Omnipotent Distillation with LLMs for Weakly-Supervised Natural Language Video Localization: When Divergence Meets Consistency

Peijun Bao*[1], Zihao Shao[2], Wenhan Yang[3], Boon Poh Ng[1], Meng Hwa Er[1], Alex C. Kot[1]

[1]Nanyang Technological University
[2]Peking University
[3]Peng Cheng Laboratory
peijun001@e.ntu.edu.sg, zh.s@pku.edu.cn, yangwh@pcl.ac.cn, {ebpng, emher, eackot}@ntu.edu.sg

Abstract

Natural language video localization plays a pivotal role in video understanding, and leveraging weakly-labeled data is considered a promising approach to circumvent the labor-intensive process of manual annotations. However, this approach encounters two significant challenges: 1) **limited input distribution**, namely that the limited writing styles of the language query, annotated by human annotators, hinder the model's generalization to real-world scenarios with diverse vocabularies and sentence structures; 2) **the incomplete ground truth**, whose supervision guidance is insufficient. To overcome these challenges, we propose an omnipotent distillation algorithm with large language models (LLM). The distribution of the input sample is enriched to obtain diverse multi-view versions while a consistency then comes to regularize the consistency of their results for distillation. Specifically, we first train our teacher model with the proposed intra-model agreement, where multiple sub-models are supervised by each other. Then, we leverage the LLM to paraphrase the language query and distill the teacher model to a lightweight student model by enforcing the consistency between the localization results of the paraphrased sentence and the original one. In addition, to assess the generalization of the model across different dimensions of language variation, we create extensive datasets by building upon existing datasets. Our experiments demonstrate substantial performance improvements adaptively to diverse kinds of language queries.

Introduction

Natural language video localization (Gao et al. 2017) is an important yet challenging task with a wide spectrum of applications in video understanding and analysis (Sreenu and Durai 2019; Qi et al. 2021; Zhu et al. 2021; Bao et al. 2023). The goal of this task is to temporally localize a video segment (i.e., start and end time) that best corresponds to a query sentence from untrimmed videos. Despite achieving impressive results, the fully-supervised natural language video localization (Liu et al. 2018; Zhang et al. 2019a,b, 2020a; Wang, Ma, and Jiang 2020; Bao and Mu 2022) requires laborious manual annotations of temporal moment boundaries, which are unscalable to the real-world setting.

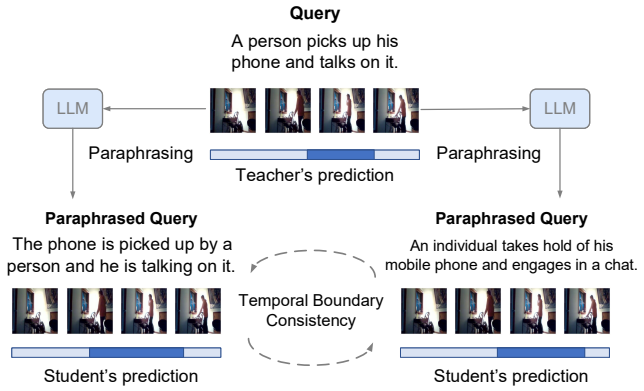

Figure 1: Existing datasets suffer from limited sentence structures and vocabulary of natural language queries *e.g.,* on the Charades-STA dataset often with the structure of "sub + pred + obj". By exploiting temporal boundary consistency, we propose a manual annotation-free omnipotent distillation algorithm with LLM, adapting the localization ability from a teacher model focusing on the vanilla language style to a student model whose input sentence queries are with rich language variations.

Due to this, the weakly-supervised setting has attracted increasing attention in recent years (Gao et al. 2019; Mithun, Paul, and Roy-Chowdhury 2019; Lin et al. 2020; Chen et al. 2020; Tan et al. 2021; Zheng et al. 2022a,b), where only video-level descriptions are available during training. However, the performance of the existing weakly-supervised methods is still unsatisfactory and lags behind the fully-supervised methods because the incomprehensive annotations do not provide sufficient supervisory signals for training. Moreover, the language queries provided by the human annotator often suffer from limitations in writing styles, such as restricted vocabulary and sentence structures. As a result, this hinders the model's capability to generalize to real-world applications, whose language queries showcase notable variations in writing styles as well as cultural nuances.

To address these issues, we propose an Omnipotent Distillation algorithm (OmniD) with the Large Language Model (LLM) as shown in Fig 1. Specifically, we first devise a bootstrapping learning framework to train a sophisticated

*Corresponding author.

teacher model. The teacher model is composed of multiple sub-models, each serving as an auxiliary model to address the shortage of effective training guidance for the other sub-models by providing pseudo-labels of temporal boundaries. An intra-model consistency distillation is designed to explicitly compensate for the localization errors of each sub-model. Instead of introducing extra manual annotations, such a design alleviates the deficient supervision problem by additional computational costs during training.

Subsequently, we capitalize on the generation capability of the large language model (Brown et al. 2020; Touvron et al. 2023; OpenAI 2023) to paraphrase each sentence query, formulating extensive rephrased sentences with diversity in vocabularies and sentence structures while maintaining the original meaning. Given the semantic equivariance of paraphrasing, the video moment described by the primary sentence query and its paraphrased version should also exhibit equivariance. To this end, we devise a semantic equivariance distillation loss to distill the teacher model to a smaller student model by encouraging the consistency of temporal boundary predictions between the paraphrased sentence and the untouched one. In this way, we transfer the localization ability from a teacher model, which emphasizes a vanilla language style, to a student model that engages with input sentences featuring diverse and intricate language variations.

In addition, existing datasets such as Charades-STA (Gao et al. 2017) and ActivityNet-Captions (Krishna et al. 2017) suffer from restricted writing styles in terms of sentence queries. For instance, the majority of the sentence queries in the Charades-STA dataset share a similar sentence structure, following the pattern of "subject-verb-object", with passive voice constructions being rare occurrences. To evaluate the model's generalization across various queries, we create totally six variants of these datasets with rich vocabulary and sentence structures.

Our contributions are summarized as follows: 1) To the best of our knowledge, we are the first to exploit the large language model for the task of natural language video localization. 2) We propose an omnipotent distillation algorithm to tackle the challenges of ineffective supervision and query multiformity. Intra-model consistency distillation and semantic equivariance distillation are crafted to tackle these challenges respectively. 3) To evaluate the generalization capability of the model across different dimensions of language variation, we construct comprehensive datasets by expanding upon existing ones. 4) Extensive experiments verify the superiority of the proposed methods in both performance and adaptability compared to state-of-the-art approaches.

Related Works

Natural Language Video Localization. The task of natural language video localization is initially introduced by Gao et al. (2017) with the goal of identifying the start and end time points of video moment based on a natural language query and an untrimmed video. Gao et al. (2017) propose a language-video localizer to identify the temporal boundary for candidate video clips. A semantic matching reinforcement learning framework is devised by Wang, Huang, and Wang (2019) to reduce the large visual-semantic discrepancy between video and language. A cross-modal attention network is proposed in (Liu et al. 2018) to highlight the essential part of visual features or query contents. Bao, Zheng, and Mu (2021) devise an event propagation network to localize video moments that are semantically related and temporally coordinated. While achieving promising localization accuracy, the fully-supervised methods rely on the manual annotations of the temporal boundaries which are labor-intensive and subjective to label.

To solve this issue, weakly-supervised natural language video localization has recently gain growing attention (Zheng et al. 2022a,b; Bao et al. 2024; Tan et al. 2021; Chen et al. 2020; Lin et al. 2020), where only the sentence query and the paired video are required for training. Early works (Mithun, Paul, and Roy-Chowdhury 2019; Tan et al. 2021) explore using joint visual-semantic embedding and text-guided attention to avoid laborious temporal boundary annotations. A latent graph co-attention (Tan et al. 2021) is proposed in to use fine-grained frame-by-word interactions to reason about correspondences between possible pairs of frames. Chen et al. (2020) devise a two-stage model to tackle the weakly-supervised natural language video localization in a coarse-to-fine manner where more precise start and end timestamps of the retrieval results are obtained during the fine stage. To the best of our knowledge, we are the first in the literature on natural language video localization to use a large language model to boost the adaptability of the localization model to diversified language queries.

Knowledge Distillation. The approaches of knowledge distillation, initially introduced in (Hinton, Vinyals, and Dean 2015), serve the purpose of compressing and accelerating models. It achieves this by transferring the knowledge amassed by a larger, intricate model to a smaller, more efficient counterpart. In recent years, knowledge distillation has seen expanded applications in various domains, including zero-shot learning (Nayak et al. 2019; Micaelli and Storkey 2019), domain adaptation (Deng, Luo, and Zhu 2019; Chen et al. 2019), and multimodal learning (Gupta, Hoffman, and Malik 2015; Wang et al. 2020; Yu, Liu, and Chan 2021). Different from these works applied in fully-/semi-supervised scenarios, we capitalize on knowledge distillation to solve the obstacles of insufficient supervision and lacking query multiformity in the weakly-supervised setting.

Large Language Model. The large language models (LLM) are transformer-based language models that contain hundreds of billions or more parameters trained on massive text data, such as GPT-3 (Brown et al. 2020), GPT-4 (OpenAI 2023), PaLM (Chowdhery et al. 2022) and LLaMA (Touvron et al. 2023). LLMs not only show a significant performance advancement but also exhibit strong capacities in in-context learning (Brown et al. 2020), that are not presented in the small-scale language models *e.g.,* BERT (Devlin et al. 2019). A milestone utilization of LLMs is ChatGPT[1] that harnesses the LLMs from the GPT series for dialogue, showcasing an impressive conversational ability. We employ the remarkable paraphrasing capability of LLM to enhance the adaptability

[1]https://chat.openai.com/

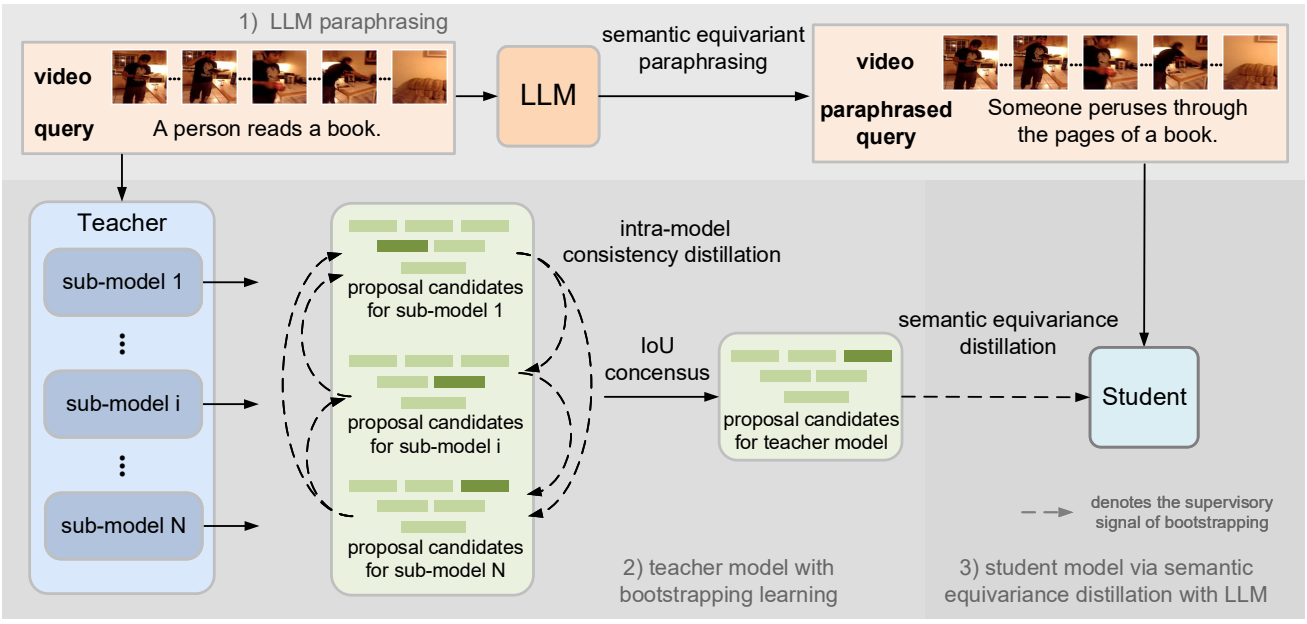

Figure 2: An overview of the proposed omnipotent distillation (OmniD) with LLM. Our method consists of 1) an LLM paraphrasing module, 2) a teacher model via bootstrapping learning with intra-model consistency distillation, and 3) a student model via semantic equivariance distillation with LLM. In the teacher model, intra-model supervision signals are utilized to mutually reduce the problem of ineffective training guidance and explicitly compensate for the errors of each sub-model. Then the student model improves the generalization ability across diversified language queries using the equivariance property of the moment proposal before/after the sentence paraphrasing. The proposal candidates highlighted in dark green correspond to those identified as positive predictions.

of natural language video localization models when dealing with a domain shift of sentence queries. We highlight that our approach eliminates additional annotation efforts to label temporal boundaries, thanks to the proposed bootstrapping distillation.

Omnipotent Distillation with LLM

Method Overview

Given a sentence query $\mathcal{Q}$ and an untrimmed video $\mathcal{V}$, natural language video localization aims to temporally localize the temporal boundary $b = (s, e)$ of a video moment described by the query $\mathcal{Q}$, where s and e denote the start and end time point of the moment respectively. In the weakly-supervised setting, the model requires the sentence-video pairs for training, without relying on the annotation of temporal boundaries. However, the weakly-supervised model is often constrained by the insufficient supervision dilemma due to incomplete annotation. Moreover, language query in the existing datasets lacks diversity in sentence structures and vocabulary. For instance, the sentence queries in the Charades-STA dataset are often with the structure of "sub + pred + obj". This limitation hampers the model's capacity to achieve robust generalization in real-world scenarios characterized by a diverse range of language variations.

To overcome these difficulties, as shown in Fig 2, we propose Omnipotent Distillation (OmniD) with the large language model by exploiting two sorts of consistency among the temporal boundary predictions. 1) *Intra-model consistency*: we first devise a bootstrapping learning algorithm with an intra-model consistency distillation to train a teacher model. This teacher model is constituted by multiple sub-models, where each sub-model serves as an assistive model to collaboratively provide training guidance to the other model, alleviating the ineffective supervision constraint. 2) *Semantic equivariance consistency*: we capitalize on the LLM (OpenAI 2023) to paraphrase the sentence query, where a wide range of rephrased sentences are formulated with diversity in vocabularies and sentence structures while keeping the semantic meaning unchanged. The localization capability of the teacher model is then transferred into a smaller, more efficient student model. This is achieved by encouraging the temporal boundary prediction of paraphrased sentences from the student model to be aligned with the prediction from the teacher model with the unaltered sentence query.

Sub-Model Construction

We use the CPL model (Zheng et al. 2022b) as the basic network architecture for the sub-model, which comprises a proposal generator and a sentence reconstructor. Here we take a brief overview of the network architecture and more details can be referred to (Zheng et al. 2022b). Specifically, the sentence query $\mathcal{Q}$ and untrimmed video $\mathcal{V}$ are first encoded to feature representation $q \in \mathbb{R}^{L_Q \times d}$ and $v \in \mathbb{R}^{L_V \times d}$ by a list of multi-head attention layers respectively. Then another list of cross-modal attention layers regress K proposal candidates

from q and v as $\{b_k = (s_k, e_k)\}$ $(k = 1 \dots K)$. We randomly mask M words $\{w_i\}$ in the sentence and enforce the sentence reconstructor to reconstruct the masked words from the k-th proposal candidate as $\hat{w}_i^k$. The cross-entropy loss can then be used to evaluate the reconstruction correctness as

$$\mathcal{L}_{rec}[k] = \sum_{i=1}^{M} \mathcal{L}_{ce}(w_i, \hat{w}_i^k) \tag{1}$$

Assume that the k^*-th proposal candidate is selected as the positive proposal that matches the sentence query. A ranking loss $\mathcal{L}_{rank}$ is further applied as in (Zheng et al. 2022b) to guarantee that the positive candidate k^* has a reconstruction loss smaller than the hard negative candidates by a specific margin. The final loss function is written as:

$$\mathcal{L}_{sub} = \mathcal{L}_{rec}[k^*] + \mathcal{L}_{rank}[k^*] \tag{2}$$

In the weakly-supervised setting, the annotations of temporal boundaries are not available, and thus the ground-truth positive proposal is unknown for training. Previous works (Lin et al. 2020; Zheng et al. 2022a,b) heuristically select the one in the proposal candidates with the smallest reconstruction loss as positive proposal k^*, formulated as

$$k^* = \text{argmin}_{k=1\dots K} \mathcal{L}_{rec}[k] \tag{3}$$

However, such selection is inherently prone to errors due to the shortage of adequate training guidance. To cope with this barrier, we propose bootstrapping learning with intra-model consensus supervisory signals as illustrated in the following subsection.

Teacher Model via Bootstrapping Learning with Intra-Model Consistency Distillation

The teacher model $\mathcal{T}$ is constituted by multiple sub-models, where each sub-model serves as an assistive model to mutually provide training guidance to the other model and alleviate the inadequate supervision issue. Specifically, the teacher model consists of N sub-models. For the i-th sub-model, each of all other sub-models is considered a reference model, which serves as a source of pseudo-labels for the temporal boundaries. We then leverage the predictions of $(N-1)$ sub-models to create a consensus-based pseudo-label for i-th sub-model, enhancing its the accuracy and reliability.

Assume that the i-th sub-model predicts P proposal candidates that correspond to the sentence query as b_{ij} where $j = 1 \dots P$ and the k^*-th proposal $\hat{b}_k$ is chosen as its prediction for i-th sub-model. The consensus score c_{ij} for j-th proposal candidate of i-th sub-model is then determined as the average intersection over union (IoU) of $(N-1)$ sub-models, written as:

$$c_{ij} = \sum_{k=1, k\neq i}^{N} \sigma(b_{ij}, \hat{b}_k) \tag{4}$$

where σ is the IoU operator.

Instead of heuristically choosing by i-th model as Eq. 3 as pseudo-labels for training which is easy to prone errors, we exploit the consensus scores from the other $(N-1)$ sub-models to determine the pseudo-label for i-th model. In more detail, the final pseudo-label for i-th sub-model is determined as the proposal candidate with the largest consensus score across the P proposal candidates, written as

$$p_i = \text{argmax}_j(c_{ij}) \tag{5}$$

where p_i-th proposal candidate of i-th model is chosen as the positive one.

The loss function of intra-model consistency distillation for bootstrapping learning can then be formulated as

$$\mathcal{L}_{boot}^{\mathcal{T}} = \sum_{i=1}^{N} \mathcal{L}_{rec}[p_i] + \mathcal{L}_{rank}[p_i] \tag{6}$$

In this way, intra-model supervision signals are leveraged to jointly relieve the problem of ineffective supervision and explicitly compensate for the errors of each sub-model. The final prediction $b^{\mathcal{T}}$ of the teacher model is determined by selecting the one with the largest consensus score of the N sub-models:

$$\begin{aligned} s_i &= \sum_{k=1, k\neq i}^{N} \sigma(\hat{b}_i, \hat{b}_k) \\ b^{\mathcal{T}} &= \hat{b}[\text{argmax}_{i=1\dots N} s_i] \end{aligned} \tag{7}$$

where σ represents the IoU operator, and $\hat{b}[k]$ denotes the selection of the entry in $\hat{b}$ with the k-th index.

Semantic Equivariance Distillation with LLM

Given a natural language query $\mathcal{Q}$, we request the large language model GPT-3.5 (OpenAI 2023) to paraphrase $\mathcal{Q}$ as $\mathcal{Q}'$. The paraphrased $\mathcal{Q}'$ maintains the intended semantic meaning while enjoying rich diversity in writing styles such as vocabularies and sentence structures. The prompts that guide the large language model with context information on paraphrasing are included in the supplementary materials (which can be referenced in the arXiv version of this paper).

Given the semantic equivariance of paraphrasing, the video moment described by both the primary sentence query and its paraphrased version should also exhibit equivariance. While ground-truth temporal boundaries are absent in the weakly-supervised setting, we exploit this equivariance property to transition the knowledge of the teacher model to a lightweight student model $\mathcal{S}$, and further enable its versatile ability in language understanding.

Assume that the student model predicts K proposal candidates $\{b'_k\}$ for the paraphrased query $\mathcal{Q}'$, the localization result of the student model with $\mathcal{Q}'$ is encouraged to be accordant to the teacher model with the unaltered query $\mathcal{Q}$. The semantic equivariance distillation loss function for the student model $\mathcal{S}$ is then formulated as

$$\mathcal{L}^{\mathcal{S}} = \mathcal{L}_{rec}[p^{\mathcal{S}}] + \mathcal{L}_{rank}[p^{\mathcal{S}}] \tag{8}$$

where the pseudo-label $p^{\mathcal{S}}$ is determined by choosing the proposal from $\{b'_k\}$ of $\mathcal{Q}'$ with the highest IoU value to the prediction of the teacher model as

$$\begin{aligned} s_i &= \sigma(\hat{b}_i, b^{\mathcal{T}}) \\ p^{\mathcal{S}} &= \text{argmax}_i s_i \end{aligned} \tag{9}$$

During inference, the teacher model is dropped and only the student model is utilized to localize the video moment based on the given language query.

Dataset Expansion

Charades-STA (Gao et al. 2017) and ActivityNet-Captions (Krishna et al. 2017) are two widely-used datasets in the task of natural language video localization. These datasets suffer from restricted writing styles in terms of sentence queries. For instance, the majority of the sentence queries in the Charades-STA dataset exhibit a consistent sentence structure, adhering to the "subject-verb-object" pattern. Additionally, passive voice constructions are infrequent occurrences in both the Charades-STA and ActivityNet-Captions. In a real-world setting, however, it is common to use assorted expressions to convey the same semantics of a video moment.

To assess the model's generalization capability across various types of queries, we extend the testing data of Charades-STA and ActivityNet-Captions by creating three variants that encompass various writing styles as follows. i) Sentence structure variant (SS): This variation focuses on changing sentence structures while being prone to maintaining the original vocabulary. ii) Vocabulary variant (VO): In this variant, we modify only the vocabulary used in the sentences while endeavoring to keep the sentence structures intact. iii) Hybrid variant (HY): Both the vocabularies and sentence structures are modified in this variant to introduce a combined effect.

To create these variant datasets, we first design a list of different prompts (available in the supplementary materials) and request the large language model (OpenAI 2023) to paraphrase the natural language sentence. Then we manually check the semantic equivariance between the untouched sentence and the paraphrased one in the testing dataset. These variations allow us to examine the model's performance across different dimensions of language variation, providing a comprehensive assessment of its generalization capabilities.

Experiments

Datasets

We validate the effectiveness of the proposed approach to the state-of-the-art methods on two benchmark datasets, namely Charades-STA (Gao et al. 2017) and ActivityNet-Captions (Krishna et al. 2017). Furthermore, we introduce six variant datasets by building upon the vanilla Charades-STA and ActivityNet-Captions datasets, which enables us to evaluate the model's localization accuracy over diverse sentences with varying structure and vocabulary. The details of the datasets are given as follows.

1) Charades-STA (Gao et al. 2017) comprises 9,848 videos showcasing daily indoor activities. The sentence queries within the original dataset exhibit an average length of 8.6 words. And the videos possess an average duration of 29.8 seconds. This dataset is initially designed for action recognition / localization (Sigurdsson et al. 2016), and subsequently extended by Gao *et al.,* (Gao et al. 2017) with language descriptions for natural language video localization. As illustrated in the previous subsections of "Dataset Expansion", we introduce three variation datasets: *Charades-STA SS*, *Charades-STA VO*, and *Charades-STA HY*. In each variant dataset, we undertake paraphrasing of the original sentence query three times, generating distinct paraphrased sentences based on the specified criteria. As a result, the total number of sentences in the variation dataset increases by a factor of 9 compared to the original one. The detailed statistics of the variant datasets are included in the supplementary materials of our arXiv version.

2) ActivityNet Captions (Krishna et al. 2017) is comprised of 19,290 untrimmed videos, encompassing a wide range of diverse and open visual content. The average duration of the video is 117.74 seconds and the average length of the description is 13.16 words in the original dataset. There are 2.4 annotated moments with a duration of 8.2 seconds in each video. Similar to the Charades-STA variants, three variation datasets are introduced *i.e., ActivityNet-Captions SS*, *ActivityNet-Captions VO*, and *ActivityNet-Captions HY* with rich language multiformity. The total sentence number is nine times as the one in the vanilla dataset. And we provide more details of the variations in the supplementary materials.

Evaluation Metrics

Following previous works (Gao et al. 2017; Lin et al. 2020; Zheng et al. 2022b), we adopt the metrics of "R@m" for evaluation. Specifically, we calculate the Intersection over Union (IoU) between the localized temporal moment and the corresponding ground truth. "R@m" is defined as the percentage of language queries having correct localization results, where the localization is correct if its IoU is larger than m. As prior works, we report the results with $m = \{0.3, 0.5, 0.7\}$ on Charades-STA and its variations, and set $m = \{0.1, 0.3, 0.5\}$ on ActivityNet-Captions and its variations.

Implementation Details

Following prevailing works, we use I3D network (Carreira and Zisserman 2017) and C3D network (Tran et al. 2015) to extract video features for Charades-STA and ActivityNet-Captions respectively. We employ pre-trained GloVe word2vec embeddings (Pennington, Socher, and Manning 2014) to extract sentence features. We set the maximum description length to 20 on both datasets. The dimensions of the hidden state for both language and visual features are set to 256. The number of video snippets is resampled to 200 on both datasets. We use the Adam optimizer (Kingma and Ba 2014) for training with a batch size of 32. The sub-model numbers N in the teacher model are set to 3. We first train the teacher model for 15 epochs, and subsequently distill it to the student model for another 15 epochs. The learning rate is set to 0.0005 for Charades-STA dataset and 0.00035 for ActivityNet-Captions dataset, respectively.

Performance Comparisons

1) Charades-STA variant datasets. Table 1 summarizes the localization accuracy of the proposed method OmniD against state-of-the-art approaches on three variant datasets of Charades-STA, *i.e., Charades-STA SS*, *Charades-STA VO*, and *Charades-STA HY*. In addition to the standard versions, we further enhance the previous methods with LLM paraphrasing for a fair comparison (indicated by a check mark in the respective column). To achieve this, we carefully reimplement each of these methods using their officially released code and incorporate training with the additional nat-

Method	LLM	Charades-STA SS			Charades-STA VO			Charades-STA HY		
		R@0.3	R@0.5	R@0.7	R@0.3	R@0.5	R@0.7	R@0.3	R@0.5	R@0.7
SCN (Lin et al. 2020)	✗	55.51	22.26	6.51	53.24	21.45	5.70	54.27	22.42	6.27
CPL (Zheng et al. 2022b)	✗	53.74	38.78	17.54	46.69	33.85	15.47	48.64	35.32	15.67
OmniD (Ours)	✗	56.91	41.41	18.64	49.78	35.66	16.47	51.80	36.57	16.81
SCN (Lin et al. 2020)	✓	57.81	24.93	7.05	58.97	25.71	8.13	56.64	23.04	6.83
CPL (Zheng et al. 2022b)	✓	60.30	44.74	20.70	58.23	42.95	19.60	59.44	43.96	20.32
OmniD (Ours)	✓	**65.66**	**49.09**	**22.93**	**64.78**	**48.63**	**22.30**	**65.09**	**49.04**	**22.83**

Table 1: Comparisons with state-of-the-art methods on three Charades-STA variation datasets.

Method	LLM	ActivityNet-Captions SS			ActivityNet-Captions VO			ActivityNet-Captions HY		
		R@0.1	R@0.3	R@0.5	R@0.1	R@0.3	R@0.5	R@0.1	R@0.3	R@0.5
SCN (Lin et al. 2020)	✗	74.29	46.46	26.92	74.41	46.44	26.96	74.36	46.51	27.01
CPL (Zheng et al. 2022b)	✗	73.03	47.90	26.86	73.58	47.87	27.13	73.19	47.98	26.94
OmniD (Ours)	✗	72.79	48.48	27.86	74.95	48.55	28.83	72.60	47.98	28.30
SCN (Lin et al. 2020)	✓	74.87	46.95	28.05	74.78	47.14	28.27	74.87	47.14	28.09
CPL (Zheng et al. 2022b)	✓	74.71	47.45	26.47	74.64	47.26	26.52	74.75	47.37	26.57
OmniD (Ours)	✓	**78.07**	**51.06**	**28.50**	**81.73**	**52.85**	**29.53**	**79.62**	**51.84**	**28.99**

Table 2: Comparisons with state-of-the-art methods on three ActivityNet-Captions variation datasets.

ural language queries paraphrased LLM. Generally, the utilization of LLM paraphrasing can boost the generalization capabilities of these methods when dealing with sentences exhibiting diverse variations. However, it is worth noting that such enhancements might be modest and occasionally unstable for their methods, For instance, LLM paraphrasing results in a mere one-point improvement for SCN within the Charades-STA HY dataset. In contrast to these approaches, the localization accuracy of our OmniD consistently and markedly improves with LLM paraphrasing. This improvement is attributed to the enforced consistency with semantic equivariance distillation, which sets our method apart.

2) ActivityNet-Captions variant datasets. Table 2 shows the performance comparison between the OmniD and the existing methods on *ActivityNet-Captions SS*, *ActivityNet-Captions VO*, and *ActivityNet-Captions HY*. Notably, OmniD consistently outperforms the state-of-the-art methods by a significant margin. For instance, our OmniD method surpasses CPL by more than 3 point on the ActivityNet-Captions VO dataset with LLM paraphrasing, in terms of "R@0.5". Moreover, the enhancement in performance achieved by incorporating LLM into OmniD is notably more consistent and pronounced compared to the effects observed when adding it to SCN and CPL. This demonstrates the efficacy of the proposed omnipotent distillation technique and the benefit of exploiting the semantic equivariance for the distillation.

3) Vanilla Charades-STA and ActivityNet-Captions datasets. We compare the localization accuracy on the original Charades-STA and ActivityNet-Captions datasets in Table 3. To ensure a fair comparison with prior research, where metrics are reported without the use of LLMs, we discard the LLM paraphrasing from our OmniD models and only use the original sentences for the distillation. Remarkably, both OmniD Teacher and Student models surpass state-of-the-art methods with a clear margin on both datasets. For instance, OmniD student model achieves more than an 8% enhancement on the Charades-STA dataset compared to the leading previous method CPL in value of R@0.7. And our teacher model is more than 5 points higher than CPL on the ActivityNet-Captions dataset in terms of R@0.3.

Ablation Studies

To investigate the effectiveness of the proposed algorithms, here we conduct ablation studies on the Charades-STA dataset and its variations.

1) The benefit of semantic equivariance distillation. Table 4 showcases the advantages of semantic equivariance distillation across three variant datasets of Charades-STA. We investigate the impact of degrading semantic equivariance distillation in the following two aspects: 1) Whether to use LLM: if we choose not to utilize the sentence queries paraphrased by the LLM, then we directly employ the original, unaltered sentences for the student model. 2) Whether to use the distillation loss Eq. 8: when we do not use Eq. 8, and instead we replace by the heuristic loss Eq. 2. The results indicate that omitting either of them (marked by a cross symbol in the respective column) noticeably leads to a decline in localization capability across all three variant datasets.

2) The advantage of intra-model consistency distillation learning. In this subsection, we delve into the benefits of bootstrapping learning through ablation studies. The summary of the ablation studies on the teacher model's localization accuracy on the original Charades-STA dataset is presented in Table 5. We refer to the complete teacher model as "full", the teacher model without intra-model consistency distillation as "full w/o icd", and the teacher model reduced

Method	Charades-STA			ActivityNet Captions		
	R@0.3	R@0.5	R@0.7	R@0.1	R@0.3	R@0.5
SCN (Lin et al. 2020)	42.96	23.58	9.97	71.48	47.23	29.22
BAR (Wu et al. 2020)	44.97	27.04	12.23	–	49.03	30.73
MARN (Song et al. 2020)	48.55	31.94	14.81	–	47.01	29.95
RTBPN (Zhang et al. 2020b)	60.04	32.36	13.24	73.73	49.77	29.63
CCL (Zhang et al. 2020c)	–	33.21	15.68	–	50.12	31.07
LCNet (Yang et al. 2021)	59.60	39.19	18.87	78.58	48.49	26.33
VCA (Wang, Chen, and Jiang 2021)	58.58	38.13	19.57	67.96	50.45	31.00
WSTAN (Wang et al. 2022)	43.39	29.35	12.28	79.78	52.45	30.01
CPL (Zheng et al. 2022b)	66.40	49.24	22.39	79.86	53.67	31.24
OmniD-Teacher (Ours)	**69.13**	**53.77**	**24.70**	**83.41**	**59.15**	**32.34**
OmniD-Student (Ours)	68.30	52.31	24.35	83.24	57.34	31.60

Table 3: Comparisons with state-of-the-art methods on the vanilla datasets.

LLM	Distill	Charades-STA SS			Charades-STA VO			Charades-STA HY		
		R@0.3	R@0.5	R@0.7	R@0.3	R@0.5	R@0.7	R@0.3	R@0.5	R@0.7
✓	✓	**65.66**	**49.09**	**22.93**	**64.78**	**48.63**	**22.30**	**65.09**	**49.04**	**22.83**
✓	✗	61.18	44.65	19.77	59.07	43.15	19.53	60.75	44.55	20.23
✗	✓	56.91	41.41	18.64	49.78	35.66	16.47	51.80	36.57	16.81
✗	✗	54.43	39.98	18.80	47.25	34.53	16.42	49.28	35.92	16.80

Table 4: Ablation studies of equivariance distillation on Charades-STA variation datasets.

Method	R@0.3	R@0.5	R@0.7
full	**69.13**	**53.77**	**24.70**
full w/o icd	67.48	51.68	23.50
full w/o bootstrap	66.50	50.44	22.86

Table 5: Ablation studies on intra-model consistency.

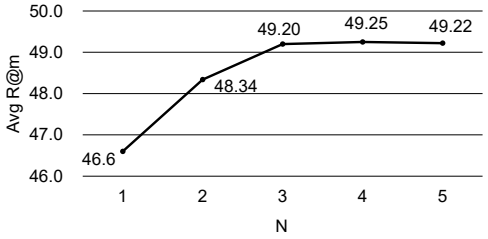

Figure 3: Ablation studies on the sub-model numbers N.

to a single sub-model without bootstrapping learning as "full w/o bootstrap." When dropping intra-model consistency distillation loss, we find that the performance decreases about one to two points. Upon removing the intra-model consistency distillation loss i.e., "full w/o icd", we observe a decline of approximately one to two points in each evaluation metric. This underscores that the enhancement of the teacher model stems not only from the ensemble effect, but further from the mutual learning among the sub-models. Moreover, downgrading the bootstrapping learning to a single sub-model results in a drop of around 3 points. This also verifies that the consensus of the N sub-models' prediction can provide useful training guidance for the single sub-model.

3) The influence of hyperparameter N in bootstrapping learning. The teacher model $\mathcal{T}$ contains N sub-models, with the hyperparameter N playing a crucial role in the bootstrapping learning. Fig 3 presents the impact of N on the $\mathcal{T}$'s localization accuracy on vanilla Charades-STA. As N increases, the average of "R@m" where m=$\{0.3, 0.5, 0.7\}$ gradually becomes larger when $N < 3$. However, beyond $N > 3$, the model's accuracy reaches saturation. This can be attributed to the fact that there is little additional complementary information provided as the number of sub-models is sufficient.

Conclusion

This paper for the first time leverages the large language model (LLM) for weakly-supervised natural language video localization. We propose omnipotent distillation with LLM to resolve the key obstacles of this task. Firstly, a bootstrapping learning framework with intra-model consistency is devised to alleviate the insufficient supervision limitation. Secondly, we capitalize on the LLM to paraphrase the language query and then distill the teacher model to an efficient student model using the semantic equivariance property of paraphrasing. To assess the generalization of the model across diverse queries, we create extensive datasets with different types of variations. Experiments demonstrate our method achieves state-of-the-art results in both adaptability and performance.

Acknowledgements

This work was carried out at the Rapid-Rich Object Search (ROSE) Lab, School of EEE, NTU, Singapore. The research is supported in part by the NTU-PKU Joint Research Institute (a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation).

References

Bao, P.; and Mu, Y. 2022. Learning Sample Importance for Cross-Scenario Video Temporal Grounding. In *ICMR*.

Bao, P.; Xia, Y.; Yang, W.; Ng, B. P.; Er, M. H.; and Kot, A. C. 2024. Local-Global Multi-Modal Distillation for Weakly-Supervised Temporal Video Grounding. In *AAAI*.

Bao, P.; Yang, W.; Ng, B. P.; Er, M. H.; and Kot, A. C. 2023. Cross-modal Label Contrastive Learning for Unsupervised Audio-Visual Event Localization. In *AAAI*.

Bao, P.; Zheng, Q.; and Mu, Y. 2021. Dense Events Grounding in Video. In *AAAI*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; and et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*.

Chen, Y.-C.; Lin, Y.-Y.; Yang, M.-H.; and Huang, J.-B. 2019. CrDoCo: Pixel-Level Domain Transfer With Cross-Domain Consistency. In *CVPR*.

Chen, Z.; Ma, L.; Luo, W.; Tang, P.; and Wong, K.-Y. K. 2020. Look Closer to Ground Better: Weakly-Supervised Temporal Grounding of Sentence in Video. *ArXiv*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; and et al. 2022. PaLM: Scaling Language Modeling with Pathways. *ArXiv*.

Deng, Z.; Luo, Y.; and Zhu, J. 2019. Cluster Alignment With a Teacher for Unsupervised Domain Adaptation. In *ICCV*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *ICCV*.

Gao, M.; Davis, L. S.; Socher, R.; and Xiong, C. 2019. WSLLN:Weakly Supervised Natural Language Localization Networks. In *EMNLP*.

Gupta, S.; Hoffman, J.; and Malik, J. 2015. Cross Modal Distillation for Supervision Transfer. In *CVPR*.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *ArXiv*, abs/1503.02531.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *ICCV*.

Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2020. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network. In *AAAI*.

Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018. Cross-modal moment localization in videos. In *ACM MM*.

Micaelli, P.; and Storkey, A. J. 2019. Zero-shot Knowledge Transfer via Adversarial Belief Matching. In *NeurIPS*.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly Supervised Video Moment Retrieval From Text Queries. In *CVPR*.

Nayak, G. K.; Mopuri, K. R.; Shaj, V.; Babu, R. V.; and Chakraborty, A. 2019. Zero-Shot Knowledge Distillation in Deep Networks. In *ICCV*.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*.

Qi, M.; Qin, J.; Yang, Y.; Wang, Y.; and Luo, J. 2021. Semantics-Aware Spatial-Temporal Binaries for Cross-Modal Video Retrieval. *TIP*.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. K. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *ECCV*.

Song, Y.; Wang, J.; Ma, L.; Yu, Z.; and Yu, J. 2020. Weakly-Supervised Multi-Level Attentional Reconstruction Network for Grounding Textual Queries in Videos. *ArXiv:2003.07048*.

Sreenu, G.; and Durai, M. A. S. 2019. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data*.

Tan, R.; Xu, H.; Saenko, K.; and Plummer, B. A. 2021. LoGAN: Latent Graph Co-Attention Network for Weakly-Supervised Video Moment Retrieval. In *WACV*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; and et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*.

Tran, D.; Bourdev, L. D.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*.

Wang, J.; Ma, L.; and Jiang, W. 2020. Temporally Grounding Language Queries in Videos by Contextual Boundary-aware Prediction. In *AAAI*.

Wang, Q.; Zhan, L.; Thompson, P. M.; and Zhou, J. 2020. Multimodal Learning with Incomplete Modalities by Knowledge Distillation. In *KDD*.

Wang, W.; Huang, Y.; and Wang, L. 2019. Language-driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model. In *CVPR*.

Wang, Y.; Deng, J.; gang Zhou, W.; and Li, H. 2022. Weakly Supervised Temporal Adjacent Network for Language Grounding. *TMM*.

Wang, Z.; Chen, J.; and Jiang, Y.-G. 2021. Visual Co-Occurrence Alignment Learning for Weakly-Supervised Video Moment Retrieval. In *ACM MM*.

Wu, J.; Li, G.; Han, X.; and Lin, L. 2020. Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos. In *ACM MM*.

Yang, W.; Zhang, T.; Zhang, Y.; and Wu, F. 2021. Local Correspondence Network for Weakly Supervised Temporal Sentence Grounding. *TIP*.

Yu, B. X. B.; Liu, Y.; and Chan, K. C. C. 2021. Multimodal Fusion via Teacher-Student Network for Indoor Action Recognition. In *AAAI*.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.-F.; and Davis, L. S. 2019a. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020a. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI*.

Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. 2019b. Cross-Modal Interaction Networks for Query-Based Moment Retrieval in Videos. In *ACM SIGIR*.

Zhang, Z.; Lin, Z.; Zhao, Z.; Zhu, J.; and He, X. 2020b. Regularized Two-Branch Proposal Networks for Weakly-Supervised Moment Retrieval in Videos. In *ACM MM*.

Zhang, Z.; Zhao, Z.; Lin, Z.; Zhu, J.; and He, X. 2020c. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. In *NeurIPS*.

Zheng, M.; Huang, Y.; Chen, Q.; and Liu, Y. 2022a. Weakly Supervised Video Moment Localization with Contrastive Negative Sample Mining. In *AAAI*.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022b. Weakly Supervised Temporal Sentence Grounding with Gaussian-based Contrastive Proposal Learning. In *CVPR*.

Zhu, W.; Lu, J.; Li, J.; and Zhou, J. 2021. DSNet: A Flexible Detect-to-Summarize Network for Video Summarization. *TIP*.