Language-free Training for Zero-shot Video Grounding

Dahye Kim[1]   Jungin Park[1]   Jiyoung Lee[2]   Seongheon Park[1]   Kwanghoon Sohn[1,3*]

[1]Yonsei University   [2]NAVER AI Lab   [3]Korea Institute of Science and Technology (KIST)

{dadaday, newrun, sam121796, khsohn}@yonsei.ac.kr   lee.j@navercorp.com
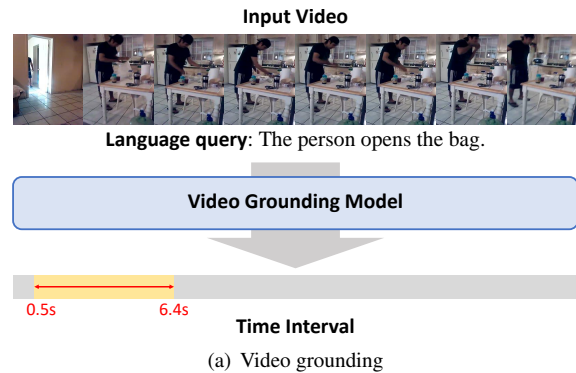
Abstract

Given an untrimmed video and a language query depicting a specific temporal moment in the video, video grounding aims to localize the time interval by understanding the text and video simultaneously. One of the most challenging issues is an extremely time- and cost-consuming annotation collection, including video captions in a natural language form and their corresponding temporal regions. In this paper, we present a simple yet novel training framework for video grounding in the zero-shot setting, which learns a network with only video data without any annotation. Inspired by the recent language-free paradigm, i.e. training without language data, we train the network without compelling the generation of fake (pseudo) text queries into a natural language form. Specifically, we propose a method for learning a video grounding model by selecting a temporal interval as a hypothetical correct answer and considering the visual feature selected by our method in the interval as a language feature, with the help of the well-aligned visual-language space of CLIP. Extensive experiments demonstrate the prominence of our language-free training framework, outperforming the existing zero-shot video grounding method and even several weakly-supervised approaches with large margins on two standard datasets.

1. Introduction

In our daily life, we surf, think, and learn through loads of videos. By extension, we wish to search for the information we want in the videos. Video grounding (also called video moment retrieval) with natural language query aims to help such video search by automatically localizing a temporal moment for various applications such as video surveillance [7] and smart video search [37, 38].

A major challenge of video grounding is the exorbitant cost of constructing time interval annotations aligned to a given text that is also collected. Although recent fully-supervised video grounding (FSVG) methods [24, 39]

---

*Corresponding author

Input Video

Language query: The person opens the bag.

Video Grounding Model

0.5s   6.4s

Time Interval

(a) Video grounding

Setting	Training		Test
	Time interval	Language query	Language query
FSVG	✓	✓	✓
WSVG	✗	✓	✓
ZSVG	✗	✗	✓

(b) Annotation types depending on the settings

Figure 1. Given a video and a language query, video grounding aims to retrieve the time interval corresponding to the language query in the video. In this paper, we address the zero-shot video grounding (ZSVG) problem which is the most challenging setting and cannot use any annotations for training.

have shown remarkable performance on the limited size of datasets [14, 19], there is still room for improvement with scale-up training. Especially in such a field, large-scale training data is required to cover numerous video domains (*e.g.*, instructional videos, movies, and so forth). However, building massive annotations as more billion scales like image-language datasets, such as LAION-5B [35], in video scale is an impractical solution.

To address the burden of annotations, researchers have proposed weakly-supervised video grounding (WSVG) methods [15, 23, 28] which use only coarse video-level descriptions for training. But they still require paired video-language data, showing limited applicability in the open world. Recently zero-shot video grounding (ZSVG) has been proposed in [30]. As illustrated in Fig. 1, ZSVG utilizes only videos to learn the video grounding model in the training stage. To learn the localizing capability in a semi-

supervised manner, [30] generates pseudo temporal event regions and corresponding pseudo sentence queries by examining noun-verb statistical co-occurrence patterns. However, pseudo sentences are built upon the composition of nouns and verbs (*e.g.*, 'flip person switch door'), which is naturally different from the form of natural language query (*e.g.*, 'person flipped the light switch near the door.'). Namely, contrived sentences with the simple composition of nouns and verbs break the structural and compositional generalization inherent in natural language that might harm the performance [22].

In this paper, we propose a novel language-free training framework for zero-shot video grounding. Our solution is to treat the visual feature as pseudo textual information while being flexible in responding to the act of forcing sentences to generate pseudo forced sentences. Specifically, we leverage an image-language pretraining model (*i.e.*, CLIP [33]) trained on large-scale web-collected data that have revealed a breakthrough in the multi-modal research field. We conjecture that text and visual features can replace each other without trouble in that CLIP provides a well-aligned visual-language semantic space.

To this end, we first generate temporal proposals that contain meaningful events from a given untrimmed video. With the visual encoder of CLIP, visual features are extracted from all the frames in the proposal. Then our learnable selection transformer takes a dominant feature that has a role of the pseudo language feature in a video grounding model instead of generating a natural sentence from the proposal. Therefore, our method is free to generate high-quality natural language form from the proposal. Moreover, since the dominant visual feature is directly used for the pseudo textual feature, our method has no need to produce textual embedding from a pseudo text label, which is a time-consuming yet necessary step for the training of the previous method [30]. Finally, the whole model is learned to predict time intervals corresponding to pseudo sentence features with generated temporal proposals as ground-truth. Our contributions are summarized three-fold:

- We introduce a language-free framework for video grounding that can be an affordable solution to effectively reduce the annotation cost.

- We validate the applicability of the pretrained visual-language model to the video-language task by providing extensive experimental analysis.

- Our language-free training framework outperforms the existing method, achieving state-of-the-art performance, and even shows comparable performance with weakly-supervised approaches on the Charades-STA [14] and ActivityNet Captions [19] datasets.

2. Related Work

2.1. Video Grounding

Video grounding is a recently proposed task [1, 14], which aims to find the best moment in a video grounded on language queries. Most of the existing methods followed fully-supervised setting [9, 22, 24, 29, 34, 44, 51, 52, 53, 54, 57, 59] to model fine-grained semantic relations of video and language. However, since such a setting requires precise annotations for the start and end timestamps, manual annotations of the temporal boundary were required, which also led to subjectivity across different annotators.

Weakly-supervised video grounding has been introduced to alleviate this burden. Existing works can be categorized into two groups. 1) Multi-instance learning [8] (MIL) based methods [15, 16, 27, 28, 41, 56] utilized similarity scores by maximizing scores between positive samples and minimizing scores between negative samples. 2) The reconstruction-based method [10, 23, 40, 50, 58] used the assumption that the video segment that best reconstructs the text query is close to the ground-truth.

However, while weakly-supervised approaches were successful in lowering the cost of temporal annotation, the cost of text query remains problematic. Several works [25, 30] considered an unsupervised setting that does not access the paired annotations. [25] proposed a deep semantic clustering network, which first aggregates distinct semantic features from the whole query set and then generates pseudo labels to provide pseudo supervision for training. [30] generated pseudo labels of temporal boundaries and corresponding query sentences. They first utilized a temporal similarity matrix to find temporal event proposals, then used an off-the-shelf object detector and fine-tuned RoBERTa [26] to make a structure-less pseudo query. However, a structure-less pseudo query, especially composed of nouns and verbs, can be interpreted in several meanings, due to systematic compositionality [5, 13] of natural language. In addition, the existence of the uninformative word in the query makes it hard for the model to distinguish the exact meaning of what the query originally intended to mean. Furthermore, inferred verbs from detected objects are loosely bonded in the sense that the verbs are not predicted directly from the video, which leads to the generation of inaccurate pseudo queries.

2.2. Language-free Paradigm

As recent trends shift from uni-modal learning to multi-modal learning, vision-language related tasks have attracted attention. Since the modality to be processed has doubled, it becomes difficult to obtain high-quality vision-language training pairs. Several works [30, 60] proposed a so-called 'language-free paradigm' to address this problem, which means training without language data in the vision-
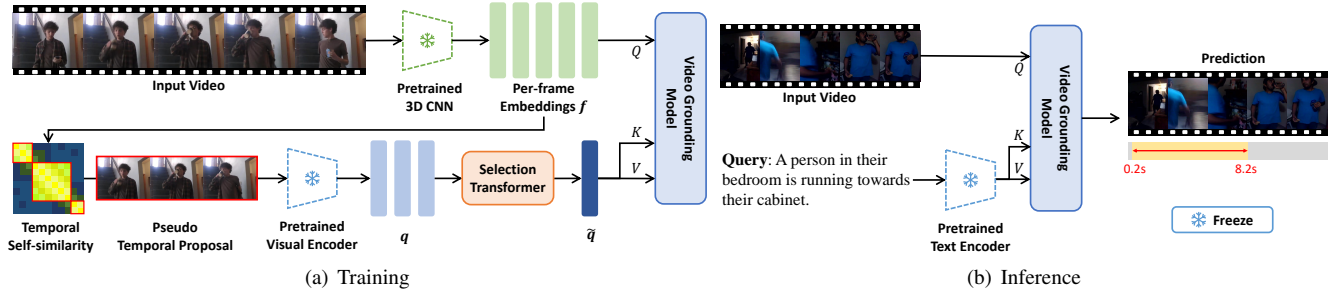
Figure 2. The overall framework of our language-free video grounding framework. In (a) training, we generate a pseudo temporal interval and corresponding pseudo language feature from the visual encoder of CLIP [33] and selection transformer to train the video grounding model. In (b) inference (test) phase, we use the video grounding model only with text encoder of CLIP.

language tasks.

One line of work [12, 17, 20, 30] presented a visual object-based approach, which utilizes an off-the-shelf object detector to make text-related pseudo labels based on detected objects. Unsupervised image captioning [12, 20] utilized an object detector to explore visual concepts in an image from unpaired image and text information. Unsupervised visual grounding [17] used detected objects as the first object proposals and then generated pseudo language queries with a pseudo-query generation module. Zero-shot video grounding [30] first detected objects from temporal event proposal as nouns, second utilized fine-tuned language model as verbs, and finally generated simplified sentence as pseudo query by composing nouns and verbs. However, the above methods heavily rely on the quality of recognized visual objects from object detectors, which has a large domain gap between the target dataset and the training dataset that the object detector has trained on. Furthermore, since the object categories were limited to the trained dataset, it was impossible to scale a wide variety of objects and rich expressions inherent in natural language [61].

Another line of work [47, 60, 61] utilized well-aligned multi-modal semantic space of the pretrained visual-language model. [61] presented prompt-based learning method for unpaired image captioning model, which utilizes the vision-language alignment established by CLIP [33]. [47, 60] proposed language-free text-to-image generation model using pretrained CLIP. Specifically, they generated a pseudo text feature directly from an image using CLIP, assuming that CLIP has learned image-text feature alignment in the joint embedding space. While we share the same spirit as the language-free text-to-image generation [60], our work is the first attempt to introduce language-free training for video grounding.

3. Language-free Video Grounding

3.1. Problem Statement and Motivation

Given an untrimmed video and a language query, video grounding aims to localize a time interval (start and end time stamps) representing the content corresponding to the query. In the zero-shot video grounding (ZSVL), the model is not allowed to access any language query and ground-truth time stamps during training. To achieve this goal, the prior work [30] generated a pseudo sentence query using a pretrained object detector and noun-verb statistics from text corpora. While they have successfully presented a baseline for zero-shot video grounding, there are still problems to be tackled: (1) they generated nouns for the pseudo query by heavily relying on the capacity of the pretrained object detector, which may have encoded inappropriate biases and has a limited number of object categories; (2) they trained the sentence query generation network and video grounding network separately, making the training procedure inefficient; (3) They assumed that simplified sentences consisting of nouns and verbs whose structural characteristics and compositional generalization inherent in natural language are ignored could be substitutes for natural language queries.

To solve the aforementioned problems, we propose a language-free framework for ZSVL, which skips doubtful sentence generation for performance improvement and lightweight training. As shown in Fig. 2(a), the training pipeline of our framework is (1) constructing temporal proposals using a pretrained video encoder, (2) generating the pseudo language feature with a selection transformer among the frame-wise visual features from pretrained CLIP, and (3) training a video grounding model that will be used to inference.

3.2. Temporal Proposal Generation

As a first step toward language-free video grounding, we should generate temporal event proposals from a video that we regard as temporal ground-truth. To detect events happening in the videos, we leverage a characteristic of visual similarity of consecutive frames. Specifically, a temporal similarity matrix is constructed to segment videos where visually similar frames are activated. Since the temporal similarity matrix reflects the temporal structure of the given

video [11, 31, 32], we utilize this information to find possible events occurring in the video.

Similar to [30], given raw video frames, we first extract the video feature from the sequence of segments using a pretrained video encoder $\mathcal{F}_v$. After obtaining extracted features f which encode the temporal structure of each segment, we construct a self-similarity matrix R of the given video as follows:

$$R_{ij} = cos(f_i, f_j) = \frac{f_i \cdot f_j}{\|f_i\|\|f_j\|}, \quad (1)$$

where R_{ij} is the cosine similarity score between pairs of segment features f_i and f_j. Then we group the segments into k dominant events by clustering the features using k-means algorithm. Also, consecutive events are merged to deal with more complex events.

3.3. Pseudo Language Feature Generation

Candidates of a language feature. To train the video grounding model, we need language queries corresponding to the generated temporal proposals. However, as mentioned in Sec. 3.1, creating a language query in a natural language form can neglect the natural property of the language and be erroneous and time-consuming. Instead, motivated by the recent success of the zero-shot text-to-image generation [60], we employ the visual encoder of the vision-language model (*i.e.* CLIP [33]) trained on large-scale image-language data using contrastive loss. Since the visual and language features are well-aligned in the semantic space, we can use the visual feature as the pseudo language feature.

Specifically, we randomly sample N frames denoted by $\{v_j\}_{j=1}^N$ in each temporal proposal and encode frame-wise features using the pretrained visual encoder $\mathcal{F}_{\text{img}}$. Thus, a set of candidates $\mathbf{q}$ for the pseudo language feature

$$\mathbf{q} = \{q_1, ..., q_N\} = \{\mathcal{F}_{\text{img}}(v_1), ..., \mathcal{F}_{\text{img}}(v_N)\}, \quad (2)$$

where q_n denotes the visual feature corresponding to v_n. However, directly using the visual feature may not be enough to represent the real language features. To this end, we intentionally perturb the features from the pretrained visual encoder using random noise following [60]:

$$q_n \leftarrow q_n + \xi\epsilon\|q_n\|_2/\|\epsilon\|_2, \quad (3)$$
$$q_n \leftarrow q_n/\|q_n\|_2, \quad (4)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the Gaussian noise, $\xi > 0$ denotes a hyperparameter to control the degree of noise, and $\|\cdot\|_2$ is L_2 normalization. We note that ViT/B-32 of CLIP image encoder is used as $\mathcal{F}_{\text{img}}$ in this work.

Pseudo language feature selection. Given encoded pseudo language feature candidates, we select a single dominant feature that is the most informative to represent the

corresponding temporal proposal. While it is natural to encode temporal information in video-language tasks, we select the pseudo language feature without the temporal modeling. Our observation is that a single dominant visual feature can be more informative to represent the corresponding query for two main reasons: 1) A video consists of consecutive frames that usually contain similar semantics from a continuous scene so that sampling a superior frame already contains important information of the video [2, 21]; 2) since a video is a collection of noisy frames due to the existence of background clutter or camera motion blur, the combination of sampled frames may contain uninformative information and be computationally inefficient.

Moreover, inappropriate temporal modeling harms the vision-language semantic space, leading to an unreliable performance at inference time where a real language query is given. One alternative solution is leveraging a pretrained video-language model (*e.g.* VideoCLIP [49]). However, the video-language model is usually pretrained on a smaller number of video-language pairs (1.1M videos in VideoCLIP [49]) than the visual-language model (400M image-text pair in original CLIP [33]). Furthermore, video-language models typically require high computation and memory costs. Therefore, we insist that incorporating the visual-language model into our work efficiently leverages the confident visual-language semantic space. We will verify this observation in Sec. 4.5.

Concretely, we formulate a selection transformer that has only simple two transformer layers for a frame selection process such that:

$$\text{ST}(\{q_1, q_2, \ldots, q_N\}) \mapsto \tilde{q}, \quad (5)$$

where ST is the selection transformer and $\tilde{q}$ denotes the pseudo language feature. To ensure the back-propagation of such transformer for an end-to-end training, we employ gumbel softmax similar to [2].

3.4. Video Grounding Model

In this section, we describe our video grounding model consisting of a video encoder and a cross-modality fusion module that learns to fuse two distinct modality features.

Video encoding. We reuse the obtained video feature f in the video grounding model with temporal positional encoding. As our goal is to regress the temporal boundaries, it is important to embed the position information.

To explicitly model the position information of each video, we apply temporal positional encoding e_{pos} of each segment as done in [43]. Then we apply bi-directional GRU [6] to further encode temporal information. A final representation of the video s is obtained by aggregating the vector that concatenates the last hidden layer of bi-directional

Method	Sup.	Charades-STA				ActivityNet Captions				
		R@0.3	R@0.5	R@0.7	mIoU	R@0.1	R@0.3	R@0.5	R@0.7	mIoU
LGI [29]	FS	72.96	59.46	35.48	51.38	-	58.52	41.51	23.07	41.13
CTRL [14]	FS	-	21.42	7.15	-	49.1	28.70	14.00	-	20.54
TGA [28]	WS	29.68	17.04	6.93	-	-	-	-	-	
CTF [4]	WS	39.8	27.3	12.9	27.3	74.2	44.3	23.6	-	32.2
SCN [23]	WS	42.96	23.58	9.97	-	74.48	47.23	29.22	-	-
WSTAN [45]	WS	43.39	29.35	12.28	-	79.78	52.45	30.01	-	-
BAR [48]	WS	44.97	27.04	12.23	-	-	49.03	30.73	-	-
MARN [40]	WS	48.55	31.94	14.81	-	-	47.01	29.95	-	-
CCL [56]	WS	-	33.21	15.68	-	-	50.12	31.07	-	-
LoGAN [41]	WS	51.67	34.68	14.54	-	-	-	-	-	-
CRM [16]	WS	53.66	34.76	16.37	-	81.61	55.26	32.19	-	-
VCA [46]	WS	58.58	38.13	19.57	38.49	67.96	50.45	31.00	-	33.15
LCNet [50]	WS	59.60	39.19	18.87	38.94	78.58	48.49	26.33	-	34.29
RTBPN [55]	WS	60.04	32.36	13.24	-	73.73	49.77	29.63	-	-
CNM* [58]	WS	60.39	35.43	15.45	-	78.13	55.68	33.33	-	-
DSCNet [25]	US	44.15	28.73	14.67	-	-	47.29	28.16	-	-
PSVL* [30]	ZS	46.17	31.29	14.17	31.24	-	44.74	30.08	14.74	29.62
Ours*	ZS	**52.95**	**37.24**	**19.33**	**36.05**	**61.35**	**47.61**	**32.59**	**15.42**	**31.85**

Table 1. Performance comparison with other methods on the Charades-STA and the ActivityNet Captions dataset. 'Sup.' refers to supervision level: WS (Weakly-supervised setting), US (Unsupervised setting, where query information utilized but not paired to videos), ZS (Zero-shot setting, where any annotation are not exploited including query information) * These works use pretrained models: ours and [58] use frozen CLIP, and [30] fine-tune RoBERTa [26].

GRU and the positional encoded video feature as follows:

$$s = \text{MLP}[\text{Bi-GRU}(\hat{f}) \oplus \hat{f}], \qquad (6)$$

where $\oplus$ is a concatenation operation and $\hat{f} = f + e_{\text{pos}}$ is a video feature that combines positional embeddings.

Cross-modality fusion module. Given obtaining the pseudo language feature $\tilde{q}$ and the encoded whole video feature s, video grounding aims to find the most related parts in the video corresponding to the given language feature. To achieve the goal, we leverage an attention mechanism proposed in [43] to enable the multi-modal interaction of the two modalities. Specifically, we obtain language-guided video feature s_{att} using multi-head attention where we denote query Q as the video feature f, key K and value V as the pseudo language feature $\tilde{q}$:

$$\text{Cross-Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}V),$$
$$(7)$$

where d_k is the dimension of K. Then, to capture more global context across the video, we additionally apply a self-attention layer after the cross-attention layer. We carefully note that cross-attention and self-attention have different roles in the fusion module, where key, query, and value in the self-attention layer are the video attention feature, and the key and value of the cross-attention layer is the pseudo

language feature. Finally, with an MLP layer, we predict the start and end time stamps of the most relevant temporal region from the condensed video feature. This process is summarized as follows:

$$(\hat{t_s}, \hat{t_e}) = \text{MLP}(\text{Self-Attention}(s_{att})), \qquad (8)$$

where $(\hat{t_s}, \hat{t_e})$ is the predicted start and end time, respectively.

3.5. Model Training and Inference

Since our method performs video grounding in the zero-shot setting, the training and inference processes are different as illustrated Fig. 2. Next, we describe training objectives to learn the video grounding model with the pseudo temporal proposal and the pseudo language query, and the inference process with the given video and real language query.

Training. Our training objective includes two loss functions, the temporal regression loss L_{reg} and the temporal attention calibration loss L_{att}:

$$L = L_{reg} + \lambda L_{att}. \qquad (9)$$

To balance each objective term, the hyper-parameter λ is used. Note that we empirically select λ to 1, which has shown less effect on training.

L_{reg}	L_{att}	R@0.3	R@0.5	R@0.7	mIoU
✓	✗	45.16	30.40	14.88	30.33
✗	✓	12.81	8.71	3.99	8.71
✓	✓	**52.95**	**37.24**	**19.33**	**36.05**

Table 2. The ablation study of different losses. ✓ means the loss term used in training.

Following previous works [29, 51], we adopt temporal regression loss L_{reg} as a smooth L1 loss between model prediction and target interval, which is given by

$$L_{reg} = smooth_{L_1}(\hat{t}_s - \tilde{t}_s) + smooth_{L_1}(\hat{t}_e - \tilde{t}_e), \quad (10)$$

where $(\tilde{t}_s, \tilde{t}_e)$ and $(\hat{t}_s, \hat{t}_e)$ denotes pseudo temporal ground-truth and model prediction, respectively.

We also adopt temporal attention calibration loss L_{att} to increase the accuracy of temporal attention since we directly regress the time intervals from temporally attended video features following [51]:

$$L_{att} = -\frac{\sum_{t=1}^{T} \tilde{a}_t \log(a_t)}{\sum_{t=1}^{T} \tilde{a}_t}, \quad (11)$$

where

$$\tilde{a}_t = \begin{cases} 1, & \text{if } \tilde{t}_s \leq t \leq \tilde{t}_e \\ 0, & otherwise. \end{cases} \quad (12)$$

Inference. Different from the training process, in the inference stage, an input is a video and its corresponding complete sentences from the test set. To deal with this difference, we extract text features from the text encoder of the pretrained vision-language model, *i.e.* text encoder of CLIP [33]. In other words, the pseudo language feature $\tilde{q}$ is replaced with the real language feature q from the real language query. Hence, our proposal generation step in Sec. 3.2 and pseudo language feature generation step in Sec. 3.3 are only leveraged to train the video grounding model.

4. Experimental Results

4.1. Datasets

In order to verify the effectiveness of our method, we conduct experiments on two datasets: Charades-STA [14] and ActivityNet Captions [19]. Since we formulate the video grounding task as a language-free setup, any annotations related to the videos have not been utilized while training, but in the test only.

Charades-STA Charades-STA was introduced by [14] from the Charades dataset [36] with the purpose of evaluating on video grounding task by annotating in a semi-automatic way. The dataset contains 12,408/3720 segment-sentence pairs and 5338/1334 videos in training and test set, respectively.

Model	R@0.3	R@0.5	R@0.7	mIoU
Ours	52.95	37.24	19.33	36.05
Ours + *temporal GT*	54.00	39.91	19.46	36.29

Table 3. Upper bound analysis using ground-truth temporal boundaries (*temporal GT*). With *temporal GT*, we directly generate the pseudo language features corresponding to GT time intervals.

Frame Selection	R@0.3	R@0.5	R@0.7	mIoU
Random	50.2	34.84	15.66	33.49
ST	**52.95**	**37.24**	**19.33**	**36.05**

Table 4. The ablation study of frame selection strategies. 'ST' denotes the proposed selection transformer.

ActivityNet Captions ActivityNet Captions was originally collected by [19] for evaluating dense video captioning, which contains 37,417/17,505/17,031 segment-sentence pairs and 10,009/4917/5044 videos in training, val_1 and val_2, respectively. Following previous works [29, 30], we evaluate our performance on the validation set since the annotation of the test set is unavailable.

4.2. Evaluation Metric

To evaluate the performance of our model, we adopt R@tIoU and mIoU (mean averaged tIoU) following previous works [14, 23] for a fair comparison. Specifically, given predicted boundaries, we compute temporal intersection over union (tIoU) with ground-truth boundaries. R@tIoU is the percentage of the predictions which are larger than the thresholds, *i.e.* {0.3, 0.5, 0.7}. mIoU is the average IoU of all the predictions.

4.3. Implementation Details

For a fair comparison, we employ I3D [3] and C3D [42] networks as video feature extractor for the Charades-STA and ActivityNet Captions datasets, respectively, following previous works [29, 30]. We set the maximum length T of the video features to 128 in both datasets. For generating pseudo language features, we use pretrained CLIP-ViT/B-32. We set $N = 9$ for frame sampling and use a low-capacity transformer [43] with 2 layers and 2 attention heads for the frame selection process, making it computationally efficient. The bidirectional GRU layers in the video encoder are 2 layers architecture with a hidden size of 256. For the cross-modality fusion module, we use multi-head attention with 3 layers and 4 heads. The dimension of their hidden state is 256. For the hyperparameters, we empirically set $k = 5$, $\xi = 0.0001$ and $\lambda = 1$. In all experiments, we train our models with a batch size of 256 using Adam [18] with a fixed learning rate of 0.0004. We provide more details in supplementary material and the code will be publicly available soon.
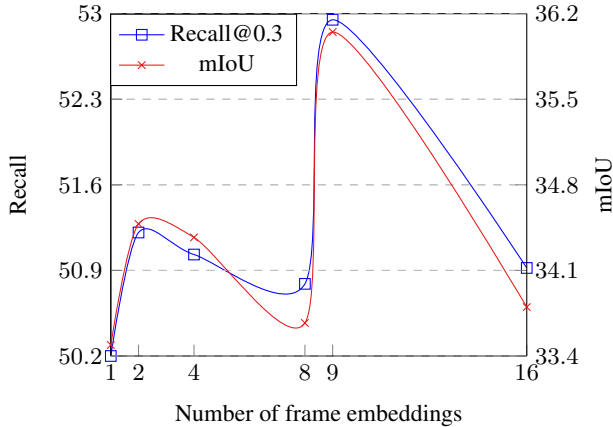
Figure 3. The ablation study of the number of frame embeddings used in frame selection process.

4.4. Comparisons to the State-Of-The-Art

Tab. 1 shows the results of our model compared to previous works in fully-supervised, weakly-supervised, unsupervised, and zero-shot conditions. The weakly-supervised (WS) methods are trained with costly annotated sentence queries, whereas the unsupervised (US) method leverages the unpaired data of videos and sentence queries in the dataset. However, zero-shot methods, including ours, leverage only videos in the dataset for training. On both Charades-STA and ActivityNet Captions datasets, we can observe that our method outperforms PSVL [30] in all metrics by large margins, demonstrating the robustness of the proposed method. Furthermore, even though our method does not use a bunch of language queries of the dataset, our method outperforms the unsupervised method [25] by a large margin. The comparisons with the weakly-supervised methods show that our method achieves comparable or even superior performance to several approaches [4, 23, 28, 40, 41, 45, 48, 56].

4.5. Analysis

To prove the excellence of our methods, we perform ablation studies and analysis from various perspectives on the Charades-STA.

Effects of different losses. We first investigate the effectiveness of using different loss terms, L_{reg} and L_{att}. As shown in Tab. 2, our model performs best when we used all loss terms, which demonstrates that using two losses is critical for training our network. We also find that the regression loss L_{reg} had a more influence on overall performance, however, training with only regression loss L_{reg} find to be inferior to the performance of baseline.

Upper bound analysis. In Tab. 3, we give the upper bound analysis to our model by replacing the pseudo tem-

Query: a person wearing a blue sweater opens a coat closet.
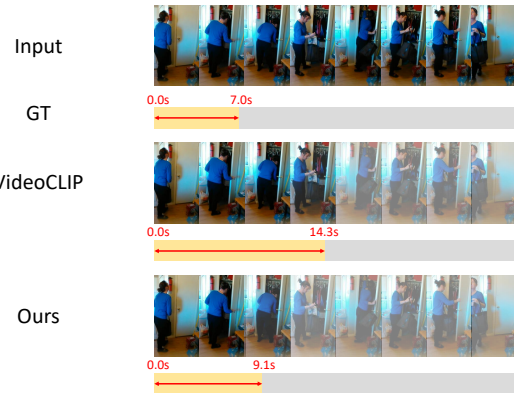
Query: a person is throwing a pillow towards the window.

Figure 4. Qualitative comparisons corresponding to the language feature encoders on the Charades-STA dataset.

poral proposals $(\tilde{t}_s, \tilde{t}_e)$ into temporal ground-truth (t_s, t_e). Replacing with temporal ground-truth leads to performance improvement which outperforms most of the existing weakly-supervised video grounding methods. However, the gain was not significant because we obtained a pseudo language feature by selecting one of the frames within the generated temporal boundaries. We carefully assume that the precise temporal location has a limited impact.

Effectiveness of the selection transformer. To investigate the importance of using the selection transformer in the pseudo language feature generation process, we replace it with a random selection module. In this analysis, we randomly sample a feature from extracted visual features in the proposal as a pseudo language feature. As shown in Tab. 4, we observe that using a selection transformer can boost the performance, suggesting the selection transformer's capability of selecting a dominant feature.

Effect of the number of frame embeddings. As shown in Fig. 3, we evaluate the effectiveness of the number of frame embeddings used in the selection transformer. We can see that the more we sample the frame embeddings, the higher the tIoU scores until the 9 frames. So, we set the $N = 9$ in all experiments. More results for the recall at different tIoU are shown in the supplementary material.
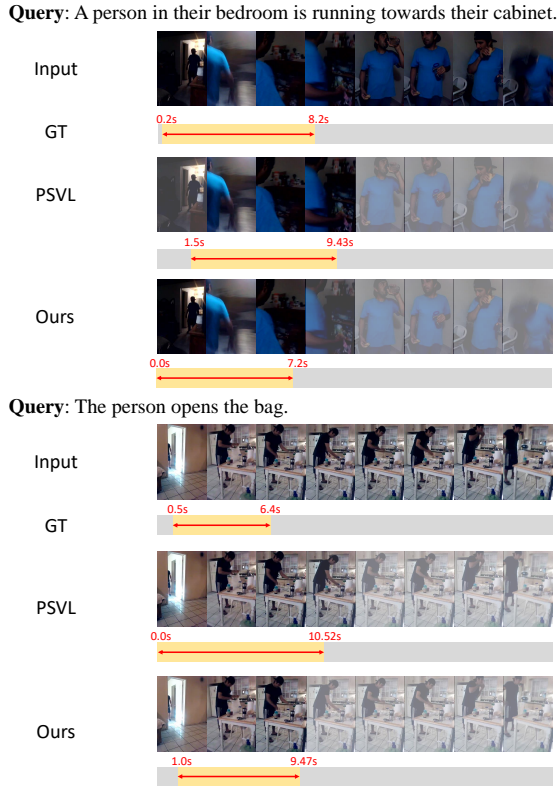
Query: A person in their bedroom is running towards their cabinet.

Query: The person opens the bag.

Figure 5. Qualitative comparisons between ours and PSVL on the Charades-STA dataset.

Effectiveness of the image-based vision-language model. In this section, we investigate the effectiveness of the image-based vision-language model for the video-language task. For this experiment, we employ pretrained video-language model [49], which has established fine-grained associations between video and text with contrastive loss, to generate a pseudo language feature. The pseudo language feature is directly obtained from the video-language model extracted from the proposal. Fig. 4 shows some qualitative results for the comparison between our method with CLIP and with its counterpart (i.e., VideoCLIP). As shown in Fig. 4, our model with CLIP can localize the better moment than with VideoCLIP, regardless of whether the given query is more static or dynamic. We observe that using an image-language model can capture semantic information from a single frame comparable to or better than the video-language model.

4.6. Qualitative Results

Fig. 5 shows some qualitative results comparing our results with the previous method [30] on the Charades-STA dataset. This example presents the temporal ground-truth boundaries and model predictions of PSVL [30] and ours, given a pair of a video and a query. The results show that the proposed method covers more of the video content related to the query, which effectively shows that our model

Query: A female child is playing with a hula hoop.
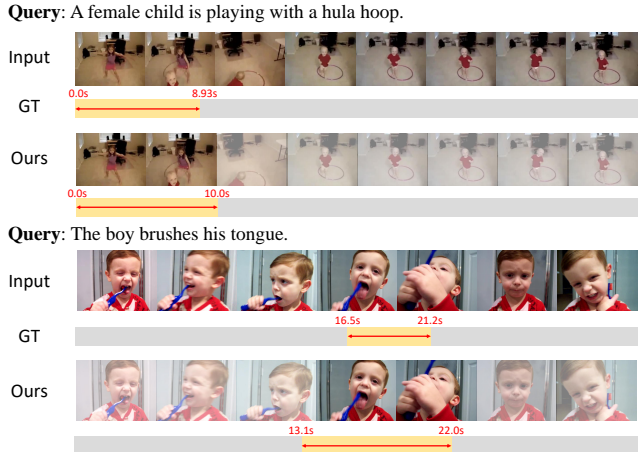

Query: The boy brushes his tongue.

Figure 6. Qualitative comparisons between ground-truth intervals and ours on the ActivityNet Captions dataset.

is qualitatively better than PSVL. Also, Fig. 6 illustrates the qualitative results on the ActivityNet Captions dataset. The more qualitative results are in the supplementary material.

5. Conclusion and Future Work

In this work, we present a novel method to train a video grounding model in a zero-shot manner without using any annotation related to paired video-sentence data. We achieve the goal by generating pseudo ground-truth of temporal locations and corresponding text features with the language-free paradigm. Primarily, we obtain a pseudo language feature from a generated proposal leveraging the well-aligned visual-language semantic space of CLIP. In contrast to the previous method of trying to make pseudo text queries into contrived language formats, we preserve the structural characteristics and compositional generalization inherent in natural language. Moreover, we develop a video grounding model based on cross- and self-attention transformers to effectively model the relationship between two modalities and the context of attended features. The experimental results demonstrate the efficacy of language-free training, achieving remarkable performances on two datasets and reducing the cost of data collection.

However, the temporal modeling is not designed in this work due to the aforementioned reasons in Sec. 3.3. As shown in the experimental results, current datasets show a limitation for temporal reasoning in the settled query. As the next step, we will investigate the new benchmark for video grounding, which should include more hard examples of causal and temporal understanding as well as more long-term videos for practical usage.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. *ICCV*, 2017.

[2] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. *CVPR*, 2022.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *CVPR*, 2017.

[4] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*, 2020.

[5] Noam Chomsky. Syntactic structures. *De Gruyter Mouton*, 2009.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[7] Robert T Collins, Alan J Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, Osamu Hasegawa, Peter Burt, et al. A system for video surveillance and monitoring. *VSAM final report*, 2000.

[8] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 1997.

[9] Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. Support-set based cross-supervision for video grounding. *ICCV*, 2021.

[10] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. *NeurIPS*, 2018.

[11] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. *CVPR*, 2020.

[12] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. *CVPR*, 2019.

[13] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 1988.

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. *ICCV*, 2017.

[15] Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong. Wslln: Weakly supervised natural language localization networks. *EMNLP*, 2019.

[16] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. *ICCV*, 2021.

[17] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. *CVPR*, 2022.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. *ICCV*, 2017.

[20] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. *ICCV*, 2019.

[21] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *CVPR*, 2021.

[22] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. *CVPR*, 2022.

[23] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. *AAAI*, 2020.

[24] 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. *CVPR*, 2021.

[25] Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. *AAAI*, 2022.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[27] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. *ECCV*, 2020.

[28] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. *CVPR*, 2019.

[29] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. *CVPR*, 2020.

[30] Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. Zero-shot natural language video localization. *ICCV*, 2021.

[31] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Sumgraph: Video summarization via recursive graph modeling. *ECCV*, 2020.

[32] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction networks for video question answering. *CVPR*, 2021.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

[34] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. *WACV*, 2020.

[35] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[36] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *ECCV*, 2016.

[37] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. *ICCV*, 2003.

[38] Cees Snoek, Kvd Sande, OD Rooij, Bouke Huurnink, J Uijlings, M van Liempt, M Bugalhoy, I Trancosoy, F Yan, M Tahir, et al. The mediamill trecvid 2009 semantic video search engine. *TRECVID workshop*, 2009.

[39] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. *ICCV*, 2021.

[40] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020.

[41] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. *WACV*, 2021.

[42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. *ICCV*, 2015.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[44] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. *CVPR*, 2021.

[45] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. Weakly supervised temporal adjacent network for language grounding. *IEEE TMM*, 2021.

[46] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. *ACM MM*, 2021.

[47] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. Clip-gen: Language-free training of a text-to-image generator with clip. *arXiv preprint arXiv:2203.00386*, 2022.

[48] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. *ACM MM*, 2020.

[49] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *EMNLP*, 2021.

[50] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE TIP*, 2021.

[51] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. *AAAI*, 2019.

[52] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. *CVPR*, 2020.

[53] 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. *CVPR*, 2019.

[54] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. *AAAI*, 2020.

[55] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. *ACM MM*, 2020.

[56] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *NeurIPS*, 2020.

[57] Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. Cascaded prediction network via segment tree for temporal video grounding. *CVPR*, 2021.

[58] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. *AAAI*, 2022.

[59] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and debias for robust temporal grounding. *CVPR*, 2021.

[60] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *CVPR*, 2022.

[61] Peipei Zhu, Xiao Wang, Lin Zhu, Zhenglong Sun, Weishi Zheng, Yaowei Wang, and Changwen Chen. Prompt-based learning for unpaired image captioning. *arXiv preprint arXiv:2205.13125*, 2022.