Modal-specific Pseudo Query Generation for Video Corpus Moment Retrieval

Minjoon Jung[1], Seongho Choi[1], Joochan Kim[1],
Jin-Hwa Kim[2,3]*, and Byoung-Tak Zhang[1,3]*
[1]Seoul National University   [2]NAVER AI Lab
[3]AI Institute of Seoul National University
{mjjung, shchoi, jckim}@bi.snu.ac.kr, j1nhwa.kim@navercorp.com, btzhang@bi.snu.ac.kr

Abstract

Video corpus moment retrieval (VCMR) is the task to retrieve the most relevant video moment from a large video corpus using a natural language query. For narrative videos, e.g., dramas or movies, the holistic understanding of temporal dynamics and multimodal reasoning is crucial. Previous works have shown promising results; however, they relied on the expensive query annotations for VCMR, i.e., the corresponding moment intervals. To overcome this problem, we propose a self-supervised learning framework: Modal-specific Pseudo Query Generation Network (MPGN). First, MPGN selects candidate temporal moments via subtitle-based moment sampling. Then, it generates pseudo queries exploiting both visual and textual information from the selected temporal moments. Through the multimodal information in the pseudo queries, we show that MPGN successfully learns to localize the video corpus moment without any explicit annotation. We validate the effectiveness of MPGN on the TVR dataset, showing competitive results compared with both supervised models and unsupervised setting models.

Figure 1: **Comparison with different supervision settings in VCMR.** (a) Fully-supervised VCMR (paired video query with timestamp), (b) Weakly-supervised VCMR (paired video query without timestamp), (c) VCMR without annotation (unlabeled video).

1 Introduction

The increased interest in video understanding has gathered attention for solving related tasks such as video captioning (Krishna et al., 2017), video question answering (Tapaswi et al., 2016; Lei et al., 2018; Kim et al., 2018), and video retrieval (Xu et al., 2016) over the past few years. Video corpus moment retrieval (VCMR) (Escorcia et al., 2019) is one of the challenging video understanding tasks, in which a model should *1) search for a related video* and *2) localize the corresponding moment* given a query sentence in a large video corpus.

Prior works have shown promising performances in VCMR using supervised (Lei et al., 2020; Zhang et al., 2020, 2021), weakly-supervised (Yoon et al.,
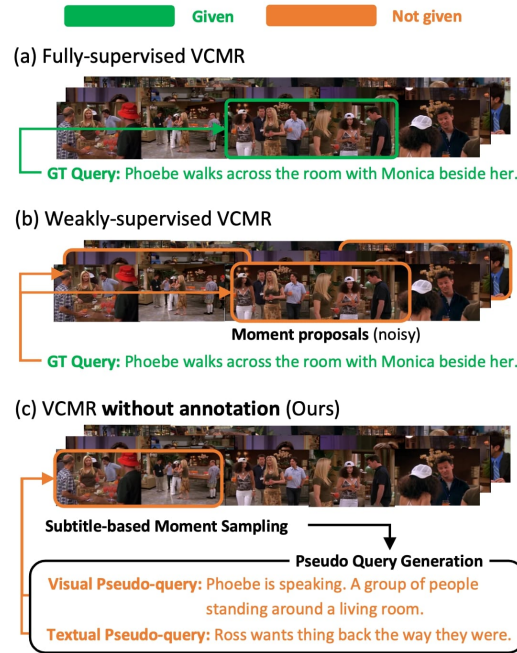
2021) settings, and pre-training (Li et al., 2020; Zhou et al., 2021) methods. Despite such accomplishments, selecting a temporal moment in a video (start time, end time) and generating the corresponding query sentence to train such models require overwhelming amounts of human labor. To annotate these videos, humans first need to understand the diverse information in the video, and then select the candidate temporal moment and generate corresponding queries.

The challenges of this VCMR are two folds: 1) Considering a large number of human annotations are required, an efficient approach is required to reduce the annotation cost. 2) These multimodal videos (e.g., drama or movies) generally contain rich interactions between characters, which widely

---

exist but has rarely been studied in the VCMR task. We introduce a novel framework to tackle these challenges: Modal-specific Pseudo Query Generation Network (MPGN). Our inspiration comes from the previous works (Nam et al., 2021; Jiang et al., 2022; Changpinyo et al., 2022) that generate pseudo queries to solve their target task in an unsupervised manner. To design our framework, we consider two research questions as follows: 1) What is a good way to select a temporal moment that can include characters' interactions? 2) What information should be considered when generating a pseudo query that sufficiently expresses the characters' interactions within the corresponding temporal moment?

First, we select a temporal moment based on where the topic of subtitles is divided considering the conversations between the characters in the target videos. Experimental results show that the proposed subtitle-based moment sampling method performs the best among competitive strategies.

For generating a pseudo query, our framework generates two modal-specific pseudo queries as follows: 1) Focusing on visual information, we extract the descriptive captions from a pre-trained image captioning model and the character names[1] from subtitles for the corresponding video frames. Then, we perform a visual-related query prompt module to generate queries that bridge the appearing character names and captions in videos. 2) Focusing on textual information, we exploit a pre-trained dialog summarization model to generate a textual query that cohesively captures the interactions among characters. Since raw subtitles are often noisy and informal, we summarize the corresponding subtitles in a temporal moment and use it as a pseudo query.

Our framework has several benefits as follows. First, our framework exploits multimodal video information to generate visual and textual pseudo queries, reducing the annotation cost for both queries and the corresponding moment intervals. Second, our framework generates high-quality modal-specific pseudo queries and shows significant performance gains.

Our contributions can be summarized as follows:

- To the extent of our knowledge, we firstly propose an unsupervised learning framework, MPGN, for the VCMR task.

- We propose the subtitle-based moment sampling method to define the temporal moment, and generate modal-specific pseudo queries exploiting both visual and textual information from the selected temporal moments.

- We experiment on the TVR benchmark to verify the effectiveness of our approach, and ablation studies validate each component of the proposed framework.

2 Related Work

2.1 Single Video Moment Retrieval

Single video moment retrieval (SVMR) aims to determine the temporal moments in a video that are related to given natural language queries. Previous works proposed remarkable progress based on fully-supervised learning (Gao et al., 2017; Mun et al., 2020; Zeng et al., 2020). However, since the annotations for SVMR are expensive, there have been attempts (Ma et al., 2020; Lin et al., 2020; Mithun et al., 2019) to address the annotation cost in a weakly-supervised manner. Unfortunately, substantial annotation costs still remain. Consequently, Liu et al. (2022) proposed DSCNet for SVMR performing without paired supervision.

Although these SVMR approaches are successful, they are unsuitable for the VCMR since they do not consider the huge computational cost involved in retrieving a video from the video corpus.

2.2 Video Corpus Moment Retrieval

Video corpus moment retrieval (VCMR) extends the number of video sources from a single (SVMR) to a collection of untrimmed videos (Escorcia et al., 2019). Previous methods have been proposed for VCMR in a supervised manner (Escorcia et al., 2019; Lei et al., 2020; Zhang et al., 2020, 2021). However, these approaches require fully-annotated data (*e.g.*, paired video query and the corresponding interval timestamps). To leverage this, Yoon et al. (2021) attempts to solve VCMR in a weakly-supervised setting where only paired videos and queries are available while the corresponding moment interval is unknown. While previous works require paired annotations for training, our framework does not require any annotation.

2.3 Pseudo Query Generation

Unsupervised image captioning methods (Laina et al., 2019; Feng et al., 2019) attempt to remove the

---

[1] We exploit the fact that the character names are annotated as speakers in the subtitles.

dependency on the paired image-sentence dataset. However, the proposed methods are not readily applicable to our video corpus. The most similar work to ours is PSVL (Nam et al., 2021), which has been proposed for the zero-shot SVMR task. They construct pseudo queries in a specific form consisting of a set of noun and verb words. However, narrative videos contain complex interactions between characters; it is inadequate to understand the videos with a limited set of noun and verb words. Unlike the previous methods, our framework can generate a pseudo query beyond these restrictions. In addition, we generate two pseudo queries that are specific to each modality.

3 Method

In this section, we introduce our framework Modal-specific Pseudo Query Generation Network (MPGN) in detail (see Figure 2). Given a video and its subtitles, we first describe how MPGN selects the candidate temporal moments. Then, we describe how MPGN generates modal-specific pseudo queries: using the visual-related prompt module for visual information and dialog summarization for textual information. We denote the generated pseudo query from each modality as visual pseudo query and textual pseudo query. Finally, we show how the generated pseudo queries are used in the training stage.

3.1 Subtitle-based Moment Sampling

MPGN samples a target temporal moment from a video to generate the corresponding modal-specific pseudo queries. Previous works have proposed to sample the temporal moments by comparing the visual similarity between adjacent frames (Nam et al., 2021; Jain et al., 2020) or sliding windows (Lin et al., 2020). However, such approaches are inappropriate for narrative videos since distinct and dissimilar visual frames can appear depending on the transitions of camera angles or speaking characters even in a single conversation. Motivated by how humans understand narrative videos, we propose a subtitle-based moment sampling method that determines the start and end timestamps from the sampled subtitles.

We denote the list of subtitles in a target video as $\mathcal{S}$. Let n be the number of subtitles, $\mathcal{S} = [s_1, s_2, ..., s_n]$. We sample the l-consecutive subtitles from S to select the temporal moment. We empirically found that if the length of the candidate

temporal moment is too short or too long, the generated pseudo queries are poor. Hence, we set the minimum number l_{min} and the maximum number l_{max}, and then, uniformly-sample l from $\{l_{min}, ..l_{max}\}$. After choosing the l, we uniformly-sample s_{start} from $\{s_1, ...s_{n-l}\}$, and s_{end} is straightforwardly determined by s_{start} and l. We summarize as follows:

$$l \sim U(l_{min}, l_{max}), \ \ l \in \mathcal{Z}$$
$$s_{start} \sim \{s_1, ...s_{n-l}\}.$$
$$s_{end} = s_{start+l}.$$

Finally, the sampled subtitles are defined as $\mathcal{S} = \{s_{start}, ..., s_{end}\}$.

3.2 Generating Modal-specific Pseudo Query

In general, the story of narrative videos can be represented through the visual (*e.g.*, *action, place*) and textual (*e.g.*, *dialog*) information related to the characters. Although they share the goal of comprehending a specific situation in a narrative video, the visual and textual-modality information can offer a different perspective. For example, if two characters are having a conversation in a video, visual features can represent that two characters are talking but cannot provide the details of the conversation. Meanwhile, textual features may provide specific details of the conversation, but not the person's location or actions. Therefore, we generate the pseudo queries for both modalities so the model can comprehensively understand the situation from diverse perspectives.

3.2.1 Visual Pseudo Query Generation

Inspired by the success of prompt engineering in vision-language tasks (Radford et al., 2021; Yao et al., 2021; Jiang et al., 2022), we adopt the visual-related prompt module to generate visual pseudo queries. To express visual information in the temporal moment, it depicts the situation of the scene by focusing on the person who appears in the temporal moment. The proposed visual-related prompt module combines this visual information to generate a visual pseudo query.

For every sampled temporal moment in a video, let a set of frames as $\mathcal{F}$ and the subtitles as $\mathcal{S}$. First, we detect the speaker name in the subtitles as shown in Fig 2-(c), and extract n unique character names $C = \{c_1, c_2, ..., c_n\}$ from $\mathcal{S}$ and generate a sentence with the character's name according to a specific template shown in the Table 1. We
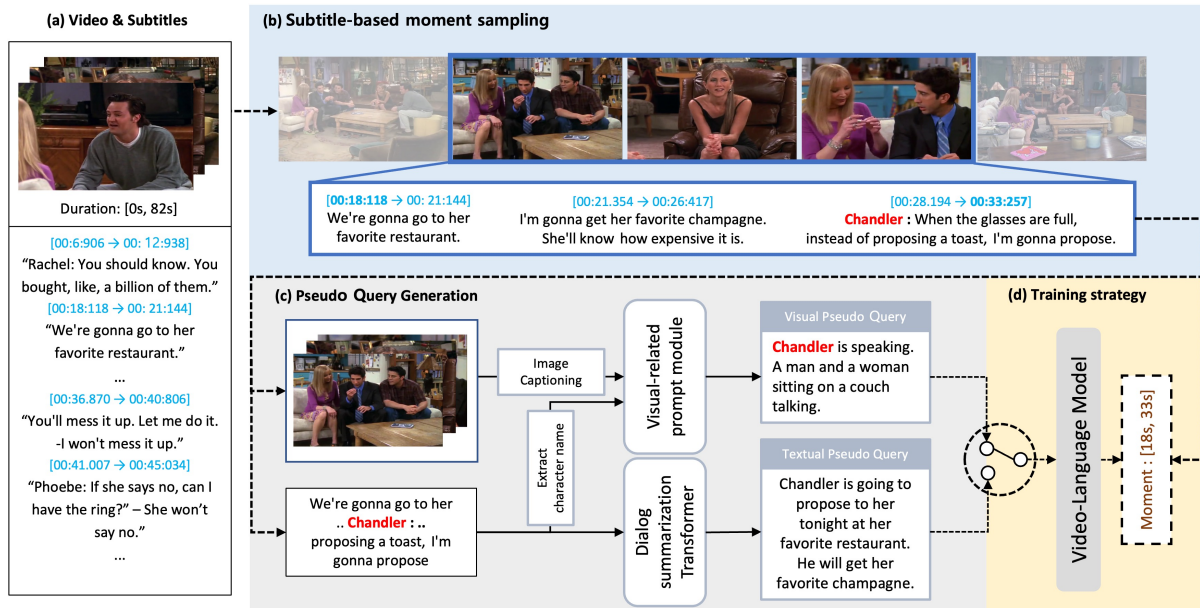
Figure 2: (a) Given a video and its aligned subtitles as input, our goal is to generate pseudo queries and train the model using them. Our framework consists of three stages: (b) define the temporal moments, (c) generate the modal-specific pseudo queries, and (d) use them for training our video-language model.

empirically found that in $n > 1$ case, the prompt "{*Character's names*} *are talking together*" shows better performance than the prompt "{*Character's names*} *having a conversation*". If we cannot identify any character name ($n = 0$) in the moment, we fill the character name with *Someone*. Then, we employ the pre-trained image captioning model (Li et al., 2022) to generate the image caption for the middle frame from $\mathcal{F}$. Finally, we concatenate these two sentences using the template for the characters' names and image caption to generate the visual pseudo query. (*e.g.*, "*Phoebe, Rachel, and Monica are talking together. A man is standing next to a woman in a living room.*")

3.2.2 Textual Pseudo Query Generation

We extract the semantic meaning from subtitles for the textual pseudo query. However, those are informal and noisy for the model to infer. Recently, Engin et al. (2021) has shown remarkable progress in video question answering by using dialog summarization. They convert the dialog into text description in several steps (per scene, whole episode) and use it to improve video-text representations in a supervised manner. Motivated by this, we denoise the subtitles by dialog summarization. To do this, we use the transformer-based BART$_{Large}$ (Lewis et al., 2019) pre-trained on the SAMSum corpus (Gliwa et al., 2019). Finally, we obtain the textual pseudo queries which capture the semantic meaning in dialog by applying a pre-trained language

Case	Template
n = 0	Someone is speaking.
n = 1	c_1 is speaking.
n > 1	c_1 and c_2 are talking together.
	c_1, c_2 and c_3 are talking together.
	etc.

Table 1: Templates for character name. c_n represent character name in a set of character's names $C = \{c_i\}_{i=1}^n$ and n represents number of characters.

model to subtitles $\mathcal{S}$.

3.3 Video-Language Model

Our video-language model consists of three components (1) video encoder, (2) query encoder, and (3) localization modules. Given a video V (with subtitle S) and query Q, we learn the representations of each input by encoders. We use two localization modules, each predicting start and end times from an input representation. The designs of the video encoder and the query encoder follow Lei et al. (2020), and the localization module consists of a 1D convolution filter. We provide more details for the video-language model in Appendix A.2.

Training Strategy To effectively utilize the modal-specific pseudo queries, we alternately train our model on pseudo queries. At each training step, we randomly (uniformly) select one of the modal-specific pseudo queries. Our training strategy can be cast as data augmentation, encouraging the

model to learn the multimodal information robustly. Note that we do not use any paired annotations (*e.g.*, *pairs of query sentences and temporal moments of a video*) in the training stage.

Inference We directly use the annotated query sentences during the inference stage without applying the visual-related prompt module for a fair comparison.

4 Experiments

4.1 Datasets and Metrics

TVR Lei et al. (2020) recently released TVR dataset, the large-scale video corpus moment retrieval dataset. TVR contains 21,973 videos from 6 TV shows, and each video is, on average, 76.2 seconds long and includes subtitles. There are five queries per video, containing an average of 13.4 words. The average length of moments in the video is 9.1 seconds. We follow the same split of the dataset as in TVR for fair comparisons. We re-emphasize that we do not use any annotations during the training stage.

Evaluation Metrics We follow the settings of previous methods (Lei et al., 2020). We evaluate the models for the VCMR task as well as its two subtasks, VR and SVMR. For SVMR and VCMR, we use Recall@k with IoU=0.7 for the main evaluation metric. For VR, we report Recall@k as the evaluation metrics.

4.2 Implementation Details

We extract 2048D RestNet-152 (He et al., 2016) and 2304 SlowFast (Feichtenhofer et al., 2019) features at 3 FPS and max pooling on frame features every 1.5 seconds. Each video feature is normalized by its L2-norm and concatenated for the final video feature. We extract the textual features via 12-layer pre-trained RoBERTa (Liu et al., 2019). Note that we fine-tune RoBERTa using only subtitles in TVR train-split with MLM objective, except the queries. As for subtitle-based moment sampling, we set the l_{min} and l_{max} to 2 and 5 respectively. We sample 130K temporal moments and each video has an average of 7 temporal moments. Each temporal moment has two pseudo queries, therefore 260K pseudo queries are generated. We train our model in an unsupervised setting with 87K pseudo queries of the same size as TVR train-split for 50 epochs with the batch size set to 128. For the supervised setting, we train our model with 260K pseudo queries and annotated queries in TVR train-

split over 70 epochs and set the same batch size as above.

All our experiments are run on a single Quadro RTX 8000. Our video-language model is optimized with AdamW and set the initial learning rate to 1.0×10^{-4}. The objective function of our model follows Zhang et al. (2021).

4.3 Quantitative Analysis

In Table 2, we compare with supervised methods, including XML (Lei et al., 2020), HAMMER (Zhang et al., 2020), ReLoCLNet (Zhang et al., 2021), HERO (Li et al., 2020), CUPID (Zhou et al., 2021), and weakly-supervised method WMRN (Yoon et al., 2021)[2]. In addition, we report the performance of the retrieval+re-ranking methods in various supervision-levels, which retrieve a set of videos by MEE (Miech et al., 2018) and then predict the temporal moment by the re-ranking method, CAL (Escorcia et al., 2019), MCN (Hendricks et al., 2017), TGA (Mithun et al., 2019) and VLANet (Ma et al., 2020). We report the performances of our framework, including (1) unsupervised settings and (2) supervised settings.

For the unsupervised and weakly-supervised methods, MPGN outperforms even when compared to the baselines in stronger supervision settings. Despite pre-training on large-scale video datasets, HERO[3] showed low performance overall. This result shows that HERO relies heavily on fine-tuning, and using subtitles instead of queries in the pre-training stage may be inappropriate. We show how the performance degrades when we use subtitles as a query instead of our pseudo queries in Section 4.4.4 As previous studies (Lei et al., 2020; Yoon et al., 2021) mentioned, retrieval+re-ranking methods show low performance since they consist of models targeting subtasks of VCMR. WMRN shows the best performance, but they generate multi-scale proposals from a large video corpus to predict temporal moments. This wasteful strategy cannot handle VCMR efficiently.

Surprisingly, MPGN outperforms the current state-of-the-art methods in supervised settings. Although the HERO and CUPID are pre-trained with a large amount of video-text pairs (136M), MPGN only uses 260K pseudo queries for training. Since

---

[2]We report results on test set for fair comparison, since no validation results are reported in WMRN (Yoon et al., 2021)

[3]We report the performance of HERO model which pre-trained on HowTo100M (Miech et al., 2020) and TV dataset (Lei et al., 2018) without fine-tuning on TVR dataset.

Sup	Method	Val			Test-public		
		R@1	R@10	R@100	R@1	R@10	R@100
Unsupervised	HERO	0.01	0.04	0.26	0.02	0.22	1.40
	MPGN (ours)	**1.24**	**4.46**	**12.01**	**1.49**	**5.93**	**15.87**
Weakly-supervised	MEE+TGA	-	-	-	0.24	1.57	5.68
	MEE+VLANet	-	-	-	0.69	3.84	10.22
	WMRN	-	-	-	**1.74**	**9.44**	**23.58**
Supervised	MEE+MCN	-	-	-	0.42	2.98	10.84
	MEE+CAL	-	-	-	0.39	2.98	11.52
	XML	2.62	9.05	22.47	3.25	12.49	29.51
	HAMMER	5.13	11.38	16.71	-	-	-
	ReLoCLNet	4.15	14.06	32.42	-	-	-
	HERO	5.13	16.26	24.55	6.21	19.34	36.66
	CUPID	5.55	17.61	25.73	7.09	20.35	40.53
	MPGN (ours)	**6.49**	**19.12**	**38.33**	**7.70**	**23.05**	**44.87**

Table 2: **Performance comparison with various models and supervision levels.** We conduct experiments on the TVR validation set and test-public set, and "-" means that the result on the metric is not reported in the original paper. "Sup" refers to supervision level: Supervised (paired video query and timestamp), Weakly-supervised (only paired video query), Unsupervised (without any annotation).

we focus on generating meaningful pseudo supervision for VCMR in this paper, we do not study pre-training tasks or model architecture that could improve performance.

4.4 Ablation Study

To investigate the importance of each component in MPGN, we conduct extensive ablation experiments. For a fair comparison, we use the same amount of pseudo queries as the original supervision for all experiments. We give detailed discussions in the subsections.

4.4.1 Effect of Modal-specific Pseudo Query

To validate the effectiveness of modal-specific pseudo query, we experiment with two baselines, including 1) a model trained on only visual pseudo queries (**VPQ**), 2) a model trained on only textual pseudo queries (**TPQ**). Our approach uses both (**VPQ + TPQ**) for training.

The results in Table 3 show that using both modal-specific queries improves model performance across all metrics. We conclude that providing both modalities information to the model helps it understand the video better.

4.4.2 Effect of Video-Language Model

We further compare our video-language model and with other baseline models, XML (Lei et al., 2020) and ReLoCLNet (Zhang et al., 2021). Also, we

report the performance of each model in both supervised and unsupervised settings in Table 4.

We confirm that the choice of video-language backbone contributes to the performance improvement and our MPGN framework is agnostic to the video-language backbones used for timestamp prediction.

4.4.3 Effect of Temporal Moment Sampling Method

We experiment our temporal moment sampling method with various l_{min} and l_{max} and compare it with the feature-based temporal moment sampling method proposed by Nam et al. (2021) in Table 5. We see that a using single subtitle (l=1) shows the lowest performance in all metrics. With this result, we safely say that a single subtitle cannot provide enough local temporal information for the model since it's grounded on a very short moment. Finally, we find the best performance when l_{min} and l_{max} were 2 and 5, respectively. We believe that using uniformly sampled temporal moments with appropriate l_{min} and l_{max} gives varying lengths compared to the fixed-size temporal moments.

Feature-based method computes all combinations of consecutive frame clusters and samples the temporal moment following a uniform distribution from them. Our approach is not only showing better performance than the feature-based method but also more efficient. To select the 100K tempo-

Method	VCMR			SVMR			VR		
	R@1	R@10	R@100	R@1	R@10	R@100	R@1	R@10	R@100
VPQ	0.73	2.57	7.61	4.86	18.24	46.74	10.21	32.65	71.34
TPQ	0.8	2.46	6.52	3.93	15.14	41.62	7.73	24.19	58.94
VPQ + TPQ	**1.24**	**4.46**	**12.01**	**5.27**	**20.44**	**48.73**	**13.93**	**38.63**	**75.36**

Table 3: **Ablation study on effect of modal-specific pseudo query.** (VCMR=Video Corpus Moment Retrieval, SVMR=Single Video Moment Retrieval, VR=Video Retrieval).

Method	Datasets		VCMR		
	P	T	R@1	R@10	R@100
XML	✓		0.8	2.84	8.17
		✓	2.62	9.95	22.47
	✓	✓	3.57	12.01	27.65
ReLoCLNet	✓		1.15	4.39	11.18
		✓	4.15	14.06	32.43
	✓	✓	5.74	18.54	38.38
Ours	✓		1.24	4.46	12.01
		✓	4.74	14.13	31.03
	✓	✓	6.49	19.12	38.33

Table 4: **Ablation study on effects of video-language models.** ✓ indicates the dataset used to train the model (P=Pseudo query, T=Training dataset in TVR datasets).

Method	l_{min}	l_{max}	VCMR		
			R@1	R@10	R@100
Feature-based	-	-	1.14	4.03	11.02
Subtitle-based	1	1	0.82	2.38	6.7
	3	3	1.16	3.54	9.78
	4	4	1.08	3.35	9.42
	2	3	1.05	3.85	9.8
	2	**5**	**1.24**	**4.46**	**12.01**
	2	7	1.02	3.52	9.46

Table 5: **Ablation study on effect of temporal moment sampling method.** If l_{min} and l_{max} are equal, we sample a fixed number of subtitles.

Method	VCMR		
	R@1	R@10	R@100
Dialog	0.15	0.47	1.68
Simplified sentence	0.3	1.24	4.62
Modal-specific	**1.24**	**4.46**	**12.01**

Table 6: **Ablation study on effect of pseudo query type.**

As shown in Table 6, our model can easily surpass other baseline methods in all metrics. We believe that a simplified sentence leads to poor results on our dataset because not only does the sentence have no textual information, but it also loses a lot of useful information since it consists of only nouns and verbs. The result of the dialog baseline achieves the lowest score across all metrics. As aforementioned in Section 4.3, it implies that it is difficult for the model to understand the videos with raw subtitles. With these experiments, we demonstrate that our generated pseudo queries represent more meaningful information in videos than other baseline methods.

4.5 Qualitative Analysis

In Figure 4, we provide two qualitative examples of moments predicted by our model and the ablation models. For the first case, our model successfully finds a temporal moment, but others do not. With this result, both of the modal-specific queries play an essential role in VCMR. Our model and VPQ localize the proper temporal moment for the second case, but TPQ fails. We hypothesize the reason is that the character name and visual information in the visual pseudo query help to find the temporal moment.

We visualize four generated pseudo queries on TVR dataset in Figure 3. In Figure 3-(a), visual pseudo query contains the speaker name and the caption of the scene, and textual pseudo query describes well what *Monica* is saying. Although most subtitles are very short and less meaningful in Figure 3-(d), our framework gets the dia-

ral moments, we take 13.48s, whereas the feature-based method consumes 328.05s.

4.4.4 Comparison with Other Types of Pseudo Query

To validate the competence of our pseudo query, we use simplified sentences and dialog, as the baselines method for our experiments. Nam et al. (2021) proposed a simplified sentence that consists of nouns and verbs for a pseudo query generation. For a fair comparison, we re-implemented their approaches as faithfully as possible, and the detailed procedure can be found in Appendix A.1. We also added a dialog baseline that uses subtitles as a pseudo query without applying a dialog summarization. Note that all the methods generate pseudo queries from the same temporal moments.
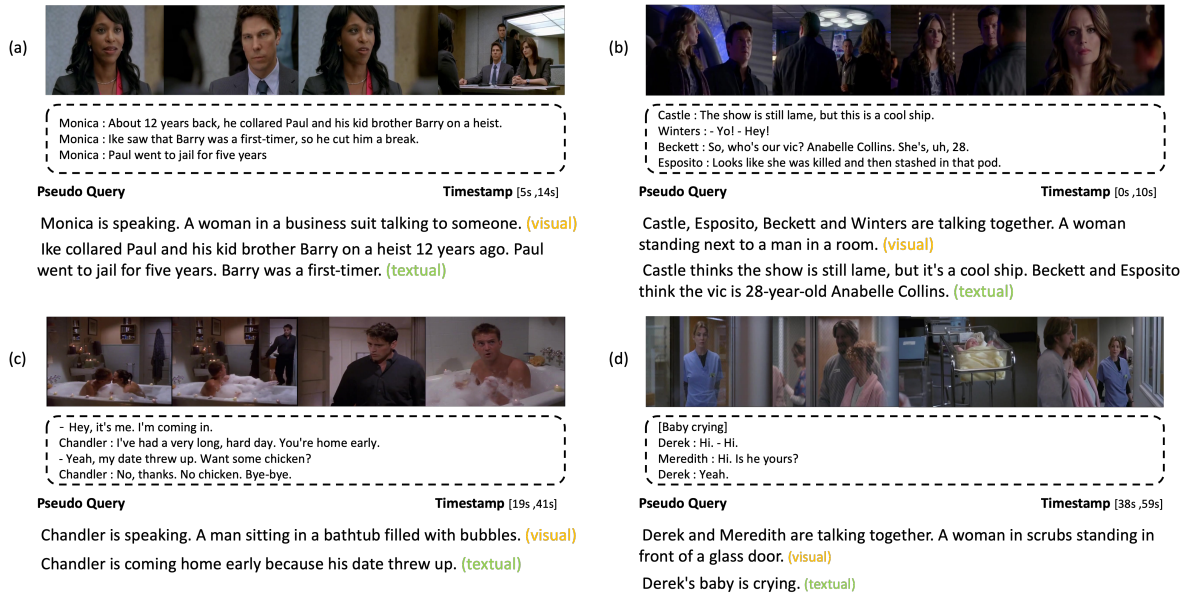
Figure 3: **Four visualization examples of pseudo queries on TVR dataset.** We show candidate temporal moment and modal-specific pseudo queries. All the pseudo queries except case (c) contain the character-centered context and describe the video moment well. (c) is a failure case due to a missing character name.
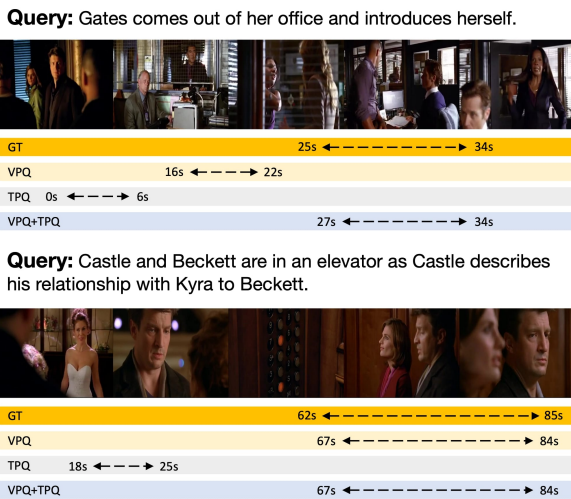

Figure 4: **Visualization of the predictions of the MPGN and the ablation models.** "GT" means ground-truth timestamp. The predictions of the models are presented below. The models trained on visual pseudo queries and text pseudo queries are called **VPQ** and **TPQ**, respectively. We denote **VPQ + TPQ** as the model trained on both pseudo queries.

log between *Meredith* and *Derek* and generates a grounded pseudo query. However, there is a case when the speaker's name is omitted from the subtitles, which is shown in Fig 3-(c). Our framework generates a textual pseudo query by substituting the situation for character *Joey*, which was missing from the subtitles, into the character *Chandler*. Although textual pseudo query describes the situation

well, incorrect character names may degrade the model performance. More examples are available in Appendix A.7.

To validate the scalability of our approach, we show the pseudo queries generated on the DramaQA (Choi et al., 2021) dataset in the Appendix A.3.

5 Conclusion

In this paper, we present a novel framework: Modal-specific Pseudo Query Generation Network (MPGN) for video corpus moment retrieval in an unsupervised manner. Our framework uses a subtitle-based temporal moment sampling method in which the timestamps (start time, end time) are determined from the sampled subtitles. After that, we generate pseudo queries from candidate temporal moments by using the visual-related prompt module and dialog summarization transformer, respectively. We can improve our model's comprehension of local temporal information and semantic meaning in multimodal videos via the pseudo queries containing the essential information in each modality. We conduct a comprehensive ablation analysis to prove the effectiveness of our approach. For future work, we plan to extend the pseudo query generation method so that it can be applied in several video understanding tasks without using manual supervision.

6 Limitation

Our framework requires the subtitles to include the name of the speaker. Therefore, it is not directly applicable to videos where the speaker is not specified (*e.g.*, YouTube videos). Also, as our framework utilizes verbal conversation between characters, it cannot guarantee performance in videos which do not include dialog (*e.g.*, videos of cooking, sports, etc.) We hypothesize that there exists some domain discrepancy between video benchmarks. We leave it as future work to extend our framework to diverse types of videos.

Acknowledgments

References

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for vqa are image captions. *arXiv preprint arXiv:2205.01883*.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. 2021. Dramaqa: Character-centered video story understanding with hierarchical qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1166–1174.

Deniz Engin, François Schnitzler, Ngoc QK Duong, and Yannis Avrithis. 2021. On the hidden treasure of dialog in video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2064–2073.

Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. 2019. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Mihir Jain, Amir Ghodrati, and Cees GM Snoek. 2020. Actionbytes: Learning from trimmed videos to localize actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1171–1180.

Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. 2022. Pseudo-q: Generating pseudo language queries for visual grounding. *arXiv preprint arXiv:2203.08481*.

Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. 2018. Multimodal dual attention memory for video story question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7414–7424.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11539–11546.

Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. 2022. Unsupervised temporal video grounding with deep semantic clustering. *arXiv preprint arXiv:2201.05307*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. 2020. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *European Conference on Computer Vision*, pages 156–171. Springer.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.

Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819.

Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. 2021. Zero-shot natural language video localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1470–1479.

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 *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*.

Sunjae Yoon, Dahyun Kim, Ji Woo Hong, Junyeong Kim, Kookhoi Kim, and Chang D Yoo. 2021. Weakly-supervised moment retrieval network for video corpus moment retrieval. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 534–538. IEEE.

Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296.

Bowen Zhang, Hexiang Hu, Joonseok Lee, Ming Zhao, Sheide Chammas, Vihan Jain, Eugene Ie, and Fei Sha. 2020. A hierarchical multi-modal encoder for moment localization in video corpus. *arXiv preprint arXiv:2011.09046*.

Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–695.

Luowei Zhou, Jingjing Liu, Yu Cheng, Zhe Gan, and Lei Zhang. 2021. Cupid: Adaptive curation of pre-training data for video-and-language representation learning. *arXiv preprint arXiv:2104.00285*.

A Appendix

We provide additional results not in the main paper due to the page limit.

A.1 Re-implementation of VerbBERT

Nam et al. (2021) proposed VerbBERT to predict the verbs from contextual nouns. We collect the dataset from the corpus that describes a person's action. Then we only select sentences that contains the word 'person' and extract only nouns and verbs from the sentences. In this step, about 10,000 sentences are remain.

For training, we randomly split the train and test dataset with a ratio of 9:1. We fine-tune the pre-trained RoBERTa model using the above sentences with MLM objective. After 20 epochs, there is no further improvement of perplexity, so we stop training as more epochs might cause over-fitting (see Figure 5). For a given sentence "person [mask] bicycle", VerbBERT will predict '[mask]' as 'ride'. To generate a pseudo query (simplified sentence) for TVR dataset, we predict the verb from detected objects and replace the 'person' with a character name. We visualize the generated simplified sentence in Figure 6.
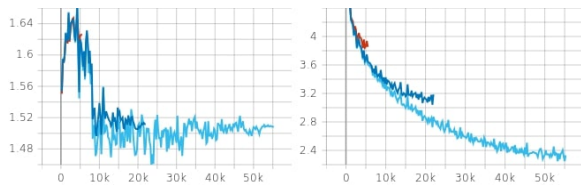

Figure 5: **Evaluation (Left) and Loss (Right) result of VerbBERT.** We use Perplexity for the evaluation metric. (epoch 1: Orange, epoch 5: Burgundy, epoch 20:Blue epoch 50: Cyan)

Figure 6: **Example sample of generated simplified sentence on TVR dataset.**

A.2 Details of Our Video-Language Model

In this section, we provide more details about the video-language model.

Model Architecture The video encoder consists of a feed-forward network and three transformer blocks for visual and subtitle representation. We apply the multimodal processing module (Gao et al., 2017) to each output of the last transformer block. The query encoder has a feed-forward network, two transformer blocks, and modularized vectors. Modularized vectors decompose the query into two query vectors each interacting with a visual and subtitle representation. Our localization module consists of two 1D-CNN layers with ReLU that predict the start and end probabilities, respectively. **Objective Function** The overall training loss follows Zhang et al. (2021) which consist of 1) video retrieval loss ($\mathcal{L}_{vr}$), 2) video moment retrieval loss ($\mathcal{L}_{vmr}$), 3) video contrastive loss ($\mathcal{L}_{vcl}$), 4) frame contrastive loss ($\mathcal{L}_{fcl}$) as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{vr} + \lambda_2 \mathcal{L}_{vmr} + \lambda_3 \mathcal{L}_{vcl} + \lambda_4 \mathcal{L}_{fcl}$$

A.3 Scalability of MPGN

Unfortunately, since TVR dataset is the only multimodal video dataset in VCMR, we cannot evaluate our framework to other benchmark To validate scalability of our framework, we generate pseudo queries on another multimodal video dataset, DramaQA which is built upon the TV drama and contains QA pairs for video question answering task. We visualize the pseudo queries generated on DramaQA dataset in Figure 7.

A.4 Statistics of Pseudo Query Dataset

In Figure 8, we show the distribution of temporal moment lengths (left) and the number of characters in temporal moments (right). The average length of the temporal moment is 12.3 seconds. We assume that including the case where the character name is omitted from the subtitle, most temporal moments include more than one character. Furthermore, visual pseudo queries and textual pseudo queries consist of an average of 14.9 words and 12.2 words, respectively.

A.5 Experiment on the Various Size of Pseudo Query Dataset

To investigate the quality of pseudo queries, we train the model on pseudo queries with different scales. We construct these subsets such that larger subsets include the smaller ones. We report AveR VCMR score (the average of R@1, 5,10, IoU=0.5) as the metric to evaluate model performance.

As shown in Figure 9, the performance increases in proportion to the scale of pseudo query dataset. However, we observed that there is no significant increase in performance after a certain point.
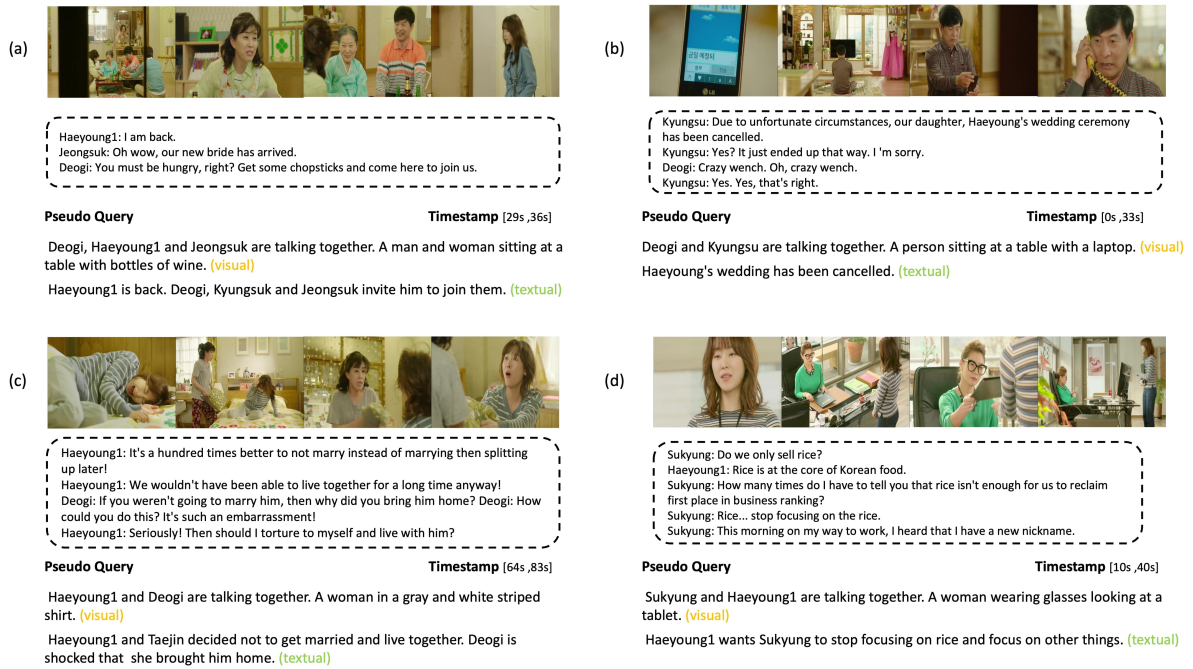
Figure 7: **Four visualization example of pseudo queries on DramaQA dataset.** All of the generated pseudo queries sufficiently describe the temporal moment of the video. In Figure7-(c), textual pseudo query well describe about the situation of *Haeyoung1* and *Taejin* weddings are canceled.
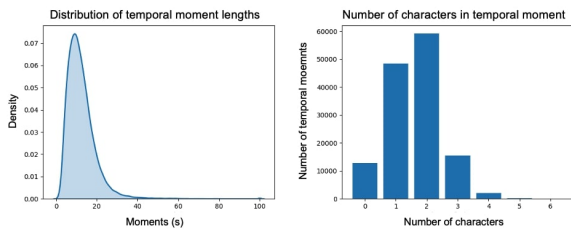
Figure 8: **Statistics of the generated pseudo queries.**

Figure 9: **AveR score on TVR according to various amount of pseudo queries.** 100% of pseudo query dataset contains 138K for training (87K queries in TVR dataset which is equivalent to 80% of the pseudo query dataset).

Method	VCMR		
	R@1	R@10	R@100
VPQ (w/o s)	0.62	2.4	6.83
VPQ	0.73	2.57	7.61

Table 7: **Ablation study on effect of speaker's name in the visual pseudo query.** "(w/o s)" means that trained with visual pseudo queries which do not contain the speaker's name.

A.6 Effect of Speaker's Name in Visual pseudo query

We report experiment with our model trained on only visual pseudo queries, which do not contain the speaker's name to investigate how the model performs.

As shown in the Table 7, the presence of the speaker's name was helpful but not essential.

A.7 More Visualization of Modal-specific Pseudo Queries

We visualize generated pseudo queries on TVR dataset in Figure 10. As we can see, most pseudo queries properly contain the multimodal information in the temporal moment of the video. However, in some cases, you can see the speaker is missing from the subtitles, such as Figure 10-(h). In these cases, it is difficult for a textual pseudo queries to completely provide textual information in subtitles.

(a)

> Chase : It could. Did you do anything to hurt it?
> Matty : Just threw the ball around the other day, for a couple hours.
> Foreman : It's probably just a muscle ache. Let us know if it gets worse.
> Matty : So soreness and stuff, that's not 'cause I'm sick?

Pseudo Query **Timestamp** [54s ,71s]

Matty, Foreman and Chase are talking together. a group of men standing next to each other (visual)

Matty has a muscle ache. He threw the ball around the other day. (textual)

(b)

> Montgomery : Beckett's the best I have.
> Montgomery : Now Castle can be a pain in the ass, but he thinks outside the box.
> Agent Mark Fallon : It's not your party anymore, Captain.
> Agent Mark Fallon : I'll have my agents escort you out.

Pseudo Query **Timestamp** [18s ,34s]

Someone is speaking. a close up of a person wearing a suit and tie (visual)

Agent Mark Fallon will escort Captain Montgomery out of the party. (textual)

(c)

> Ryan : Kind of Crocodile Dundee meets Indiana Jones meets Ted Nugent.
> Beckett : Now, according to the Meeks, he's violent and highly unstable. What's more, is that this guy's hell bent on bagging a Bigfoot.
> Esposito : I found him on YouTube.

Pseudo Query **Timestamp** [0s ,11s]

Esposito, Beckett and Ryan are talking together. a woman standing next to a window in a room. (visual)

According to Ryan and Beckett, this guy is violent and unstable and hell bent on bagging a Bigfoot. Esposito found him on YouTube. (textual)

(d)

> Rachel : Ross said my name. I can't just pretend it didn't happen. I don't know.
> Rachel : Monica, what should I do?
> Monica : Just do the right thing.

Pseudo Query **Timestamp** [37s ,48s]

Monica and Rachel are talking together. a woman in a red dress talking to another woman (visual)

Ross said Rachel's name. Monica advises Rachel to do the right thing. (textual)

(e)

> Castle : And then you slapped her hard enough to cut her cheek.
> Travis : Look, I didn't mean... Look, I just... I couldn't... I couldn't see straight.
> Travis : It's like I wasn't even myself anymore.
> Travis : I was so mad. I just... I didn't even know what I was doing.
> Beckett : - And so you killed her. - No! No. I didn't kill her

Pseudo Query **Timestamp** [56s ,72s]

Castle, Travis and Beckett are talking together. a man with blonde hair standing in front of a wall (visual)

Travis slapped her hard enough to cut her cheek. Beckett thinks Travis killed her (textual)

(f)

> Nadia : is the only thing that has ever made me special.
> Chase : Maybe so.
> Chase : But if you want to be special, you're going to have to be alone. Hey!

Pseudo Query **Timestamp** [21s ,56s]

Nadia and Chase are talking together. a woman sitting at a table with a cup of coffee. (visual)

Chase and Nadia discuss what makes Nadia special. (textual)

(b)

> The Director : Peel the onion. First of all, he's good-looking. I think my character needs a bit more of a reason than that.

Pseudo Query **Timestamp** [52s ,62s]

Someone is speaking. a man standing in front of a red chair. (visual)

The Director says that he's good-looking, but my character needs more of a reason than that. (textual)

(h)

> - Five american bucks. Damn it. Okay, so how do we get?
> Robin : The truth out of her.

Pseudo Query **Timestamp** [0s ,5s]

Robin is speaking. a man and a woman standing next to each other. (visual)
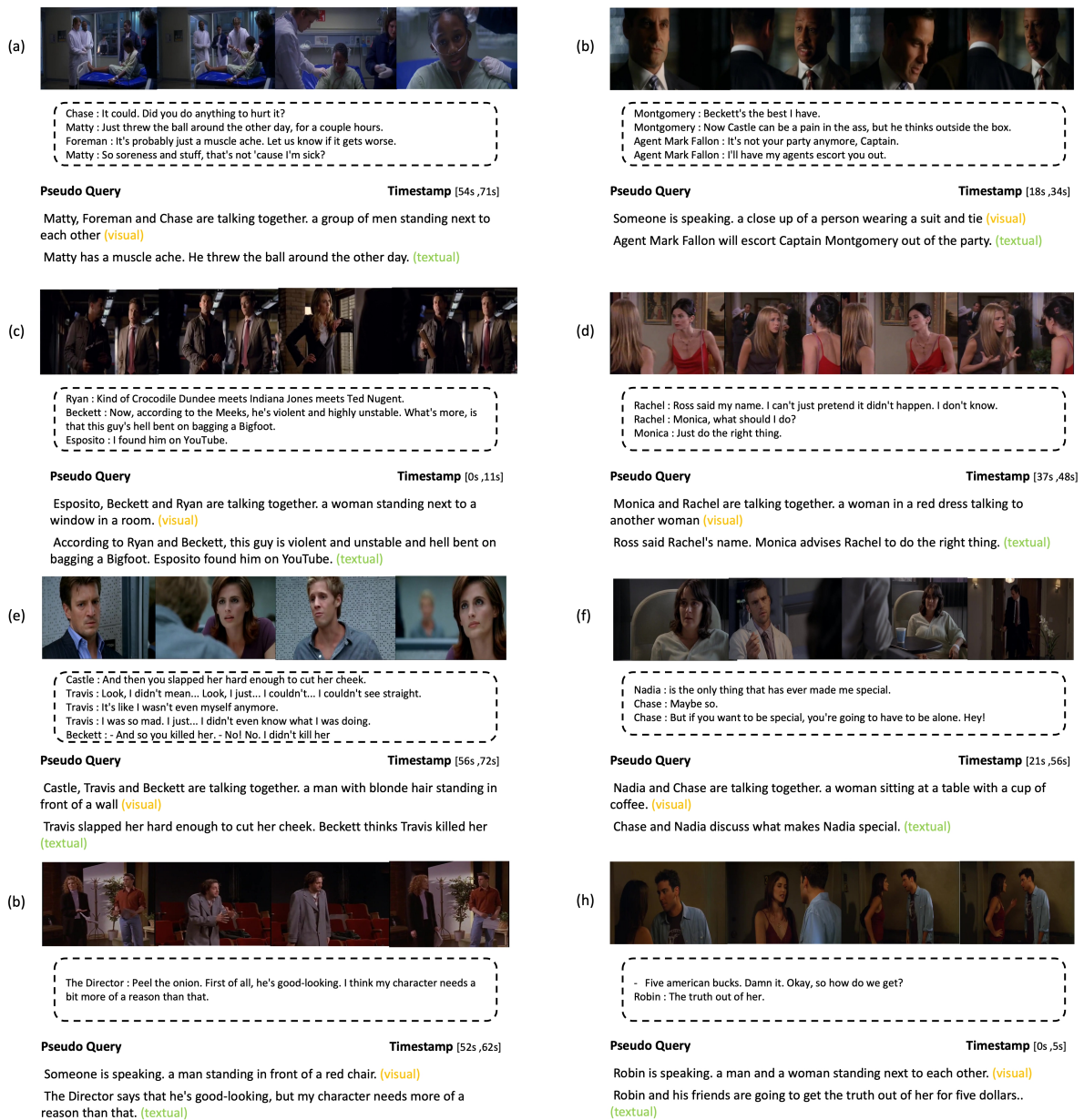
Robin and his friends are going to get the truth out of her for five dollars.. (textual)

Figure 10: **A visualization examples of pseudo queries on TVR dataset.** All of generated pseudo queries contain a character-centered context and well describe the multimodal information in the temporal moment.