Exploiting Intrinsic Multilateral Logical Rules for Weakly Supervised Natural Language Video Localization

Zhe Xu, Kun Wei, Xu Yang, Cheng Deng*
School of Electronic Engineering, Xidian University, Xi'an, China
zhexu@stu.xidian.edu.cn, {weikunsk,xuyang.xidian,chdeng.xd}@gmail.com

Abstract

Weakly supervised natural language video localization (WS-NLVL) aims to retrieve the moment corresponding to a language query in a video with only video-language pairs utilized during training. Despite great success, existing WS-NLVL methods seldomly consider the complex temporal relations enclosing the language query (*e.g.*, between the language query and sub-queries decomposed from it or its synonymous query), yielding illogical predictions. In this paper, we propose a novel plug-and-play method, Intrinsic Multilateral Logical Rules, namely IMLR, to exploit intrinsic temporal relations and logical rules for WS-NLVL. Specifically, we formalize queries derived from the original language query as the nodes of a directed graph, *i.e.*, intrinsic temporal relation graph (ITRG), and the temporal relations between them as the edges. Instead of directly prompting a pre-trained language model, a relation-guided prompting method is introduced to generate ITRG in a hierarchical manner. We customize four types of multilateral temporal logical rules (*i.e.*, identity, inclusion, synchronization, and succession) from ITRG and utilize them to train our model. Experiments demonstrate the effectiveness and superiority of our method on the Charades-STA and ActivityNet Captions datasets.

1 Introduction

Natural language video localization (NLVL) (Gao et al., 2017; Chen and Jiang, 2019; Xu et al., 2023b; Wang et al., 2023; Zhang et al., 2023) aims to locate the temporal interval that semantically corresponds to a language query in an untrimmed video (Zhang et al., 2023; Xu et al., 2023a; Lan et al., 2023). Due to its promising applications in various fields such as video editing and intelligent surveillance, NLVL has attracted increasing research interest in the last few years. Traditional NLVL models are trained under full supervision (Gao et al., 2017), *i.e.*, the accurate start and end timestamps are utilized for training. Despite remarkable success, it is extremely time-consuming and expensive to obtain the boundary annotations. To this end, weakly supervised natural language video localization (WS-NLVL) (Mithun et al., 2019) is introduced to reduce the annotation cost by utilizing only the video-language pairs for training.

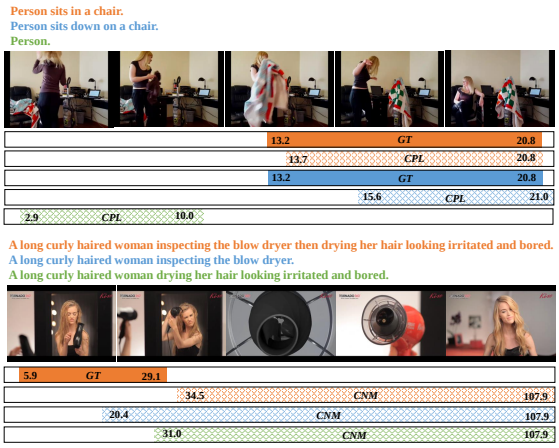

Figure 1: Examples of illogical predictions of start-of-the-art WS-NLVL methods, *i.e.*, CPL (Zheng et al., 2022b) and CNM (Zheng et al., 2022a). Ground-truths and predictions are represented by filled rectangles and grid lines, respectively. Best viewed in color.

Early methods of WS-NLVL are based on multiple instance learning (MIL) (Mithun et al., 2019; Chen et al., 2022), where each segment of the input video is treated as an instance and the input video as a bag of instances. The predictions of instances are aggregated to form the bag-level prediction. Recently, reconstruction-based methods (Zheng et al., 2022a,b; Huang et al., 2023; Cao et al., 2023; Lv et al., 2023; Yoon et al., 2023) are proposed to solve the task through joint learning with the reconstruction loss, assuming that the proposal corresponding to the language query should best reconstruct it.

*Corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4511–4521

Despite great success, existing WS-NLVL methods directly fuse the video features with word or sentence-level language features, which seldomly consider the complex temporal relations enclosing the language query and thus yield illogical predictions. As the first example shown in Fig.1, given an untrimmed video and a language query "Person sits in a chair", the localization result is supposed to be identical to that of its synonymous query "Person sits down on a chair" and be included by that of its sub-query "Person". However, as represented by grad lines in Fig.1, a start-of-the-art WS-NLVL method CPL (Zheng et al., 2022b) yields predictions far from logical. Fig.1 also presents another example of a language query "A long curly haired woman inspecting the blow dryer then drying her hair looking irritated and bored". The temporal intervals of its sub-queries, *i.e.*, "A long curly haired woman inspecting the blow dryer" and "A long curly haired woman drying her hair looking irritated and bored", should be temporally successive and integrated to that of the original language query. Unfortunately, the existing methods fail to capture such complex temporal relations.

To tackle the issue above, we propose a novel framework, dubbed IMLR, to exploit intrinsic temporal relations and logical rules for WS-NLVL. Concretely, we first propose to understand the complex language query by generating its intrinsic temporal relation graph (ITRG). In this graph, the original language query and queries derived from it are formulated as the nodes while the temporal relations between them are represented as the edges. We systematically consider all possible temporal relations for WS-NLVL and customize four types of relations with corresponding multilateral temporal logical rules (MTLRs), *i.e.*, identity, inclusion, synchronization, and succession. Since it is challenging to generate ITRG by manually designing rules or directly prompting (Brown et al., 2020) a pre-trained language model (LM), we then introduce a relation-guided prompting method to generate ITRG in a hierarchical manner. Finally, MTLRs from the graph are utilized to train our model of boundary-aware transformer. We conduct experiments to integrate our method with two recent WS-NLVL approaches, *i.e.*, CNM (Zheng et al., 2022a) and CPL (Zheng et al., 2022b), verifying the effectiveness and superiority of our method on the Charades-STA and ActivityNet Captions datasets.

Our contributions can be summarized as follows:

- We propose to exploit intrinsic temporal relations and customize multilateral logical rules for WS-NLVL.
- We present a novel framework by first generating an ITRG via relation-guided prompting and then leveraging MTLRs from ITRG to supervise the training of our model.
- We verify the effectiveness and superiority of our method on the Charades-STA and ActivityNet Captions datasets.

2 Related Works

2.1 Fully Supervised Natural Language Video Localization

Fully supervised natural language video localization (FS-NLVL) methods utilize accurate temporal boundary annotations for training, which can be divided into two categories: (1) proposal-based methods (Gao et al., 2017; Chen and Jiang, 2019) and (2) proposal-free methods (Yuan et al., 2019; Chen et al., 2020a; Mun et al., 2020; Chen et al., 2020b; Xu et al., 2022; Jang et al., 2023; Li et al., 2023; Zhang et al., 2020a; Wang et al., 2023). To be specific, proposal-based methods first generate a set of candidate proposals for a given video and then select the most relevant one. Since proposal-based methods need to compare all the proposals with the language query, their computational costs are extremely expensive. Therefore, proposal-free methods are proposed to treat NLVL as a regression problem and directly predict the time interval.

Despite remarkable success, it is time-consuming and expensive to obtain the temporal boundary annotations for FS-NLVL.

2.2 Weakly Supervised Natural Language Video Localization

Weakly supervised natural language video localization (WS-NLVL) (Mithun et al., 2019; Ma et al., 2020; Huang et al., 2021; Lin et al., 2020; Wu et al., 2020) is proposed to reduce the expensive cost of annotating temporal boundary for FS-NLVL. TGA (Mithun et al., 2019) presents text-guided attention to highlight video segments relevant to a language query and obtain a single text-dependent video feature. The network is trained by minimizing the distance between the text-dependent video feature and the language feature. VLANet (Ma et al., 2020) proposes a video-language alignment network to

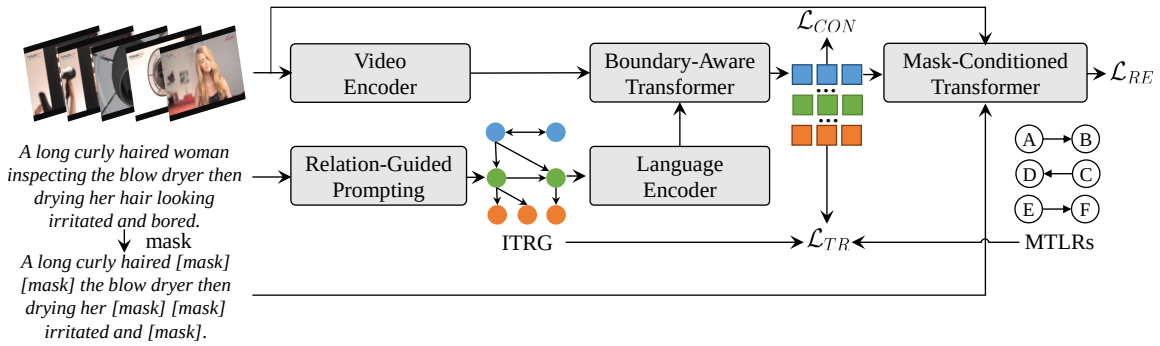

Figure 2: Overall framework of our method. Given an untrimmed video and a language query, we first generate an intrinsic temporal relation graph (ITRG) by relation-guided prompting. Then, two encoders and a boundary-aware transformer are employed to predict the keypoint and boundary offsets of the moment. The model parameters are jointly optimized using temporal relation loss $\mathcal{L}_{TR}$ with multilateral temporal logical rules (MTLRs), contrastive loss $\mathcal{L}_{CON}$, and reconstruction loss $\mathcal{L}_{RE}$.

prune out irrelevant proposals and consider various attention flows to learn multi-modal alignment. CPL (Zheng et al., 2022b) proposes contrastive proposal learning to use multiple Gaussian functions to generate both positive and negative proposals from the same video. Huang et al. (Huang et al., 2023) propose the first self-training-based method for WS-NLVL, which includes a pair of mutually learned teacher and student networks with weak and strong augmentation.

Although great progress has been made, existing WS-NLVL methods seldomly consider the complex temporal relations enclosing the language query and thus yield illogical predictions. In this paper, we propose a novel plug-and-play method to tackle the issue.

2.3 Temporal Relation Modeling in Video

There are several works leveraging the temporal relations for video understanding. TRN (Zhou et al., 2018) proposes to learn and reason about temporal dependencies between video frames at multiple time scales for action recognition. TRM (Zheng et al., 2023) uses phrase-level predictions to refine the sentence-level prediction for FS-NLVL, where accurate boundary annotations are utilized to mine the temporal relations between the phrase and sentence. CRM (Huang et al., 2021) explores the temporal order of different sentences in a paragraph for WS-NLVL, which is the most relevant work to our method. However, it requires additional temporal order annotations for training and is only suitable for sentences with multiple neighbor sentences in a paragraph. In contrast, we focus on the intrinsic temporal relations within a single sentence and systematically consider four types of multilateral temporal relations.

2.4 Prompt-Based Learning

Prompt-based learning (Liu et al., 2023; Brown et al., 2020; Wei et al., 2022; Li and Liang, 2021; Yao et al., 2023) is a new paradigm originated in natural language processing, which aims to perform prediction tasks by directly prompting a pre-trained model instead of finetuning a separate model checkpoint. Brown et al. (Brown et al., 2020) demonstrate that scaling up LMs greatly improves task-agnostic few-shot performance without any gradient updates or fine-tuning. Chain-of-thought prompting (Wei et al., 2022) improves the ability of LMs to perform complex reasoning with series of intermediate reasoning steps. In this paper, we introduce relation-guided prompting to generate ITRG for WS-NLVL, which models the complex intrinsic temporal relations of a language query.

3 Methodology

Given an untrimmed video V and a natural language query Q, NLVL aims to learn a model f that can predict the time interval $I = (t^s, t^e)$ in the video corresponding to the language query, *i.e.*, $f : (V, Q) \rightarrow I$, where t^s and t^e represent the start and end time, respectively. Different from FS-NLVL that utilizes accurate temporal boundaries $\{V^m, Q^m, I^m\}_{m=1}^M$ for training, only video-language pairs $\{V^m, Q^m\}_{m=1}^M$ are available for WS-NLVL. Here M denotes the number of samples in the training set.

Fig.2 shows the overall framework of our method for WS-NLVL, which consists of relation-guided prompting, video encoder, language en-

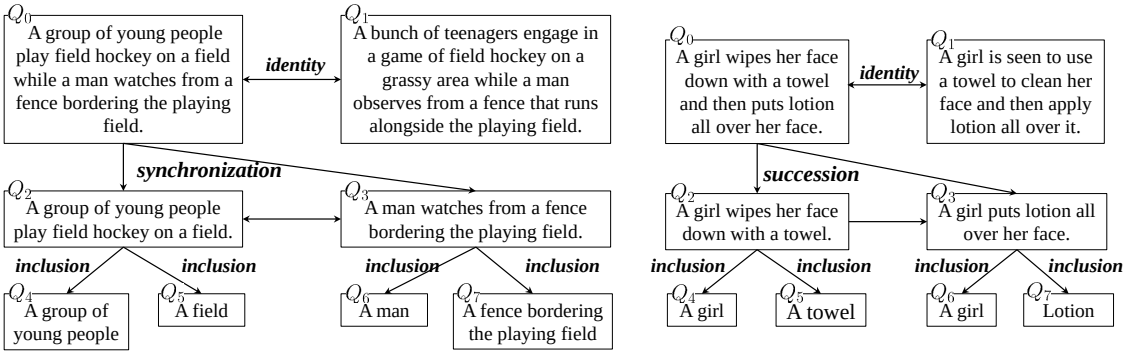

Figure 3: Two examples of ITRG. Q_i represents the index of node in its BFS ordering enumeration and Q_0 is the given language query.

coder, boundary-aware transformer, and mask-conditioned transformer. Given a language query, we first propose relation-guided prompting to generate its intrinsic temporal relation graph (ITRG). Then the video encoder, language encoder, and boundary-aware transformer are utilized to predict the keypoint and boundary offsets of the target moment. We formalize the multilateral temporal logical rules (MTLRs) for the temporal relation loss $\mathcal{L}_{TR}$, which is utilized together with the contrastive loss $\mathcal{L}_{CON}$ and reconstruction loss $\mathcal{L}_{RE}$ to train our model. In the following, we elaborate the main components of our method: (1) Definition of MTLRs, (2) Relation-guided prompting for ITRG generation, (3) WS-NLVL with MTLRs, and (4) Training and Inference.

3.1 Definition of MTLRs

Formally, given a complex language query, its ITRG is a directed graph $\mathcal{G}$. Each node $Q_i \in \mathcal{G}$ represents a query that is either the original query or a query derived from it (*e.g.*, its synonymous query or a sub-query decomposed from it) and the edge between queries denotes the temporal relation.

Fig.3 shows two examples of ITRG, where we enumerate the nodes of $\mathcal{G}$ with breadth-first-search (BFS) ordering and Q_0 is the given language query. For a language query Q_0 "A group of young people play field hockey on a field while a man watches from a fence bordering the playing field", we can obtain its synonymous query Q_1 "A bunch of teenagers engage in a game of field hockey on a grassy area while a man observes from a fence that runs alongside the playing field". Moreover, Q_0 can be decomposed into two sub-queries that temporally overlap with each other, *i.e.*, Q_2 "A group of young people play field hockey on a field" and Q_3 "A man watches from a fence bordering the playing field". In addition, Q_4 "A group of young people" Q_5 "A field" and Q_6 "A man" Q_7 "A fence bordering the playing field" can be extracted from Q_2 and Q_3, respectively. In the second example of a given language query Q_0 "A girl wipes her face down with a towel and then puts lotion all over her face", we can similarly generate its ITRG except that its sub-queries Q_2 "A girl wipes her face down with a towel" and Q_3 "A girl puts lotion all over her face" temporally occur in succession.

We systematically consider all possible intrinsic temporal relations for WS-NLVL and customize four types of relations with multilateral temporal logical rules (MTLRs) among different language queries, *i.e.*, *Identity*, *Inclusion*, *Synchronization*, and *Succession*.

Definition 1. (Identity) Given an untrimmed video V, a language query Q, and its synonymous query Q_{id}, the localization results of Q and Q_{id} are supposed to be identical with each other:

$$f(V, Q) = f(V, Q_{id}), \tag{1}$$

where f denotes the localization model.

Definition 2. (Inclusion) Given an untrimmed video V, a language query Q, and a noun phrase query Q_{np} extracted from Q, the localization result of Q should be included by that of Q_{np}:

$$f(V, Q) \subseteq f(V, Q_{np}). \tag{2}$$

Definition 3. (Synchronization) Given an untrimmed video V, a language query Q, and sub-queries $Q_{sy_1}, \cdots, Q_{sy_N}$ that are decomposed from Q and describe events temporally overlap with each other, the localization results of $Q, Q_{sy_1}, \cdots, Q_{sy_N}$ should subject to:

$$f(V, Q) = \bigcap_{n=1}^{N} f(V, Q_{sy_n}). \tag{3}$$

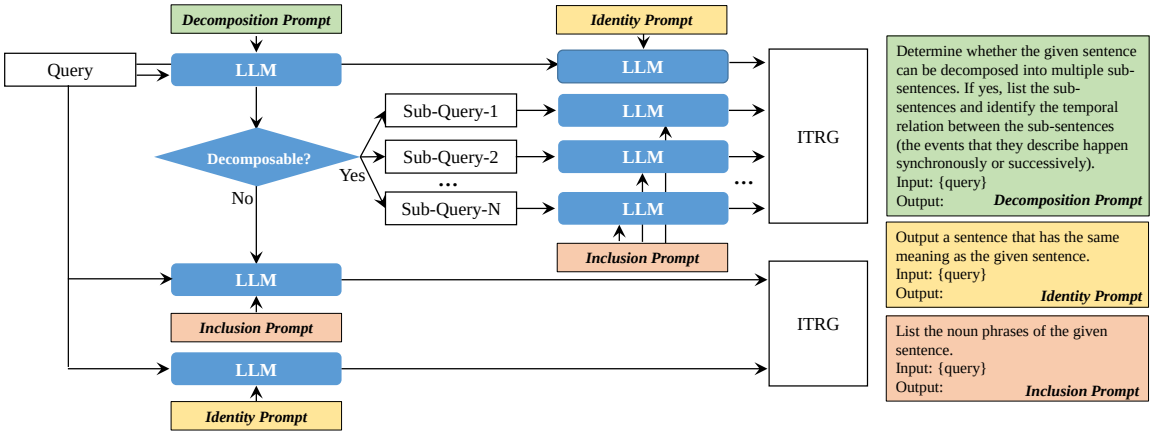

Figure 4: Illustration of relation-guided prompting for generating ITRG. We introdude three types of prompt templates, *i.e.*, decomposition prompt, identity prompt, and inclusion prompt, to hierarchically generate ITRG.

Definition 4. (Succession) Given an untrimmed video V, a language query Q, and sub-queries $Q_{su_1}, \cdots, Q_{su_N}$ that are decomposed from Q and describe events successively happen, the localization results of $Q, Q_{su_1}, \cdots, Q_{su_N}$ should satisfy:

$$f(V,Q) = \bigcup_{n=1}^{N} f(V, Q_{su_n}), \\ f(V, Q_{su_n}) \leq f(V, Q_{su_{n+1}}). \tag{4}$$

3.2 Relation-Guided Prompting for ITRG Generation

Due to the complex nature of the language query in NLVL, it is challenging to generate ITRG by manually designing rules. A naive solution is leveraging the powerful language understanding ability of a pre-trained LM of by prompting. However, directly prompting LM with few input-output pairs can not accurately capture the complex temporal relations enclosing a query.

To this end, we propose a relation-guided prompting method to generate ITRG in a hierarchical manner. Fig.4 shows the pipeline of our method. Specifically, we introduce three types of prompt templates, *i.e.*, decomposition prompt, identity prompt, and inclusion prompt. The decomposition prompt is utilized to determine whether the given language query can be decomposed into different sub-queries and list the sub-queries with their relations if applicable. The identity prompt outputs the synonymous query of the given language query. The inclusion prompt extracts the noun phrases of a query. The results of different prompts are integrated to generate ITRG.

3.3 WS-NLVL with MTLRs

Encoders. Given an untrimmed video, a language query, and its ITRG $\mathcal{G}$, we first extract video and query features using pre-trained models. Specifically, the video feature $V = \{v_1, v_2, \cdots, v_{L_V}\} \in \mathbb{R}^{L_V \times D_V}$ is encoded with pre-trained 3D convolutional network (Carreira and Zisserman, 2017; Tran et al., 2015), where L_V denotes the number of segments per video and D_V is the dimension of video segment feature. The query features $\{Q_i = \{w_{i,1}, w_{i,2}, \cdots, w_{i,L_Q}\} \in R^{L_Q \times D_Q}\}_{i=0}^{G-1}$ are obtained using pre-trained Glove (Pennington et al., 2014) model, where L_Q, D_Q, and G represent the number of words per sentence, the dimension of query feature, and the number of nodes in $\mathcal{G}$, respectively.

Then, the video and query features are embedded into a common latent space:

$$V \leftarrow \mathbf{W_V} V, \quad Q_i \leftarrow \mathbf{W_Q} Q_i, \tag{5}$$

where $\mathbf{W_V} \in \mathbb{R}^{D_H \times D_V}$ and $\mathbf{W_Q} \in \mathbb{R}^{D_H \times D_Q}$ are the learnable matrices and D_H is the dimension of the latent space.

Boundary-Aware Transformer. We introduce a boundary-aware transformer to perform multi-modal interaction and predict the keypoint and boundary offsets of the moment. To be specific, a learnable [CLASS] token v_{cls} is added at the end of the video feature, *i.e.*, $\hat{V} = \{v_1, v_2, \cdots, v_{L_V}, v_{cls}\} \in \mathbb{R}^{(L_V+1) \times D_H}$. We fuse the query features $\{Q_i \in \mathbb{R}^{L_Q \times D_H}\}_{i=0}^{G-1}$ and $\hat{V}$ by a dual transformer to obtain the hidden features $\{H_i = \{h_{1,i}, h_{2,i}, \cdots, h_{L_V,i}, h_{cls,i}\} \in \mathbb{R}^{(L_V+1) \times D_H}\}_{i=0}^{G-1}$:

$$H_i = \mathbf{D}(\hat{V}, \mathbf{E}(Q_i)), \tag{6}$$

where $\mathbf{E}$ and $\mathbf{D}$ represent the transformer encoder and decoder, respectively.

Different from existing methods that directly predict the center and width of the moment, we propose to predict the keypoint and boundary offsets using a prediction head $\mathbf{P}$:

$$k_i, \tilde{s}_i, \tilde{e}_i = \mathbf{P}(h_{cls,i}), \tag{7}$$

where k_i, $\tilde{s}_i$, and $\tilde{e}_i$ represent the normalized keypoint coordinate, start offset, and end offset, respectively. This models the intrinsic temporal relations of the moment from an architecture perspective. The start and end timestamps thus are $s_i = k_i - \tilde{s}_i$ and $e_i = k_i + \tilde{e}_i$. The center and width of the target moment can be re-formulated as $c_i = k_i + (\tilde{e}_i - \tilde{s}_i)/2$ and $w_i = \tilde{s}_i + \tilde{e}_i$.

After obtaining the predictions of different language queries, we apply the MTLRs in Sec.3.1 to guide the model training. Formally, the identity loss $\mathcal{L}_{ID}$ enforces the predicted temporal intervals of the given language query Q_0 to be identical to that of its synonymous query Q_{id}:

$$\mathcal{L}_{ID} = \|s_0 - s_{id}\|_F^2 + \|e_0 - e_{id}\|_F^2, \tag{8}$$

where s_0, e_0 and s_{id}, e_{id} represent the predictions of Q_0 and Q_{id}, respectively. $\|\cdot\|_F$ denotes the Frobenius norm.

The inclusion loss $\mathcal{L}_{IN}$ encourages the predictions of Q_i to be included by that of a noun phrase Q_{np} extracted from it, which can be formulated as:

$$\begin{aligned}\mathcal{L}_{IN} = \mathop{\mathbb{E}}_{\forall(Q_i,Q_{np})\in\mathcal{G}} &[\max(s_{np} - s_i + \gamma, 0) \\ &+ \max(e_i - e_{np} + \gamma, 0)],\end{aligned} \tag{9}$$

where γ is a hyperparameter.

The synchronization loss $\mathcal{L}_{SY}$ is employed to enforce the predictions of Q_0 to be consistent with the intersection of $Q_{sy_1}, \cdots, Q_{sy_N}$ that are decomposed from Q_0 and temporally overlap:

$$\begin{aligned}\mathcal{L}_{SY} = \mathop{\mathbb{E}}_{\forall(Q_0,Q_{sy_{1,\cdots,N}})\in\mathcal{G}} &[\|s_0 - \max(s_{sy_{1,\cdots,N}})\|_F^2 \\ &+ \|e_0 - \min(e_{sy_{1,\cdots,N}})\|_F^2].\end{aligned} \tag{10}$$

Moreover, the succession loss $\mathcal{L}_{SU}$ is utilized to model the temporal relations among Q_0 and $Q_{su_1}, \cdots, Q_{su_N}$ that are decomposed from Q_0 and successively happen:

$$\begin{aligned}\mathcal{L}_{SU} = \mathop{\mathbb{E}}_{\forall(Q_0,Q_{su_{1,\cdots,N}})\in\mathcal{G}} &[\|s_0 - \min(s_{su_{1,\cdots,N}})\|_F^2 \\ &+ \|e_0 - \max(e_{su_{1,\cdots,N}})\|_F^2 + \|e_{su_n} - s_{su_{n+1}}\|_F^2].\end{aligned} \tag{11}$$

Method	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
TGA	32.14	19.94	8.84	-
SCN	42.96	23.58	9.97	-
CTF	39.8	27.3	12.9	27.3
BAR	44.97	27.04	12.23	-
LoGAN	51.67	34.68	14.54	-
VLANet	45.24	31.83	14.17	-
CRM*	53.66	34.76	16.37	-
VCA	58.58	38.13	19.57	38.49
LCNet	59.60	39.19	18.87	38.94
RTBPN	60.04	32.26	13.24	-
Huang et.al.	69.16	52.18	23.94	45.20
SCANet*	68.04	50.85	24.04	-
CCR	68.59	50.79	23.75	44.66
CNM (ori.)	60.04	35.15	14.95	38.11
CNM (rep.)	59.31	35.37	14.91	38.01
CNM+Ours	63.06 (**+3.75**)	36.53 (**+1.16**)	16.45 (**+1.54**)	39.94 (**+1.93**)
CPL (ori.)	66.40	49.24	22.39	43.48
CPL (rep.)	66.58	49.55	22.88	43.65
CPL+Ours	70.42 (**+3.84**)	51.87 (**+2.32**)	24.67 (**+1.79**)	45.64 (**+1.99**)

Table 1: Performance comparison with state-of-the-art methods on Charades-STA. * indicates additional paragraph-level annotations are utilized.

The overall temporal relation loss $\mathcal{L}_{TR}$ is summarized as:

$$\mathcal{L}_{TR} = \lambda_1\mathcal{L}_{ID} + \lambda_2\mathcal{L}_{IN} + \lambda_3\mathcal{L}_{SY} + \lambda_4\mathcal{L}_{SU}, \tag{12}$$

where λ_1, λ_2, λ_3, and λ_4 are weighting coefficients.

3.4 Training and Inference

We build our method upon two recent WS-NLVL models, *i.e.*, CNM (Zheng et al., 2022a) and CPL (Zheng et al., 2022b), that are publicly accessible. In addition to the temporal relation loss $\mathcal{L}_{TR}$, a contrastive loss $\mathcal{L}_{CON}$ and a reconstruction loss $\mathcal{L}_{RE}$ are also employed for training. To be specific, $\mathcal{L}_{CON}$ is utilized to contrast between positive and negative proposals while $\mathcal{L}_{RE}$ measures the differences between reconstructed query using positive proposal and original query. For inference, we follow the same pipeline as the base methods.

4 Experiments

4.1 Datasets

Following the common practice of previous works (Mithun et al., 2019; Ma et al., 2020; Zheng et al.,

Method	IoU@0.1	IoU@0.3	IoU@0.5	mIoU
WS-DEC	62.71	41.98	23.34	-
SCN	71.48	47.23	29.22	-
CTF	74.2	44.3	23.6	32.2
CCL	-	50.12	31.07	-
BAR	-	49.03	30.73	-
CRM*	81.61	55.26	32.19	
VCA	67.96	50.45	31.00	33.15
LCNet	78.58	48.49	26.33	34.29
RTBPN	73.73	49.77	29.63	-
Huang et.al.	82.10	58.07	36.91	41.02
SCANet*	83.62	56.07	31.53	-
CCR	80.32	53.21	30.39	-
CNM (ori.)	79.74	54.61	30.26	36.59
CNM (rep.)	80.53	54.77	29.18	36.43
CNM+Ours	83.91 (**+3.38**)	55.97 (**+1.20**)	31.90 (**+2.72**)	38.14 (**+1.71**)
CPL (ori.)	79.86	53.67	31.24	-
CPL (rep.)	79.46	52.27	29.93	35.93
CPL+Ours	82.67 (**+3.21**)	54.01 (**+1.74**)	30.57 (**+0.64**)	37.68 (**+1.75**)

Table 2: Performance comparison with state-of-the-art methods on ActivityNet Captions. * indicates additional paragraph-level annotations are utilized.

2022a,b), we evaluate the proposed method on two public datasets: ActivityNet Captions (Krishna et al., 2017) and Charades-STA (Gao et al., 2017).

ActivityNet Captions. The ActivityNet Captions dataset contains 37,417, 17,505, and 17,031 video-language pairs for training, validation, and testing, respectively. The average length of language queries and the average duration of videos are 13.48 words and 117.6 seconds, respectively. We use the validation set for evaluation since the testing set is not publicly available.

Charades-STA. The Charades-STA dataset is built on Charades dataset for NLVL. It contains 12,408 and 3,720 video-language pairs for training and testing, respectively. The average length of language queries and the average duration of videos are 8.6 words and 29.8 seconds respectively.

4.2 Evaluation Metrics

We follow previous works (Zheng et al., 2022a; Huang et al., 2023) to utilize IoU@n and mIoU to measure the performance of WS-NLVL:

IoU@n is referred to as the percentage of test samples whose top-1 Intersection over Union (IoU) with ground-truth (GT) is higher than n. We report $n = \{0.1, 0.3, 0.5\}$ for the ActivityNet Captions dataset and $n = \{0.3, 0.5, 0.7\}$ for the Charades-STA dataset in our experiments.

mIoU denotes the mean IoU of all test samples.

4.3 Implementation Details

For data pre-processing, we follow previous works (Zheng et al., 2022b; Huang et al., 2023; Yoon et al., 2023) to use pre-trained C3D (Tran et al., 2015) and I3D (Carreira and Zisserman, 2017) models to extract video features for the ActivityNet Captions and Charades-STA datasets, respectively. Glove (Pennington et al., 2014) is employed to extract word features of the language query. We set the maximum description length to 20 and the maximum number of segments per video to 200. The vocabulary sizes for the ActivityNet Captions and Charades-STA datasets are 8,000 and 1,111, respectively. Moreover, we prompt Phi-2 [1], a publicly available LM, to generate the ITRG.

Our approach is implemented with PyTorch (Paszke et al., 2019) and optimized by ADAM (Kingma and Ba, 2014) optimizer with a learning rate of 0.0004. γ in Eq.9, $\lambda_1, \lambda_2, \lambda_3$ and λ_4 in Eq.12 are determined by the grid search and set to 0.15, 20, 20, 10, and 10, respectively. For the base models, we use the hyperparameters from their official implementation [2][3].

4.4 Comparisons with the State-of-the-Art

We compare our method with the following state-of-the-art ones: TGA (Mithun et al., 2019), WS-DEC (Duan et al., 2018), SCN (Lin et al., 2020), CTF (Chen et al., 2020c), BAR (Wu et al., 2020), VLANet (Ma et al., 2020), RTBPN (Zhang et al., 2020b), LoGAN (Tan et al., 2021), CRM (Huang et al., 2021), VCA (Wang et al., 2021), LCNet (Yang et al., 2021), CNM (Zheng et al., 2022a), CPL (Zheng et al., 2022b), Huang et al. (Huang et al., 2023), SCANet (Yoon et al., 2023), and CCR (Lv et al., 2023). Notably, it is unfair to directly compare CRM and SCANet with other methods since they requires additional paragraph-level annotations for training.

Tab.1 and Tab.2 present the results on the Charases-STA and ActivityNet Captions datasets, respectively. We integrate our method with two most recent WS-NLVL models that release their

[1]https://huggingface.co/microsoft/phi-2
[2]https://github.com/minghangz/cpl
[3]https://github.com/minghangz/cnm

Method	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
Base Model	66.58	49.55	22.88	43.65
+BAT	68.14	50.63	24.38	44.79
+ITR	69.02	51.03	24.54	45.04
+BAT+ITR	**70.42**	**51.87**	**24.67**	**45.64**

Table 3: Effectiveness of each component of our method. BAT and ITR represent the boundary-aware transformer and intrinsic temporal relation, respectively.

Method	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
Base Model	68.14	50.63	24.38	44.79
+$\mathcal{L}_{ID}$	69.11	51.05	24.18	45.12
+$\mathcal{L}_{IN}$	69.52	**51.68**	**24.64**	**45.36**
+$\mathcal{L}_{SY}$	**70.72**	51.08	24.50	45.33
+$\mathcal{L}_{SU}$	70.33	50.89	24.53	45.13

Table 4: Gain of each type of temporal logical rule.

source codes, *i.e.*, CNM and CPL. CNM (ori.) and CPL (ori.) denote the results reported in their original papers, while CNM (rep.) and CPL (rep.) represent our reproduced ones. As shown in the tables, our method can largely improve the performances of the base models and even outperform those with additional annotations used during training.

4.5 Ablation Study

Effectiveness of each component of our method. We perform an ablation study to investigate the effectiveness of each component of our method. Four variants of our model are compared: (1) Base Model, (2) +BAT, (3) +ITR, and (4) +BAT+ITR. To be specific, Base Model directly predicts the center and width of the target moment utilizing the contrastive loss $\mathcal{L}_{CON}$ and reconstruction loss $\mathcal{L}_{RE}$. +BAT denotes the method with a boundary-aware transformer, which predicts the keypoint and boundary offsets. +ITR represents the method trained with intrinsic temporal relation modeling, *i.e.*, additional temporal relation loss $\mathcal{L}_{TR}$. +BAT+ITR is our full model with both the boundary-aware transformer and intrinsic temporal relation modeling.

Tab.3 shows the results of different models on the Charades-STA dataset. Adding both the BAT and ITR can improve the base model by a large margin, *e.g.*, 1.56% and 2.44% in terms of IoU@0.3. The best performance is obtained when they are simultaneously utilized, verifying the effectiveness of the components of our method.

Method	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
VLSNet	70.46	54.19	35.22	50.02
+$\mathcal{L}_{TR}$	**72.40**	**56.54**	**36.51**	**51.93**

Table 5: Generality of our method.

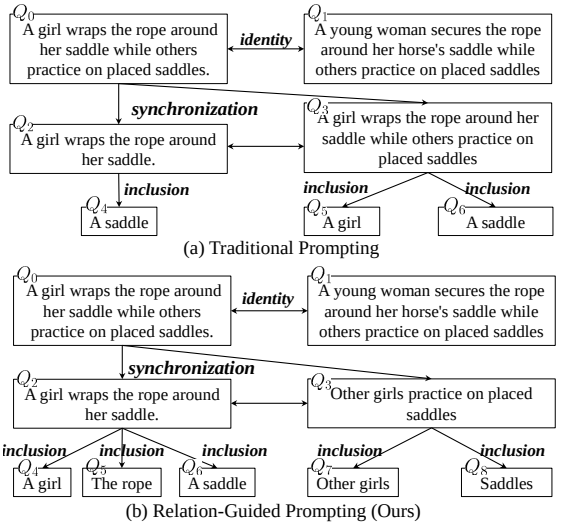

Figure 5: Comparison between our relation-guided prompting and traditional few-shot prompting.

Gain of each type of temporal logical rule. We customize four types of multilateral temporal logical rules (*i.e.*, identity, inclusion, synchronization, and succession) and utilize them to train our model. In this experiment, we investigate the influence of them and present the results on the Charades-STA dataset in Tab.4. Each type of relation can significantly improve the performance of the base model, verifying their effectiveness.

Effectiveness of the relation-guided prompting. We propose relation-guided prompting to generate ITRG in a hierarchical manner. To verify its effectiveness, in this experiment, we compare our method with traditional few-shot prompting (Brown et al., 2020) that directly prompts LM using a few input-output pairs (20 prompts in our experiment). We present an example in Fig.5. Given a language query "A girl wraps the rope around her saddle while others practice on placed saddles.", traditional prompting fails to decompose it into synchronous sub-queries while our method can successfully capture the complex intrinsic temporal relations.

Generality of the method. We focus on the WS-NLVL task since it requires only video-language pairs for training and thus is more suitable for real-world scenarios. To investigate the gener-

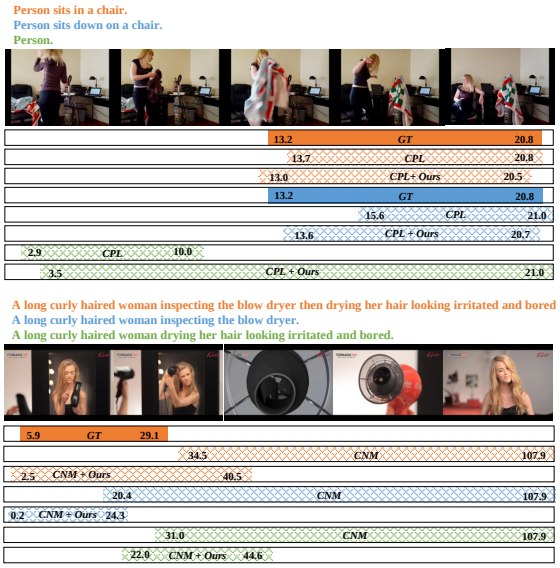

Figure 6: Qualitative results on the Charades-STA (top) and ActivityNet Captions (bottom) datasets.

ality of our method, we integrate our method with a fully-supervised method, VSLNet (Zhang et al., 2020a). Tab.5 shows the results on the Charades-STA dataset, verifying the generality of our method.

4.6 Qualitative Results

We present some qualitative results in Fig.6. Our method can improve the performance of both the CPL and CNM models. In addition, our method yields more logical predictions than both the base models, demonstrating the effectiveness of boundary-aware transformer and intrinsic temporal relation modeling.

5 Conclusion and Discussion

In this paper, we propose to exploit intrinsic temporal relations and multilateral logical rules for WS-NLVL. Language queries derived from the given one are formalized as the nodes of a directed graph and the temporal relations between them as the edges. We introduce relation-guided prompting to hierarchically generate the graph by leveraging a pre-trained LM. Four types of multilateral temporal logical rules (*i.e.*, identity, synchronization, succession, and inclusion) are customized to guide the training of our models. Extensive experiments on the Charades-STA and ActivityNet Captions datasets demonstrate the effectiveness and superiority of our method.

Limitations. Since only the video-language pairs are available during training, the performance of our method is still lag from several state-of-the-art fully supervised approaches trained with accurate annotations.

Acknowledgments

Our work is supported in part by the National Key R&D Program of China (No. 2023YFC3305600), National Natural Science Foundation of China (62132016 and 62201436), and Fundamental Research Funds for the Central Universities (ZDRC2102).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

Meng Cao, Fangyun Wei, Can Xu, Xiubo Geng, Long Chen, Can Zhang, Yuexian Zou, Tao Shen, and Daxin Jiang. 2023. Iterative proposal refinement for weakly-supervised video grounding. In *CVPR*, pages 6524–6534.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308.

Jiaming Chen, Weixin Luo, Wei Zhang, and Lin Ma. 2022. Explore inter-contrast between videos via composition for weakly supervised temporal sentence grounding. In *AAAI*, volume 36, pages 267–275.

Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020a. Rethinking the bottom-up framework for query-based video localization. In *AAAI*, volume 34, pages 10551–10558.

Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. 2020b. Learning modality interaction for temporal sentence localization and event captioning in videos. In *ECCV*, pages 333–351. Springer.

Shaoxiang Chen and Yu-Gang Jiang. 2019. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, volume 33, pages 8199–8206.

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

Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. *NeurIPS*, 31.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275.

Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *ICCV*, pages 7199–7208.

Yifei Huang, Lijin Yang, and Yoichi Sato. 2023. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *CVPR*, pages 18908–18918.

Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. 2023. Knowing where to focus: Event-aware transformer for video grounding. In *ICCV*, pages 13846–13856.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *ICCV*, pages 706–715.

Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. 2023. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–33.

Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. 2023. Momentdiff: Generative video moment retrieval from random to real. *arXiv preprint arXiv:2307.02869*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, volume 34, pages 11539–11546.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):1–35.

Zezhong Lv, Bing Su, and Ji-Rong Wen. 2023. Counterfactual cross-modality reasoning for weakly supervised video moment localization. In *ACM MM*, pages 6539–6547.

Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. 2020. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *ECCV*, pages 156–171. Springer.

Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *CVPR*, pages 11592–11601.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *CVPR*, pages 10810–10819.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 2021. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *WACV*, pages 2083–2092.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497.

Jing Wang, Aixin Sun, Hao Zhang, and Xiaoli Li. 2023. Ms-detr: Natural language video localization with sampling moment-moment interaction. *ACL*.

Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. 2021. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *ACM MM*, pages 1459–1468.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837.

Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. 2020. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *ACM MM*, pages 1283–1291.

Zhe Xu, Da Chen, Kun Wei, Cheng Deng, and Hui Xue. 2022. Hisa: Hierarchically semantic associating for video temporal grounding. *IEEE Trans. on Image Process.*, 31:5178–5188.

Zhe Xu, Kun Wei, Erkun Yang, Cheng Deng, and Wei Liu. 2023a. Bilateral relation distillation for weakly supervised temporal action localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):11458–11471.

Zhe Xu, Kun Wei, Xu Yang, and Cheng Deng. 2023b. Point-supervised video temporal grounding. *IEEE Trans. Multimed.*, 25:6121–6131.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Sunjae Yoon, Gwanhyeong Koo, Dahyun Kim, and Chang D Yoo. 2023. Scanet: Scene complexity aware network for weakly-supervised video moment retrieval. In *ICCV*, pages 13576–13586.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, volume 33, pages 9159–9166.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. Span-based localizing network for natural language video localization. In *ACL*, pages 6543–6554.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2023. Temporal sentence grounding in videos: A survey and future directions. *IEEE Trans. Pattern Anal. Mach. Intell.*

Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. 2020b. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACM MM*, pages 4098–4106.

Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. 2022a. Weakly supervised video moment localization with contrastive negative sample mining. In *AAAI*, volume 36, pages 3517–3525.

Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. 2022b. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *CVPR*, pages 15555–15564.

Minghang Zheng, Sizhe Li, Qingchao Chen, Yuxin Peng, and Yang Liu. 2023. Phrase-level temporal relationship mining for temporal sentence localization. In *AAAI*.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2018. Temporal relational reasoning in videos. In *ECCV*, pages 803–818.