Hierarchical Semantic Correspondence Networks for Video Paragraph Grounding

Chaolei Tan Zihang Lin Jian-Fang Hu Wei-Shi Zheng Jianhuang Lai
1School of Computer Science and Engineering, Sun Yat-sen University, China
2Guangdong Province Key Laboratory of Information Security Technology, China
3Key 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
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
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 $\mathcal{V} = \{v_i\}_{i=1}^{N^F}$ where $N^F$ is the number of frames. And the paragraph is represented as a set of sentences, i.e., $\mathcal{S} = \{s_i\}_{i=1}^{N^S}$ where $N^S$ is the number of sentences. The output of VPG can be formulated as $T = \{(t_s, t_e)\}_{i=1}^{N^S}$, where $(t_s, t_e)$ 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 $\mathcal{W} = \{w_i\}_{i=1}^{N^W}$, where $N^W$ 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 \( N^V \) 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 \( F^V = \{f^V_i\}_{i=1}^{N^V} \in \mathbb{R}^{N^V \times D} \), where \( D \) indicates the feature dimension.

**Text Encoder.** For the paragraph query consisting of \( N^W \) 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 \( F^W = \{f^W_i\}_{i=1}^{N^W} \in \mathbb{R}^{N^W \times D} \). 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 \( V^W = F^V \) and \( Q^W(f) = F^W \). 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 \( M^{VW(k)} \in \mathbb{R}^{(N^V + N^W) \times D} \) by concatenation and projection. Then pairwise semantic similarities are computed in \( M^{VW(k)} \) as:

\[
 s_{ij}^{VW(k)} = \frac{\phi(k)^T \rho(k)}{\|\phi(k)\| \|\rho(k)\| \|M^{VW(k)}\|} \sigma_{vw}(\cdot)^{-1}
\]

where \( \phi(k) \) and \( \rho(k) \) 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 L2 norm. Afterwards, we utilize another linear function \( \psi(k)(\cdot) \) to rearrange the multi-modal semantics of \( M^{VW(k)} \) as follows:

\[
 H^{VW(k+1)} = \text{Softmax}(s^{VW(k)}) \psi(k) \left( M^{VW(k)} \right)
\]

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 \( V^{VW(k+1)} \) and textual output \( Q^{VW(k+1)} \) from \( H^{VW(k+1)} \). 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 learn the global tokens of textual sentences as follows:

\[
 \Omega^S_i = 1 \frac{1}{|Z^S_i|} \sum_{j \in Z^S_i} Q^{VW(C)}(k)
\]

where \( Z^S_i \) 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 \( Q^{V^S(0)} \) as:

\[
 Q^{V^S(0)} = \text{Cross-Atttn} ( \Omega^S, Q^{VW(C)}, Q^{VW(C)})
\]

We further employ a self-attention layer on \( V^{VW(C)} \) to obtain \( V^{V^S(0)} \), 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 \( M^{V^S(0)} \in \mathbb{R}^{(N^V + N^S) \times D} \) by concatenating and projecting the sentence-level multi-modal inputs \( V^{V^S(0)} \) and \( Q^{V^S(0)} \). A stack of multi-modal self-attention layers and MLPs are used to obtain the visual and textual output \( V^{V^S(C)} \) and \( Q^{V^S(C)} \), where \( C_2 \) is the number of sentence-level layers. \( V^{S(C)} \) and \( Q^{V^S(C)} \) 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 \( Q^{V^S(C)} \) to obtain the initial paragraph-level textual representation \( Q^{V^P(0)} \). A self-attention layer is also employed on \( V^{V^S(C)} \) to obtain \( V^{V^P(0)} \). Again, we form a video-paragraph representation \( M^{V^P(0)} \). Then \( C_3 \) multi-modal self-attention layers are iteratively employed on \( M^{V^P(0)} \) for the paragraph-level encoder output \( V^{V^P(C)} \) and \( Q^{V^P(C)} \). At the paragraph level, we learn to establish the correspondence between video content and the highest-level global semantics of the paragraph, which helps to

18976
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:

\[
A_V^\ell = \left( \mathbf{Q}^V(d) \mathbf{V}^V(d) \right)^T \left( \mathbf{r}^\ell \right)^{-1}
\]

\[
g^\ell = \left[ \mathbf{Q}^V(d) , \text{Softmax} ( A_V^\ell \mathbf{V}^V(d) ) \right]
\]

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

### 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 \( g^P \) and then use a two-layer MLP predictor to obtain the grounding results of the paragraph, which is denoted as \( \mathbf{T}^P \in \mathbb{R}^{1 \times 2} \). 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 \( g^P \) and \( g^S \) to an LSTM [11] module as follows:

\[
\{ \mathbf{g}_i^{SP} \}_{i=1}^{N^S} = \text{LSTM} \left( \left\{ \left[ \mathbf{g}_i^P ; \mathbf{g}_i^S \right] \right\}_{i=1}^{N^S} \right)
\]

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

**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 \( \mathbf{g}^W \in \mathbb{R}^{N^S \times M \times D} \) by reformatting the word-level tokens distributed in the paragraph (i.e., \( g^W \in \mathbb{R}^{N^W \times D} \)) 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 \( g^W \) and \( g^{SP} \) into an LSTM module to learn the word-level contexts as follows:

\[
\{ \mathbf{g}_{i,j}^{WS} \}_{i,j=1}^{N^W} = \text{LSTM} \left( \left\{ \left[ \mathbf{g}_{i,j}^W ; \mathbf{g}_{i,j}^{SP} \right] \right\}_{i,j=1}^{N^W} \right)
\]

where \( g^{WS} \) is the word-level grounding features learned in the context of the sentence it belongs to. We also employ a two-layer MLP on \( g^{WS} \) to acquire word-level grounding results \( \mathbf{T}^W \in \mathbb{R}^{N^S \times M \times 2} \), 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 \( g^{WS} \in \mathbb{R}^{N^W \times D} \). Then we jointly utilize multi-level features for grounding by forming \( [g^{WS}, g^{SP}] \). A two-layer MLP predictor is employed to obtain the final output for VPG, i.e., \( \mathbf{\hat{T}} \in \mathbb{R}^{N^S \times 2} \).

### 3.5. Training Loss

**Encoder Loss.** To guide the learning of hierarchical multimodal 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 \( \mathcal{L}_{\text{enc}} \) is formulated as:

\[
\mathcal{L}_{\text{enc}} = \mathcal{L}_{\text{enc}}^{VW} + \mathcal{L}_{\text{enc}}^{VS} + \mathcal{L}_{\text{enc}}^{VP}
\]

where \( \mathcal{L}_{\text{enc}}^{VW} \), \( \mathcal{L}_{\text{enc}}^{VS} \), and \( \mathcal{L}_{\text{enc}}^{VP} \) are computed on the visual-textual semantic similarity matrices at different levels, i.e., \( A_V^W \), \( A_V^S \), and \( A_V^P \) 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
levels of decoder, which is formulated as:

$$L_{dec} = L_{dec}^P + L_{dec}^S + L_{union}^W + L_{subset}^W$$  (10)


where the paragraph-level decoding loss $L_{dec}^P$ and sentence-level decoding loss $L_{dec}^S$ 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, $L_{union}^W$ constrains the temporal union of timestamps corresponding to all words within a sentence is close to the timestamp of that sentence, and $L_{subset}^W$ 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 $\hat{T}$, we also employ a localization loss $L_{grd}$, 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:

$$L_{total} = L_{enc} + L_{dec} + L_{grd}$$  (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 $\tau^\ell$ 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 SVPT [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.

<table>
<thead>
<tr>
<th>Level</th>
<th>R@0.3</th>
<th>R@0.5</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL</td>
<td>49.27</td>
<td>30.14</td>
<td>32.66</td>
</tr>
<tr>
<td>SL + PL</td>
<td>53.73</td>
<td>32.67</td>
<td>35.18</td>
</tr>
<tr>
<td>SL + WL</td>
<td>54.79</td>
<td>38.76</td>
<td>37.46</td>
</tr>
<tr>
<td>SL + PL + WL</td>
<td>59.74</td>
<td>42.00</td>
<td>40.61</td>
</tr>
</tbody>
</table>

**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.

**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.

<table>
<thead>
<tr>
<th>Decoder</th>
<th>R@0.3</th>
<th>R@0.5</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td>53.56</td>
<td>38.55</td>
<td>37.57</td>
</tr>
<tr>
<td>SD + PD</td>
<td>55.41</td>
<td>38.90</td>
<td>38.33</td>
</tr>
<tr>
<td>SD + WD</td>
<td>56.78</td>
<td>40.21</td>
<td>39.10</td>
</tr>
<tr>
<td>SD + PD + WD</td>
<td>59.74</td>
<td>42.00</td>
<td>40.61</td>
</tr>
</tbody>
</table>
A young man with bleached white hair sits at a piano. He begins playing the piano enthusiastically. The young man then finishes the song, gets up and leaves smiling.

1. A woman and man begin dancing with one another in front of the large crowd. They continue dancing with one another while others watch and end by bowing and taking tips.

2. A young man with bleached white hair sits at a piano. He begins playing the piano enthusiastically. The young man then finishes the song, gets up and leaves smiling.

3. The two continue dancing with one another while others watch and end by bowing and taking tips.

4. She gets two kiwis. She gets a cutting board. She gets a plate. She cleans up. She peels and chops the kiwis. She cleans up.

5. She peels and chops the kiwis. The kiwis are rinsed to remove any dirt. The kiwis are then peeled and chopped into smaller pieces. The chopped kiwis are then cleaned up.

6. The peeling and chopping of the kiwis are completed. The kiwis are then cleaned up.


