DPHANet: Discriminative Parallel and Hierarchical Attention Network for Natural Language Video Localization

Ruihan Chen†, Junpeng Tan†, Zhijing Yang*, Xiaojun Yang, Qingyun Dai, Yongqiang Cheng, and Liang Lin

***Abstract*—Natural Language Video Localization (NLVL) has recently attracted much attention because of its practical significance. However, the existing methods still face the following challenges: 1) When the models learn intra-modal semantic association, the temporal causal interaction information and contextual semantic discriminative information are ignored, resulting in the lack of intra-modal semantic context connection; 2) When learning fusion representations, existing cross-modal interaction modules lack hierarchical attention function to extract inter-modal similarity information and intra-modal self-correlation information, resulting in insufficient cross-modal information interaction; 3) When the loss function is optimized, the existing models ignore the correlation of causal inference between the start and end boundaries, resulting in inaccurate start and end boundary calibrations. To conquer the above challenges, we proposed a novel NLVL model, called Discriminative Parallel and Hierarchical Attention Network (DPHANet). Specifically, we emphasized the importance of temporal causal interaction information and contextual semantic discriminative information and correspondingly proposed a Discriminative Parallel Attention Encoder (DPAE) module to infer and encode the above critical information. Besides, to overcome the shortcomings of the existing cross-modal interaction modules, we designed a Video-Query Hierarchical Attention (VQHA) module, which can perform cross-modal interaction and intra-modal self-correlation modeling in a hierarchical manner. Furthermore, a novel deviation loss function was proposed to capture the correlation of causal inference between the start and end boundaries and force the model to focus on the continuity and temporal causality in the video. Finally, extensive experiments on three benchmark datasets demonstrated the superiority of our proposed DPHANet model, which has achieved about 1.5% and 3.5% average performance improvement and about 2.5% and 7.5% maximum performance improvement on the Charades-STA and TACoS datasets respectively.**

***Index Terms*—Cross-modal retrieval, natural language video localization, video moment localization, video understanding.**

R. Chen, J. Tan, and X. Yang are with the School of Information Engineering, Guangdong University of Technology, Guangzhou, 510006, China. (e-mail: 2112103075@mail2.gdut.edu.cn; tjeepgdut@foxmail.com; yangxj18@gdut.edu.cn.).

Z. Yang is with the School of Information Engineering, Guangdong University of Technology, Guangzhou, 510006, China, and also with the Guangdong Provincial Key Laboratory of Intellectual Property and Big Data, Guangzhou, 510665, Guangdong, China. (e-mail: yzhj@gdut.edu.cn).

Q. Dai is with the Guangdong Provincial Key Laboratory of Intellectual Property and Big Data, Guangdong Polytechnic Normal University, Guangzhou, 510665, China. (e-mail: dqy@gpnu.edu.cn).

Y. Cheng is with the Faculty of Technology, University of Sunderland, Sunderland, SR6 0DD, UK. (e-mail: yongqiang.cheng@sunderland.ac.uk).

L. Lin is with the School of Data and Computer Science, Sun Yat-sen University, China. (e-mail: linliang@ieee.org).

†The first two authors share equal contributions.

*Corresponding author.

Fig. 1. Visualization of multiple critical information in the modality encoding and interaction stages of the NLVL task. (a) Temporal causal interaction information: Common-sense causal interaction relationships exist in the self-modal contexts of both video and text, e.g., there exist dependencies between "sitting down on a sofa" and "watching television". (b) Discriminative information: The key objects in the video such as sofa, glass, television, and the key words in the sentence such as "sofa", "glass" and "television", are often distinctive from the background. (c) Intra-modal self-correlation information: There are correlations between different video frames for the same objects in the video, and there are also certain dependencies between different words in the sentences.

I. Introduction

VIDEO content analysis has received increasing attention from both academia and industry, which has stimulated the research and application of novel video understanding tasks, such as video retrieval [1], [2] and video question answering [3], [4]. As a classic example of cross-modal information retrieval, video retrieval retrieves the semantically most relevant videos in the trimmed video dataset based on textual sentence queries. However, videos often contain redundant and irrelevant content, that is, only a small fraction of the video clips are semantically relevant to the query [5], [6]. For example, for a long untrimmed surveillance video, only a few short key clips are of interest. To localize these clips, we have to spend several hours manually browsing through the entire video. This process is inefficient and labor-intensive [7].

In view of this, the Temporal Action Localization (TAL) [8] task has been proposed to address the aforementioned problem, which aims to localize video segments (i.e., start and end timestamps) that are semantically relevant to the predefined limited human action query in untrimmed videos. Nevertheless, it is only able to localize within a set of predefined action categories and cannot cover the rich and complex scenarios of the real-world. Therefore, Natural Language Video Localization (NLVL) [9], [10] has garnered significant attention due to its flexibility and practical value. The objective of NLVL is to accurately localize video segments within untrimmed videos that semantically match an arbitrary given natural language query. The main difference between NLVL and TAL is that the query of NLVL can be arbitrary free natural language sentences, whereas the query of TAL can only be predefined lists of a limited number of human action categories. In other words, NLVL can cover more complex and diverse real-world scenarios than TAL and has a higher application value.

Inspired by the TAL, most of the early NLVL works adopted a two-stage ranking approach [9], i.e., sampling candidate video segments by sliding windows in the first stage, and then matching and ranking the query with each candidate segment for relevance in the second stage. Obviously, since the locations and durations of the target moments are unknown and diverse, intensive and overlapping sampling is necessary to achieve high localization performance, which results in inefficiency and high computation resource consumption.

Therefore, proposal-free NLVL methods [11], [12] have been proposed to directly localize the target moments without generating candidate proposals. Specifically, proposal-free NLVL methods focus on cross-modal interactions between text queries and video frames. It can directly compute the probabilities that each frame becomes the start and end boundaries of the target moment. According to the format of moment boundaries, proposal-free NLVL methods can be further divided into regression-based [13] and span-based methods [11], both essentially replace the inefficient two-stage ranking strategy with fine-grained cross-modal interaction. Although the above methods have achieved good performance, there are still following limitations that need to be further addressed.

As shown in Fig. 1, there are some key information in the videos and texts that affect the performance of NLVL, which are neglected in most existing NLVL methods. On one hand, in the analysis of the characteristics of the original multi-modal data, the text description often contains causal interaction. This causal interaction is manifested in the video as long-range semantic association dependence. On the other hand, both video and text modalities contain rich contextual semantic discriminative information, which is very important for distinguishing the foreground from the background. In other words, the role of discriminative information in distinguishing between foreground and background can be seen as a coarse categorization. Moreover, in the cross-modal interaction stage, many existing methods directly and simply capture fine-grained cross-modal interactions and thus learn fusion representations. However, adequate interaction between video and text requires a hierarchical relationship, which includes both inter-modal similarity information and intra-modal self-correlation information. i.e., the hierarchical relationship between video and text is essentially the complementary and collaborative relationship between their intra-modal self-correlation and cross-modal interaction information, which are stacked together to build this hierarchy. This hierarchical information facilitates modalities to complement each other. In addition, many previous cross-modal interaction modules focus only on fine-grained modeling and ignore the importance of coarse-grained interactions. However, only performing coarse-grained modeling would ignore the semantic correlation between key video frames and key words, while only performing fine-grained modeling may bury meaningful global video-text interaction into trivial details, and both schemes would affect the quality of the learned cross-modal fusion representations. Finally, the previous loss functions focus only on the localization information of the two frames at the target moment and overlook the correlation of causal inference between the boundaries. Due to the continuity and long duration of the video, focusing only on two frames at the target moment can be severely affected by the sparsity of the positive samples (i.e., the two frames corresponding to the target moment), resulting in a large localization bias.

In this paper, we propose a novel NLVL framework called Discriminative Parallel and Hierarchical Attention Network (DPHANet) as shown in Fig. 2, where Discriminative Parallel Attention represents parallel extraction and encoding of two key information using the Discriminative Parallel Attention Encoder (DPAE) module, and Hierarchical Attention represents hierarchical cross-modal interaction and self-modal modeling using the Video-Query Hierarchical Attention (VQHA) module. More specifically, it first models the temporal causal interaction information and encodes discriminative information by our proposed DPAE module. Then, a VQHA module is designed to hierarchically perform cross-modal interaction and intra-modal self-correlation modeling with different sub-modules in our proposed hierarchical structure. Moreover, we propose a novel deviation loss function that forces the model to focus on continuity and temporal causality in the video and to capture the correlation of causal inference between the start and end boundaries. Finally, extensive experiments on three benchmark datasets demonstrate the superiority of our proposed DPHANet model, which has achieved about 1.5% and 3.5% average performance improvement and about 2.5% and 7.5% maximum performance improvement on the Charades-STA and TACoS datasets respectively.

In summary, the main contributions of our proposed method are as follows:

- We highlight the importance of temporal causal interaction information and discriminative information in the encoding stage for high-precision NLVL, and correspondingly propose a DPAE module to fully capture and exploit the above-mentioned critical information.
- We design a VQHA module to hierarchically perform cross-modal interaction and intra-modal self-correlation modeling, and finally learn high-quality cross-modal fusion representations.
- In view of the severe sparsity of the positive samples of the localization loss, we propose a novel deviation loss

$\mathcal{L}_{dev}$, to force the model to focus on the continuity and temporal causality in the video.
- We conducted extensive experiments on three benchmark datasets to demonstrate the superiority of our proposed DPHANet framework, which outperforms many state-of-the-art NLVL methods.

II. Related work

In this section, we first briefly review related work on video retrieval and temporal action localization, which are closely related to NLVL. Then, we describe in detail the tasks and recent research progress of NLVL.

A. Video Retrieval

As a typical example of cross-modal retrieval, video retrieval aims to search for the semantically most relevant videos in the trimmed video dataset based on textual sentence queries. Most early video retrieval methods projected textual sentences and videos into a common subspace and then ranked the similarity [14]. However, these methods obviously ignore the characteristic of video modality that is rich in temporal information. To address this problem, recent methods focus on high-quality video encoding and multi-modal fusion to take full advantage of the unique characteristics of video modality. Mithun et al. [15] improved the performance of cross-modal retrieval by fully mining and fusing the multi-modal cues available in the video, such as different visual characteristics, audio, etc. In addition, Liang et al. [16] designed a Local-Global Context Aware Transformer (LOCATER) to solve the problems of missing long-term context capture and visual-linguistic misalignment, thus improving performance. Moreover, Xu et al. [17] developed a Spatiotemporal Decouple-and-Squeeze Contrastive Learning (SDS-CL) framework to contrast multi-level features to learn more abundant representations and capture spatiotemporal specific information. Although video retrieval can only retrieve the whole trimmed videos, its advanced ideas on multi-modal learning and fusion are still very enlightening to NLVL.

B. Temporal Action Localization

Different from video retrieval, temporal action localization requires localizing the exact moments of action instances in the untrimmed videos. Early works used sliding window strategy and hand-crafted features to perform action localization. However, these methods are inefficient and crude when facing complex actions or variable videos. To solve this problem, recent methods follow a two-stage pipeline. For example, Gao et al. [8] proposed a Temporal Unit Regression Network (TURN) model for jointly predicting action proposals and refining temporal boundaries through temporal coordinate regression with contextual information. In addition, Wu et al. [18] designed a transformer-based architecture called TransRMOT to utilize text queries as a semantic cue to guide the prediction of multi-object tracking. Besides, Hui et al. [19] proposed a language-aware spatial-temporal collaboration framework to simultaneously recognize the described actions and provide undisturbed spatial features, better facilitating spatial-temporal collaboration. Moreover, Xu et al. [20] proposed a novel Pyramid Self-attention Polymerization Learning (PSPL) framework to learn multi-level action representations that contain abundant and complementary semantic information via contrastive learning covering coarse-to-fine granularity. Although temporal action localization has achieved good performance, its predefined query list leads to its inability to cover complex scenarios in the real-world.

C. Natural Language Video Localization

Due to the inherent limitations of video retrieval and temporal action localization, NLVL methods have gained lots of attention. Specifically, NLVL can localize semantically relevant target moments in untrimmed videos based on natural language text queries. Early NLVL works primarily employed a two-stage ranking manner, i.e., proposal-based manner. They sample candidate segments and then match them to the queries to retrieve the most relevant segments. For example, Gao et al. [9] proposed a Cross-modal Temporal Regression Localizer (CTRL), which generates candidate segments by sliding windows and jointly models text queries and video candidates. Although these pioneering methods can perform NLVL, they must densely and overlappingly sample to obtain the desired localization results. In view of this, proposal-free methods have been proposed to eliminate the need to generate candidate proposals and directly localize them to the target moment. Depending on the form of the boundaries, it can be further divided into regression-based and span-based methods. The regression-based methods directly calculate the timestamps of the predicted moments and compare them with the target moments for optimization. For example, Ghosh et al. [13] designed three different predictors to directly regress the target moments. The span-based methods aim at calculating the probabilities of each frame being the target moment boundaries. For example, Zhang et al. [11] addressed NLVL from a span-based Question Answering (QA) perspective, and reduced the gap between QA and NLVL with a Query-Guided Highlighting (QGH) strategy.

Furthermore, in the task of NLVL, the quality of the cross-modal interaction module is critical to the performance of NLVL, and many previous methods have been devoted to designing and improving this module. However, most existing methods focus only on fine-grained interactions and ignore the importance of coarse-grained interactions. For example, Zhang et al. [6] proposed a bi-directional attention mechanism to build bi-directional fine-grained information interaction. Only performing fine-grained modeling and lacking coarse-grained modeling may bury meaningful global video-text interaction into trivial details, thus limiting NLVL performance. In addition, many previous methods only perform the cross-modal interactions and lack intra-modal self-correlation modeling to construct a hierarchical structure. For instance, Zhang et al. [11] used a Context-Query Attention (CQA) module to capture the cross-modal interactions between visual and textural features. Focusing only on cross-modal interactions would miss some intra-modal self-correlation information, thus resulting in inadequate interaction and limited performance.

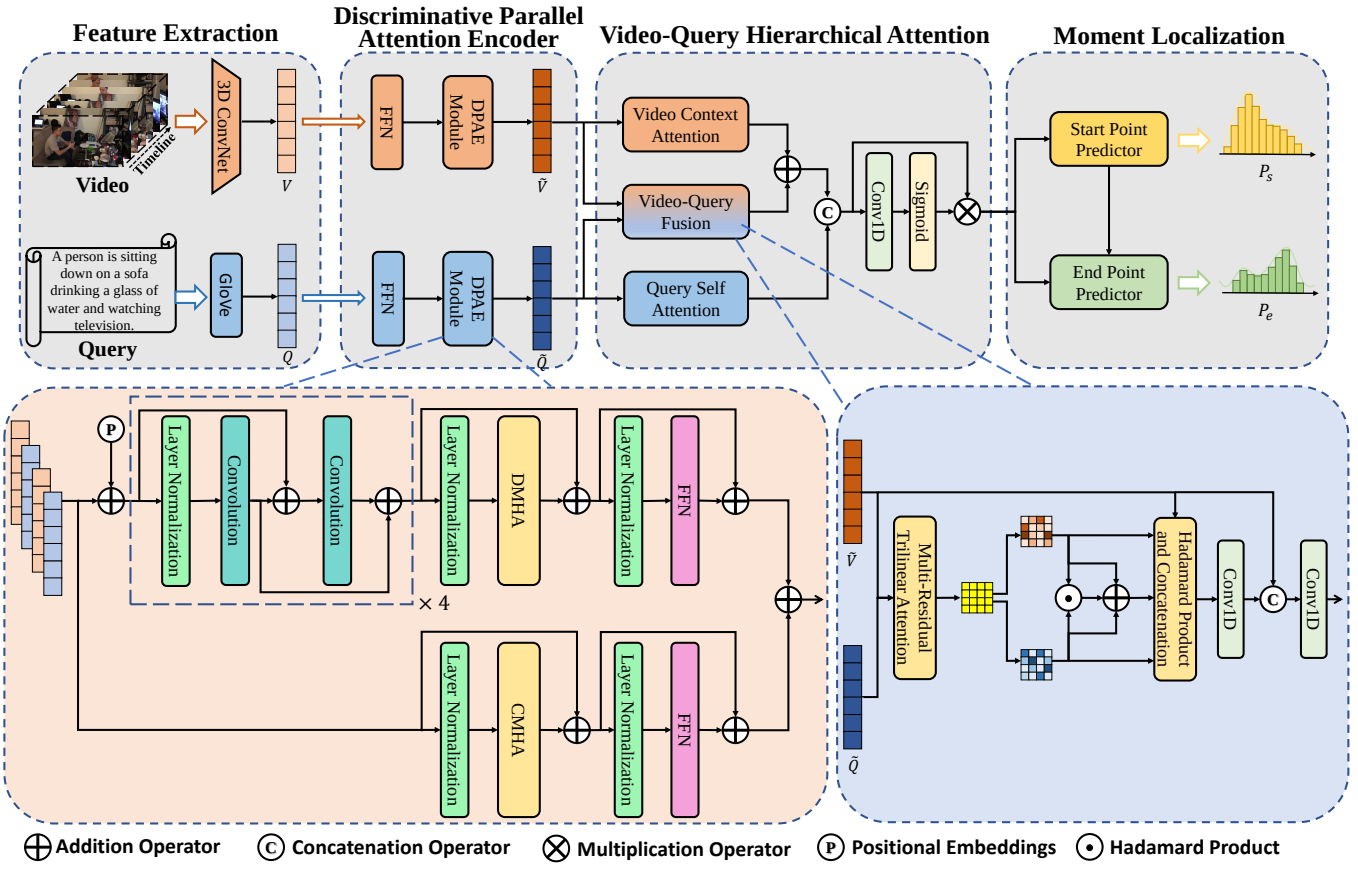


Fig. 2. Our proposed DPHANet framework. It comprises four components: 1) the Feature extraction module employs two independent extraction networks to extract features; 2) the DPAE module utilizes two parallel streams to model temporal causal interaction information and encode discriminative information; 3) the VQHA module performs cross-modal interaction and intra-modal self-correlation modeling in a hierarchical manner; 4) Moment localization module predicts the probabilities of frames to start and end points.

Although the above methods have achieved good performance, they still have some shortcomings that need to be addressed as follows: 1) The existing methods ignore the temporal causal interaction information and contextual semantic discriminative information. 2) The previous methods lack a hierarchical mechanism for cross-modal interaction and intra-modal self-correlation information. 3) The previous loss functions ignore the causal inference between the start and end boundaries of the target moment. In this work, we will design our method to address these issues aiming to achieve high-performance NLVL.

III. The Proposed Model

In this section, we first introduce the preliminary of our NLVL work. After that, we describe the components of our proposed DPHANet framework in detail.

A. Preliminaries

As mentioned earlier, the goal of the NLVL is to localize a temporal segment (i.e., start and end timestamps) that semantically corresponds to the specific natural language description query, in the given untrimmed video. Concretely, we denote the untrimmed video as $V = \{v_t\}_{t=1}^{T}$, and the natural language query as $Q = \{q_j\}_{j=1}^{m}$, where v_t and T represent the t-th image frame and the total number of frames in V, q_j and m indicate the j-th word and the total number of words in Q, respectively. For the above purpose, we describe each training instance as a video-query-timestamps tuple $\{V, Q, \tau_s, \tau_e\}$, where τ_s and τ_e represent the ground-truth start and end timestamps. And then, by representing the desired start and end timestamps as (t_s, t_e), we can formulate the above task as follows:

$$M_{\mathrm{NLVL}}(V, Q) \rightarrow (t_s, t_e), t_s < t_e, \tag{1}$$

where M_{NLVL} is the NLVL model. During the training stage, our proposed model follows an end-to-end manner. In the evaluation stage, our objective is to predict the correct (t_s, t_e) that is as close as possible to (τ_s, τ_e).

B. Our Proposed Model

In this section, we give details of our proposed DPHANet model, as shown in Fig. 2, which consists of the following components: 1) **Feature extraction** module employs two independent extraction networks to efficiently extract features from the raw video frame data and text data. 2) **Discriminative Parallel Attention Encoder** utilizes two parallel streams to model temporal causal interaction information and encode contextual semantic discriminative information for enhancing

representation learning. 3) **Video-Query Hierarchical Attention** module hierarchically performs cross-modal interaction and intra-modal self-correlation modeling to learn high-quality video-query fusion representations. 4) **Moment localization** module predicts the probabilities of frames belonging to start or end points, and localizes the target moments directly. Next, we introduce the compositions and roles of these components in detail.

1) Feature Extraction: For the original untrimmed video V and the raw text query Q, we utilize two independent extraction networks to extract features.

Video feature extraction: We cut the untrimmed video V into n fixed-length clips, and extract its features $\mathbf{V} \in \mathbb{R}^{n \times d_v}$ directly using an off-the-shelf pre-trained 3D ConvNet [21], where d_v denotes the visual feature dimension.

Query feature extraction: For each text query Q, we can transform it into the word embeddings $\mathbf{Q} \in \mathbb{R}^{m \times d_q}$ with the off-the-shelf pre-trained GloVe [22], where d_q indicates the word embedding dimension.

2) Discriminative Parallel Attention Encoder: In order to improve the accuracy of localization, we need to further encode the features of the video and query to fully capture the underlying information and model the temporal information. More specifically, the critical step to high-precision localization is to sufficiently extract the contextual semantic discriminative information and capture the long-range causal interactions between the contexts. Some examples of these two key pieces of information are shown in Fig. 1. For example, considering the query "A person is sitting down on a sofa drinking a glass of water and watching television" and the corresponding video shown in Fig. 1, the desired moment in the video is the entire process of the person performing these behaviors. To achieve high precision localization, on one hand, we should model the causal interaction behavior in contexts, e.g., there exist dependencies between "sitting down on a sofa" and "watching television", and we should capture the latent semantic correlation between them to enhance context interaction. On the other hand, we should also focus on the discriminative information within self-modality, e.g., for video, we should focus on those key frames when key objects appear such as sofa and TV, which are often distinctive from background moments and have a great impact on the localization performance. Although this discriminative information can help us to distinguish between foreground and background, it is unable to identify the start and end points for an action accurately and independently based on discriminative information alone. The role of discriminative information in distinguishing between foreground and background can be seen as a coarse categorization. For example, moments in which key objects are not present can be directly considered as background, while the remaining moments can be coarsely considered as foreground, but that coarse foreground needs to be further distinguished in a fine-grained way. More specifically, this fine-grained distinguishing can be accomplished with the temporal causal interaction information. In summary, discriminative information and temporal causal interaction information can complement each other to accomplish accurate moment localization.

However, the existing methods, e.g., BiLSTM [23], model the context interactions inadequately and ignore the importance of discriminative information inevitably, which limits the performance of localization. Furthermore, the multi-head attention mechanism [24] has been widely used in the field of machine translation and has demonstrated that it is effective for capturing long-range dependencies of video and text. However, the classical multi-head attention module cannot satisfy our need for more adequate temporal causal interaction modeling and contextual semantic discriminative information extraction. In view of this, we propose a novel two-stream encoder module named Discriminative Parallel Attention Encoder (DPAE), as shown in Fig 2.

Specifically, we design two parallel streams to capture adequate temporal causal interaction and contextual semantic discriminative information, respectively. We first employ two linear layers to map video features $\mathbf{V}$ and text features $\mathbf{Q}$ to the same dimension d, i.e, $V' \in \mathbb{R}^{n \times d}$ and $Q' \in \mathbb{R}^{m \times d}$. Taking video modality as an example, for the first stream, which captures the contextual semantic discriminative information, we first employ a position encoding to encode the position information in the video frame sequences, i.e.,

$$F_{s1}^{pe} = \mathrm{PE}\left(V'\right), \tag{2}$$

where $\mathrm{PE}(\cdot)$ denotes the learnable position encoding function, and $F_{s1}^{pe} \in \mathbb{R}^{n \times d}$ is the video feature representation with encoded position information. After that, in order for the model to focus on the objects or local locations with discriminative information in the video, we employ a stacked convolutional block, and each layer of the convolutional block utilizes layer normalization and residual connection. Specifically, each layer is formulated as follows:

$$\begin{aligned} F_{s1}^{cb} &= \mathrm{Convblock}\left(F_{s1}^{pe}\right) \\ &= \mathrm{Conv1D}\left(\mathrm{Conv1D}\left(\mathrm{LN}\left(F_{s1}^{pe}\right)\right) + F_{s1}^{pe}\right) \\ &\quad + \mathrm{Conv1D}\left(\mathrm{LN}\left(F_{s1}^{pe}\right)\right), \end{aligned} \tag{3}$$

where F_{s1}^{cb} represents the features with rich contextual semantic discriminative information after passing through the stacked convolutional block, and $\mathrm{LN}\left(\cdot\right)$ denotes the layer normalization. Specifically, by stacking multiple convolutional blocks, the model can gradually extract increasingly abstract, high-level features from the data. This enables the model to understand features such as shape, texture, and high-level semantics of objects, and improves its ability to perceive specific objects or specific local features, thus making it easier to focus on key objects and specific local regions in the video that are rich in discriminative information.

In order to maintain the consistency of temporal information in both streams and considering the difference in the tasks of these two streams, we design different improved multi-head attention modules for each stream to capture the temporal context information and fit their respective tasks. Specifically, for the first stream, after the convolutional block, a discriminative multi-head attention block (DMHA) is employed to capture the temporal context information. In particular, we replace Q_i^V with $Q_i^V + K_i^V$ in the classic multi-head attention to force the model to focus on the contextual relevance

information within a modality. In other words, it is essentially equivalent to improving the projection space collaboration of Q_i^V and K_i^V while using K_i^V to enhance the extraction of self-modal special information, so as to enhance the ability of the model to capture long-range dependence and discriminative information. It can be formulated as follows:

$$\begin{aligned} h_i^{s1} &= \mathrm{DMHAttn}\left(Q_i^V, K_i^V, V_i^V\right) \\ &= \mathrm{Softmax}\left(\frac{\left(Q_i^V + K_i^V\right)\left(K_i^V\right)^{\mathrm{T}}}{\sqrt{d_K}}\right) V_i^V, \end{aligned} \tag{4}$$

where Q_i^V, K_i^V and V_i^V denote the i-th linear projections of query, key, and value of video samples, d_K is the dimension of key, and h_i^{s1} represents the i-th head attention in the DMHA. Finally, by adding layer normalization and residual connection, the output of the DMHA F_{s1}^{mha} and the output of the first stream F_{s1}^{out} are respectively as follows:

$$\begin{aligned} F_{s1}^{mha} &= \mathrm{DiscMultiHead}\left(\mathrm{LN}\left(F_{s1}^{cb}\right)\right) + F_{s1}^{cb} \\ &= \left[h_1^{s1}; h_2^{s1}; \cdots h_H^{s1}\right] + F_{s1}^{cb}, \end{aligned} \tag{5}$$

$$F_{s1}^{out} = \mathrm{FFN}\left(\mathrm{LN}\left(F_{s1}^{mha}\right)\right) + F_{s1}^{mha}. \tag{6}$$

For the second stream, which aims at modeling the causal interaction behavior in temporal contexts, we design a causal multi-head attention block (CMHA) for construction but no position encoding and convolutional block. CMHA and DMHA are similar but slightly different. Specifically, on the basis of DMHA, CMHA further normalizes the V_i^V as follows:

$$\Phi\left(V_i^V\right) = \frac{\left(V_i^V\right)^e}{1 + \log\left(\left\|\left(V_i^V\right)^e\right\|_{\mathrm{F}}\right)}, \tag{7}$$

$$\begin{aligned} h_i^{s2} &= \mathrm{CMHAttn}\left(Q_i^V, K_i^V, V_i^V\right) \\ &= \mathrm{Softmax}\left(\frac{\left(Q_i^V + K_i^V\right)\left(K_i^V\right)^{\mathrm{T}}}{\sqrt{d_K}}\right) \Phi\left(V_i^V\right), \end{aligned} \tag{8}$$

where $\Phi(\cdot)$ is our proposed novel normalization function for the value linear projections, and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix. In this way, it can reduce the interference of abnormal sample data, and is essentially equivalent to reconstructing the projected feature space of V_i^V, thus enhancing the intra-modal representation learning with quality and robustness. More specifically, CMHA allows the model to focus on different locations and features at the same time, thus capturing the complex spatio-temporal correlations between video frames. In addition, each attention head can focus on different aspects, such as motion direction, spatial location, etc., thus modeling behaviors in multiple dimensions and enabling the model to understand the causal relationships between different behaviors in a more comprehensive way. Furthermore, our proposed novel normalization function can reconstruct the projected feature space, allowing the model to focus more on semantic continuity and causality. After that, similar to the first stream, the output of the CMHA F_{s2}^{mha} and the output of the second stream F_{s2}^{out} are respectively as follows:

$$\begin{aligned} F_{s2}^{mha} &= \mathrm{CausMultiHead}\left(\mathrm{LN}\left(V'\right)\right) + V' \\ &= \left[h_1^{s2}; h_2^{s2}; \cdots; h_H^{s2}\right] + V', \end{aligned} \tag{9}$$

$$F_{s2}^{out} = \mathrm{FFN}\left(\mathrm{LN}\left(F_{s2}^{mha}\right)\right) + F_{s2}^{mha}. \tag{10}$$

Finally, by weighted addition between the results of the two-stream outputs, we can obtain high-quality video modality representations $\widetilde{V}$ with rich contextual semantic discriminative information and causal interaction information, i.e.,

$$\widetilde{V} = \lambda_{s1} F_{s1}^{out} + \lambda_{s2} F_{s2}^{out}, \tag{11}$$

where λ_{s1} and λ_{s2} are the balancing parameters.

It is worth noting that the text modality representations $\widetilde{Q}$ can also be obtained by the same module above. As mentioned before, this two-stream structure can adequately capture temporal causal interaction and contextual semantic discriminative information, respectively. It can solve the problem of inadequate extraction and modeling for key information by single-stream structure. To further verify this conclusion, we performed the corresponding ablation experiments in subsection IV-C.

3) Video-Query Hierarchical Attention: After feature encoding, we need to model the cross-modal interaction to learn video-query fusion representations. To this end, context-query attention (CQA) [25] is a simple yet effective method, which captures the fine-grained visual-text interaction directly. However, the classic CQA module suffers from the following two major shortcomings: 1) It focuses only on the fine-grained modeling and ignores the importance of coarse-grained interactions; 2) the fusion representations learned by CQA lack hierarchical relationship for inter-modal similarity information and intra-modal self-correlation information. Specifically, both fine-grained and coarse-grained modeling are important for cross-modal interactions. Only performing coarse-grained modeling would ignore the semantic correlation between key video frames and key words, while only performing fine-grained modeling may bury meaningful global video-text interaction into trivial details, and both schemes would affect the quality of the learned cross-modal fusion representations, thus reducing NLVL performance. On the contrary, performing joint coarse- and fine-grained two-level interaction can well overcome the above problems to learn high-quality fusion representations enriched with cross-modal detail information and global clues, thus improving NLVL performance.

In view of this, we propose a novel cross-modal interaction module, named Video-Query Hierarchical Attention (VQHA), which contains Video-Query Fusion, Video Context Attention, and Query Self-Attention. Specifically, inspired by CQA, the Video-Query Fusion sub-module first calculates the similarity scores $\mathcal{S} \in \mathbb{R}^{n \times m}$ between fine-grained cross-modal features by our proposed multi-residual trilinear attention, i.e.,

$$\begin{aligned} \mathcal{S} &= \mathrm{MultiResTriAttn}\left(\widetilde{V}, \widetilde{Q}\right) \\ &= \widetilde{V} W_1 + W_2 \widetilde{Q}^T + \widetilde{V}\widetilde{Q}^T + (W_2 \widetilde{Q}^T) \odot (\widetilde{V}\widetilde{Q}^T) \end{aligned}, \tag{12}$$

where $\mathrm{MultiResTriAttn}(\cdot)$ is our proposed multi-residual trilinear attention, W_1 and W_2 are the learnable weights, and $\odot$ is the element-wise multiplication. This method is different from trilinear attention by adding more residual connections and a more fine-grained attention item. Specifically, the main

reason for adding more residual connections is that it allows the model to learn more fine-grained similarity information while retaining the previously learned similarity information, enabling the model to learn more comprehensive similarity information. In addition, this more fine-grained similarity information is obtained through $(W_2\widetilde{Q}^T)\odot(\widetilde{V}\widetilde{Q}^T)$. It can capture the fine-grained attention between query projection and cross-modal similarity and merge it into the learned similarity information, thus enhancing the fine-grained of similarity. Then the video-to-query ($\mathcal{A}_{V2Q}$) and query-to-video ($\mathcal{A}_{Q2V}$) attention weights can be computed as:

$$\begin{aligned}\mathcal{A}_{V2Q} &= \mathcal{S}_r \cdot \widetilde{Q} \in \mathbb{R}^{n\times d},\\ \mathcal{A}_{Q2V} &= \mathcal{S}_r \cdot \mathcal{S}_c^{\mathrm{T}} \cdot \widetilde{V} \in \mathbb{R}^{n\times d},\end{aligned} \tag{13}$$

where $\mathcal{S}_r$ and $\mathcal{S}_c$ denote the row- and column-wise normalization of $\mathcal{S}$. In addition, in order to further encode both fine-grained and coarse-grained interactions, we further construct the following two items:

$$\begin{aligned}\mathcal{G} &= \mathcal{A}_{V2Q} + \mathcal{A}_{Q2V} + \mathcal{A}_{V2Q} \odot \mathcal{A}_{Q2V},\\ \mathcal{L} &= \mathcal{G} - \widetilde{V},\end{aligned} \tag{14}$$

where $\mathcal{G}$ and $\mathcal{L}$ are the fine-grained and coarse-grained interaction information, respectively. This coarse-grained interaction information represents a global or large-scale cross-modal interaction clue, i.e., the interactions between the segment-level and the phrase-level. Finally, the output of Video-Query Fusion sub-module $V^Q \in \mathbb{R}^{n\times d}$ can be written as:

$$\begin{aligned}V^q &= \mathrm{FFN}\left(\left[\widetilde{V}; \mathcal{A}_{V2Q}; \widetilde{V}\odot\mathcal{A}_{V2Q}; \widetilde{V}\odot\mathcal{A}_{Q2V}; \mathcal{G}\right]\right),\\ V^Q &= \mathrm{FFN}\left(\left[\widetilde{V}; \mathcal{L}; V^q\right]\right),\end{aligned} \tag{15}$$

where $\odot$ is the element-wise multiplication. In this way, we can obtain high-quality representations with fine-grained and coarse-grained interaction information, thus achieving efficient and adequate cross-modal interaction. However, we still need to model the intra-modal self-correlation information in order to obtain a higher-quality hierarchical fusion representation. Although the 3D ConvNet has modeled the video representations, it lacks the extraction of intra-modal self-correlation information to obtain more adequate self-modal modeling, which limits its performance. In view of this, we further propose two sub-modules Video Context Attention and Query Self-Attention to extract intra-modal self-correlation information, while forming a hierarchical structure with Video-Query Fusion to construct the VQHA module.

Specifically, for convenience, our proposed Video Context Attention sub-module uses the same module structure as the Video-Query Fusion sub-module, changing only the inputs, to obtain the video context self-modal representations V^V. Then we fuse the V^Q and V^V by weighted addition, and denote it as $V^{QV} = V^Q + \mu V^V$, where μ is the trade-off parameter.

Notably, unlike video modality, most of the natural language queries in NLVL datasets are relatively simple and short, so they carry relatively limited semantic information. In this case, intra-modal self-correlation information within the text modality can be well captured using only the self-attention mechanism, which has been demonstrated to be simple and effective in many previous works. In contrast, due to the limitations of the semantic information carried by natural language queries, the use of the same sub-module as the Video-Query Fusion for Query Self-Attention sub-module does not bring significant performance gains but increases the computational consumption of the model. Therefore, for the Query Self-Attention sub-module, we directly encode $\widetilde{Q}$ into sentence representations h_q by self-attention mechanism. Then h_q is concatenated with each element in V^{QV} to obtain the final high-quality cross-modal fusion representations $\widetilde{V}^Q = [\widetilde{v}_1^q, \cdots, \widetilde{v}_n^q]$, where $\widetilde{v}_i^q = [v_i^{qv}; h_q]$.

In fact, the hierarchical relationship between video and text is essentially the complementary and collaborative relationship between their intra-modal self-correlation and cross-modal interaction information, which are stacked together to build this hierarchy. Stacking these complementary features with each other constructs a hierarchical structure that helps to improve the quality of the fusion representations. It is worth noting that these three types of information are of equal importance, so the hierarchical structure constructed has no sorting requirements or level of ranking.

4) Moment Localization Module: Finally, we present a moment localization module to predict the start and end boundaries directly. It is worth noting that we adopt a proposal-free strategy, which overcomes the high complexity and inefficiency of the proposal-based methods. Firstly, inspired by Ref. [10], we employ a Query-Guided Highlighting (QGH) strategy to enhance our cross-modal fusion representations, which regards the target moment as the foreground and the rest as the background and further extends the boundary. It can be expressed as follows:

$$\begin{aligned}S' &= \sigma\left(\mathrm{FFN}\left(\widetilde{V}^Q\right)\right),\\ \widehat{V}^Q &= S' \cdot \widetilde{V}^Q,\end{aligned} \tag{16}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function, $S' \in \mathbb{R}^n$ indicates the highlight score, and $\widehat{V}^Q \in \mathbb{R}^{n\times d}$ represents the final fusion representations after QGH processing. Correspondingly, the QGH loss function can be calculated as follows:

$$\mathcal{L}_{qgh} = f_{CE}\left(S', Y_h\right), \tag{17}$$

where $Y_h \in \mathbb{R}^n$ denotes a binary sequence vector, whose elements are equal to 1 if the moment belongs to the foreground and 0 otherwise. After that, we construct the moment localization module with two Transformer blocks and two FFNs to calculate the start (P_s) and end (P_e) probability distributions for every frame as follows:

$$\begin{aligned}\widehat{V}_s^Q &= \mathrm{TRM_s}\left(\widehat{V}^Q\right), P_s = \mathrm{Softmax}\left(W_s\left[\widehat{V}_s^Q; \widehat{V}^Q\right] + b_s\right),\\ \widehat{V}_e^Q &= \mathrm{TRM_e}\left(\widehat{V}_s^Q\right), P_e = \mathrm{Softmax}\left(W_e\left[\widehat{V}_e^Q; \widehat{V}^Q\right] + b_e\right),\end{aligned} \tag{18}$$

where $\widehat{V}_s^Q$ and $\widehat{V}_e^Q$ denote the outputs of Transformer blocks, while $\mathrm{TRM_s}(\cdot)$ and $\mathrm{TRM_e}(\cdot)$ are two Transformer blocks. W_s and W_e are the learnable weights of FFNs, and b_s and b_e are the corresponding bias.

TABLE I
THE STATISTICS OF NLVL BENCHMARK DATASETS. WHERE N_q, $\bar{N}_q$, $\bar{L}_v$, AND $\bar{L}_m$ DENOTE VOCABULARY SIZE, WORD AVERAGE NUMBER, VIDEO AVERAGE LENGTH, AND MOMENT AVERAGE LENGTH, RESPECTIVELY.

Dataset	Division	Videos	Annotations	N_q	$\bar{N}_q$	$\bar{L}_v$	$\bar{L}_m$
Charades-STA	train	5,338	12,408	1,303	7.22	30.59s	8.22s
	val	-	-				
	test	1,334	3,720				
ActivityNet	train	10,009	37,421	12,460	14.78	117.61s	36.18s
	val	-	-				
	test	4,917	17,505				
TACoS	train	75	10,146	2,033	10.05	287.14s	5.45s
	val	27	4,589				
	test	25	4,083				

C. Loss Function

The total loss function of our model consists of three parts, i.e., localization loss $\mathcal{L}_{loc}$, QGH loss $\mathcal{L}_{qgh}$, and our proposed deviation loss $\mathcal{L}_{dev}$, where QGH loss is shown in Eq. (17).

The localization loss $\mathcal{L}_{loc}$ is the frequently-used and critical loss for proposal-free localization. Following the previous methods, we compare the probability distributions P_s and P_e obtained by the moment localization module with the one-hot label vectors Y_s and Y_e of start and end boundaries, and calculate the cross-entropy between them, which can be calculated as follows:

$$\mathcal{L}_{loc} = \frac{1}{2}\left[f_{CE}\left(P_s, Y_s\right) + f_{CE}\left(P_e, Y_e\right)\right], \tag{19}$$

where $f_{CE}(\cdot)$ denotes the cross-entropy loss function. However, the above localization loss function still has a limitation, i.e., it focuses only on the localization information at the start and end timestamp positions, but ignores the correlation of causal inference between the start and end boundaries of the target moment. This may cause the localization results for some samples to have significantly deviated from the start or end boundaries, thus resulting in inaccurate start and end boundary calibrations. To overcome the above problem, we propose a novel deviation loss $\mathcal{L}_{dev}$, which forces the deviations of the start and end boundaries to be as close as possible to capture the causal inference between them. Therefore, the above deviation loss $\mathcal{L}_{dev}$ can be formulated as follows:

$$\mathcal{L}_{dev} = f_{CE}\left(P_e - P_s, Y_e - Y_s\right). \tag{20}$$

Finally, by combining the above three loss functions in Eqs. (17), (19), and (20), we can train our model in an end-to-end manner by the following overall objective function:

$$\mathcal{L} = \mathcal{L}_{loc} + \mathcal{L}_{qgh} + \omega\mathcal{L}_{dev}, \tag{21}$$

where ω is the balancing parameter.

In the inference stage, the predicted boundaries are determined by maximizing the joint probability as follows:

$$t_s, t_e = \arg\max_{t_s, t_e} P_s\left(t_s\right) P_e\left(t_e\right), s.t.\ t_s \leq t_e. \tag{22}$$

IV. EXPERIMENTS

A. Experiment Setup

1) Dataset: We conducted experiments and analyses on three public datasets, i.e., Charades-STA [9], ActivityNet Captions [26], and TACoS [27], which have been widely used in the NLVL domain. Therefore, we summarize the statistics of the three datasets in Table I, and describe the specific information as follows:

Charades-STA: It contains 6,672 videos of indoor activities and 16,128 corresponding natural language temporal annotations, which was extended from the original Charades dataset [28] by Gao et al. [9]. Specifically, the average length of the videos and the average length of the temporal moments are 30.59 seconds and 8.22 seconds, respectively. Correspondingly, the vocabulary size of words and the average number of words in the natural language temporal annotations are 1,303 and 7.22. Following the setup of the previous methods, the training set consists of 12,408 moment-annotation pairs and the remaining 3,720 pairs form the test set.

ActivityNet Captions: It is a large dataset and widely used in previous NLVL works, which contains about 20,000 open videos collected from ActivityNet [29] and their corresponding temporal annotations. The average length of each video is 117.61 seconds, while the average length of each moment is 40.18 seconds. Similarly, the vocabulary size of words and the average number of words in sentence annotations are 12,460 and 14.78, respectively. Following Ref. [30], 37,421 moment-annotation pairs are selected to construct the training set, while the other 17,505 pairs are for testing.

TACoS: It contains 127 long videos constructed from Max Planck Institute for Informatics (MPII) cooking composite activities [31], all of which are cooking activities, and the lengths of the videos (287.14 seconds on average) are long while the lengths of target moments are short, making it a difficult dataset. In addition, there are two versions of TACoS available for experiments, i.e., TACoS$_{org}$ and TACoS$_{tan}$. Specifically, the training set, validation set, and test set of TACoS$_{org}$ from Gao et al. [9] contain 10,146, 4,589, and 4,083 moment-annotation pairs, respectively. While the TACoS$_{tan}$ from Zhang et al. [32] utilizes 9,790, 4,436, and 4,001 moment-annotation pairs to construct the training set, validation set, and test set, respectively. We followed the above setup and conducted the experiments separately.

2) Implementation Details: Following the previous methods, for the untrimmed video V, we first downsample or zero-pad it into fixed-length n, and then extract its features using an off-the-shelf 3D ConvNet pre-trained on Kinetics dataset [21]. Specifically, the parameter size of 3D ConvNet is 79M, and the Kinetics dataset has 400 human action classes with more than 400 examples for each class. Moreover, the maximal feature lengths n are set to 100, 256, and 300 for Charades-STA, ActivityNet Captions, and TACoS, respectively. Where downsampling is performed when the video length is longer than n, and otherwise zero-padding is performed. For the textual query Q, we lowercase all its words and then initialize them with a fixed 300-dimensional GloVe embedding (840B tokens, 2.2M vocab, cased, 300d vectors trained on the Common Crawl). For

the model parameters, the dimension of hidden layers is set to 128, while the kernel size of the convolution layer and the head size of multi-head attention are set to 7 and 16, respectively. In addition, the balancing parameters λ_{s1} and λ_{s2} in Eq. (11) are set to 0.7 and 0.9 respectively according to subsection IV-D. Similarly, the balancing parameter ω for the loss function in Eq. (21) is set to 0.001 according to the experimental analysis. Furthermore, we set the trade-off parameter μ as 0.01. During the training stage, for all datasets, the batch size and epoch are set to 16 and 100 with an early stopping strategy. Adam optimizer [33] is employed to optimize parameters, with the linear decay of learning rate with an initial 0.001 and gradient clipping of 1.0. To prevent over-fitting, we adopt Dropout [34] with a dropout ratio of 0.2. All our experiments are conducted on a server equipped with four NVIDIA TITAN RTX GPUs.

3) Evaluation Metrics: In this paper, we adopt the standard evaluation metrics in NLVL, i.e., "Rank@n, IoU=μ" and "mIoU", which are widely used in this field, to evaluate the localization performance. Specifically, IoU denotes the Intersection over Union between the predicted moment and ground truth, n denotes the top-n samples and μ denotes the threshold, respectively. Therefore, the "Rank@n, IoU=μ" represents the percentage of queries in which at least one of the IoU between the top-n localization moments and ground-true is greater than μ, while mIoU denotes the average IoU of all test samples. More specifically, for all datasets, we set the n as 1 and use $\mu \in \{0.3, 0.5, 0.7\}$.

B. Comparison With State-of-The-Arts

1) Baselines: To verify the performance of our proposed method, we compare it experimentally with several state-of-the-art methods on the above three datasets, which include seven proposal-based methods, seven proposal-free methods, and two other methods. Specifically, these methods are as follows:

- *Proposal-Based Methods:* Cross-modal Temporal Regression Localizer (CTRL) [9], Interaction-Integrated Network (I2N) [35], Semantic Conditioned Dynamic Modulation (SCDM) [36], Contextual Boundary-aware Prediction (CBP) [37], Fast Video Moment Retrieval (FVMR) [38], Cross-modal Dynamic Networks (CDN) [39], Progressive Localization Network (PLN) [40].
- *Proposal-free Methods:* Dense Regression Network (DRN) [41], Boundary Proposal Network (BPNet) [42], Graph-FPN with Dense Predictions (GDP) [43], Cross Interaction Multi-Head Attention (CI-MHA) [12], Query-Controlled Temporal Convolution (PEARL) [44], Video Span Localizing Network (VSLNet) [10], Multimodal, Multichannel, and Dual-step Capsule Network (M^2DCapsN) [45].
- *Other Methods:* Multi-Agent Boundary-Aware Network (MABAN) [46], Text-Visual Prompting (TVP) [47].

It is worth noting that our proposed DPHANet model belongs to the proposal-free method, but we compare it with different types of NLVL methods to fully demonstrate its superior performance.

TABLE II
PERFORMANCE EVALUATION RESULTS ON CHARADES-STA USING C3D/VGG/I3D FEATURES.

Feature	Methods	Rank@1, IoU=μ			mIoU
		μ=0.3	μ=0.5	μ=0.7	
C3D	CTRL [9] (ICCV'17)	-	23.63	8.89	-
	CBP [37] (AAAI'20)	-	36.80	18.87	35.74
	GDP [43] (AAAI'20)	54.54	39.47	18.49	-
VGG	TVP [47] (CVPR'23)	65.92	44.39	21.51	-
	CDN [39] (TMM'22)	-	45.24	26.99	-
	PLN [40] (TOMM'23)	68.60	56.02	35.16	49.09
I3D	DRN [41] (CVPR'20)	-	53.09	31.75	-
	VSLNet [10] (ACL'20)	70.46	54.19	35.22	-
	BPNet [42] (AAAI'21)	65.48	50.75	31.64	46.34
	SCDM [36] (TPAMI'22)	-	54.44	33.43	-
	FVMR [38] (ICCV'21)	-	55.01	33.74	-
	CI-MHA [12] (SIGIR'21)	69.87	54.68	35.27	-
	PEARL [44] (WACV'22)	<u>71.90</u>	53.50	<u>35.40</u>	<u>51.20</u>
	I2N [35] (TIP'21)	-	<u>56.61</u>	34.14	-
	MABAN [46] (TIP'21)	-	56.29	32.26	-
	M^2DCapsN [45] (TNNLS'23)	-	55.03	31.61	-
	Ours	**72.07**	**58.17**	**37.77**	**52.47**

Note: The experimental result data for our comparison methods are taken from the original papers. For convenience and accuracy, if the original papers of some methods did not perform experiments on specific datasets, we will omit the results of the corresponding experiments.

2) Performance Analysis: Tables II, III, and IV report the localization performance evaluation results of our proposed method and the above state-of-the-art methods on the Charades-STA, ActivityNet Captions, and TACoS datasets, respectively, which are expressed in terms of "Rank@1, IoU=μ" and "mIoU". For clarity, the best result for each item is shown in bold, while the second-best result is underlined. Specifically, according to the results shown in the tables, we have the following conclusions:

For the Charades-STA dataset, our proposed DPHANet model consistently outperforms all the baseline methods in all metrics, especially in the high-precision strict metric "Rank@1, IoU=0.7", which is about 2.5% higher than the best baseline. In particular, our proposed DPHANet framework belongs to the proposal-free method, but compared with those proposal-free baseline methods, i.e., DRN, BPNet, GDP, CI-MHA, PEARL, VSLNet, and M^2DCapsN, it still achieves a significant improvement. One possible reason is that our proposed DPAE module can encode temporal causal interaction information and contextual semantic discriminative information to enhance representation learning and further facilitate the high-precision NLVL, demonstrating the superiority of our proposed method.

For the ActivityNet Captions dataset, the superiority of our proposed method on this dataset is less. One possible reason is that the ActivityNet Captions dataset is open-world oriented and also has a large standard deviation of the length of its complex and variable target moments, which can introduce more noise, indistinguishable clues, and subjective biases, thus affecting the stability of NLVL. However, although our proposed method exhibits less superiority on this dataset compared to those on the other two datasets, it still shows a competitive performance, demonstrating its excellent perfor-

TABLE III
PERFORMANCE EVALUATION RESULTS ON ACTIVITYNET CAPTIONS USING C3D FEATURES.

Methods	Rank@1, IoU=μ			mIoU
	μ=0.3	μ=0.5	μ=0.7	
SCDM [36] (TPAMI'22)	54.80	36.75	19.86	-
CBP [37] (AAAI'20)	54.30	35.76	17.80	36.85
FVMR [38] (ICCV'21)	60.63	45.00	26.85	-
PLN [40] (TOMM'23)	59.65	45.66	29.28	**44.12**
DRN [41] (CVPR'20)	-	45.45	24.36	-
BPNet [42] (AAAI'21)	58.98	42.07	24.69	42.11
GDP [43] (AAAI'20)	56.17	39.27	-	39.80
CI-MHA [12] (SIGIR'21)	61.49	43.97	25.13	-
VSLNet [10] (ACL'20)	**63.16**	43.22	26.16	43.19
M^{2}DCapsN [45] (TNNLS'23)	61.41	**47.03**	**29.99**	-
MABAN [46] (TIP'21)	-	42.42	24.34	-
TVP [47] (CVPR'23)	60.71	43.44	25.03	-
Ours	59.87	43.29	27.37	43.78

TABLE IV
PERFORMANCE EVALUATION RESULTS ON TACOS USING C3D FEATURES.

Dataset	Methods	Rank@1, IoU=μ			mIoU
		μ=0.3	μ=0.5	μ=0.7	
TACoS$_{org}$	CTRL [9] (ICCV'17)	18.32	13.30	-	-
	SCDM [36] (TPAMI'22)	26.11	21.17	-	-
	CBP [37] (AAAI'20)	27.31	24.79	19.10	21.59
	DRN [41] (CVPR'20)	-	23.17	-	-
	BPNet [42] (AAAI'21)	25.96	20.96	14.08	19.53
	GDP [43] (AAAI'20)	24.14	-	-	16.18
	VSLNet [10] (ACL'20)	29.61	24.27	20.03	24.11
	Ours	**37.17**	**29.24**	**20.22**	**28.77**
TACoS$_{tan}$	I2N [35] (TIP'21)	31.47	29.25	-	-
	FVMR [38] (ICCV'21)	41.48	29.12	-	-
	CDN [39] (TMM'22)	43.09	32.82	-	-
	PLN [40] (TOMM'23)	43.89	31.12	-	29.70
	PEARL [44] (WACV'22)	42.94	32.07	18.37	31.08
	M^{2}DCapsN [45] (TNNLS'23)	46.41	32.58	-	-
	Ours	**47.01**	**34.12**	**23.59**	**33.95**

mance in NLVL.

For the TACoS dataset, its long video length, short duration of the target moments, dense and indistinguishable activities make this dataset a very challenging one. Therefore, considering the results on Charades-STA and ActivityNet Captions together, most methods have poorer localization performance on TACoS. However, in this case, our proposed method still significantly outperforms all baseline methods in all metrics on both TACoS$_{org}$ and TACoS$_{tan}$, demonstrating that it can understand the semantic information of the activities in the videos well, and thus achieving excellent performance even on very challenging datasets. More specifically, the "Rank@1, IoU=0.3" result of our proposed method on TACoS$_{org}$ is about 7.5% substantially higher than the best baseline, while the results of other metrics on TACoS$_{org}$ and TACoS$_{tan}$ also show a significant improvement ranging between 2% and 5% against the best baseline. One possible reason for this excellent performance is that our proposed VQHA module can hierarchically perform cross-modal interaction and intra-modal self-correlation modeling to more fully capture the dense and indistinguishable activities in the videos from this challenging dataset and further learn the high-quality fusion representations, thus greatly enhancing the localization performance.

Overall, our proposed method outperforms the baseline methods in most cases on benchmark datasets, which demonstrates the superiority and outstanding performance of our method in NLVL. Meanwhile, it also highlights the plausibility and effectiveness of our designed modules and loss function, and we will verify it further by the ablation experiments in the next subsection.

C. Ablation Study

To validate the effectiveness of key modules of our proposed framework, which includes DPAE, VQHA, and deviation loss $\mathcal{L}_{dev}$, etc., we conducted the ablation studies on the Charades-STA, TACoS, and ActivityNet Captions datasets as follows. Specifically, we have the following ablation models:

1) ***DPHA***$_{base}$: It represents the baseline of our proposed DPHANet framework. Concretely, we removed all the key modules, i.e., DPAE, VQHA, and deviation loss $\mathcal{L}_{dev}$, and replaced them with some classic components from existing methods to construct the most basic baseline for comparison.

2) ***DPAE***: We discarded or replaced the submodules of the DPAE module to verify whether the introduction of semantic discriminative information and temporal causal interaction information is effective, i.e., the effectiveness of the DPAE module. The details are as follows:

- *Single-stream TRM:* We replaced the whole DPAE module with a single-stream Transformer.
- *TRM + CMHA:* We kept the two-stream structure of DPAE and replaced the DMHA with a standard Transformer.
- *TRM + DMHA:* We maintained the two-stream structure of DPAE and replaced the CMHA with a standard Transformer.
- *Two-stream TRM:* We retained the two-stream structure of DPAE and replaced the DMHA and CMHA with two standard Transformer, i.e., the two-stream standard Transformer.

3) ***VQHA***: We discarded or replaced the submodules of VQHA module to verify whether the introduction of hierarchical attention and the combination of coarse-grained and fine-grained interactions are effective, i.e., the effectiveness of VQHA module. The details are as follows:

- *CQA + w/o. Hierarchical:* We replaced the whole VQHA module with a CQA module, and removed the hierarchical mechanism, i.e., discarding the intra-modal self-correlation information.
- *CQA + Hierarchical:* We replaced the whole VQHA module with a CQA module, and maintained the hierarchical mechanism.
- *w/o. Hierarchical:* We removed the hierarchical mechanism, and kept the combination of coarse-grained and fine-grained interactions.
- *w/o. MultiResTriAttn:* We only replaced the multi-residual trilinear attention with the trilinear attention.

4) ***Loss***: We analyzed the influence of our proposed $\mathcal{L}_{dev}$ to the NLVL performance, to verify the effectiveness of our loss function. The details are as follows:

- *w/o.* $\mathcal{L}_{dev}$*:* We removed the $\mathcal{L}_{dev}$ in the overall objective function and constructed the overall objective function using only $\mathcal{L}_{loc}$ and $\mathcal{L}_{qgh}$.

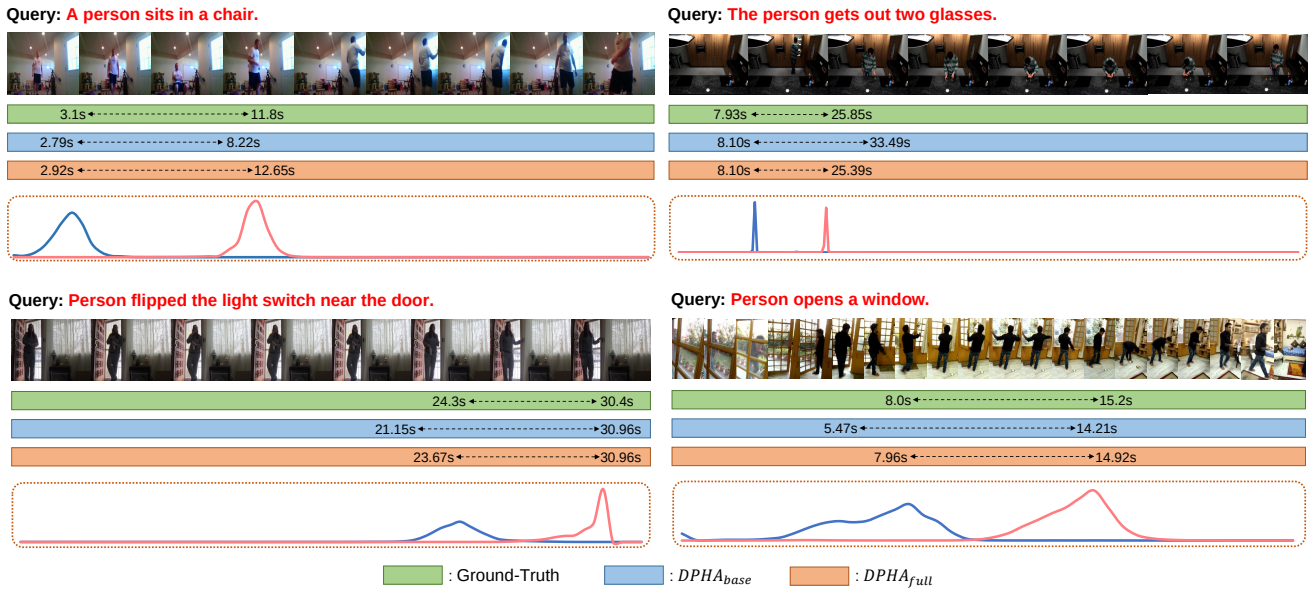


Fig. 3. Visualization results of the NLVL by our proposed DPHANet model.

TABLE V
ABLATION STUDIES OF OUR PROPOSED DPHANET MODEL.

Ablation	Method	Charades-STA				TACoS				ActivityNet Captions			
		μ=0.3	μ=0.5	μ=0.7	mIoU	μ=0.3	μ=0.5	μ=0.7	mIoU	μ=0.3	μ=0.5	μ=0.7	mIoU
$DPHA_{base}$	CQA+TRM	68.44	54.01	35.38	49.62	41.04	29.74	19.47	29.97	55.27	38.58	23.32	40.70
DPAE	Single-stream TRM	70.54	54.68	36.77	51.10	43.51	31.37	20.44	31.86	57.23	41.59	25.54	41.49
	TRM + CMHA	71.72	54.76	37.12	51.52	45.46	33.49	22.57	33.19	58.76	42.16	26.54	43.23
	TRM + DMHA	71.10	55.43	37.50	51.66	45.86	33.27	21.54	32.97	58.58	42.40	26.64	43.02
	Two-stream TRM	70.43	54.46	36.69	51.00	43.81	31.89	20.74	31.69	58.00	41.36	25.86	42.80
VQHA	CQA + w/o.Hierarchical	70.40	53.82	36.02	50.86	42.66	31.72	20.84	31.09	57.70	41.31	25.23	41.54
	CQA + Hierarchical	71.77	55.27	36.69	51.79	43.74	32.24	22.42	32.30	58.40	42.28	26.59	42.67
	w/o. Hierarchical	70.86	54.84	36.99	51.10	44.91	31.77	21.57	32.64	58.57	42.01	26.65	43.06
	w/o. MultiResTriAttn	71.51	56.16	37.28	51.95	45.36	32.34	20.29	32.15	58.60	42.39	26.92	43.04
Loss	w/o. $\mathcal{L}_{dev}$	70.70	53.95	36.37	50.87	43.79	31.62	21.42	32.12	58.02	41.55	25.87	42.67
$DPHA_{full}$	DPHA	**72.07**	**58.17**	**37.77**	**52.47**	**47.01**	**34.12**	**23.59**	**33.95**	**59.87**	**43.29**	**27.37**	**43.78**

5) ***DPHA***$_{full}$: This model represents our complete DPHANet framework.

All the ablation study results on the Charades-STA, TACoS, and ActivityNet Captions datasets are reported in Table V. Based on these results, we have the following observations:

1) The DPHA$_{full}$ model outperforms all the ablation models in all cases, demonstrating the rationality and excellent performance of the combination of different components in our complete DPHANet framework.
2) All ablation models except DPHA$_{base}$ have achieved considerable improvements over the DPHA$_{base}$ model, demonstrating the excellent performance of our various components.
3) For the DPAE module, we can find that DPHA$_{full}$ gains about 2% performance improvement over Single-stream TRM on the Charades-STA dataset, and up to about 3% on the TACoS dataset. In addition, TRM + CMHA and TRM + DMHA both outperform Single-stream TRM and Two-stream TRM. The main reason is that the DMHA and CMHA sub-modules can respectively capture and encode the contextual semantic discriminative information and temporal causal interaction information better compared to previous classical encoder modules, thus improving both representation quality and localization performance. The above experimental results demonstrate that the introduction of semantic discriminative information and temporal causal interaction information and the two-stream structure of DPAE are effective.
4) For the VQHA module, we have similar observations to those on the DPAE module. Specifically, for hierarchical mechanism, we can find that CQA + Hierarchical outperforms CQA + w/o.Hierarchical, while DPHA$_{full}$ outperforms w/o.Hierarchical. Both of them demonstrate the effectiveness of hierarchical mechanism, i.e., adding brings performance improvement and removing reduces performance. Similarly, for the combination of coarse-grained and fine-grained interactions, we can find that

w/o. Hierarchical outperforms CQA + w/o.Hierarchical, while DPHA$_{full}$ outperforms CQA + Hierarchical. Both of them demonstrate the effectiveness of the combination of coarse-grained and fine-grained interactions. The main reason is that the VQHA module can hierarchically perform cross-modal interaction and intra-modal self-correlation modeling, and combine both coarse-grained and fine-grained interactions, to obtain high-quality fusion representations. Furthermore, compared with w/o. MultiResTriAttn, DPHA$_{full}$ gains about 1% performance improvement, proving the effectiveness of our proposed multi-residual trilinear attention.

5) For the deviation loss $\mathcal{L}_{dev}$, its observations are similar to those on the DPAE module and VQHA module, demonstrating that $\mathcal{L}_{dev}$ can capture the correlation of causal inference between the start and end boundaries to enhance localization performance. More specifically, to further demonstrate the role of $\mathcal{L}_{dev}$, we analyzed the impact of different values of ω on the localization results in subsection IV-D. From the results, we can find that the introduction of $\mathcal{L}_{dev}$ has a certain improvement on the localization performance, especially at $\omega = 0.001$, where the localization achieves the best results.

Finally, to show the efficiency and performance of our method in practical scenarios, we further recorded the model parameter size and decoding time for our method and baseline. Specifically, in terms of the model parameter size, our proposed DPHA$_{full}$ model is 4.8M, 5.8M and 18.5M on the Charades-STA, TACoS and ActivityNet Captions datasets, while DPHA$_{base}$ model is 4.0M, 5.0M and 17.7M, respectively. Moreover, in terms of the decoding time, our proposed DPHA$_{full}$ model is 21.7s, 28.2s and 69.6s on the Charades-STA, TACoS and ActivityNet Captions datasets, while DPHA$_{base}$ model is 18.5s, 20.5s and 51.5s, respectively in the same computing environment. From the results, we can find that our proposed method is efficient, i.e., the model parameter size is low and the decoding time is fast, which is comparable to the baseline. In addition, we can observe from Table V that our proposed method obtains a significant performance improvement with similar efficiency as the baseline, which further proves its superiority.

D. Hyperparameter Analysis

In this subsection, we analyze the impact of different hyperparameters on the NLVL performance. For the hyperparameter ω, which balances the role of $\mathcal{L}_{dev}$ at the overall objective function, we conducted the experiments on the Charades-STA dataset by gradually adjusting its value from 0 to 1. The results are shown in Fig. 6, from the results, we can find that the introduction of $\mathcal{L}_{dev}$ has a certain boosting effect on the NLVL performance, especially at $\omega = 0.001$, where the localization achieves the best results. In addition, it is worth noting that there is a balance, i.e., the performance decreases when the value of ω is too large.

Similarly, for the hyperparameters λ_{s1} and λ_{s2}, which balance the contributions of two different streams at DPAE module, we conducted the experiments on the Charades-STA dataset by gradually adjusting their values from 0.1 to 1. The results are shown in Fig. 7, which includes two metrics "Rank@1, IoU=0.7" and "mIoU". From the results, we can find that the localization performance is insensitive to these two parameters, demonstrating the stability of our model. In addition, we can find that our method achieves the best results when $\lambda_{s1} = 0.7$ and $\lambda_{s2} = 0.9$. Therefore, for convenience, we set the values of these two parameters to 0.7 and 0.9 for all experiments, respectively.

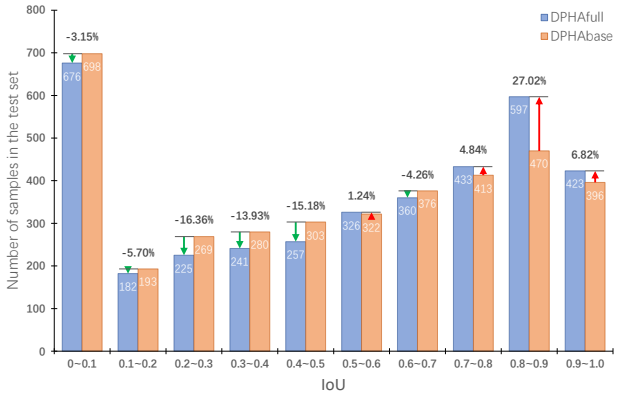


Fig. 4. Histogram of the number of localization results on the test set under different IoUs, on the Charades-STA dataset.

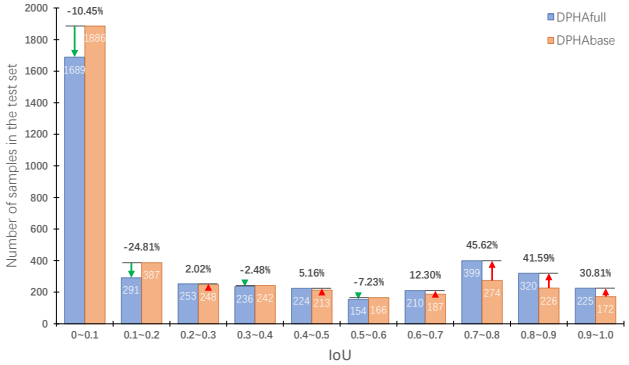


Fig. 5. Histogram of the number of localization results on the test set under different IoUs, on the TACoS dataset.

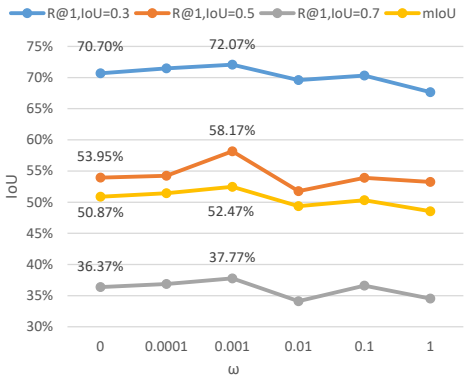


Fig. 6. The "Rank@1, IoU=μ" and "mIoU" results of our proposed method with different ω on the Charades-STA dataset.

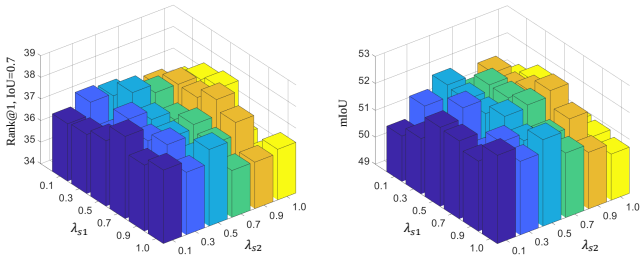


Fig. 7. The "Rank@1, IoU=0.7" and "mIoU" results of our proposed method with different λ_{s1} and λ_{s2} on the Charades-STA dataset.

E. Qualitative Results

To further demonstrate the effectiveness of our proposed DPHANet model and to better understand it, we visualized the experimental results of the NLVL process, IoU histograms, and multi-residual trilinear attention scores on the benchmark datasets, and performed the corresponding qualitative analysis.

First, we visualized four samples of the NLVL process of our proposed model in Fig. 3, which contains two ablation models, DPHA$_{base}$ and DPHA$_{full}$, to demonstrate the effectiveness of our proposed module. Where the blue and pink curves represent the probability distributions of the start and end moment boundaries, respectively. From Fig. 3, we can observe that the probability distributions of the boundaries coincide well with ground-true, i.e., the probability distributions are all clustered in the vicinity of the ground-true boundaries, while the probability distributions at non-target moments are all extremely small. More specifically, for the first video-query pair, our proposed model understands the semantics of the query and the video well and returns the precise moment corresponding to the process of a man going from sitting to a chair and then leaving, based on the focus on discriminative information (i.e., objects: man and chair, action: sit). To some extent, it can demonstrate that our proposed DPAE module allows the model to focus on moments in the video that contain discriminative information and ignore irrelevant and redundant information, thus improving the performance in NLVL tasks.

Furthermore, to demonstrate the performance of our proposed method on high-precision localization, we plotted the IoU histograms of the predicted results on the Charades-STA and TACoS datasets, which contain two ablation models, DPHA$_{base}$ and DPHA$_{full}$, as shown in Fig. 4 and Fig. 5, respectively. From the results, we can observe that our proposed method is more focused on high-precision localization than baseline, i.e., it has more samples at high IoU, which highlights the effectiveness of our model in high-precision NLVL. In addition, we can observe that there are many test samples with IoU scores between 0 and 0.1. The main reason is that there are some high difficulty samples in the datasets. Specifically, some of them have long total video length and short target moment length, some of them have a lot of noise, and some of them have low quality text query, which make most of the methods perform poorly for these samples.

In addition, VQHA is a very critical component of our model which learns high-quality fusion representations. Therefore, in order to verify its effectiveness and analyze how it works, we also visualized the proposed multi-residual trilinear attention scores in VQHA through a heat map, and visualized the trilinear attention scores in CQA as a comparison, as shown in Fig. 8. Specifically, each small square in the heat map represents the attention score between the corresponding video frame and word, whereas darker colors represent higher attention scores, i.e., more relevant. Our model also pays more attention to those parts with high attention scores, because they are rich in cross-modal similarity information and are often where the potential boundaries are located. As shown in Fig. 8, we can observe that "takes" and "pillow" receive high attention scores in our model because the VQHA module focuses on those objects or actions that are critical for NLVL. Meanwhile, "takes" and "pillow" have higher attention scores near the target moment than at non-target moments, demonstrating that our model can perform well with the bidirectional understanding of text and video modalities and capture the cross-modal similarity between them. It is worth noting that "person" and "bed" are present in most moments of the video, but they both receive a low attention score. The reason is that they are not discriminative information and have less impact on localization, so our model classifies them as redundant information and ignores their role in achieving more accurate and efficient NLVL. However, for trilinear attention, we can observe that it incorrectly focuses on the redundant information "bed" and neglects the discriminative information "pillow", which leads to a lower localization performance.

F. Failure Cases Analysis

In this subsection, we show the representative failure cases of our proposed DPHANet in Fig. 9, and further analyze its regarding limitations. For the first case shown in Fig. 9, we can find that our method cannot accurately localize the start boundary. It ignores the subtle action of covering the face with the hand during awakening, and focuses on the action of rising up. The main reason is that our model considers the rising up action as discriminative information and over-amplifies the role of the rising up action in the awakening procedure, while ignoring the detail of covering the face with hand. For the second case shown in Fig. 9, we can find that our model cannot correctly identify the accurate start point of "washing the clothes", which incorrectly encompasses the process of a person undressing. One possible reason is that our model focuses on temporal causal interaction information and misinterprets undressing as a necessary antecedent to washing, which ultimately leads to over-interpreted mislocalization. We expect that in future work, we can improve our model to obtain more controllable correlation information attention and interaction capabilities, to address the above limitations and further improve NLVL performance.

V. Conclusion

In this paper, we proposed a novel DPHANet framework for NLVL. Firstly, we highlighted the importance of temporal causal interaction information and contextual semantic discriminative information in the encoding stage for NLVL and correspondingly designed a DPAE module to fully capture

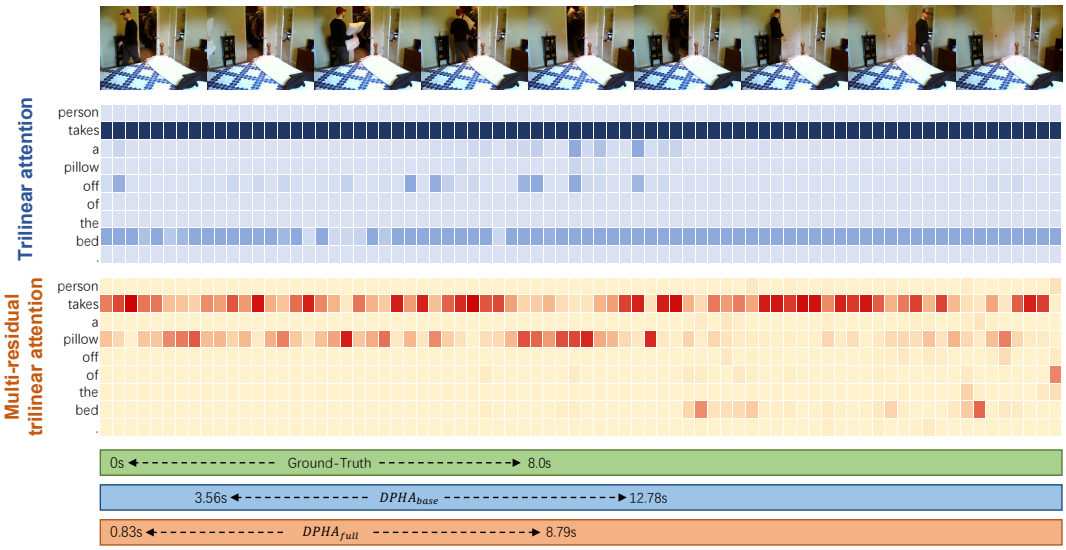


Fig. 8. Visualization results of the proposed multi-residual trilinear attention and the previous trilinear attention.

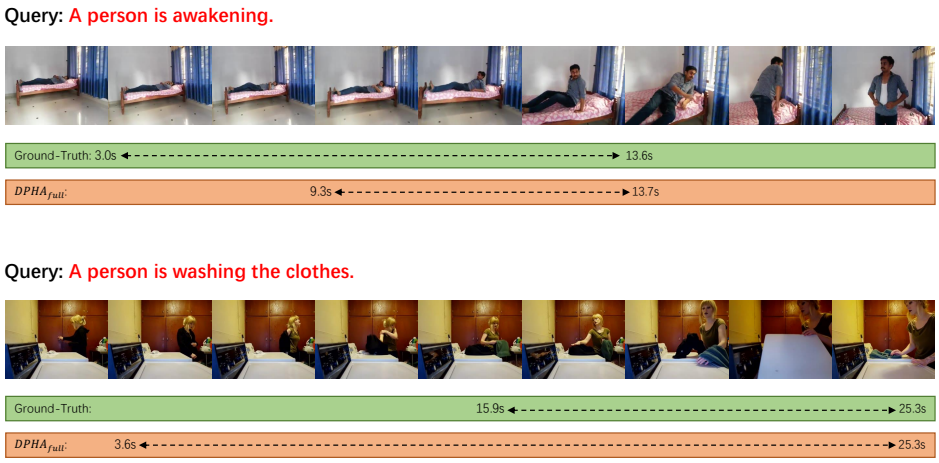


Fig. 9. Visualization results of two failure cases.

and exploit the above critical information. In addition, a VQHA module was proposed to hierarchically perform cross-modal interaction and intra-modal self-correlation modeling, to obtain high-quality cross-modal fusion representations. Furthermore, to overcome the shortcomings of the localization loss, we proposed a novel deviation loss function, to force the model to focus on the continuity and temporal causality in the video. Finally, extensive experiments on three benchmark datasets demonstrated the superiority of our proposed DPHANet model, which outperformed many state-of-the-art methods.

ACKNOWLEDGMENTS

This work is supported in part by National Natural Science Foundation of China (62373112), Guangdong Basic and Applied Basic Research Foundation (nos. 2021A1515011341, 2023A1515012561, 2021A1515011141), and Guangdong Provincial Key Laboratory of Intellectual Property and Big Data under Grant (no. 2018B030322016).

REFERENCES

[1] W. Wang, J. Gao, X. Yang, and C. Xu, "Many hands make light work: Transferring knowledge from auxiliary tasks for video-text retrieval," *IEEE Transactions on Multimedia*, 2022.

[2] L. Wang, M. Zareapoor, J. Yang, and Z. Zheng, "Asymmetric correlation quantization hashing for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 24, pp. 3665–3678, 2021.

[3] J. Wang, B.-K. Bao, and C. Xu, "Dualvgr: A dual-visual graph reasoning unit for video question answering," *IEEE Transactions on Multimedia*, vol. 24, pp. 3369–3380, 2021.

[4] R. Shen, N. Inoue, and K. Shinoda, "Text-guided object detector for multi-modal video question answering," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1032–1042.

[5] H. Tang, J. Zhu, M. Liu, Z. Gao, and Z. Cheng, "Frame-wise cross-modal matching for video moment retrieval," *IEEE Transactions on Multimedia*, vol. 24, pp. 1338–1349, 2021.

[6] Z. Zhang, Z. Zhao, Z. Zhang, Z. Lin, Q. Wang, and R. Hong, "Temporal textual localization in video via adversarial bi-directional interaction networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 3306–3317, 2020.

[7] M. Huang, C. Qing, J. Tan, and X. Xu, "Context-based adaptive multi-modal fusion network for continuous frame-level sentiment prediction,"

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.

[8] J. Gao, Z. Yang, K. Chen, C. Sun, and R. Nevatia, "Turn tap: Temporal unit regression network for temporal action proposals," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3628–3636.

[9] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5267–5275.

[10] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," *arXiv preprint arXiv:2004.13931*, 2020.

[11] H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou, and R. S. M. Goh, "Natural language video localization: A revisit in span-based question answering framework," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 8, pp. 4252–4266, 2021.

[12] X. Yu, M. Malmir, X. He, J. Chen, T. Wang, Y. Wu, Y. Liu, and Y. Liu, "Cross interaction network for natural language guided video moment retrieval," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1860–1864.

[13] S. Ghosh, A. Agarwal, Z. Parekh, and A. Hauptmann, "Excl: Extractive clip localization using natural language descriptions," *arXiv preprint arXiv:1904.02755*, 2019.

[14] D. Lin, S. Fidler, C. Kong, and R. Urtasun, "Visual semantic search: Retrieving videos via complex textual queries," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2657–2664.

[15] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 19–27.

[16] C. Liang, W. Wang, T. Zhou, J. Miao, Y. Luo, and Y. Yang, "Local-global context aware transformer for language-guided video segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[17] B. Xu, X. Shu, J. Zhang, G. Dai, and Y. Song, "Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[18] D. Wu, W. Han, T. Wang, X. Dong, X. Zhang, and J. Shen, "Referring multi-object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 633–14 642.

[19] T. Hui, S. Liu, Z. Ding, S. Huang, G. Li, W. Wang, L. Liu, and J. Han, "Language-aware spatial-temporal collaboration for referring video segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[20] B. Xu and X. Shu, "Pyramid self-attention polymerization learning for semi-supervised skeleton-based action recognition," *arXiv preprint arXiv:2302.02327*, 2023.

[21] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[22] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[23] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Proceedings, Part II 15*. Springer, 2005, pp. 799–804.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[25] A. W. Yu, D. Dohan, Q. Le, T. Luong, R. Zhao, and K. Chen, "Fast and accurate reading comprehension by combining self-attention and convolution," in *International conference on learning representations*, vol. 2, no. 1, 2018.

[26] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[27] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013.

[28] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 510–526.

[29] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *2015 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2015, pp. 961–970.

[30] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9159–9166.

[31] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*. Springer, 2012, pp. 144–157.

[32] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 870–12 877.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] K. Ning, L. Xie, J. Liu, F. Wu, and Q. Tian, "Interaction-integrated network for natural language moment localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 2538–2548, 2021.

[36] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2725–2741, 2022.

[37] J. Wang, L. Ma, and W. Jiang, "Temporally grounding language queries in videos by contextual boundary-aware prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 168–12 175.

[38] J. Gao and C. Xu, "Fast video moment retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1523–1532.

[39] G. Wang, X. Xu, F. Shen, H. Lu, Y. Ji, and H. T. Shen, "Cross-modal dynamic networks for video moment retrieval with text query," *IEEE Transactions on Multimedia*, vol. 24, pp. 1221–1232, 2022.

[40] Q. Zheng, J. Dong, X. Qu, X. Yang, Y. Wang, P. Zhou, B. Liu, and X. Wang, "Progressive localization networks for language-based moment localization," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2, pp. 1–21, 2023.

[41] R. Zeng, H. Xu, W. Huang, P. Chen, M. Tan, and C. Gan, "Dense regression network for video grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 287–10 296.

[42] S. Xiao, L. Chen, S. Zhang, W. Ji, J. Shao, L. Ye, and J. Xiao, "Boundary proposal network for two-stage natural language video localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 2986–2994.

[43] L. Chen, C. Lu, S. Tang, J. Xiao, D. Zhang, C. Tan, and X. Li, "Rethinking the bottom-up framework for query-based video localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 551–10 558.

[44] L. Zhang and R. J. Radke, "Natural language video moment localization through query-controlled temporal convolution," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 682–690.

[45] N. Liu, X. Sun, H. Yu, F. Yao, G. Xu, and K. Fu, "M^{2}dcapsn: Multimodal, multichannel, and dual-step capsule network for natural language moment localization," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[46] X. Sun, H. Wang, and B. He, "Maban: Multi-agent boundary-aware network for natural language moment retrieval," *IEEE Transactions on Image Processing*, vol. 30, pp. 5589–5599, 2021.

[47] Y. Zhang, X. Chen, J. Jia, S. Liu, and K. Ding, "Text-visual prompting for efficient 2d temporal video grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 794–14 804.

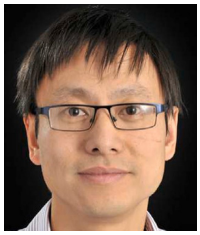
Ruihan Chen received his B.S. degree in the School of Information Engineering at Guangdong University of Technology, in 2021. He is currently pursuing an M.S. degree in the School of Information Engineering at Guangdong University of Technology since 2021. His current research interests include cross-modal retrieval and video understanding.

Junpeng Tan received his M.S. degree in the School of Information Engineering at Guangdong University of Technology, in 2021. He is currently pursuing a Ph.D. degree in the School of Electronics and Information Engineering at the South China University of Technology since 2021. He has published over 20 peer-reviewed conference/journal papers including AAAI, ISBI, ICIP, and T-NNLS, T-AFFC, T-MM, T-ASLP, PR, INS, ESWA, NCAA, etc. His research interests include information retrieval, medical image processing, and generative artificial intelligence.

Zhijing Yang received the B.S and Ph.D. degrees from the Mathematics and Computing Science, Sun Yat-sen University, Guangzhou China, in 2003 and 2008, respectively. He was a Visiting Research Scholar in the School of Computing, Informatics and Media, University of Bradford, U.K, between July-Dec, 2009, and a Research Fellow in the School of Engineering, University of Lincoln, U.K, between Jan. 2011 to Jan. 2013. He is currently a Professor and Vice Dean at the School of Information Engineering, Guangdong University of Technology, China. He has published over 80 peer-reviewed journal and conference papers. His research interests include image retrieval, artificial intelligence-generated content, machine learning, and pattern recognition.

Xiaojun Yang received the B.S. and M.S. degrees from High-Tech Institute, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree from HighTech Institute in 2010 by a coalition form with Tsinghua University, Beijing, China. He is currently an Associate Professor and Hundred Young Talent ProgramTalent Program of Guangdong University of Technology, Guangzhou, China. His research interests include complex systems control, artificial intelligence, and machine learning.

Qingyun Dai received the Ph.D. degree in communication and electronic system from the South China University of Technology, Guangzhou, China, in 2001. She is currently a Professor with Guangdong Polytechnic Normal University, Guangzhou. She is also the Leader of the Guangdong Provincial Key Laboratory of Intellectual Property & Big Data. Her research interests include blockchain and big data security, and the application of intellectual property.



Yongqiang Cheng is currently a Professor with the Faculty of Technology at the University of Sunderland. His research interests include digital healthcare technologies, AI, UAV, control theory and applications, embedded system, secure communication, and data mining.

Liang Lin (Fellow, IEEE) is a full professor at Sun Yat-sen University. From 2008 to 2010, he was a postdoctoral fellow at the University of California, Los Angeles. From 2016–2018, he led the SenseTime R&D teams to develop cutting-edge and deliverable solutions for computer vision, data analysis and mining, and intelligent robotic systems. He has authored and coauthored more than 100 papers in top-tier academic journals and conferences (e.g., 15 papers in TPAMI and IJCV and 60+ papers in CVPR, ICCV, NIPS, and IJCAI). He has served as an associate editor of IEEE Trans. Human-Machine Systems, The Visual Computer, and Neurocomputing and as an area/session chair for numerous conferences, such as CVPR, ICME, ACCV, and ICMR. He was the recipient of the Annual Best Paper Award by Pattern Recognition (Elsevier) in 2018, the Best Paper Diamond Award at IEEE ICME 2017, the Best Paper Runner-Up Award at ACM NPAR 2010, Google Faculty Award in 2012, the Best Student Paper Award at IEEE ICME 2014, and the Hong Kong Scholars Award in 2014. He is a Fellow of IEEE, IAPR, and IET.