Hierarchical Semantic Correspondence Networks for Video Paragraph Grounding

Chaolei Tan[1] Zihang Lin[1] Jian-Fang Hu[1,2,3*] Wei-Shi Zheng[1,2,3] Jianhuang Lai[1,2,3]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{tanchlei,linzh59}@mail2.sysu.edu.cn, hujf5@mail.sysu.edu.cn,
wszheng@ieee.org, stsljh@mail.sysu.edu.cn

Abstract

Video Paragraph Grounding (VPG) is an essential yet challenging task in vision-language understanding, which aims to jointly localize multiple events from an untrimmed video with a paragraph query description. One of the critical challenges in addressing this problem is to comprehend the complex semantic relations between visual and textual modalities. Previous methods focus on modeling the contextual information between the video and text from a single-level perspective (i.e., the sentence level), ignoring rich visual-textual correspondence relations at different semantic levels, e.g., the video-word and video-paragraph correspondence. To this end, we propose a novel Hierarchical Semantic Correspondence Network (HSCNet), which explores multi-level visual-textual correspondence by learning hierarchical semantic alignment and utilizes dense supervision by grounding diverse levels of queries. Specifically, we develop a hierarchical encoder that encodes the multi-modal inputs into semantics-aligned representations at different levels. To exploit the hierarchical semantic correspondence learned in the encoder for multi-level supervision, we further design a hierarchical decoder that progressively performs finer grounding for lower-level queries conditioned on higher-level semantics. Extensive experiments demonstrate the effectiveness of HSCNet and our method significantly outstrips the state-of-the-arts on two challenging benchmarks, i.e., ActivityNet-Captions and TACoS.

1. Introduction

As a fundamental problem that bridges the gap between computer vision and natural language processing, Video Language Grounding (VLG) aiming to localize the video segments corresponding to given natural language queries, has been drawing increasing attention from the community in these years. Early works in the field of VLG mainly focused on addressing Video Sentence Grounding (VSG) [1, 8], whose goal is to localize the most relevant moment with a single sentence query. Recently, Video Paragraph Grounding (VPG) is introduced in [2]. It requires to jointly localize multiple events via a paragraph query consisting of several temporally ordered sentences. Rather than grounding each event independently, VPG needs to further exploit the contextual information between the video and the textual paragraph, which helps to avoid ambiguity and achieve more precise temporal localization of video events.

*Corresponding author

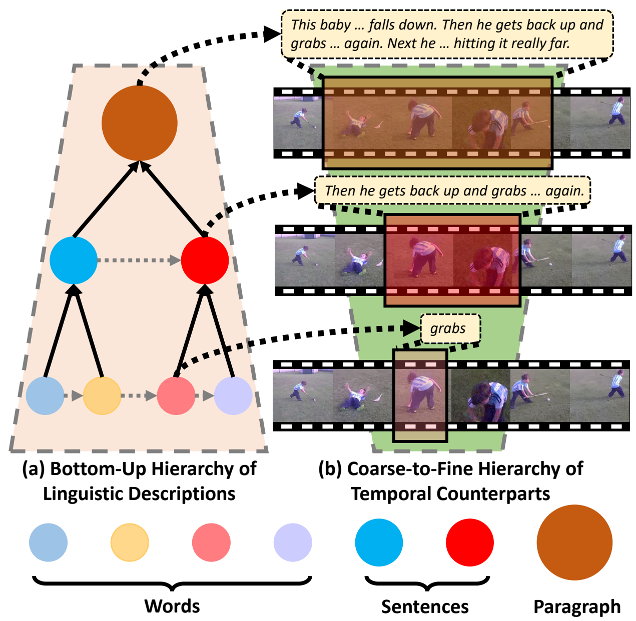

Figure 1. (a) The bottom-up hierarchy for linguistic semantics is composed of textual words, sentences, and the paragraph. (b) The coarse-to-fine hierarchy for temporal granularity consists of visual counterparts for each level of linguistic query.

Previous VPG works [2,12,26] commonly explore correlations of events by modeling the video-text correspondence from a single semantic level (*i.e*., the sentence level). However, they neglect the rich visual-textual correspondence at other semantic levels, such as the word level and paragraph level, which can also provide some useful information for grounding events in the video. Considering the grounding of "The man stops and hits the ball far away", the semantic relations between video content and the word "hits" is crucial in determining the end time of the event. Besides, when we consider the paragraph as a whole, then grounding the holistic paragraph in the video first is beneficial to suppress the irrelevant events or backgrounds, which eases the further grounding of sentences.

To be more general, we observe that there naturally exist two perspectives of hierarchical semantic structures in tackling VPG, which is intuitively illustrated in Figure 1. On the language side, Figure 1 (a) shows that the semantics of paragraph query can be divided into an inherent three-level hierarchy consisting of words, sentences, and the holistic paragraph in a bottom-up organization. On the video side, Figure 1 (b) shows that the temporal counterparts of different levels of queries also form a three-level granularity hierarchy with temporally nested dependencies from the top down. By relating the video content to different levels of query semantics for multi-level query grounding, the model is enforced to capture more complex relations between events by reasoning about their interconnections at different semantic granularities, and exploit richer temporal clues to facilitate the grounding of events in the video.

Motivated by the above observations, we propose a novel framework termed as Hierarchical Semantic Correspondence Network (HSCNet) for VPG. Our HSCNet is designed as a multi-level encoder-decoder architecture in order to leverage hierarchical semantic information from the two perspectives. *On the one hand*, we learn the hierarchical visual-textual semantic correspondence by gradually aligning the visual and textual semantics into different levels of common spaces from the bottom up. Concretely, we construct a hierarchical multi-modal encoder on top of the linguistic semantic hierarchy. It comprises three semantic levels of visual-textual encoders. Each encoder receives the semantic representation from a lower level and continues to establish visual-textual correspondence at a higher level via iterative multi-modal interactions. *On the other hand*, we utilize richer cross-level contexts and denser supervision by progressively grounding multiple levels of queries from coarse to fine. Specifically, we construct a hierarchical progressive decoder on top of the temporal granularity hierarchy, which also comprises three levels of decoders. The lower-level queries are grounded by finer temporal boundaries conditioned on contextual knowledge from higher-level queries, which eases the learning of multi-level localization that provides diverse temporal clues to promote fine-grained video paragraph grounding.

We evaluate the proposed HSCNet on two challenging benchmarks, i.e., ActivityNet-Captions and TACoS. Extensive ablation studies validate the effectiveness of the method. Our contributions can be summarized as follows:

- We investigate and propose a novel hierarchical modeling framework for Video Paragraph Grounding (VPG). To the best of our knowledge, it's the first time in the problem of VPG that hierarchical visual-textual semantic correspondence is explored and multiple levels of linguistic queries can be grounded.
- We design a novel encoder-decoder architecture to learn multi-level visual-textual correspondence by hierarchical semantic alignment and progressively perform finer grounding for lower-level queries.
- Experiments demonstrate that our proposed HSCNet achieves new state-of-the-art results on the challenging ActivityNet-Captions and TACoS benchmarks, remarkably surpassing the previous approaches.

2. Related Work

Video Sentence Grounding. Video Sentence Grounding (VSG) is introduced by [1, 8], which aims to determine the start and end timestamps of the most relevant video segment depicted by a textual sentence query. Existing methods can be roughly grouped into two categories, *i.e*., proposal-based methods and proposal-free methods. Most VSG approaches [1, 3, 5, 8, 10, 20, 31, 32, 36, 39, 40] fall into the proposal-based framework, where candidate segments are generated and then selected by the query matching scores. Although the proposal-based methods perform well in most cases, they suffer from overly expensive computation cost and time consumption, which prevents their applications in more realistic scenarios. More lately, proposal-free methods [4, 22, 33, 34, 37] are developed to tackle VSG by modeling the cross-modal interactions to directly predict the timestamps of the target moment. Despite the above progress, VSG approaches are essentially limited to localizing the single event described by a single sentence, lacking the capability of understanding more complicated paragraph texts with multiple consecutive sentences.

Video Paragraph Grounding. Video Paragraph Grounding (VPG) is a recently emerging task introduced by [2]. It requires to simultaneously determine the start and end timestamps of multiple video segments according to the given paragraph description. Bao *et al.* [2] proposed a Dense Events Propagation Network (DepNet) to effectively capture temporal contexts of multiple events via an aggregation-and-propagation mechanism. Shi *et al.* [26] proposed an end-to-end transformer network to conduct text-conditioned temporal regression. Jiang *et al.* [12] proposed a contrastive encoder to learn the video-paragraph

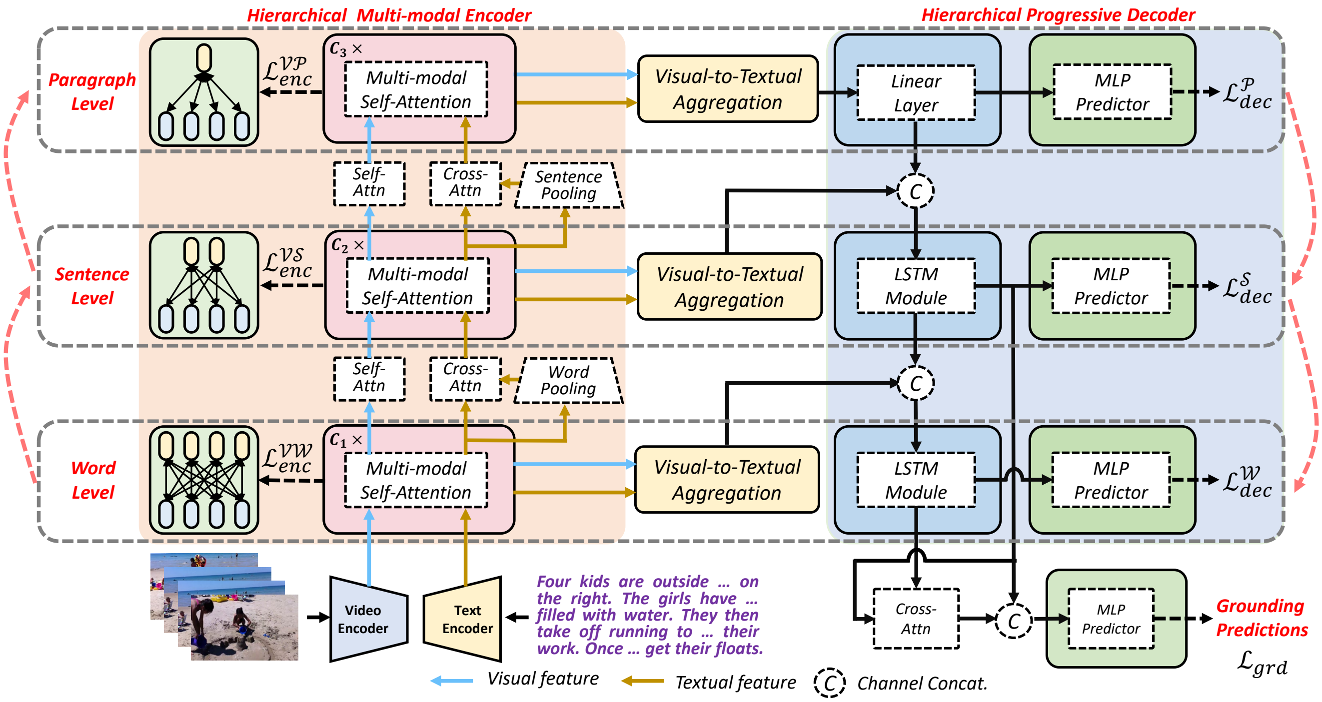

Figure 2. The overall architecture of Hierarchical Semantic Correspondence Network (HSCNet). It mainly consists of a hierarchical multi-modal encoder and a hierarchical progressive decoder. Multi-level semantic correspondence is gradually established from the bottom up in the encoder, while multiple levels of queries are grounded from the top down in the decoder.

matching among sample pairs and explored the semi-supervised setting in VPG. However, existing methods neglect to utilize the hierarchical semantic correspondence between visual and textual modalities. This weakness limits the performance of these methods.

Hierarchical Vision-Language Learning. With the recent progress of Vision-Language Pre-training (VLP), a series of works [9, 15–17, 35, 38] have started to investigate the hierarchical learning of vision-language representation. HERO [15] hierarchically encodes multi-modal inputs for local-global alignment. OSCAR [16] introduces object semantics to align texts and images in a shared space. X-VLM [35] proposes to perform multi-grained alignment between texts and visual concepts. The goal of these VLP methods is to obtain a hierarchical feature representation for various downstream tasks using pre-trained object detectors, off-the-shelf parsers, or additional backbones. In contrast, our approach is free of any additional external components (e.g., object detectors) and aims to establish the hierarchical semantic correspondence between visual and textual modalities for better video paragraph grounding.

3. Methodology

3.1. Overview

Given an untrimmed video and a paragraph query, VPG aims to jointly localize the temporal boundaries of events depicted by the temporally ordered sentences in the paragraph. Specifically, we represent the video as $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^{N^F}$, where N^F is the number of frames. And the paragraph is represented as a set of sentences, *i.e.*, $\mathcal{S} = \{\mathbf{s}_i\}_{i=1}^{N^{\mathcal{S}}}$, where $N^{\mathcal{S}}$ is the number of sentences. The output of VPG can be formulated as $T = \{(t_s, t_e)_i\}_{i=1}^{N^{\mathcal{S}}}$, where $(t_s, t_e)_i$ represents the starting and ending time of the i-th event.

Existing methods mainly achieve dense grounding of events by modeling the cross-modal interactions between $\mathcal{V}$ and $\mathcal{S}$. However, we note that the paragraph can also be semantically parsed as a whole $\mathcal{P}$ or a set of words $\mathcal{W} = \{\mathbf{w}_i\}_{i=1}^{N^{\mathcal{W}}}$, where $N^{\mathcal{W}}$ is the number of words. Motivated by the above considerations, we propose a novel framework that explores the hierarchical semantic relations between $\mathcal{V}$ and $\{\mathcal{P}, \mathcal{S}, \mathcal{W}\}$ to achieve fine-grained cross-modal understanding for high-quality grounding results.

An overview of the model is illustrated in Figure 2. We firstly encode the input video and paragraph into feature representations by a video encoder and text encoder. Then the visual and textual features are forwarded into a hierarchical multi-modal encoder to learn a hierarchical semantic representation of the two modalities. We then employ a hierarchical decoder that progressively leverages the hierarchical semantics to conduct multi-level localization from coarse to fine, which benefits the visual-textual correspondence learning by richer supervision and contexts. For the final prediction, we jointly utilize multi-level semantics to conduct fine-grained temporal localization for VPG.

3.2. Feature Encoder

Video Encoder. We uniformly sample $N^{\mathcal{V}}$ short clips from the video and each clip consists of a fixed number of consecutive frames. Then, we utilize a pre-trained 3D CNN backbone followed by a three-layer self-attention network [28] to extract clip-level visual features as $\mathbf{F}^{\mathcal{V}} = \{\mathbf{f}_i^{\mathcal{V}}\}_{i=1}^{N^{\mathcal{V}}} \in \mathbb{R}^{N^{\mathcal{V}} \times D}$, where D indicates the feature dimension.

Text Encoder. For the paragraph query consisting of $N^{\mathcal{W}}$ words, we convert each word into a vector embedding and then employ a three-layer self-attention network to construct word-level textual features as $\mathbf{F}^{\mathcal{W}} = \{\mathbf{f}_i^{\mathcal{W}}\}_{i=1}^{N^{\mathcal{W}}} \in \mathbb{R}^{N^{\mathcal{W}} \times D}$. Here, the extracted textual features are embedded so that they have the same dimension as visual features.

3.3. Hierarchical Multi-Modal Encoder

We construct a hierarchical encoder on top of three levels of semantic relations (*i.e*., video-word, video-sentence, and video-paragraph), which are naturally derived from the inherent hierarchical structure in the text query. Multi-modal inputs flow through the semantic hierarchy of encoders in a bottom-up manner, establishing visual-textual semantic correspondence at different levels along the way: (1) The word-level encoder encourages to align video content with diverse linguistic details, such as verbs that deliver action dynamics, nouns that deliver entity categories, or other useful contextualized fragments. (2) The sentence-level encoder is responsible for event-centric semantic relations between the video content and sentences, which helps to recognize and reason about the activity concepts. (3) The paragraph-level encoder conducts visual-textual reasoning at the highest level of abstraction, which takes effect in connecting video content with the global semantics of the paragraph.

Word-level Encoder. To capture delicate cross-modal dependencies between the video and paragraph, we construct a word-level encoder to learn the low-level semantic relations between visual and textual modalities. We obtain the initial visual and textual input as $\mathbf{V}^{\mathcal{VW}(0)} = \mathbf{F}^{\mathcal{V}}$ and $\mathbf{Q}^{\mathcal{VW}(0)} = \mathbf{F}^{\mathcal{W}}$. Then in each $(k+1)$-th layer, we employ a multi-modal self-attention mechanism based on semantic similarities. Concretely, we transform the multi-modal inputs into a shared representation $\mathbf{M}^{\mathcal{VW}(k)} \in \mathbb{R}^{(N^{\mathcal{V}}+N^{\mathcal{W}})\times D}$ by concatenation and projection. Then pairwise semantic similarities are computed in $\mathbf{M}^{\mathcal{VW}(k)}$ as:

$$\mathbf{s}_{ij}^{\mathcal{VW}(k)} = \frac{\boldsymbol{\phi}^{(k)}\left(\mathbf{M}_i^{\mathcal{VW}(k)}\right)^{\mathrm{T}} \boldsymbol{\rho}^{(k)}\left(\mathbf{M}_j^{\mathcal{VW}(k)}\right)}{\left\|\boldsymbol{\phi}^{(k)}\left(\mathbf{M}_i^{\mathcal{VW}(k)}\right)\right\| \left\|\boldsymbol{\rho}^{(k)}\left(\mathbf{M}_j^{\mathcal{VW}(k)}\right)\right\|} (\sigma_{vw}^{(k)})^{-1} \quad (1)$$

where $\boldsymbol{\phi}^{(k)}(\cdot)$ and $\boldsymbol{\rho}^{(k)}(\cdot)$ represent different linear projection functions learned by the network, $\sigma^{(k)}$ is a scalar parameter that automatically controls the sharpness of the above scoring function. And $\|\cdot\|$ indicates calculating the value of vector L_2 norm. Afterwards, we utilize another linear function $\boldsymbol{\psi}^{(k)}(\cdot)$ to rearrange the multi-modal semantics of $\mathbf{M}^{\mathcal{VW}(k)}$ as follows:

$$\mathbf{H}^{\mathcal{VW}(k+1)} = \mathrm{Softmax}(\mathbf{s}^{\mathcal{VW}(k)}) \boldsymbol{\psi}^{(k)}\left(\mathbf{M}^{\mathcal{VW}(k)}\right) \quad (2)$$

After the multi-modal self-attention layer, we employ two unshared MLPs that are expert in capturing modality-specific information to obtain the visual output $\mathbf{V}^{\mathcal{VW}(k+1)}$ and textual output $\mathbf{Q}^{\mathcal{VW}(k+1)}$ from $\mathbf{H}^{\mathcal{VW}(k+1)}$. To ensure favorable correspondence learning between the video and textual words, we stack C_1 multi-modal layers to learn the complex multi-modal interactions, *i.e*., $k \in \{0, 1, \cdots C_1 - 1\}$. Through understanding the interconnections between the video and words, the model captures the most subtle visual-linguistic semantic relations, which fosters fine-grained video paragraph grounding.

Sentence-level Encoder. We construct the sentence-level semantic learning on the foundation of word-level semantics. First of all, we employ a word pooling operation to earn the global tokens of textual sentences as follows:

$$\boldsymbol{\Omega}_i^{\mathcal{VS}} = \frac{1}{\left|\mathcal{I}_i^{\mathcal{S}}\right|} \sum_{j \in \mathcal{I}_i^{\mathcal{S}}} \mathbf{Q}_j^{\mathcal{VW}(C_1)} \quad (3)$$

where $\mathcal{I}_i^{\mathcal{S}}$ denotes the set of indices of words within the i-th sentence. Then we use the global tokens as query vectors in the cross-attention mechanism [28] to induce the initial sentence-level textual representation $\mathbf{Q}^{\mathcal{VS}(0)}$ as:

$$\mathbf{Q}^{\mathcal{VS}(0)} = \text{Cross-Attn}\left(\boldsymbol{\Omega}^{\mathcal{VS}}, \mathbf{Q}^{\mathcal{VW}(C_1)}, \mathbf{Q}^{\mathcal{VW}(C_1)}\right) \quad (4)$$

We further employ a self-attention layer on $\mathbf{V}^{\mathcal{VW}(C_1)}$ to obtain $\mathbf{V}^{\mathcal{VS}(0)}$, which enables the reasoning of the word-level visual semantics and makes it adapted for the subsequent sentence-level correspondence learning.

Likewise, we construct a video-sentence representation $\mathbf{M}^{\mathcal{VS}(0)} \in \mathbb{R}^{(N^{\mathcal{V}}+N^{\mathcal{S}})\times D}$ by concatenating and projecting the sentence-level multi-modal inputs $\mathbf{V}^{\mathcal{VS}(0)}$ and $\mathbf{Q}^{\mathcal{VS}(0)}$. A stack of multi-modal self-attention layers and MLPs are used to obtain the visual and textual output $\mathbf{V}^{\mathcal{VS}(C_2)}$ and $\mathbf{Q}^{\mathcal{VS}(C_2)}$, where C_2 is the number of sentence-level layers. $\mathbf{V}^{\mathcal{VS}(C_2)}$ and $\mathbf{Q}^{\mathcal{VS}(C_2)}$ jointly reflect the semantic relations between the video and sentences. The sentence-level multi-modal interactions enable to establish the visual-textual correspondence at a higher level than the words, which is significant for the grounding of events in the video.

Paragraph-level Encoder. Analogously, we also employ a sentence pooling operation followed by a cross-attention operation on $\mathbf{Q}^{\mathcal{VS}(C_2)}$ to obtain the initial paragraph-level textual representation $\mathbf{Q}^{\mathcal{VP}(0)}$. A self-attention layer is also employed on $\mathbf{V}^{\mathcal{VS}(C_2)}$ to obtain $\mathbf{V}^{\mathcal{VP}(0)}$. Again, we form a video-paragraph representation $\mathbf{M}^{\mathcal{VP}(0)}$. Then C_3 multi-modal self-attention layers are iteratively employed on $\mathbf{M}^{\mathcal{VP}(0)}$ for the paragraph-level encoder output $\mathbf{V}^{\mathcal{VP}(C_3)}$ and $\mathbf{Q}^{\mathcal{VP}(C_3)}$. At the paragraph level, we learn to establish the correspondence between video content and the high-level global semantics of the paragraph, which helps to highlight the meaningful content related to the text query and suppress the irrelevant events or backgrounds.

Visual-to-Textual Semantic Aggregation. To exploit the visual-textual correspondence established in the encoder for multi-level grounding, we aggregate the visual contents into the textual queries based on the textual-to-visual semantic relevance, which can be formulated as follows:

$$\mathbf{A}^{\mathcal{V}\ell} = \left(\overline{\mathbf{Q}}^{\mathcal{V}\ell(d)}\right)\left(\overline{\mathbf{V}}^{\mathcal{V}\ell(d)}\right)^{\mathrm{T}}(\tau^{\ell})^{-1} \tag{5}$$

$$\mathbf{g}^{\ell} = \left[\overline{\mathbf{Q}}^{\mathcal{V}\ell(d)}; \mathrm{Softmax}\left(\mathbf{A}^{\mathcal{V}\ell}\right)\overline{\mathbf{V}}^{\mathcal{V}\ell(d)}\right] \tag{6}$$

where $(\ell, d) \in \{(\mathcal{W}, C_1), (\mathcal{S}, C_2), (\mathcal{P}, C_3)\}$. ℓ and d indicate the semantic level and depth of the last layer in each visual-textual encoder, respectively. $\overline{\mathbf{Q}}^{\mathcal{V}\ell(d)}$ and $\overline{\mathbf{V}}^{\mathcal{V}\ell(d)}$ are obtained by employing L2 normalization on the features in $\mathbf{Q}^{\mathcal{V}\ell(d)}$ and $\mathbf{V}^{\mathcal{V}\ell(d)}$, respectively. $\mathbf{A}^{\mathcal{V}\mathcal{W}}$, $\mathbf{A}^{\mathcal{V}\mathcal{S}}$, and $\mathbf{A}^{\mathcal{V}\mathcal{P}}$ are cross-modal semantic similarity matrices. τ^{ℓ} is the temperature and $[;]$ denotes the channel concatenation operation. $\mathbf{g}^{\ell}$ is the multi-modal grounding feature that contains query semantics in both visual and textual modalities, *i.e.*, $\mathbf{g}^{\ell}$ jointly represents what the textual query conveys and what relates to the query in the video at semantic level ℓ.

3.4. Hierarchical Progressive Decoder

In the existing works [2, 12, 26], sentence-based query decoders are commonly employed for video paragraph grounding. Due to the neglect of hierarchical modeling, these methods fail to access the multi-level contextual information or the potential multi-level supervision provided by grounding diverse levels of queries. In this work, we develop a hierarchical decoder that progressively performs finer grounding for lower-level queries conditioned on the contextual knowledge associated with higher-level queries. The decoder utilizes rich contexts for multi-level query grounding, which provides diverse guidance to facilitate the hierarchical visual-textual correspondence learning.

Paragraph-level Decoder. In the beginning, we employ a linear layer on $\mathbf{g}^{\mathcal{P}}$ and then use a two-layer MLP predictor to obtain the grounding results of the paragraph, which is denoted as $\widetilde{\mathbf{T}}^{\mathcal{P}} \in \mathbb{R}^{1\times 2}$. Inspired by the principle of Multiple Instance Learning (MIL) [6, 21], we define the temporal union of sentence-wise timestamps as the ground-truth annotation for the holistic paragraph, which is approximately to say, video content relevant to any one of the sentences should be considered related to the holistic paragraph.

Sentence-level Decoder. At the sentence level, we aim to localize the sentence-wise timestamps utilizing sentence-level semantics. Note that we have obtained the paragraph-level grounding information, which helps to highlight the meaningful video content associated with the global semantics of the paragraph. To utilize the paragraph-level contextual knowledge, we feed $\mathbf{g}^{\mathcal{S}}$ and $\mathbf{g}^{\mathcal{P}}$ to an LSTM [11] module as follows:

$$\{\mathbf{g}_i^{\mathcal{S}\mathcal{P}}\}_{i=1}^{N^{\mathcal{S}}} = \mathrm{LSTM}\left(\{[\mathbf{g}_i^{\mathcal{S}}; \mathbf{g}^{\mathcal{P}}]\}_{i=1}^{N^{\mathcal{S}}}\right) \tag{7}$$

where $\mathbf{g}^{\mathcal{P}}$ serves as a conditional context vector that facilitates the learning of sentence-level grounding. Then a two-layer MLP predictor is employed on $\mathbf{g}^{\mathcal{S}\mathcal{P}}$ to acquire temporal boundaries for each sentence, *i.e.*, $\widetilde{\mathbf{T}}^{\mathcal{S}} \in \mathbb{R}^{N^{\mathcal{S}}\times 2}$.

Word-level Decoder. For the word-level decoding, we perform temporal localization with respect to each individual word, which stimulates the learning of fine-grained grounding. To this end, we first obtain $\widetilde{\mathbf{g}}^{\mathcal{W}} \in \mathbb{R}^{N^{\mathcal{S}}\times M\times D}$ by reformatting the word-level tokens distributed in the paragraph (*i.e.*, $\mathbf{g}^{\mathcal{W}} \in \mathbb{R}^{N^{\mathcal{W}}\times D}$) into word-level tokens distributed in single sentences, where M is the padding size and we set it as the length of the longest sentence in the paragraph. Then we input $\widetilde{\mathbf{g}}^{\mathcal{W}}$ and $\mathbf{g}^{\mathcal{S}\mathcal{P}}$ into an LSTM module to learn the word-level contexts as follows:

$$\left\{\widetilde{\mathbf{g}}_{i,j}^{\mathcal{W}\mathcal{S}}\right\}_{j=1}^{N_i^{\mathcal{W}}} = \mathrm{LSTM}\left(\left\{[\widetilde{\mathbf{g}}_{i,j}^{\mathcal{W}}; \mathbf{g}_i^{\mathcal{S}\mathcal{P}}]\right\}_{j=1}^{N_i^{\mathcal{W}}}\right) \tag{8}$$

where $\widetilde{\mathbf{g}}^{\mathcal{W}\mathcal{S}}$ is the word-level grounding features learned in the context of the sentence it belongs to. We also employ a two-layer MLP on $\widetilde{\mathbf{g}}^{\mathcal{W}\mathcal{S}}$ to acquire word-level grounding results $\widetilde{\mathbf{T}}^{\mathcal{W}} \in \mathbb{R}^{N^{\mathcal{S}}\times M\times 2}$, which is supervised by an approximate word-level grounding loss.

Grounding Prediction. To jointly exploit multi-level semantics for final prediction, we first use a cross-attention layer to select word-level semantics highly relevant to the sentence's grounding, which is denoted as $\mathbf{g}^{\mathcal{W}\mathcal{S}} \in \mathbb{R}^{N^{\mathcal{S}}\times D}$. Then we jointly utilize multi-level features for grounding by forming $[\mathbf{g}^{\mathcal{W}\mathcal{S}}; \mathbf{g}^{\mathcal{S}\mathcal{P}}]$. A two-layer MLP predictor is employed to obtain the final output for VPG, *i.e.*, $\widehat{\mathbf{T}} \in \mathbb{R}^{N^{\mathcal{S}}\times 2}$.

3.5. Training Loss

Encoder Loss. To guide the learning of hierarchical multi-modal interactions, we employ multi-level semantic alignment loss in the encoder, which is constructed based on the correspondence relationships between visual and textual modalities at different semantic levels. The encoder loss $\mathcal{L}_{enc}$ is formulated as:

$$\mathcal{L}_{enc} = \mathcal{L}_{enc}^{\mathcal{V}\mathcal{W}} + \mathcal{L}_{enc}^{\mathcal{V}\mathcal{S}} + \mathcal{L}_{enc}^{\mathcal{V}\mathcal{P}} \tag{9}$$

where $\mathcal{L}_{enc}^{\mathcal{V}\mathcal{W}}$, $\mathcal{L}_{enc}^{\mathcal{V}\mathcal{S}}$, and $\mathcal{L}_{enc}^{\mathcal{V}\mathcal{P}}$ are computed on the visual-textual semantic similarity matrices at different levels, *i.e.*, $\mathbf{A}^{\mathcal{V}\mathcal{W}}$, $\mathbf{A}^{\mathcal{V}\mathcal{S}}$, and $\mathbf{A}^{\mathcal{V}\mathcal{P}}$ derived from eq.5. The alignment loss is computed as the negative log-likelihood of the sum of semantic similarities after softmax operation.

Decoder Loss. We employ multiple levels of localization loss on the intermediate grounding results given by different

Table 1. Comparison with state-of-the-arts on ActivityNet-Captions and TACoS.

Method	ActivityNet-Captions				TACoS			
	R@IoU=0.3	R@IoU=0.5	R@IoU=0.7	mIoU	R@IoU=0.1	R@IoU=0.3	R@IoU=0.5	mIoU
DRN [34]	-	45.45	24.36	-	-	-	23.17	-
2D-TAN [39]	59.45	44.51	26.54	-	47.59	37.29	-	-
BPNet [30]	58.98	42.07	24.69	42.11	-	25.96	20.96	19.53
CBLN [19]	66.34	48.12	27.60	-	49.16	38.98	27.65	-
MMN [29]	65.05	48.59	29.26	-	51.39	39.24	26.17	-
SLP [18]	-	52.89	32.04	-	-	42.73	32.58	-
Beam Search [7]	62.53	46.43	27.12	-	48.46	38.14	25.72	-
3D-TPN [39]	67.56	51.49	30.92	-	55.05	40.31	26.54	-
DepNet [2]	72.81	55.91	33.46	-	56.10	41.34	27.16	-
PRVG [26]	78.27	61.15	37.83	55.62	61.64	45.40	26.37	29.18
SVPTR [12]	78.07	61.70	38.36	55.91	67.91	47.89	28.22	31.42
HSCNet	**81.89**	**66.57**	**44.03**	**59.71**	**76.28**	**59.74**	**42.00**	**40.61**

levels of decoder, which is formulated as:

$$\mathcal{L}_{dec} = \mathcal{L}_{dec}^{\mathcal{P}} + \mathcal{L}_{dec}^{\mathcal{S}} + \underbrace{\mathcal{L}_{union}^{\mathcal{W}} + \mathcal{L}_{subset}^{\mathcal{W}}}_{\text{weakly-supervised } \mathcal{L}_{dec}^{\mathcal{W}}} \tag{10}$$

where the paragraph-level decoding loss $\mathcal{L}_{dec}^{\mathcal{P}}$ and sentence-level decoding loss $\mathcal{L}_{dec}^{\mathcal{S}}$ both consist of a L1 distance loss and a GIoU [24] loss supervised by the ground-truth timestamps. For word-level decoding, since there is no ground-truth supervision provided, we design a weakly-supervised loss for approximation of word-level localization. Specifically, $\mathcal{L}_{union}^{\mathcal{W}}$ constrains the temporal union of timestamps corresponding to all words within a sentence is close to the timestamp of that sentence, and $\mathcal{L}_{subset}^{\mathcal{W}}$ encourages each word to be grounded as a temporal subset of the timestamps corresponding to the sentence it belongs to.

Grounding Loss. For the final grounding predictions $\widehat{T}$, we also employ a localization loss $\mathcal{L}_{grd}$, which is computed as the sum of L1 distance and GIoU Loss between the predicted and ground-truth timestamps.

In total, we jointly minimize the encoder loss, decoder loss, and grounding loss for end-to-end model training:

$$\mathcal{L}_{total} = \mathcal{L}_{enc} + \mathcal{L}_{dec} + \mathcal{L}_{grd} \tag{11}$$

We provide more implementation details about the training loss functions in the supplementary materials.

4. Experiments

4.1. Datasets and Evaluation Metrics

ActivityNet-Captions. ActivityNet-Captions [14] is originally collected for dense video captioning and is later introduced into VSG and VPG. The training, val_1, and val_2 sets include 37417, 17505, and 17031 annotated sentences, respectively. On average, each paragraph consists of 4.08 sentences and the duration of annotated moments is 36.2 seconds. Following the previous work [2], We adopt val_2 as the testing set.

TACoS. TACoS is manually collected from MPII Cooking Composite Activities dataset [25]. Each video is annotated with diverse paragraph descriptions at different granularities. On average, each video has a duration of 4.79 minutes and each paragraph consists of 8.75 sentences in total. There are 10146, 4589, and 4083 annotated sentences for training, validation, and testing sets, respectively.

Evaluation Metrics. We adopt the recall with an IoU threshold of m to evaluate grounding performance under various precision requirements, which is denoted as R@m. And m is set to be $\{0.3, 0.5, 0.7\}$ on ActivityNet-Captions and $\{0.1, 0.3, 0.5\}$ on TACoS, respectively. We also adopt mIoU to evaluate the overall grounding performance of the model. Following the previous work [2], reported evaluation metrics are averaged over all sentences in the dataset.

4.2. Implementation Details

We uniformly sample 256 and 512 video clips for ActivityNet-Captions and TACoS, respectively. The length of each video clip is set to be 16 on all datasets. For fair comparison with the previous works [2, 12, 26], we adopt the same backbones for feature extraction, *i.e.*, we employ the pre-trained C3D [27] model without fine-tuning to extract visual features for video clips and employ pre-trained Glove [23] model to extract word-level features for the paragraph. The depth of encoder layers $\{C_1, C_2, C_3\}$ is set as $\{1, 1, 1\}$ and $\{3, 3, 3\}$ on ActivityNet-Captions and TACoS. We train the model by Adam [13] optimizer without weight decay. The learning rate is set as 0.0001 on all datasets and the batch size is set as 32 and 16 for ActivityNet-Captions and TACoS, respectively. Following [28], we implement multi-modal self-attention layers in a multi-head fashion.

The temperature τ^ℓ is empirically set to be 0.2 for all levels. The hidden size D is set to be 256 in all settings.

4.3. Comparison with state-of-the-arts

To show the superiority of our proposed method, we compare it with the existing state-of-the-art VPG approaches including DepNet [2], PRVG [26], and SVPTR [12] on ActivityNet-Captions and TACoS benchmarks. Two baselines (*i.e*., Beam Search and 3D-TPN) proposed in [2] are also reported for reference. For more comprehensive comparison, we also compare our method with the state-of-the-art VSG approaches, including DRN [34], 2D-TAN [39], BPNet [30], CBLN [19], MMN [29] and SLP [18].

As shown in Table 1, our HSCNet outperforms the existing state-of-the-arts in all evaluation metrics by a significant margin, which shows the superiority of our method with hierarchical modeling. Concretely, our method achieves an mIoU performance of 59.71% and 40.61% on ActivityNet-Captions and TACoS, which exceeds the state-of-the-art VPG approaches [2, 12, 26] without hierarchical modeling by 3.80% and 9.19%, respectively. This verifies that explicitly exploring hierarchical visual-textual correspondences (*i.e*., video-word, video-sentence, and video-paragraph) is beneficial for video paragraph grounding. We also note that previous methods perform worse and achieve lower recall rates on TACoS than on ActivityNet-Captions. The reason is that TACoS is more challenging due to its longer videos and more complicated paragraphs. The more complex structure of videos and paragraphs in TACoS deteriorates the performance of previous methods, while strengthening the advantages of our method in which hierarchical modeling is utilized to handle complicated visual-textual semantic relations. Specifically, our HSCNet can bring a remarkable improvement up to 13.78% in R@0.5 on TACoS, which further validates the superiority of our method.

4.4. Quantitative Analysis

In this section, we conduct extensive ablation studies on TACoS to verify the effectiveness of our model designs.

Impact of Hierarchical Modeling We study the impact of the proposed hierarchical modeling by removing certain levels (e.g., word level and paragraph level) from both of the hierarchical encoder and decoder. The results are presented in Table 2. By converting our hierarchical model into a single-level (sentence-level) model, we observe a clear degradation in performance up to 7.95% of mIoU, which indicates the importance of hierarchical modeling in VPG. Moreover, we find that modeling the word level or paragraph level consistently improves the system performance, and simultaneously modeling all the semantic levels performs the best, which demonstrates that the formulated multiple semantic levels complement well with each other.

Impact of Alignment Loss. We conduct extensive experiments to investigate the influences of visual-textual alignment loss at different semantic levels. As shown in Table 3, we can see that each level of semantic alignment brings some gains to the model performance. More specifically, the word-level alignment loss has the greatest impact on the performance, which brings an improvement of 12.64% (row 1 vs. row 4), 11.93% (row 2 vs. row 6), 5.68% (row 3 vs. row 7), and 6.15% (row 5 vs. row 8) in mIoU. This is because the word-level correspondence captures the fine-grained semantic relations between visual and textual modalities, which is crucial for obtaining more accurate event boundary prediction.

Table 2. Ablation study on hierarchical modeling. We abbreviate the modeling of word level, sentence level, and paragraph level as WL, SL, and PL, respectively.

Level	R@0.3	R@0.5	mIoU
SL	49.27	30.14	32.66
SL + PL	53.73	32.67	35.18
SL + WL	54.79	38.76	37.43
SL + PL + WL	**59.74**	**42.00**	**40.61**

Table 3. Ablation study on semantic alignment loss in the encoder.

$\mathcal{L}_{enc}^{\mathcal{W}}$	$\mathcal{L}_{enc}^{\mathcal{S}}$	$\mathcal{L}_{enc}^{\mathcal{P}}$	R@0.3	R@0.5	mIoU
✗	✗	✗	29.98	12.81	22.38
✗	✗	✓	36.34	17.14	25.56
✗	✓	✗	48.94	31.18	32.71
✓	✗	✗	52.23	34.25	35.02
✗	✓	✓	51.66	32.86	34.46
✓	✗	✓	55.66	35.85	37.49
✓	✓	✗	56.80	38.82	38.39
✓	✓	✓	**59.74**	**42.00**	**40.61**

Table 4. Ablation Study on different levels of decoders. We abbreviate the word-level, sentence-level, and paragraph-level decoder as WD, SD, and PD, respectively.

Decoder	R@0.3	R@0.5	mIoU
SD	53.56	38.55	37.57
SD + PD	55.41	38.90	38.33
SD + WD	56.78	40.21	39.10
SD + PD + WD	**59.74**	**42.00**	**40.61**

Impact of Decoders. As shown in Table 4, we verify the effectiveness of the design of hierarchical decoder. For the baseline model in row 1, it is obtained by discarding the paragraph-level decoder and word-level decoder. In row 2 and row 3, we observe that enabling the paragraph-level or word-level decoder is beneficial to improve the performance. Row 4 indicates that jointly employing all levels of decoder is most effective to boost the performance, with up to 3.04% gains in mIoU compared with the baseline.

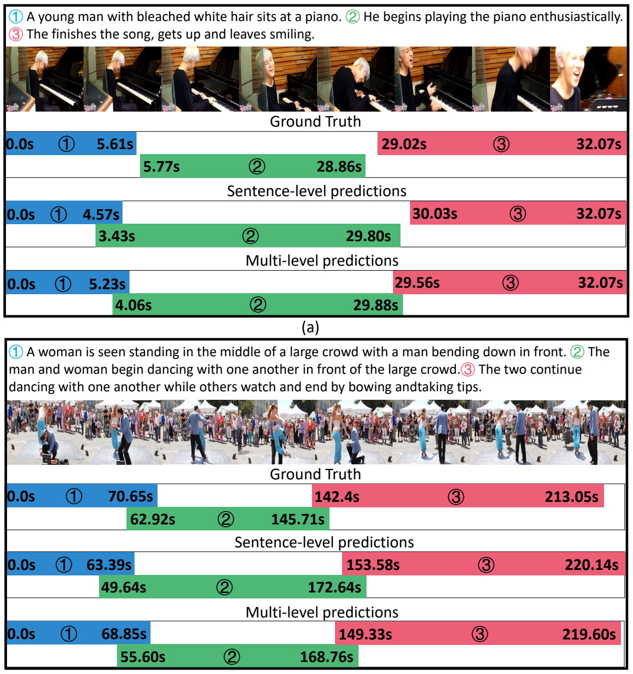

Figure 3. Comparison of single-level (sentence-level) predictions and the final multi-level predictions.

4.5. Qualitative Analysis

In this section, we would like to present some visualization results about HSCNet for further qualitative analysis.

Comparison of single-level and multi-level predictions. As shown in Figure 3, we visualize the model's prediction results and compare them with the ground-truth and the intermediate grounding results given by the sentence-level decoder. It can be seen that the final predictions inferred via multi-level semantic features are close to the ground-truth and have finer boundaries than the sentence-based single-level predictions, which verifies the efficacy of jointly utilizing multiple levels of semantics in the hierarchy.

Demonstration of the hierarchy understanding ability. We visualize the grounding results given by our hierarchical model for multiple levels of linguistic queries in Figure 4, which helps to intuitively demonstrate the hierarchy understanding ability of HSCNet towards the video-text semantic correspondence. For the high-level paragraph understanding, we can see the holistic paragraph is grounded by a lengthy moment spanning plenty of video content mentioned by the text query. For the sentence-level understanding, a series of textual sentences are successfully grounded by a set of moments located within the paragraph's duration. These event moments consistently follow the same temporal order as their sentences. Furthermore, we also present detailed results for word-level understanding of "She peels and chops the kiwis". It's clear that we can observe success cases where two consecutive verbs "peels" and "chops"

Figure 4. Demonstration of the model's hierarchy understanding ability towards the video-text semantic correspondence.

are exactly grounded by two temporally ordered moments, whose temporal union approximately spans across the full event. Meanwhile, we also observe some failure cases, e.g., the word "kiwis" is only grounded by the second half of the event's temporal interval, while it actually lasts for the entire period. The reason might be that when the fruit is peeled, it is heavily occluded by the knife and hands, making it difficult to be identified.

5. Conclusion

In this paper, we propose a novel Hierarchical Semantic Correspondence Network (HSCNet) to explicitly explore the hierarchical semantic structures in tackling the problem of Video Paragraph Grounding (VPG). Specifically, we design our HSCNet as a three-level encoder-decoder architecture. The encoder is employed to learn different levels of visual-textual correspondence along a semantic hierarchy from the bottom up, and the decoder progressively performs finer temporal grounding for lower-level queries conditioned on higher-level queries from coarse to fine. The combination of our hierarchical encoder and decoder enables the model to achieve fine-grained vision-language understanding and precise query-based localization of events in the video. Extensive experiments validate the effectiveness of our HSCNet and demonstrate that our method achieves new state-of-the-art results on two challenging benchmarks, i.e., ActivityNet-Captions and TACoS.

Acknowledgements. This work was supported partially by the NSFC (U21A20471, U22A2095, 62076260, 61772570), Guangdong Natural Science Funds Project (2020B1515120085, 2023B1515040025), Guangdong NSF for Distinguished Young Scholar (2022B1515020009, 2018B030306025), and the Key-Area Research and Development Program of Guangzhou (202007030004). We would like to thank Bing Shuai for the helpful discussions with us.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 1, 2

[2] Peijun Bao, Qian Zheng, and Yadong Mu. Dense events grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 920–928, 2021. 1, 2, 5, 6, 7

[3] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP)*, pages 162–171, 2018. 2

[4] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10551–10558, 2020. 2

[5] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8199–8206, 2019. 2

[6] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997. 5

[7] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. *ACL 2017*, page 56, 2017. 6

[8] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 1, 2

[9] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 2022. 3

[10] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. *In Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1984–1990, 2019. 2

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[12] Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. Semi-supervised video paragraph grounding with contrastive encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2466–2475, 2022. 2, 5, 6, 7

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 6

[15] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020. 3

[16] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 3

[17] Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4395–4405, 2022. 3

[18] Daizong Liu and Wei Hu. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 6, 7

[19] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021. 6, 7

[20] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 15–24, 2018. 2

[21] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 5

[22] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 2

[23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 6

[24] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[25] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *European conference on computer vision*, pages 144–157. Springer, 2012. 6

[26] Fengyuan Shi, Limin Wang, and Weilin Huang. End-to-end dense video grounding via parallel regression. *arXiv preprint arXiv:2109.11265*, 2021. 2, 5, 6, 7

[27] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 6

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 6

[29] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2613–2623, 2022. 6, 7

[30] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2986–2994, 2021. 6, 7

[31] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069, 2019. 2

[32] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[33] Yitian Yuan, Tao Mei, and Wenwu 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*, volume 33, pages 9159–9166, 2019. 2

[34] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 2, 6, 7

[35] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*. PMLR, 2022. 3

[36] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019. 2

[37] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 2

[38] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 3

[39] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877, 2020. 2, 6, 7

[40] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 655–664, 2019. 2