后的技术同样值得关注。音频描述（Audio Description，AD，也被称为视频描述）就是其中一种，它通过解说屏幕上正在发生的事情，让视障人士得以体验电影或电视节目。然而，制作电影解说脚本并非易事，通常需要专业的解说员精心制作，其高昂成本 (Lakritz and Salway, 2006) 阻碍了有解说电影的制作，限制了视障观众体验电影的机会。

为解决这个问题，已有工作尝试将此过程自动化，通过构建电影解说数据集，以支持自动解说生成研究，包括 MPII-MD 数据集 (Rohrbach et al., 2015) 和 M-VAD 数据集 (Torabi et al., 2015)，具有与电影视觉内容对齐的、镜头级别的解说或电影剧本。基于这些数据集，不同的自动电影解说方案被提出 (Rohrbach et al., 2017)。

然而，已有的这些基准存在局限。首先，任务设计与实际电影解说场景存在差距。这些任务主要专注为几秒钟长的镜头生成单句解说，而无法支持为较长的剧情生成连贯解说，对于视障人士理解电影而言，后者是至关重要的。而且，这些镜头的时间戳是精心标注的，这对于现实生活中待解说的新电影来说很难获取。同时，这些任务通过一些简化手段，将极具特色的电影解说任务降格为普通视频描述任务，比如将角色名字替换为 SOMEONE，丢掉了角色与剧情的联系。还有，这些基准在评测时使用基于 n-grams 的指标来评估生成的解说，那么语义正确、但表述与参考不一致的解说就会受到过度惩罚，尤其电影解说往往只有一句参考。此外，现有的数据集都是英文的，而世界上约有五分之一的人以中文为母语，其中有超过 1700 万视障人士 (Yu and Bu, 2021)。因此，构建一个中文电影解说基准是有必要的。

* 通讯作者

图 1: 来自电影《夏洛特烦恼》的数据样例（英文翻译为方便阅读而提供）。

针对现有电影解说基准的局限性，本工作提出了一个新基准，包含 101 部中文电影，命名为 **Movie101**。这些电影收集自西瓜视频的无障碍影院[1]，这里电影是带有解说的重制版。我们通过自动处理和人工校正，从原始视频中获取了解说和演员台词，还爬取了与电影相关的附加信息。Movie101 包含总计 92 小时的 30,174 个解说片段，数据样例如图 1 所示。根据我们的调查，没有演员说话的时候往往就会加入解说（见 Appendix A），为了实现切合实际的自动电影解说，我们提出了**电影片段解说**（Movie Clip Narrating, MCN）任务，模型被要求在没有台词的地方生成解说。对于一部未曾被解说的新电影而言，这种设定具有潜在的好处，既然演员台词的时间戳容易获取（例如利用电影剧本或 OCR 和 ASR 等手段获取），我们很容易知道模型需要在哪生成解说。对于 MCN 任务，我们重新组织了 Movie101 数据集，将两个演员对话之间的解说片段合并为一个较长的片段，以实现对真实场景电影解说的模拟，得到了 14,109 个不同长度的长片段。为了让观众能够准确理解剧情，生成的解说中应包含角色名。此外，为了更好地评估模型生成解说的质量，我们组织了人工评测，并据此设计了一个专用于电影解说的新指标，即电影解说评分（Movie Narration Score, MNScore），它与人工评测具有最高的一致性。除了 MCN 任务，我们的数据集还支持**时序解说定位**（Temporal Narration Grounding, TNG）任务，该任务要求模型根据文本解说在电影中定位目标片段的起止时间。对于这两个任务，我们测试了现有方法的表现，并提出了我们融入辅助的外部知识的改进模型。除了 MCN 和 TNG 任务外，Movie101 还可以潜在地支持其他电影理解任务，如视觉问答和动作识别等。

本文的主要贡献如下：1) 我们提出了一个新的电影理解基准 Movie101，包含大量与视频对齐的中文文本解说。2) 我们提出了两项主要任务，MCN 和 TNG，以及一个新的解说评价指标 MNScore，其中 MCN 更符合实际电影解说场景的需求，而 MNScore 与人类评估更一致。3) 我们对已有模型进行了基准测试，并分别为 MCN 和 TNG 提出了外部知识增强的改进模型。我们期望 Movie101 基准能够激发电影解说与理解的更多新探索。

2 相关工作

数据集. 目前支持自动解说生成任务的数据集包括 M-VAD (Torabi et al., 2015) 和 MPII-MD (Rohrbach et al., 2015)，它们被合并为 LSMDC (Rohrbach et al., 2017)。M-VAD 是基于解说自动分割与对齐方法收集的，包含来自 92 部 DVD 的 47K 个视频片段，平均长度为 6.2 秒，每个片段都有一个对齐的叙述。MPII-MD 包含来自 94 部电影的 68K 个视频片段，平均时长为 3.9 秒，其中大约一半的片段

[1] https://www.ixigua.com/channel/barrier_free

带有配对的剧本，另一半带有配对的解说。除了电影，电视节目也可以支持自动解说生成。Lei et al. (2020) 提出了 TV Show Caption (TVC)，是 TV Show Retrieval (TVR) 的一个变体。它包含平均时长为 9.1 秒的 11K 个短视频和 26K 描述视觉、对话和字幕的文本。这些都是英文数据集。

视频描述（Image Captioning）. 作为经典的视觉语言任务，视频描述任务要求模型为给定的视频生成自然语言描述。常规视频描述的解决方案经历了从预设计模板 (Kojima et al., 2002; Guadarrama et al., 2013) 到利用深度神经网络进行序列到序列生成 (Pasunuru and Bansal, 2017) 的不同阶段。该任务的一个挑战性变体是密集视频描述 (Krishna et al., 2017)，它要求为长时间多事件视频生成多句描述。主流方法为两阶段生成方法，即先在视频中进行事件检测，然后为每个事件单独生成描述 (Krishna et al., 2017; Park et al., 2019; Rohrbach et al., 2014; Xiong et al., 2018)。最近，一些工作避免了事件检测阶段，直接基于视频生成段落描述，例如 One-stage Paragraphing Model (OVP) (Song et al., 2021)，取得了有竞争力的效果。基于此，我们提出了我们的外部知识增强的电影解说模型。能够区分不同角色的视频描述更为实用。 Park et al. (2020) 尝试要求模型生成如 `PERSON1`、`PERSON2` 之类的标签来区分不同的人，从而实现有角色意识的电影解说。然而，它无法生成具体的角色名称，在实用性上有所欠缺。

时序定位（Temporal Sentence Grounding, TSG）. 时序定位任务旨在根据自然语言查询在视频中定位特定时间片段 (Gao et al., 2017)。主流的方法是采用两步流程，首先通过滑动窗口生成大量的时间片段候选项，然后根据它们与查询句的语义相似性进行排名。后续的工作试图通过增强视频和查询之间的跨模态交互 (Liu et al., 2021; Li et al., 2022) 或引入新的检测头 (Lei et al., 2021; Zhang et al., 2020a) 来提高定位性能。对于交互方法，Liu et al. (2021) 采用了迭代对齐网络 (IA-Net) 多步迭代地进行模态内和模态间的特征交互。Li et al. (2022) 将视频和查询分解成多个结构化层次，并在它们之间学习细粒度的语义对齐。在这项工作中，我们基于 IA-Net 模型，将外部知识融入进来。

3 Dataset

3.1 数据收集

电影获取. 据我们所知，提供中文无障碍电影的平台寥寥无几，西瓜视频的无障碍影院就是其中之一。该平台在线提供了 100 多部无障碍电影，并且仍在上新，可以支持我们的数据集进一步扩展。我们从西瓜视频收集到了迄今为止可用的全部 101 部电影，并爬取了尽可能多的附加信息，如标题、简介、类型、导演、演员等。我们特别强调演员，包括重要角色的演员姓名、角色姓名、演员肖像、角色排名以及其它信息。我们期望这些信息有利于电影解说和一般的电影理解任务。

解说与台词提取. 由于平台上电影的台词和解说分别为视频字幕和音频形式，因此我们利用 OCR 和自动语音识别（ASR）工具进行转录。对于台词，我们使用开源 OCR 工具 PaddleOCR[2]以 2.4 FPS 的帧率从字幕中提取文本，并手动去除每部电影开头和结尾的无关字幕。对于解说，我们从电影中提取音轨，并利用 iFlyTek[3]提供的 ASR 服务，该服务可以检测音频中的语音并将其转录为文本。此外，该服务支持识别不同的说话人，这有助于区分解说者和演员。然而，ASR 服务并不完美，其输出包含错误，如错误字符、不合理的句子断开、误将解说识别为电影对话等。因此，我们招募人类标注者进一步纠正 ASR 转录错误，并手动去除非解说文本，以提高数据质量。我们还删除了开头（例如，电影剧情简介，演员介绍）和结尾的摘要解说。为了保证解说的连贯性，

[2] https://github.com/PaddlePaddle/PaddleOCR

[3] https://www.xfyun.cn/doc/asr/lfasr/API.html

表 1: Movie101 和其他电影解说、时序定位数据集的比较（* 表示统计基于汉字）。

Task	Dataset	Video num.	Text num.	Avg. video len.	Avg. text len.	Avg. actions	Avg. role names
Narrating	M-VAD	47K	47K	6.2 sec.	10.8	-	-
	MPII-MD	68K	68K	3.9 sec.	9.6	1.4	0.37
	TVC	109K	262K	9.1 sec.	13.4	1.9	0.75
	Movie101-N	14K	14K	20.4 sec.	80.7*	12.3	2.0
Grounding	Charades-STA	10K	16K	31 sec.	7.2	1.1	0
	ActivityNet	20K	72K	118 sec.	13.5	2.1	0.02
	TVR	22K	109K	76 sec.	13.4	1.9	0.75
	Movie101-G	101	30K	6,144 sec.	47.3*	6.9	1.1

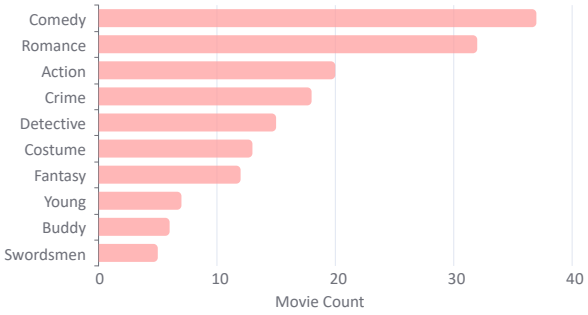


图 2: 电影类别分布。

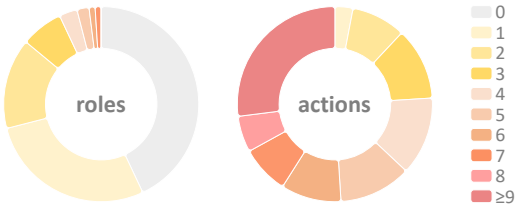


图 3: Movie101 中每个解说中的角色名数量和动作数量分布。

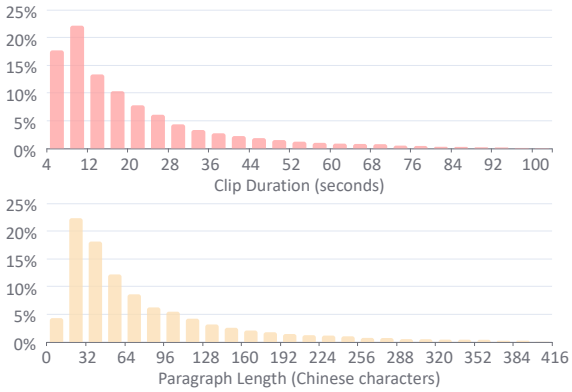


图 4: Movie101-N 中解说片段的时长和长度分布。

我们进一步将 ASR 输出的零散片段组织解说段落。具体而言，如果两个片段的时间间隔小于 1 秒，我们将其合并。段落长度阈值为 100 个字符，以限制过度合并，避免片段过长。我们也考虑了标点符号，例如，在中文中，一个句号可能意味着解说段落的结束。更详细的数据质量描述可以在 Appendix B 中找到。

Movie101-N 与 Movie101-G. 对于现实生活中的电影解说，我们期望模型能在两段演员对话之间的间隙进行解说。因此，我们重新组织 Movie101 以适应这种任务形式。具体而言，我们首先将 Movie101 中的独立台词合并为对话，其中两行之间的时间间隔小于 5 秒的视为属于同一对话。然后，我们将两个相邻对话之间的所有解说片段合并为一个长段落。这样，我们得到了 **Movie101-N**，其解说段落由对话分隔，很好地模拟了实际的解说场景。同时，利用 Movie101 中丰富的视频-文本配对，我们创建了另一个变体数据集来支持时序定位任务，命名为 **Movie101-G**，其中解说被当做查询，与之对齐的视频为目标片段。对于验证和测试，我们分别精心选择了类型丰富的 10 部电影。

3.2 数据统计

电影属性. Movie101 包含 101 部电影，涉及 41 种类型（一部电影最多可以属于 4 种类型）和总计 645 个角色。图 2 显示了最热门的前 10 种类型的电影数量，其中喜剧、爱情和动作位列前三。

片段属性. Movie101 包含总计 30,174 个短解说片段，平均时长为 11.0 秒，平均长度为 47.3 个字。从解说中，我们使用爬取的附加数据和中文词性标注工具箱 HanLP[4] 识别角色

[4] https://github.com/hankcs/HanLP

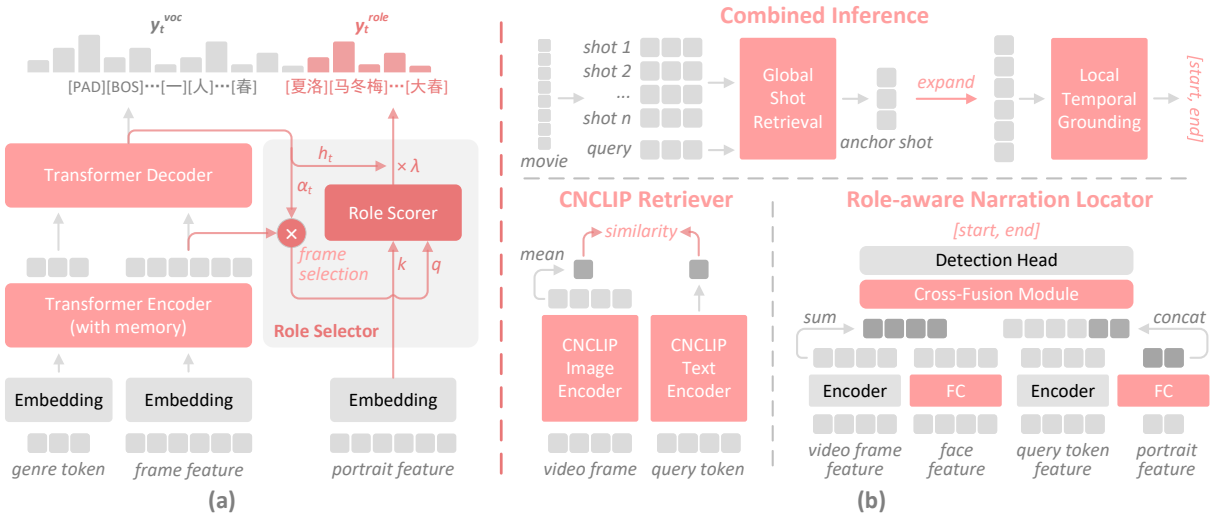


图 5: 我们的模型框架。(a) 用于 MCN 任务的 RMN；(b) 用于 TNG 任务的全局镜头检索和局部时序定位。

名和动作。图 3 显示了单个片段中角色名和动作的数量分布。用于解说任务的变体数据集 Movie101-N 包含平均长度为 20.4 秒、80.7 个字符的 14,109 个长解说片段。表 1 的比较表明，Movie101-N 包含比现有电影解说数据集显著更长的视频和文本描述，而 图 4 中的长度分布表明片段长度变化多端。Movie101-G 包含 30,174 个待定位的片段，平均视频长度为 6,144 秒，同样远超已有 TSG 数据集。

4 电影片段解说

4.1 任务描述

为了帮助视障人士在看电影时能够跟上电影剧情，我们首先提出了一个电影片段解说（Movie Clip Narrating, MCN）任务，目标是在给定 Movie101-N 中的视频片段后生成与剧情相关的段落解说。此外，不同电影类型的解说风格可能会有所不同；角色肖像和角色名是模型准确描述剧情主体的重要外部知识。因此，我们也提供这些信息以支持 MCN 任务。

4.2 方法

对于 MCN 任务，我们提出了一种基于 Transformer (Vaswani et al., 2017) 和编-解码器架构的模型——Role-pointed Movie Narrator (RMN)。模型接受包括视频、电影类型、角色名和演员肖像在内的多模态输入，其中，编码器主要对视频剪辑进行编码，解码器生成解说，如 图 5 (a) 所示。

对于编码器，视频被编码成帧级图像特征；为了强调角色，我们从每一帧中提取面部特征，并根据面部检测的置信度得分将特征按序拼接到相应的帧特征上；电影类型也被编码为一系列可学习的类型特征。视频和类型表征拼接，由交叉编码器编码。然后，我们遵循 OVP 模型 (Song et al., 2021)，使用动态记忆库来优化视频表征，该记忆库在每个解码时间步更新。

对于解码器，在原始的 Transformer 解码器之外，我们还让模型能够通过指针网络 (Gu et al., 2016)，在逐字生成的过程中，直接选择电影演员表中的整个角色名进行生成。在解码的第 t 个时间步，利用解码器隐藏状态 h_t，我们首先计算常规词汇表上的各单词的预测得分 y_t^{voc}。然后，我们设计了一个角色选择器模块，在该电影的角色名列表中获取各个角色名的预测得分 y_t^{role}。具体地，根据解码器当前在各视频帧上的注意力分布 α_t，我们对帧级别视频特征进行加权求和，得到一个具备上下文意识的视频特征。然后，我们用加权特征作为 query，角色肖像特征作为 key，计算出各角色名的得分 y_t^{role}。最后，在第 t 个解码时间步，下一词

表 2：以人类评估为参考，评测指标对候选解说的评估的准确性。(Acc.：准确性；Info.：信息含量；Qual.：语言质量)

Metric	Acc.	Info.	Qual.	Overall
CIDEr	86.7	83.0	82.0	87.0
BLEU@4	85.0	82.0	80.3	86.0
METEOR	87.0	82.7	82.7	87.0
CLIPScore	39.7	40.0	38.0	39.0
BERTScore	88.0	84.7	**87.7**	90.3
EMScore	40.3	42.3	41.3	41.3
DIV	51.7	51.3	57.3	54.7
PPL	43.0	46.7	45.0	45.3
RoleF1	33.0	31.0	28.3	32.3
MNScore	**90.3**	**86.7**	86.3	**92.0**

预测的概率分布为：

$$y_t = \text{softmax}([y_t^{voc}; \lambda y_t^{role}]) \tag{1}$$

这里 [;] 指拼接；λ 是一个门控系数，由 h_t 计算（h_t 可以用于决定什么时候该生成角色名，什么时候该生成常规词汇）。

4.3 评测

现有的电影解说基准直接像常规视频描述一样，采用基于 n-grams 的评测指标，例如 CIDEr、BLEU 和 METEOR，然而这些标准存在缺陷，例如，它们常常低估语义正确但在文本上不一致的语句，这已被广泛报道 (Zhang et al., 2020b; Shi et al., 2022)。对于电影解说，一个电影片段可以有多种表述方式，却只有一个参考解说。因此，仅仅通过文本比较不足以衡量一个解说段落的质量。

为了更好地评估 MCN 任务中生成的解说，我们开展了人工评测，以研究人类对不同解说的偏好。我们随机选择了 30 个电影片段，每个片段有 5 个候选解说，其中 3 个是不同模型生成的，2 个是通过对参考解说进行退化获得的。接下来，我们招募了 10 个标注者，从准确性、信息含量和语言质量三个维度对每个视频的候选解说进行排序。准确性定义了解说描述视频内容的准确性，特别是角色、动作和物体；信息含量定义了解说描述视频内容的丰富程度；语言质量则由解说的流畅度和语法正确性决定。

基于人类评估的结果，我们研究了一系列客观指标，包括：(1) 基于深度神经网络的最新视频描述评价指标，包括 CLIPScore (Hessel et al., 2021)、BERTScore (Zhang et al., 2020b) 和 EMScore (Shi et al., 2022)，它们已经被验证在视频描述评估上的表现优于 ngram-based 的指标；(2) 语言质量评测指标，包括 n-grams 多样性 (Shetty et al., 2017) (DIV) 和语言建模困惑度 (PPL)；(3) 角色名预测的 F1 分数 (RoleF1)。对于一个视频的任意两个候选解说，我们使用人类评分作为参考，以确定这些指标是否能够正确判断两个候选的优劣，其判断准确性越高，表明其评估表现与人类更一致。最后，我们确定了一个新的度量标准——电影解说评分（Movie Narration Score, MNScore），定义如下：

$$mns = \frac{1 \cdot ems + 4 \cdot berts + 1 \cdot rf1}{6} \times 100 \tag{2}$$

其中 mns, ems, $berts$ 和 $rf1$ 分别指的是 MNScore，EMScore，BERTScore 和 RoleF1。如表 2 所示，BERTScore 在解说评估准确性方面超过了基于 ngram 的度量，而我们新提出的 MNScore 则与人类评估达到了最好的一致性。有关候选解说和上述指标的实现，更多细节请参见 Appendix C。

4.4 实验

实现细节. 模型通过下一词预测任务训练，使用最大似然估计 (MLE) 目标进行优化。对于视频，我们使用在大规模图像-文本对上预训练的 CLIP (Radford et al., 2021) 模型和在 HowTo100M (Miech et al., 2019) 上预训练的 MIL-NCE (Miech et al., 2020) 模型，以 1 帧/秒的速度提取帧级别的 CLIP 和 S3D 特征，其维度分别为 512 和 1024，并将它们拼接在一起。对于视频帧中出现的人脸和角色肖像，我们使用在 MS1M (Guo et al., 2016) 上预训练

表 3: Movie101-N 上电影片段解说任务的表现 (f_v: 视频中的面部特征; g: 电影类型)。

Model	f_v	g	EMScore	BERTScore	RoleF1	MNScore
Vanilla Transformer			0.153	0.150	0	12.55
OVP			**0.155**	0.159	0	13.18
RMN			0.153	0.185	0.195	18.13
	✓		0.154	0.186	**0.240**	18.97
	✓	✓	0.154	**0.188**	0.238	**19.07**

表 4: 一阶段模型在 Movie101-GSR（temp）上的全局镜头检索表现。

Model	Recall@1	Recall@5	Recall@10
CNCLIP	25.98	54.91	66.99

的 Arcface 模型 (Deng et al., 2019) 来提取面部特征。视频帧中足够的人脸时，特征由零向量填充。

结果与分析. 我们选择 Vanilla Transformer (Zhou et al., 2018) 和 OVP(Song et al., 2021) 作为 MCN 任务的基线。如表 3 所示，RMN 显著超越基线，尤其在 RoleF1 上。这表明我们的模型学习了如何在指针网络的帮助下利用外部知识生成角色名。为了验证我们的 RMN 模型中的电影类型和面部表特征的贡献，我们也通过逐步加入这些特征进行了消融研究。结果显示，从视频帧中提取的面部特征为模型的角色意识带来了显著提升，这表明使用面部特征来链接视频内容和演员肖像对于角色相关的解说的生成是有益的。定性结果可以在 Appendix D 中找到。

5 时序解说定位

5.1 任务描述

为了帮助人们在观看电影时快速找到感兴趣的片段，AI 需要理解用户的意图并定位目标片段。为了实现此目标，我们提出了时序解说定位（Temporal Narration Grounding，TNG）任务。给定一段解说作为 query，TNG 的目标是在整部电影中预测相应片段的起止时间。

表 5: 二阶段模型在 Movie101-LTG(temp) 上的局部时序定位表现（f_v 和 f_t 分别指向视频和文本中表示添加面部特征）。

模型	f_v	f_t	Rank@1		Rank@5	
			IoU0.3	IoU0.5	IoU0.3	IoU0.5
2D-TAN			25.85	18.60	52.17	43.82
IA-NET			25.16	17.98	57.11	42.68
RNL	✓		26.64	19.01	**59.63**	44.51
RNL		✓	16.98	19.57	57.18	42.86
RNL	✓	✓	**27.54**	**20.22**	59.52	**45.69**

5.2 方法

在有限的计算资源下，已有时序定位模型难以处理整部电影的输入。因此，我们为 TNG 任务提出了一个两阶段框架，第一阶段通过全局镜头检索（Global Shot Retrieval, GSR）粗略地定位目标片段，第二阶段通过局部时序定位（Local Temporal Grounding）确定目标片段的精确时间戳，如图 5（b）所示。

全局镜头检索. 为了找到目标的大致位置，我们将其视为一个文本-视频检索子任务。我们将电影均分成 20 秒长的镜头，与文本 query 最匹配的镜头将被用作第二阶段进一步定位的锚点镜头。为了训练这样一个检索系统，我们构建了一个临时数据集 Movie101-GSR(temp)。具体来说，将电影切分成镜头后，我们根据镜头和标注解说的时间重叠来识别它们是否可以被认为是一个对齐的视频-文本对[5]。

我们对中文视觉-语言预训练（VLP）模型 ChineseCLIP (Yang et al., 2022)（CNCLIP）进行微调，使其实现从图像-文本到视频-文本

[5]对于一个镜头和一个解说，如果重叠时间大于二者之一的持续时间一半，它们就被认为是对齐的。

表 6: 我们提出的两阶段方法在 Movie101-G 上的联合推理表现。

Model	k-way re-ranking	Rank@1				Rank@5			
		IoU0.1	IoU0.3	IoU0.5	IoU0.7	IoU0.1	IoU0.3	IoU0.5	IoU0.7
CNCLIP+RNL	1	**18.69**	**11.65**	**6.66**	**15.38**	35.79	29.77	22.68	**14.87**
	2	18.17	10.53	5.99	14.45	36.98	**30.98**	**26.28**	13.56
	3	17.18	10.05	5.47	13.96	**37.91**	30.23	25.37	13.33

能力的迁移。具体来说，我们用 CNCLIP 的视觉编码器单独编码多个视频帧，并对CLS标记进行平均池化作为视频特征。然后我们在 Movie101-GSR(temp) 上进行视频和文本特征的对比学习，以对 CNCLIP 进行微调。

局部时序定位. 在第一阶段获取锚定镜头后，我们进一步在锚定镜头周围的 200 秒窗口（往前往后各扩充 90 秒）内定位目标片段。在 200 秒长的电影片段中进行时序定位，理解不同角色的动作是至关重要的。因此，基于最先进的 TSG 模型 IA-Net (Liu et al., 2021)，我们提出了 Role-aware Narration Locator (RNL)。

我们使用双向 GRU (Chung et al., 2014) 编码视觉特征，以获取时间上下文感知的帧表示 V。我们还从帧中提取面部特征，并使用全连接（FC）层对它们进行编码以过滤关键面部信息 F。然后，我们通过对 V 和 F 求和来得到最终的视觉表示。对于文本编码，为了将文本查询中的角色名称与视频中的角色联系起来，我们从与角色名称对应的肖像中提取面部特征，用 FC 层得到视觉表示，然后将其拼接在查询的文本 token 表示序列的后面。在训练期间，对于一个训练样例，在每个训练周期，我们随机构建一个覆盖目标片段的 200 秒长的窗口来模拟长片段。我们还构建了一个带有固定窗口的临时数据集 Movie101-LTG(temp)，作为测试集以单独评估第二阶段模型的性能。

5.3 实验

实现细节. 对于全局镜头检索，我们使用平均召回率 R@n ($n \in 1, 5, 10$) 来评估全部电影的平均检索性能。对于局部时序定位，我们按照之前的工作 (Zhang et al., 2020a)，使用 “R@n, IoU@m” 作为评价指标，即在前 n 个预测候选片段中，至少有一个片段与参考片段的时间 IoU 大于 m 的测试样例的占比。对于全局镜头检索，我们对 CNCLIP-huge 在 Movie101-GSR(temp) 进行微调。对于局部时序定位，我们在 Movie101-LTG(temp) 上对两个代码可获取的最新 TSG 模型 2D-TAN (Zhang et al., 2020a) 和 IA-Net(Liu et al., 2021) 进行了基准测试。在我们的 RNL 模型中，视频帧、面部特征和文本特征提取器分别由预训练的 MIL-NCE, Arcface（与 MCN 任务中的一样）和 BERT-base-Chinese (Devlin et al., 2019) 提取。

结果与分析. 表 4 和表 5 分别展示了全局镜头检索和局部时间定位模型的性能。通过角色感知的视频和文本编码，我们的 RNL 超越了基线模型，表明区分不同角色的动作对于电影叙述的定位至关重要。此外，我们进行了消融研究以验证角色感知编码的有效性。如表 5 所示，在视频和文本表示中加入角色信息都能够提升模型表现，兼具两者的 RNL 模型达到了最好的性能。表 6 显示了通过全局镜头检索和局部时间定位联合推理的性能。我们还展示了 k 路重排序的性能，即第一阶段检索到的前 k 个镜头分别在第二阶段用作锚点镜头，得到所有预测候选后根据置信度重新排序。实验结果显示，k 路重排提高了 Rank@5 的性能，但损害了 Rank@1 的性能。定性结果可以在 Appendix D 中查看。

6 结论

在这项工作中，我们提出了 Movie101，一个中文大规模电影理解基准。为了帮助视障人

士享受电影，我们针对自动电影描述提出了更契合实际的电影片段解说任务，并设计了一个符合人类偏好的评估指标 MNScore。Movie101 还支持时间叙述定位任务，比以前的 TSG 基准更具挑战性。此外，我们实验验证了电影理解中类型和角色等外部知识的重要性。然而，我们的模型与人类专业解说之间仍然存在显著的差距。我们仍需要进一步的研究，以帮助视障人士通过 AI 享受电影。

局限性

对视障人士来说，为整部电影提供连贯解说非常重要。在这项工作中，我们通过将电影片段解说任务中的零散片段重新组织为段落，将更长的视频片段作为输入，向这个目标迈出了一步。然而，如何确保电影中不同片段之间的描述连贯性，在这项工作中并没有被研究。这需要模型具有更高级别的理解能力，以处理整部电影并衔接不同情节。我们将这个问题留给未来的研究。

伦理声明

我们提出了一个新基准 Movie101，旨在支持提高视障人士无障碍访问的新技术探索。我们的工作有两个潜在的伦理问题，分别涉及数据来源和众包服务。我们声明如下：

数据源. 我们收集的电影来自公开可用的西瓜视频，网站的服务协议允许爬取[6]。考虑到版权问题，我们只会发布电影的 URL 列表。此外，我们的数据源不包含任何涉及隐私的或冒犯性内容。

众包服务. 我们通过社交媒体招募了 20 名中国大学生（12 名女性和 8 名男性）。对于 ASR 输出清洗，工作人员在观看电影的同时被要求纠正叙述文本中的错误。每部电影大约需要 2 小时，薪酬 50 元人民币（7.40 美元）。对于审查清洗后的数据，每部电影大约需要 30 分钟，薪酬 25 元人民币（3.70 美元）。我们提供的酬劳在中国是公平和合理的，尤其在这项工作简单而有趣的情况下。在标注工作开始之前，我们在任务文档中介绍了数据未来的用途，并确保每个人都知晓。

[6] https://www.ixigua.com/robots.txt

致谢

本工作部分受中国国家重点研发计划（编号：2020AAA0108600）和中国国家自然科学基金（编号：62072462）支持。

References

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5277–5285. IEEE Computer Society.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities

using semantic hierarchies and zero-shot recognition. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 2712–2719. IEEE Computer Society.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer.

Yuan He, Aiping Nie, Jinzhi Pei, Zhi Ji, Jun Jia, Huifeng Liu, Pengfei Wan, Mingli Ji, Chuntao Zhang, Yanni Zhu, et al. 2020. Prevalence and causes of visual impairment in population more than 50 years old: The shaanxi eye study. *Medicine*, 99(20).

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50:171–184.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 706–715. IEEE Computer Society.

James Lakritz and Andrew Salway. 2006. The semi-automatic generation of audio description from screenplays. *Dept. of Computing Technical Report CS-06-05, University of Surrey.*

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11846–11858.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.

Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. 2022. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3022–3031. IEEE.

Daizong Liu, Xiaoye Qu, and Pan Zhou. 2021. Progressively guide to attend: An iterative alignment framework for temporal sentence grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9302–9311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9876–9886. IEEE.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.

Jae Sung Park, Trevor Darrell, and Anna Rohrbach. 2020. Identity-aware multi-sentence video description. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 360–378. Springer.

Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. 2019. Adversarial inference for multi-sentence video description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6598–6608. Computer Vision Foundation / IEEE.

Ramakanth Pasunuru and Mohit Bansal. 2017. Multi-task video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1273–1283, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 184–195. Springer.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3202–3212. IEEE Computer Society.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123:94–120.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4155–4164. IEEE Computer Society.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2022. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17929–17938.

Yuqing Song, Shizhe Chen, and Qin Jin. 2021. Towards diverse paragraph captioning for untrimmed videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11245–11254. Computer Vision Foundation / IEEE.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *ArXiv preprint*, abs/2107.02137.

Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Using descriptive video services to create a large data source for video annotation research. *ArXiv preprint*, abs/1503.01070.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yilei Xiong, Bo Dai, and Dahua Lin. 2018. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 468–483.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *ArXiv preprint*, abs/2211.01335.

Chun Yu and Jiajun Bu. 2021. The practice of applying ai to benefit visually impaired people in china. *Communications of the ACM*, 64(11):70–75.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020a. Learning 2d temporal adjacent networks for moment localization with natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12870–12877. AAAI Press.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA,*

June 18-22, 2018, pages 8739–8748. IEEE Computer Society.

A 解说分布

”没有演员说话的时候”指的是**任何**没有文本对话进行的片段，无论视频中是否有演员出现。例如，一个天空的特写，或者画面中有演员但其保持沉默。我们调查了收集的 101 部电影中的对话和解说，通过合并演员的台词，我们得到了总共 15,307 段对话，构成了 15,206 个对话间隙，总时长为 99.4 小时。我们收集的 30,174 个解说片段填补了 95.3% 的对话间隙（数量），覆盖了 92.9% 的时长。因此，可以合理地假设，没有台词的地方即需要解说。

B 数据集质量阐述

我们采用了两阶段的注释过程以确保解说的质量。在第一阶段，我们聘请一组标注者根据我们的指南清洗数据。在第二阶段，另一组标注者进一步检查并纠正错误。我们将零碎的解说片段合并成段落的启发式方法是基于我们的观察经验设计的。我们进一步对解说质量进行了人工评估。在随机抽样的 300 个片段中，（1）对于解说转录的准确性，96.7%与原始 AD 在文本上保持一致；（2）对于段落的连贯性，90%保持完整和连贯的语义，7.7%应与上下文合并，2.3%应被分割为多个段落。因此，解说质量足够好以支持下游任务。

C 实现细节

候选解说. 在 Section 4.3 中，我们为每一个抽样的电影片段提供 5 个不同的候选解说供人类评测者进行排序。这些候选是通过以下方式获取的：

1. 由 Vanilla Transformer (Zhou et al., 2018) 生成；
2. 由 OVP 模型 (Song et al., 2021) 生成；
3. 由我们提出的 RMN 模型生成；
4. 通过移除和替换角色名对参考解说进行退化；
5. 通过替换名词和动词对参考解说进行退化。

指标实现. 对于基于 CLIP 的指标，包括 CLIPScore 和 EMScore，我们在我们的数据集上对 ChineseCLIP-huge (Yang et al., 2022) 进行微调，方法与在 Section 5.2 中的相同。对于每一个电影片段和生成的解说，CLIPScore 是用均匀采样的 10 帧的平均池化特征和整体文本特征计算的，而 EMScore 是用采样得到的每一帧各自的特征和文本 token 特征计算的。对于 BERTScore，我们使用 BERT-base-Chinese (Devlin et al., 2019) 进行计算，并根据基线对 BERTScore 进行缩放[7]。对于 DIV，我们按照Shetty et al. (2017) 计算 1-gram diversity 和 2-gram diversity，并对它们进行平均。对于 PPL，我们按照 HuggingFace 的计算方法[8]，使用 Causal Ernie 3.0 模型 (Sun et al., 2021) 获取每个解说的困惑度。对于 RoleF1，我们从参考解说和生成的解说中提取角色名字。我们通过 Recall 测量生成的解说如何覆盖电影片段中出现的角色；考虑到这些生成的角色名字也可能来自模型的幻觉，例如来自其它电影，我们也考虑 Precision。最后，我们用 Precision 和 Recall 计算 F1 得分。

超参数和计算. 我们在表 7 中详细说明了模型训练的关键超参数和计算负担。对于每个模型，结果都来自单次运行。

D 定性结果

D.1 电影片段解说

图 6 显示了 MCN 任务的定性结果，包括基线模型和我们提出的 RMN 模型的生成结果，以及以前的指标和我们提出的 MNScore 的评估结果。Vanilla Transformer 和 OVP 能正确提到一些动作，但不能生成正确的角色名，因为测试集中这些角色在训练期间从未出现过。

[7] https://github.com/Tiiiger/bert_score/blob/master/journal/rescale_baseline.md

[8] https://huggingface.co/spaces/evaluate-metric/perplexity

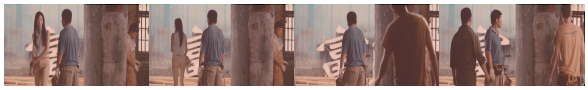
表 7: 模型训练的关键超参和计算代价。

Model	Batch size	Learning rate	Training epochs	GPU hours / epoch
VT	150	$1e-4$	≤ 100	∼3min on single RTX 2080ti
OVP	56	$1e-4$	≤ 100	∼40min on single RTX 3090
RMN	56	$1e-4$	≤ 100	∼1h on single RTX 3090
CNCLIP	16	$2e-6$	≤ 1	∼1h on 4 RTX A6000 nodes
2D-TAN	64	$1e-4$	≤ 30	∼40min on single RTX 3090
IA-Net	64	$8e-4$	≤ 15	∼20min on single RTX 3090
RNL	64	$8e-4$	≤ 15	∼20min on single RTX 3090

然而，借助 Role Selector 模块，我们的 RMN 可以很好地将视频片段中的角色与相应的角色名联系起来。此外，这些例子表明我们新提出的 MNScore 更符合人类的评估。

D.2 时序解说定位

图 7 显示了我们提出的两阶段方法的定性结果。通过全局镜头检索，我们从整部电影中得到一个接近目标片段的锚定镜头，这进一步帮助了局部时序定位来定位最终目标。



GT: 说完想说的话叶薰才离开，又不甘心地回头看了眼屋内。黄达躲在墙后面听到了叶薰讲的所有话，叶薰离开后他才从墙后走出来。(After saying what she want to say YeXun leaves only to look back at the house again reluctantly. HuangDa hides behind the wall and hears all that YeXun speaks, and he comes out from behind the wall only after YeXun leaves.)

VT: 他一边回头一边往外走，一边一边回头看了看，又看了看自己一眼，又看了看自己的背影，又低头没有说话，(As he turns back and walks out the door, he looks back, looks at himself again, looks at his back again, and looks down again without speaking,)

OVP: 他站在门口的毒贩已经停下了脚步，看着这里，他转过身来看着脚下的两个人，(The drug dealer he is standing in the doorway stop and look at the place, and he turn to look at the two men at his feet,)

RMN: 黄达还是一副模样，他稍微落寞的样子，一时间还算回到，黄达跟在黄达身后，他先是一屁股坐在椅子上，(HuangDa is still the same, slightly melancholy, for a while ... come back, HuangDa follows behind HuangDa, he first buttocks in the chair.)

Candidate	CIDEr	BLEU@4	METEOR	MNScore	Manual Ranking↓
VT	**14.30**	**6.24**	**11.75**	21.65	3.87
OVP	0.01	0	6.70	17.46	4.37
RMN	3.67	0	7.43	**27.22**	**3.63**

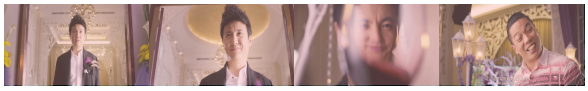
GT: 现场大屏幕上的数字转了起来，黄达和主持人转身看向大屏幕，数字转动了一会儿之后停了下来，(The numbers on the big screen turn up, and HuangDa and the host turn to look at the big screen, the numbers turns for a while and then stops.)

VT: 他们在台上观察着，时间间的位置上台下的观音室内，(They watch from the stage, the position between time on the stage in the chamber of the observer,)

OVP: 第二天，三人来到现场，孟云和余飞一起看着屏幕上的选择题，三人离开了。(The next day, the three come to the scene, MengYun and YuFei look at the multiple choice questions on the screen together, the three leave.)

RMN: 黄达看着台下的电脑，这时余飞和黄达拉着手来到台球厅，他们相互打量着这一切，(HuangDa looks at the computer under the stage, at this time, YuFei and HuangDa come to the billiard room with hands holding, they survey all this each other)

Candidate	CIDEr	BLEU@4	METEOR	MNScore	Manual Ranking↓
VT	0.01	0	4.98	9.47	4.33
OVP	**31.28**	**7.95**	**16.65**	24.65	3.87
RMN	13.12	0	10.75	**34.76**	**3.53**



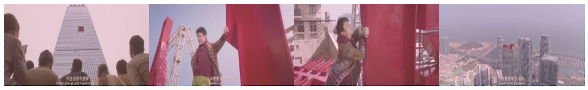
GT: 在大家的嘲笑声中夏洛也尴尬的笑了一下并坐在席位上，(In the sound of everyone's ridicule, XiaLuo smiles awkwardly and sits in the seat,)

VT: 他看着面前的菜刀，脸上满是泪水，他面面带笑意，(He looks at the chopper in front of him, his face is full of tears, and he has a smile on his face,)

OVP: 李行表情极其尴尬，他看着场上的人若有所思。(LiXing's expression is extremely embarrassed, and he looks at the people as if in thought.)

RMN: 夏洛看着手机，似乎陷入了沉思。马小丽缓缓开口，(XiaLuo looks at the phone and seems to be in deep thought. MaXiaoli slowly opens her mouth,)

Candidate	CIDEr	BLEU@4	METEOR	MNScore	Manual Ranking↓
VT	18.61	0	6.48	12.09	4.60
OVP	**47.52**	0	**8.73**	16.50	4.10
RMN	31.05	0	7.02	**24.57**	**3.30**



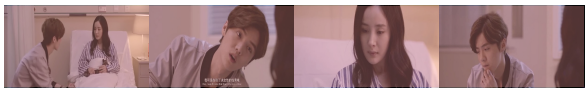
GT: 王多鱼站在保险公司大厦最顶层，穿着红色衣服绿裤衩，一只手背在后面，另一只手扶着巨大的"瘦"字，双腿交叉带着一脸享受闭上双眼，倚靠在"瘦"字上面。镜头缓缓上升拉伸，王多鱼变得越来越渺小，最后完全看不见了。(WangDuoyu stands at the top of the insurance company building, wearing red clothes and green pants, one hand behind the back, the other hand holding the huge "thin" character, legs crossing, with a face of enjoyment, closing eyes, leaning on the "thin" character. The camera slowly rises, WangDuoyu becomes smaller and smaller, and finally completely invisible.)

VT: 在众人的高举行下，张彪也在这一场，下面的高楼下，阿俊也摔倒在地上。(In the crowd hold under the high, ZhangBiao is also in this scene, below the high floor, Arjun also falls to the ground.)

OVP: 随着飞机的轰鸣声，继续朝下抓捕，而在楼梯上的临时，他选择了一个高挑的身影，这时，焦急的他选择了一个按钮，大厦向下飞去。(With the roar of the plane, continues capture downward, while on the stairs of the temporary, he chooses a tall figure, at this point, anxious he chooses a button, the building flies downward.)

RMN: 随后，王多鱼在天台上朗着升机，踏上了行程，登基本的装饰演员，王多鱼独自在空中，美丽在云大楼里摆着各种姿势，王多鱼顺着绳索向上攀爬 (Subsequently, WangDuoyu boards on the rooftop ... , embarking on a trip, ... decorative actors, WangDuoyu is in the air alone, posing in a variety of positions beautifully in the cloud building; WangDuoyu climbs upward along the rope)

Candidate	CIDEr	BLEU@4	METEOR	MNScore	Manual Ranking↓
VT	0	0	0	7.90	4.73
OVP	0	0	4.02	9.31	3.90
RMN	**4.97**	**7.06**	**11.84**	**29.83**	**3.20**



GT: 他凑近小星，得意地捶捶自己胸口，林冲又皱着眉撑起下巴。(He comes close to Xiaoxing, proudly pounding his chest. LinChong then frowns and props up his chin.)

VT: 林佳看着他们，点头，点头。(LinJia looks at them, nods her head, nods her head.)

OVP: 林佳看着丽丽，又看向自己说道。(LinJia looks at Lili, and then looks at herself and says.)

RMN: 陈姗姗抬起头，看着林冲的背影有些害怕，(ChenShanshan lifts her head and looks at LinChong's back with some fear.)

Candidate	CIDEr	BLEU@4	METEOR	MNScore	Manual Ranking↓
VT	0.67	0	9.64	19.67	4.83
OVP	5.57	0	**11.62**	26.83	3.73
RMN	**11.89**	0	9.80	**29.77**	**3.37**

GT: 桃子边说话边拦下一辆出租车，然后坐上车快速离去了。黄达一个人愣在原地。(Taozi stops a cab as she talks, then gets in and quickly leaves. HuangDa freezes alone.)

VT: 卢小鱼看到了他的眼神，他低头看着他，然后低下头，(LuXiaoyu sees the look in his eyes, and he looks down at him, then lowers his head,)

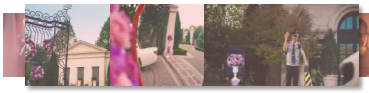
OVP: 江丰回过头来看着他，然后叹了口气，(JiangFeng looks back at him, then sighs,)

RMN: 黄达听到这话愣住了，他回头一看，(HuangDa freezes when he hears this, and he looks back,)

Candidate	CIDEr	BLEU@4	METEOR	MNScore	Manual Ranking↓
VT	**3.38**	0	6.87	16.72	4.13
OVP	0.22	0	5.92	15.73	4.30
RMN	0.04	0	**9.11**	**27.26**	**3.12**

图 6: MCN 任务的定性结果。(GT: 参考解说; VT: Vanilla Transformer) 在解说文本中，绿色和红色字符分别表示正确和错误生成的角色名。在表格中，绿色指标表示指标对候选项的排序与人类排序一致，而红色表示不一致。

Query: 影片开始挂满鲜花的欧式大铁门缓缓打开，铁门内是一座欧式建筑，一辆小汽车从门外行驶而进，它穿过摆满花束的院子中，一名保安在礼堂前一手拿起路障，一手指挥着汽车向前，这辆车没有停下。 (At the beginning of the film, a large European-style iron gate full of flowers slowly opens. Inside the iron gate is a European-style building. A car drives in through the gate, and crosses the courtyard full of flowers. A security guard holds a barricade in front of the auditorium with one hand and directs the car forward with the other; this car does not stop.)

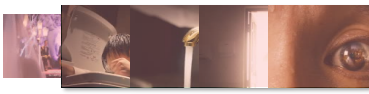


GT 182s – 200s
anchor shot 177s – 197s
prediction 183s – 205s

Query: 时间又过了一天，一只纸飞机穿过学校的楼顶，只见秋雅一个人孤零零的站在楼顶，这时袁华慢慢走了过来，秋雅看了一眼袁华便皱起眉头，一边摸着自己的小辫子，一边低下头，袁华吐了嘟嘴，含着眼泪询问：(Another day, a paper airplane flies through the roof of the school building. QiuYa is standing alone on the roof, when YuanHua slowly walks over. QiuYa takes a look at YuanHua and then frowns, while touching her pigtails, while lowering her head, YuanHua spits out his mouth and asks with tears.)

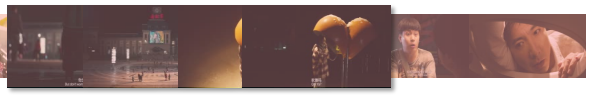
GT 2095s – 2121s
anchor shot 2097s – 2117s
prediction 2102s – 2122s

Query: 不知过了多久夏洛醒了过来，他发现洗手间的水龙头关不上了，镜子也碎了，他看着镜子中的自己说道，他打开卫生间的门，一束耀眼的光芒刺向他，他用手臂挡住眼睛，缓缓抬起头，他的瞳孔里看到的是教室的场景。(A long time later, XiaLuo wakes up. He finds the bathroom faucet can not be turned off, and the mirror is also broken. He says while looking at himself in the mirror. He opens the bathroom door, and a dazzling light shines towards him. He blocks his eyes with his arm, and slowly raises his head. In his pupils is the scene of the classroom.)



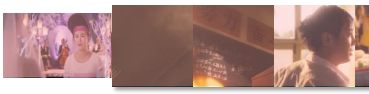
GT 725s – 748s
anchor shot 677s – 697s
prediction 724s – 746s

Query: 马冬梅拉着行李走向西虹市的车站，车站上方的大屏上播放着夏洛的相约98，画面一转时间来到了高考后，袁华穿着格子衫，脖子上挂着围巾耐着寒冷，在一个顶部积雪的公共电话亭里拨打着电话，(MaDongmei walks towards the Station of Xihong City with her luggage. The large screen above the station is playing XiaLuo's *meeting 1998*. The scene switches to the time after the college entrance exams, YuanHua is wearing a plaid shirt with a scarf around his neck to withstand the cold, dialing the phone in a public phone booth with snow on top.)



GT 3526s – 3548s
anchor shot 3557s – 3577s
prediction 3528s – 3548s

Query: 火苗瞬间点燃了教室的窗帘，同学们乱作一团，纷纷拿出书本灭火，校长还拿出一瓶墨水泼向燃烧的窗帘，这时夏洛的妈妈跑进教室，夏洛缓缓转过头，看到是母亲来了，缓缓走向他，突然间就跪倒在地上，一把抱住他的大腿。(The fire instantly ignits the curtains in the classroom, and the students are in a mess, and all taking out books to put out the fire. The school principal also takes out a bottle of ink throwing to the burning curtains. At this time, XiaLuo's mother runs into the classroom, XiaLuo slowly turns his head, seeing his mother coming, slowly walking towards her, and suddenly falling to her knees with a hug on her thighs.)



GT 1085s – 1111s
anchor shot 1077s – 1097s
prediction 1089s – 1110s

Query: 夏洛的记者发布会开始了，一群记者围着夏洛，夏洛的两边站着秋雅和张扬，大家都喜笑颜开，只有夏洛，面对记者一言不发。曾经他弹着吉他，冬梅在一旁举着灯牌静静聆听，在家弹吉他时，冬梅在旁边拖着地，还有他最爱吃的茴香打卤面，这时夏洛突然举起左手示意大家安静，(XiaLuo's press conference begins. A group of reporters surround XiaLuo. On either side of XiaLuo stand QiuYa and ZhangYang. Everyone is happy and smiling, only Xia Luo faces the reporters without saying a word. Once when he plays the guitar, Dongmei in the side holding a light sign quietly listening. When playing the guitar at home, Dongmei is mopping the floor next to him, and there is also his favorate Dalu noodles. At this time, XiaLuo suddenly raises his left hand to signal everyone to be quiet.)

GT 4775s – 4806s
anchor shot 4777s – 4797s
prediction 4777s – 4803s

图 7: 电影《夏洛特烦恼》中 TNG 任务的定性结果。