MovieUN: A Dataset for Movie Understanding and Narrating

Qi Zhang*, **Zihao Yue***, **Anwen Hu, Ziheng Wang, Qin Jin**†
School of Information, Renmin University of China
{zhangqi1996, yzihao, anwenhu, wzh2019201417, qjin}@ruc.edu.cn

Abstract

Automatic movie narration generation and narration grounding are very important to provide a true movie experience for the blind and visually impaired. To tell the movie story well, it is necessary to mention plot-related details (such as character names) and keep the narrations in a plot coherent. Taking these two points into consideration, we construct a Chinese large-scale video benchmark from 101 movies for **Movie U**nderstanding and **N**arrating (**MovieUN**) to support the Movie Clip Narrating (MCN) task and Temporal Narration Grounding (TNG) task. We split movies in MovieUN into movie clips according to plots, and pair them with corresponding narrations provided by the movie narrators. Ultimately, the TNG task involves 3,253 long video clips totaling 179 hours. The MCN task contains 33,060 video clips totaling 105 hours. We benchmark state-of-the-art video captioning models and temporal grounding models in MCN and TNG tasks, respectively. Furthermore, to accurately comprehend plots of different characters, we propose methods to incorporate portraits of actors as external knowledge in both tasks. The experiment results demonstrate the effectiveness of our proposed methods. The dataset and codes are released at https://github.com/yuezih/MovieUN.

1 Introduction

The estimated number of visually impaired people worldwide is 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 AD are constructed to support the research on automatic AD generation, including the MPII-MD dataset (Rohrbach et al., 2015) and M-VAD dataset (Torabi, 2015), with shot-level ADs or scripts aligned to the movie visual contents. Consequently, different solutions for automatic movie narrating have been proposed based on these datasets (Rohrbach et al., 2017).

However, these efforts mainly focus on generating a single sentence narration for a shot of a few seconds, which gaps with the actual movie narration scenario. Shot-level single-sentence narration generation has some obvious deficiencies. First, it does not consider whether the shot contains something worth narrating and just describes all shots indiscriminately. Second, it does not consider the context of the shot, and thus if simply splicing all the shots together to form the narration of the plot segment, there will be apparent redundancy and incoherence problems. We consider that a high-quality narration should be generated from a plot segment (which we define as a movie clip) rather than the smallest unit of shot in a movie. Therefore, in this paper, we focus on the movie clip narrating task. Moreover, these existing datasets are all in English. However, 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; Wang and Wu, 2021). Contrary to the huge demand, the accessibility of Chinese media content (Zhu et al., 2020) to visu-

*Equal contribution. †Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1873–1885
December 7-11, 2022

Figure 1: Samples of the annotation formats from the movie *Goodbye Mr. Loser*.

ally impaired audiences is minimal. For Chinese movie narrating, the English datasets are of little help due to cultural differences and poor translation quality, given the complexity of discourse in movie narrations. Therefore, building a Chinese movie narration dataset is urgent and meaningful.

Intending to support research on clip-level movie narrating and to fill in the gap in the Chinese movie narration datasets, in this work, we propose a new Chinese movie narration dataset for movie understanding and narrating, named MovieUN, where movies are collected from the barrier-free channel on Xigua Video platform[1]. We crawl 101 movies with ADs in total. As the narrations are in audio format, we leverage the speech recognition tools to automatically transcribe the narration audio. We then design a manual correction procedure to filter the errors based on the automatic transcriptions. We crawl rich meta information relevant to the movie as well. Finally, MovieUN contains 101 movies with 33,060 clip-level narrations totaling 105 hours, with data samples as shown in Fig. 1. Aiming at assisting visually impaired audiences to experience a movie, we believe that two technologies are essential, namely automatic narration generation that generates descriptions of movie content, and automatic movie clip grounding that locates a clip in the movie based on user interest. Therefore, in this paper, we propose two tasks based on MovieUN, the Movie Clip Narrating (MCN) task and the Temporal Narration Grounding (TNG) task. We first benchmark the performance of existing methods on both tasks, and further propose our improved models via incorporating auxiliary external knowledge. In addition to MCN and TNG tasks, MovieUN can also potentially support other tasks such as Visual Question Answering.

The main contributions of this paper are as follows: (1) We propose a new movie narration dataset, MovieUN, which provides a large number of aligned narrations in Chinese. (2) We propose two primary tasks, MCN and TNG, to support the research on clip-level movie understanding and narrating. (3) We benchmark state-of-the-art models on MovieUN and propose improved models enhanced by external knowledge for MCN and TNG, respectively. We expect our proposed MovieUN and the benchmark tasks can contribute to the final realization of the narrating and understanding of a whole movie.

2 Related Works

Datasets. Existing datasets to support the automatic narration generation task include M-VAD (Torabi, 2015) and MPII-MD (Rohrbach et al., 2015). M-VAD, which is collected based on an automatic AD segmentation and alignment method, contains 47k videos from 92 DVDs, with an average length of 6.2s, each with an aligned narration. MPII-MD contains 68k videos from 94 movies with an average length of 3.9s, about half of which come with paired scripts and the other half with paired ADs. In addition to movies, TV shows

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

are also good data sources for automatic narration generation. Lei et al. propose TV Show Caption (TVC), a variant of TV Show Retrieval (TVR) (Lei et al., 2020). 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. Video captioning (VC), as a classic vision and language task, has received much research attention in recent years. Early works identify visual objects from videos and generated descriptions using pre-designed templates (Kojima et al., 2002; Guadarrama et al., 2013). With the advancement of deep learning, efficient video feature aggregation and high-quality description generation have been made possible (Pasunuru and Bansal, 2017). Compared to the traditional video captioning task, dense video captioning is more challenging as it requires generating descriptions of multiple sentences for multiple events in a long video. 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; Senina, 2014; Xiong, 2018). Recently, some works avoid event detection and generate paragraph descriptions directly based on the video (Song et al., 2021), obtaining competitive performance compared to previous works. Inspired by the one-stage paragraph generation work, we propose our knowledge-enhanced movie narration generation model. Identity-aware video description that distinguishes different persons is more practical in real applications. Park et al. (Park et al., 2020) attempts 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 task aims to localize the moment in a video based on a natural language sentence (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 performs final localization based on the similarity between each candidate and 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., 2020). 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 learns fine-grained semantic alignment among them. For detection heads, Zhang et al. (2020) proposes a two-dimensional map, one dimension indicates the starting time of a moment, and the other indicates the end time, to model the temporal relations. Lei et al. (2021) adopt a transformer encoder-decoder model to directly generate moment proposals, which uses a detr-based (Carion et al., 2020) detection head. In this work, we propose incorporating external knowledge based on the IA-Net model structure.

Table 1: MovieUN dataset overal statistics.

Movies	Texts/Avg. text len.	Clips/Avg. len(s).	Long clips/Avg. len(s).
101	33,060/45.3	33,060/11.4	3,253/197.6

3 MovieUN Dataset

3.1 Data Collection

Movie Acquisition. As far as we know, there are only a handful of platforms that provide accessible movies in Chinese. Compared with the platforms such as Netflix or Apple TV, each of which provides thousands of accessible media resources, the Chinese platforms only offer a small number of movies with AD. 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 potential general movie understanding tasks.

Movie Narration Extraction. As the movie narration or AD is only available in the audio format from the platform, we therefore leverage the automatic speech recognition tools to transcribe the audio description. We extract the audio track from

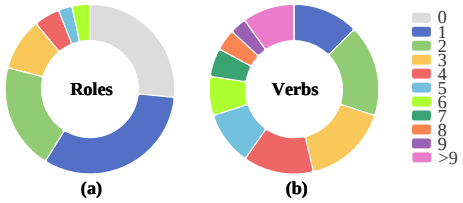


Figure 2: The number distribution of role names and verbs in each paragraph.

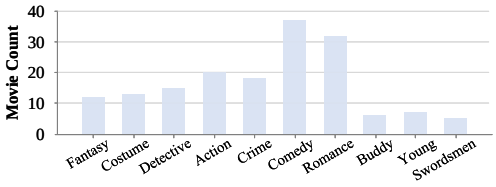


Figure 3: Top 10 movie genres in MovieUN.

the movie and utilize the Automatic Speech Recognition (ASR) service provided by iFlyTek[2], 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 dialogues, etc. We therefore 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 clips at the beginning (e.g., movie synopsis, cast introductions) and the summary narration at the end.

We then further organize the narrations at the clip level. We merge two narration sentences if their temporal gap is less than one second. we also apply a paragraph-length threshold set to 100 to limit over-merging to avoid excessively long paragraphs. We take punctuation into account as well, for example, a period in Chinese is likely to mean the end of a narrative paragraph. We further elaborate on data quality in Appendix B. The various annotation formats are illustrated in Fig. 1.

3.2 Dataset Statistics

Table 1 shows the overall statistics of MovieUN. MovieUN collects a total of 33,060 narration paragraphs, which are annotated with START and END timestamps. The average length of narration paragraphs is 45.3 Chinese characters, and the average duration of corresponding movie clips is 11.4 seconds. Meanwhile, MovieUN provides 3,253 long movie clips as well, with an average duration of 197.6 seconds to support the localization task.

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

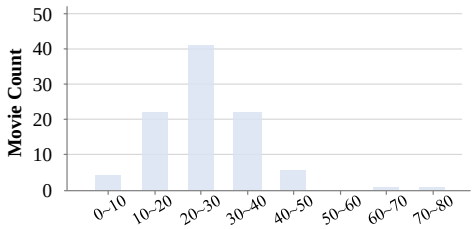


Figure 4: The number distribution of portraits in each movie.

Paragraph Properties. We tagged all words in the dataset adopting the Chinese Part-Of-Speech (POS) and Named Entity Recognition (NER) tagger toolbox HanLP[3]. Figure 2(a) presents the distribution of the number of role names in each paragraph. The vast majority of paragraphs contain at least one role name. Figure 2(b) shows the distribution of number of verbs in each paragraph. 71% of paragraphs involve at least three verbs.

Movie Properties. We analyze the genres of all 101 movies in MovieUN, which involves 41 different genres in total. A movie can have up to four genre tags. Figure 3 shows the top 10 genres, with Comedy/Romance/Action ranking in the top 3, which illustrates the diversity of movies. Meanwhile, Figure 4 presents the distribution of the number of portraits in each movie. Specifically, most movies contain at least 20 portraits of different roles.

4 Movie Clip Narrating

4.1 Task Description

In order to help the visually impaired keep up with the plot of the movie, it is crucial to describe the visual contents when no actors are speaking. Besides, ensuring the descriptions within a plot are coherent plays an important role in understanding the overall story. Thus, we first propose a Movie Clip Narrating (MCN) task, which aims to generate a plot-related paragraph description given a video clip. MovieUN contains 33k (movie clip, narration) pairs to support the MCN task. We name this set MovieUN-N. As shown in Table 2, compared with existing movie-related narrating datasets, MovieUN-N contains much longer video clips and text descriptions, which indicates that MCN tasks on MovieUN-N are more challenging

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

Table 2: MovieUN versus Video Captioning and Temporal Sentence Grounding datasets. * indicates statistics based on Chinese characters.

Task	Dataset	Text type	Videos/Texts	Avg. video len.	Avg. text len.	Avg. actions	Avg. role names	Portraits
MCN	M-VAD	sentence	47k/47k	6.2 sec.	10.8	-	-	0
	MPII-MD	sentence	68k/68k	3.9 sec.	9.6	1.4	0.37	0
	TVC	sentence	109k/262k	9.1 sec.	13.4	1.9	0.75	0
	MovieUN-N	paragraph	33k/33k	11.4 sec.	45.3*	7.0	1.58	2,533
TNG	Charades-STA	sentence	10k/16k	31 sec.	7.2	1.1	0.0	0
	AcitivityNet	sentence	20k/72k	118 sec.	13.5	2.1	0.02	0
	TVR	sentence	22k/109k	76 sec.	13.4	1.9	0.75	0
	MovieUN-G	paragraph	3k/33k	198 sec.	45.3*	7.0	1.58	2,533

for models to comprehend key visual contents and generate coherent narrations. 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 MovieUN-N to support the MCN task.

4.2 Proposed Model

For the MCN task, we propose two models, namely Multimodal Movie Narrator (MMN) and Role-pointed Movie Narrator (RMN).

Multimodal Movie Narrator. With multimodal inputs including video, movie genres (e.g., action, romance), role names, and actor portraits, we first propose an end-to-end Transformer-based model, namely Multimodal Movie Narrator (MMN), as shown in Fig. 5 (a). Taking into account the frame-level visual information, especially the face characteristics, and temporal variance, the video clip is embedded into a sequence of frame-level features. To emphasize the characters, we extract face features from each frame and concatenate them to corresponding frame feature sequentially based on confidence scores of face detection. With learnable genre embeddings, genres are represented as a sequence of genre features. For each role name of the movie, we sum its textual representations and visual portrait representations as external knowledge. After multimodal representation, we apply a Transformer Encoder to learn the cross-modal relationship and a Transformer Decoder to generate narrations. At each decoding step, 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.

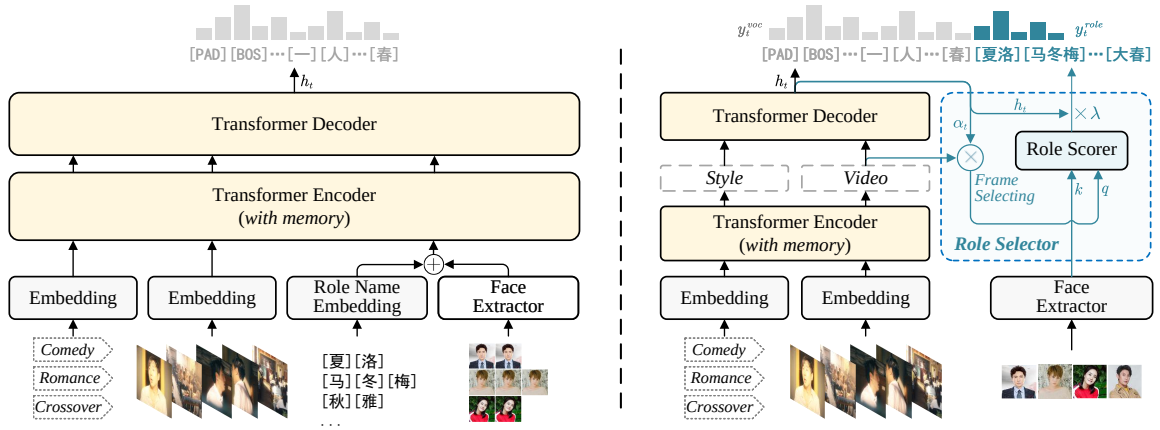
Role-pointed Movie Narrator. While MMN introduces role names as inputs, it still generates a person name token by token. This can distract the model from describing visual contents to organizing person names, which is not the key target of narration generation. Thus, we further propose a Role-pointed Movie Narrator (RMN), which can directly choose a complete role name from the movie cast according to context via a Pointer Network (Gu et al., 2016), as shown in Fig. 5(b). During encoding, different from MMN, only genre and video representations are fed into the Transformer Encoder. At the decoding step t, with the decoder output 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 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:

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

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

4.3 Experiments

Task Settings. We design two task settings to examine the narrating ability of models, namely movie-seen and movie-unseen. For the movie-seen setting, we randomly select 80% of the clips as the training set from each movie, and the remaining 10% and 10% as the validation and test sets. In this setting, clips from train/val/test splits can belong to an identical movie. For the movie-unseen setting, we select 81 movies as the training set, 10 movies each for the validation and test sets. Therefore, in this setting, models are tested on video clips from movies that have never been seen during training.



(a) Multimodal Movie Narrator

(b) Role-pointed Movie Narrator

Figure 5: The architectures of Multimodal Movie Narrator (MMN) and Role-pointed Movie Narrator (RMN) for the Movie Clip Narrating (MCN) task.

Table 3: Movie Clip Narrating Performance on MovieUN-N.

Setting	Method	Accuracy			Diversity		
		BLEU@4↑	METEOR↑	CIDEr↑	Div@1↑	Div@2↑	Rep@4↓
Movie-seen	Vanilla Transformer	3.89	11.06	20.43	60.65	72.19	12.19
	OVP	3.61	10.65	23.35	73.71	83.67	5.88
	MMN	**3.95**	**11.38**	**28.55**	**77.84**	**88.24**	**2.60**
	RMN	3.71	11.18	24.95	75.04	86.89	3.07
Movie-unseen	Vanilla Transformer	1.13	7.60	5.45	67.91	79.17	7.43
	OVP	1.19	7.76	6.08	74.02	84.31	5.83
	MMN	1.11	8.21	6.59	77.46	**89.64**	**2.03**
	RMN	**1.73**	**8.91**	**10.15**	**78.88**	89.16	2.18

Implementation Details. We evaluate the qualities of narration generation in terms of both accuracy and diversity. Following previous works (Song et al., 2021), we use BLEU@4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and CIDEr (Vedantam et al., 2015) to measure accuracy, and n-gram diversity (Div@n) (Shetty et al., 2017) and n-gram repetition (Rep@n) (Xiong, 2018) to measure diversity. Given the ground truth, both MMN and RMN are trained by the maximum likelihood estimation (MLE) objective. For videos, we use CLIP (Radford et al., 2021) pretrained 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. For faces in videos frames and portraits, we use the Arcface model (Deng et al., 2019) pre-trained on MS1M (Guo et al., 2016) to extract face features. More details about model input can be found in Appendix C.1.

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 baselines. As shown in Table 3, firstly, in movie-seen or movie-unseen settings, both MMN and RMN outperform baselines in terms of narration accuracy and diversity. This validates the effectiveness of our models for narration generation. Secondly, RMN outperforms MMN in the movie-unseen setting but underperforms it in the movie-seen setting. This shows that compared with MMN, RMN is poorer at remembering some empirical knowledge (e.g., role names) but better at describing video clips of a new movie with the help of the pointer network.

To verify the contribution of different-modality representations in our MMN and RMN models, we perform an ablation study by progressively adding these representations as input. As shown in Table 4, adding representations of each modality brings performance improvement for MMN and RMN. In particular, for RMN, face features extracted from video frames bring significant gains in narration accuracy and diversity. This shows that using face features to bridge the video content and external

Table 4: Ablation study about multimodal inputs. (*n*: role names, *p*: face features from external portraits, *v*: face features from video frames, *g*: movie genres.) MMN and RMN are tested in the movie-seen setting and movie-unseen setting, respectively.

Model	Multimodal input				CIDEr↑	Rep@4↓
	n	*p*	*v*	*g*		
MMN	✓				25.36	4.44
	✓	✓			26.64	3.35
	✓	✓	✓		27.47	2.79
	✓	✓	✓	✓	**28.55**	**2.60**
RMN	✓	✓			7.78	8.94
	✓	✓	✓		**10.66**	3.62
	✓	✓	✓	✓	10.15	**2.18**

actor portraits is critical for generating role-related narrations. More qualitative results can be found in Appendix D.1.

5 Temporal Narration Grounding

5.1 Task Description

To help visually impaired people revisit clips of interest to them during movie entertainment, an AI agent should be able to understand users' intentions and locate the target movie clips. To achieve this goal, we propose a Temporal Narration Grounding (TNG) task. With a paragraph narration as the query, TNG aims to predict its starting and ending time in an untrimmed video. Based on *MovieUN*, we build *MovieUN-G* to support the TNG task.

Concretely, we first split each movie into 200-second clips and discard ones without narrations. Then, we adjust the clip boundaries to ensure that all narrations in them are complete. After such processing, MovieUN-G contains 3,253 long video clips. Finally, follow the movie-unseen setting in the MCN task, we split the 3,253 long clips into 2,611/328/314 for training, validation, and test. A clip contains an average of 10 queries.

There are two major differences between MovieUN-G and existing Temporal Sentence Grounding (TSG) datasets. As shown in Table 2, firstly, text queries in MovieUN-G contain much more actions and role names, which suggests that understanding the correspondence between roles and actions is crucial for the TNG task. To aid in learning such relationships, we additionally provide portraits of the characters, which are ignored in previous datasets. Secondly, video clips of MovieUN-G are much longer than ones in existing datasets while the target scale (duration proportion of the target interval in a video clip) is smaller, as shown in Figure 7, which suggests that the model should understand longer video content at a finer-grained level in order to locate shorter target intervals. In general, the TNG task based on MovieUN-G is more challenging than existing Temporal Sentence Grounding tasks.

5.2 Proposed Model

Comprehending the actions of different roles is critical for Temporal Narration Grounding. Therefore, based on state-of-the-art TSG model IA-Net (Liu et al., 2021), we propose a Role-aware Narration Locator (RNL), as shown in Figure 6.

Role-aware Video Encoding. Each video clip is firstly embedded into a sequence of frame features $V = \{v_1, v_2, \cdots, v_L\}$ with CLIP and MIL-NCE as MMN/RMN, where L is the total number of video frames. Similar to the solution for the MCN task, we extract face features from the frames and perform dimension reduction by a fully-connected layer to filter key face information. The sequence of face representations for the video clip is noted as $F = \{f_1, f_2, \cdots, f_L\}$. Finally, by adding V and F, the video clip is encoded into a sequence of role-aware video representations $\hat{V} = \{\hat{v}_1, \hat{v}_2, \cdots, \hat{v}_L\}$.

Role-aware Text Encoding. The paragraph query is tokenized into N tokens and then represented as a sequence of contextualized text features $Q = \{q_1, q_2, \cdots, q_N\}$ by the Bert-Base-Chinese model (Devlin et al., 2019). To relate a role name in the text query with a character in the video, it is important to introduce external vision knowledge about the appearance of the actor playing the role. Therefore, we firstly extract role names in the text query by Named Entity Recognition with HanLP. By leveraging metadata in MovieUN, we relate these role names with portraits of actors. Then, each portrait is represented as a face feature and fed to a fully-connected layer. By adding all tokens in role names with corresponding portrait representations, the text query is encoded into a sequence of role-aware text representations $\hat{Q} = \{\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_k\}$.

Grounding Module. IA-Net consists of two main components: cross-fusion module and detection head. Video representations $\hat{V}$ and text representations $\hat{Q}$ are firstly fed to the cross-fusion module, which iteratively interacts inter- and intra-modalities in multiple steps for semantic alignment. Then, the anchor-based detection head is applied

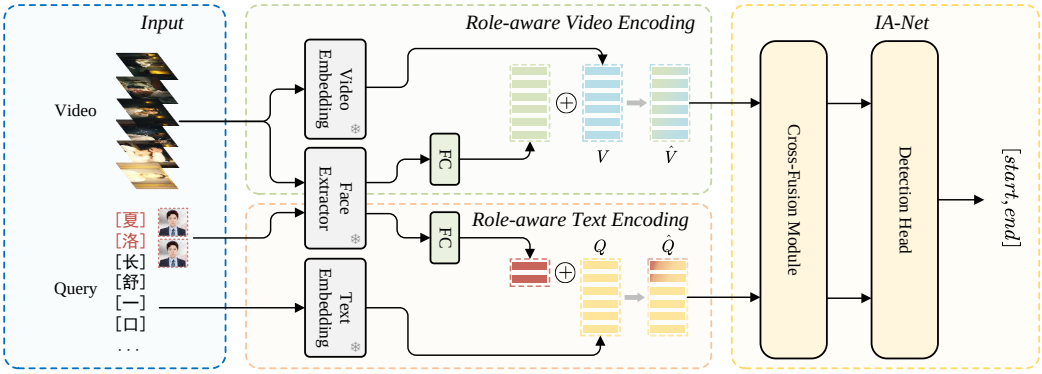


Figure 6: The architecture of Role-aware Narration Locator (RNL) for the Temporal Narration Grounding task.

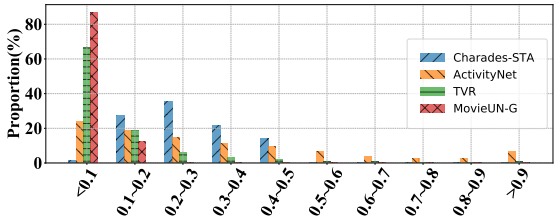


Figure 7: Target scale distribution of different Temporal Grounding datasets.

Table 5: Temporal Narration Grounding performance on MovieUN-G validation and test set.

Set	Method	mIoU	IoU=0.3		IoU=0.5	
			R@1	R@5	R@1	R@5
val	2D-TAN	20.63	29.86	58.47	23.65	48.72
	IA-Net	22.43	32.84	63.56	24.35	49.63
	RNL	**24.17**	**34.82**	**67.26**	**26.97**	**53.10**
test	2D-TAN	16.66	23.98	54.11	18.56	43.25
	IA-Net	20.38	30.20	59.91	22.14	45.66
	RNL	**22.21**	**32.32**	**64.03**	**24.59**	**49.18**

to predict the confidence scores of all anchors and corresponding temporal offsets.

5.3 Experiments

Implementation Details. Following previous works (Zhang et al., 2020; Liu et al., 2021), we use "R@n, IoU@m" and mIoU (the average temporal IoU) as the metrics. "R@n, IoU@m" is defined as the percentage of at least one of top-n proposals having a larger temporal IoU than m with the ground truth. Meanwhile, we benchmark two code-released state-of-the-art Temporal Grounding models (Zhang et al., 2020; Liu et al., 2021) on our MovieUN-G dataset. The training and inference settings of these models (Zhang et al., 2020; Liu et al., 2021) are the same as their officially released ones on the ActivityNet benchmark. In order to make a fair comparison in experimental evaluations, our model settings are the same as IA-Net, in which the maximum number of video frames is set to 200, and two complementary losses are used: a binary cross-entropy loss for confidence evaluation and a smooth L1 loss for regressing the IoU.

Table 6: Ablation study of role-aware encoding. video-face and text-face refer to adding face features to video representations and text representations, respectively.

Method	video-face	text-face	mIoU
IA-Net	-	-	20.38
RNL	✓		21.43
		✓	21.87
	✓	✓	22.21

Results & Analysis. Table 5 shows the grounding performance of our model RNL and state-of-the-art Temporal Sentence Grounding models 2D-TAN (Zhang et al., 2020) and IA-Net (Liu et al., 2021). Firstly, IA-Net outperforms 2D-TAN in the TNG task with the help of an Iterative Alignment Network, which can better encode complicated intra- and inter-modality relations. Secondly, RNL further outperforms IA-Net by introducing role-aware video and text encoding. This shows 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 6, adding face features to either video or text representations outperforms the baseline IA-Net. RNL with both role-aware video encoding and role-aware text encoding achieves the best performance. More qualitative results can be found in Appendix D.2.

6 Conclusion

In this work, we propose MovieUN, a Chinese large-scale video benchmark for movie understanding and narrating. To assist visually impaired people in enjoying movies, MovieUN supports two

challenging tasks, namely Movie Clip Narrating (MCN) and Temporal Narration Grounding (TNG). In both tasks, we design role-aware models that outperform strong baselines. Furthermore, our experiments validate the importance of movie genres and external actor portraits for movie understanding and narrating. 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.

Limitation

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 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 is not studied in this work. This requires models to process the whole movie and a higher-level comprehending ability to link different plots. We leave this to study in the future.

7 Acknowledgement

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).

References

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proc. of ECCV*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proc. of CVPR*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: temporal activity localization via language query. In *Proc. of ICCV*.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proc. of ACL*.

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 *Proc. of ICCV*.

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 *Proc. of ECCV*.

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*.

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*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proc. of ICCV*.

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. *Proc. of NeurIPS*.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Proc. of ECCV*.

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 *Proc. of CVPR*.

Daizong Liu, Xiaoye Qu, and Pan Zhou. 2021. Progressively guide to attend: An iterative alignment framework for temporal sentence grounding. *ArXiv preprint*.

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 *Proc. of CVPR*.

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 *Proc. of ICCV*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Jae Sung Park, Trevor Darrell, and Anna Rohrbach. 2020. Identity-aware multi-sentence video description. In *Proc. of ECCV*.

Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. 2019. Adversarial inference for multi-sentence video description. In *Proc. of CVPR*.

Ramakanth Pasunuru and Mohit Bansal. 2017. Multi-task video captioning with video and entailment generation. In *Proc. of ACL*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proc. of CVPR*.

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*.

Anna Senina. 2014. Coherent multi-sentence video description with variable level of detail. *ArXiv preprint*.

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 *Proc. of ICCV*.

Yuqing Song, Shizhe Chen, and Qin Jin. 2021. Towards diverse paragraph captioning for untrimmed videos. In *Proc. of CVPR*.

Atousa Torabi. 2015. Using descriptive video services to create a large data source fo. *ArXiv preprint*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proc. of CVPR*.

Min Wang and Dan Wu. 2021. Ict-based assistive technology as the extension of human eyes: technological empowerment and social inclusion of visually impaired people in china. *Asian Journal of Communication*.

Yilei Xiong. 2018. Move forward and tell: A progressive generator of video descriptions. *ArXiv preprint*.

Chun Yu and Jiajun Bu. 2021. The practice of applying ai to benefit visually impaired people in china. *Communications of the ACM*.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proc. of AAAI*.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. In *Proc. of CVPR*.

Yutao Zhu, Ruihua Song, Zhicheng Dou, Jian-Yun Nie, and Jin Zhou. 2020. ScriptWriter: Narrative-guided script generation. In *Proc. of ACL*.

A USABILITY ASSESSMENT OF ENGLISH DATASETS

To deal with the lack of Chinese movie narration datasets, one might consider a straightforward solution, which is to translate the available English movie narration datasets. We therefore conduct an experiment to verify whether Chinese movie narration data can be obtained simply by translating existing English datasets. Concretely, we automatically translate the English narrations in LSMDC (containing M-VAD, MPII-MD mentioned in Related Works) into Chinese via Youdao's text translation API [4], and then perform paragraph merging to obtain an average length of 11.4s (the same as our MovieUN-N), constructing a translation-based Chinese movie clip narration dataset LSMDC-Ch containing 23,504 clips. We then train our baseline model OVP on LSMDC-Ch and test it on MovieUN-N. As expected, it performs worse than the same model trained on MovieUN-N under the movie-unseen setting, as shown in Table 7, which indicates that it is necessary to construct a Chinese movie narration dataset to support follow-up research.

Table 7: Movie Clip Narrating Performance of the OVP model trained on LSMDC-Ch and MovieUN-N.

Training Set	BLEU@4	METEOR	CIDEr
LSMDC-Ch	0.68	6.87	2.81
MovieUN-N	**1.19**	**7.76**	**6.08**

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 is 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 in good quality to support our two tasks.

[4] https://ai.youdao.com/product-fanyi-text.s

C ADDITIONAL IMPLEMENTATION DETAILS

C.1 Movie Clip Narrating

The maximum length of the paragraph narration is set to 100. For each video clip, up to 25 frames are selected as input. For each frame, at most 3 face features are concatenated to the frame representation, if there are less than 3 face features, zero vectors are padded. Up to 4 genres are used as input. For MMN, at most 30 role names with portraits are fed to Transformer Encoder. For RMN, the maximum length of the external role name dictionary is set to 10.

D QUALITATIVE RESULT

D.1 Movie Clip Narrating

Qualitative results in the movie-seen setting are shown in Fig. 8(a) and (b). All of these models are able to correctly generate role names token by token because these roles also appear in the training set. Compared to VT and OVP, MMN describes more relevant actions for different characters.

Fig. 8(c) and (d) show the generation results of baselines and our proposed RMN model in the movie-unseen setting. Under this setting, VT and OVP can correctly mention some actions but fail to generate correct role names because these characters have never appeared during training. However, with the help of the Role Selector module, our RMN could well relate characters in video clips with their role names.

Overall, while our models outperform strong baselines in this task, the quality of auto-generated narrations is still far from expert annotations.

D.2 Temporal Narration Grounding

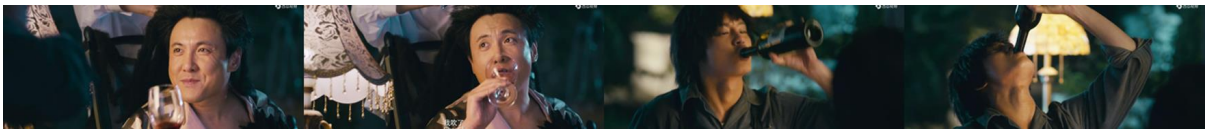
Figure 9 shows qualitative results of our RNL model and strong baseline IA-Net. While IA-Net applies an Iterative Alignment Network to encoding intra- and inter-modality relationships, it still performs poorly at relating actions to different characters since it does not know the appearance of characters. By emphasizing face features in videos and relating role names in the text query with actor portraits, our Role-aware Narration Locator can better locate the target clip.

VT: 她转身走到柜子前，拿起手机看了眼陈姗姗，随后转身离开了，她走到桌子上，她把手中的药瓶放在桌子上，她拿起桌子上的药瓶和杯子，
(*She turns around and walks to the cabinet, picks up her phone, looks at Chen Shanshan, then turns around and leaves. She walks to the table; she puts the medicine bottle in her hand on the table; she picked up the medicine bottle and the cup on the table,...*)
OVP: 叶薰笑着点了点头，随后转身走向厨房，随后她先将桌子上的茶杯递给陈姗，随后拿起桌上的一瓶红酒放在茶几上，她看着桌上的瓶子
(*Ye Xun smiles and nods, then turns towards the kitchen, then she first hands the teacup on the table to Chen Shan, then picks up a bottle of red wine on the table and puts it on the coffee table; she looks at the bottle on the table...*)
MMN: 张代晨微笑着看了看陈姗姗，随后转身离开，陈姗姗和张代晨来到浴室，她走到柜子前拿起一瓶酒，冲着柜子打了个招呼，陈姗姗走到
(*Zhang Daichen smiles, looking at Chen Shanshan, then turns around and leaves. Chen Shanshan and Zhang Daichen come to the bathroom; she walks to the cabinet and picks up a bottle of wine, and snaps at the cabinet; Chen Shanshan walked to...*)
GT: 张代晨带陈姗姗走进屋子，随后拿着药膏走向冰箱，她从冰箱里拿出一瓶饮料，然后又转身去拿杯子，
(*Zhang Daichen takes Chen Shanshan into the house, then walks to the refrigerator with the ointment; she takes out a bottle of drink from the refrigerator, then turns around to get a cup.*)

(a)

VT: 黄达看到叶薰的身影，叶薰看着他，叶薰笑笑了笑，转身离开了酒店，黄达看着叶薰，黄达坐在椅子上的椅子上。
(*Huang Da looks at the figure of Ye Xun, and Ye Xun looks at him; Ye Xun smiles, turns around, and leaves the hotel; Huang Da looks at Ye Xun; Huang Da sits in the chair on the chair.*)
OVP: 黄达说完之后，转身离开了酒店，叶薰看着黄达离开的背影，双手叉腰，默默拿起一只手在腰间，黄达看着他离开的背影，然后双手叉
(*Huang Da finishes speaking and turns around and leaves the hotel; Ye Xun looks at the back of Huang Da while he is leaving, crosses her arms, and silently picks up a hand on the waist; Huang Da looks at his back, then crosses his arms...*)
MMN: 叶薰说完话，转身离开，黄达看到之后立马起身离开，他穿着一身杂志坐在沙发上，从后座开始了酒吧，叶薰一把拉住了他。
(*Ye Xun finishes speaking and turns around to leave; Seeing this, Huang Da immediately gets up and leaves; he sits on the sofa wearing a magazine, and from the back seat starts the bar, Ye Xun pulls a hand on him.*)
GT: 叶薰说完之后转身离开，黄达尴尬地笑了笑，他看着叶薰离开的背影沉下了脸，随后离开了酒吧现场，叶薰回到桌前继续听外国人讲述着自己的事情。
(*After Ye Xun finishes speaking, she turns around to leave; Huang Da smiles awkwardly, becomes unhappy while looking at Ye Xun's back as she leaving, and then leaves the bar; Ye Xun returns to the table and continues to listen to the foreigner tell about himself.*)

(b)



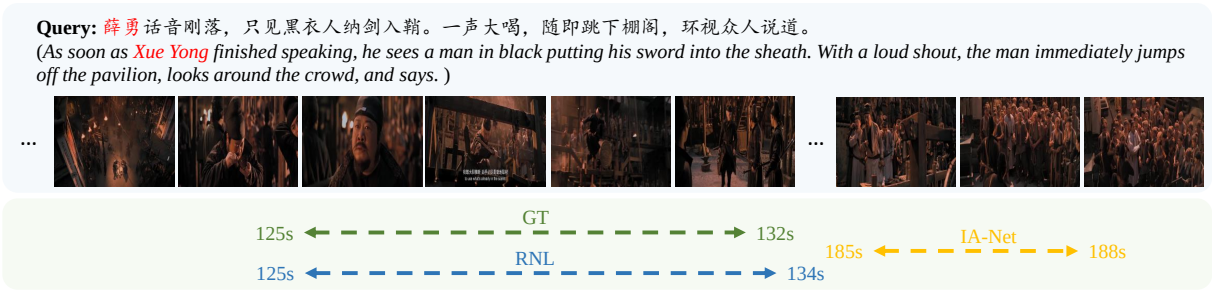
VT: 酒吧老板喝下酒，酒杯一饮而尽。
(*The bar owner drinks his wine and drains his glass in one go.*)
OVP: 酒吧老板端起酒杯喝了一口，酒杯满面，酒杯满足地喝了一口，
(*The bar owner picks up his glass and takes a sip; the glass is full...; the glass takes a sip satisfied...*)
RMN: 夏洛笑着点了点头，随后又拿起酒杯喝了一口，
(*Xia Luo smiles and nods, then picks up a glass and takes a sip.*)
GT: 袁华端起红酒杯邀请夏洛和他碰一杯，两人都一口饮完了杯中的酒。
(*Yuan Hua lifts his red wine glass and invites Xia Luo to clink glasses with him, and both of them finished their glasses in one gulp.*)

(c)

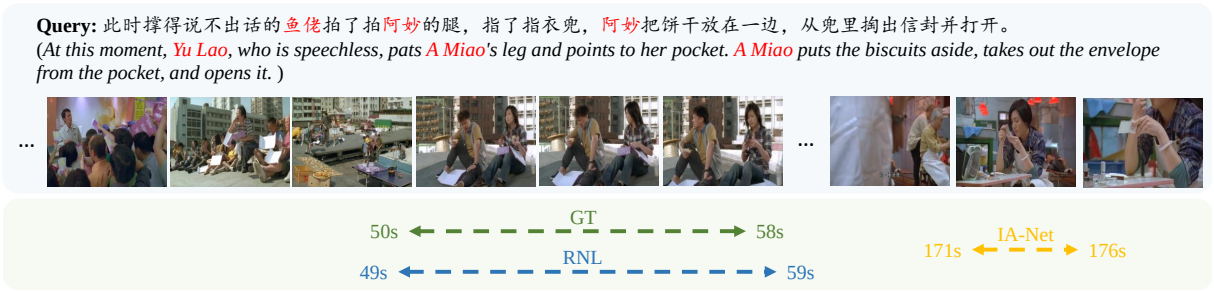
VT: 大家纷纷举起大家，大家看到大家，大家纷纷起身离开，
(*Everyone lifts up everyone, everyone sees everyone, everyone gets up and leaves...*)
OVP: 查理的话音刚落，宾客们的助理便被保安拦了下来，
(*As soon as Charlie's words leave his mouth, the guests' assistants are stopped by guards, ...*)
RMN: 王多鱼一脸严肃地站在宾馆门口，一时间掌声响彻云霄。
(*Wang Duoyu stands in front of the hotel with a serious face, and loud applause rings out.*)
GT: 所有人都站在会议室笑着鼓掌，王多鱼一头雾水。
(*Everyone stands in the conference room, laughing and applauding, and Wang Duoyu is bewildered.*)

(d)

Figure 8: Qualitative Results of MCN task on MovieUN-N. (VT: Vanilla Transformer; GT: the ground truth). Green denotes the correctly generated role names, while red denotes the wrong ones.



(a)



(b)

(c)

(d)

Figure 9: Qualitative Results of TNG task on MovieUN-G. The words in red represent the role names.