Learning Multi-Scale Video-Text Correspondence for Weakly Supervised Temporal Article Grounding

Wenjia Geng[1], Yong Liu[1], Lei Chen[3*], Sujia Wang[1], Jie Zhou[2], Yansong Tang[1]

[1]Shenzhen International Graduate School, Tsinghua University
[2]Department of Automation, Tsinghua University
[3]University of Science and Technology Beijing
{gengwj22@, liuyong23@, wsj22@}mails.tsinghua.edu.cn, chenlei2022@ustb.edu.cn,
{jzhou@, tang.yansong@sz.}tsinghua.edu.cn

Abstract

Weakly Supervised temporal Article Grounding (WSAG) is a challenging and practical task in video understanding. Specifically, given a video and a relevant article, whose sentences are at different semantic scales, WSAG aims to localize corresponding video segments for all "groundable" sentences. Compared to other grounding tasks, *e.g.,* localizing one target segment with respect to a given sentence query, WSAG confronts an essential obstacle rooted in the intricate multi-scale information inherent within both textual and visual modalities. Existing methods overlook the modeling and alignment of such structured information present in multi-scale video segments and hierarchical textual content. To this end, we propose a Multi-Scale Video-Text Correspondence Learning (MVTCL) framework, which enhances the grounding performance in complex scenes by modeling multi-scale semantic correspondence both within and between modalities. Specifically, MVTCL initially aggregates video content spanning distinct temporal scales and leverages hierarchical textual relationships in both temporal and semantic dimensions via a semantic calibration module. Then multi-scale contrastive learning module is introduced to generate more discriminative representations by selecting typical contexts and performing inter-video contrastive learning. Through the multi-scale semantic calibration architecture and supervision design, our method achieves new state-of-the-art performance on existing WSAG benchmarks.

Introduction

Video Grounding (Anne Hendricks et al. 2017; Gao et al. 2017; Chen et al. 2020) aims to localize target segments from an untrimmed video with respect to the given language query, which is fundamental to various multi-modal tasks, such as video question answering (Le et al. 2020; Zhu et al. 2017), video context retrieval (Gabeur et al. 2020), and video storytelling (Li et al. 2019).

Early works in Video Grounding mainly focus on single sentence grounding (Ma et al. 2020; Song et al. 2020), which aims to localize the most relevant video segment with a single sentence query. However, the majority of real-world languages consist of multiple sentences, and simply grounding sentence individually ignores rich contextual and semantic information within the sentences. As a remedy to these limitations, the concept of multi-sentence grounding (Huang et al. 2021; Bao, Zheng, and Mu 2021; Wang et al. 2021) has been introduced, which requires jointly localizing multiple sentences. Chen et al (Chen et al. 2022) identifies certain unrealistic assumptions prevalent in existing multi-sentence grounding methods: all query sentences can be grounded in the video, and query sentences are at the same semantic scales. Illustrated in Figure 1, the wiki article consists of high-level summaries, *e.g,* sentence (1) and corresponding low-level details, *e.g,* sentence [1.1]. Besides, some sentences, *e.g.*, sentence [1.3] has no corresponding video segments in the whole video. To address these two unrealistic assumptions, they introduce a more realistic and challenging grounding task: Temporal Article Grounding. To further alleviate the need for extensive manual annotations of the large-scale training set, they consider a more meaningful setting: Weakly-Supervised Temporal Article Grounding (WSAG), in which there are no temporal annotations for each sentence in the training data. As shown in Figure 1, given a multi-scale article and a relevant video, WSAG aims to localize corresponding video segments for all "groundable" sentences.

In addressing WSAG, Chen et al (Chen et al. 2022) introduce DualMIL, a method that extends multiple instance learning into a two-level framework encompassing "sentence-level" and "segment-level" representations. However, it overlooks the inherent hierarchical structures in the visual and text modalities and their complex correspondence, resulting in a loss of useful prior knowlege for precise grounding. As demonstrated in Figure 1, if we find that sentence (2) has a high similarity score with video segment $[5\tau - 9\tau]$, it is highly likely that sentence (2) also semantically related to sub video segment within it, *e.g.,* $[5\tau - 6\tau]$.

To be more general, a primary challenge inherent in WSAG is that the richness of video content and the variety of textual granularity lead to complex hierarchies of grounded video segments, which is intuitively illustrated in Figure 1. On the single-sentence level, the sentence may semantically correspond to multi-scale video segments, for instance, the broader temporal scope $[5\tau - 9\tau]$ encompasses multiple ground truth proposals related to sentence (2), while a nar-

*Corresponding author.

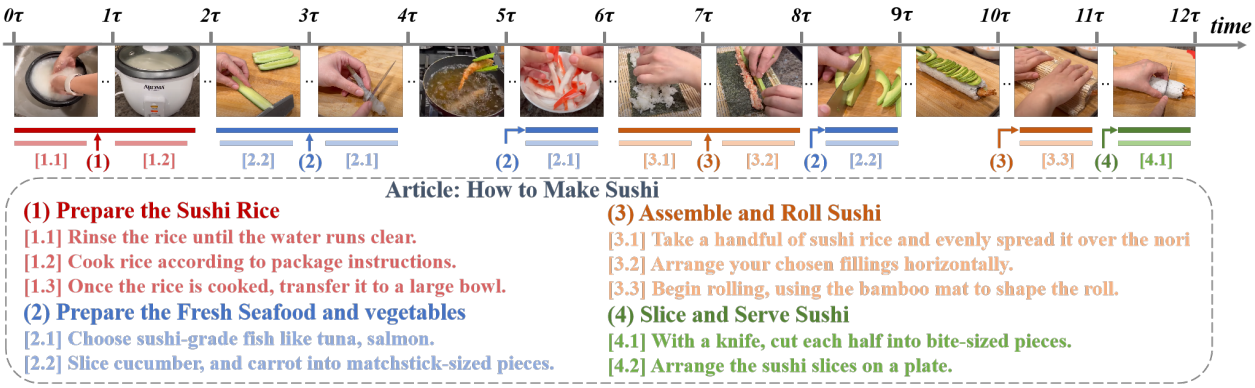

Figure 1: The overview of weakly supervised temporal article grounding (WSAG). WSAG aims to localize sentences in the article within a video. Sentences in the article exhibit distinct semantic scales, depicted in the figure using varying shades of color. Additionally, the color bar below the video corresponds to the ground truth localized segments for each color-coded sentence. The symbol 'τ' denotes the time interval between displayed video frames.

rower temporal scope $[5\tau-6\tau]$ in it contains a single ground truth proposal. On the multi-sentence level, the article's sentences exhibit multi-scale granularities, both semantically and temporally. For example, sentence [1.1] is the low-level detail of sentence (1) semantically, and $[0\tau-1\tau]$ (sentence [1.1]'s GT proposal) is included in a larger temporal scale $[0\tau-2\tau]$ (sentence (1)'s GT proposal) temporally.

Motivated by the above observations, we present a novel approach termed Multi-Scale Video-Text Correspondence Learning (**MVTCL**). MVTCL utilizes the visual content of different time spans and the structural relationship of text to obtain better video-article correspondence, and such perspective is manifested in both the network architecture and the supervisory information. In terms of network architecture, we introduce the concept of Multi-Scale Semantic Calibration (**MSC**), which capitalizes on preexisting knowledge about the aggregation of multi-scale visual and language information for every proposal. MSC comprises two primary procedures: visual content semantic integration and hierarchical language semantic suppression. Firstly, MSC integrates information from multiple temporal scales of video segments with sentences, thereby infusing prior visual knowledge spanning different scales into each proposal. After that, MSC introduces the prior information regarding the high-level sentence's correlation with various video segments into the corresponding low-level sentences. On the supervisory guidance, Multi-Scale Contrastive Loss (**MCL**) is introduced to select the most typical video segment proposals to represent the whole video when conducting contrastive inter-video learning. Specifically, MCL selects video segment proposals of multiple scales, ensuring that proposals within the same scale are non-overlapping and encompass the entirety of the video's content. The proposals selected by MCL not only have comprehensive multi-scale context aggregated by MSC but also reduce abundance due to its non-overlapping design.

Our contributions are summarized as follows:

- We delve into the intricate multi-scale structure present in both video and textual modalities within weakly supervised temporal article grounding. Then we introduce Multi-Scale Video-Text Correspondence Learning, which helps to achieve a more precise video-text correspondence within complex scenes.
- Experiments on WSAG datasets show that the proposed MVTCL outstrips existing WSAG methods by significant margins. Furthermore, our comprehensive ablation studies illuminate the individual efficacy of each component within MVTCL.

Related Work

Weakly Supervised Video Grounding Temporal sentence grounding is firstly introduced by MCN and Tall (Anne Hendricks et al. 2017; Gao et al. 2017), which aims to localize the target video segment from an untrimmed video with respect to a given sentence query. Then, due to the labor-intensive ground-truth annotation procedure, weakly supervised temporal video grounding become a popular and more practical setting.

Weakly supervised video grounding methods can be divided into two categories: single-sentence grounding and multi-sentence grounding. Early weakly supervised works (Chen et al. 2020; Song et al. 2020; Ma et al. 2020) mainly focus on single-sentence grounding. These methods can basically be grouped into two categories: MIL-based (Tan et al. 2021; Gao et al. 2019) and reconstruction-based (Lin et al. 2020; Zheng et al. 2022; Chen and Jiang 2021). TGA (Mithun, Paul, and Roy-Chowdhury 2019) is a typical MIL-based method, which treats the whole video as a bag of instances with bag-level annotations and the predictions for video segment proposals are aggregated as the bag-level prediction, then it learns the video-text alignment in the video-level by maximizing the similarity score of the matching video-text pair while minimizing the similarity score of video and other irrelevant text. As for reconstruction-based method, they attempt to reconstruct the given sentence query based on the selected video segments, and then use the reconstruction result as the supervision information.

All of the above weakly supervised methods stick to single-sentence grounding, however, a more realistic setting is to jointly ground multiple consecutive sentences. Some recent work like DepNet, WSTAN, CRM (Bao, Zheng, and Mu 2021; Wang et al. 2021; Huang et al. 2021) have explored such setting. Compared to single sentence grounding, these works usually utilize the relations between sentences to achieve higher accuracy in grounding. For example, CRM explores two cross-sentence relational constraints: temporal ordering and semantic consistency in a paragraph description of video activities. They claim that the temporal order of the events in the article and video is the same and the semantic information of related things in the article is consistent. However, such assumptions do not hold in complex scenes *e.g.,* article grounding.

Multi-Scale Video-Language Learning Many grounding methods (Hou et al. 2022; Bao, Zheng, and Mu 2021; Ding et al. 2022) utilize the multi-scale relations to get better grounding results, such relations mainly lie on video-level and language-level. For example, HSCNet (Tan et al. 2023) explores multi-level visual-textual correspondence by learning hierarchical semantic alignment and utilizes dense supervision by grounding diverse levels of queries including word-level, sentence-level and paragraph-level. Besides, other video understanding tasks such as moive understanding (Huang et al. 2020) and action recognition (Tang et al. 2019) have considered the multiple semantic scale issue.

However, all grounding methods neglect that there may exist different semantic-scale within the same level (such as sentence level). Chen et al (Chen et al. 2022) points out that all query sentences for the same video may have different semantic scales, thus propose a more challenging and realistic task: weakly supervised temporal article grounding (WSAG), which aims to localize the corresponding video segments for each "groundable" sentence in an article containing different semantic scale sentences.

Method

Problem Definition

Given an untrimmed video *V* and the relevant article *A* with multi-scale sentences, WSAG aims to jointly localize the temporal boundaries of events depicted by the sentences in the article. Specifically, each article is organized as $A = \left\{ s_1^h, s_1^{l_1}, ..., s_{n_1}^{l_1}; s_2^h, ...; s_m^h, ..., s_{n_m}^{l_m} \right\}$, where s_k^h is the *k*-th high-level summary, and $s_i^{l_k}$ is the *i*-th corresponding low-level details of s_k^h. Totally, there are *m* high-level summaries, and each summary s_k^h has n_k low-level details. WSAG needs to predict all possible temporal locations for all groundable sentences. The grounding result for high-level sentence s_k^h can be represented by $T(s_k^h) = \{(t_s, t_e)_i\}_{i=1}^{N(s_h^k)}$, where $(t_s, t_e)_i$ represents the starting time and end time of the *i*-th corresponding video segments for s_k^h, and $N\left(s_h^k\right)$ denotes the total video segments number of s_k^h. Due to the hierarchical relations between high-level summaries and low-level details, the grounding results of corresponding low-level sentence $T(s_i^{l_k})$ is the subset of $T(s_k^h)$ temporally.

Feature Extraction

Visual Encoding. For the untrimmed video *V*, we uniformly sample *N* short clips from it and each clip consists of a fixed number of consecutive frames. Then, we use standard backbone networks, *e.g.,* a frozen S3D (Xie et al. 2018) to extract clip-level features $\left\{f_i^V\right\}_{i=0}^{N-1}$, where $f_i^V \in R^{d^V}$ denotes the i-th clip feature and d^V is the feature dimension. Therefore, a candidate proposal can be identified by its start and end clips, specifically, m_{ij} represents the candidate proposal which starts from $clip_i$ and ends at $clip_j$, then all possible candidate proposals can be organized into a 2D temporal map *M* which consists of different time scales' candidate proposals.

Based on the 2D temporal map *M*, we extract each proposal feature by averaging all inside clip features, then a few conv-layers are used to further encode the context. Finally, we get the 2D visual feature map, denoted as $F^M \in R^{N \times N \times d^V}$, where d^V is the feature dimension. And the element in F^M is denoted as f_{ij}^M, which is the feature of candidate proposal m_{ij}.

Language Encoding. For each sentence $S_i = \left\{w_j^i\right\}$ in the article *A*, we first follow existing methods (Chen et al. 2022) to generate its textual embedding for each word by the GloVe (Pennington, Socher, and Manning 2014) word2vec model, then we sequentially feed the word embeddings into a bidirectional LSTM (Hochreiter and Schmidhuber 1997) network, and use its last hidden state as the feature representation of the input sentence, denoted as $F^{S_i} \in R^{d^S}$, where *i* is the position number of the sentence and R^{d^S} is the sentence feature dimension. After obtaining the video feature and all sentence features, both these features are fed into two respectively learnable MLP layers to get the same feature dimensions d^h, denoted as $\widetilde{F}^M \in R^{N \times N \times d^h}$ and $\widetilde{F}^{S_i} \in R^{d^h}$, and the element in $\widetilde{F}^M$ denotes as $\widetilde{f}_{ij}^M$.

Multi-Scale Semantic Calibration

We construct the Multi-Scale Semantic Calibration (**MSC**), which performs multi-scale semantic calibration on both visual and text modalities. Specifically, on the multi-scale visual information level, MSC interacts and fuses visual information from different scales through the Visual Content Aggregation Module. This enables the MSC to perceive both coarse-grained and fine-grained visual information. On the multi-scale language information level, MSC utilizes Hierarchical Language Semantic Suppression to exploit the hierarchical relationship that exists among multi-scale sentences both semantically and temporally, which effectively constrains the temporal distribution relationship of these two-scale sentences (the video segment corresponding to s_k^h include that of $s_i^{l_k}$)[1].

Visual Content Aggregation. To better utilize and aggregate visual context at different scales, we first select the most representative multi-scale visual features. In practice, we utilize three different granularities of visual features in

[1] s_k^h represents high-level summaries and s_k^h denotes its low-level details

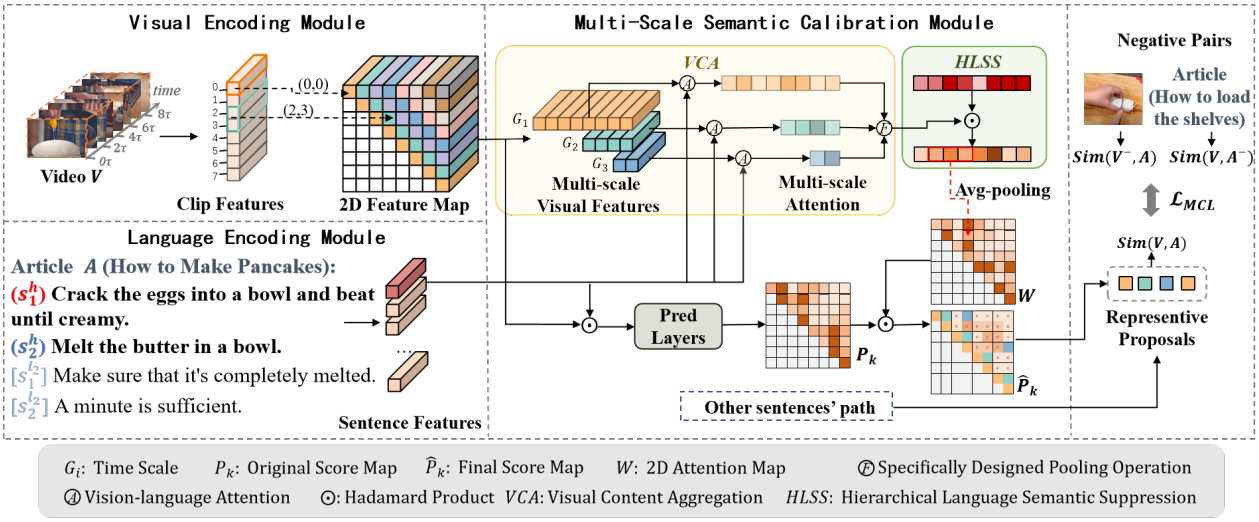

Figure 2: The overall architecture of our method. A 2D feature map is first extracted from input video, and the multi-scale visual features are selected to calculate vision-language attention with sentence features, then visual content aggregation is conducted through specifically designed pooling operation. After that, multi-scale language context is further utilized by hierarchical language semantic suppression to get the final attention map W. Given W, we conduct semantic calibration on original score map P_k to get the final score map $\hat{P}_k$, where P_k is obtained by basic multi-modal fusion operation. Finally, multi-scale contrastive loss is adopted to learn the most discrimitive features.

$\widetilde{F}^M$, denoted as $F^{G_i} \in R^{(N/G_i)\times d^h}$, where G_i represent different time scales and N/G_i represent the number of visual features at this scale. For each time granularity:

$$F^{G_i} = \left[f^M_{i(i+G_i-1)}\right]_{i=0,G_i,\ldots,N-G_i} \tag{1}$$

where [] represents the concatenation operation. Then pairwise semantic similarities across visual context at different scales are computed as:

$$Sim(S_i, G_j) = \frac{\widetilde{F}^{S_i} \cdot F^{G_j}}{\left\|\widetilde{F}^{S_i}\right\| \cdot \|F^{G_j}\|} \tag{2}$$

The final attention weight at scale G_j is obtained as below:

$$W^{S_i,G_j} = Softmax(Sim(S_i, G_j)) \tag{3}$$

In the end, we adopt a specifically designed average pooling operation[2] to get multi-scale visual context aggregated attention $W^{S_k} \in R^{1\times N}$.

Hierarchical Language Semantic Suppression. After modeling the multi-scale visual context, we further adopt hierarchical language semantic suppression to utilize the multi-scale language context. Specifically, the high-level summaries and low-level details in an article naturally have "hierarchical" relations. This relation manifests in two ways: 1) The high-level sentence contains corresponding low-level details semantically, 2) The video clip grounded by high-level summaries temporally includes the video clip grounded by corresponding low-level details. Due to such relation, our assumption is that if a candidate proposal has a low attention score with a high-level summary, then its attention score with corresponding low-level details should be low too. So we take the Hadamard product operation to complete semantic suppress:

$$\widehat{W}^{s_i^{l_k}} = W^{s_k^h} \odot W^{s_i^{l_k}} \tag{4}$$

Such operation introduces the prior knowledge of high-level sentence s_k^h into the corresponding low-level sentence $s_i^{l_k}$, effectively suppressing the occurrence of an unreasonable situation where a candidate proposal has a low attention score with s_k^h but a high attention score with $s_i^{l_k}$.

Semantic Calibration on Multimodal Score Map. At first, we acquire the original score map through multi-modal matching. Specifically, we fuse the mapped two features by Hadamard product:

$$f_{ij,k} = \widetilde{f}^M_{ij} \odot \widetilde{F}^{S_i} \tag{5}$$

then we reorganize the single element $f_{ij,k}$ into 2D format $F_k \in R^{N\times N\times d^h}$, which represents the fused feature between sentence S_k and all candidate proposals M. Then the fused features are fed into serval conv-layers and a classifier to predict the original score map $\{P_k\}$, where $P_k \in R^{N\times N}$ represents the original matching score between sentence S_k and all candidate proposals.

After obtaining the original score map, we calculate the element in 2D attention map as below:

$$W^{S_k}_{ij} = avgpool(\widehat{W}^{S_k}[i], \widehat{W}^{S_k}[i+1], ..., \widehat{W}^{S_k}[j]) \tag{6}$$

where $W^{S_k}_{ij}$ denotes the final attention score between sentence S_k and candidate proposal m_{ij}, then we reorganize $W^{S_k}_{ij}$ into 2D format $W^{S_k} \in R^{N\times N}$. Finally, we utilize the

[2]Specific procedures are detailed in the appendix.

aggregated multi-scale visual-linguistic information to calibrate the original score map:

$$\widehat{P}_k = W^{S_k} \odot P_k \tag{7}$$

where $\widehat{P}_k \in R^{N \times N}$ and element in $\widehat{P}_k$ is denoted as p_{ij}.

Multi-Scale Contrastive Loss

Previous methods (Zhang et al. 2020; Wang et al. 2021) often select high-score proposals from all candidates in the 2D score map to represent the whole video. However, we claim that employing the entirety of proposals is not efficient, as it encompasses a lot of redundant information, *e.g.,* numerous candidate proposals exhibit temporal overlaps, which could potentially impede the model's ability to learn the most effective representations through the process of contrastive learning. Based on such consideration, MVTCL utilizes multi-scale contrastive loss (MCL) to enhance WSAG's representation learning through the multi-scale and non-overlapping proposal selection strategy. Such design yields two advantages: 1) The selected proposals are consistent with those in MSC, which can effectively represent the whole video due to the semantic calibration before. 2) The absence of content overlap among proposals of the same scale reduces redundancy in information a lot.

Video-Article Similarity with Multi-Scale Structure. Given the score map $\widehat{P}_k$ after multi-scale semantic calibration, we first select the corresponding proposals at different scales. Similar to the operation in MSC, for each scale G_i, we choose the corresponding proposal set:

$$P_k^{G_i} = \left\{ p_{i(i+G_i-1)} \right\}_{0,G_i,\ldots,N-G_i} \tag{8}$$

where $P_k^{G_i}$ represent the set of representive elements at scale G_i and the set size is N/G_i. Then we use the above multi-scale proposal set to calculate the similarity score $Sim(V, A)$ between video *V* and article *A*. Firstly, we calculate the matching score $Sim(V, S_k)$ between each sentence and video, which is the average of similarity scores among the $top\text{-}k_2$ proposals at all scales:

$$Sim(V, S_k) = avg(top\text{-}k_2 \max_{ij} p_{k,ij}^{all}) \tag{9}$$

where $p_{k,ij}^{all}$ is the element of set at all scales $P_k^{all} = P_k^{G_1} \cup P_k^{G_2} \cup \cdots \cup P_k^{G_L}$, and *L* is the number of different scales. Since not all sentences in the article can be grounded in the video, we consider the average of similarity scores among the $top\text{-}k_1$ sentences as the matching score between the article and video:

$$Sim(V, A) = avg(top\text{-}k_1 \max_{k} Sim(V, S_k)) \tag{10}$$

Such video-article similarity score effectively captures the hierarchical structure present in video-text context.

Inter-Video Contrastive Loss. After obtaining the video-article matching score, we perform contrastive learning between positive and negative samples across videos. Specifically, we consider the matching video-article pair as positive pair *(V, A)*, then we randomly replace the video or article in the matching pair with video or article from other tasks[3] to get negative pair (V^-, A) and (V, A^-) respectively. The inter-video contrastive loss is calculated below:

$$\begin{aligned} \mathcal{L}_{MCL}^{ij} = &\max(0, \alpha - sim(V, A)_i + sim(V^-, A)_j) \\ &+ \max(0, \alpha - sim(V, A)_i + sim(V, A^-)_j) \end{aligned} \tag{11}$$

where *i,j* denotes pair index and the final inter-video contrastive loss $\mathcal{L}_{MCL} = \sum_i \sum_j \mathcal{L}_{MCL}^{ij}$.

Model Training and Inference

Training. To better stimulate the ability of each module, we adopt a multi-stage training strategy. In stage 1, we use inter-video contrastive loss without multi-scale structure $\mathcal{L}_{CL}$ to supervise training. The calculation of $\mathcal{L}_{CL}$ is similar to $\mathcal{L}_{MCL}$, the difference is that we use all candidate proposals instead of multi-scale proposals when calculating video-sentence similarity score. In stage 2, we use multi-scale contrastive loss $\mathcal{L}_{MCL}$ to continue training.

Inference. During inference, given the video and a relevant article, we first predict the semantic calibrated score map for each sentence and all candidate proposals, then we conduct non-maximum suppression (NMS) (Bodla et al. 2017) to filter out proposals which have high overlap with other proposals but lower scores. At last, we combine all proposals from different sentences based on their similarity score.

Experiments

Datasets

YouwikiHow. YouwikiHow consists of 47K untrimmed videos. Each video includes a corresponding multi-scale article consisting of high-level summaries and low-level details. YouwikiHow has a total of 1,398 wikiHow tasks, and each task has an average of 33.88 long-term videos. Besides, there are 20.8 sentences for each video on average. We use YouwikiHow as training dataset.

CrossTask. Since YouwikiHow has no time annotations, we use the same test set in DualMIL. Specifically, CrossTask (Zhukov et al. 2019) has 4.7K videos and 18 primary tasks. For each video, it has temporal boundaries corresponding to the predefined task-specific steps. Later, the step is linked manually to the wikiHow article.

Evaluation Metrics

Following the setting in previous work (Chen et al. 2022), we evaluate our model by computing *Recall@K* (**R@K**) over different IoU thresholds (0.1/0.3/0.5). It's defined as the recalls of all GT annotations within top-K candidate proposals based on given IoU and top-K candidate proposals are selected in all proposal-sentence pairs. Besides, we also use *Recall@K meet Constraint* **(RC@K)** as a supplementary metric for low-level sentences. The calculation is similar to **R@K**, except filter out proposals that don't meet the temporal constraint.[4]

[3]There are different tasks in the training data and video content across different tasks displays significant diversity.

[4]This constraint is the assumption about multi-scale sentences in WSAG: the temporal grounding results of low-level sentences should be inside its high-level manual annotations.

Model	MM Pretrain	R@50(IoU)			R@100(IoU)			RC@50(IoU)		
		0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
RandomGuess(Wang et al. 2021)	×	19.55	5.22	1.67	33.05	10.46	3.88	7.23	1.87	0.66
WSTAN-full(Wang et al. 2021)	×	27.22	15.98	7.42	36.52	20.97	9.92	13.75	10.59	9.25
WSTAN-base(Wang et al. 2021)	×	16.41	2.36	0.49	16.51	2.36	0.49	6.96	0.82	0.19
DualMIL(Chen et al. 2022)	×	40.11	23.08	10.07	54.32	31.30	13.97	10.71	8.09	6.96
MIL_NCE-max(Miech et al. 2020)	✓	33.48	12.01	4.89	39.71	14.30	5.87	11.85	3.12	1.06
MIL_NCE-avg(Miech et al. 2020)	✓	42.86	24.26	12.88	56.77	32.04	16.98	16.40	7.71	3.87
MIL-NCE+WSTAN	✓	32.30	18.03	8.61	50.30	28.81	14.04	13.44	10.49	9.09
MVTCL(Ours)	×	48.54	28.80	12.91	64.62	37.06	16.57	16.31	12.32	10.6

Table 1: erformance comparison with other methods. "MM Pretrain" denotes these models use large-scale multimodal pretraining features. Red text and blue text represent the best and second-best results, respectively.

Experimental Settings

For a fair comparison, we use pretrained S3D extractor (Xie et al. 2018) to extract video clip features on both datasets. The initial video clip number is set to 256, and the sampled clip number *N* is set to 16. The channel numbers of sentence feature d^S and video proposal feature d^V are set to 300 and 512 respectively, and the hidden features are set to 256.

During training, we use Adam (Kingma and Ba 2014) with an initial learning rate of 0.0003, the batch size of 128 as optimization algorithm. Besides, we adopt the ReduceLROnPlateau strategy to prevent overfitting. For each training sample, we randomly sample 20 sentences if the articles have more than 20 sentences, and we truncated or padded each sentence to a maximum length of 25 words. Such random sampling strategy saves GPU memory and reduces the model to overfit to moment prior of sentences in the article. For MSC and MCL module, the number of different scales *L* is set to 3. And the scale size of G_1, G_2, G_3 is set to 1,2,4 respectively. When calculating MCL, the selected number of proposals k_2 and sentences k_1 is set to 5 and 6 respectively, and the predefined margin α is set to 0.3.

Performance Comparisons

In this section, we compare our method with state-of-the-art method, including:

- **Weakly Supervised Temporal Grounding Methods.** WSTAN (Wang et al. 2021) is a weakly supervised paragraph grounding method, we compare three different settings of WSTAN: *Base, Full, RandomGuess*. DualMIL (Chen et al. 2022) is the previous state-of-the-art WSAG method.
- **Pretrained Large-Scale Multimodal Video-Text Retrieval Models.** MIL-NCE (Miech et al. 2020) is a large-scale pretrained multi-modal model for video-text retrieval. To apply MIL-NCE for WSAG, we use the same proposal generation strategy in this paper, then we adopt max-pooling or avg-pooling for clip features within the candidate proposals. Table 1 reports the *zero-shot* results using MIL-NCE.
- **Two-stage Method.** A characteristic of WSAG is that not all sentences can be grounded, so we first use MIL-NCE to retrieve all sentences that can be grounded for the video, the WSTAN is trained to select corresponding video segments.

Strategies		R@50 (IoU)			R@100 (IoU)		
MSC	MCL	0.1	0.3	0.5	0.1	0.3	0.5
×	×	26.60	14.98	6.48	44.05	24.81	10.82
✓	×	37.93	21.79	10.13	52.10	29.61	13.94
×	✓	36.74	16.77	8.22	53.15	24.57	12.31
✓	✓	**48.54**	**28.80**	**12.91**	**64.62**	**37.06**	**16.57**

Table 2: Ablations(%) on the effectiveness of each part

Visual Scale			SS	R@50 (IoU)			R@100 (IoU)		
G_1	G_2	G_3		0.1	0.3	0.5	0.1	0.3	0.5
✓	×	×	✓	34.99	18.37	9.20	42.6	21.75	11.24
×	✓	×	✓	29.32	15.70	8.26	40.17	21.04	11.00
×	×	✓	✓	32.32	18.27	8.38	45.84	25.44	11.61
✓	✓	×	✓	33.12	18.86	8.94	45.63	25.96	12.09
✓	×	✓	✓	32.62	17.86	8.60	46.16	24.73	11.96
×	✓	✓	✓	32.21	18.49	9.11	44.05	24.93	12.06
✓	✓	✓	×	27.39	15.64	7.49	42.23	24.47	11.79
✓	✓	✓	✓	**37.93**	**21.79**	**10.13**	**52.10**	**29.61**	**13.94**

Table 3: Ablations(%) on MSC, where "SS" denotes hierarchical language semantic suppress.

From the results in Table 1, we can draw following results: 1) Our proposed method outperforms all existing methods, including large-scale pretrained video-text retrieval models and before state-of-the-art WSAG method DualMIL. For example, DualMIL performs about 8 and 10 points lower than MVTCL in terms of R@50 (IoU=0.1) and R@100 (IoU=0.1) respectively. 2) Even though MIL-NCE gets relatively good *zero-shot* results for WSAG, it still has some limitations: it needs numerous training data and its performance is greatly affected by the clip features processing way.

Figure 3: Qualitative prediction examples. The first row shows the ground-truths for the given article queries (high-level sentences), and the second and third row shows the grounding results of DualMIL and our MVTCL in the corresponding color.

Ablation Study

In this section, we conduct comprehensive ablation studies to analyze the effectiveness of our proposed method.

Effectiveness of Each Strategy. Table 2 shows the performance comparisons of our proposed full model MVTCL with respect to main module ablations. MVTCL (full) outperforms all ablation models by a large margin, and two modules both contribute a lot to the improvement of model. This demonstrates that MSC and MCL both are critical to temporal article grounding .

Ablations of Multi-Scale Semantic Calibration (MSC). MSC contains two main parts: Visual Content Aggregation and Hierarchical Language Semantic Suppress. Table 3 reports the impact of each part and the different scales in visual content aggregation during stage 1. Firstly, we study the impact of visual content aggregation with hierarchical language semantic suppress, for three different granularity scales $\{G_1, G_2, G_3\}$, we explore the different combinations within them. We find that single-scale strategies basically perform worse than other fusion strategies, which is reasonable because such fusion way has less information than others. The model performs best when using all scales, which shows the effectiveness of the design of multi-scale visual context aggregation. Besides, we can see that the performance drops significantly without hierarchical language semantic suppress (penultimate line), which claims the importance of this part.

Ablations of Multi-Scale Contrastive Loss (MCL). The impact of combination of different scales is reported in Table 4. Similar to the ablation of MSA, both single-scale loss and the combination of two scales are explored. From Table 4, we have the following observations: 1) Within all possible combinations, the combination of all three scales $\{G_1, G_2, G_3\}$ performs best. 2) Within single-scale type loss, finer granularity leads to better results. This is reasonable because the most fine-grained video segment proposal usually contains intricate details for wach video when conducting inter-video contrastive learning.

Loss Scale			R@50 (IoU)			R@100 (IoU)		
G_1	G_2	G_3	0.1	0.3	0.5	0.1	0.3	0.5
✓	×	×	47.56	27.64	12.23	63.64	36.35	16.03
×	✓	×	44.91	21.2	10.67	56.59	26.93	13.66
×	×	✓	42.76	19.80	9.75	59.40	27.63	13.87
✓	✓	×	48.33	28.79	**13.00**	64.7	37.16	16.51
✓	×	✓	48.47	28.06	12.99	**64.86**	**37.50**	16.22
×	✓	✓	44.26	20.83	10.51	56.88	26.91	13.55
✓	✓	✓	**48.54**	**28.8**	12.91	64.62	37.06	**16.57**

Table 4: Ablations(%) on multi-scale contrastive loss.

Qualitative Analysis

We display a typical example of weakly supervised article grounding in Figure 3, which shows the grounding results of all high-level sentences. We can see that MVTCL have better grounding results than DualMIL, which can reflect in two aspects: 1) MVTCL can deal with more complex query sentences: for sentence "Pour tablespoons to the tip of a large spoon or from the hot griddle or greased frying pan.", DualMIL does not recall any of its GT video segments while MVTCL still performs well. 2) MVTCL can handle scattered multiple segments: for example, sentence "Cook the other side until golden and remove" has two corresponding video segments, DualMIL grounds it completely wrong while MVTCL also gives a good result. Such case proves the effectiveness of our well-designed MVTCL framework.

Conclusion

In this paper, we propose a novel Multi-Scale Video-Text Correspondence Learning framework to explicitly explore the complex multi-scale relations of language level and video level in a realistic grounding task WSAG. Extensive experiments validate the effectiveness of our MVTCL and demonstrate that our method achieves new state-of-the-art results on exsiting WSAG benchmarks.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No. 62206153, 62306031), and CCF-Tencent Rhino-Bird Open Research Fund.

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, 5803–5812.

Bao, P.; Zheng, Q.; and Mu, Y. 2021. Dense events grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 920–928.

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, 5561–5569.

Chen, L.; Niu, Y.; Chen, B.; Lin, X.; Han, G.; Thomas, C.; Ayyubi, H.; Ji, H.; and Chang, S.-F. 2022. Weakly-supervised temporal article grounding. *arXiv preprint arXiv:2210.12444*.

Chen, S.; and Jiang, Y.-G. 2021. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8425–8435.

Chen, Z.; Ma, L.; Luo, W.; Tang, P.; and Wong, K.-Y. K. 2020. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*.

Ding, X.; Wang, N.; Zhang, S.; Huang, Z.; Li, X.; Tang, M.; Liu, T.; and Gao, X. 2022. Exploring language hierarchy for video grounding. *IEEE Transactions on Image Processing*, 31: 4693–4706.

Gabeur, V.; Sun, C.; Alahari, K.; and Schmid, C. 2020. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 214–229. Springer.

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

Gao, M.; Davis, L. S.; Socher, R.; and Xiong, C. 2019. Wslln: Weakly supervised natural language localization networks. *arXiv preprint arXiv:1909.00239*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hou, Z.; Zhong, W.; Ji, L.; Gao, D.; Yan, K.; Chan, W.-K.; Ngo, C.-W.; Shou, Z.; and Duan, N. 2022. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. *arXiv preprint arXiv:2209.10918*.

Huang, J.; Liu, Y.; Gong, S.; and Jin, H. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7199–7208.

Huang, Q.; Xiong, Y.; Rao, A.; Wang, J.; and Lin, D. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 709–727. Springer.

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

Le, T. M.; Le, V.; Venkatesh, S.; and Tran, T. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9972–9981.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2019. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22(2): 554–565.

Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11539–11546.

Ma, M.; Yoon, S.; Kim, J.; Lee, Y.; Kang, S.; and Yoo, C. D. 2020. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 156–171. Springer.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9879–9889.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11592–11601.

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

Song, Y.; Wang, J.; Ma, L.; Yu, Z.; and Yu, J. 2020. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*.

Tan, C.; Lin, Z.; Hu, J.-F.; Zheng, W.-S.; and Lai, J. 2023. Hierarchical Semantic Correspondence Networks for Video Paragraph Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18973–18982.

Tan, R.; Xu, H.; Saenko, K.; and Plummer, B. A. 2021. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2083–2092.

Tang, Y.; Ding, D.; Rao, Y.; Zheng, Y.; Zhang, D.; Zhao, L.; Lu, J.; and Zhou, J. 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.

Wang, Y.; Deng, J.; Zhou, W.; and Li, H. 2021. Weakly supervised temporal adjacent network for language grounding. *IEEE Transactions on Multimedia*, 24: 3276–3286.

Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, 305–321.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12870–12877.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15555–15564.

Zhu, L.; Xu, Z.; Yang, Y.; and Hauptmann, A. G. 2017. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124: 409–421.

Zhukov, D.; Alayrac, J.-B.; Cinbis, R. G.; Fouhey, D.; Laptev, I.; and Sivic, J. 2019. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3537–3545.