Boundary-Aware Temporal Dynamic Pseudo-Supervision Pairs Generation for Zero-Shot Natural Language Video Localization

Xiongwen Deng[1,2], Haoyu Tang[2*], Han Jiang[1], Qinghai Zheng[3], Jihua Zhu[1*]

[1]School of Software Engineering, Xi'an Jiaotong University
[2]School of Software, Shandong University
[3]College of Computer and Data Science, Fuzhou University
V1ncent@stu.xjtu.edu.cn, tanghao258@sdu.edu.cn

Abstract

Zero-shot Natural Language Video Localization (NLVL) aims to automatically generate moments and corresponding pseudo queries from raw videos for the training of the localization model without any manual annotations. Existing approaches typically produce pseudo queries as simple words, which overlook the complexity of queries in real-world scenarios. Considering the powerful text modeling capabilities of large language models (LLMs), leveraging LLMs to generate complete queries that are closer to human descriptions is a potential solution. However, directly integrating LLMs into existing approaches introduces several issues, including insensitivity, isolation, and lack of regulation, which prevent the full exploitation of LLMs to enhance zero-shot NLVL performance. To address these issues, we propose **BTDP**, an innovative framework for **B**oundary-aware **T**emporal **D**ynamic **P**seudo-supervision pairs generation. Our method contains two crucial operations: 1) Boundary Segmentation that identifies both visual boundaries and semantic boundaries to generate the atomic segments and activity descriptions, tackling the issue of insensitivity. 2) Context Aggregation that employs the LLMs with a self-evaluation process to aggregate and summarize global video information for optimized pseudo moment-query pairs, tackling the issue of isolation and lack of regulation. Comprehensive experimental results on the Charades-STA and ActivityNet Captions datasets demonstrate the effectiveness of our BTDP method.

Introduction

Natural Language Video Localization (NLVL) is a critical and challenging task in video understanding(Tang et al. 2021a; Hu et al. 2023a; Yan et al. 2024), which aims to localize specific moment in an untrimmed video based on the corresponding language query. Most methods focus on fully supervised settings (Tang et al. 2021b; Jiang, Yizhang, and Mu 2024). However, due to the high precision required by localization tasks(Tang et al. 2024; Jiang et al. 2024), the exorbitant cost of annotations has significantly limited their practical application. To reduce the reliance on timestamp annotations, researchers have proposed weakly supervised methods (Chen et al. 2022; Bao et al. 2024), which only require the correspondence between the video and the language query. Nonetheless, annotation cost remains substantial in large-scale model training (Feng et al. 2019).

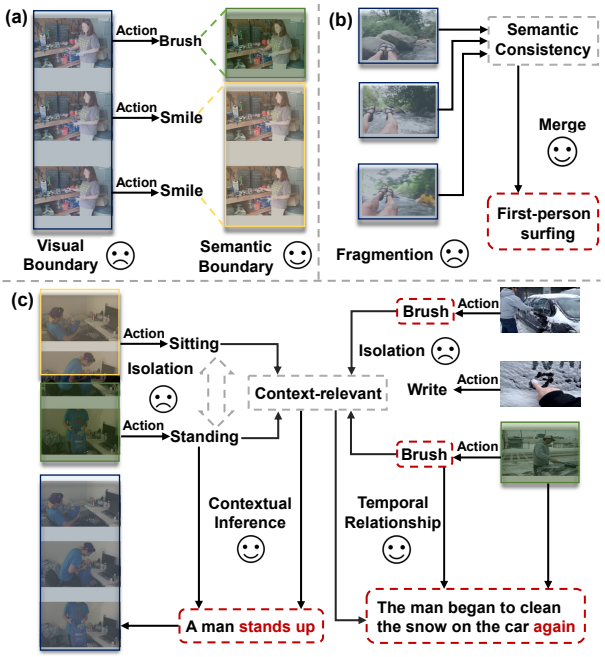

Figure 1: Issues of directly integrating LLMs to existing approaches: (a) **Insensitivity** to semantic changes when relying solely on visual features. (b) **Isolation** of segments with the same semantic meaning due to visual-based segmentation. (c) **Isolation** of contextual and temporal coherence due to the absence of global video information.

To completely eliminate the need for manual annotations, the zero-shot NLVL method, automatically generating pseudo moment-query pairs to train the localization model, has recently gained research attention. Current mainstream zero-shot methods (Nam et al. 2021) typically generate video proposals based on single visual features, and then construct pseudo queries using nouns detected by object detectors and verbs inferred by language models based on these nouns.

However, these methods often produce overly simplistic

*Corresponding Authors.

pseudo queries. Taking LoVe (Lu et al. 2024) as an example, the query format for a video proposal depicting "Person opens the entry way door" might be rendered as "door, doorway, entry, person, room". Such a reduction to the non-human and isolated term of queries presents a challenge for localization models, as they are trained to understand complex and variable language queries in open-world scenarios. This limitation leads to a critical question: **How can we generate accurate pseudo queries that more closely mimic human descriptions under zero-shot settings?** The recent advancements and applications of Large Language Models (LLMs) (Achiam et al. 2023; Team et al. 2023; Liu et al. 2024a; Li et al. 2024) and their integration with visual models (Li et al. 2023; Chen et al. 2024; Huang et al. 2024; Wu and Xie 2024) offer a potential solution: **Applying LLMs to the NLVL task to accurately generate human-like descriptive pseudo queries.**

Nonetheless, directly integrating LLMs into the existing approaches can introduce several issues: **(1) Insensitivity**: Existing approaches tend to segment only visually distinct scenes while overlooking segments where visual changes are subtle but semantic changes occur. As shown in Figure 1(a), a woman starts to smile, but the overall visual features of the scene do not change significantly. Existing approaches struggle to accurately segment in such cases, leading to generated query texts that do not match the video content. **(2) Isolation**: Existing approaches do not take into account the contextual influence, treating each segment independently. As shown in Figure 1(b), during rapid camera movement, although the semantic context remains unchanged, segmentation based solely on visual features still occurs, leading to fragmented video segments. Furthermore, as illustrated in Figure 1(c), the isolation between segments results in an absence of contextual information, preventing the LLMs from inferring the entire sequence of dynamic activities from the static image captions and from establishing temporal relationships between pseudo queries. **(3) Lack of regulation**: Similar to the challenges LLMs encounter in numerous complex text generation tasks (Madaan et al. 2024; Shinn et al. 2024), the difficulty in devising clear instructions and precise criteria means that the initial results generated by LLMs often struggle to meet task requirements accurately. These results frequently fall short of the expected criteria for high-quality pseudo queries, necessitating further refinement and adjustment.

To address these issues, we propose a Boundary-aware Temporal Dynamic Pseudo-supervision pairs (BTDP) generation framework. Our approach comprises two essential components: boundary segmentation and context aggregation. To tackle the issue of insensitivity, the boundary segmentation module is designed to establish both visual and semantic boundaries. Initially, it segments the video based on visual features to create visual boundaries. From these segments, keyframes are extracted, and visual language models are employed to generate image captions and tags. Subsequently, a LLM analyzes the semantic changes within these visual segments to determine semantic boundaries and generate corresponding temporal dynamic descriptions, resulting in more accurate atomic segments. To tackle the issue of isolation, the context aggregation module integrates global video information to create contextually relevant queries. It leverages the LLM to merge atomic segments into final segments and summarize them into dynamic activity descriptions, incorporating temporal and logical relationships. Furthermore, to tackle the issue of lack of regulation, we have incorporated a self-evaluation mechanism within the context aggregation process. This mechanism, along with multiple instructions and criteria set within the prompts, ensures that the LLMs generate outputs that closely align with predefined standards.

Our contributions are summarized as follows: (1) We propose a BTDP framework for zero-shot NLVL that leverages LLMs and visual-language models to generate boundary-aware temporal dynamic pseudo-supervision pairs through boundary segmentation and context aggregation. (2) We specifically design a prompt template that includes instructions and criteria tailored for the zero-shot NLVL task and introduce a self-evaluation mechanism, which effectively enhances the quality of the generated pseudo-supervision pairs. (3) We have conducted extensive experiments on Charades-STA and ActivityNet Captions, and the experimental results demonstrate the advantages of our method.

Related Work

Fully/Weakly Supervised NLVL Methods

Fully supervised NLVL methods train a localization model with the manually annotated moment-query pairs. Those methods can be broadly categorized into two types: proposal-based methods (Gao et al. 2021; Yuan et al. 2022; Liu et al. 2022), and proposal-free methods (Li, Guo, and Wang 2021; Nan et al. 2021; Xu et al. 2022). Proposal-based methods first generate extensive candidate video moments, which are subsequently ranked based on their alignment with the language query. In contrast, proposal-free methods directly predict the boundaries of the target moment within the video by regressing from the query. Despite progress in this area, the high cost of manual annotations remains a significant barrier to the development of those techniques.

To alleviate this issue, weakly supervised NLVL methods have emerged. These methods only require video-query pairs, thereby eliminating the need for detailed temporal boundary annotations. Early methods (Mithun, Paul, and Roy-Chowdhury 2019; Huang et al. 2021; Wang et al. 2021) used sliding windows to generate possible proposals and selected the most probable proposal from them. More recent approaches (Lv, Su, and Wen 2023; Cao et al. 2023; Kim et al. 2024) employ a learnable proposal generator to produce Gaussian proposals and optimize the quality through masked query reconstruction. Nevertheless, in this setting, the annotation cost of queries cannot be entirely avoided.

Zero-shot NLVL Methods

To eliminate manual annotation costs, several zero-shot NLVL methods have been proposed, which train video localization models with the automatically generated pseudo moment-query pairs from raw videos. PSVL (Nam et al.

2021) is the first method in this field. Specifically, it constructed pseudo queries with nouns detected from images and verbs inferred by language models based on these nouns. Building upon this, CORONET (Holla and Lourentzou 2024) further enriched the extracted video and pseudo query features using common-sense information, while LoVe (Lu et al. 2024) retrieved relevant concepts from an open concept pool to enrich the query content, and validated their accuracy and relevance within the video context, thus enhancing the alignment between visual and textual modalities. Although these methods have improved the query content and the consistency between textual and visual information, the generated pseudo queries remain relatively simplistic, and do not fully align with the human-generated language in open-world settings. Moreover, even when augmented with LLMs to strengthen the generation of their queries, these methods still encounter issues of insensitivity and isolation, resulting in the absence of semantic changes and global contextual information.

In contrast, our method leverages LLMs to generate human-like descriptive pseudo queries rather than single words. Moreover, we perform the boundary segmentation and context aggregation operations to identify semantic changes and summarize global contextual information, thereby alleviating issues of insensitivity and isolation.

LLMs for Video Understanding

Due to the language modeling capabilities of LLMs (Liu et al. 2024b), researchers have integrated LLMs with visual models to address many complex tasks in video understanding (Wang et al. 2023, 2024). For example, (Wang et al. 2022) proposed to employ LLMs to transform static frame-level captions into cohesive summaries of entire videos for video captioning, and (Zhang et al. 2023a) presented leveraging short-term visual captioners and LLMs to aggregate dense, short-term video clip descriptions for reasoning.

Note that some methods have also leveraged LLMs in the NLVL task, such as ChatVTG (Qu et al. 2024) and VTimeLLM (Huang et al. 2024). These methods designed a variety of inferencing and fine-tuning strategies to directly improve the moment localization ability of multimodal large language models (MLLMs). However, their performance is constrained to the limited video understanding ability of MLLMs. Different from them, our BTDP method enriches the language processing abilities of LLMs to generate high-quality pseudo query-moment pairs, so that any traditional NLVL models can be trained without further annotations.

The Proposed Method

Overview

As illustrated in Figure 2, our BTDP method automatically generates pseudo-supervision pairs within two stages: Boundary Segmentation and Context Aggregation. 1) Boundary Segmentation: We first segment the video into visual boundaries based on visual changes. Subsequently, we extract keyframes from each visual segment and generate coarse-grained caption and fine-grained tags as their representations. Then, we input each visual segment and its corresponding keyframes representations into the LLM. The LLM further divides semantic boundaries based on semantic changes in keyframe representations, obtaining atomic segments of the video. 2) Context Aggregation: The atomic segments and their activity descriptions of the entire video are fed into a LLM for the generation of temporal dynamic activity descriptions and their corresponding moment based on the global information of the video. To ensure that the LLM can generate high-quality pseudo-supervision pairs, we set multiple instructions and criteria in the prompts to facilitate the LLM's understanding of the task. When dealing with more challenging task of context aggregation, the LLM undergoes an iterative self-evaluation process to make the results more aligned with the set criteria. These generated pseudo-supervision pairs are used to train a video localization model. Next, the descriptions of each stage will be provided in detail.

Boundary Segmentation

To alleviate the issue of insensitivity, our method starts with visual and semantic boundary segmentation of the video. Specifically, we first segment the video into visual boundaries based on visual changes. Subsequently, we further divide semantic boundaries based on semantic changes in keyframe representations. After completing the boundary segmentation, we generate atomic segments and their corresponding activity descriptions.

Visual Boundary Detection. Given a video of N frames $\mathrm{V} = \{\mathrm{F}_1, \ldots, \mathrm{F}_N\}$, we first perform visual boundary segmentation based on a core assumption: significant changes in visual features may correspond to changes in the content of video. To this end, we utilize the PySceneDetect[1], which allows for automatic video splitting. To achieve accurate visual boundary segmentation, it identifies visual boundaries by detecting significant pixel changes between neighboring frames. As a result, the video is segmented into multiple visually independent segments, obtaining the visual boundary segmentation $\mathrm{V} = \{\mathrm{V}_b^{(1)}, \ldots, \mathrm{V}_b^{(i)}\}$. The visual segment $\mathrm{V}_b^{(i)}$ constitutes the initial segmentation of our framework.

Keyframes Extraction. With the initial segments with the visual boundary, we perform keyframe extraction to enhance processing efficiency and accurately capture the main content of each visual segment. Videos often contain numerous highly similar frames (Himakunthala et al. 2023; Hu et al. 2023b), leading to redundant visual information, while keyframes can effectively represent the core content of the video. By analyzing the content of keyframes, we can infer the main content of the entire video segment, thereby avoiding resource waste caused by unnecessary dense uniform sampling. To achieve this, we choose the Katna[2] tool for keyframe extraction. Katna employs advanced image analysis techniques, incorporating various factors such as LUV color space, brightness, contrast, and blurriness, as well as the k-means clustering algorithm to identify and ex-

[1]https://www.scenedetect.com/

[2]https://katna.readthedocs.io/

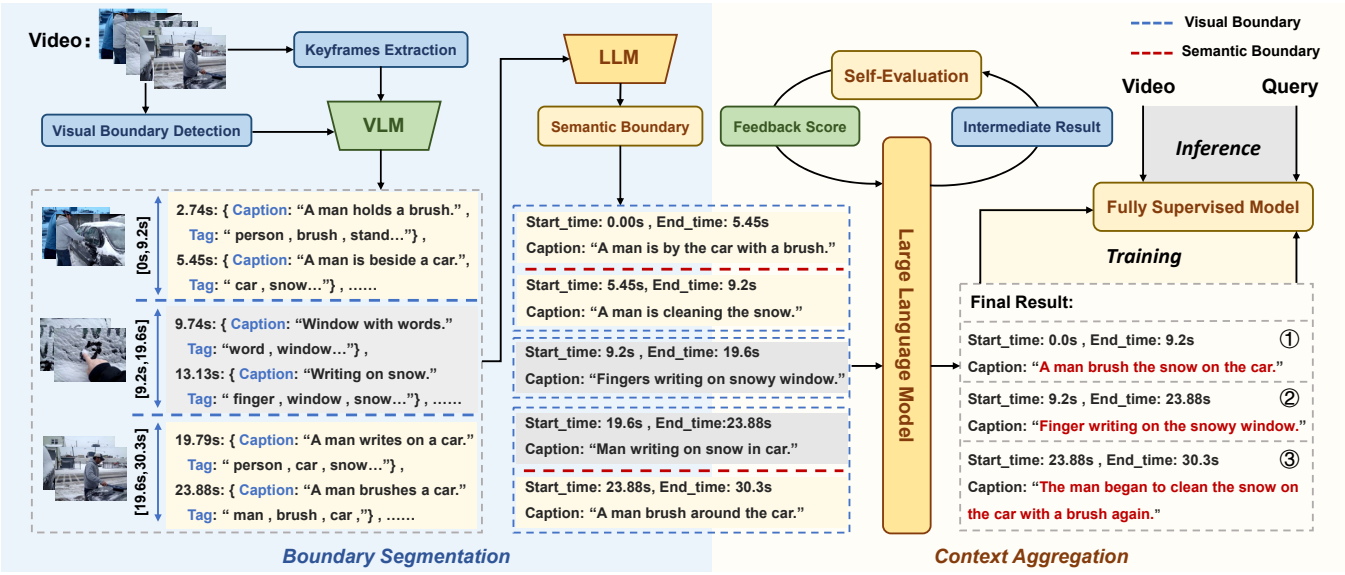

Figure 2: Illustration of our BTDP framework. The boundary segmentation step first segments the video into atomic segments based on visual and semantic boundaries, and then the context aggregation based on a LLM with the self-evaluation mechanism summarizes the atomic segments to obtain moment-query pairs. Finally, we use these pairs for the model training.

tract keyframes F_k. Following this step, the keyframes of the visual segments $\mathrm{V}_b^{(i)} = \{\mathrm{F}_k^{(1)}, \ldots, \mathrm{F}_k^{(n)}\}$ can be obtained.

Keyframes Representation. To generate diverse and varied descriptions for keyframes, we adopt a multi-granular representation generation approach. At the coarse-grained level, we utilize image captioning model to produce broad captions for each keyframe, summarizing the primary content depicted in the image. As for the fine-grained level, we use the RAM (Zhang et al. 2023b) model to identify the detailed tags for each frame, encompassing dimensions such as objects, actions, attributes, and more. These detailed tags complement the coarse-grained representations. By integrating both coarse and fine-grained representations, we achieve a comprehensive and detailed descriptions of each keyframe:

$$\left(\mathrm{C}_k^{(n)}, \mathrm{T}_k^{(n)}\right) = \mathrm{VLM}\left(\mathrm{F}_k^{(n)}\right). \tag{1}$$

Where C represents the captions of the keyframes, and T represents the tags extracted from the keyframes.

Semantic Boundary. After visual boundary segmentation and keyframes representations, we observe that segmenting video based solely on visual features is often insufficient to fully separate all video segments. This is the issue of insensitivity we previously mentioned. Sometimes, videos cannot be split by visual features, and further segmentation is required according to semantic changes. Consequently, the visual segment can be described as:

$$\mathrm{V}_b^{(i)} = \left\{\left(\mathrm{F}_k^{(1)}, \mathrm{F}_k^{(2)}\right), \ldots, \left(\mathrm{F}_k^{(n-1)}, \mathrm{F}_k^{(n)}\right)\right\}, \tag{2}$$

since the semantic information between each pair of parentheses is different, they are separated. We leverage a LLM to identify semantic boundaries within the visual segments. With its superior nature language processing abilities, the LLM can deduce dynamic activities based on the descriptions of static descriptions of keyframes, much like humans do. We input each visual segment and its corresponding keyframes descriptions into the LLM and instruct it to infer all semantically distinct activities of each visual segment. We set multiple instructions and criteria in the prompt to facilitate the LLM's better understanding of the task. Subsequently, the LLM performs semantic boundary segmentation based on the changes of keyframes representations:

$$(\mathrm{Atom}_1, \ldots, \mathrm{Atom}_k) = \mathrm{LLM}\left(\mathrm{V}_b^{(i)}\right). \tag{3}$$

Through dual visual and semantic boundary segmentation, we ultimately obtain the video atomic segment Atom_k and its dynamic description.

Context Aggregation

To alleviate the issue of isolation between segments and the deficiency of contextual information, we continue to use the LLM to analyze the global information of the video and perform context aggregation. By examining the temporal and logical relationships between atomic segments, the LLM merges semantically similar segments to summarize and generate queries for each merged segment. Through this process, we accomplish context aggregation, obtaining non-redundant final segments with new context-related queries, thus generating high-quality pseudo-supervision pairs. Given the complexity of this task, we have designed a self-evaluation prompt template in addition to setting criteria, which ensures that the generated pseudo-supervision pairs better align with our criteria.

Self-Evaluation. For dealing with complex text tasks, it is often challenging to ensure that the output of LLMs meets

the expected criteria in a single iteration. Therefore, a "**task init** $\rightarrow$ **feedback** $\rightarrow$ **iterate**" self-evaluation mechanism has been proposed, inspired by human thinking processes. This mechanism has been proven to significantly enhance the ability of LLMs to handle complex text tasks. In this mechanism, the LLM first generates an initial output, then evaluates the initial output according to predefined criteria and provides scores and detailed feedback. Based on feedback, the LLM refines its output to achieve higher quality and better align with our criteria.

Prompt Template. We provide the prompt template to better illustrate the details of our self-evaluation process. In this template, the criteria for initialization and feedback scoring are developed based on our observations and analysis of poor quality results:

(1) task init: <EXAMPLE & VIDEO>+ You are an excellent video analyst. Your task is to merge and summarize the initial atomic segments of the video and the description of each segment and generate the final segments and corresponding descriptions. The criteria for the final results are: (1) Non-redundancy: Ensure that each segment describes a unique activity, and integrate segments involving the same activity. (2) Singularity: Each segment should describe only one activity, avoiding the mixture of multiple activities within a single segment. (3) Integrity: Adjust segments based on the overall information of the video, unify segment descriptions, and emphasize temporal order and logical relationships. (4) Non-fragmentation: Avoid the presence of short segments (less than 5 seconds), ensuring that each segment is not fragmented content. (5) Accuracy: Based on global video information, ensure that the description of each segment is accurate, allowing no uncertainty.

(2) feedback: Provide atomic segments from the video and the final result obtained by context aggregation. Your task is to give scores and feedback according to the provided criteria : (1) Non-redundancy, (2) Singularity, (3) Integrity, (4) Non-fragmentation, (5) Accuracy. Here are some examples of scoring criteria: <EXAMPLE>.

(3) iterate: <FEEDBACK & SCORE>+ Okay, let's use this feedback to improve the response.

Result. After the iterative process of the aforementioned three steps, once the feedback score meets the predefined threshold, it indicates that the generated content has achieved the desired quality level. At this point, the obtained pseudo-supervision pairs are considered the final result:

$$(\mathrm{P}_1, \ldots, \mathrm{P}_n) = \mathrm{LLM}\left(\mathrm{V}_{\mathrm{Atom}}\right), \quad \text{if SCORE} > \mathrm{t}, \tag{4}$$

where P_n represents the final generated pseudo-supervision pair, $\mathrm{V}_{\mathrm{Atom}}$ represents the overall video description composed of atomic events, and t denotes the threshold score.

We summarize our pseudo-supervision pairs generation pipeline including boundary segmentation and combined summary in Algorithm 1.

Algorithm 1: Pseudo queries generation

Input: Training videos
Output: Pseudo-supervision pairs

for each training video **do**
  perform visual boundary segmentation
  for each visual segment **do**
    extract keyframes
    represent keyframes(**Eq.**1)
    perform semantic boundary segmentation(**Eq.**2,3)
  end for
  perform iterative context aggregation(**Eq.**4)
  if SCORE >t **then**
    end iteration
  end if
  return result
end for

Training and Inference

After obtaining pseudo-supervision pairs, we can directly use them to train the fully supervised models. We choose the EMB (Huang et al. 2022) model as our video localization model. The model first performs video-text alignment and then constructs a flexible elastic boundary according to the timestamp difference between the predicted endpoint and the manually marked endpoint, achieving excellent performance. The total loss function can be expressed as:

$$\mathcal{L} = \sum_{i=1}^{b} \mathcal{L}_{\mathrm{EMB}}(\mathrm{V}_i, \mathrm{Q}_i, [\hat{\tau}_s, \hat{\tau}_e]_i), \tag{5}$$

Where V_i is the video, Q_i is the pseudo query, $[\hat{\tau}_s, \hat{\tau}_e]_i$ is the corresponding start and end time of the query, $\mathcal{L}_{\mathrm{EMB}}$ is the localization loss function used in EMB, and b is the batch-size.

After completing the training, we directly uses the trained model for inference, without the need to process the query for splitting words as in the previous approaches, the whole process is more concise and efficient.

Experiment

Datasets

ActivityNet Captions: ActivityNet Captions includes approximately 20,000 open-domain videos. The average video length is 117.61 seconds. The dataset is divided into 37,421 sentence-moment pairs for training and 17,505 for test.

Charades-STA: Charades-STA consists of a collection of 9,848 untrimmed videos, each with an average length of 30.59 seconds. The dataset includes 12,408 event-query pairs for training and 3,720 for test.

Evaluation Metrics

Following previous methods, we adopt "R@k" and"mIoU" as our evaluation metrics. Specifically, "R@k" measures the proportion of instances that achieve at least one top-1 result with an IoU exceeding k compared to the ground truth. Meanwhile, "mIoU" represents the mean IoU for all instances relative to the ground truth. In our experiments, we consistently set k to the values $\{0.3, 0.5, 0.7\}$.

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
SCN (Lin et al. 2020)	Weakly	42.96	23.58	9.97	-	47.23	29.22	-	-
CRM (Huang et al. 2021)		53.66	34.76	16.37	-	55.26	32.19	-	-
CNM (Zheng et al. 2022)		60.39	35.43	15.45	-	55.68	33.33	-	-
VTimeLLM† (Huang et al. 2024)	Fine-tuning	55.30	34.30	14.70	34.60	44.80	29.50	14.20	31.40
PSVL (Nam et al. 2021)	Zero-shot	46.47	31.29	14.17	31.24	44.74	30.08	14.74	29.62
CORONET (Holla and Lourentzou 2024)		50.98	33.18	16.48	33.06	45.43	28.27	12.81	30.88
LoVe (Lu et al. 2024)		47.74	<u>34.62</u>	<u>20.16</u>	32.97	<u>49.26</u>	**31.45**	<u>15.27</u>	<u>33.25</u>
ChatVTG† (Qu et al. 2024)		<u>52.69</u>	33.01	15.89	<u>34.87</u>	40.67	22.49	9.42	27.21
BTDP (ours)	Zero-shot	**58.31**	**40.00**	**20.94**	**39.05**	**50.64**	<u>30.56</u>	**17.54**	**36.55**

Table 1: Performance comparison on the Charades-STA and ActivityNet Captions datasets. † indicates that its method uses LLMs. The best performance is highlighted in bold and the second-best is underlined.

Implementation Details

We employ PySceneDetect to segment the visual boundary and Katna to extract the keyframes. For the keyframes representation, we use the Qwen-VL (Bai et al. 2023) for coarse-grained level and the RAM (Zhang et al. 2023b) for fine-grained level. We utilize Gemini (Team et al. 2023) as our LLM for segmenting semantic boundaries and performing context aggregation. During LLM reasoning, we set the temperature to 0.7. To ensure diversity due to the randomness of each output generated by the LLM, we combine multiple results to form the final training set. Finally, we use the resulting pseudo-supervision pairs set to train the EMB (Huang et al. 2022) model. All experiments were conducted on a single NVIDIA A100 GPU.

Performance Comparison

As in Table 1, our BTDP method is compared with several baseline methods with different settings on the Charades-STA and ActivityNet-Captions dataset. As we can see, on the Charades-STA, our approach demonstrates superior performance compared to all zero-shot methods, achieving the best results across all metrics and exhibiting a 4.18% improvement in mIoU over the second-best method. On the ActivityNet Captions, our BTDP still outperforms other zero-shot methods in R@0.3, R@0.7, and mIoU, while maintaining comparable results in terms of R@0.5. Furthermore, without any annotations, our BTDP method even surpasses some weakly-supervised methods over some metrics, demonstrating the high quality of the generated pseudo-supervision pairs.

Ablation Study

To explore the effectiveness of each module in our method, we conduct extensive ablation studies on Charades-STA.

Effectiveness of boundary segmentation and context aggregation. As shown in Table 2, "BS" denotes the application of semantic boundary segmentation, while "CA" represents the context aggregation process. When "BS" is removed, LLM inference directly generates a single description for each visual segment, treating the visual boundaries as final segments without further refinement. When "CA" is removed, the results after boundary segmentation are directly used as the final pseudo-supervision pairs for the training of the localization model.

BS	CA	R@0.3	R@0.5	R@0.7	mIoU
✗	✗	49.52	33.33	15.58	33.95
✓	✗	51.77	34.87	15.88	34.81
✗	✓	54.25	35.65	17.90	36.71
✓	✓	**57.07**	**38.47**	**20.16**	**38.19**

Table 2: Ablation study of the boundary segmentation and context aggregation processes.

Method	R@0.3	R@0.5	R@0.7	mIoU
BTDP w/o SE	57.07	38.47	20.16	38.19
BTDP w/ SE	**58.31**	**40.00**	**20.94**	**39.05**

Table 3: Ablation study of the self-evaluation processes, "SE" means the self-evaluation process.

The comparison between the first and second rows, as well as the third and fourth rows, elucidates the impact of boundary segmentation. This component significantly enhances performance across all metrics, with a notable improvement in mIoU when "BS" is incorporated, demonstrating its critical role in addressing the insensitivity issue by refining the alignment between the query and the visual content. Conversely, the comparison between the first and third rows, and the second and fourth rows, underscores the contribution of the context aggregation process. The employment of "CA" results in a marked increase in performance, particularly in R@0.3 and mIoU, suggesting that the aggregation and summarization of context alleviate semantic redundancy and enhance temporal and logical relationships, leading to the improved overall performance.

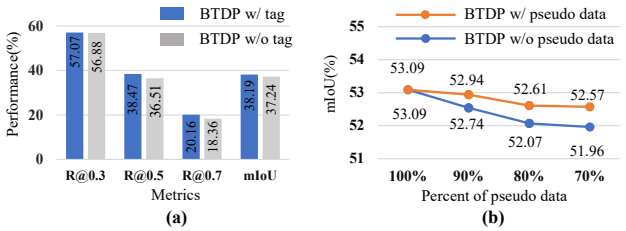

Figure 3: (a) Ablation study on the use of tags. (b) Experiment of reducing annotation cost.

Effectiveness of self-evaluation. Table 3 demonstrates the effectiveness of self-evaluation on the LLM. Specifically, the integration of self-evaluation leads to an average performance increase of 1.11% across all metrics, with a maximum improvement of 1.53% for R@0.5. This indicates that self-evaluation can effectively alleviate the issue of lack of regulation. Through iterative generation, self-evaluation makes the generated results more consistent with our criteria, improves the quality of the generated pseudo-supervised pairs, and thus enhances the performance.

Effectiveness of tags. Figure 3(a) demonstrates the effectiveness of using tags. As we can see, incorporating tags leads to performance improvements across all four metrics. The addition of tags provides finer-grained information, which enhances the accuracy and richness of the pseudo queries, resulting in improved overall performance.

Reducing annotation cost. Figure 3(b) illustrates the effectiveness of pseudo-supervision pairs in reducing annotation costs by comparing two scenarios: one where missing annotations are left unaddressed, and another where they are supplemented with pseudo-supervision pairs. The results show that by filling in the gaps with pseudo data, the model maintains a high level of accuracy even as the proportion of manually annotated data decreases. In contrast, when no supplementation is provided, performance declines significantly. This demonstrates that our method of generating pseudo-supervision pairs effectively reduces the need for extensive manual labeling, leading to substantial annotation cost savings.

Qualitative Results

To more effectively evaluate the performance of our method in generating pseudo-supervision pairs, we visualize the pseudo queries produced for sample videos and assess the qualitative results of the model in temporal localization. Additionally, we compare our approach with previous methods, highlighting the superiority of our approach.

The visualization of the pseudo queries, as shown in Figure 4(a), demonstrates that the queries generated by our method closely align with the ground truth (GT) timestamps and provide detailed descriptions that rival the quality of manual annotations. This indicates that our method is capable of producing high-quality pseudo-supervision pairs, which can be used to train models without requiring extensive human annotation efforts. We also present a qualitative comparison of video localization results, as illustrated

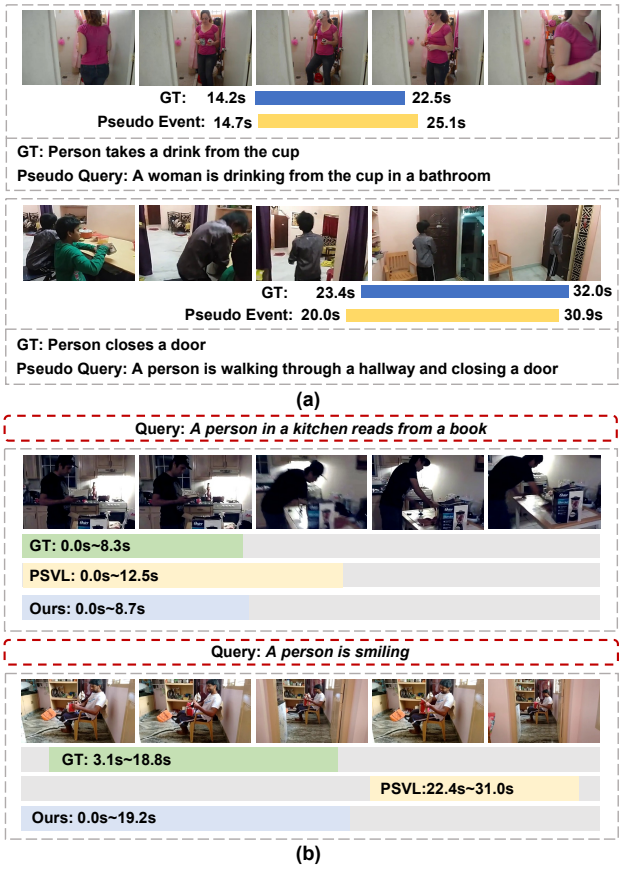

Figure 4: (a) Visualization of the pseudo-supervision pairs on the Charades-STA dataset. (b) Qualitative comparisons with PSVL on the Charades-STA dataset.

in Figure 4(b). For a given input video and corresponding query sentence, we provide the GT timestamp, the result of PSVL, and the result of our method. In these comparisons, our approach consistently achieves more accurate localization than the PSVL baseline, which further indicates the effectiveness of our method.

Conclusion

In this paper, we introduce a method that uses LLMs to generate boundary-aware temporal dynamic pseudo-supervision pairs for zero-shot natural language video localization. Our approach addresses challenges such as insensitivity, isolation, and lack of regulation through boundary segmentation, context aggregation, and self-evaluation. These processes improve text-to-visual alignment, consolidate the global video information, and ultimately produce pseudo-supervision pairs that approach the quality of manual annotations. Extensive experiments on the Charades-STA and ActivityNet Captions datasets validate the effectiveness of our method. In the future, we will focus on refining the query-segment correspondence, such as fine-tuning boundaries to create higher-quality pseudo-supervision pairs.

Acknowledgments

This work was supported in part by the Alibaba Group through Alibaba Innovative Research Program, No.21169774; in part by the National Natural Science Foundation (NSF) of China, No.62206156, No.62276155, No.72004127, and No.62206157; in part by the NSF of Shandong Province, No.ZR2024QF104 No.ZR2021MF040 and No.ZR2022QF047; in part by the Key R&D Program of Shandong Province, China (Major Scientific and Technological Innovation Projects), No.2022CXGC020107.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Bao, P.; Shao, Z.; Yang, W.; Ng, B. P.; Er, M. H.; and Kot, A. C. 2024. Omnipotent Distillation with LLMs for Weakly-Supervised Natural Language Video Localization: When Divergence Meets Consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 747–755.

Cao, M.; Wei, F.; Xu, C.; Geng, X.; Chen, L.; Zhang, C.; Zou, Y.; Shen, T.; and Jiang, D. 2023. Iterative proposal refinement for weakly-supervised video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6524–6534.

Chen, J.; Luo, W.; Zhang, W.; and Ma, L. 2022. Explore inter-contrast between videos via composition for weakly supervised temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 267–275.

Chen, T.-S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.-w.; Jeon, B. E.; Fang, Y.; Lee, H.-Y.; Ren, J.; Yang, M.-H.; et al. 2024. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13320–13331.

Feng, Y.; Ma, L.; Liu, W.; and Luo, J. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4125–4134.

Gao, J.; Sun, X.; Xu, M.; Zhou, X.; and Ghanem, B. 2021. Relation-aware video reading comprehension for temporal language grounding. *arXiv preprint arXiv:2110.05717*.

Himakunthala, V.; Ouyang, A.; Rose, D.; He, R.; Mei, A.; Lu, Y.; Sonar, C.; Saxon, M.; and Wang, W. Y. 2023. Let's Think Frame by Frame with VIP: A Video Infilling and Prediction Dataset for Evaluating Video Chain-of-Thought. *arXiv preprint arXiv:2305.13903*.

Holla, M.; and Lourentzou, I. 2024. Commonsense for Zero-Shot Natural Language Video Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2166–2174.

Hu, Y.; Wang, K.; Liu, M.; Tang, H.; and Nie, L. 2023a. Semantic collaborative learning for cross-modal moment localization. *ACM Transactions on Information Systems*, 42(2): 1–26.

Hu, Z.; Wang, Z.; Song, Z.; and Hong, R. 2023b. Dual Video Summarization: From Frames to Captions. In *IJCAI*, 846–854.

Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14271–14280.

Huang, J.; Jin, H.; Gong, S.; and Liu, Y. 2022. Video activity localisation with uncertainties in temporal boundary. In *European Conference on Computer Vision*, 724–740. Springer.

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.

Jiang, H.; Tang, H.; Yan, M.; Zhang, J.; Xu, M.; Hu, Y.; Zhu, J.; and Nie, L. 2024. Revisiting Unsupervised Temporal Action Localization: The Primacy of High-Quality Actionness and Pseudolabels. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5643–5652.

Jiang, H.; Yizhang, Y.; and Mu, Y. 2024. Transferable Video Moment Localization by Moment-Guided Query Prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2516–2524.

Kim, S.; Cho, J.; Yu, J.; Yoo, Y.; and Choi, J. Y. 2024. Gaussian Mixture Proposals with Pull-Push Learning Scheme to Capture Diverse Events for Weakly Supervised Temporal Video Grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2795–2803.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, K.; Guo, D.; and Wang, M. 2021. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1902–1910.

Li, Z.; Liu, F.; Wei, Y.; Cheng, Z.; Nie, L.; and Kankanhalli, M. 2024. Attribute-driven Disentangled Representation Learning for Multimodal Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9660–9669. ACM.

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.

Liu, D.; Qu, X.; Zhou, P.; and Liu, Y. 2022. Exploring motion and appearance information for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1674–1682.

Liu, F.; Liu, Y.; Chen, H.; Cheng, Z.; Nie, L.; and Kankanhalli, M. 2024a. Understanding Before Recommendation:

Semantic Aspect-Aware Review Exploitation via Large Language Models. *ACM Transactions on Information Systems*, 1–26.

Liu, X.; Sun, Y.; Cheng, R.; Xia, L.; Abumarshoud, H.; Zhang, L.; and Imran, M. A. 2024b. Knowledge-Assisted Privacy Preserving in Semantic Communication. *arXiv preprint arXiv:2410.18418*.

Lu, Y.; Quan, R.; Zhu, L.; and Yang, Y. 2024. Zero-Shot Video Grounding With Pseudo Query Lookup and Verification. *IEEE Transactions on Image Processing*, 33: 1643–1654.

Lv, Z.; Su, B.; and Wen, J.-R. 2023. Counterfactual cross-modality reasoning for weakly supervised video moment localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6539–6547.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

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.

Nam, J.; Ahn, D.; Kang, D.; Ha, S. J.; and Choi, J. 2021. Zero-shot natural language video localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1470–1479.

Nan, G.; Qiao, R.; Xiao, Y.; Liu, J.; Leng, S.; Zhang, H.; and Lu, W. 2021. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2765–2775.

Qu, M.; Chen, X.; Liu, W.; Li, A.; and Zhao, Y. 2024. ChatVTG: Video Temporal Grounding via Chat with Video Dialogue Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1847–1856.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Tang, H.; Hu, Y.; Wang, Y.; Zhang, S.; Xu, M.; Zhu, J.; and Zheng, Q. 2024. Listen as you wish: Fusion of audio and text for cross-modal event detection in smart cities. *Information Fusion*, 110: 102460.

Tang, H.; Zhu, J.; Liu, M.; Gao, Z.; and Cheng, Z. 2021a. Frame-wise cross-modal matching for video moment retrieval. *IEEE Transactions on Multimedia*, 24: 1338–1349.

Tang, H.; Zhu, J.; Wang, L.; Zheng, Q.; and Zhang, T. 2021b. Multi-level query interaction for temporal language grounding. *IEEE Transactions on Intelligent Transportation Systems*, 23(12): 25479–25488.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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.

Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X.; Li, X.; Chen, G.; Chen, X.; Wang, Y.; et al. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Wang, Y.; Meng, X.; Liang, J.; Wang, Y.; Liu, Q.; and Zhao, D. 2024. Hawkeye: Training video-text llms for grounding text in videos. *arXiv preprint arXiv:2403.10228*.

Wang, Z.; Li, M.; Xu, R.; Zhou, L.; Lei, J.; Lin, X.; Wang, S.; Yang, Z.; Zhu, C.; Hoiem, D.; et al. 2022. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35: 8483–8497.

Wu, P.; and Xie, S. 2024. V?: Guided Visual Search as a Core Mechanism in Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13084–13094.

Xu, Z.; Chen, D.; Wei, K.; Deng, C.; and Xue, H. 2022. HiSA: Hierarchically semantic associating for video temporal grounding. *IEEE Transactions on Image Processing*, 31: 5178–5188.

Yan, M.; Liu, F.; Sun, J.; Sun, F.; Cheng, Z.; and Han, Y. 2024. Behavior-Contextualized Item Preference Modeling for Multi-Behavior Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 946–955. ACM.

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2022. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. *IEEE transactions on pattern analysis and machine intelligence*, 44: 2725–2741.

Zhang, C.; Lu, T.; Islam, M. M.; Wang, Z.; Yu, S.; Bansal, M.; and Bertasius, G. 2023a. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*.

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.

Zhang, Y.; Huang, X.; Ma, J.; Li, Z.; Luo, Z.; Xie, Y.; Qin, Y.; Luo, T.; Li, Y.; Liu, S.; et al. 2023b. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*.

Zheng, M.; Huang, Y.; Chen, Q.; and Liu, Y. 2022. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3517–3525.