Deep Video Understanding with Video-Language Model

Runze Liu
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
liurz@smail.nju.edu.cn

Yaqun Fang
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
fangyq@smail.nju.edu.cn

Fan Yu
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
yf@smail.nju.edu.cn

Ruiqi Tian
Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada
ruiqitian@outlook.com

Tongwei Ren*
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
rentw@nju.edu.cn

Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
gswu@nju.edu.cn

ABSTRACT

Pre-trained video-language models (VLMs) have shown superior performance in high-level video understanding tasks, analyzing multi-modal information, aligning with Deep Video Understanding Challenge (DVUC) requirements. In this paper, we explore pre-trained VLMs' potential in multimodal question answering for long-form videos. We propose a solution called Dual Branches Video Modeling (DBVM), which combines knowledge graph (KG) and VLMs, leveraging their strengths and addressing shortcomings. The KG branch recognizes and localizes entities, fuses multimodal features at different levels, and constructs KGs with entities as nodes and relationships as edges. The VLM branch applies a selection strategy to adapt input movies into acceptable length and a cross-matching strategy to post-process results providing accurate scene descriptions. Experiments conducted on the DVUC dataset validate the effectiveness of our DBVM.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Deep video understanding; Pre-trained video-language model; Knowledge graph; Cross-matching strategy

ACM Reference Format:
Runze Liu, Yaqun Fang, Fan Yu, Ruiqi Tian, Tongwei Ren*, and Gangshan Wu. 2023. Deep Video Understanding with Video-Language Model. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3581783.3612863

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612863

1 INTRODUCTION

Movies, as a special kind of video, usually contain more complex and dense semantic as a result of their long duration and the extensive care taken in the design and editing of all their shots. The Deep Video Understanding Challenge (DVUC) aims to perform deep analysis and understanding on movies, including human-centric interactions and relationships, and the descriptions and sentiments of movie clips. Specifically, DVUC 2023 requires to answer two types and five types questions on movie-level and scene-level, respectively. As compared to the DVUCs in the past years, DVUC 2023 still suffers the challenges in localizing entities and recognizing relationships between entities in long-form videos, which requires to fuse multiple modalities of information, such as visual, audio and speech, and reasoning upon them. Meanwhile, DVUC 2023 firstly describes questions and answer options in natural-language, which brings in new challenges in question and answer option understanding.

Recently, pre-trained video-language models (VLMs) attract much attention for significant performance in high-level video understanding tasks, such as video question answering (QA) [2, 12] and video-text matching [3, 4], which derives from their strong ability in cross-modality content alignment. Nevertheless, current studies of applying pre-trained VLMs in QA task have many constraints, *e.g.*, short videos (up to five minutes) and answer options (up to six words), and single modality (video frame). It is still uncertain whether pre-trained VLMs can achieve excellent performance in DVUC, which contains long videos (90 minutes on average) and answer options (more than ten words) and requires multi-modality in QA.

In this paper, we propose an integrated method named *Dual Branches Video Modeling* (DBVM) to handle the challenges in DVUC 2023. We divide the queries in DVUC 2023 into two categories: structural queries (Q1 of movie-level and Q1 and Q2 of scene-level) and non-structural queries (Q2 of movie-level and Q3, Q4, Q5 and Q6 of scene-level). The former has clear language description and can be answered well with the result of movie structurization, and the latter cannot be easily answered with the result of movie structurization and requires more understanding of the description of both questions and answer options. To structural questions, we

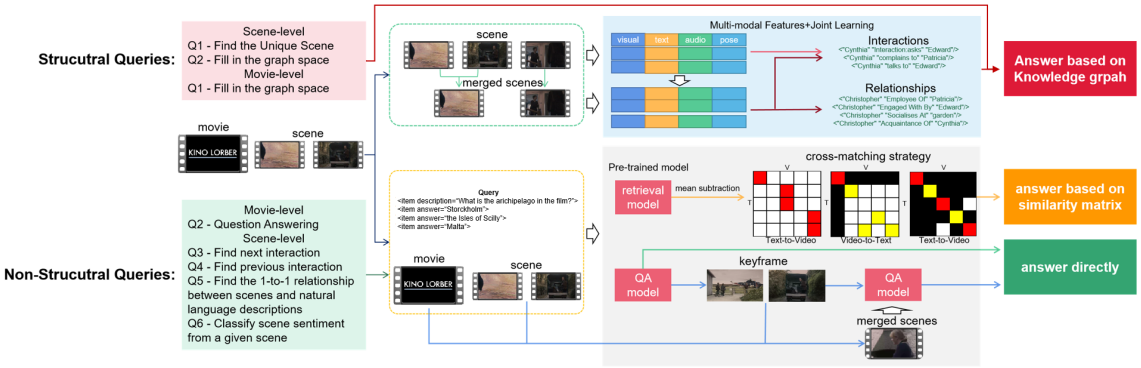

Figure 1: The framework of the proposed DBVM method.

follow the solution in our previous work [13], in which knowledge graphs (KGs) are pre-constructed on movies or clips, and questions are answered according to interactions and relationships used to form the KGs. To non-structural questions, we explore the potential of pre-trained VLMs by finetuning XClip [4] and Clip4Clip [3] for description-scene matching and adapting the input of SeViLA [12] for QA.

The contributions of our work are summarized as follows:

1) We explore the potential of pre-trained VLMs in long-form video analysis and validate their effectiveness.

2) We design an input selection strategy for SeViLA, a pre-trained VLM, to handle long-form video analysis in natural-language movie QA.

3) We design a cross-matching strategy to combine text-to-video (T2V) and video-to-text (V2T) results of pre-trained VLMs to improve description-scene matching accuracy.

2 PRELIMINARY

Text-To-Video Retrieval. Clip4clip model [3] firstly applies CLIP model to text-to-video retrieval task. XClip [4] adapts the pre-trained language-image models to video recognition directly, and brings in a cross-frame attention mechanism to exchange information across frames explicitly. LF-ViLA model [9] uses multimodal temporal contrastive loss and hierarchical temporal window attention mechanism to model long-range relationships and reduce computation cost.

Video Question Answering. Due to its ability of capturing temporal relationships, LF-ViLA model [9] after fine-tune is applied in video QA task. MIST [2] decomposes traditional dense spatialtemporal self-attention into cascaded segment and region selection modules. Visual concepts at different granularities are then processed efficiently through an attention module. To resolve the situation uniform frame sampling mismatches partly-relevance about video input with language query, SeViLA model [12] leverages a single image-language model to tackle both temporal keyframe localization and question answering on videos.

3 OUR METHOD

As shown in Figure 1, the queries in DVUC 2023 are divided into two categories: structural queries and non-structural queries. To structural queries, which are formulated as <subject, predicate, object> triplets, we generate KGs for movies and scenes based on our previous work in DVUC 2022 [13], and perform KG search to answer these queries. To non-structural queries, which involve only descriptions, we utilize multiple VLMs, namely Clip4clip [3] and XClip [4] for video-text matching and SeViLA [12] for QA.

3.1 Structural Query Answering

3.1.1 Multi-modal Feature Extraction. We merge scenes by using LGSS [7]. Extracting and combining various features (visual, text, audio, and pose), we use two branches to understand videos: predicting interactions based on single scene features and predicting relationships based on features of merged scenes. We use a video language transformer [11] to learn video-text interactions and extract features used to train the prediction model.

3.1.2 Knowledge Graph Construction. Features extracted from the above steps demonstrate the significance of relationships and interactions among entities, serving as essential elements for constructing KGs. By fusing and encoding these features, we compute similarity to target relationships, interactions, and sentiments, ultimately obtaining triplets comprising entities and predicates to construct KGs. Given that Q1 of movie-level and Q1 and Q2 of scene-level exhibit two distinctive characteristics, including their unique structural forms and close relevance to specific entities, we solve these queries by traversing KGs generated for each movie and counting the number of matches to the queries' descriptions.

3.2 Non-Structural Query Answering

We further divide the non-structural queries into three sub-categories: 1) Q2 of movie-level requires to find the best-matched answer option to a query, which can be solved with video QA model; 2) Q5 of scene-level requires to find the best-matched video to a scene description, which can be solved with video-text retrieval model; 3) Q3, Q4 and Q6 of scene-level also requires to find the best-matched answer option to a query but each answer options can be reconstructed to a scene description, which can be solved with either video QA model or video-text retrieval model.

Q2 of Movie-Level. Given the limitation of GPU memory, we use a coarse-to-fine strategy here. We first feed a query, all answer

Table 1: Ablation experiments on description-and-scene retrieval queries. We calculate *Accuracy* on T2V and V2T results with some combination of mean cross-strategy. *MS* and *CS* in subscript denote mean subtraction and cross-strategy, respectively. The best results in each row are highlighted in BOLD.

	XClip					*Clip4clip*				
	$V2T$	$V2T_{MS}$	$T2V$	$T2V_{CS}$	$T2V_{MS+CS}$	$V2T$	$V2T_{MS}$	$T2V$	$T2V_{CS}$	$T2V_{MS+CS}$
CallousedHands	0.50	0.40	0.50	0.60	0.60	0.70	0.60	0.60	0.60	**0.70**
ChainedforLife	0.60	0.70	0.50	**0.70**	**0.70**	0.70	0.60	**0.70**	**0.70**	**0.70**
LibertyKid	0.70	0.80	0.70	0.80	**1.00**	0.90	0.90	0.90	0.90	0.90
LikeMe	0.60	0.60	0.60	0.70	0.60	0.80	0.80	0.70	0.70	**0.80**
LosingGround	0.20	0.50	0.40	0.40	0.50	0.40	0.60	0.50	**0.70**	**0.70**
All	0.52	0.60	0.54	0.64	0.68	0.70	0.70	0.68	0.72	**0.76**

options, and a whole movie to SeViLA [12], which is set to output the best-matched video keyframe and answer to the query. Then, we replace the whole movie with the movie segment with half previous scene and half successive scene around the output keyframe, in which the scenes are provided by DVUC, and feed them to SeViLA again. The output of SeViLa is treated as the final result of Q2 of movie-level.

Q5 of Scene-Level. We used Clip4clip [3] with the training dataset to generate a similarity matrix between 10 descriptions and 10 scenes. After performing mean subtraction on each T2V vector in the similarity matrix of the retrieval model, we identified the T2V results where the descriptions did not share a mapping scenario with other descriptions. Subsequently, we removed the corresponding rows and columns containing these results. We then followed the same process to single out the V2T results and treated the similarity matrix in a similar manner. We iterated through the above steps until the similarity matrix could no longer be reduced in size, and then selected the latest T2V result as the final outcome. Our cross-matching strategy is based on the obvious fact that if two or more descriptions select the same scene, their scores would be at most 0.5 since they cannot both be correct. Moreover, we were motivated to design this cross-matching strategy by the weakly dominant position of V2T result and the outstanding performance of retrieval models on description-and-scene retrieval queries as shown in Table 1.

Q3, Q4 and Q6 of Scene-Level. While next-and-previous interaction prediction and sentiment retrieval queries are not typical video QA queries, they share similar characteristics in form with video QA. Both QA models and video-text retrieval models possess the capability to comprehend temporal and semantic information [9]. Hence, we utilize SeViLA to address next-to-previous interaction prediction and sentiment retrieval queries. The motivation behind employing a video QA model for these queries is their close resemblance to natural language forms. This allows us to enhance the queries with prompts, such as "immediately" and "in a short time", in their descriptions to improve performance and understanding.

Compared to last year's approach, the new method incorporates pre-trained VLMs to tackle non-structural queries, with a primary emphasis on capturing the semantics of movies and scenes. By leveraging the enhanced capabilities of these introduced pre-trained VLMs, the new method demonstrates more effective understanding and processing of video-related information.

4 EXPERIMENTS

4.1 Dataset and Experimental Settings

We validate our method on the HLVU dataset [1] and the Kinolorber dataset, totaling 24 movies. Among these, 14 movies are used for training, five for validation, and five for main task. In DVUC 2023, the evaluation metrics of structural query answering and non-structural query answering are *Mean Reciprocal Rank* (MRR) and *Accuracy* (Acc), respectively. Specifically, *MRR* is calculated as $MRR = \frac{1}{N_T}\sum_{i=1}^{N_T}\frac{1}{RANK_i}$, here N_T is the number of total queries and $RANK_i$ is the rank of the correct answer in the answer list of the ith query; *ACC* is calculated as $Acc = \frac{N_C}{N_T}$, here N_C is the number of correct answers.

All experiments are conducted on Xeon 2.40GHz CPU, 64GB memory and one GeForce RTX 3090 GPU.

4.2 Ablation Study

Due to page limitation, we only present the experiment results on non-structural query answering, as our improvements through incorporating pre-trained VLMs has been applied to this aspect.

4.2.1 Description-Scene Retrieval. We divide 10 description-scene pairs into groups and calculated the *Accuracy* score for each group, as well as the average *Accuracy* score across all groups. All the proposed methods, including mean subtraction, and cross-matching strategy, were utilized.

To validate effectiveness of each step, we conduct experiments with different step settings, as shown in Table 1. In the V2T direction, we observe an increase in the average score. In the T2V direction, we also achieve an average score increase at each step of our method, with significant improvements in scores across all groups. It confirms the reliability of the mean subtraction and cross-matching strategy we designed.

4.2.2 Comparative experiments on QA and Retrieval models with different settings. **QA model with different parameters.** We configured varying numbers of keyframes to be used in the localizer module of SeViLA based on different features of sentiment retrieval queries and next-to-previous interaction prediction queries. As depicted in Table 2, we calculate *Accuracy* of sentiment retrieval

Table 2: Comparative experiments on sentiment queries. KF in subscript denotes number of keyframes. The best results of two types scene-level queries (Next-and-Previous Interaction Prediction (Q3 and Q4) and Sentiment Retrieval (Q6)) in each row are highlighted in BOLD, respectively.

	Next-and-Previous Interaction Prediction (Q3 and Q4)					Sentiment Retrieval (Q6)				
	$XClip_{V2T}$	$Clip4clip_{V2T}$	$QA_{KF=1}$	$QA_{KF=2}$	$QA_{KF=4}$	$XClip_{V2T}$	$Clip4clip_{V2T}$	$QA_{KF=1}$	$QA_{KF=2}$	$QA_{KF=4}$
CallousedHands	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	0.50	0.50	**0.67**	**0.67**	**0.67**
ChainedforLife	0.25	0.25	0.25	0.25	**0.38**	**0.67**	**0.67**	0.33	0.33	0.17
LibertyKid	0.13	0.00	0.25	0.25	**0.38**	**0.50**	**0.50**	**0.50**	0.33	0.33
LikeMe	0.13	**0.38**	0.25	0.25	0.25	**0.50**	**0.50**	0.17	0.17	0.17
LosingGround	0.13	0.13	**0.75**	**0.75**	**0.75**	**0.50**	0.17	**0.50**	**0.50**	0.33
All	0.18	0.20	0.35	0.35	**0.40**	**0.53**	0.47	0.43	0.40	0.33

query results of QA model with different parameters to evaluate its performance. We observed a decrease in performance for sentiment retrieval queries as the number of keyframes increased. This decline could be attributed to each keyframe containing an independent sentiment, and combining keyframes with different sentiments leading to a confused result. Conversely, as shown in Table 2, we noticed an increase in performance as the number of keyframes increased for next-to-previous interaction prediction queries. This observation suggests that multiple keyframes are more applicable for such queries as they require keyframes from different timestamps.

Comparison of QA model with Retrieval model. Both QA and Retrieval models possess the capability to capture temporal and semantic information, which is why we apply them to both next-and-previous interaction prediction queries and sentiment retrieval queries. As indicated in Table 2, the retrieval model exhibits better performance on sentiment queries, while the QA model performs better on next-to-previous interaction prediction queries. We attribute this observation to the retrieval model considering all frames of the scene, making it unable to determine the duration of the interaction pair accurately. The QA model shows larger variance as it is unclear how many frames are needed to assess sentiment accurately.

4.3 Comparison with DVUC 2022

Similar to ablation study, we only present the performance comparison with the methods in DVUC 2022 on non-structural query answering. As shown in Table 3, our method has resulted in significant improvements across all these problems compared to our work in DVUC 2022 [13], that is attributed to the capability of pre-trained VLMs in capturing semantic relations. However, our method performs worse than E-VG [8] on Q3 and Q4 of scene-level, because E-VG extracts timestamp information which is useful in answering next-and-previous interaction prediction queries. Our performance is also worse than DVU-SQL [10] on Q5 and Q6 of scene-level, because DVU-SQL is trained on an external and large movie dataset to handle movie related task better.

4.4 Discussion

Noise subtask. To address the subtasks with three types of noise, we denoise on scenes with video noise by clip timestamps with frame loss directly.

Table 3: Performance comparison with DVUC 2022. The best and second best results in each column are highlighted in BOLD and UNDERLINE, respectively.

	Scene-Q3	Scene-Q4	Scene-Q5	Scene-Q6
E-VG [8]	**0.63**	**0.69**	-	-
DVU-SQL [10]	-	-	**1.00**	**0.61**
HERO TVQA [5]	0.21	0.26	0.65	0.19
Graphen [6]	0.25	0.31	0.15	0.14
Nanjing U. [13]	0.29	0.26	0.22	0.14
Ours	<u>0.40</u>	<u>0.40</u>	<u>0.76</u>	<u>0.53</u>

Efficiency. In our previous work [13], feature extraction occupies over 20GB GPU memory and takes about five days to process a single movie. Its extremely low efficiency leads to the failure in answering all queries in structural query answering of main task and subtasks. However, thanks to the pre-trained VLMs, we finish all non-structural query answering in only one day, significantly improving the efficiency and feasibility of the method.

5 CONCLUSION

We proposed a novel DBVM method that utilizes pre-trained VLMs to capture temporal and semantic information from movies and scenes. Specifically, we analyzed the impact of different parameter and prompt settings in applying pre-trained VLMs. Moreover, we designed an input selection strategy for SeViLa and a cross-matching strategy for XClip and Clip4clip to handle long-form video analysis and improve description-scene matching accuracy, respectively. Extensive experiments validated the effectiveness of our method.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation of China (62072232), the Fundamental Research Funds for the Central Universities (021714380026) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. HLVU: A New Challenge to Test Deep Understanding of Movies the Way Humans do. In *International Conference on Multimedia Retrieval*. 355–361.
[2] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2023. MIST: Multi-modal Iterative Spatial-Temporal Transformer for Long-form

Video Question Answering. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14773–14783.

[3] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.

[4] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*. Springer, 1–18.

[5] Penggang Qin, Jiarui Yu, Yan Gao, Derong Xu, Yunkai Chen, Shiwei Wu, Tong Xu, Enhong Chen, and Yanbin Hao. 2022. Unified QA-aware knowledge graph generation based on multi-modal modeling. In *The 30th ACM International Conference on Multimedia*. 7185–7189.

[6] Raksha Ramesh, Vishal Anand, Zifan Chen, Yifei Dong, Yun Chen, and Ching-Yung Lin. 2022. Leveraging Text Representation and Face-head Tracking for Long-form Multimodal Semantic Relation Understanding. In *The 30th ACM International Conference on Multimedia*. 7215–7219.

[7] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A local-to-global approach to multi-modal movie scene segmentation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10146–10155.

[8] Siyang Sun, Xiong Xiong, and Yun Zheng. 2022. Two stage Multi-Modal Modeling for Video Interaction Analysis in Deep Video Understanding Challenge. In *The 30th ACM International Conference on Multimedia*. 7040–7044.

[9] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. 2022. Long-form video-language pre-training with multimodal temporal contrastive learning. *Advances in neural information processing systems* 35 (2022), 38032–38045.

[10] Chen-Wei Xie, Siyang Sun, Liming Zhao, Jianmin Wu, Dangwei Li, and Yun Zheng. 2022. Deep Video Understanding with a Unified Multi-Modal Retrieval Framework. In *The 30th ACM International Conference on Multimedia*. 7055–7059.

[11] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996* (2021).

[12] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-Chained Image-Language Model for Video Localization and Question Answering. *arXiv preprint arXiv:2305.06988* (2023).

[13] Beibei Zhang, Yaqun Fang, Tongwei Ren, and Gangshan Wu. 2022. Multimodal Analysis for Deep Video Understanding with Video Language Transformer. In *The 30th ACM International Conference on Multimedia*. 7165–7169.

[14] Beibei Zhang, Fan Yu, Yaqun Fang, Tongwei Ren, and Gangshan Wu. 2021. Hybrid improvements in multimodal analysis for deep video understanding. In *ACM Multimedia Asia*. 1–5.