Improving Cross-Modal Visual Answer Localization in Chinese Medical Instructional Video Using Language Prompts

Zineng Zhou[1,2,3], Jun Liu[1,2,3], Shuang Cheng[1,2,3], Haiyong Luo[1,2,3](✉),
Yang Gu[1,2,3], and Jian Ye[1,2,3]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{zhouzineng22s,liujun22s,chengshuang22s,guyang,jye}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Beijing Key Laboratory of Mobile Computing and Pervasive Device, Beijing, China
yhluo@ict.ac.cn

Abstract. The growing popularity of video content for acquiring knowledge highlights the need for efficient methods to extract relevant information from videos. Visual Answer Localization (VAL) is a solution to this challenge, as it identifies video clips that can provide answers to user questions. In this paper, we explore the VAL task using the Chinese Medical instructional video dataset as part of the CMIVQA track1 shared task. However, VAL encounters difficulties due to differences between visual and textual modalities. Existing VAL methods use separate video and text encoding streams, as well as cross encoders, to align and predict relevant video clips. To address this issue, we adopt prompt-based learning, a successful paradigm in Natural Language Processing (NLP). Prompt-based learning reformulates downstream tasks to simulate the masked language modeling task used in pre-training, using a textual prompt. In our work, we develop a prompt template for the VAL task and employ the prompt learning approach. Additionally, we integrate an asymmetric co-attention module to enhance the integration of video and text modalities and facilitate their mutual interaction. Through comprehensive experiments, we demonstrate the effectiveness of our proposed methods, achieving first place in the CMIVQA track1 leaderboard with a total score of 0.3891 in testB.

Keywords: VAL · Prompt · Cross-modal fusion · Data pre-processing

1 Introduction

The emergence of video content has led people to increasingly adopt video formats for acquiring knowledge. However, due to the typically lengthy nature of video clips, extracting knowledge from them can be a time-consuming and tedious process. Therefore, finding efficient methods to retrieve relevant information from videos is important.

Z. Zhou, J. Liu and S. Cheng—Equal contribution.

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14304, pp. 221–232, 2023.
https://doi.org/10.1007/978-3-031-44699-3_20

Visual Answer Localization (VAL) is an emerging task that corresponds to this issue. Its objective is to identify the video clips that can answer the user's question. The process of VAL involves analyzing the visual and subtitle components of a video to identify segments that contain relevant information. Recently, a new task temporal answer localization in the Chinese Medical instructional video is proposed. The datasets for this task have been collected from high-quality Chinese medical instructional channels on the YouTube website. These datasets have been manually annotated by medical experts. In this paper, we explored the VAL task in Chinese Medical dataset, which is the shared task in CMIVQA track1.

The VAL task presents challenges due to significant disparities between the visual and textual modalities [1]. Previous research has been conducted in related tasks like video segment retrieval [2] and video question answering [3]. However, it does not work well to directly transfer these methods due to difference in tasks [4]. Existing VAL methods typically employ a two-stream model to separately encode video and text, and utilize a cross encoder to align the modalities [4,5]. They then use cross-modal representations to predict the relevant video clips. The effectiveness of these methods relies on pre-trained language models, such as Deberta, but there is a noticeable discrepancy between the finetuning process of the VAL task and the pre-training of language models. The pre-training phase utilizes Masked Language Modeling, while the downstream VAL tasks involve token prediction.

We adopt language prompt to resolve this issue. Prompt-based learning is a novel paradigm in NLP that has achieved great success. In contrast to the conventional“pre-train, finetune” paradigm that involves adapting pre-trained models to downstream tasks through objective engineering, the “pre-train, prompt predict” paradigm reformulates the downstream tasks to simulate the masked language modeling task optimized during the original pre-training, utilizing a textual prompt. This paradigm aligns the downstream tasks more closely with the pre-training tasks, thereby enabling better retention of acquired knowledge. Notably, under low-resource conditions, it surpasses the “pre-train, finetune” paradigm, and has demonstrated promising results across various NLP tasks, including Question Answering and Text Classification.

In our work, we developed a prompt template for the VAL task and utilized the prompt learning approach. To enhance the integration of video and text modalities, we employ an asymmetric co-attention module to foster their mutual interaction. Our comprehensive experiments demonstrate the effectiveness of our proposed methods, which achieved the first place on the leaderboard on CMIVQA track1 with a total score of 0.3891 in testB.

2 Related Work

2.1 Visual Answer Localization

Visual answer localization is an important task in cross-modal understanding [4,5]. This task involves identifying the video clips that correspond to the user's query [6]. The current methods in(VAL) primarily employ sliding windows to generate multiple segments and rank them based on their similarity to the query. Alternatively, some methods use a scanning-and-ranking approach. They sample candidate segments through the sliding window mechanism and integrate the query with each segment representation using a matrix operation Some approaches directly predict the answer without the need for segment proposals [7]. In the latest work [5,8], the subtitle and query are inputted into a pretrained language model. Subsequently, a cross encoder is utilized to interact with the visual modality. In this paper, we utilize the prompt technique to improve the model's comprehension of the task, achieving this by employing a prompt to transform VAL into a MLM task.

2.2 Prompt Based Learning

Prompt-based learning is an emerging strategy that enables pre-trained language models to adapt to new tasks without additional training, or by training only a small subset of parameters. The manual prompt involves creating an intuitive template based on human understanding. The early use of prompts in pre-trained models can be traced back to GPT-1/2 [9,10]. These studies demonstrated that by designing suitable prompts, language models (LMs) could achieve satisfactory zero-shot performance in various tasks, including sentiment classification and reading comprehension. Subsequent works [11–13] further explored the use of prompts to extract factual or commonsense knowledge from language models (LMs). PET [14] is a semi-supervised training technique that rephrases the input in completion format using a prompt to enhance the model's comprehension of the task. It subsequently annotates the unsupervised corpus with multiple models tailored to single prompts, and ultimately trains the classifier on the enlarged corpus. And our approach is inspired by the ideas introduced in the PET technique.

3 Method

This section begins with the presentation of our data preprocessing approach, which aims to reduce noise in the model inputs. Subsequently, we will introduce our novel model architecture, emphasizing the significant components of prompt construction and cross-modal fusion. Ultimately, we will provide a detailed explanation of our loss design and training techniques.

3.1 Task Formalization

The Chinese Medical Instructional Video Question-Answering task aims to provide a comprehensive solution by addressing medical or health-related question (Q) in conjunction with Chinese medical instructional video (V) and their corresponding set of subtitles ($S = [T_i]_{i=1}^{r}$), where r denotes the number of subtitle spans. The primary objective is to accurately determine the start and end timepoints of the answer $[\hat{V}_s, \hat{V}_e]$ within the video V.

This task endeavors to develop advanced algorithms and systems capable of comprehending questions posed in Chinese medical instructional videos and effectively retrieving the corresponding answers. Moreover, it incorporates a subtitle timeline table (STB) that precisely maps each subtitle span to its corresponding timeline span in the subtitle set S. By functioning as a Look-up $Table$, this timeline table facilitates seamless mapping between frame span timepoints $[\hat{V}_s, \hat{V}_e]$ and accurate target answers$[V_s, V_e]$, thereby ensuring the provision of accurate target answers. Ultimately, the task can be mathematically represented as:

$$\left[\hat{V}_s, \hat{V}_e\right] = \mathrm{f}(Q, V, S) \tag{1}$$

$$[V_s, V_e] = \mathrm{STB}(\left[\hat{V}_s, \hat{V}_e\right]) \tag{2}$$

3.2 Data Preprocess

Before inputting the data into the model, a comprehensive data analysis was conducted. It was discovered that the subtitle information provided in the dataset was incomplete, as certain videos lacked subtitles. Upon closer examination, it was determined that the absence of subtitles was primarily attributed to the lack of audio subtitles (soft subtitles) in these videos.

However, it was observed that the video content itself contained subtitles, referred to as hard subtitles. In order to address this issue, Optical Character Recognition (OCR)[1] technology was employed to extract the subtitle text from the videos and supplement the missing captions. Additionally, for cases where subtitle information lacked both hard and soft subtitles, the corresponding subtitle was filled with a space. This method was implemented to enhance the integrity of the data.

3.3 Model Architecture

In this shared task, a novel method (MutualSL) [5] is employed as the baseline, demonstrating superior performance compared to other state-of-the-art (SOTA) approaches across various public VAL datasets. To further enhance the VAL capability for Chinese medical videos, an extended version of the baseline method is utilized by integrating prompts that activate powerful language comprehension and representation capabilities offered by large-scale language models for downstream tasks. Additionally, asymmetric co-attention [15] is incorporated to improve the model's cross-modal interaction capability.

The structure of the model is depicted in Fig. 1. During the initial stage, diverse feature extractors are employed to extract representations from both the input text and video frame sequences. Subsequently, the model combines

[1] https://github.com/YaoFANGUK/video-subtitle-extractor.

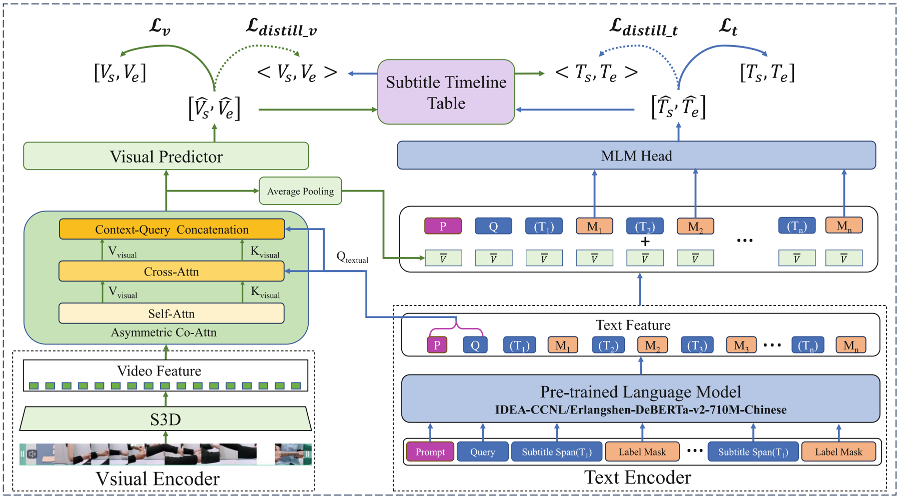

Fig. 1. The proposed cross-modal prompt model comprises separate feature extractors for video and text. Video features are enriched through Asymmetric Co-Att and the Visual Predictor, yielding $[\hat{V}_s, \hat{V}_e]$ predictions. Text and video features are combined using MLM Head and a broadcast mechanism, resulting in $[\hat{T}_s, \hat{T}_e]$ predictions. The final outcome considers four losses, with $< V_s, V_e >$ as the pseudo-label generated by text for video, and $< T_s, T_e >$ as the pseudo-label generated by video for text.

asymmetric co-attention with both video features and text-query features. To facilitate cross-modal interaction, the model employs a broadcast mechanism to combine the deep video features extracted by Asymmetric Co-Attn with the text features extracted by Deberta-V2-large [16] resulting in the final fused features being text-based. Finally, both the fused textual features and visual features are individually processed by their corresponding Predictors to obtain the final result.

Visual Feature Extraction. In contrast to the baseline model, we use separable 3D CNN (S3D) [17] pretrained on Kinetic 400 [18], which has better integration of spatial-temporal features and enhanced generalization capability, to extract video features instead of Two-Stream Inflated 3D ConvNets (I3D) [19]. Specifically, first, we extract video keyframes from video V using FFmpeg, and then S3D extracts video features from the video frames:

$$\mathbf{V} = \mathrm{S3D}(V) \tag{3}$$

Here, $\mathbf{V} \in \mathbb{R}^{k \times d}$, where d represents the dimension and k represents the length of the video.

Text Input Template. The main objective of text encoder is to provide us with high-quality question and subtitle information representation. we still follow the baseline approach by using the pre-trained Chinese language model Deberta-V2-large as the text encoder. However, simply putting in subtitle information does not fully activate the large model's understanding of the language task, so we employ prompt-based techniques to reconstruct the input text.

We introduce a reconstruction process by adding a prompt (P) before the problem and inserting the [Mask] (M) token in the middle of each subtitle segment to predict the result. Input template is defined as T

$$T = \{P, Q, T_1, M_1, T_2, M_2, ..., T_n, M_n\} \tag{4}$$

$$\mathbf{T} = \text{Deberta-V2-large}(T) \tag{5}$$

Here, $\mathbf{T} \in \mathbb{R}^{n \times d}$, where d represents the dimension and n represents the length of the text. We went through a lot of experiments and ended up with the best performing prompt templates. Finally Our prompt P is set as "请根据视频和字幕判断问题对应的答案在哪个位置".

Cross-Modal Fusion. To improve the semantic representation of video features and capture interactions between visual and textual information, we employ asymmetric co-attention. This mechanism consists of three components: a self-attention (SA) layer, a cross-attention (CA) layer, and the Context-Query Concatenation (CQA).

In the self-attention layer, the video features $\mathbf{V}$ extracted by S3D are utilized to capture internal dependencies within the visual information. This process yields enhanced visual features, denoted as $\mathbf{V}_{\text{visual}}^{SA}$, and attention keys, represented as $\mathbf{K}_{\text{visual}}^{SA}$.

$$\mathbf{V}_{\text{visual}}^{\text{SA}}, \mathbf{K}_{\text{visual}}^{\text{SA}} = \text{LN}(\text{SA}(\mathbf{V})) \tag{6}$$

Next, we incorporate textual features $\mathbf{Q}_{\text{textual}}$ obtained from the prompt and question into the visual features. The cross-attention layer plays a crucial role in integrating these textual features with the visual features $\mathbf{V}_{\text{visual}}^{SA}$ and $\mathbf{K}_{\text{visual}}^{SA}$. This integration facilitates the fusion of information from both modalities, enabling a comprehensive understanding of the video content. The output of the cross-attention layer is represented as $\mathbf{V}_{\text{visual}}^{CA}$ and $\mathbf{K}_{\text{visual}}^{CA}$, capturing the cross-modal interactions and enriching the semantic representation of the visual features.

$$\mathbf{V}_{\text{visual}}^{CA}, \mathbf{K}_{\text{visual}}^{CA} = \text{LN}(\text{CA}(\mathbf{V}_{\text{visual}}^{SA}, \mathbf{K}_{\text{visual}}^{SA})) \tag{7}$$

Finally, the outputs of the cross-attention layer, $\mathbf{V}_{\text{visual}}^{CA}$ and $\mathbf{K}_{\text{visual}}^{CA}$, along with the textual features $\mathbf{Q}_{\text{textual}}$, are concatenated and fed into the Context-Query Concatenation layer. This layer combines the contextual information from the video and the query, resulting in a text-aware video representation, $\mathbf{V}_{\text{visual}}^{CQA}$, that captures the interplay between visual and textual elements.

$$\mathbf{V}_{\text{visual}}^{CQA} = \text{Conv1d}(\text{Concat}[\mathbf{Q}_{\text{textual}}, \mathbf{V}_{\text{visual}}^{CA}, \mathbf{K}_{\text{visual}}^{CA}]) \tag{8}$$

Regarding the textual modality, we employ global averaging to pool the visual features $\mathbf{V}^{CQA}_{\text{visual}}$, resulting in the representation $\overline{V}^{CQA}_{\text{visual}}$. Finally, we combine $\overline{V}^{CQA}_{\text{visual}}$ with $\mathbf{T}_{\text{Deberta}}$ extracted by the Deberta-v2 model through summation to obtain the ultimate output of the textual features $\overline{\mathbf{T}}$.

$$\overline{V}^{CQA}_{\text{visual}} = \text{AvgPool}(\mathbf{V}^{CQA}_{\text{visual}}) \tag{9}$$

$$\overline{\mathbf{T}} = \{\overline{V}^{CQA}_{\text{visual}} + \mathbf{T}^{i}_{Deberta}\}^{n}_{i=1} \tag{10}$$

Visual Predictor. To address the current task, we adhere to the Visual Predictor approach established by the baseline, which includes separate start and end predictors. Each predictor is composed of a unidirectional LSTM model and a FNN. The $\mathbf{V}^{CQA}_{\text{visual}}$) features are inputted into the LSTM model, followed by the utilization of the feedforward layer to calculate the logarithm of the predicted time point logits $\{\hat{\mathbf{V}}_{\mathbf{s}}, \hat{\mathbf{V}}_{\mathbf{e}}\}$, encompassing both the start and end time points.

$$\hat{\mathbf{V}}_{\mathbf{s}} = \text{FNN}(\text{LSTM}_{\text{start}}(\mathbf{V}^{CQA}_{\text{visual}})) \tag{11}$$

$$\hat{\mathbf{V}}_{\mathbf{e}} = \text{FNN}(\text{LSTM}_{\text{end}}(\mathbf{V}^{CQA}_{\text{visual}})) \tag{12}$$

Prompt-Based Prediction. Figure 2 illustrates the “prompt,predict” paradigm. Our Input template is T with n “[mask]” tokens. We aim to predict the category words “始” and “末” using the textual prompt T. This process is similar to masked language modeling during the pre-training stage. Let $\mathbf{T}_s$ represent the probability of the “始” token and $\mathbf{T}_e$ represent the probability of the “末” token of all mask. Additionally, $[\hat{T}_s, \hat{T}_e]$ represent the probabilities of the ground truth being predicted as “始” and “末” , respectively.

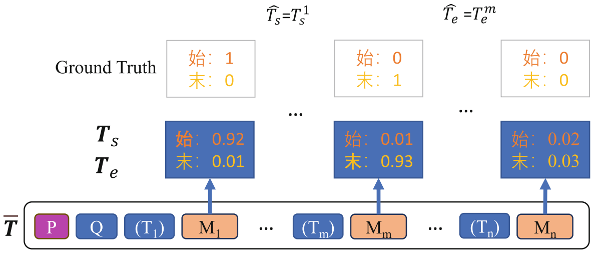

Fig. 2. Illustration of the “prompt, predict” paradigm.

Loss Function. In order to optimize the logits of the Visual Predictor and the Prompt-based Prediction, we utilize the Cross-Entropy function (CE). To enhance the model’s robustness, we employ a subtitle timeline Look-up Table, which generates pseudo-labels $\langle V_s, V_e \rangle$ for videos based on text prediction results and $\langle T_s, T_e \rangle$ for texts based on video prediction results. Additionally, we introduce the rdrop loss to further improve the model's robustness and enhance its generalization capabilities.

Finally, our loss function is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{v}} + \mathcal{L}_{\text{t}} + \mathcal{L}_{\text{distill\_v}} + \mathcal{L}_{\text{distill\_t}} + \beta \times \mathcal{L}_{\text{Rdrop}} \tag{13}$$

The loss terms are defined as follows: $\mathcal{L}_{\text{v}}$ represents the loss between the predicted video features and the true labels, $\mathcal{L}_{\text{distill\_v}}$ represents the loss between the predicted video features and the pseudo labels, $\mathcal{L}_{\text{t}}$ represents the loss between the predicted text features and the true labels, $\mathcal{L}_{\text{distill\_t}}$ represents the loss between the predicted text features and the pseudo labels, and $\mathcal{L}_{\text{Rdrop}}$ represents the loss of rdrop. Additionally, β represents the weight of the rdrop loss.

4 Experiments

4.1 Dataset and Metrics

NLPCC Shared Task 5 involves a dataset of 1628 Chinese Medical Instructional Videos with annotated question-answer pairs tied to video sections and divided into training (2936 examples) and two test sets (491 and 510 examples). This dataset, sourced from YouTube's Chinese medical channels and annotated by experts, includes videos, audios, and both types of Chinese subtitles. The data extraction process converts everything to Simplified Chinese.

Performance is evaluated using two metrics: Intersection over Union (IoU) and mean IoU (mIoU) [20], assessing video frame localization as a span prediction task. The examination includes "$R@n, IoU = \mu$" and "$mIoU$", with experiments using $n = 1$ and $\mu \in 0.3, 0.5, 0.7$ for evaluation.

4.2 Experiment Details

We executed a range of thorough experiments to verify the pipeline's efficiency. All tests maintained consistent training tactics and dataset arrangements for accurate comparisons. Particularly, the AdamW optimizer was used in our training regimen with an initial learning rate of 8e-6 and a 10% linear warmup. We divided the training and validation sets at a 0.9:0.1 ratio from the officially given annotated data, ensuring uniform dataset splits. Performance assessment of the top-performing model occurred on the validation set using the official testA sets. It is noteworthy that each epoch's training time was optimized to be only 30 min.

4.3 Experimental Results and Analysis

In this study, we have evaluated the impact of text feature extraction, visual feature extraction, data preprocessing schemes, asymmetric co-attention model setting and prompt setting on the Visual Answer Localization task. The experimental results summarized in Table 1 and Table 2 provide insights into the performance of various methods, which can be analyzed in the following sections.

Table 1. Impact of text and visual feature extraction, and data preprocessing schemes on Visual Answer Localization task performance. Visual Feature setting is based on DeBERTa-v2-710M-Chinese; Data Preprocess setting is based on DeBERTa-v2-710M-Chinese and S3D.

Method	Valid Set				TestA Set			
	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU
Baseline	60.52	43.13	26.58	43.41	56.71	40.65	23.58	40.28
Text Feature Setting								
Macbert-large	62.86	45.30	25.73	44.63	57.37	40.93	22.06	40.12
RoBERTa-large	58.74	42.07	27.90	41.57	55.74	40.07	19.12	38.31
DeBERTa-v2	**62.37**	**44.93**	**27.64**	**44.98**	**56.92**	**41.13**	**23.54**	**40.53**
Visual Feature Setting								
+ S3DG	61.64	43.25	26.81	43.90	56.63	40.55	23.81	40.33
+ Resnet151	59.14	42.49	25.99	42.54	52.10	40.49	23.12	38.57
+ **S3D**	**62.68**	**44.34**	**29.06**	**45.36**	**56.68**	**41.34**	**25.52**	**41.18**
Data Preprocess Setting								
+ Soft Caption	62.83	44.40	29.09	45.44	57.03	41.49	25.56	41.36
+ Hard Caption	63.70	44.88	29.30	45.96	57.86	42.03	25.69	41.86
+ **Both Caption**	**64.17**	**45.13**	**29.42**	**46.24**	**58.18**	**42.17**	**25.71**	**42.02**

Text Feature Setting. Among Chinese-Mac- bert-large [21], Chinese-RoBERTa-large [22], and DeBERTa-v2-710M-Chinese [16], we observe that the DeBERTa-v2-710M-Chinese model performs the best in IoU scores and mIoU for Valid Set and TestA Set, surpassing the baseline model. This proves its superior effectiveness in text feature extraction for the Visual Answer Localization task.

Visual Feature Setting. When evaluating different visual feature extraction schemes, we find that incorporating S3D into the DeBERTa-v2-710M model yields the highest mIoU on the Valid Set (45.36) and shows consistent improvement in the TestA Set (41.18), exceeding the baseline by 1.2%. This demonstrates S3D's suitability compared to S3DG and Resnet151, which scored lower than the baseline, showcasing their lower efficacy in visual feature extraction.

Data Preprocess. The blend of soft and hard caption extraction schemes outperforms the baseline model in mIoU scores for Valid Set (46.24) and TestA Set (42.02), substantiating the benefit of using audio and OCR-based techniques for caption extraction. A combination of both techniques results in the biggest improvement, hinting the advantage of using both audio and visual information to improve model performance in Visual Answer Localization tasks.

Model Evaluation. When examining the impact of different model settings, adding the Asy-Co-Att mechanism results in a significant improvement in performance across all IoU thresholds on both the validation and TestA sets, as compared to the baseline De-S3D-DP model. This indicates the mechanism effectively captures visual-textual interactions and refines video feature semantics.

While the addition of Rdrop also improves upon the base model, it doesn't provide the same significant gains as Asy-Co-Att. However, combining Asy-Co-Att and Rdrop attains the best performance, highlighting their complementary benefits.

Table 2. Impact of Model Settings and Prompt Configurations on De-S3D-DP for Visual Answer Localization. De-S3D-DP denotes the method which separately employs DeBERTa-v2-710M-Chinese and S3D models to extract textual and visual modality features, and optimizes the text through the use of both soft and hard subtitles. Asy-Co-Att refers to the asymmetric co-attention mechanism. Both (A&R) indicates that both the Asy-Co-Att and RDrop methods are employed simultaneously.

Method	Valid Set				TestA Set			
	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU
De-S3D-DP	64.17	45.13	29.42	46.24	58.18	42.17	25.71	42.02
Model Setting								
+ Asy-Co-Att	65.61	46.23	34.29	48.71	60.17	43.53	26.98	43.56
+ Rdrop	64.87	46.11	31.07	47.35	60.61	42.72	25.43	42.92
+ **Both(A&R)**	**67.28**	**48.01**	**35.14**	**50.14**	**60.08**	**44.57**	**27.43**	**44.03**
Text Prompt Setting Based on best method								
+ Prompt$_1$	65.10	52.64	37.15	51.63	60.25	45.49	27.19	44.31
+ **Prompt$_2$**	**66.14**	**54.03**	**39.20**	**53.12**	**60.82**	**46.94**	**27.71**	**45.16**

Text Prompt Setting. We analyzed two prompt configuration schemes: Prompt$_1$, which constructs text input without a [Mask] (M) token for predictions; and Prompt$_2$, which includes the [Mask] (M) token for downstream prediction. Prompt$_2$ performs better across all IoU thresholds and datasets. This consistency with the pretraining task seems to enhance the model's Visual Answer Localization abilities.

In summary, our analysis indicates that certain pre-trained language models (e.g., Macbert-large) and data preprocessing techniques (e.g., combining soft and hard captions) can significantly improve the performance of the Visual Answer Localization task. Besides, based on the De-S3D-DP model above, the results in Table 2 suggest that incorporating both the Asy-Co-Att mechanism and Rdrop method, along with the Prompt$_2$ configuration, leads to the most significant improvements in performance for the Visual Answer Localization task.

5 Conclusion

This research is dedicated to the challenge of gleaning pertinent information from videos through Visual Answer Localization (VAL). Our focus was the VAL task, utilizing the Chinese Medical instructional video dataset in the CMIVQA track1 shared task. The inability of existing methods to effectuate a smooth transfer from related tasks is recognized. To surmount these challenges, the Prompt-based learning paradigm from Natural Language Processing (NLP) was employed by us. This approach recalibrates downstream tasks to emulate the masked language modeling task employed during pre-training. A prompt template customized for the VAL task was developed and the prompt learning approach institutionalized. Furthermore, an asymmetric co-attention module was initiated to augment the integration of video and text data.

The efficiency of our methods was illustrated by our experiments, culminating in us achieving the topmost place on the CMIVQA track1 leaderboard, with an aggregate score of 0.3891 in testB. Prompt-based learning is proven to hold superiority over traditional pre-training and fine-tuning methods, especially under low-resource conditions. To conclude, our research propels VAL techniques forward and lays out functional solutions for valuing knowledge from videos.

Acknowledgement. This work was supported in part by the National Key Research and Development Program under Grant 2020YFB2104200 the National Natural Science Foundation of China under Grant 62261042 and 62002026, the Key Research Projects of the Joint Research Fund for Beijing Natural Science Foundation and the Fengtai Rail Transit Frontier Research Joint Fund under Grant L221003, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA28040500, and the Key Research and Development Project from Hebei Province under Grant 21310102D.

References

1. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Temporal sentence grounding in videos: a survey and future directions. arXiv preprint arXiv:2201.08071 (2022)
2. Tang, H., Zhu, J., Liu, M., Gao, Z., Cheng, Z.: Frame-wise cross-modal matching for video moment retrieval. IEEE Trans. Multimedia **24**, 1338–1349 (2021)
3. Lei, J., Yu, L., Bansal, M., Berg, T.L.: TVQA: localized, compositional video question answering. arXiv preprint arXiv:1809.01696 (2018)
4. Li, B., Weng, Y., Sun, B., Li, S.: Towards visual-prompt temporal answering grounding in medical instructional video. arXiv preprint arXiv:2203.06667 (2022)
5. Weng, Y., Li, B.: Visual answer localization with cross-modal mutual knowledge transfer. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
6. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5803–5812 (2017)
7. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. arXiv preprint arXiv:2004.13931 (2020)
8. Li, B., Weng, Y., Sun, B., Li, S.: Learning to locate visual answer in video corpus using question. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
9. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
10. Radford, A., Jeffrey, W., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
11. Petroni, F., et al.: Language models as knowledge bases? arXiv preprint arXiv:1909.01066 (2019)

12. Talmor, A., Elazar, Y., Goldberg, Y., Berant, J.: oLMpics-on what language model pre-training captures. Trans. Assoc. Comput. Linguist. **8**, 743–758 (2020)
13. Liu, J., Cheng, S., Zhou, Z., Gu, Y., Ye, J., Luo, H.: Enhancing multilingual document-grounded dialogue using cascaded prompt-based post-training models. In: Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, Toronto, Canada, pp. 44–51. Association for Computational Linguistics (2023)
14. Schick, T., Schütze, H.: Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676 (2020)
15. Li, C., et al.: mPLUG: effective and efficient vision-language learning by cross-modal skip-connections. arXiv preprint arXiv:2205.12005 (2022)
16. Zhang, J., et al.: Fengshenbang 1.0: being the foundation of Chinese cognitive intelligence. CoRR, abs/2209.02970 (2022)
17. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 305–321 (2018)
18. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
19. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
20. Gupta, D., Attal, K., Demner-Fushman, D.: A dataset for medical instructional video classification and question answering. Sci. Data **10**(1), 158 (2023)
21. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting pre-trained models for Chinese natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pp. 657–668. Association for Computational Linguistics (2020)
22. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for Chinese BERT. arXiv preprint arXiv:1906.08101 (2019)