Generating Structured Pseudo Labels for Noise-resistant Zero-shot Video Sentence Localization

Minghang Zheng[1], **Shaogang Gong**[2], **Hailin Jin**[3], **Yuxin Peng**[1,4], **and Yang Liu**[1,5*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Queen Mary University of London, [3]Adobe Research
[4] National Key Laboratory for Multimedia Information Processing, Peking University
[5] National Key Laboratory of General Artificial Intelligence, BIGAI
{minghang, pengyuxin, yangliu}@pku.edu.cn
s.gong@qmul.ac.uk, hljin@adobe.com

Abstract

Video sentence localization aims to locate moments in an unstructured video according to a given natural language query. A main challenge is the expensive annotation costs and the annotation bias. In this work, we study video sentence localization in a zero-shot setting, which learns with only video data without any annotation. Existing zero-shot pipelines usually generate event proposals and then generate a pseudo query for each event proposal. However, their event proposals are obtained via visual feature clustering, which is query-independent and inaccurate; and the pseudo-queries are short or less interpretable. Moreover, existing approaches ignores the risk of pseudo-label noise when leveraging them in training. To address the above problems, we propose a **S**tructure-based **P**seudo **L**abel generation (SPL), which first generate free-form interpretable pseudo queries before constructing query-dependent event proposals by modeling the event temporal structure. To mitigate the effect of pseudo-label noise, we propose a noise-resistant iterative method that repeatedly re-weight the training sample based on noise estimation to train a grounding model and correct pseudo labels. Experiments on the ActivityNet Captions and Charades-STA datasets demonstrate the advantages of our approach. Code can be found at https://github.com/minghangz/SPL.

1 Introduction

Video sentence localization, which aims to localize the most salient video segments from an untrimmed video given a free-form nature language query, has attained increasing attention due to its potential applications in video surveillance (Collins et al., 2000), robot manipulation (Kemp et al., 2007), etc. The free-form natural language queries allow the model to be flexibly adapted to the requirements of different practical applications.

*Corresponding author

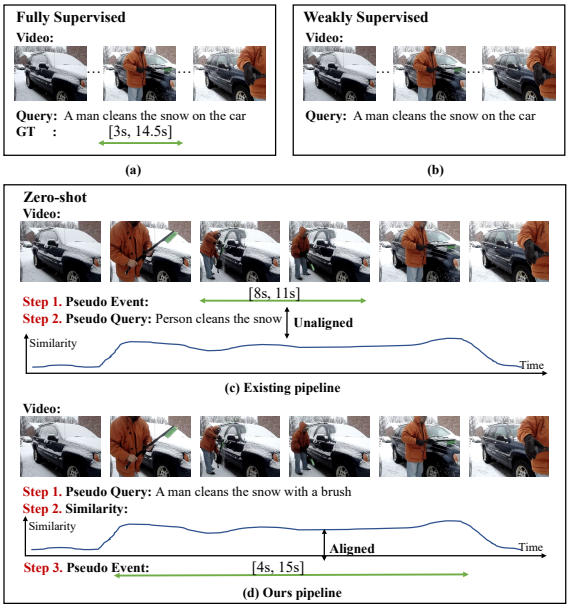

Figure 1: (a) Training data for fully-supervised models. (b) Training data for weakly-supervised models. (c) The zero-shot models are trained with videos only. Existing pipeline may generate unaligned pseudo event-query pairs. (d) We construct query-dependent event proposals by modeling the event temporal structure.

In recent years, the performance of video sentence localization has been improved with the help of advanced deep learning techniques and massively annotated data. However, the high annotation cost and the annotation bias still prevent the practical application of these models. On the one hand, the process of generating descriptions for the events in the video and labeling the corresponding events with the exact start and end timestamps are labor-intensive. On the other hand, many methods tend to capture the annotation bias (both in the query and timestamps) in the dataset, thus affecting the robustness of these models (Yuan et al., 2021; Otani et al., 2020). As shown in Figure 1, although the weakly supervised approaches do not require the timestamps annotation, the annotation costs of natural language queries are still unavoidable and

they still suffer language-related annotation bias (e.g. query style and structure, etc). Therefore, in this work, we study the video sentence localisation in a zero-shot setting, i.e. only video data is needed for training without any manual annotation[1].

Existing zero-shot video sentence localization approaches (Nam et al., 2021; Wang et al., 2022; Kim et al., 2023) follow the same pipeline, i.e. looking for event proposals in the video, and then generating pseudo queries for the events. They either construct a simple subject-verb-object pseudo query by detecting possible verbs and nouns in the video or directly use the CLIP (Radford et al., 2021) features of video frames to serve as the query text features, assuming the visual and text feature spaces are well aligned. However, there are three problems in this pipeline. Firstly, their pseudo queries are either too simple (simple subject-verb-object structure) or less interpretable (only given as features), which makes it potentially difficult to generalize the model to the real queries. Besides, they usually generate nouns and verbs by pre-trained object detectors or image-text pre-trained models, where temporal structured information is absent. Secondly, as shown in Figure 1(c), though they encourage the pseudo queries to have high semantic relevance to the proposal, they ignore the pseudo queries might also have a high score to the time-span out of the proposal, leading to miss-alignment between the pseudo queries and proposals, which may result in the model learning the incorrect visual and text alignment. Thirdly, existing methods train the model directly using pseudo-labels, ignoring the risk of noise in the generated start and end timestamps. They may fit the noise during training, resulting in poor test performance.

To tackle these problems, we propose a novel **S**tructure-base **P**seudo **L**abel generation pipeline (SPL) to generate flexible and generalizable pseudo-labels and reduce the noise in the pseudo-labels during training. Firstly, to generate free-form pseudo-queries, we sample video frames and generate captions using a pre-trained image caption model. The queries from the caption model are more diverse and flexible than those simple subject-verb-object pseudo queries. Secondly, to generate reasonable events for pseudo-queries, we consider the temporal structure of an event, i.e. the relevance between the query and the content in the event should be high, while the relevance outside the event should be low. Specifically, we enumerate event proposals and select the one with the largest gap between the semantic relevance to the query within the event and outside the event, and use the gap value as the quality of the pseudo query. To prevent too many queries from describing the same event, we use non-maximum suppression to filter out the pseudo-queries whose events have a high IoU with others and keep the top-K pseudo-query-event pairs based on their quality. Finally, to mitigate the effect of pseudo-label noise when training a fully supervised model using our pseudo-query-event pairs, we propose a noise-resistant iterative method. We repeatedly re-weight each training sample based on our noise estimation from the model's prediction, and continuously refine the temporal labels during training. Our pipeline shows significant performance advantages on the Charades-STA and ActivityNet Captions datasets.

Our contributions are: (1) We propose a novel model learning process for zero-shot video sentence localization, which generates free-form pseudo query candidates first, and then generates pseudo events according to the temporal structure of an event. (2) We propose a sample re-weight and pseudo-label refinement method to reduce the effect of pseudo-label noise on the model. (3) Experiments on Charades-STA and ActivityNet Captions demonstrate the advantages of our method.

2 Related Works

2.1 Fully/Weakly Supervised Video Temporal Localization

The fully supervised methods(Gao et al., 2017; Wang et al., 2021; Zhao et al., 2021; Zhou et al., 2021; Huang et al., 2022; Zhang et al., 2020, 2021; Zheng et al., 2023) usually train a model with the annotations of start and end timestamps for each video and query. However, the high cost of manual annotation limits the scalability of fully supervised methods. Moreover, as studied in (Yuan et al., 2021; Otani et al., 2020), the annotation bias in the dataset may also affect the robustness of these models. To reduce the annotation cost, the weakly supervised methods (Lin et al., 2020; Zheng et al., 2022b,a; Yang et al., 2021; Huang et al., 2021; Mithun et al., 2019) train the model with only the videos and annotated queries. However, the weakly

[1]In this work, we follow the definition of 'zero-shot video sentence localization' in previous works (Nam et al., 2021; Wang et al., 2022; Kim et al., 2023), which may be different from the zero-shot setting in other tasks.

supervised methods still suffer the language-related annotation bias, and the annotation costs of natural language queries are also unavoidable. Therefore, in this work, we study the video sentence localization using only video data (without any manual annotation), which is more practical but also more challenging.

2.2 Zero-shot Video Temporal Localization

In the zero-shot setting, only the video data are required during training. Existing zero-shot methods (Nam et al., 2021; Kim et al., 2023; Wang et al., 2022; Gao and Xu, 2021) follow the same pipeline, i.e. search event proposals in the video, and then generate pseudo queries for the events. PSVL (Nam et al., 2021) first discovers the temporal event proposals and then generates simplified pseudo queries by detecting nouns in the video and discovering appropriate verbs with those nouns. Gao et al. (Gao and Xu, 2021) directly generate pseudo query features in the pre-trained visual language feature space. However, the pseudo queries in existing methods are either too simple or less interpretable. Besides, the existing pipeline does not take the temporal structure of an event into account, which may lead to unaligned pseudo-events and queries. Moreover, they ignore the risk of pseudo-label noise when leveraging them in model training. In this paper, we generate free-form interpretable pseudo queries and construct query-dependent event proposals by modeling the event temporal structure and propose a noise-resistant method to mitigate the effect of pseudo-label noise.

2.3 Learning with Noisy Labels

Many works have explored how to train models with noisy labels on the tasks such as image classification (Han et al., 2018; Li et al., 2020), object detection (Li et al., 2020, 2022b), et al. Some approaches correct the noisy labels by learning from a small set of clean samples (Xiao et al., 2015; Veit et al., 2017), or learning with hard or soft labels using the model predictions (Tanaka et al., 2018; Yi and Wu, 2019; Li et al., 2020, 2022b). Some approaches re-weight or select training samples by estimating the noise in each sample (Li et al., 2020; Arazo et al., 2019; Chen et al., 2019). Existing noisy label image classification methods mostly assume noisy labels in different pixels are i.i.d, which is not realistic in the video sentence localization task, where pseudo label noise is likely to be introduced near the boundary of the events. To the best of our knowledge, we make the first attempt to reduce label noise introduced by pseudo labels in the video sentence localization task by iterative sample re-weighting and pseudo-label refinement.

3 Approach

The overview of our model design is illustrated in Figure 2. Our method is divided into four steps. In the first step, we generate pseudo queries for a given video. To obtain realistic free-form nature language queries, we sample video frames and generate captions using a pre-trained image caption model, which will serve as our pseudo query candidates. In the second step, we generate pseudo-event proposals for each pseudo query. As the events described by the query should have a certain structure, i.e. the relevance to the query in the event should be high, while the relevance to the query outside the event should be low, we calculate the similarity between each query candidate and each video frame. Then, for each pseudo query, we select the best event proposal with the largest gap between the correlation within the event and the correlation outside the event, and the gap is served as the quality of that pseudo query. As a good pseudo query should not be too general (e.g. 'there is a person' is a bad query), it should be significantly more relevant to the corresponding event in the video than other video segments. Thus, in the third step, we will only keep the top-k high-quality proposal-query pairs. In the last step, we will train a fully supervised model using the filtered pseudo-query-event pair. To reduce the noise in the pseudo-labels, we propose to estimate the noise and then re-weight each sample, while refining the pseudo-labels during training.

3.1 Pseudo Query Generation

In this step, we will generate free-form natural language queries based on the video. The pseudo queries in previous works are usually too simple (simple subject-verb-object structure) or unspecified (only given as features), which have a large gap between the real nature language queries. Thus, we propose to generate free-form nature language queries using a pre-trained image caption model based on the video frames.

Specifically, given a video V, we first uniformly sample N frames $v_1, v_2, ..., v_N$. Then, we use a pre-trained BLIP model (Li et al., 2022a) to gener-

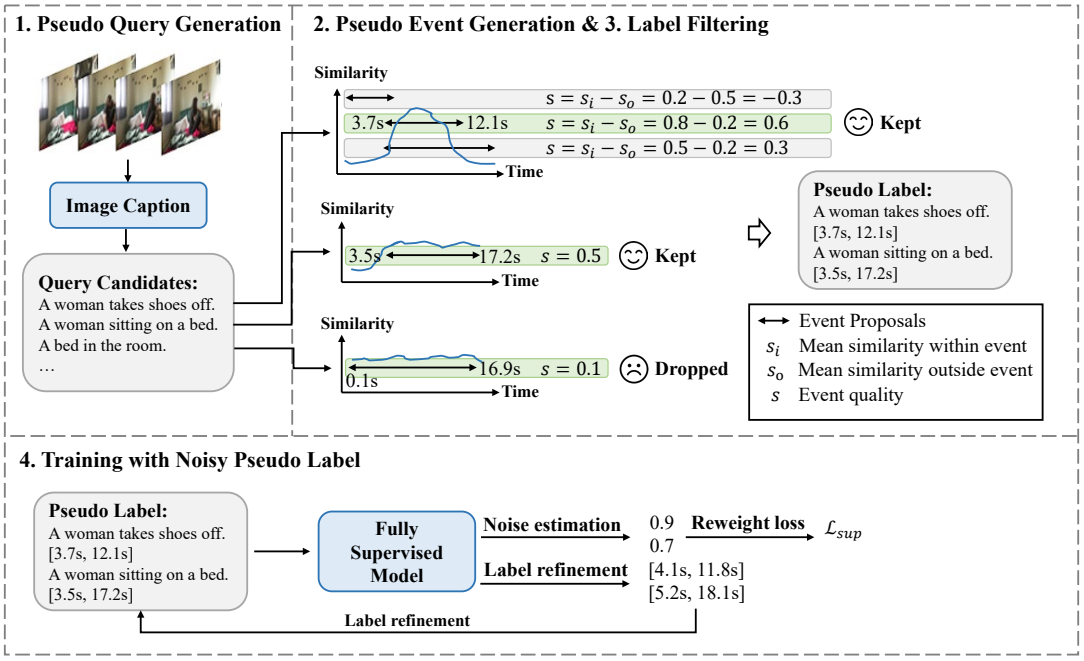

Figure 2: SPL overview. We first generate free-form pseudo queries by generating captions using a pre-trained image caption model. Then, we generate pseudo event proposals for each pseudo query and filter the pseudo query-event pairs by modeling the event temporal structure. Finally, we train a fully supervised model using the filtered pseudo query-event pair, and propose a sample re-weight and pseudo label refinement method to mitigate the effect of pseudo-label noise.

ate captions for each frame. As a video frame may be rich in content, we generate multiple queries for the same frame to ensure that the description of the frame is as complete as possible. Then, the captions $c_1, c_2, ..., c_M$ serve as our pseudo query candidates, where M is the number of captions in the video. Note that in this step, M will usually be large in order to ensure that the candidate queries contain as many meaningful queries as possible. However, this can also lead to a large number of low-quality queries in the candidates, which will be filtered out in the method described in Sec. 3.3

3.2 Pseudo Event Generation

In this step, we generate pseudo-events (i.e. start and end timestamps) for each pseudo-query candidate. Existing methods usually generate query-independent pseudo-events first, and then generate pseudo-queries for those pseudo-events. They ignore the temporal structure of real events, i.e. video within the event should be highly correlated with the query, while video outside the event should be lowly correlated with the query. Therefore, we take full account of in-event and out-of-event relevance to the query to produce high-quality pseudo-events.

Specifically, for each pseudo-query candidate $c_1, ..., c_M$, we use pre-trained BLIP text encoder to extract text features $F^c = [f_1^c, ..., f_M^c] \in \mathbb{R}^{M \times D}$, where D is the feature dimension. Then, for each video frames $v_1, ..., v_N$, we can also use the BLIP image encoder to extract image features $F^v = [f_1^v, ..., f_N^v] \in \mathbb{R}^{N \times D}$. As the BLIP text and image feature space are well aligned, We can directly use the cosine similarity of the text and image features to measure the relevance of the query and the video frame:

$$S = \frac{F^c F^{v\intercal}}{\|F^c\| \|F^v\|} \in \mathbb{R}^{M \times N} \quad (1)$$

We believe that the most relevant event for a given query should satisfy the requirement that videos within the event have a high relevance to the query and videos outside the event have a low relevance to the query. Therefore, we use the sliding window to enumerate the possible event proposals $p_1, p_2, ..., p_{N_p}$, where N_p is the number of event proposals. Then, we calculate the average similarity within each event and the average similarity outside each event, and use the difference between them as the quality for each event proposal:

$$Q_{ik} = \frac{1}{\|p_k\|} \sum_{j \in p_k} S_{ij} - \frac{1}{N - \|p_k\|} \sum_{j \notin p_k} S_{ij} \quad (2)$$

where Q_{ik} is the quality of the k-th event proposal

Algorithm 1: Pseudo label generation

Input : Training videos
Output : Pseudo query-event pairs

```
for each training video do
    Generate image captions for video frames using BLIP model
    Calculate the similarity between captions and video frames by Eq.(1)
    for each pseudo query (caption) do
        Calculate event quality by Eq.(2)
        Keep the best event by Eq.(3)
    for each the query-event pairs do
        Calculate query quality by Eq.(4)
        Keep top-K query-event via NMS
```

to the i-th query candidate, S_{ij} is the relevance of the i-th query and the the j-th frame, and $\|p_k\|$ is the number of frames in the event proposal p_k.

Finally, to ensure that the event to each query is unique, we select the highest quality event proposal as the pseudo-event label for the i-th query:

$$e_i = p_{\hat{k}}, \hat{k} = \arg\max_k Q_{ik} \tag{3}$$

3.3 Label Filtering

Due to the uneven quality of the large number of pseudo query-event pairs, we will filter them further. We believe that a good query-event pair should not be too general, so the relevance to the video within the corresponding event should be as high as possible, while the relevance to the video outside the event should be as low as possible. This means that the quality of the best event proposal for each query candidate in Eq.(2) can also be used to evaluate the quality of that query-event pair.

Specifically, we define the quality of the i-th query-event pair as:

$$Q_i^c = max_{k=1}^{N_p} Q_{ik} \tag{4}$$

We do not want too many queries describing the same event in the video, so we will further filter out those query-event pairs whose events have a high IoU between others using Non-maximum suppression. Finally, we will keep the top-K query-event pairs in order of quality Q^c for a video.

We summarise our pseudo label generation pipeline including the pseudo event generation and label filtering in Algorithm 1.

3.4 Training with Noisy Pseudo Label

In this step, we can use the generated pseudo-queries with their corresponding events to train any of the fully supervised video sentence localization models. Considering the performance, we chose the recent open-source model EMB (Huang et al., 2022). EMB conducts a proposal-based video-text alignment first, and then constructs elastic boundaries with the timestamps between the predicted endpoints and the manually labeled endpoints. EMB requires the model to select the endpoints in these elastic boundaries and thus models the uncertainty of the temporal boundaries.

However, most of the existing fully supervised models are designed for clean training data and may not be robust enough for pseudo-labels that contain a lot of noise. Therefore, we design a sample re-weight and label refinement method to reduce the effect of label noise on the fully supervised model.

Sample Re-weight. It has been shown neural networks are trained to fit clean data first and then to fit the noise (Han et al., 2018; Yu et al., 2019). Therefore, the confidence of a model in its prediction can reflect the noise in the sample. That is, if the model is more confident in its predictions and the predictions are close to the training labels, there is relatively less noise in the data.

Specifically, we use the video-text matching score given by EMB between its prediction and pseudo query as the confidence s_i^{conf} for the i-th training sample. Then, we calculate the interaction-over-union (IoU) between the prediction and pseudo label s_i^{iou}. The higher s^{conf} and s^{iou}, the lower the noise, and the greater the weight of the training sample should be. Therefore, we define the sample weights as:

$$w = \alpha \frac{1}{1 - s^{iou}} + (1 - \alpha) \frac{1}{1 - s^{conf}} \tag{5}$$

where α is a hyper-parameter to balance the effects of s^{iou} and s^{conf}. Finally, we use the same loss function as EMB and re-weight different samples loss using w. By sample re-weight, we can reduce the negative impact of noisy labels, but we prefer that noisy labels also provide useful training signals, so we further propose a label refinement method to correct the noisy labels.

Label Refinement. Since the pseudo-events may not be accurate enough, we design a label refinement procedure, so that the model can update higher-quality pseudo-labels during training.

During training, if the model is confident in its prediction, it is possible that the prediction is the true label. Besides, we also believe that the true label should not differ too much from the pseudo-label, so we will also consider the IoU between the prediction and the pseudo-label to prevent the model from being overconfident in the wrong prediction. Specifically, we can obtain the visual-text matching scores s_k^m for the k-th proposal to the query in EMB as well as its IoU s_k^{iou} with the pseudo label. We will select the $\hat{k}$-th proposal as the refined pseudo-label for the next epoch model training, where $\hat{k} = \arg\max_k(\beta s_k^m + (1-\beta)s_k^{iou})$ and β is a hyper-parameter. In this way, if the model has sufficient confidence in the prediction of the correct label, it is possible to refine the noisy label to the correct one.

The overall loss function is formulated as:

$$\mathcal{L} = \sum_{i=1}^{B} w_i \mathcal{L}_{loc}(V_i, c_i, \hat{e_i}) \tag{6}$$

where w_i is the weight for the i-th pseudo query-event pair for training, V_i is the video, c_i is the pseudo query, $\hat{e_i}$ is the refined pseudo event, $\mathcal{L}_{loc}$ is the localization loss function used in EMB (Huang et al., 2022), and B is the batch-size.

4 Experiments

To evaluate our method, we conduct experiments on the Charades-STA (Gao et al., 2017) and ActivityNet Captions (Krishna et al., 2017) dataset.

4.1 Datasets

ActivityNet Captions. ActivityNet Captions (Caba Heilbron et al., 2015; Krishna et al., 2017) was originally collected for video captioning, which contains 20K videos. There are 37,417/17,505/17,031 video-query pairs in the train /val_1/val_2 split. We follow previous works and report the performance on the val_2 split.

Charades-STA. Charades-STA (Gao et al., 2017) was built upon the Charades dataset. There are 12,408/3,720 video-query pairs in the train/test split. We report the performance on the test split.

4.2 Evaluation Metrics

We follow the evaluation metrics 'R@m' and 'mIoU' in the previous work (Nam et al., 2021), where m is the predefined temporal Intersection over Union (IoU) threshold. In particular, 'R@m' means that the percentage of predicted moments that have the IoU value larger than m. 'mIoU' represents the average Intersection over Union.

4.3 Implementation Details

We use the BLIP model(Li et al., 2022a) to generate captions for the video. We sample an image every 8 and 16 frames and use BLIP to generate 10 and 5 captions for each image on the Charades-STA and ActivityNet Captions datasets respectively. For each video, we only keep the top-10 and top-5 pseudo queries for Charades-STA and ActivityNet Captions datasets respectively. We train the EMB (Huang et al., 2022) model using our pseudo labels and keep the training hyper-parameters consistent. The hyper-parameters in sample re-weight and label refinement are $\alpha = \beta = 0.75$.

4.4 Comparison with Other Methods

Table 1 shows the performance comparison of our SPL to other methods on Charades-STA and ActivityNet Captions datasets respectively. As we can see, on the Charades-STA dataset, we led in all metrics, e.g. the mIoU is 4.42% higher than the second place (Kim et al., 2023). On the ActivityNet Captions dataset, we obtained the best performance for R@0.3 and mIoU. On the other hand, we outperform some of the weakly supervised methods without using any human annotation, proving the quality of the pseudo-labels we generated.

4.5 Experiments on Annotation Bias

In Table 2, we empirically investigate how the performance of different methods are affected by the annotation bias on the Charades-CD dataset (Yuan et al., 2021). Charades-CD re-partitioned the Charades-STA dataset to obtain the test_iid (independent and identically distributed (IID)) and test_ood (out-of-distribution (OOD)) splits.

As we can see, the fully supervised method EMB (Huang et al., 2022) shows a significant drop (7.79%) in performance on test_ood split, which indicates that EMB relies on the annotation bias in the training data. Our method is not affected by the annotation bias, and hence there is no significant drop in performance on the test_ood split. As the Charades-CD dataset is constructed considering only the bias in the timestamps, the degradation of the weakly supervised method CPL (Zheng et al., 2022b) is also not significant, but their overall performance is worse even with the help of annotated queries. This proves the quality of the pseudo-labels we generated.

Method	Sup.	Charades-STA				ActivityNet Captions			
		R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
2D-TAN (Zhang et al., 2020)	fully	-	39.81	23.25	-	58.75	44.05	27.38	-
EMB (Huang et al., 2022)		**72.50**	58.33	39.25	**53.09**	**64.13**	44.81	26.07	**45.59**
MGSL-Net (Liu et al., 2022)		-	**63.98**	**41.03**	-	-	**51.87**	**31.42**	-
CRM (Huang et al., 2021)	weakly	53.66	34.76	16.37	-	55.26	32.19	-	-
CNM* (Zheng et al., 2022a)		60.39	35.43	15.45	-	55.68	**33.33**	-	-
CPL (Zheng et al., 2022b)		**66.40**	**49.24**	**22.39**	-	**55.73**	31.37	-	-
Gao et al.* (Gao and Xu, 2021)	no	46.69	20.14	8.27	-	46.15	26.38	11.64	-
PSVL* (Nam et al., 2021)		46.47	31.29	14.17	31.24	44.74	30.08	14.74	29.62
PZVMR* (Wang et al., 2022)		46.83	33.21	18.51	32.62	45.73	31.26	**17.84**	30.35
Kim et al.* (Kim et al., 2023)		52.95	37.24	19.33	36.05	47.61	**32.59**	15.42	31.85
SPL* (ours)	no	**60.73**	**40.70**	**19.62**	**40.47**	**50.24**	27.24	15.03	**35.44**

Table 1: Evaluation Results on the Charades-STA Dataset and ActivityNet Captions Dataset. *These works use pre-trained models: ours uses BLIP (Li et al., 2022a), CNM, PZVMR, and Kim et al. use CLIP (Radford et al., 2021), PSVL fine-tune RoBERTa (Liu et al., 2019), Gao et al. uses VSE++ (Faghri et al., 2017).

Method	Sup.	mIoU		
		iid	ood	drop
EMB (Huang et al., 2022)	fully	**55.44**	**47.65**	7.79
CPL (Zheng et al., 2022b)	weakly	35.29	33.28	2.01
SPL (ours)	no	41.32	39.61	**1.71**

Table 2: Experiment on annotation bias on Charades.

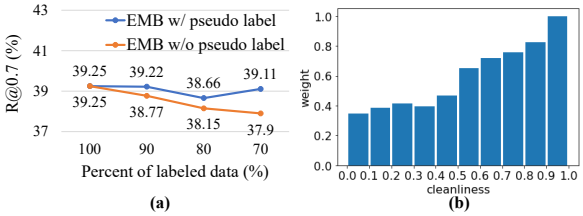

Figure 3: (a) Experiment of reducing annotation cost on the Charades-STA dataset. (b) The average weights (scaled to [0, 1]) assigned to the samples with different cleanliness.

Event	Query	Model	R@0.5	R@0.7	mIoU
PSVL	PSVL	PSVL	31.29	14.17	31.24
PSVL	PSVL	Ours	29.62	15.70	33.45
PSVL	Ours	Ours	36.94	19.30	38.31
Ours	Ours	Ours	**40.70**	**19.62**	**40.47**

Table 3: Compare with PSVL (Nam et al., 2021)'s pipeline on the Charades-STA dataset.

4.6 Reducing Annotation Cost

Our pseudo-label generation method can reduce the cost of manual annotation. In practice, considering the balance between performance and annotation cost, we can manually annotate a portion of data and use our generated pseudo-labels for the remainder. In Figure 3(a), We train a fully supervised model using partially annotated data and augment missing data with our generated pseudo-labels. As we can see, supplementing data with pseudo-labels improves performance compared with training without pseudo-labels, and when only using 70% of the manually annotated data, the model performance drops by just 0.14%. This shows the practical application of our approach in reducing annotation costs and improving annotation efficiency.

4.7 Ablation Studies

To verify the effectiveness of our method, we conduct ablation studies on the Charades-STA dataset.

Compare with existing pipeline. In Table 3, we compare the pipeline used in exsisting methods and our method. We train our localization model with PSVL (Nam et al., 2021)'s and our pseudo queries and events respectively. As we can see in Table 3, (1) even with the same query from PSVL, our noise-robust localization model still show clear performance advantages; (2) our pseudo queries and temporally structured events demonstrate significant performance improvements, which proves the effectiveness of our pipeline. Besides, we calculate the variances of pseudo-query features in our method and PSVL. The variances are 0.88 and 0.67 respectively, which demonstrates our pseudo-queries are more flexible and diverse.

Effectiveness of pseudo event generation. Table 4 shows the performance of different ways of generating pseudo-events. 'Naive' means randomly generating events; 'Expand' means expanding the frame where the query is generated until the similar-

Event generation	R@0.5	R@0.7	mIoU
Naive	28.31	11.99	33.59
Expand	30.62	15.24	35.23
PSVL	36.94	19.30	38.31
SPL	**40.70**	**19.62**	**40.47**

Table 4: Effectiveness of pseudo event generation.

Label filter	R@0.5	R@0.7	mIoU
Random	24.52	12.20	34.12
Similarity	32.45	16.72	31.85
SPL	**40.70**	**19.62**	**40.47**

Table 5: Effectiveness of label filtering.

Queries per video	R@0.5	R@0.7	mIoU
1	26.96	12.98	32.32
5	35.59	18.41	37.70
10	**40.70**	**19.62**	**40.47**
20	40.43	19.49	39.84

Table 6: Different number of pseudo-queries per video.

Reweight	Refine	R@0.5	R@0.7	mIoU
✗	✗	38.74	18.71	39.38
✗	✓	39.68	**20.13**	40.07
✓	✗	39.76	19.78	39.91
✓	✓	**40.70**	19.62	**40.47**

Table 7: Effectiveness of reducing label noise.

ity falls below a certain threshold. 'PSVL' means the pseudo-events used in (Nam et al., 2021). It can be found that our method takes into account the temporal structure of the event, and therefore has the best performance.

Effectiveness of label filtering. Table 5 shows the performance of different ways of selecting pseudo labels. 'Random' means randomly selecting K pseudo labels for a video; 'Similarity' means selecting top-K pseudo labels with the highest average similarity within the event. As we can see, our method requires not only a high similarity within the event but also a low similarity outside the event to prevent the query from being too general and therefore having the best performance.

Number of training queries. Table 6 shows the performance trained with different number of pseudo-labels generated for a video. As we can see, when the number of pseudo-labels is small, increasing the number of pseudo-labels improves the performance. However, when the number of pseudo-labels is too large, the number of incorrect pseudo-labels also increases and therefore has a negative impact on the model.

Effectiveness of reducing label noise. Table 7 shows the effectiveness of sample re-weight and pseudo-label refinement. As we can see, both the sample re-weight and pseudo label refinement improve the performance. In addition, to intuitively demonstrate the effect of label re-weight, we construct a noise-controlled training set by randomly offsetting the temporal annotations in the Charades-STA dataset. Figure 3(b) shows the average weights assigned to the samples with different cleanliness (IoU with true label). As we can see, the cleaner the sample is, the greater the weight assigned to it, which demonstrates that our re-weight method indeed estimates the noise in the sample.

Choices of hyper-parameters α and β. In Figure 4, we compared the performance of using different values of hyper-parameters α and β in sample re-weight and label refinement. As we can see, when α or β is small, our sample re-weight and label refinement overly relies on the confidence of the model's output. This can have a negative impact when the model's output confidence is not accurate. As α and β gradually increase to 0.75, the model performance also gradually improves. When α and β are both 1, we do not re-weight samples or refine the labels, which exacerbates the impact of label noise on the model and leads to a decrease in performance.

4.8 Qualitative Results

Figure 5 shows some qualitative results on the Charades-STA dataset. In Figure 5(a), we show some pseudo queries and pseudo events from the Charades-STA and ActivityNet Captions datasets respectively. As we can see, we generate the free-form nature language query for the video and the pseudo-event is also correct. In Figure 5(b), we show some predictions of our model on the Charades-STA dataset. As we can see, the knowledge learned from the pseudo-labels can be gener-

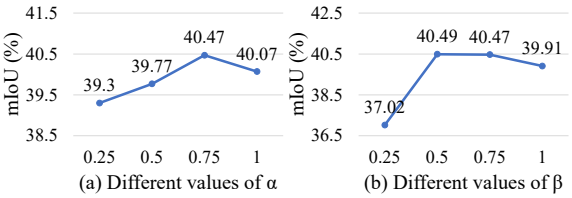

Figure 4: Choices of hyper-parameters α and β.

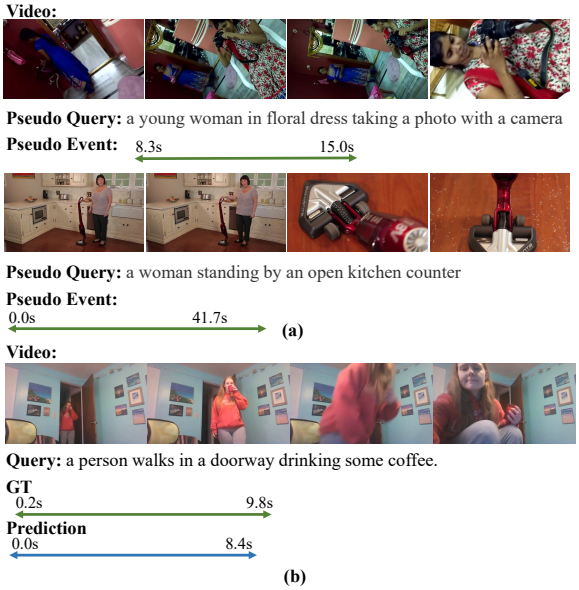

Figure 5: (a) Pseudo-labels on the Charades-STA and ActivityNet Captions datasets respectively. (b) Our predictions on the Charades-STA dataset.

alized to real queries.

5 Conclusion

In this work, we introduce a novel model SPL for zero-shot video sentence localization. We first generate free-form interpretable pseudo queries for video frames and construct query-dependent event proposals by modeling the event temporal structure. To mitigate the effect of pseudo-label noise, we propose an iterative sample re-weight and pseudo-label refinement method during training. Experiments on the Charades-STA and ActivityNet Captions datasets show the advantages of our method.

6 Limitations

In this work, we propose a structure-based pseudo-label generation method for zero-shot video sentence localization and propose a noise-resistant method to reduce the effect of pseudo-label noise. The limitations of our work are: (1) although we generate free-form natural language queries, the distribution of generated queries may still differ from the distribution of queries in the dataset (e.g. queries on the Charades-STA dataset usually start with 'person'), which may degrade the performance during testing; (2) our pseudo label refinement can correct the noisy event labels, but there is no mechanism to correct noisy queries. These can be studied as future works.

7 Acknowledgements

This work was supported by the grants from the Zhejiang Lab (NO.2022NB0AB05), National Natural Science Foundation of China (61925201,62132001,U22B2048), CAAI-Huawei MindSpore Open Fund, Alan Turing Institute Turing Fellowship, Veritone and Adobe. We thank MindSpore[2] for the partial support of this work, which is a new deep learning computing framework.

References

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR.

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

Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. 2019. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*, pages 1062–1070. PMLR.

Robert T Collins, Alan J Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, Osamu Hasegawa, Peter Burt, et al. 2000. A system for video surveillance and monitoring. *VSAM final report*, 2000(1-68):1.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

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

Junyu Gao and Changsheng Xu. 2021. Learning video moment retrieval without a single annotated video. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1646–1657.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

[2]https://www.mindspore.cn/

Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. 2022. Video activity localisation with uncertainties in temporal boundary. In *European Conference on Computer Vision*, pages 724–740. Springer.

Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7199–7208.

Charles C Kemp, Aaron Edsinger, and Eduardo Torres-Jara. 2007. Challenges for robot manipulation in human environments [grand challenges of robotics]. *IEEE Robotics & Automation Magazine*, 14(1):20–29.

Dahye Kim, Jungin Park, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. 2023. Language-free training for zero-shot video grounding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2539–2548.

R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. 2017. Dense-captioning events in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

Hengduo Li, Zuxuan Wu, Chen Zhu, Caiming Xiong, Richard Socher, and Larry S Davis. 2020. Learning from noisy anchors for one-stage object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10588–10597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Shuai Li, Chenhang He, Ruihuang Li, and Lei Zhang. 2022b. A dual weighting label assignment scheme for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9387–9396.

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

Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. 2022. Memory-guided semantic learning network for temporal sentence grounding. *arXiv preprint arXiv:2201.00454*.

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

Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601.

Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. 2021. Zero-shot natural language video localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1470–1479.

Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. 2020. Uncovering hidden challenges in query-based video moment retrieval. In *BMVC*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5552–5560.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. 2017. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847.

Guolong Wang, Xun Wu, Zhaoyuan Liu, and Junchi Yan. 2022. Prompt-based zero-shot video moment retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 413–421.

Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. 2021. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7026–7035.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699.

Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. 2021. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing*, 30:3252–3262.

Kun Yi and Jianxin Wu. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7025.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR.

Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. 2021. A closer look at temporal sentence grounding in videos: Dataset and metric. In *Proceedings of the 2nd International Workshop on Human-centric Multimedia Analysis*, pages 13–21.

Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. 2021. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12669–12678.

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

Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. 2021. Cascaded prediction network via segment tree for temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4197–4206.

Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. 2022a. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. 2022b. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Minghang Zheng, Sizhe Li, Qingchao Chen, Yuxin Peng, and Yang Liu. 2023. Phrase-level temporal relationship mining for temporal sentence localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

H. Zhou, C. Zhang, Y. Luo, Y. Chen, and C. Hu. 2021. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8441–8450, Los Alamitos, CA, USA. IEEE Computer Society.