EventFormer: A Node-graph Hierarchical Attention Transformer for Action-centric Video Event Prediction

Qile Su*
Beihang University
Beijing, China
zy2423342@buaa.edu.cn

Shoutai Zhu*
Beihang University
Beijing, China
shoutaizhu@buaa.edu.cn

Shuai Zhang
University of Science and Technology Beijing
Beijing, China
u202140073@xs.ustb.edu.cn

Baoyu Liang
Beihang University
Beijing, China
liangbaoyu96@buaa.edu.cn

Chao Tong†
School of Computer Science and Engineering
State Key Laboratory of Virtual Reality Technology and Systems
Beihang University
Beijing, China
tongchao@buaa.edu.cn

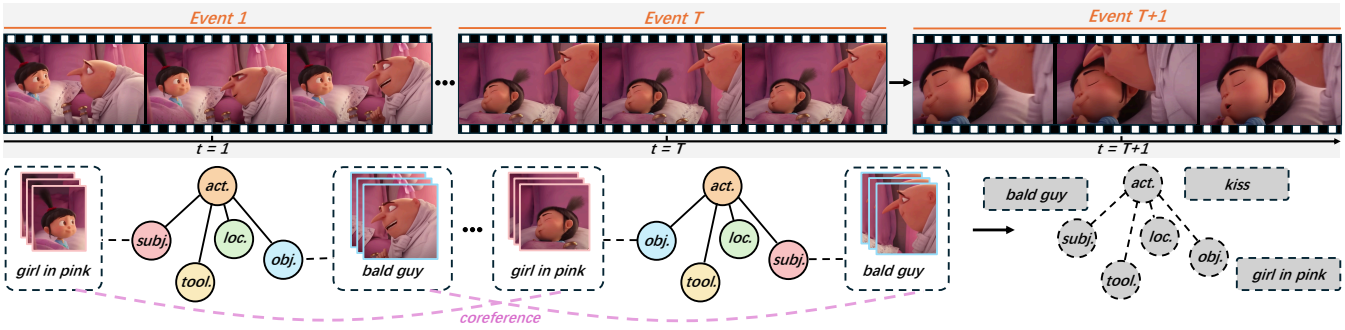

Figure 1: A video event graph chain C is a series of observed historical video events, represented as video event graphs, where each node in the graph contains both visual and textual representations, and the edges denote argument roles. Given a C, the proposed AVEP task requires the model to predict the verb of future events along with their corresponding arguments.

Abstract

Script event induction, which aims to predict the subsequent event based on the context, is a challenging task in NLP, achieving remarkable success in practical applications. However, human events are mostly recorded and presented in the form of videos rather than scripts, yet there is a lack of related research in the realm of vision. To address this problem, we introduce AVEP (Action-centric Video Event Prediction), a task that distinguishes itself from existing video prediction tasks through its incorporation of more complex logic and richer semantic information. We present a large structured dataset, which consists of about 35K annotated videos and more than 178K video clips of event, built upon existing video event datasets to support this task. The dataset offers more fine-grained annotations, where the atomic unit is represented as a multimodal event argument node, providing better structured representations of video events. Due to the complexity of event structures, traditional visual models that take patches or frames as input are not well-suited for AVEP. We propose EventFormer, a node-graph hierarchical attention based video event prediction model, which can capture both the relationships between events and their arguments and the coreferencial relationships between arguments. We conducted experiments using several SOTA video prediction models as well as LVLMs on AVEP, demonstrating both the complexity of the task and the value of the dataset. Our approach outperforms all these video prediction models. We will release the dataset and code for replicating the experiments and annotations.

CCS Concepts

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

*Both authors contributed equally to this research.
†Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '25, Dublin, Ireland.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755556

Keywords

Video Event Prediction, Event Graph, Hierarchical Attention, Computer Vision

ACM Reference Format:

Qile Su, Shoutai Zhu, Shuai Zhang, Baoyu Liang, and Chao Tong. 2025. EventFormer: A Node-graph Hierarchical Attention Transformer for Action-centric Video Event Prediction. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3746027.3755556

1 Introduction

Event prediction has been a key research problem in the field of artificial intelligence [57] for a long time, which can not only be used for security surveillance to anticipate potentially harmful behaviors [39] but also play a significant role in embodied intelligence [32]. The task of event prediction can trace back to Granroth-Wilding and Clark [11], in which it was first formally defined as script event induction.

It is worth noting that in real life and practical applications, human events are often recorded in multimodal forms. However, there is a lack of research on event prediction in the realm of vision. Several similar video prediction tasks have been proposed, such as action anticipation and visual reasoning, while video event prediction differs from these tasks in the following ways: **Richer semantic information** [25, 47]. An event typically comprises subject, action, object, location, and relationship between arguments, with action representing a higher semantic level concept such as "chase" and "discuss" rather than low-level verbs like "run" or "speak". For event participants, the descriptions tend to be more detailed. For instance, a participant may be referred to as "a black-haired woman wearing a blue coat" rather than simply "a woman". **Higher-order Markov chains** [22, 55]. Event prediction can be modeled as a higher-order Markov chain, in which the influence of historical events on future ones does not necessarily diminish with increasing temporal distance. In contrast, action prediction and video frame generation are often highly dependent on adjacent time steps. Therefore, we believe that incorporating event prediction into video models is highly meaningful.

Compared with the existing event prediction tasks in NLP, video event prediction differs in following aspects: **Ambiguous coreference relationship** [25, 48]. In text events, objects are typically referred to by the same noun or pronoun, making them easily identifiable within the context. However, in video events, due to the changes in camera angles and variations in appearance, coreference relationships are often difficult to capture. **More complex logic** [16, 47]. The involvement of different entities in various events makes event prediction tasks non-first-order logical, posing significant challenges to reasoning. For instance, in pairs of events such as ["Mr. Bean chases the thief" , "The thief escapes"] and ["Mr. Bean chases the thief" , "Mr. Bean fails to catch the thief"], the latter can be simplified to ["chase", "fail"], while the non-first-order logic means that the subject of the event may change in chains.

To fill in the gap in event prediction within the field of computer vision, in this paper, we introduce a novel video prediction task, which we term *action-centric video event prediction (AVEP)* as shown in Figure 1. We drew inspiration from the MCNC [11] task in NLP, as well as action anticipation tasks [7, 12] in computer vision, to design the AVEP task in a more scientifically grounded manner. To better address the semantic richness of video events, we propose to represent a video event in the form of a multimodal graph. Instead of following the MCNC task, we no longer provide the model with predefined options but directly predict the events. While this significantly increases the difficulty of the task, it aligns more closely with the original intent of event prediction.

To support the proposed task, we have constructed the AVEP dataset, which contains around 35K videos with 178K video events, based on existing video event datasets including VidSitu[40], Epic-kitchens[7] and VidEvent[25]. For each video, based on the textual descriptions of the events, we identify the event arguments and their corresponding image slices through a novel annotation procedure involving different labeling steps and a validation stage, and represent these in the form of event graphs.

Considering existing visual models are based on frames or patches that are not well suited for the event structure, we propose an approach that adapts to the event graph structure by introducing a node-graph hierarchical attention Transformer to capture the relationships between historical events at the event graph level and the argument node level. This enables the model to analyze the factors that influence future events based on past occurrences, thus facilitating accurate predictions. Furthermore, to address the ambiguity of coreference relationships in video events, we introduce a novel coreference encoding method to represent the same object appearing across different video events. This enables the model to effectively recognize coreferential objects across temporal dimensions. Finally, we pre-train the model using random masking across the entire dataset, followed by post-training for the AVEP task. This training strategy enhances the model's ability to capture and understand the relationships between events.

The contributions of this paper are as follows: 1) We introduce the action-centric video event prediction (AVEP) task, which fills in the gap in event prediction in computer vision. 2) We introduce the AVEP dataset, built upon several existing datasets, which contains approximately 35K videos along with 178K events annotated through a novel procedure to construct event graphs. 3) We propose a node-graph hierarchical attention and co-reference encoding mechanism to address the challenges in video event prediction. 4) Our proposed method outperforms existing SOTA models in related prediction tasks and open-source LVLMs in visual reasoning, establishing a strong baseline for the AVEP task.

2 Related Work

In this section, we will introduce other related work that has been mentioned previously in the following sections.

2.1 Script Induction

The concept of script knowledge in Artificial Intelligence, along with early knowledge-based methods to learn scripts were introduced by Schank and Abelson (1977) [41] and Mooney and DeJong (1985) [29]. One particularly influential work in this area is that of Chambers and Jurafsky (2008) [4], which formulates the script induction problem as learning coherent event chains and predicting

the missing event. Several follow-up works (Chambers and Jurafsky, 2009 [5]; Jans et al., 2012 [17]; Pichotta and Mooney, 2014 [33]; Rudinger et al., 2015 [38]) employ and extend Chambers's methods for learning narrative chains. Considering the uncertainty in the prediction results, Granroth-Wilding and Clark [11] proposed a multiple-choice variation, called MCNC, to simplify the evaluation process. To address these challenges, numerous studies have proposed effective solutions. Seminal works [53, 58] utilize graphs to structurally represent events instead of only texts, and some [53] leverage attention mechanisms to analyze the influencing factors of historical events. However, most of these studies only focus on the text modality and do not extend to the visual modality, while most events appear in form of videos in the real world. Hence, exploring event prediction in the visual domain is highly meaningful.

2.2 Video Action Prediction

In the realm of vision, several similar prediction tasks such as Action Anticipation, VLEP [27], and VidEvent [25] have already been proposed. These tasks are fundamental to many real world applications in area like robots and transportation [21]. While these studies have made significant contributions to the field of video event prediction, there remains ample room for further research and improvement, particularly in terms of modeling event structures and designing more effective task formulations.

Several excellent datasets have been developed to facilitate these video prediction tasks. VidSitu [40] is a large-scale dataset containing diverse videos from movies depicting complex situations (a collection of related events). VLEP [23] is a multimodal commonsense next-event prediction dataset. VidEvent [25] is a dataset designed to advance the understanding of complex event structures in videos. Ego4D [12] contains thousands of hours of egocentric video footage recorded in varied environments. EPIC-KITCHENS [7] is a large-scale dataset focusing on kitchen-based activities. However, although these datasets cover videos from various types of activities, they lack structured annotations describing the events within the videos, making them unsuitable for event prediction tasks. Therefore, a dataset specifically designed for detailed and structural descriptions of video events is needed.

In terms of recent model algorithms, many researchers have explored various models to address these video prediction tasks. Girdhar [9] presents a new model called Anticipative Video Transformer and a self-supervised future prediction loss for action anticipation. And Liang [25] proposes to use two separate Transformers to model video events and textual events respectively. Recently, some [19] start to use large language models to help predict the actions in text modality. Existing prediction methods have demonstrated the effectiveness of the Transformer architecture in predicting action sequences. However, most current approaches are not well suited for the inherent structural nature of events. Therefore, we modify a node-graph hierarchical attention Transformer architecture to incorporate event structure into the model.

2.3 Graph-based Representations

Compared to caption-based anticipation methods [30, 50] that produce unstructured textual descriptions, graph-based representations offer explicit and compositional modeling of event structure and have been widely explored for video understanding [34]. The Action Genome dataset [18] decomposes each action into a spatio-temporal scene graph of objects and their relationships, and the Egocentric Action Scene Graphs (EASG) [36] extends this idea to first-person videos by providing temporally evolving graphs of verbs and interacting objects. Building on these representations, prior methods have applied GNNs and attention mechanisms for fine-grained event reasoning. ViGAT [10] employs Graph Attention Network blocks to model spatial and temporal dependencies among objects, while dynamic scene graph generation frameworks like TEMPURA [6] leverage transformer-based sequence modeling to learn object interactions over time and mitigate long-tail biases in video relation data. However, these works focus on recognizing or detecting structured graphs in observed videos, whereas our approach aims to predict future event graphs by using the node-graph hierarchical attention to capture the relationships between historical events.

3 AVEP Task

In this section we formulate the AVEP task, which aims to predict the next potential event based on a series of observed historical video events. We extensively draw upon the definitions and evaluation frameworks of the natural language event task MCNC [11] and video prediction-related tasks [25, 27] to ultimately define the form of the video event prediction task.

3.1 Formal Definition

We provide the following definition to describe the task:

Definition 1 (Video Event). In the area of NLP, an event is typically defined as a occurrence that involves one or more participants, occurring in a specific time and place. We adopt an NLP-inspired perspective to define a video event. A video event E is characterized as an action or occurrence within a video clip V containing T frames, involving one or more participants. The video clip is treated as an atomic unit—further division would lead to the loss of essential event arguments. We retain the original definition of an event as Chambers [4], in which an event E can be represented as $E = (trg_t, ARG_t)$, where trg_t is a word or span expressing the event and ARG_t denotes the arguments of the event. To incorporate video-based descriptions of events, we extend trg_v and ARG_v using frames or frame slices related with the arguments, f_{trg} and f_{ARG}, to describe the event's trigger and the corresponding arguments, in which $trg_v = (f_{trg0}, f_{trg1}, \ldots, f_{trgn}; trg_t)$ and $ARG_v = (f_{ARG0}, f_{ARG1}, \ldots, f_{ARGn}; ARG_t)$.

Definition 2 (Event Graph). To better represent the structural relationships between event arguments, we introduce a directed graph structure to represent an event, called an event graph $\mathcal{G}_E = (\mathcal{V}, \mathcal{E})$, as shown in Figure 1. In such a graph, the nodes $\mathcal{V} = \{trg_v; ARG_v\}$ represent the event trigger and arguments, and the edges $e = rel(v_i, v_j)$ represent the relationships between these arguments. Here, $v \in \mathcal{V}, e \in \mathcal{E}$, with the function rel is designed to define the argument roles of nodes in an event, such as $subj.$, $obj.$, and $loc.$.

Definition 3 (Event Chain). Followed Granroth-Wilding and Clark [11], in which the observed historical event sequence is represented in the form of an event chain $C = (E_1, E_2, \ldots, E_n)$, we

define an event chain through a sequence of event graphs $C = (\mathcal{G}_{E_1}, \mathcal{G}_{E_2}, \ldots, \mathcal{G}_{E_n})$.

3.2 Action-centric Video Event Prediction Task Definition

As shown in Figure 1, its formalized expression is as follows: Given an event graph chain $C = (\mathcal{G}_{E_1}, \mathcal{G}_{E_2}, \ldots, \mathcal{G}_{E_n})$ representing a sequence of observed historical events, AVEP requires the model to predict trg_t and ARG_t within the possible future event. Each event graph chain C consists of t event graphs $\mathcal{G}_{E_i} = (\mathcal{V}_i, \mathcal{E}_i)$, where each event graph is composed of m multimodal argument nodes $v_{ij} \in \mathcal{V}_i$ and their corresponding edges $e_{ij} \in \mathcal{E}_i$. To more accurately evaluate the model's performance in real-world applications, we no longer provide five candidate future events as in the MCNC task. Instead, following the design of the action anticipation task, we require the model to directly predict the future event's trg_t from the verb vocabulary of the entire dataset and further predict the arguments ARG_t associated with trg_t.

4 Dataset

We expand existing event-level video datasets to construct AVEP dataset, a large-scale dataset for video event prediction, which structurally represents video events using event graphs. Moreover, we divide the AVEP into first-person perspective and third-person perspective to cater to the needs of different downstream tasks and real-world applications.

The proposed dataset meets the requirements of video event prediction tasks in terms of task support, semantic richness, structural clarity, and practical applicability. We further describe the detailed data and data analysis, along with the overview of the annotation process which involves different labeling steps and a validation stage.

We also provide a detailed description of all the datasets utilized in the expansion process to construct the AVEP dataset in the Appendix.

4.1 Construction Procedure

Our dataset construction process consists of three main stages: data collection, integration and expansion, and manual verification. First, we conducted a comprehensive survey of existing event-related video datasets, evaluating them based on their structural support for event representation and their suitability for event prediction tasks. Based on this evaluation, we selected appropriate datasets as the foundation for constructing the AVEP dataset.

In the integration and expansion stage, we standardized all textual annotations across datasets into a unified event-structured format. Additionally, we developed an automatic argument annotation tool based on the GDINO [28] model to generate bounding box annotations for event arguments in videos, followed by cropping the corresponding regions.

Finally, during the data verification stage, three annotators reviewed and refined the automatically generated argument annotations to ensure accuracy and consistency. A detailed description of the dataset construction process is provided in the Appendix.

4.2 Data Statistics

AVEP dataset contains more than 178K video event graphs of complex structures, rich hierarchy, and logical evolutionary. We provide statistics on the splits of the dataset in Appendix. To validate these characteristics, we performed a detailed statistical analysis of the dataset from multiple perspectives.

In our dataset, event chains range in length from 3 to 15, which is sufficient to cover most event reasoning scenarios. Furthermore, each event involves 2 to 5 arguments, with an average of 2.8 arguments per event, aligning well with the definitions used in previous studies.

The dataset comprises a total of 2284 unique verbs. Our analysis of the verb distribution reveals a relatively balanced spread. Notably, the top 30 most frequently occurring verbs collectively constitute less than 40% of the entire set, with no single verb overwhelmingly dominating the distribution. This level of lexical diversity ensures comprehensive coverage of the vast majority of events encountered in real-world applications.

To evaluate whether the dataset sufficiently captures the rich semantics of events, we analyzed the distribution of nouns in the annotations of *ARGs*. The results indicate that the dataset contains over 6,000 unique nouns. Excluding commonly used terms such as "man" and "woman", which frequently appear in annotations referring to people, we found that the remaining nouns are relatively evenly distributed. This balanced distribution ensures that our dataset aligns with the semantic richness observed in real-world applications.

We provide a detailed dataset statistics in the Appendix.

5 Method

In this section, we will provide a detailed introduction of the EventFormer model designed for video event prediction, along with the baseline models we developed for the AVEP task based on this framework.

5.1 EventFormer

In a standard Transformer model, the basic unit of input is the tokens, and all tokens operate at the same hierarchical level within the architecture. For the AVEP task, the model needs to attend not only to the key events in the historical event chain, but also to the key arguments within historical events. However, taking either node or graph embeddings as tokens, this single-level structure fails to capture the complex relationships between nodes and graphs at the same time. To address this limitation, we propose EventFormer, a temporal prediction framework incorporating a node-graph hierarchical attention mechanism tailored for graph-structured data, as shown in Figure 2.

5.1.1 Overview. We basically follow the architecture of Transformer [45] to construct EventFormer, while extending a novel **Node-graph Hierarchical Attention** mechanism to adapt to the complex structure of events. The EventFormer is composed of a stack of N layers, in which each layer has two sub-layers, taking an event graph chain C as input. The first is a Node-graph Hierarchical Attention mechanism module network, and the second is a position-wise fully connected feed-forward network. Following the

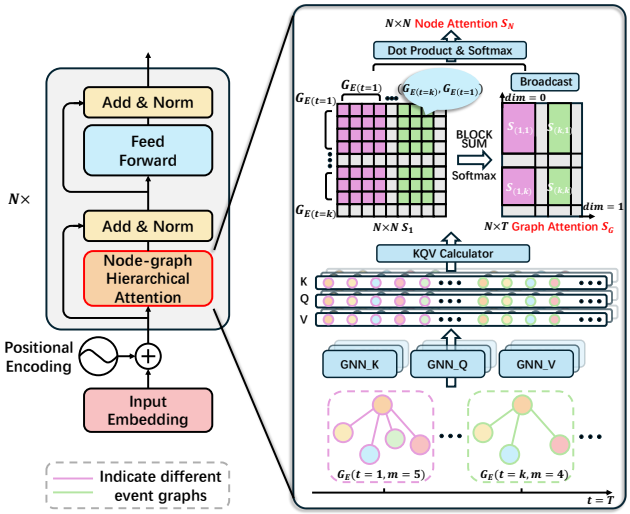

Figure 2: The architecture of EventFormer. On the left is the overall model framework, which follows the same architecture as Transformer. On the right is a magnified view of Node-graph Hierarchical Attention, providing detailed explanations of its implementation. We include calculation examples for these components in the Appendix.

Transformer encoder architecture, we similarly employ a residual connection [15] around each of the two sub-layers, followed by layer normalization. To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of dimension d.

5.1.2 Node-graph Hierarchical Attention. To better accommodate the structure of events, we introduce the node-graph hierarchical attention mechanism, as shown in Figure 2. This mechanism enables the model to capture the attention at both the node level and the graph level, including node-to-node interactions within the same graph, node-to-node interactions across different graphs, and the interactions between different graphs.

We treat an event graph G_E as a token, which is composed of m units of arguments, while applying several **graph neural networks**(GNN) to generate the KQV values instead of linear layers. This allows the model to build a more comprehensive event representation by capturing both internal argument interactions and the relationships between them.

$$K = GNN_K(G_E), Q = GNN_Q(G_E), V = GNN_V(G_E), KQV \in R^{N\times d}, \tag{1}$$

where N represents the number of all nodes in C and d denotes the feature dimension. The GNN can be substituted with various basic graph neural network models [20, 46, 52], and we will compare the performance of these models in the Experiment Section.

With the KQV generated by multi-head GNNs, we apply the same attention score method at the node level, as follows:

$$S_1 = \text{softmax}(\frac{QK^T}{\sqrt{d}}), S_1 \in R^{N\times N}. \tag{2}$$

We further partition the S_1 according to the sequence of graphs in C, and assemble them into blocks to represent the node level attention scores between each graph pairs $(G_{E(t=i)}, G_{E(t=j)})$, which indicates the node attention scores of $G_{E(t=j)}$ towards $G_{E(t=i)}$. Then we apply a BlockSum function on the blocks to generate a $N \times T$ S_G, representing the Graph Attention between graph pairs.

$$S_G = \text{softmax}(\text{BlockSum}(S_1); dim = 1), S_G \in R^{N\times T}. \tag{3}$$

After that, we use S_G to compute Node Attention S_N as follows:

$$S_N = \text{softmax}(S_1 \cdot \text{Broadcast}(S_G)), S_N \in R^{N\times N}, \tag{4}$$

in which we employ a broadcasting mechanism to expand S_G to the same dimension as S_1 based on the number of nodes in each graph, enabling element-wise multiplication. With the attention score S_N and the calculated V, we update the node features $\mathcal{F}$ of each event graph $G_{E(t)}$ in chain C with the latent representation generated by the Node-graph Hierarchical Attention module,

$$\mathcal{F} = S_N \times V, \mathcal{F} \rightarrow C. \tag{5}$$

5.2 Video Event Prediction Model

In this section, we elaborate on how EventFormer is applied to the video event prediction task, where the model predicts future event graphs $\hat{G_E}$ based on the input event graph chain C. Additionally, we introduce a novel coreference encoding mechanism to address the ambiguity in coreference relationships within video events.

5.2.1 Overview. As shown in Figure 3, the Video Event Prediction Model is composed of four main components: Video and Text Embedding, Coreference Encoding, EventFormer and Prediction Head. We utilize a pre-trained CLIP [35] model as an embedding model and keep its parameters frozen during training. For each event graph G_{Ei}, we utilize the CLIP model to generate textual features F_T and video features F_V for each node $v_{ij} = (f_0, f_1, \ldots, f_n; trg/ARG)$, in which we apply pooling to the extracted frame representations to generate the video features. We represent the features of nodes F_{ij} in event graph by concatenating the textual and video features:

$$F_{ij} = \text{Concat}(F_{Tij}, F_{Vij}), \tag{6}$$

in which i denotes the graph number and j denotes the node number. Then we introduce a coreference encoding mechanism to capture coreferential relationships within C, enhancing the model's ability to distinguish coreferential nodes. This mechanism CE will be explained in detail in the next subsection:

$$F'_{ij} = \text{CE}(F_{ij}) + F_{ij}. \tag{7}$$

At this point, we obtain the multimodal features for each node in the video event chain. The C are then input into the EventFormer model, yielding a hidden representation C' of the event chain in the feature space. Using this hidden representation, we predict future event $\hat{G_E}$ through a prediction head composed of multiple fully connected layers. As shown in Figure 3, to enable event prediction, we apply a learnable graph mask to conceal future events within the event chain as G_{masked}. For prediction, the prediction head of model Head utilizes the node features of the masked graph to generate the logits of verb $logits_{verb}$ and the predicted noun embeddings E_{noun},

$$logits_{verb}, E_{noun} = \text{Head}(G_{masked}). \tag{8}$$

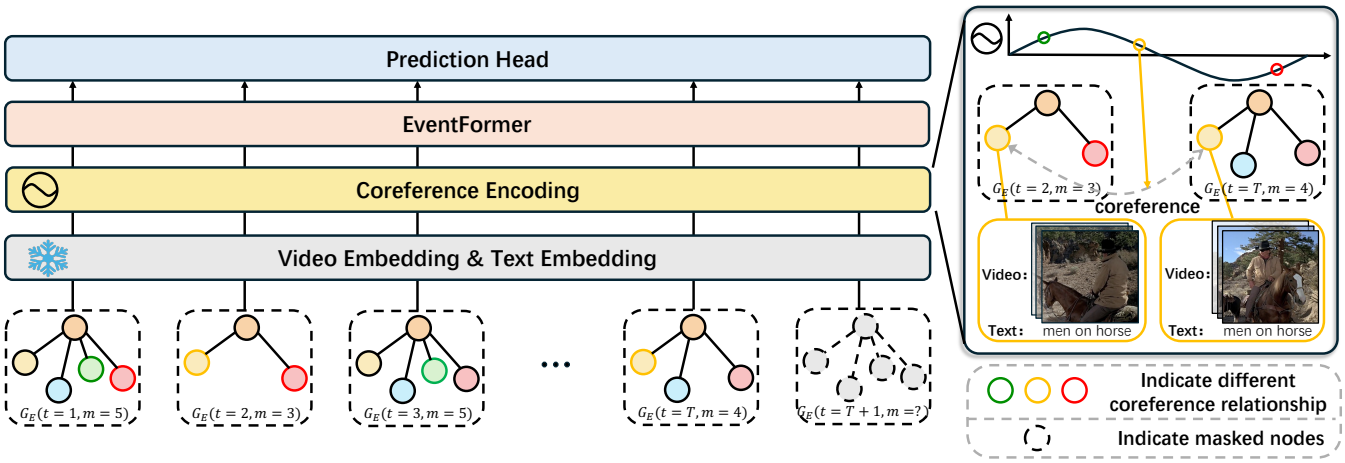

Figure 3: The architecture of Video Event Prediction Model based on EventFormer. The left side of the image shows the overall framework of the model. The frozen symbols indicate that the model's parameters are frozen during training. The right side of the image provides an detailed view of the Coreference Encoding module. The same colored borders represent coreference relationships, while the dashed borders indicate that future events are replaced by the learned mask.

5.2.2 Coreference Encoding. Event graph nodes with coreferential relationships typically share identical textual annotations, yet they may exhibit significant differences in the visual modality. For instance, as shown in Figure 3, the same "man on a horse" appears in two different event graphs. However, due to variations in scene and perspective, it is not immediately evident from the video alone that they refer to the same individual.

Therefore, we introduce a novel coreference encoding mechanism to equip the model with the ability to capture the coreference relationships between nodes. Before feeding the video event chain C into the EventFormer model, we incorporate coreference encodings into embeddings of nodes with coreferential relationships. The corefernece encodings have the same dimension d as the node embeddings, so that the two can be summed. There are many choices of coreference encodings, learned and fixed [8, 43, 45]. In this work, we use sine and cosine functions of different frequencies:

$$\mathrm{CE}_{(index,2k)} = s \cdot sin(index/10000^{2k/d}), \tag{9}$$

$$\mathrm{CE}_{(index,2k+1)} = s \cdot cos(index/10000^{2k/d}), \tag{10}$$

in which $index$ denotes the index of $ARGs$ in the video event chain C, k denotes the embedding dimension and s denotes the scaling factor of the coreference encoding. By directly adding the coreference encodings to node embeddings, we obtain a video event chain that incorporates coreferential relationships. We demonstrate the effectiveness of the coreference encoding through experiments, as shown in the Table 3 in the Experimental Section.

5.3 Training

This section describes the training regime for our models.

Loss Function. To predict verbs and nouns in future video events, we employ a combination of weighted cross-entropy loss, focal loss, and mean squared error (MSE) loss, denoted as $\mathcal{L}_{ce}$, $\mathcal{L}_{focal}$, and $\mathcal{L}_{mse}$, respectively. For verb prediction, we utilize cross-entropy loss, formulated as:

$$\mathcal{L}_{ce} = \mathrm{CrossEntropy}(logits_{verb}, gt), \tag{11}$$

where gt represents the ground truth labels. Given that the target nouns to be predicted have appeared in past events, we leverage the similarity score s between the predicted noun embeddings and historical noun embeddings to identify noun nodes in future events. This similarity-based approach is incorporated into the focal loss to enhance noun prediction:

$$s = \mathrm{CosineSimilarity}(E_{noun}, ARGs), \tag{12}$$

$$\mathcal{L}_{focal} = \mathrm{FocalLoss}(s, gt). \tag{13}$$

Furthermore, to ensure that the predicted noun embeddings closely approximate the ground truth noun embeddings, we introduce an additional MSE loss term. Finally, the overall objective function is defined as a weighted sum of the three loss components:

$$\mathcal{L} = \alpha\mathcal{L}_{ce} + \beta\mathcal{L}_{focal} + \gamma\mathcal{L}_{mse}. \tag{14}$$

Two-stage Training. We adopt a two-stage training strategy to train our Eventformer on AVEP task. We first apply a random masking approach to the input event graph sequence, randomly masking certain event graphs and training the model to reconstruct them during the pre-training phase. This pre-training strategy enables the model to better capture relationships between events and enhances its reasoning ability for event inference. Then we proceed with a post-training phase, where the pre-trained model undergoes a few additional training epochs on the specific task. In this phase, the masking is fixed on the last event graph in the sequence, allowing the model to be fine-tuned on AVEP. Table 4 in Experimental demonstrates that this training strategy effectively improves the model's performance on AVEP task.

Table 1: The comparative experiment with other SOTA models.

Perspective	Set	Method	Parameter	Verb		Noun			Verb-Noun		
				Top1	Top5	P	R	F1	P	R	F1
third person	val	VidEvent	44.1M	18.97	**45.61**	32.62	34.39	33.48	6.19	6.45	6.32
		LLaVA-Video	7B	5.67	21.32	44.50	41.95	43.19	3.13	2.95	3.04
		Qwen2.5-VL	72B	2.05	16.82	**52.74**	**51.71**	**52.22**	1.43	1.40	1.41
		Ours	**7.3M**	**22.32**	45.16	33.10	44.90	38.11	**7.67**	**10.88**	**9.00**
		Human*	-	37.42	60.65	61.05	55.18	57.97	25.09	21.04	22.89
	test	VidEvent	44.1M	18.44	43.96	36.49	37.06	36.77	5.85	5.91	5.88
		LLaVA-Video	7B	6.27	10.30	44.96	41.64	43.24	3.57	3.30	3.43
		Qwen2.5-VL	72B	2.00	16.23	**52.23**	**54.31**	**53.25**	1.21	1.26	1.23
		Ours	**7.3M**	**22.71**	**45.21**	42.93	50.11	46.24	**7.67**	**7.71**	**7.69**
		Human*	-	38.46	63.08	64.80	52.22	57.83	27.15	22.22	24.44
first person	test	InAViT	157.2M	29.78	64.00	-	26.67	-	-	**8.69**	-
		LLaVA-Video	7B	3.49	5.23	**25.42**	45.50	32.62	1.40	2.50	1.79
		Qwen2.5-VL	72B	1.16	25.34	23.56	**55.66**	**33.11**	0.46	1.08	0.65
		Ours	**7.3M**	**29.94**	**65.25**	18.88	23.35	20.88	**6.43**	7.67	**7.00**

6 Experiments

In this section, we employ SOTA models [25, 37] from existing prediction tasks, along with LVLMs[2, 26], to conduct a series of comparative experiments, validating the complexity of our proposed task and the significance of our dataset. All LVLMs used in our experiments were evaluated in a zero-shot setting. In addition, we design both comparative and ablation studies to assess the effectiveness of our proposed approach.

6.1 Setups

We use CLIP as our node feature embedding model, which is frozen when training. To assess the impact of different GNN architectures on performance, we performed experiments on GCN, GAT, and GIN implementations from the Deep Graph Library (DGL) [49]. The dimension of hidden feature d in our model is 1024. Based on the experimental results, we set the loss weights $[\alpha, \beta, \gamma]$ to $[1.0, 1.0, 0.5]$. Our model is optimized by Adam, and the learning rate and weight decay are $1e-5$ and $1e-6$. Setting the maximum training step to 300, we select the performance of the best epoch as our final results. Dropout ratio and batch size are 0.3 and 64. The maximum sequence length is 50. Training is conducted on 2×3090 GPUs and takes $\sim$ 3 days.

We measure future event verb prediction using the Acc@K. For noun argument prediction in future events, we evaluate performance using recall, precision, and F1-score metrics. We also introduce a verb-noun metrics to to assess the model's predictive capability in future event prediction, where the verb-noun recall, precision and F1-score represent the recall, precision and F1-score of noun prediction, conditioned on the correct prediction of the verb.

6.2 Comparative Experiments

Considering the inherent ambiguity in AVEP, we provide human evaluation results as a credible reference point for assessing model performance. To this end, we simulate human performance on the task, offering a benchmark that reflects the upper bound of predictive accuracy. Further details are in the Appendix.

Comprehensive details of the baseline models used for comparison are provided in the Appendix. As shown in Table 1, for verb prediction, our proposed model significantly outperforms all other models in terms of ACC@1 by **4.27**, and achieves comparable performance to the best-performing model on ACC@5. Notably, the LVLM demonstrates poor performance in verb prediction, suggesting that current LVLMs still lack reliable capabilities for video-based event reasoning. For the prediction of noun arguments, our method outperforms *VidEvent* by achieving improvements of **13.05** in Recall, **6.44** in Precision and **9.47** in F1-score. Interestingly, LVLMs exhibit strong performance in noun prediction. We attribute this to their robust visual recognition capabilities, which allow them to accurately identify and associate characters and objects appearing in the video. Compared to other metrics, the Verb-Noun metric provides a more comprehensive reflection of a model's capability in future event prediction. Our proposed method achieves the best results on all three metrics across both the validation and test sets.

We also evaluated our model's performance on first-person AVEP. While our method still outperforms existing SOTA models in verb prediction, it does not surpass them in noun or verb-noun prediction. This is primarily because, to ensure a fair and realistic comparison, we strictly followed the input format of the original InAViT model, which includes the initial frames of the future events. As a result, the input frames may already contain visual clues about the upcoming nouns, making the noun prediction task easier.

6.3 Ablation Experiments

We conducted a series of ablation studies to systematically evaluate the effectiveness of the proposed method.

6.3.1 Ablation on GNN. We conducted experiments on the third-person AVEP using different GNN architectures and linear layers. As shown in Table 2, the linear layer performed the worst across all metrics, indicating that incorporating GNNs can effectively enhance

Table 2: The ablation experiment on different GNNs.

Perspective	Set	Method	Verb		Noun			Verb-Noun		
			Top1	Top5	P	R	F1	P	R	F1
third person	val	EventFormer-Linear	11.38	30.21	30.15	41.1	34.78	4.02	6.34	4.92
		EventFormer-GCN	14.96	38.99	29.41	41.72	34.50	4.56	6.89	5.49
		EventFormer-GAT	16.00	39.58	31.09	**47.74**	**37.66**	4.13	**10.13**	5.87
		EventFormer-GIN	**19.20**	**41.67**	**32.95**	42.13	36.98	**6.71**	9.93	**8.01**
	test	EventFormer-Linear	15.31	36.35	38.53	39.23	38.88	3.92	4.22	4.06
		EventFormer-GCN	14.97	39.79	36.34	**49.86**	42.04	4.47	5.02	4.73
		EventFormer-GAT	13.61	40.76	**42.55**	47.51	**44.89**	4.13	4.67	4.38
		EventFormer-GIN	**19.20**	**42.33**	41.68	48.62	44.88	**6.58**	**9.67**	**7.83**

Table 3: The ablation experiment on Coreference Encoding.

Perspective	Set	Method	Verb		Noun			Verb-Noun		
			Top1	Top5	P	R	F1	P	R	F1
third person	val	Ours-w/ 0.5CE	15.92	40.03	32.30	38.99	35.33	4.62	8.88	6.08
		Oursr-w/o CE	16.52	41.07	31.83	41.17	35.902	5.73	**10.18**	7.33
		Ours-w/ CE	**19.20**	**41.67**	**32.95**	**42.13**	**36.98**	**6.71**	9.93	**8.01**
	test	Ours-w/ 0.5CE	16.02	41.05	34.71	39.44	36.92	4.65	8.63	6.04
		Ours-w/o CE	16.96	**43.23**	40.22	42.53	41.34	6.33	7.62	6.92
		Ours-w/ CE	**19.20**	42.33	**41.68**	**48.62**	**44.88**	**6.58**	**9.67**	**7.83**

Table 4: The ablation experiment on two-stage training.

Perspective	Set	Method	Verb		Noun			Verb-Noun		
			Top1	Top5	R	P	F1	R	P	F1
third person	val	Ours(one-stage train)	19.20	41.67	32.95	42.13	36.98	6.71	9.93	8.01
		Ours(two-stage train)	22.32	45.16	33.10	44.90	38.11	7.67	10.88	9.00
		Improvement	+3.12	+3.49	+0.15	+2.77	+1.13	+0.96	+0.95	+0.99
	test	Ours(one-stage train)	19.20	42.33	41.68	48.62	44.88	6.58	9.67	7.83
		Ours(two-stage train)	22.71	45.21	42.93	50.11	46.24	7.67	7.71	7.69
		Improvement	+3.51	+2.88	+1.25	+1.49	+1.36	+1.09	-1.96	-0.14

token representation in graph-structured inputs. Among the three basic GNN models, GIN achieved the best overall performance on most metrics. In particular, it significantly outperformed GAT and GCN in verb prediction and verb-noun prediction from the event structure. Meanwhile, GAT showed competitive performance in noun prediction, suggesting its relative strength in capturing localized node-level information.

6.3.2 Ablation on Coreference Encoding. We conducted an ablation study on coreference encoding (CE) on the third-person AVEP, as shown in Table 3. Specifically, we evaluated three configurations: (1) without CE (2) with CE using a scaling factor of 0.5, and (3) with a scaling factor of 1.0. Among them, the model with CE at $scale = 0.5$ performed the worst. In contrast, introducing coreference encoding ($scale = 1.0$) led to consistent improvements across all evaluation metrics compared to the no-CE setting. Notably, verb prediction accuracy (ACC@1) improved by 2.68, and the F1-score for verb-noun prediction increased by 0.68. These results indicate that the proposed CE method helps the model better capture the semantic relationships among event arguments, thereby enhancing its ability to predict future events.

6.3.3 Ablation on Two-stage Training. Finally, we conducted an ablation study on the two-stage training strategy, as shown in Table 4. Compared to training the model directly on the AVEP task, pretraining the model with a random masking strategy followed by task-specific post-training led to substantial performance improvements across most evaluation metrics. Although results on the test set show a slight drop in the verb-noun Precision and F1-score under the two-stage training setup, the significant gains in the majority of other metrics still demonstrate that this training strategy effectively enhances the model's capability in future event prediction.

7 Conclusion

We introduce a novel task, Action-centric Video Event Prediction (AVEP), which requires models to predict the most probable future

event trigger and its associated arguments based on a given video event chain. To facilitate research on this task, we construct the AVEP dataset based on several existing video datasets. To address the challenges of this task, we propose a node-graph hierarchical attention Transformer coupled with a coreference encoding mechanism, which jointly captures the structural relationships among events and arguments as well as the coreferential dependencies between arguments across different events. Our method achieves SOTA performance on AVEP task, providing a strong baseline for further research.

Acknowledgments

This study is partially supported by National Natural Science Foundation of China (62176016, 72274127), National Key R&D Program of China (No. 2021YFB2104800), Guizhou Province Science and Technology Project: Research on Q&A Interactive Virtual Digital People for Intelligent Medical Treatment in Information Innovation Environment (supported by Qiankehe[2024] General 058), Capital Health Development Research Project(2022-2-2013), Haidian innovation and translation program from Peking University Third Hospital (HDCXZHKC2023203), and Project: Research on the Decision Support System for Urban and Park Carbon Emissions Empowered by Digital Technology - A Special Study on the Monitoring and Identification of Heavy Truck Beidou Carbon Emission Reductions.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] https://arxiv.org/abs/2502.13923

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.

[4] Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*. 789–797.

[5] Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 602–610.

[6] Jen-Hao Cheng, Vivian Wang, Huayu Wang, Huapeng Zhou, Yi-Hao Peng, Hou-I Liu, Hsiang-Wei Huang, Kuang-Ming Chen, Cheng-Yen Yang, Wenhao Chai, et al. 2025. Tempura: Temporal event masked prediction and understanding for reasoning in action. *arXiv preprint arXiv:2505.01583* (2025).

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2020. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (2020), 4125–4141.

[8] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*. PMLR, 1243–1252.

[9] Rohit Girdhar and Kristen Grauman. 2021. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 13505–13515.

[10] Nikolaos Gkalelis, Dimitrios Daskalakis, and Vasileios Mezaris. 2022. ViGAT: Bottom-up event recognition and explanation in video using factorized graph attention network. *IEEE Access* 10 (2022), 108797–108816.

[11] Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18995–19012.

[13] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19383–19400.

[14] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6047–6056.

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

[16] Lukas Helff, Wolfgang Stammer, Hikaru Shindo, Devendra Singh Dhami, and Kristian Kersting. 2024. V-LoL: A Diagnostic Dataset for Visual Logical Learning. arXiv:2306.07743 [cs.AI] https://arxiv.org/abs/2306.07743

[17] Bram Jans, Steven Bethard, Ivan Vulic, and Marie-Francine Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*. ACL; East Stroudsburg, PA, 336–344.

[18] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10236–10247.

[19] Sanghwan Kim, Daoji Huang, Yongqin Xian, Otmar Hilliges, Luc Van Gool, and Xi Wang. 2024. Palm: Predicting actions through language models. In *European Conference on Computer Vision*. Springer, 140–158.

[20] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[21] Yu Kong and Yun Fu. 2022. Human action recognition and prediction: A survey. *International Journal of Computer Vision* 130, 5 (2022), 1366–1401.

[22] Dao Xuan Ky and Luc Tri Tuyen. 2018. A Higher order Markov model for time series forecasting. *International Journal of Applied Mathematics and Statistics* 57, 3 (2018), 1–18.

[23] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. What is More Likely to Happen Next? Video-and-Language Future Event Prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8769–8784.

[24] Cheng Li and Kris M Kitani. 2013. Model recommendation with virtual probes for egocentric hand detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2624–2631.

[25] Baoyu Liang, Qile Su, Shoutai Zhu, Yuchen Liang, and Chao Tong. 2025. VidEvent: A Large Dataset for Understanding Dynamic Evolution of Events in Videos. arXiv:2506.02448 [cs.CV] https://arxiv.org/abs/2506.02448

[26] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 5971–5984.

[27] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. 2020. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 704–721.

[28] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*. Springer, 38–55.

[29] Raymond J Mooney and Gerald DeJong. 1985. Learning schemata for natural language processing.. In *Ijcai*. 681–687.

[30] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2017. Text-guided attention model for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[31] Katsuyuki Nakamura, Hiroki Ohashi, and Mitsuhiro Okada. 2021. Sensor-Augmented Egocentric-Video Captioning with Dynamic Modal Attention. In *ACM International Conference on Multimedia (MM)*.

[32] Fabio Persia, Daniela D'Auria, and Giovanni Pilato. 2020. Fast learning and prediction of event sequences in a robotic system. In *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*. IEEE, 447–452.

[33] Karl Pichotta and Raymond Mooney. 2014. Statistical script learning with multi-argument events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 220–229.

[34] Haiyi Qiu, Minghe Gao, Long Qian, Kaihang Pan, Qifan Yu, Juncheng Li, Wenjie Wang, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. 2025. STEP: Enhancing

Video-LLMs' Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training. arXiv:2412.00161 [cs.CV] https://arxiv.org/abs/2412.00161
[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.
[36] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. 2024. Action scene graphs for long-form understanding of egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18622–18632.
[37] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. 2024. Interaction region visual transformer for egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 6740–6750.
[38] Rachel Rudinger, Vera Demberg, Ashutosh Modi, Benjamin Van Durme, and Manfred Pinkal. 2015. Learning to predict script events from domain-specific text. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*. 205–210.
[39] Shakila Khan Rumi, Ke Deng, and Flora Dilys Salim. 2018. Crime event prediction with dynamic features. *EPJ Data Science* 7, 1 (2018), 43.
[40] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5589–5600.
[41] Roger C Schank and Robert P Abelson. 1975. Scripts, plans and goals. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence. IJCAI*, Vol. 1.
[42] Michel Silva, Washington Ramos, Joao Ferreira, Felipe Chamone, Mario Campos, and Erickson R Nascimento. 2018. A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2383–2392.
[43] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.
[44] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1207–1216.
[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[46] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. GRAPH ATTENTION NETWORKS. *stat* 1050 (2018), 4.
[47] Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. 2025. Fostering Video Reasoning via Next-Event Prediction. arXiv:2505.22457 [cs.CV] https://arxiv.org/abs/2505.22457
[48] Liming Wang, Shengyu Feng, Xudong Lin, Manling Li, Heng Ji, and Shih-Fu Chang. 2021. Coreference by Appearance: Visually Grounded Event Coreference Resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, Maciej Ogrodniczuk, Sameer Pradhan, Massimo Poesio, Yulia Grishina, and Vincent Ng (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 132–140. doi:10.18653/v1/2021.crac-1.14
[49] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2019. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv preprint arXiv:1909.01315* (2019).
[50] Teng Wang, Jinrui Zhang, Junjie Fei, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, and Shanshan Zhao. 2023. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677* (2023).
[51] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4581–4591.
[52] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).
[53] Shuang Yang, Fali Wang, Daren Zha, Cong Xue, and Zhihao Tang. 2020. NarGNN: Narrative graph neural networks for new script event prediction problem. In *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 481–488.
[54] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. In *CVPR*.
[55] He Zhao and Richard P. Wildes. 2021. Review of Video Predictive Understanding: Early Action Recognition and Future Action Prediction. arXiv:2107.05140 [cs.CV] https://arxiv.org/abs/2107.05140
[56] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. 2019. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. *arXiv preprint arXiv:1712.09374* (2019).
[57] Liang Zhao. 2021. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–37.
[58] Jianming Zheng, Fei Cai, Yanxiang Ling, and Honghui Chen. 2020. Heterogeneous graph neural networks to predict what happen next. In *Proceedings of the 28th International Conference on Computational Linguistics*. 328–338.

A Dataset

A.1 Dataset Composition

Here, we provide a detailed description of all the datasets utilized in the expansion process to construct the AVEP dataset.

EPIC-KITCHENS-100 [7]. EPIC-KITCHENS-100 is a collection of 100 hours, $20M$ frames, $90K$ actions in 700 variable-length videos, capturing long-term unscripted activities in 45 environments, using head-mounted cameras, which initially builed for the STA task.

VidSitu [40]. VidSitu, a large-scale video understanding dataset comprising $29K$ richly annotated 10-second movie clips, structures videos as a series of interconnected events, with each event containing a verb and multiple entities assigned to specific semantic roles. Clips in VidSitu are drawn from a large collection of movies ($\sim 3K$) and have been chosen to be both complex ($\sim$ 4.2 unique verbs within a video) as well as diverse ($\sim$ 200 verbs have more than 100 annotations each).

VidEvent [25]. VidEvent is a large-scale dataset containing over $23,000$ well-labeled events, featuring detailed event structures, broad hierarchies, and logical relations extracted from movie recap videos, which is initially builed for video understanding.

A.2 Construction Procedure

We further describe the detailed overview of the annotation process in this subsection. The entire AVEP dataset construction process can be divided into three main stages: data collection, integration and expansion, and manual verification.

We conducted a comprehensive survey of existing event-related video datasets and categorized them based on first-person and third-person perspectives as shown in Table 5. We evaluated the completeness of event structures based on whether they satisfy the requirements of event triggers and arguments. Datasets lacking annotated actions or containing only a limited set of action categories were considered insufficient for meeting the event trigger criteria. Similarly, datasets that failed to annotate action-related nouns or had a restricted variety of noun annotations were deemed inadequate for fulfilling the event argument requirements. Building upon the datasets that meet the event structure requirements, we further evaluated their suitability for the AVEP task. AVEP requires each sample in the dataset to contain at least three events, with clearly defined event boundaries and explicit temporal relationships between events.

Considering both the dataset's support for the task and its scale, we selected three datasets, VidSitu, VidEvent and Epic-Kichens as the foundation datasets. We then expanded them to construct a dataset that effectively supports the AVEP task following the pipeline detailed in the subsequent sections.

The video datasets are composed of continuous video event segments, which are derived by partitioning the entire video, along

Table 5: A non-exhaustive summary on the related video datasets.

Perspective	Dataset	Event Structure		Event Prediction Surpport
		trigger	arguments	
Third Person	ActivityNet(2015) [3]	✓	✗	✗
	Youtube8M(2016) [1]	✗	✗	✗
	AVA(2018) [14]	✓	✗	✗
	HACS(2019) [56]	✓	✗	✗
	COIN(2019) [44]	✗	✗	✗
	Vatex(2020) [51]	✓	✓	✗
	VidSTG(2020) [54]	✓	✓	✗
	VLEP(2020) [23]	✓	✓	✗
	VidSitu(2021) [40]	✓	✓	✓(only after processing)
	VidEvent(2025) [25]	✓	✓	✓
First Person	EgoAction(2013) [24]	✓	✗	✗
	DoMSEV(2018) [42]	✓	✓	✗
	Epic-Kitchens(2020) [7]	✓	✓	✓(only after processing)
	MMAC Captions(2021) [31]	✓	✓	✗
	Ego4D(2022) [12]	✓	✓	✓(only after processing)
	Ego-Exo4D(2024) [13]	✓	✓	✓(only after processing)

with corresponding textual annotations of event arguments or sentences. As shown in Figure 4, we can divide the entire annotation process into two parts: automatic program annotation and manual review annotation. First, we manually organized the text annotations of all datasets into a format based on event arguments. For the unannotated data with single-sentence descriptions, we utilized an argument segmentation model to automatically extract the event arguments. The event boundaries in the annotations were then used to segment the videos into a series of events. Next, we employ the DGINO model [28] to sequentially label the bounding box corresponding to each argument in every frames of an event video clip, obtaining the visual representation for each argument. After that, three professional annotators conduct a review and correction of the automated bounding box annotations for each argument.

Three annotators spent five months correcting the data, with an average rate of around 600 samples per week. On average, each video clip required approximately 1-2 minutes to complete the full annotations. Additionally, around 35% of the initially generated bounding boxes required manual correction to ensure accuracy. After the correction phase, an additional month was spent to cross-validation through random sampling.

Along with the argument roles, text argument annotations, and video argument annotations , we treat the text and video argument annotations as node information, while the argument roles serve as the edge relationships between the nodes, thus constructing a multimodal graph representation of the events. By arranging all the events in the video sequentially, we finally obtain the video event graph chain.

A.3 Data Statistics

AVEP is a dataset that contains more than 178K video event graphs of complex structures, rich hierarchy, and logical evolutionary. To validate these characteristics, we performed a statistical analysis of the dataset from multiple perspectives, as shown in Figure 5.

Table 6: Statistics on splits of AVEP dataset.

	train	Val	Test	Total
Videos	29793	1326	4145	35264
Events	154022	6615	17771	178208
Tri.&Args.	434758	19985	43382	498125

The dataset comprises a total of 35K video clips, with an average video length of 15 seconds. Each clip contains historical segments averaging 12 seconds in duration, followed by future events. To ensure temporal independence between events, we maintain a minimum gap of 20 frames between the last historical event and subsequent future events, effectively preventing any potential frame leakage in our dataset construction.

The dataset comprises a total of 2, 284 unique verbs. Our analysis of the verb distribution reveals a relatively balanced spread. Notably, the top 30 most frequently occurring verbs collectively constitute less than 40% of the entire set, with no single verb overwhelmingly dominating the distribution. This level of lexical diversity ensures comprehensive coverage of the vast majority of events encountered in real-world applications.

To evaluate whether the dataset sufficiently captures the rich semantics of events, we analyzed the distribution of nouns in the annotations of *Args*. The results indicate that the dataset contains over 6,000 unique nouns. After excluding commonly used terms such as "man" and "woman," which frequently appear in annotations referring to people, we found that the remaining nouns are relatively evenly distributed. This balanced distribution ensures that our dataset aligns with the semantic richness observed in real-world applications. Additionally, we analyzed the total number of *Args* present in each video event graph. The results show that the average of the event graphs contain 2.8 arguments.

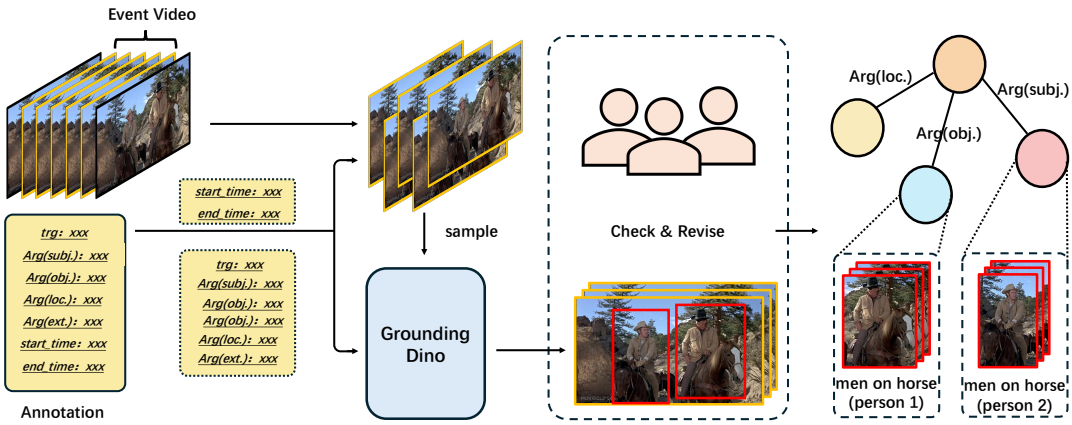

Figure 4: The annotation pipeline of AVEP. We employ the pre-trained Grounding Dino model to localize the corresponding regions in each frame based on textual annotations. Subsequently, three annotators manually verify and refine the bounding boxes according to the textual descriptions. Finally, the selected regions are cropped to generate a sequence of images, which serve as the visual representations of event graph nodes.

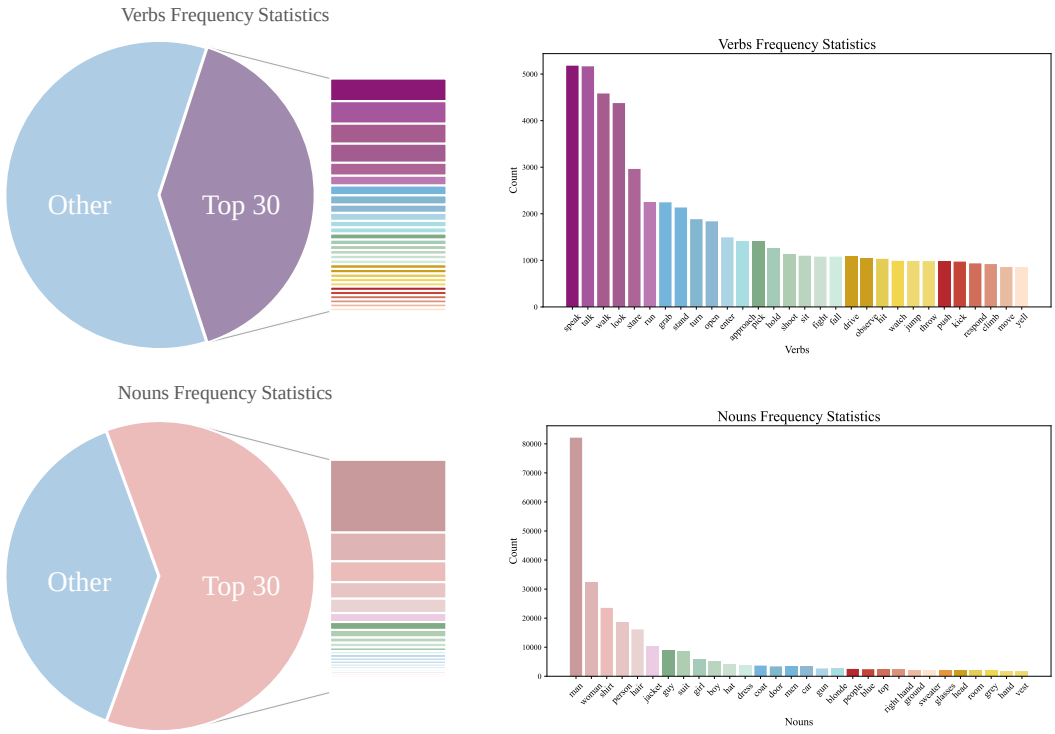

Figure 5: The data statistics on verbs and nouns in the AVEP dataset. Given the large number of verb categories (2, 284 in total), we analyzed the distribution by calculating the proportion of the top 30 most frequent verbs versus all others, and also reported the frequency of each verb within the top 30. A similar analysis was conducted for the nouns in the dataset.

A.4 Impact of Bounding Box Quality

To provide a higher-quality dataset, we manually refined the automatically generated bounding boxes. Although this process reduced noise and improved overall data quality, we remain concerned about how bounding box quality impacts model performance in this task. To investigate this, we conducted experiments using the original, uncorrected bounding boxes. The results are presented in Table 7.

The results show that using the original, uncorrected data indeed leads to a performance drop. Notably, the impact on noun prediction is more significant than that on verbs, which aligns with our intuition. Overall, however, the decline across all evaluation metrics remains under 10%, indicating that the model can still be effectively trained using the original data. This also demonstrates the robustness of our method to noise in the bounding boxes.

B Task Comparison

To better position our task within the landscape of event reasoning and prediction, we compare it with related tasks along several key dimensions, as shown in Table 8:

- **Modelities.** Our task involves both video and text inputs, making it a multimodal task. This contrasts with many existing approaches that focus on either visual or textual input alone.
- **Form of historical context.** nstead of using raw video clips, we employ structured event chains to represent historical context. This structure provides richer and more explicit logical relationships between events, thereby imposing higher demands on the model's reasoning capabilities. Moreover, representing event chains as event graph chains enables finer-grained modeling of internal event structures.
- **Granularity.** Events in our task are defined at a higher semantic level than low-level actions. Unlike simple action units, events encapsulate complex interactions and relationships, making the task more challenging and semantically rich.
- **Task formulation.** Many prior works adopt a multiple-choice format to ensure determinism and simplify evaluation. However, such formats deviate from real-world application scenarios. In contrast, we formulate our task as an open classification problem, and supplement it with human evaluation to provide a more realistic and credible measure of model performance.

C A Node-graph Hierarchical Attention Mechanism Example

To provide a clearer explanation of this attention mechanism, we illustrate the attention computation process using a sequence of three event graphs, each consisting of two nodes, as shown in Figure 6. Different colors are used to represent different event graphs.

In the attention matrix, the red box indicates the attention scores between node pairs across different event graphs. To derive a graph-level attention score, we sum the scores within each matrix block, producing a 6×3 event graph-level attention matrix. By applying a softmax function in the dimension of graphs, we obtain the attention scores on the nodes for each event graph in the sequence, *GRAPH ATTENTION*.

Furthermore, to integrate graph and node attention, we broadcast the computed *GRAPH ATTENTION* back to their original shape 6×6 and perform an element-wise multiplication with the attention matrix. This operation yields the final *NODE ATTENTION*, capturing both intra- and inter-event dependencies effectively.

D Human Evaluation

Considering the inherent ambiguity in the AVEP task, we provide human evaluation results as a credible reference point for assessing model performance. To this end, we developed a dedicated web interface and invited several domain experts to participate in the evaluation. Each participant was asked to watch the video segment corresponding to the historical events of a given sample. To prevent data leakage, we trimmed the videos to exclude any frames related to future events. Based on the observed video content, the evaluators were then asked to predict the arguments of the future event. To ensure fair and meaningful evaluation, we provided a predefined vocabulary and reference options, minimizing the influence of lexical variation on prediction quality. On average, each sample took about 30 seconds to complete, and the entire human evaluation process spanned approximately two weeks.

E Baselines

VidEvent[25] VidEvent is an event prediction model, which is proposed to predict events more likely to happen in five candidates. VidEvent is composed of two braches of vision and text. Each branch contains an encoder for video or text to extract the corresponding features, followed by a transformer encoder to make inner interactions among events within one modal. Through this two-branch architecture, VidEvent presents a strong capability in comprehensing event chains, thus achieving extraordinary results in Script event induction task.

InAViT[37] InAViT is a Transformer variant to model interactions by computing the change in the appearance of objects and human hands due to the execution of the actions and use those changes to refine the video representation. Specifically, it models interactions between hands and objects using Spatial Cross-Attention (SCA) and further infuses contextual information using Trajectory Cross-Attention to obtain environment-refined interaction tokens. Through these, InAViT achieves the SOTA performance in STA task.

Video-LlaVA[26] In this approach, the authors unify visual representation into the language feature space to advance the foundational LLM towards a unified LVLM. As a result, they establish a simple but robust LVLM baseline, Video-LLaVA, which learns from a mixed dataset of images and videos, mutually enhancing each other. Video-LLaVA achieves superior performances on a broad range of 9 image benchmarks across 5 image question-answering datasets and 4 image benchmark toolkits.

Qwen2.5-VL[2] Qwen2.5-VL is a comprehensive series of large language models (LLMs) designed to meet diverse needs. Compared to previous iterations, Qwen2.5 has been significantly improved during both the pre-training and post-training stages. In terms of pre-training, the authors have scaled the high-quality pre-training datasets from the previous 7 trillion tokens to 18 trillion tokens. This provides a strong foundation for common sense, expert knowledge, and reasoning capabilities. In terms of post-training, they implement intricate supervised finetuning with over 1 million samples, as well as multistage reinforcement learning. This model currently represents the state-of-the-art (SOTA) in video reasoning tasks.

Table 7: The impact of bounding box quality on AVEP task.

Set	Verb		Noun			Verb-Noun		
	Top1	Top5	P	R	F1	P	R	F1
val	22.13(-0.19)	44.93(-0.23)	31.29(-1.81)	42.98(-1.92)	36.22(-1.89)	7.40(-0.27)	10.03(-0.85)	8.59(-0.41)
test	21.45(-1.26)	44.32(-1.11)	40.96(-1.76)	48.35(-1.89)	44.35(-1.89)	7.32(-0.35)	7.69(-0.02)	7.50(-0.19)

Table 8: A non-exhaustive summary on the related prediction tasks.

Task	Modalities	Form of historical context	Granularity	Task formulation
MCNC	text	event chain	event	Multiple Choice
VLEP	video	video clip	event	Multiple Choice
VidEvent	text + video	event chain	event	Multiple Choice
NEP	video	video clip	event	Multiple Choice
Next Action Anticipation	video	frames	action	Word-based Classification
LTA	video	video clip	action	Word-based Classification
STA	video	video clip	action	Detection&Classification&Regression
AVEP(ours)	text + video	event graph chain	event	Word-based Classification

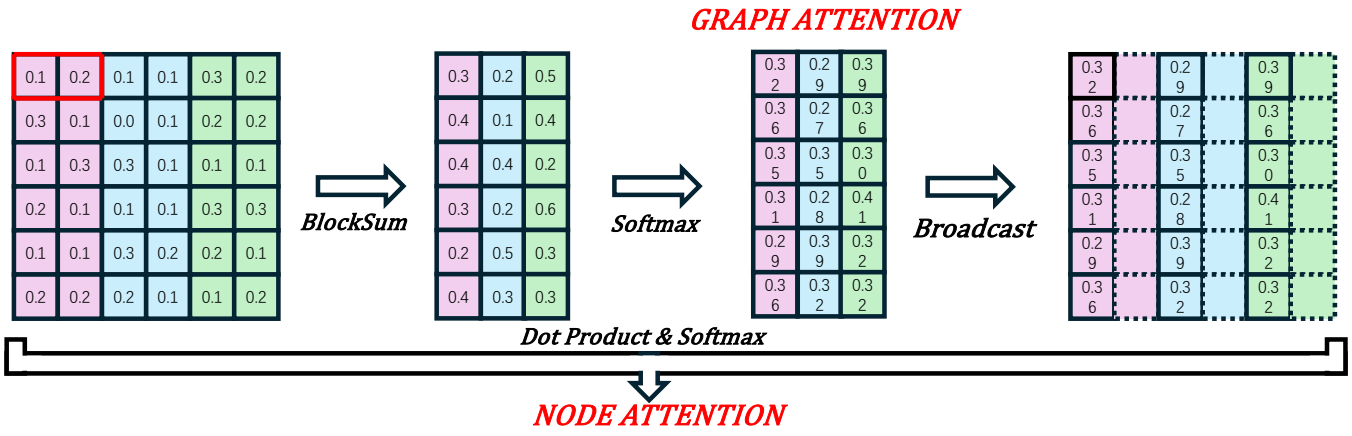

Figure 6: An example illustrating how the attention mechanism is calculated.

F LVLM Experiment details

We compared the performance of state-of-the-art Large Vision-Language Models (LVLMs) in video understanding tasks on the AVEP task under two different parameter scales 7B and 72B. Below, we provide detailed descriptions of the experimental setup and results.

Prompts.

- X_{system}: *"As an event prediction expert, please assist me in completing a video event prediction task."*
- X_0: *"I will provide you with a video segment that describes four sequential events and a list of all characters appearing in the video. Your task is to predict the future event that may occur. Please select the top five most probable verbs from the provided list of verbs and select the relevant argument(s) from the provided list of arguments."*
- X_1: *"Here is the video segment describing four events." @Video*
- X_2: *"Here is the verb list, please choose verbs from it. Answer with the selected verbs without output any other words." @Verb Vocabulary*
- X_3: *"Here is the argument list, please select the relevant argument(s) from the provided list. Answer with the selected arguments separated with "|" without output any other words." @ Argument List*

An Example.

System: As an event prediction expert, please assist me in completing a video event prediction task.

User: I will provide you with a video segment that describes four sequential events and a list of all characters appearing in the video. Your task is to predict the future event that may occur. Please select the top five most probable verbs from the provided list of verbs and select the relevant argument(s) from the provided list of arguments.

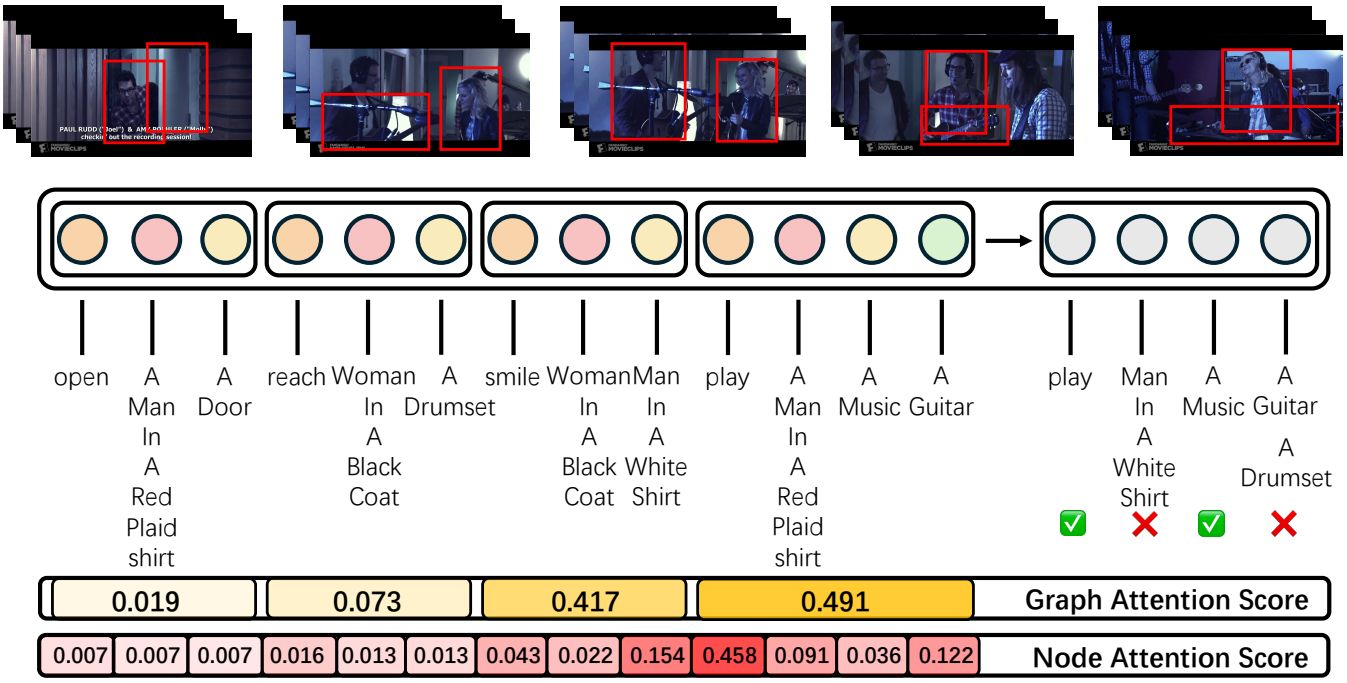

Figure 7: An example illustrating the node-graph hierarchical attention scores and the event prediction result.

Assistant: Sure, I can help you with that. Please provide me with the video segment and the list of arguments and verbs.

User: Here is the video segment describing four events. @Video

Assistant: Based on the video segment, here are the four events that have occurred: 1. A woman is walking down a hallway. 2. The woman enters a bathroom. 3. Another woman is in the bathroom holding a pregnancy test. 4. The first woman takes the pregnancy test from the second woman. Given these events, here are the top five most probable verbs and relevant characters for the future event: 1. Verb: "react" - Arguments: Second woman (the one who was holding the pregnancy test) - The second woman may react to the results of the pregnancy test, either positively or negatively. 2. Verb: "talk" - Arguments: Both women - The two women may talk about the results of the pregnancy test and what it means for their future. 3. Verb: "cry" - Arguments: Either woman - Depending on the results of the pregnancy test, either woman may cry due to happiness, sadness, or shock. 4. Verb: "hug" - Arguments: Both women - The two women may hug each other in support or celebration after receiving the results of the pregnancy test. 5. Verb: "leave" - Arguments: First woman - The first woman may leave the bathroom after taking the pregnancy test from the second woman, possibly to give her some privacy or to process the information.

User: Here is the verb list, please choose verbs from it. Answer with the selected verbs without output any other words." @Verb Vocabulary

Assistant: react talk cry hug leave

User: Here is the argument list, please select the relevant argument(s) from the provided list. Answer with the selected arguments separated with "|" without output any other words." @ Argument List

Assistant: woman with brown hair | cwoman with blonde hair

G Frequent Questions and Answers

This section addresses frequently asked questions that may arise from our main paper and experimental setup. We hope this QFA helps clarify certain design choices and implementation details.

Q1: How dose AVEP differs from VLEP and VidEvent?

A1: VLEP is an outstanding contribution to video event prediction, as it is the first to introduce a multiple-choice formulation for this task. It focuses on verb prediction by providing the subject of the event directly, which allows for a more targeted evaluation of a model's understanding of event dynamics and its ability to anticipate future developments. However, one limitation of VLEP is that it considers only a single historical event as the context for predicting a future event. In real-world scenarios, future events are often shaped by a sequence of preceding events rather than by an isolated instance. Therefore, modeling event prediction based on event chains can offer a more realistic and informative setting. VidEvent is also an excellent work in the field of video understanding, which first proposes a comprehensive video understanding framework composed of several sub-tasks, including event prediction. However, event prediction is not the main focus of this work, and the formulation simply borrows the design of the MCNC task. This multiple-choice format, while convenient, falls short in rigorously evaluating a model's true understanding of events. As shown in their experimental results, models tend to choose options with higher overlap in arguments with the historical context, rather than demonstrating genuine reasoning capabilities. In contrast, our

proposed AVEP task requires the model to accurately predict all arguments of the future event, which forces it to learn the underlying logical relationships between events rather than relying on shallow similarity-based cues. That said, VLEP and VidEvent have made a significant contribution to the field by laying foundational insights and offering a valuable perspective for subsequent research.

Q2: The parameter size of the SOTA model EventFormer in the experimental results is $7.3M$. Why was increasing the model's parameter size not considered to achieve better experimental results?

A2: In our method, we proposed a node-graph hierarchical attention mechanism that significantly improved the model's performance on the task. However, since the current model is primarily designed to validate the effectiveness of our proposed approach, the implementation of the attention mechanism still contains some computational redundancies that could be optimized. Therefore, increasing the model size would have negatively impacted computational efficiency. We made a deliberate trade-off between performance and efficiency. However, we are willing to develop a more optimized implementation of the attention mechanism in the future to further improve the model's efficiency. Notably, our $7.3M$ model already outperforms other larger-scale SOTA models, which demonstrates that our proposed method can more effectively capture complex event relationships and perform event prediction. We look forward to exploring the model's full potential in future work.

Q3: Why did not explore more complex GNN architectures to test whether they could further improve the model's performance?

A3: As shown in the experimental results, different GNN architectures can indeed influence the overall performance of the model. However, it's important to note that in our approach, the GNN module is mainly used to replace Linear later as a module for computing QKV in attention, rather than serving as the core of the model. Therefore, there is no strong justification for significantly increasing the complexity of this component. Meanwhile, considering the scale of the event graph, we suppose that the basic GNNs possess the capability in performing effective graph embedding. Moreover, we suggest that increasing the model's depth by adding more layers can be a more effective way to enhance its capacity and performance, compared to simply using more complex GNNs. However, our work only lays a solid foundation for this research direction, and we warmly welcome further exploration and improvements that may unlock more potential from the model.

H Visualization

We visualize the model's node-graph hierarchcal attention scores and prediction results, as shown in the Figure 7. In this example, the historical event chain consists of events *[open, a man in a red plaid shirt, a door]*, *[reach, woman in a black coat, a drumset]*, *[smile, woman in a black coat, man in a white shirt]*, *[play, a man in a red plaid shirt, a music, a guitar]*, while the ground-truth future event is *[play, woman in a black coat, a music, a drumset]*. The model, however, predicted *[play, man in a white shirt, a music, a guitar]*. At the graph-level attention, the third and fourth events received the highest attention scores 0.417 and 0.491, which aligns well with our intuition. At the node level, the verb node *play* and the noun node *man in a white shirt* received the highest attention scores 0.458 and 0.154. This contributed to the model correctly predicting the future event's verb, but incorrectly predicting the associated noun.