Commonsense for Zero-Shot Natural Language Video Localization

Meghana Holla[1], Ismini Lourentzou[2],

[1]Department of Computer Science, Virginia Tech
[2]School of Information Sciences, University of Illinois at Urbana - Champaign
mmeghana@vt.edu, lourent2@illinois.edu

Abstract

Zero-shot Natural Language-Video Localization (NLVL) methods have exhibited promising results in training NLVL models exclusively with raw video data by dynamically generating video segments and pseudo-query annotations. However, existing pseudo-queries often lack grounding in the source video, resulting in unstructured and disjointed content. In this paper, we investigate the effectiveness of commonsense reasoning in zero-shot NLVL. Specifically, we present CORONET, a zero-shot NLVL framework that leverages commonsense to bridge the gap between videos and generated pseudo-queries via a commonsense enhancement module. CORONET employs Graph Convolution Networks (GCN) to encode commonsense information extracted from a knowledge graph, conditioned on the video, and cross-attention mechanisms to enhance the encoded video and pseudo-query representations prior to localization. Through empirical evaluations on two benchmark datasets, we demonstrate that CORONET surpasses both zero-shot and weakly supervised baselines, achieving improvements up to 32.13% across various recall thresholds and up to 6.33% in mIoU. These results underscore the significance of leveraging commonsense reasoning for zero-shot NLVL.

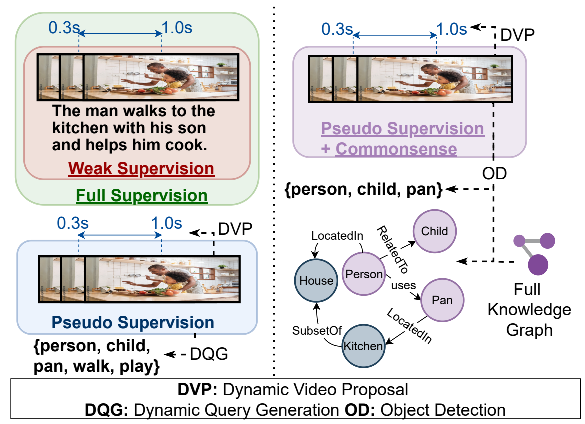

Figure 1: NLVL tasks under various supervision settings. Color-coded boxes show the expected annotations at each supervision level. Full supervision: Temporal Video Annotations + Text Queries; Weak Supervision: Text Queries; Pseudo-Supervision: Only Raw Videos. Makes use of DVP + DQG; CORONET (Ours, right) Only Raw Videos. DVP + OD + video-grounded commonsense knowledge subgraph.

Introduction

Natural Language Video Localization (NLVL) is a fundamental multimodal understanding task that aims to align textual queries with relevant video segments. NLVL is a core component for various applications such as video moment retrieval (Cao et al. 2022), video question answering (Qian et al. 2023; Lei et al. 2020a), and video editing (Gao et al. 2022). Prior works have primarily explored supervised (Zeng et al. 2020; Wang, Ma, and Jiang 2020; Soldan et al. 2021; Liu et al. 2021; Yu et al. 2020; Gao et al. 2021) or weakly supervised (Mun, Cho, and Han 2020; Zhang et al. 2020, 2021) NLVL methodologies, relying on annotated video-query data to various extents.

Obtaining annotated data for NLVL is a labor-intensive process that requires video samples paired with meticulous annotations of video moments and corresponding textual descriptions. Figure 1 illustrates the annotation requirements for different levels of supervision in NLVL. Fully supervised methods demand fine-grained moment span annotations, while weakly supervised methods typically rely on query descriptions alone. Nevertheless, both still heavily rely on paired video-language data, which limits practicality in open-domain settings.

Recent works formulate zero-shot NLVL, which aims to dynamically generate video moments and their corresponding queries, eliminating the need for paired video-query data (Nam et al. 2021; Kim et al. 2023). Nonetheless, existing approaches have certain limitations. On one hand, recent methods generate pseudo-queries using off-the-shelf object detectors for objects (nouns) and text-based language models for actions (verbs), resulting in noisy pseudo-queries that lack grounding in the video content (Nam et al. 2021). On the other hand, language-free methods remove pseudo-queries entirely by utilizing vision-language models pretrained on large-scale image and text datasets (Kim et al. 2023). However, eliminating textual information entirely may lead to missing out on important semantic nuances.

Visual (video) and textual (query) modalities provide very distinct but complementary types of information; videos provide spatial and physical information, while queries provide situational and contextual information. Existing works focus on complex vision-language interactions for observed video-query pairs in an attempt to bridge this gap (Nam et al. 2021; Mun, Cho, and Han 2020). However, in the zero-shot/pseudo-supervised setting, where queries are in a simpler form without structural information, finding common ground between modalities becomes crucial for effective cross-modal interactions. Commonsense knowledge, which encompasses general knowledge about the world and relationships between concepts, has proven valuable in various tasks (Fang et al. 2020; Ding et al. 2021; Yu et al. 2021; Li, Niu, and Zhang 2022; Maharana and Bansal 2021; Cao et al. 2022). By incorporating commonsense information, NLVL models could potentially bridge the semantic gap between video and text modalities, enhancing the cross-modal understanding and performance in zero-shot NLVL.

To this end, this work introduces **C**omm**O**nsense ze**R**o sh**O**t la**N**guage vid**E**o localiza**T**ion (CORONET), a zero-shot NLVL model that leverages commonsense knowledge to enhance the pseudo-query generation and cross-modal localization of video moments. We introduce a Commonsense Enhancement Module to enrich the encoded video and query representations with rich contextual information and employ external commonsense knowledge from ConceptNet (Speer, Chin, and Havasi 2017) to extract relevant relationships between a predefined set of concepts, mined from the input videos. Our primary objective is to investigate the potential benefits and challenges of leveraging commonsense for zero-shot NLVL. By jointly incorporating commonsense knowledge, we show that our model effectively bridges the gap between visual and linguistic modalities.

The contributions of this work are summarized as follows: **(1)** We introduce CORONET[1], a zero-shot NLVL framework that utilizes external commonsense knowledge to enrich cross-modal understanding between the visual and natural language components of pseudo-query generation. To the best of our knowledge, we are the first to incorporate commonsense information in zero-shot natural language video localization. **(2)** CORONET extracts knowledge subgraphs that can be employed to enrich vision-language understanding effectively and an accompanying commonsense enrichment module that can be easily integrated into video localization. **(3)** We provide empirical evidence of the effectiveness of our approach, demonstrating improvements up to 32.13% across various recall thresholds and up to 6.33% in mIoU. Extensive ablation studies thoroughly investigate the impact of commonsense on zero-shot NLVL performance.

Related Work

Natural Language Video Localization (NLVL)

Previous works on NLVL can be categorized into proposal-based (Liu et al. 2022b; Gao et al. 2021; Soldan et al. 2021; Xiao et al. 2021; Yang et al. 2021; Gao and Xu 2021; Yu et al. 2020; Wu et al. 2022) and proposal-free approaches (Rodriguez et al. 2020; Chen and Jiang 2020; Mun, Cho, and Han 2020; Zeng et al. 2021; Zhao et al. 2021; Zhang and Radke 2022; Rodriguez-Opazo et al. 2021). Proposal-based methods employ a generate-and-rank strategy, *i.e.* generating candidate video moments and subsequently ranking them based on their alignment with the given textual query. In contrast, proposal-free methods directly regress on the untrimmed video, estimating the boundaries of the target video segment based on the query.

The majority of NLVL works are fully supervised, with proposal-free methods primarily focusing on segment localization or regression accuracy (Zeng et al. 2020; Wang, Ma, and Jiang 2020; Rodriguez-Opazo et al. 2021), while proposal-based concentrating on improving the quality of the proposed video moment candidates (Xiao et al. 2021). To effectively capture cross-modal relationships, several works transform either the video or query modalities, or both, into graphs and perform graph matching (Soldan et al. 2021; Rodriguez-Opazo et al. 2021; Zeng et al. 2021; Chen and Jiang 2020). Some proposal-free works utilize convolutions to capture long-span dependencies within videos (Li, Guo, and Wang 2021) or as a form of cross-modal interaction (Zhang and Radke 2022; Chen and Jiang 2020). Moreover, there exist works that reframe NLVL into a generative task (Li et al. 2023) or traditional NLP tasks such as multiple-choice reading comprehension (Gao et al. 2021) and dependency parsing (Liu et al. 2021).

Weakly Supervised and Zero-shot NLVL Methods

Fully supervised methods achieve impressive performance but require laborious fine-grained video segment annotations corresponding to queries that are often prohibitively expensive for adapting to new domains. To address this challenge, weakly supervised methods have emerged, which operate with paired video-query data but without the need for precise video segment span annotations (Huang et al. 2021; Zhang et al. 2020; Ma et al. 2020; Lei, Berg, and Bansal 2021). Many weakly supervised approaches leverage contrastive learning to improve visual-textual alignment (Zhang et al. 2020, 2021; Ma et al. 2020). Recent work employs graph-based methodologies to capture contextual relationships between frames (Tan et al. 2021) and iterative approaches for fine-grained alignment between individual query tokens and video frames (Wang, Zhou, and Li 2021).

Despite requiring fewer annotations, the effort involved in acquiring queries is still substantial. Unsupervised iterative approaches (Liu et al. 2022a) and zero-shot NLVL (ZS-NLVL) (Nam et al. 2021) address this issue. ZS-NLVL aims to train an NLVL model using raw videos alone in a self-supervised setting, by generating video moments and corresponding pseudo-queries dynamically. Pseudo-query generation is critical in zero-shot localization methods, although limited work has been done in this direction. Nam et al. (2021) introduce pseudo-query generation for video localization, and subsequently, Jiang et al. (2022) for language grounding in images. Nam et al. (2021) consider a pseudo-query to be an unordered list of nouns and verbs, obtained from an off-the-shelf object detector and a fine-tuned language model (LM) that predicts the most probable verbs conditioned on the nouns. While the objects are grounded in

[1] Code available at https://github.com/PLAN-Lab/CORONET

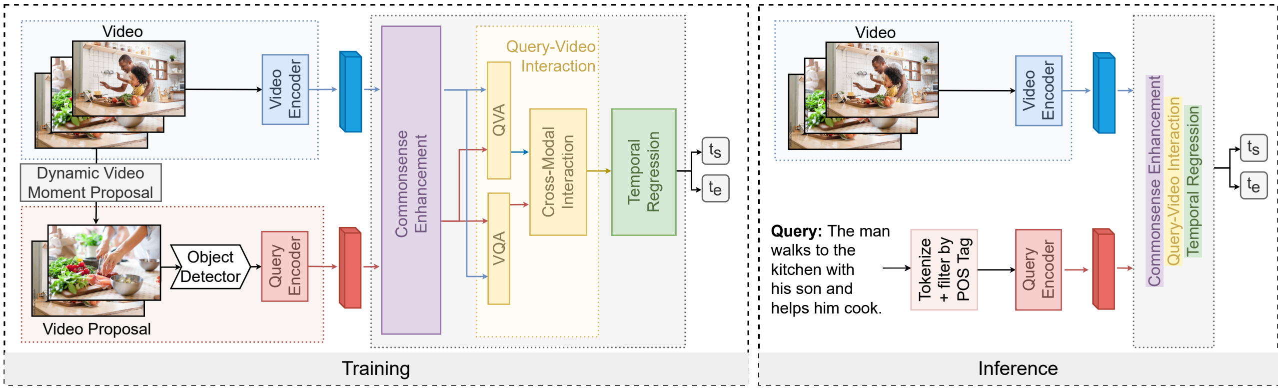

Figure 2: CORONET consists of a Video Encoder and a Query Encoder, the proposed Commonsense Enhancement, a Cross-Modal (video-query) Interaction, and a Temporal Regression module. During training, CORONET utilizes a Dynamic Video Moment Proposal module to extract a video moment span V_{span} and an off-the-shelf object detector to detect objects (nouns) in V_{span}. During inference, the given natural language query is converted to a simplified query using a part-of-speech tagger.

the video segment, the generation of verbs is not, potentially introducing irrelevant verbs and resulting in noisy pseudo-queries. Moreover, explicit verb-noun co-occurrences may encourage the localization model to learn spurious latent relationships and co-occurrence patterns between noun and verb data. Kim et al. (2023) propose a language-free approach that leverages the aligned visual-language space of a pretrained CLIP model. A limitation is primarily relying on visual and temporal cues for video grounding but not fully capturing higher-level contextual knowledge and implicit relationships often conveyed through natural language. This could hinder the model's ability to understand and localize events in videos when additional context is necessary. In contrast, CORONET enriches the extracted video and pseudo-query features with commonsensical information. By considering spatiotemporal, causal, and physical relations w.r.t. the visual information, our model reasons beyond video cues and grounds pseudo-query information in the video.

Commonsense in Video-Language Tasks

Recent video-language research has shifted towards enhancing reasoning capabilities rather than solely focusing on recognition. Datasets such as Video2Commonsense (Fang et al. 2020), Something Something (Goyal et al. 2017), Violin (Liu et al. 2020), SUTD-TrafficQA (Peng et al. 2021), and VLEP (Lei et al. 2020b) emphasize commonsense reasoning. Metrics have also been proposed to evaluate the commonsense reasoning abilities of video-language models (Shin et al. 2021; Park et al. 2022). Commonsense has also been incorporated into tasks such as video captioning (Yu et al. 2021), video question answering (Li, Niu, and Zhang 2022), and visual story generation (Maharana and Bansal 2021). Existing methods enhance query-based video retrieval using a co-occurrence graph of concepts mined from the target video moment (Wu et al. 2022; Cao et al. 2022). However, both are proposal-based fully supervised approaches that rely on fine-grained annotations and the quality of candidate video moments, let alone solely exploit the internal relations between the detected visual objects through a co-occurrence graph of entities as opposed to using external knowledge sources. In contrast, we utilize structured knowledge sources such as ConceptNet (Speer, Chin, and Havasi 2017) to encode commonsense information and leverage explicit relations spanning spatial, temporal, and physical aspects. This allows us to access information beyond what visual and textual cues can provide.

Commonsense for Zero-Shot NLVL

Problem Formulation

We denote an input video as V, and its grounding annotations as (Q, V_{span}), where Q is the query representation and $V_{\text{span}} = (t_s, t_e)$ is the corresponding video moment span annotation, with t_s and t_e representing the start and end timestamps, respectively. Learning to localize a video moment conditioned on a query entails maximizing the expected log-likelihood of the model parameterized by θ. In its typical setting, this can be formulated as follows:

$$\theta^* = \arg\max_{\theta} \mathbb{E}\left[\log p_\theta\left(V_{\text{span}}|V, Q\right)\right]. \quad (1)$$

In the zero-shot setting, the goal is to learn this task without parallel video-query annotations. Hence, the query and video moment annotations are derived from V, using a dynamic video moment proposal method followed by a pseudo-query generation mechanism. Formally, $V_{\text{span}}=f_{\text{span}}(V)$ and $Q=f_{pq}(V_{\text{span}})$, where f_{span} and f_{pq} are video moment proposal and pseudo-query generation mechanisms, respectively. Given that f_{span} and f_{pq} are responsible for generating the annotations, the performance of the localization model heavily depends on the quality of these modules. Existing methods face challenges in aligning Q to V_{span} due to noise introduced by ungrounded pseudo-query generation mechanisms. To address this, we simplify f_{pq} while augmenting cross-modal understanding by leveraging external information in the form of a commonsense graph $G_C(C, E)$ with n_c nodes, where $C = \{c_1, c_2, \ldots, c_{n_C}\}$ are the concept node vector representations and E is the set of weighted directed edges, respectively. Accordingly, learning can be formulated as

$$\theta^* = \arg\max_{\theta} \mathbb{E}\left[\log p_\theta\left(V_{\text{span}} | V, Q, G_C\right)\right]. \quad (2)$$

Figure 2 shows both training and inference flows.

Pseudo-supervised Setup

CORONET first processes a raw video with a video moment proposal f_{span} module that extracts important video segments capturing key events, and a pseudo-query generation f_{pq} that generates text query annotations corresponding to the extracted video segments.

Dynamic Video Moment Proposal (f_{span}). We adopt the dynamic video moment proposal approach proposed by Nam et al. (2021). Specifically, f_{span} primarily comprises a k-means clustering mechanism that groups semantically similar and temporally proximal video frame features together to extract atomic moments. To obtain frame features, we consider the columns of a frame-wise similarity matrix derived from the CNN features of individual frames. We enforce temporal proximity by concatenating the frame index to the features. Composite video moments are then formed by combining neighboring atomic moments, and a subset of all possible combinations is sampled uniformly at random. The resulting set of video moments corresponds to V_{span}.

Pseudo-query Generation (f_{pq}). The pseudo-query is constructed as a collection of objects present in the video. To generate the pseudo-query, we employ an off-the-shelf object detector, enabling the extraction of pertinent objects in V_{span}. We adopt a top-k strategy to sample the k most probable object predictions associated with the query Q.

Video Encoder. We uniformly sample T frames from V and extract their CNN (*e.g.*, I3D (Qian et al. 2023)) features. These features are contextually encoded using a video encoder ϕ_v to yield frame features $\phi_v(V) = \{v_1, v_2, \ldots, v_T\}$ where $v_i \in \mathbb{R}^d$, and d is the common video/query encoding dimension. We implement ϕ_v as a GRU-based encoder.

Query Encoder. Our pseudo-query Q, composed of up to k tokens, is encoded using a query encoder ϕ_q that generates query embeddings $\phi_q(Q) = \{q_1, q_2, \ldots, q_k\}$, for the top-$k$ detected objects extracted from the pseudo-query generation. Here, $q_i \in \mathbb{R}^d$ and d is the common video/query encoding dimension. We implement ϕ_q as a bi-directional GRU-based encoder preceded by a trainable embedding layer.

Commonsense Enhancement Module

To enrich the encoded video and query features with information grounded in commonsensical knowledge, we introduce a Commonsense Enhancement Module (CEM), pictorially described in Figure 3. This enhancement helps inject necessary information into video and query representations, which can not just help bridge the gap between the available visual and textual cues but also provide rich information to the downstream span localization module.

CEM includes a set $C = \{c_1, c_2, \ldots, c_{n_C}\}$ of n_C concept vectors, where $c_i \in \mathbb{R}^d$ and d is the concept feature dimension (same dimension as $\forall v_i \in V$ and $\forall q_i \in Q$). In general,

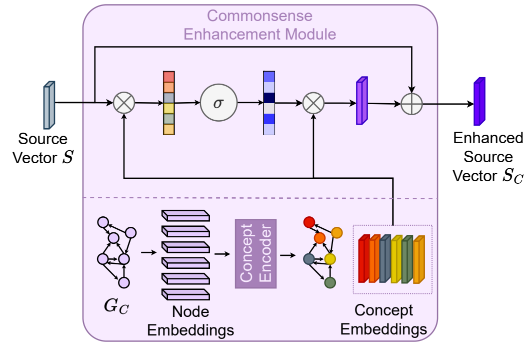

Figure 3: CORONET Commonsense Enhancement Module (CEM). CEM comprises a concept encoder and an enhancement mechanism that uses the previously encoded concept vectors to update a given input vector (video/query vectors). The concept encoder employs a Graph Convolution Network for encoding the nodes (concepts) of G_C.

given source feature vectors $S = \{s_1, s_2, \ldots, s_n\}$ with individual feature vectors $s_{i \in [1,n]} \in \mathbb{R}^d$, the enhanced feature vectors S_C are obtained using a commonsense enhancement mechanism ϕ_C. We implement this commonsense enhancement step ϕ_C as a cross-attention mechanism that enriches source input features, attending over S guided by the commonsense concept vectors C, *i.e.*,

$$S_C = S + \phi_C(S) = S + \sigma\left(\frac{SW_Q(CW_K)^T}{\sqrt{d}}\right) CW_V, \quad (3)$$

where σ is a softmax activation, W_Q, W_K, W_V are trainable matrices and d is the common dimension of the vectors S and C. In our setting, the source feature vectors S are either video V or pseudo-query Q features. We build separate enhancement mechanisms for V and Q, *i.e.*, the projection matrices W_Q, W_K, W_V are not shared between Q and V. We elaborate more on the rationale in the appendix. The enriched video and pseudo-query features are denoted as $V_C = \phi_{C_{\text{vid}}}(V)$ and $Q_C = \phi_{C_{\text{pq}}}(Q)$, respectively.

Concept Encoder. The concept vectors C mentioned above are feature representations that internally form the nodes of the commonsense graph, G_C. Accordingly, graph G_C is represented as a matrix, where $G_{C(i,j)}$ represents the total number of directed relational edges between $c_i, c_j \in C$ that start at c_i and end at c_j. To encode the commonsense information, we employ Graph Convolutional Networks (GCN) (Hammond, Vandergheynst, and Gribonval 2011). The concept encoder is composed of L graph convolution layers, each of which performs a convolution step

$$C^{(l+1)} = \sigma\left(AC^{(l)}W^{(l)}\right), \quad (4)$$

where $C^{(l)}$ are node (concept) features and $W^{(l)}$ trainable weight matrix of layer $l \in [1, L]$, σ is a nonlinear activation function, and A is the adjacency matrix obtained by normalizing graph G_C with the degree matrix D. Since G_C is a directed graph, normalization can be formulated as $A = D^{-1}G_C$.

Commonsense Information. We use ConceptNet (Speer, Chin, and Havasi 2017), a popular knowledge graph that provides information spanning various types of relationships such as physical, spatial, behavioral, *etc.* To ensure that the ConceptNet information utilized is relevant to themes found in the video data, we consider the set of objects available in pseudo-queries and include the top-k most frequently occurring objects to be the seed concept set C. We extract the ConceptNet subgraph that includes all edges incident between the concepts in C. We filter the edge types based on a pre-determined relation set R, which is compiled to involve relations that are relevant to the nature of the video localization task, *e.g.*, spatial (*AtLocation*, *etc.*) and temporal (*HasSubevent*, *etc.*) relations are useful for video understanding, while *RelatedTo* and *Synonym* are fairly generic relations that add little information to the localization task. Table 1 shows the relations included in G_C.

Cross-Modal Interaction Module. The commonsense enriched video and query features, V_C and Q_C, are fused with a multi-modal cross-attention mechanism. We employ a two-step fusion process. First, Query-guided Video Attention (QVA) is applied to attend over video V_C, and Video-guided Query Attention (VQA) attends over query Q_C guided by video V_C, resulting in updated features V'_C and Q'_C, respectively. Both QVA and VQA utilize Attention Dynamic Filters (Rodriguez et al. 2020) that adaptively modify video features, dynamically adjusting them in response to the query, and vice versa. Next, the attended features are fused using a cross-attention mechanism over V'_C guided by Q'_C, resulting in localized video features $V_{C_{\text{loc}}}$.

Temporal Regression Module. The final step involves a regression layer that approximates $\hat{V}_{\text{span}}$. We employ attention-guided temporal regression to estimate the span of the target video moment. To find important temporal segments relevant to the query, the fused features $V_{C_{\text{loc}}}$ are temporally attended and span boundaries are localized using a regressor implemented as a Multi-Layer Perceptron (MLP).

$$a_i = \sigma\left(W_1 V_{C_{\text{loc}_i}} + b_1\right) \tag{5}$$

$$v_{\text{ta}} = \sum_{i=1}^{T} a_i V_{C_{\text{loc}_i}} \tag{6}$$

$$[\hat{t}_s, \hat{t}_e] = W_2 v_{\text{ta}} + b_2. \tag{7}$$

Here, W_1 and b_1 are the weight matrix and bias vector of the temporal attention MLP, σ represents the sigmoid activation function, $V_{C_{\text{loc}_i}}$ stands for the i-th localized video feature, v_{ta} represents the temporally attended video feature vector, W_2 and b_2 denote the weight matrix and bias vector of the regression MLP, and $[\hat{t}_s, \hat{t}_e]$ correspond to the start and end timestamps of the predicted video span $\hat{V}_{\text{span}}$.

Training and Inference

The training objective is $\mathcal{L}_{loc} = \mathcal{L}_{treg} + \lambda\mathcal{L}_{ta}$, where λ is a balancing hyperparameter, $\mathcal{L}_{ta}$ is a temporal attention guided loss and $\mathcal{L}_{treg}$ is the regression loss. The temporal attention-guided loss is defined as

$$\mathcal{L}_{ta} = \frac{\sum_{i=1}^{T} g_i \log(a_i)}{\sum_{i=1}^{T} g_i}, \tag{8}$$

where a_i is the attention weight for video frame v_i and g_i is the attention mask for v_i, that is assigned to 1 if v_i is inside the target video segment, and 0 otherwise. This objective encourages the model to produce higher attention weights for video segments that are relevant to the query. On the other hand, $\mathcal{L}_{treg}$ dictates the video span boundary regression and is the sum of smooth ℓ_1 distances between start and end timestamps of the ground truth and predicted spans, *i.e.*,

$$\mathcal{L}_{treg} = \text{smooth}\ell_1(t_s, \hat{t}_s) + \text{smooth}\ell_1(t_e, \hat{t}_e). \tag{9}$$

Here, t_s and t_e represent the ground truth start and end timestamps and $\hat{t}_s$ and $\hat{t}_e$ the predicted start and end timestamps, respectively. The integration of a smoothing mechanism enhances training stability and improves the model's ability to handle outliers. Finally, during inference, we employ an off-the-shelf part-of-speech tagger to extract nouns from the text input query and feed them as query input to the trained CORONET video localizer.

Category	Relations
Spatial	AtLocation, LocatedNear
Temporal	HasSubevent, HasFirstSubevent, HasLastSubevent, HasPrerequisite
Functional	UsedFor
Causal	Causes
Motivation	MotivatedByGoal, ObstructedBy
Other	CreatedBy, MadeOf
Physical	HasA, HasProperty, Antonym, SimilarTo

Table 1: Relations in the Commonsense Enhancement Module (CEM) grouped by categories.

Experiments

Experimental Setup. Consistent with prior zero-shot NLVL research, we evaluate on Charades-STA (Gao et al. 2017) and ActivityNet-Captions (Heilbron et al. 2015; Krishna et al. 2017). Note that we only utilize the video components of the dataset during training. Query and video span annotations are only used for evaluation purposes. We compare CORONET against several zero-shot (Nam et al. 2021; Kim et al. 2023), weakly supervised (Mithun, Paul, and Roy-Chowdhury 2019; Chen et al. 2020; Lin et al. 2020; Duan et al. 2018; Gao et al. 2019) and fully supervised (Gao et al. 2017; Mun, Cho, and Han 2020) baselines. We evaluate performance with the mean temporal Intersection over Union (*mIoU*) for the predicted video moment spans and recall at specific threshold values (*R@k*), which is defined as the percentage of video span predictions with IoU value of at least k, where $k = \{0.3, 0.5, 0.7\}$, following prior works.

Approach	Supervision	Charades-STA R@0.3	Charades-STA R@0.5	Charades-STA R@0.7	Charades-STA mIoU	ActivityNet-Captions R@0.3	ActivityNet-Captions R@0.5	ActivityNet-Captions R@0.7	ActivityNet-Captions mIoU
CTRL (Gao et al. 2017)	Full	-	21.42	7.15	-	28.70	14.00	-	20.54
LGI (Mun, Cho, and Han 2020)	Full	72.96	59.46	35.48	51.38	58.52	41.51	23.07	41.13
TGA (Mithun et al. 2019)	Weak	29.68	17.04	6.93	-	-	-	-	-
WSTG (Chen et al. 2020)	Weak	39.80	27.30	12.90	27.30	44.30	23.60	-	32.20
SCN (Lin et al. 2020)	Weak	42.96	23.58	9.97	-	47.23	29.22	-	-
WS-DEC (Duan et al. 2018)	Weak	-	-	-	-	41.98	23.34	-	28.23
WSLLN (Gao et al. 2019)	Weak	-	-	-	-	42.80	22.70	-	32.20
PSVL† (Nam et al. 2021)	None	46.63	30.84	13.57	31.09	43.03	25.14	10.96	30.77
LFVL† (Kim et al. 2023)	None	<u>49.50</u>	<u>34.39</u>	<u>16.95</u>	**33.19**	43.34	25.17	**13.10**	**31.67**
CORONET	None	49.21	**34.60**	**17.93**	32.73	**46.05**	<u>28.19</u>	<u>12.84</u>	<u>31.11</u>
CORONET$_{250}$	None	**50.98**	33.18	16.48	<u>33.06</u>	<u>45.43</u>	**28.27**	12.81	30.88

Table 2: Localization accuracy compared to zero-shot, weakly, and fully supervised baselines. † indicates reproduction with official checkpoints and/or implementations. Best-performing method is highlighted in bold and second-best is underlined.

Experimental Results

Table 2 illustrates a comparative analysis of CORONET against baselines. We compare CORONET using two G_C versions with varying the number of concepts in the commonsense module, *i.e.* $n_C \in \{300, 250\}$. We represent these configurations as CORONET and CORONET$_{250}$, respectively. CORONET outperforms the fully supervised CTRL baseline and all of the weakly supervised baselines by significant margins. In addition, CORONET surpasses the PSVL zero-shot baseline across various configurations, with a particularly strong performance in the higher recall regime (R@0.7). For instance, for the Charades-STA dataset, CORONET consistently outperforms PSVL, yielding gains of up to 32.13% in higher recall scenarios. Similarly, on the same dataset, CORONET achieves recall enhancements of up to 5.78% over LFVL for R@0.7. In the context of the ActivityNet-Captions dataset, CORONET also outperforms PSVL across all metrics, showcasing performance improvements ranging from 7.02% to 17.15%. Notably, CORONET substantially outperforms LFVL on ActivityNet-Captions in terms of R@0.3 and R@0.5 (up to 12% for R@0.5), while maintaining comparable results in terms of R@0.7 and $mIoU$. Considering that ActivityNet-Captions represents a challenging benchmark encompassing diverse video themes, our findings highlight that leveraging commonsense information effectively helps integrate diverse visual-linguistic themes, outperforming methods that rely on pre-trained large-scale vision-language models.

Furthermore, since $mIoU$ for all models is close to 30%, an increase in $mIoU$ corresponds to a proportional increase in model predictions with recall above 0.3 (R@0.3). The performance of CORONET with lower *vs.* higher G_C sizes highlights a pattern of exclusivity between overall localization performance (*i.e.*, $mIoU$) and precision of accuracy (higher recall regimes, *e.g.*, R@0.7). This dichotomy between better recall at higher regimes and increased average localization highlights the trade-off between being able to generalize to a diverse set of videos and accurately localizing the moment in a specific video. Higher concept set sizes may provide wider levels of information to accommodate different types of videos better but may impede the model's capability to ground the exact video moment accurately.

Relations	R@0.3	R@0.5	R@0.7	mIoU
S	39.63	26.53	11.81	26.03
T	44.98	28.08	13.93	29.63
ST	49.84	30.35	15.16	32.38
F	49.21	34.60	17.93	32.73
F-ST	47.98	29.26	14.29	30.77
All	49.42	34.03	17.99	32.85

Table 3: CORONET with Spatial (S), Temporal (T), and Spatial and Temporal (ST) relations, the customized set of Filtered relations (F) mentioned in Table 1, F without the spatial/temporal relations (F-ST) and All relation types.

Ablation Studies

Comprehensive ablation studies, found in the appendix, provide further insights into the importance of commonsense in zero-shot NLVL. Specifically, we evaluate (1) the influence of various relation types, (2) the relative significance of commonsense in augmenting video or query features, (3) the potential usefulness of auxiliary commonsense information, and (4) the best approach for injecting commonsense.

Which Relation Types Are Most Important? We analyze the contribution and relative importance of relation types used in building G_C. Given that video localization requires spatiotemporal understanding, we hypothesize that relations falling in spatial/temporal categories are essential. Accordingly, we examine the performance of CORONET with 1) only spatial relations (**S**), 2) only temporal relations (**T**), and 3) both spatial and temporal relations (**ST**). We also evaluate CORONET with 4) a bigger subset of relations (as given in Table 1) to accommodate domain invariance (**F**), 5) the relation set mentioned in the previous configuration excluding spatial and temporal relation types (**F-ST**), and 6)

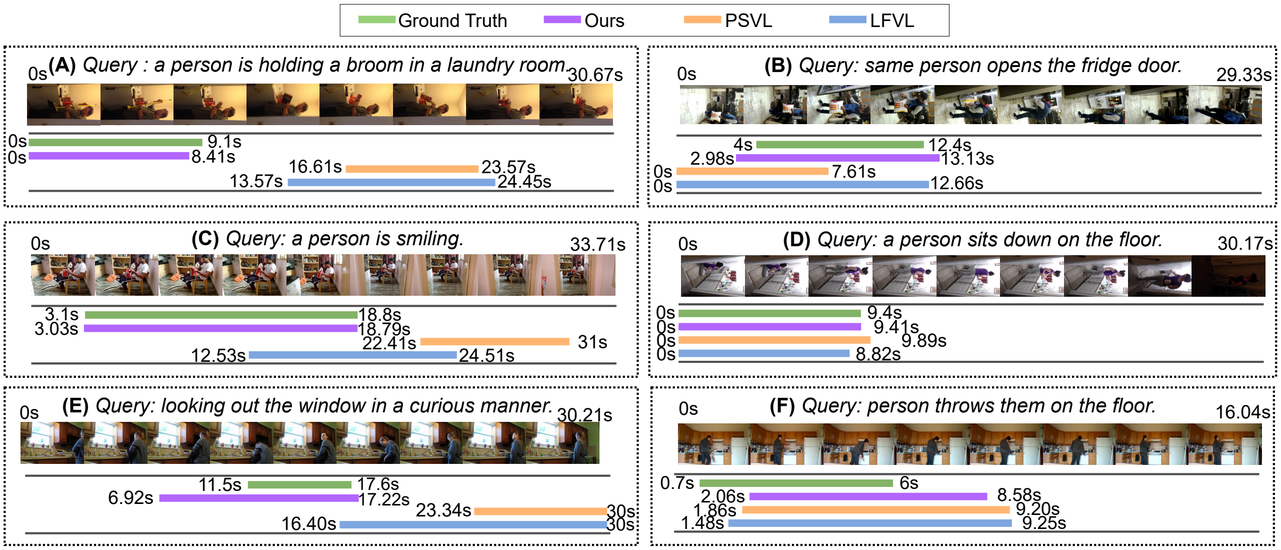

Figure 4: Qualitative inference results on examples from Charades-STA test data. Video span timestamps predicted by CORONET (purple lines), PSVL (orange lines), and LFVL (blue lines), juxtaposed with ground truth timestamps (green lines).

all relations in ConceptNet (**All**).

Table 3 enumerates the results across all CORONET configurations. The performance drops significantly across all metrics with **S**, where only spatial relations are considered. The temporal relation set **T** performs much better than **S**, highlighting the higher importance of temporal commonsense than spatial commonsense for precise localization. Despite the poor recall performance on **S**, spatial relations are valuable to the localization process, which is supported by a further improved performance with **ST**, which considers both spatial and temporal relations. The considerable drop in performance of **F-ST**, a filtered set that excludes spatial and temporal relation types, as compared to **F** and **ST**, further emphasizes the importance of spatial and temporal commonsense information for accurate localization. Finally, performance is considerably high with all relations included (**All**), but not significantly better than previous configurations that use a far smaller G_C and are hence more resource-efficient. Overall, **F** seems to provide a good balance between recall at various levels and mean localization accuracy.

Qualitative Results

We present qualitative results w.r.t. CORONET's localization performance along with the PSVL and LFVL baselines. In Figure 4, we showcase a few examples from the Charades-STA test split, accompanied by ground-truth video and query annotations and the localization results of the three models. Examples (A)-(C) show how CORONET accurately localizes the video moment, while PSVL and LFVL perform poorly. Upon inspection, PSVL and LFVL localize succeeding but semantically different events from the target segment in (A) and (C). On the other hand, in (B), PSVL and LFVL are seen to localize temporally preceding events along with the target event jointly. This example is challenging since the query includes "same person", which requires the ability to contextualize and distinguish preceding events from the target event. The inability of PSVL and LFVL baselines to localize the target segment alone (*i.e.*, isolating it from the preceding contextual segments), and conversely the accurate localization performance of CORONET, shows that CORONET can effectively contextualize and distinguish "same person" as a co-referenced entity by leveraging commonsense. Our results corroborate previous works that demonstrate the usefulness of commonsense in co-reference resolution (He et al. 2022; Ravi et al. 2023).

Example (D) highlights a case where all three models can localize accurately, while (E) and (F) show examples where none of the three models performs exceptionally well. However, it is important to note that CORONET localizes closest to the ground truth – it captures the ground truth event, but also jointly localizes the preceding event, which is semantically similar to the target event, *i.e.*, the person is looking outside the window. In contrast, PSVL and LFVL localize video segments that showcase events that are very distinct from the target query, *i.e.*, the person in the frame is looking away from the window (*e.g.*, LFVL localizes an event where the person is looking at the camera, whereas PSVL localizes the succeeding event where the person is looking down). Finally, all three models perform similarly in example (F). A deeper analysis reveals that each model localizes neighboring events in addition to the target event.

Conclusion

In this paper, we integrate commonsense in zero-shot natural language video localization to reduce noise in pseudo-queries and enhance cross-modal grounding between video and query modalities. Experimental results demonstrate the impact of commonsense relational information in enriching video and query representations, resulting in improved recall and localization performance within the zero-shot setting.

Acknowledgements

This material is based upon work supported in part by the U.S. DARPA KMASS Program #HR001121S0034, the Amazon – Virginia Tech Initiative for Efficient and Robust Machine Learning, and the Amazon Alexa Prize TaskBot Challenge 2. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements - either expressed or implied - of Amazon, DARPA, or the U.S. Government.

Note: The arXiv version of this paper includes supplementary material such as implementation details and additional ablation studies. https://arxiv.org/abs/2312.17429.

References

Cao, S.; Wang, B.; Zhang, W.; and Ma, L. 2022. Visual Consensus Modeling for Video-Text Retrieval. In *AAAI*, 167–175.

Chen, S.; and Jiang, Y.-G. 2020. Hierarchical Visual-Textual Graph for Temporal Activity Localization via Language. In *ECCV*, 601–618.

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:2001.09308*.

Ding, M.; Chen, Z.; Du, T.; Luo, P.; Tenenbaum, J.; and Gan, C. 2021. Dynamic Visual Reasoning by Learning Differentiable Physics Models from Video and Language. In *NeurIPS*, 887–899.

Duan, X.; Huang, W.; Gan, C.; Wang, J.; Zhu, W.; and Huang, J. 2018. Weakly Supervised Dense Event Captioning in Videos. In *NeurIPS*, 3062–3072.

Fang, Z.; Gokhale, T.; Banerjee, P.; Baral, C.; and Yang, Y. 2020. Video2Commonsense: Generating Commonsense Descriptions to Enrich Video Captioning. In *EMNLP*, 840–860.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. TALL: Temporal Activity Localization via Language Query. In *ICCV*.

Gao, J.; Sun, X.; Xu, M.; Zhou, X.; and Ghanem, B. 2021. Relation-aware Video Reading Comprehension for Temporal Language Grounding. In *EMNLP*, 3978–3988.

Gao, J.; and Xu, C. 2021. Fast Video Moment Retrieval. In *ICCV*, 1503–1512.

Gao, M.; Davis, L.; Socher, R.; and Xiong, C. 2019. WSLLN:Weakly Supervised Natural Language Localization Networks. In *EMNLP-IJCNLP*, 1481–1487.

Gao, Y.; Luo, Z.; Chen, S.; and Zhou, W. 2022. End-to-end Multi-task Learning Framework for Spatio-Temporal Grounding in Video Corpus. In *CIKM*, 3958–3962.

Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; Hoppe, F.; Thurau, C.; Bax, I.; and Memisevic, R. 2017. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *ICCV*, 5842–5850.

Hammond, D. K.; Vandergheynst, P.; and Gribonval, R. 2011. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30: 129–150.

He, K.; Mao, B.; Zhou, X.; Li, Y.; Gong, T.; Li, C.; and Wu, J. 2022. Knowledge Enhanced Coreference Resolution via Gated Attention. In *IEEE BIBM*, 2287–2293.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. 2015. ActivityNet: A Large-scale Video Benchmark for Human Activity Understanding. In *CVPR*.

Huang, J.; Liu, Y.; Gong, S.; and Jin, H. 2021. Cross-Sentence Temporal and Semantic Relations in Video Activity Localisation. In *ICCV)*, 7199–7208.

Jiang, H.; Lin, Y.; Han, D.; Song, S.; and Huang, G. 2022. Pseudo-Q: Generating Pseudo Language Queries for Visual Grounding. In *CVPR*, 15513–15523.

Kim, D.; Park, J.; Lee, J.; Park, S.; and Sohn, K. 2023. Language-free Training for Zero-shot Video Grounding. In *WACV*, 2539–2548.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Niebles, J. C. 2017. Dense-Captioning Events in Videos. In *ICCV*.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting Moments and Highlights in Videos via Natural Language Queries. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *NeurIPS*, 11846–11858.

Lei, J.; Yu, L.; Berg, T.; and Bansal, M. 2020a. TVQA+: Spatio-Temporal Grounding for Video Question Answering. In *ACL*, 8211–8225.

Lei, J.; Yu, L.; Berg, T.; and Bansal, M. 2020b. What is More Likely to Happen Next? Video-and-Language Future Event Prediction. In *EMNLP*, 8769–8784.

Li, J.; Niu, L.; and Zhang, L. 2022. From Representation to Reasoning: Towards Both Evidence and Commonsense Reasoning for Video Question-Answering. In *CVPR*, 21273–21282.

Li, K.; Guo, D.; and Wang, M. 2021. Proposal-free Video Grounding with Contextual Pyramid Network. In *AAAI*, 1902–1910.

Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023. MomentDiff: Generative Video Moment Retrieval from Random to Real. In *NeurIPS*.

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

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021. Context-Aware Biaffine Localizing Network for Temporal Sentence Grounding. In *CVPR*, 11235–11244.

Liu, D.; Qu, X.; Wang, Y.; Di, X.; Zou, K.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022a. Unsupervised Temporal Video Grounding with Deep Semantic Clustering. In *AAAI*, 1683–1691.

Liu, D.; Qu, X.; Zhou, P.; and Liu, Y. 2022b. Exploring Motion and Appearance Information for Temporal Sentence Grounding. In *AAAI*, 1674–1682.

Liu, J.; Chen, W.; Cheng, Y.; Gan, Z.; Yu, L.; Yang, Y.; and Liu, J. 2020. Violin: A Large-Scale Dataset for Video-and-Language Inference. In *CVPR*, 10900–10910.

Ma, M.; Yoon, S.; Kim, J.; Lee, Y.; Kang, S.; and Yoo, C. D. 2020. VLANet: Video-Language Alignment Network for Weakly-Supervised Video Moment Retrieval. In *ECCV*, 156–171.

Maharana, A.; and Bansal, M. 2021. Integrating Visuospatial, Linguistic, and Commonsense Structure into Story Visualization. In *EMNLP*, 6772–6786.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly Supervised Video Moment Retrieval from Text Queries. In *CVPR*, 11592–11601.

Mun, J.; Cho, M.; and Han, B. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In *CVPR*.

Nam, J.; Ahn, D.; Kang, D.; Ha, S. J.; and Choi, J. 2021. Zero-Shot Natural Language Video Localization. In *ICCV*, 1470–1479.

Park, J. S.; Shen, S.; Farhadi, A.; Darrell, T.; Choi, Y.; and Rohrbach, A. 2022. Exposing the Limits of Video-Text Models through Contrast Sets. In *NAACL-HTL*, 3574–3586.

Peng, H.; Huang, H.; Xu, L.; Li, T.; Liu, J.; Rahmani, H.; Ke, Q.; Guo, Z.; Wu, C.; Li, R.; Ye, M.; Wang, J.; Zhang, J.; Liu, Y.; He, T.; Zhang, F.; Liu, X.; and Lin, T. 2021. The Multi-Modal Video Reasoning and Analyzing Competition. In *ICCV Workshops (ICCVW)*, 806–813.

Qian, T.; Cui, R.; Chen, J.; Peng, P.; Guo, X.; and Jiang, Y.-G. 2023. Locate before Answering: Answer Guided Question Localization for Video Question Answering. *IEEE Transactions on Multimedia*.

Ravi, S.; Tanner, C.; Ng, R.; and Shwartz, V. 2023. What happens before and after: Multi-Event Commonsense in Event Coreference Resolution. In *EACL*, 1708–1724.

Rodriguez, C.; Marrese-Taylor, E.; Saleh, F. S.; LI, H.; and Gould, S. 2020. Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention. In *WACV*.

Rodriguez-Opazo, C.; Marrese-Taylor, E.; Fernando, B.; Li, H.; and Gould, S. 2021. DORi: Discovering Object Relationships for Moment Localization of a Natural Language Query in a Video. In *WACV*, 1079–1088.

Shin, M.; Kim, J.; Choi, S.; Heo, Y.; Kim, D.; Lee, M. S.; Zhang, B.; and Ryu, J. 2021. CogME: A Novel Evaluation Metric for Video Understanding Intelligence. *ArXiv:2107.09847*.

Soldan, M.; Xu, M.; Qu, S.; Tegner, J.; and Ghanem, B. 2021. VLG-Net: Video-Language Graph Matching Network for Video Grounding. In *ICCV*, 3224–3234.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*, 4444–4451.

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*, 2083–2092.

Wang, J.; Ma, L.; and Jiang, W. 2020. Temporally Grounding Language Queries in Videos by Contextual Boundary-Aware Prediction. In *AAAI*, 12168–12175.

Wang, Y.; Zhou, W.; and Li, H. 2021. Fine-grained Semantic Alignment Network for Weakly Supervised Temporal Language Grounding. In *Findings of EMNLP*, 89–99.

Wu, Z.; Gao, J.; Huang, S.; and Xu, C. 2022. Learning Commonsense-aware Moment-Text Alignment for Fast Video Temporal Grounding. *ArXiv:2204.01450*.

Xiao, S.; Chen, L.; Zhang, S.; Ji, W.; Shao, J.; Ye, L.; and Xiao, J. 2021. Boundary Proposal Network for Two-Stage Natural Language Video Localization. In *AAAI*.

Yang, X.; Feng, F.; Ji, W.; Wang, M.; and Chua, T.-S. 2021. Deconfounded Video Moment Retrieval with Causal Intervention. In *ACM SIGIR*, 1–10.

Yu, W.; Liang, J.; Ji, L.; Li, L.; Fang, Y.; Xiao, N.; and Duan, N. 2021. Hybrid Reasoning Network for Video-based Commonsense Captioning. In *ACM MM*, 5213–5221.

Yu, Z.; Song, Y.; Yu, J.; Wang, M.; and Huang, Q. 2020. Intra- and Inter-modal Multilinear Pooling with Multitask Learning for Video Grounding. *Neural Processing Letters*, 52(3): 1863–1879.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense Regression Network for Video Grounding. In *CVPR*.

Zeng, Y.; Cao, D.; Wei, X.; Liu, M.; Zhao, Z.; and Qin, Z. 2021. Multi-Modal Relational Graph for Cross-Modal Video Moment Retrieval. In *CVPR*, 2215–2224.

Zhang, H.; Sun, A.; Jing, W.; Nan, G.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021. Video Corpus Moment Retrieval with Contrastive Learning. In *ACM SIGIR*, 685–695.

Zhang, L.; and Radke, R. J. 2022. Natural Language Video Moment Localization Through Query-Controlled Temporal Convolution. In *WACV*, 2524–2532.

Zhang, Z.; Zhao, Z.; Lin, Z.; zhu, j.; and He, X. 2020. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. In *NeurIPS*, 18123–18134.

Zhao, Y.; Zhao, Z.; Zhang, Z.; and Lin, Z. 2021. Cascaded Prediction Network via Segment Tree for Temporal Video Grounding. In *CVPR*, 4197–4206.