Movie101: A New Movie Understanding Benchmark

Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, Qin Jin*
School of Information, Renmin University of China
{yzihao, zhangqi1996, anwenhu, zhangliang00, zihengwang, qjin}@ruc.edu.cn

Abstract

To help the visually impaired enjoy movies, automatic movie narrating systems are expected to narrate accurate, coherent, and role-aware plots when there are no speaking lines of actors. Existing works benchmark this challenge as a normal video captioning task via some simplifications, such as removing role names and evaluating narrations with ngram-based metrics, which makes it difficult for automatic systems to meet the needs of real application scenarios. To narrow this gap, we construct a large-scale Chinese movie benchmark, named **Movie101**. Closer to real scenarios, the Movie Clip Narrating (MCN) task in our benchmark asks models to generate role-aware narration paragraphs for complete movie clips where no actors are speaking. External knowledge, such as role information and movie genres, is also provided for better movie understanding. Besides, we propose a new metric called Movie Narration Score (MNScore) for movie narrating evaluation, which achieves the best correlation with human evaluation. Our benchmark also supports the Temporal Narration Grounding (TNG) task to investigate clip localization given text descriptions. For both two tasks, our proposed methods well leverage external knowledge and outperform carefully designed baselines. The dataset and codes are released at `https://github.com/yuezih/Movie101`.

1 Introduction

The estimated number of visually impaired people worldwide was about 285 million by 2020, according to reports (He et al., 2020). While regulations are in place to ensure increased access for these audiences to experience the culturally dominant movies and TV shows on popular media platforms, technologies that provide them with genuine experience are becoming increasingly important. Audio description (AD, also known as video description) is a form of such technology intended for visually impaired audiences to experience the movie or TV show by hearing what is happening on-screen. However, producing movie narration scripts is not trivial, often requiring a professional writer to oversee the original movie. The high cost of narration generation (Lakritz and Salway, 2006) greatly hinders the production of movies with AD and thus limits the opportunities for visually impaired users to experience movies.

To address this issue, attempts have been carried out to automate AD production. Datasets of movies with ADs are constructed to support the research on automatic AD generation, including the MPII-MD dataset (Rohrbach et al., 2015) and M-VAD dataset (Torabi et al., 2015), with shot-level ADs or scripts aligned to the visual contents of movie. Consequently, different solutions for automatic movie narrating have been proposed based on these datasets (Rohrbach et al., 2017).

However, existing benchmarks suffer from several limitations. Firstly, there is a gap between the designed tasks and the actual movie narration scenario. These tasks mainly focus on generating single-sentence narrations for shots of a few seconds. They can not support the generation of coherent narrations for longer plots, which is critical for the visually impaired to better understand the movie, and the timestamps of these shots are carefully annotated, which are difficult to obtain for new movies in real application. Meanwhile, these tasks treat the very distinctive movie narrating task as a normal video captioning task through some simplifications such as replacing role names with `SOMEONE`, resulting in the inability to connect roles to plots. Secondly, these benchmarks evaluate the generated narrations with ngram-based metrics, which can over-penalize a semantically correct but textually inconsistent narration, especially when there is only one reference available. In addition, these existing datasets are all in English. However,

*Corresponding Author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4669–4684
July 9-14, 2023 ©2023 Association for Computational Linguistics

Figure 1: Data samples from the movie *Goodbye Mr. Loser*. (English translations are provided for easy reading.)

about one-fifth of the world's population speaks Chinese as their mother tongue, of whom more than 17 million are visually impaired (Yu and Bu, 2021). Therefore, building a Chinese movie narration benchmark is necessary.

Intending to address the limitations of the existing narrating benchmarks, in this work, we propose a new benchmark with 101 Chinese movies for movie understanding, named **Movie101**. We collect the movies from the barrier-free channel on Xigua Video platform[1], where normal movies are remastered with ADs. Through automatic process and manual correction, we obtain the ADs and actor lines from the raw videos. We crawl rich meta-information relevant to the movies as well. Finally, Movie101 contains 30,174 narration clips totaling 92 hours, with data samples as shown in Fig. 1. As our investigation shows that narrations mostly occur at those times when no actors are speaking (see Appendix A), to achieve realistic movie narrating, we propose the Movie Clip Narrating (MCN) task that requires a model to narrate where there are no lines. It brings a potential benefit for identifying where to narrate in an unlabeled new movie, since the timestamps of the actor lines are easily accessible[2]. Meanwhile, in order for the audience to accurately comprehend the role-related plots, concrete role names should be contained in the generated narration. For the MCN task, we reorganize the Movie101 dataset, merging the narration clips between two actor dialogues into a longer clip, to simulate real-scenario movie narrating. We thus obtain 14,109 long clips of variable length for narration generation. Moreover, to better evaluate the quality of model-generated narrations, we conduct human evaluations and design a new metric specific to movie narrating, namely Movie Narration Score (MNScore), which well aligns with human evaluation. In addition to the MCN task, our dataset also supports the Temporal Narration Grounding (TNG) task, which asks a model to locate target clips in the movie according to some text descriptions. For both tasks, we benchmark the performance of existing methods, and further propose our improved models by incorporating auxiliary external knowledge. In addition to MCN and TNG tasks, Movie101 can also potentially support other movie understanding tasks such as visual question answering and action recognition, etc.

The main contributions of this paper are as follows: 1) We propose a new benchmark for movie understanding, Movie101, with a large number of video-aligned text descriptions in Chinese. 2) We propose two primary tasks, MCN and TNG, and a new narrating evaluation metric MNScore, where MCN is more in line with the needs of actual movie narrating, while MNScore is more consistent with human evaluation. 3) We benchmark state-of-the-art models and propose improved models enhanced by external knowledge for MCN and TNG, respectively. We expect our proposed Movie101 benchmark can inspire more explorations on narrating and understanding a whole movie.

2 Related Works

Datasets. Existing datasets to support the automatic narration generation task include M-VAD (Torabi et al., 2015) and MPII-MD (Rohrbach et al., 2015), which are merged into LSMDC (Rohrbach et al., 2017). M-VAD, which is collected based on an automatic AD segmentation and alignment method, contains 47K videos from 92 DVDs,

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

[2] The timestamps of the lines can be obtained from the movie script or by automatic methods such as OCR and ASR.

with an average length of 6.2s, each with an aligned narration. MPII-MD contains 68K videos from 94 movies with an average duration of 3.9s, about half of which come with paired scripts and the other half with paired ADs. In addition to movies, TV shows are also good data sources for automatic narration generation. Lei et al. (2020) propose TV Show Caption (TVC), a variant of TV Show Retrieval (TVR). It contains 11K short videos averaging 9.1s in length, and 26K captions describing the visual content, dialogues, and subtitles. All the existing datasets are in English.

Video Captioning. As a classic vision and language task, the video captioning task requires a model to generate natural language descriptions for given videos. Solutions for normal video captioning go through stages from pre-designed templates (Kojima et al., 2002; Guadarrama et al., 2013) to sequence-to-sequence generation with deep neural networks (Pasunuru and Bansal, 2017). A challenging variant for this task is dense video captioning (Krishna et al., 2017), which requires the generation of multi-sentence descriptions for long multi-event videos. The two-stage generation approach, which firstly performs proposal detection on the video and then generates descriptions for each proposal separately, has been the dominant approach (Krishna et al., 2017; Park et al., 2019; Rohrbach et al., 2014; Xiong et al., 2018). Recently, some works avoid event detection and generate paragraph descriptions directly based on the video, such as the one-stage paragraphing model (OVP) (Song et al., 2021), obtaining competitive performance compared to previous works, inspired by which we propose our knowledge-enhanced movie narrating model. Identity-aware video description that distinguishes different persons is more practical in real applications. Park et al. (2020) attempt to achieve role-aware movie narrating by distinguishing different people using labels such as `PERSON1`, `PERSON2`, etc. However, it fails to generate concrete role names and falls short in terms of practicality.

Temporal Sentence Grounding. The temporal sentence grounding (TSG) task aims to localize the moment in a video based on a natural language query (Gao et al., 2017). A two-step pipeline has been the mainstream approach, which first produces a large number of moment candidates via sliding windows, then ranks them with their similarity to the query sentence. The following works try to improve the grounding performance by enhancing interaction between video and query modalities (Liu et al., 2021; Li et al., 2022) or introducing novel detection heads (Lei et al., 2021; Zhang et al., 2020a). Specifically, for interaction methods, Liu et al. (2021) adopt an Iterative Alignment Network (IA-Net) to iteratively interact inter- and intra-modal features within multiple steps. Li et al. (2022) explicitly decompose video and query into multiple structured hierarchies and learn fine-grained semantic alignment among them. In this work, we propose to incorporate external knowledge based on the IA-Net model structure.

3 Dataset

3.1 Data Collection

Movie Acquisition. To the best of our knowledge, there are only a handful of platforms that provide accessible movies in Chinese. The barrier-free channel of Xigua Video is one such platform that provides over 100 accessible movies online, and new movies are still being released that can support further expansion of our dataset. From Xigua Video, we collect all 101 movies available to date and crawl as much meta information as possible for each movie, including title, introduction, genres, directors, actors, etc. We emphasize actors in particular, including actor names, role names, actor portraits, role rankings, and other information about important roles. We expect such information can benefit the movie narrating task and general movie understanding tasks.

Narrations and Lines Extraction. As the movie lines and narrations are only available in the subtitle and audio format respectively from the platform, we therefore leverage OCR and automatic speech recognition (ASR) tools for transcription. For lines, we extract text from subtitles by open-source OCR toolkit PaddleOCR[3] at 2.4 FPS, and manually remove the irrelevant subtitles from the beginning and the end of each movie. For narrations, we extract the audio track from the movie and utilize the ASR service provided by iFlyTek[4], which detects the speech in the audio and transcribes it into text. In addition, the service supports identifying different speakers, which helps discriminate the narrator from the actors. However, the ASR service is not perfect, and its outputs contain errors such as wrong characters, unreasonable sentence breaking, and misidentification of narrations as movie

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

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

Table 1: Movie101 and other Movie Narrating and Temporal Sentence Grounding datasets. (* indicates statistics based on Chinese characters.)

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

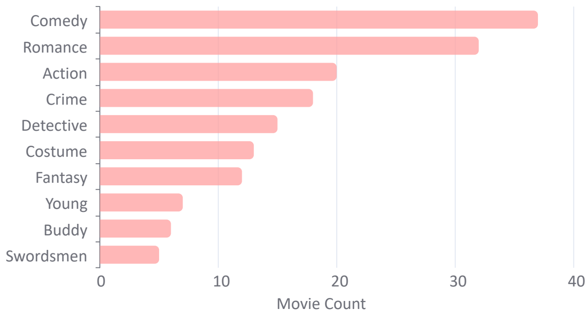


Figure 2: Distribution of movie genres.

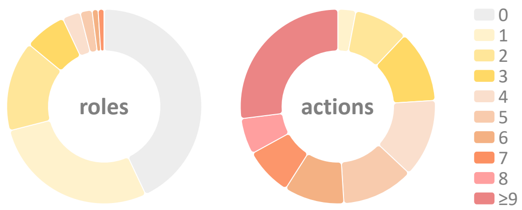


Figure 3: Distribution of the number of role names and the number of actions in each narration in Movie101.

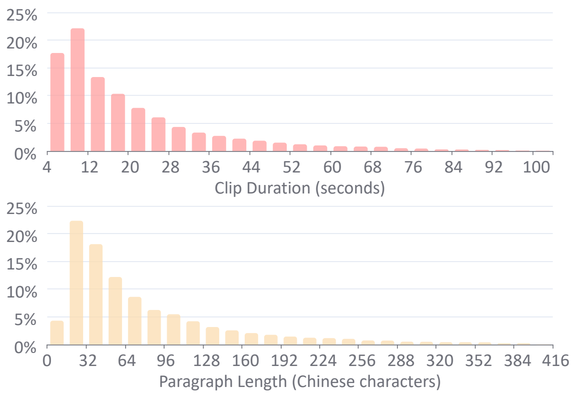


Figure 4: Length and duration distribution of narration clips in Movie101-N.

dialogues, etc. Therefore, we recruit human annotators to further correct the ASR transcription errors and remove non-narration texts manually to improve the data quality. We also delete the irrelevant fragments at the beginning (e.g., movie synopsis, cast introductions) and the summary narration at the end. For coherency, we further organize the narration fragments at the clip level. We merge every two fragments if their temporal gap is less than 1 second. we also apply a paragraph-length threshold of 100 characters to limit over-merging to avoid excessively long clips. We take punctuation into account as well, for example, a period in Chinese is likely to mean the end of a narrative paragraph. Further detailed descriptions of data quality can be found in Appendix B.

Movie101-N and Movie101-G. For real-life movie narrating, models are expected to narrate in the breaks between different actor dialogues. Thus, we reorganize Movie101 to fit this task format. Concretely, we first merge the independent lines in Movie101 into dialogues, where two lines with a temporal gap shorter than 5 seconds are considered to belong to one dialogue. Then, we merge all the narration clips between two adjacent dialogues into a long paragraph. In this way, we obtain **Movie101-N** with narration paragraphs separated by dialogues, which well simulates the practical narrating challenge. Meanwhile, with rich video-text pairs in Movie101, we create another variant dataset to support the temporal grounding tasks, named **Movie101-G**, where narrations are taken as queries and aligned videos serve as targets. For validation and testing, we carefully select 10 movies of different genres for each respectively.

3.2 Dataset Statistics

Movie Properties. Movie101 contains 101 movies, involving 41 genres (a movie can belong to up to 4 genres) and 645 roles in total. Fig. 2 shows the numbers of movies in the top 10 most popular genres, with comedy, romance, and action in the top 3.

Clip Properties. Movie101 contains a total of

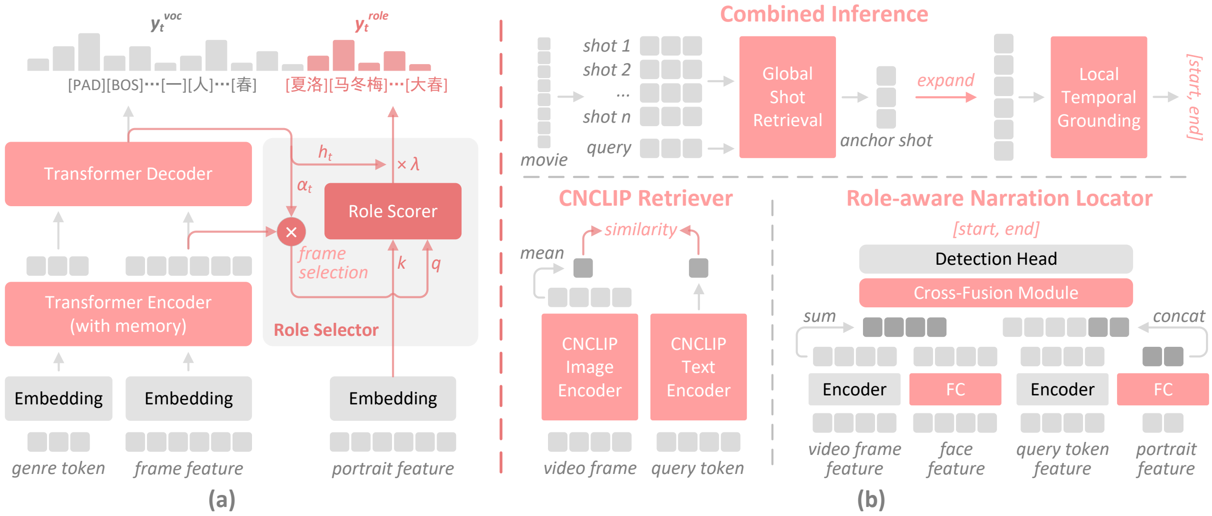


Figure 5: The frameworks of our models. (a) Role-pointed Movie Narrator (RMN) for the Movie Clip Narrating task, and (b) Global Shot Retrieval + Local Temporal Grounding for the Temporal Narration Grounding task.

30,174 short narrated clips with an average duration of 11.0 seconds and an average length of 47.3 Chinese characters. From the narrations, we locate role names and action words with the metadata and the Chinese Part-Of-Speech (POS) tagging toolbox HanLP[5], to detail role and action content in narrations. Fig. 3 shows the distribution of the number of role names and actions per single clip. The narrating variant dataset Movie101-N contains 14,109 long narration clips of an average length of 20.4 seconds and 80.7 characters. The comparison in Table 1 shows that Movie101-N contains much longer video clips and text descriptions than existing movie narrating datasets, while the length distribution in Fig. 4 indicates that the clip length varies a lot. Movie101-G contains 30,174 clips to be located from 101 movies. The average video length of 6,144 seconds also greatly exceeds existing TSG datasets.

4 Movie Clip Narrating

4.1 Task Description

In order to help the visually impaired keep up with the plot in the movie, we first propose a Movie Clip Narrating (MCN) task, which aims to generate a plot-related paragraph description given a clip in Movie101-N. Besides, the narration styles may vary across different genres of movies. The role portraits are important external knowledge for a model to accurately describe the subject of actions. Thus, we also provide this information in Movie101-N to support the MCN task.

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

4.2 Proposed Method

For the MCN task, with multimodal inputs including video, movie genres, role names, and actor portraits, we propose a Transformer-based (Vaswani et al., 2017) model with an encoder-decoder framework, namely Role-pointed Movie Narrator (RMN), where the encoder mainly encodes video clips and the decoder generates narrations, as shown in Fig. 5 (a).

On the encoder side, taking into account the frame-level visual information, the video clip is embedded into a sequence of frame-level features. To emphasize the roles, we extract face features from each frame and concatenate them to the corresponding frame feature sequentially based on the confidence scores of face detection. With learnable genre embeddings, genres are also represented as a sequence of genre features. After video and genre representation, we apply a Transformer encoder to perform cross-encoding. Then, we follow the One-stage Video Paragraphing model (OVP) (Song et al., 2021) to use a dynamic memory bank to refine the video-part representations, which updates at each decoding step.

On the decoder side, in addition to the Transformer decoder, we enable the model to directly choose a complete role name from the movie cast according to context during token-by-token generation via a pointer network (Gu et al., 2016). At the decoding step t, with the decoder hidden state h_t, we first calculate the token scores y_t^{voc} among normal vocabulary. Then we design a Role Selector module to get the name scores among external

Table 2: Accuracy of metrics in terms of their assessment of the candidate narrations against human assessment. (Acc.: accuracy; Info.: informativeness; Qual.: textual quality)

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**

role vocabulary. Concretely, with the decoder's video-part attention distribution α_t, we perform a weighted summation among video representations to get a context-filtered video feature. Then the role scores y_t^{role} are computed with the context-filtered video feature as query and portrait features as key. Finally, the prediction distribution at step t is calculated as follows:

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

where $[;]$ means concatenation, λ is a gate computed from h_t, $f()$ is the softmax function.

4.3 Evaluation

Existing movie narration benchmarks directly adopt ngram-based metrics including CIDEr, BLEU, and METEOR as in normal video captioning. However, there are pitfalls for these metrics, such as underestimating semantically correct but textually inconsistent phrases, which have been widely reported (Zhang et al., 2020b; Shi et al., 2022). For movie narrating, a movie clip can be narrated in multiple expressions, while there is only one reference. Thus, text matching is inadequate to measure the quality of a narration paragraph.

To better evaluate the generated narrations in the MCN task, we conduct a manual evaluation to investigate how humans assess different narrations. We randomly select 30 movie clips, each with 5 candidate narrations, of which 3 are derived from the predictions of different models and 2 are obtained by disturbing the ground truth narrations. Next, we recruit 10 annotators to individually rank the candidates for each video in terms of accuracy, informativeness, and textual quality. Accuracy defines how the narration accurately describes the video, especially roles, actions, and objects; informativeness defines how richly the narration reveals the video content; textual quality is determined by the narration fluency and grammatical correctness.

With the human evaluation results, we investigate a wide range of objective metrics as follows: (1) State-of-the-art video captioning metrics based on deep neural networks including CLIPScore (Hessel et al., 2021), BERTScore (Zhang et al., 2020b) and EMScore (Shi et al., 2022), which are reported outperforming ngram-based metrics in video captioning evaluation; (2) Textual quality metrics including n-grams diversity(Shetty et al., 2017) (DIV) and causal language model perplexity (PPL); (3) F1 score of role name generation (RoleF1). For every two candidate narrations of a video, we use human ranking as a reference to determine whether these metrics correctly judge which of the two candidates is better or worse, and the accuracy is used for evaluating metrics' correlation with human judgment. Finally, we settle on a new metric Movie Narration Score (MNScore) as follows:

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

where mns, ems, $berts$ and $rf1$ refer to MNScore, EMScore, BERTScore and RoleF1, respectively. As shown in Table 2, BERTScore outperforms ngram-based metrics in narration evaluation accuracy, while our new proposed MNScore achieves the best alignment with human evaluation. More details about the implementation of the candidate narrations and the above metrics are presented in Appendix C.

4.4 Experiments

Implementation Details. In our proposed method, models are trained with next-token language modeling by the maximum likelihood estimation (MLE) objective. For videos, we use CLIP (Radford et al., 2021) pre-trained on large-scale image-text pairs and MIL-NCE (Miech et al., 2020) pre-trained on HowTo100M videos (Miech et al., 2019) to extract frame-level CLIP and S3D features with dimensions of 512 and 1024, respectively, at 1 FPS, and further concatenate them. For faces in video frames and portraits, we use the Arcface model (Deng et al., 2019) pre-trained on MS1M (Guo et al., 2016) to extract face features. When there are insufficient faces detected within a frame, the

Table 3: Movie Clip Narrating Performance on Movie101-N. (f_v: face features from the video; g: movie genres)

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
RMN	✓		0.154	0.186	**0.240**	18.97
RMN	✓	✓	0.154	**0.188**	0.238	**19.07**

Table 4: Global Shot Retrieval performance of the first-stage model on Movie101-GSR(temp).

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

face feature extractor compensates by substituting the extracted features with zero-vectors.

Results & Analysis. We choose Vanilla Transformer (Zhou et al., 2018) and state-of-the-art video paragraphing model OVP (Song et al., 2021) as the MCN baselines.

As shown in Table 3, RMN outperforms the baselines by a large margin, especially on RoleF1. This indicates that our model learns to generate role names from external knowledge with the help of the pointer network. To verify the contribution of the genre and face representations in our RMN model, we also perform an ablation study by progressively adding these representations as input. From the results, face features extracted from video frames bring significant gains in role awareness, which shows that using face features to bridge the video content and external actor portraits is beneficial for generating role-related narrations. Qualitative results can be found in Appendix D.

5 Temporal Narration Grounding

5.1 Task Description

To help people locate clips of interest during movie entertainment, an AI agent should be able to understand users' intentions and locate the target clips. To achieve this goal, we propose the Temporal Narration Grounding (TNG) task. Given a clip narration as the query, TNG aims to predict the starting and ending time of the clip in the whole movie.

5.2 Proposed method

Existing temporal sentence grounding models can hardly handle an entire movie input with limited computational resources. Thus, we propose a two-stage framework for the TNG task, with global shot retrieval to coarsely locate the target clip in the first stage and local temporal grounding to finalize the precise timestamp of the target clip in the second stage, as shown in Fig. 5 (b).

Table 5: Local Temporal Grounding performance of the second-stage models on Movie101-LTG(temp). (f_v and f_t refer to adding face features to the video and text representations, respectively.)

Model	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**

Global Shot Retrieval. To find the approximate location of the target, we treat it as a text-video retrieval subtask. We divide a movie into 20s-long shots, and the shot with the highest similarity to the text query will be used as the anchor for further grounding in the second stage. For training such a retrieval system, we construct a temporary dataset Movie101-GSR(temp). Concretely, after cutting the movie into shots, each shot and each annotated narration in Movie101 are judged with the temporal overlap whether they can be considered as an aligned video-text pair.[6]

We build the retrieval model by transferring a Chinese Vision-Language Pre-training (VLP) model ChineseCLIP (Yang et al., 2022) (CNCLIP) from image-text to video-text. Specifically, the shot frames are separately encoded as image features by the visual encoder of CNCLIP, and the final video feature is obtained by performing mean pooling over the `CLS` tokens of all frames. We then perform contrastive learning between the video and text features on Movie101-GSR(temp) to fine-tune the modified CNCLIP.

Local Temporal Grounding. After obtaining the anchor shot in the first stage, we further lo-

[6]A shot and a narration with a temporal overlap larger than half of the duration of either the shot or the narration are regarded as aligned.

Table 6: Combined inference performance of our proposed two-stage method on Movie101-G.

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

calize the target clip within a 200-second window around the anchor shot. This requires the temporal sentence grounding in a 200s-long movie clip, where comprehending the actions of different roles is critical. Therefore, based on the state-of-the-art TSG model IA-Net (Liu et al., 2021), we propose Role-aware Narration Locator (RNL). With a bidirectional GRU (Chung et al., 2014) visual encoder, we encode the input frame features to get temporal context-aware frame representations V. We in addition extract face features from the frames and encode them with a fully connected (FC) layer to filter key face information F. Then we finalize the visual representation by summing V and F. For text encoding, to relate role names in the text query with roles in the video, we extract face features from the portraits that correspond to the role names and encode them as visual token representations with a FC layer, which are then concatenated to the query's textual token representation sequence. During training, for each target, we randomly select a 200s-long clip window that covers the target in each training epoch. We also construct a temporary dataset Movie101-LTG(temp) with fixed window to separately evaluate the second-stage model performance.

5.3 Experiments

Implementation Details. For Global Shot Retrieval, we use average Recall@n ($n \in 1, 5, 10$) to evaluate the retrieval performance on all movies. For Local Temporal Grounding, following previous works (Zhang et al., 2020a), we use "R@n, IoU@m" as metrics, which are defined as the percentage of at least one of top-n proposals having a larger temporal IoU than m with the ground truth. We fine-tune CNCLIP-huge on our Movie101-GSR(temp) for Global Shot Retrieval, and benchmark two code-released state-of-the-art temporal grounding models 2D-TAN (Zhang et al., 2020a) and IA-Net(Liu et al., 2021) on Movie101-LTG(temp) for Local Temporal Grounding. In our RNL model, the video frame, face, and text feature extractors are pre-trained MIL-NCE, Arcface (same as in the MCN task) and BERT-base-Chinese (Devlin et al., 2019), respectively.

Results & Analysis. Table 4 and Table 5 show the performance of models on Global Shot Retrieval and Local Temporal Grounding, respectively. Our RNL outperforms baselines by introducing role-aware video and text encoding, indicating that distinguishing actions of different roles is critical for grounding movie narration. Furthermore, we perform an ablation study to verify the effectiveness of role-aware encoding. As shown in Table 5, adding face features to either video or text representations outperforms our base method IA-Net. RNL with both role-aware video and text encoding achieves the best performance. Table 6 shows the performance of combined inference by Global Shot Retrieval and Local Temporal Grounding. We in addition show the performance of k-way re-ranking, where the top-k shots retrieved in the first stage are respectively used as the anchors in the second stage, and all predictions obtained are re-ranked with their confidence scores. The experimental results show that k-way re-ranking improves Rank@5 performance but harms Rank@1 performance. Qualitative results can be found in Appendix D.

6 Conclusion

In this work, we propose Movie101, a Chinese large-scale video benchmark for movie understanding. To assist visually impaired people in enjoying movies, we propose a more realistic Movie Clip Narrating task to address the automatic movie description issue and design a human-preference-compatible metric MNScore for narrating evaluation. Movie101 also supports the Temporal Narration Grounding task, which is more challenging than the previous TSG benchmarks. Furthermore, our experiments validate the importance of external knowledge including genres and roles for movie understanding. However, there is still a significant gap between our models and expert annotations. This reveals that further research endeavors are still needed to help visually impaired people enjoy movies by AI.

Limitations

Keeping narration coherent within a movie is crucial for visually impaired people to enjoy the movie. In this work, we move a step forward for this target by setting the ground-truth texts in the Movie Clip Narrating task as narration paragraphs and providing longer video clips as inputs. However, how to ensure description coherence across different clips within a movie has not been studied in this work. This requires a higher-level comprehending ability of models to process the whole movie and connect different plots. We leave this to our future investigation.

Ethics Statement

We propose Movie101, a new benchmark to support exploring technologies to benefit the accessibility of the visually impaired. There are two potential ethical issues with our work, regarding data source and crowdsourcing services, respectively. We state each of them as follows:

Data Source. The collected movies are publicly available from Xigua Video, and are allowed to be crawled according to the service contract of the website[7]. Considering the copyright issue, we will only release the url list of movies. Besides, our data source does not contain any information that names or uniquely identifiable individuals or offensive content.

Crowdsourcing Services. We recruited 20 Chinese college students (12 females and 8 males) via social media. For ASR outputs cleaning, workers were required to correct errors in the narration text while watching the movie. For each movie, it took about 2 hours with a payment of 50 RMB ($7.40 USD). To review corrections, for each movie, it took about 30 minutes with a payment of 25 RMB ($3.70 USD). Our payment is fair and reasonable in China, especially since the work is easy and fun. Before the annotation works began, we introduced the future use of the data in the task document to ensure that everyone was informed.

Acknowledgements

This work was partially supported by the National Key R&D Program of China (No.2020AAA0108600) and the National Natural Science Foundation of China (No. 62072462).

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

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 Narration Distribution

Clips where ‘no actors are speaking’ refer to ANY scene wherein no verbal dialogue is being employed by the actors, regardless of whether they are visually present or absent. This definition encompasses, for example, a scene focused solely on a depiction of the sky. We detail the dialogues and narrations in the 101 collected movies. By merging the actor lines, we obtain a total of 15,307 dialogues, constituting 15,206 dialogue gaps with a total duration of 99.4 hours. The 30,174 narration clips we collect fill in 95.3% of the dialogue gaps in terms of quantity and cover 92.9% in terms of duration. Therefore, it is reasonable to assume that where there are no lines, there is a need for narration.

B Dataset Quality Description

We adopt a two-stage annotation process to ensure the quality of the narrations. In the first stage, a group of workers is recruited to clean the data according to our guidelines. In the second stage, another group of workers further checks and corrects the annotation data. Our heuristics used to divide the paragraphs are designed based on our observation experience. We further conduct a manual evaluation of the narration quality. Of the randomly sampled 300 paragraphs, (1) in terms of narration recognition, 96.7% are textually consistent with original ADs; (2) as for the paragraph coherence, 90% maintain complete and coherent semantics, 7.7% should be merged with contexts, and 2.3% should be divided into multiple paragraphs. Thus, the narration is of good quality to support downstream tasks.

C Implementation Details

Candidate Narrations. In Section 4.3, We provide 5 different candidate narrations for each sampled movie clip for human evaluators to rank. These candidates are created as follows:

1. generated by the Vanilla Transformer (Zhou et al., 2018);

Table 7: Key hyperparameters and computational burden for models training.

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

2. generated by the OVP model (Song et al., 2021);
3. generated by our proposed RMN model;
4. generated by disturbing the ground truth with role name removal and replacement;
5. generated by disturbing the ground truth with nouns and verbs replacement.

Metrics Implementation. For CLIP-based metrics including CLIPScore and EMScore, we fine-tune ChineseCLIP-huge (Yang et al., 2022) on our dataset in the same way as in Section 5.2. For each movie clip and generated narration, CLIPScore is calculated with the mean pooled feature of 10 uniformly selected frames and the overall text feature, while EMScore is calculated with all selected frame features and textual token features. For BERTScore, we use the BERT-base-Chinese (Devlin et al., 2019) model checkpoint to calculate, and rescale the raw BERTScore with baseline[8]. For DIV, we calculate 1-gram diversity and 2-gram diversity following Shetty et al. (2017), and average them. For PPL, we obtain the perplexity of each narration with the causal Ernie 3.0 model (Sun et al., 2021) following the calculation of HuggingFace[9]. For RoleF1, we extract role names from the ground truth and the generated narration. We measure how the generated narration covers the roles appearing in the movie clip by Recall; given that these generated role names may also come from the model's hallucination, for example from a wrong movie, we also take Precision into account. Finally, we calculate the F1 score with Precision and Recall.

Hyperparameters and Computation. We detail the key hyperparameters and computational burden for the models training in Table 7. For each model, the results are derived from a single run.

D Qualitative Result

D.1 Movie Clip Narrating

Fig. 6 shows the qualitative results of the MCN task, including the generation results of baselines and our proposed RMN model, and the evaluation results of previous metrics and our proposed MNScore. Vanilla Transformer and OVP can correctly mention some actions but fail to generate correct role names because these roles never appear during training. However, with the help of the Role Selector module, our RMN could well relate roles in video clips with their role names. In addition, these cases demonstrate that our newly proposed MNScore evaluates more consistently with humans.

D.2 Temporal Narration Grounding

Fig. 7 shows the qualitative results of our proposed two-stage method. Through Global Shot Retrieval, we obtain an anchor shot near the target clip from the whole movie, which further helps Local Temporal Grounding to locate the final target.

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

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

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**

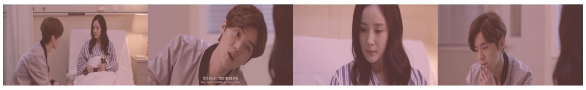
GT: 王多鱼站在保险公司大厦最顶层，穿着红色衣服绿裤衩，一只手背在后面，另一只手扶着巨大的“瘦”字，双腿交叉带着一脸享受闭上双眼，倚靠在“瘦”字上面。镜头缓缓上升拉伸，王多鱼变得越来越渺小，最后完全看不见了。(WangDuoyu stands at the top of the insurance company building, wearing red clothes and green pants, one hand behind the back, the other hand holding the huge “thin” character, legs crossing, with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



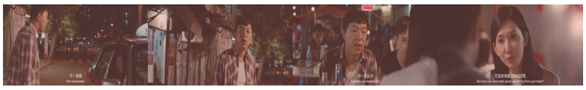
GT: 他凑近小星，得意地捶捶自己胸口，林冲又皱着眉撑起下巴。(He comes close to Xiaoxing, proudly pounding his chest. LinChong then frowns and props up his chin.)

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**

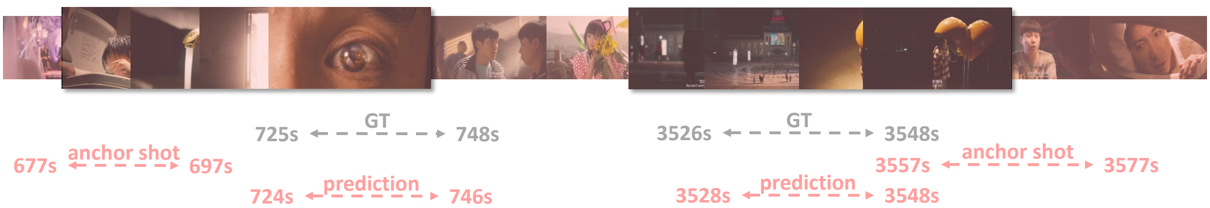
Figure 6: Qualitative Results of the MCN task. (GT: the ground truth; VT: Vanilla Transformer). In the narration texts, green and red characters denote the correctly and wrongly generated role names, respectively. In the tables, metrics in green indicate that the ranking of candidates by the metric is consistent with human ranking, while red indicates inconsistency.

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.)

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.)

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.)

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.)



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.)

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.)

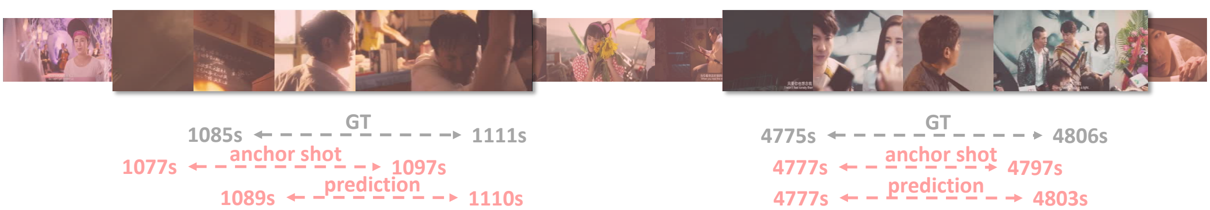


Figure 7: Qualitative results of the TNG task from the movie *Goodbye Mr. Loser*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section: Limitation

☑ A2. Did you discuss any potential risks of your work?
Section: Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section: Abstract, Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section: Introduction, Dataset, Movie Clip Narrating, Temporal Narration Grounding, Appendix

☑ B1. Did you cite the creators of artifacts you used?
Section: Introduction, Dataset, Movie Clip Narrating, Temporal Narration Grounding, Appendix

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section: Dataset, Ethics Statement

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section: Dataset, Ethics Statement

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section: Ethics Statement

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section: Introduction, Dataset, Temporal Narration Grounding, Appendix

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section: Introduction, Dataset

C ☑ Did you run computational experiments?

Section: Appendix

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section: Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section: Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section: Appendix

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section: Dataset, Movie Clip Narrating, Temporal Narration Grounding, Appendix

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Introduction, Dataset, Movie Clip Narrating, Ethics Statement

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We report key information about the requirements for human annotators. (Section: Dataset, Movie Clip Narrating)

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section: Ethics Statement

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section: Ethics Statement

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section: Ethics Statement