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.

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 (**?**), video question answering (**??**), and video editing (**?**). Prior works have primarily explored supervised (**??????**) or weakly supervised (**???**) 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 **??** 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.

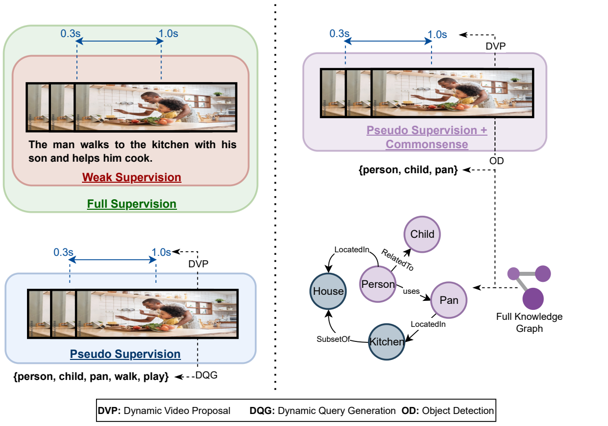


Figure 1: NLVL tasks with different supervision settings. Color-coded boxes enclose the annotations components expected at each supervision level. **Full supervision**: Temporal Video Annotations + Text Queries; **Weak Supervision**: Text Queries; **Pseudo-Supervision**: Only Raw Videos. Makes use of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

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 (**??**). 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 (**?**). On the other hand, language-free methods remove pseudo-queries entirely by utilizing vision-language models pretrained on large-scale image and text datasets (**?**). 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 (**??**).

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 (**??????**). 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 (**?**) 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 (**????????**) and proposal-free approaches (**???????**). 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 (**???**), while proposal-based concentrating on improving the quality of the proposed video moment candidates (**?**). To effectively capture cross-modal relationships, several works transform either the video or query modalities, or both, into graphs and perform graph matching (**????**). Some proposal-free works utilize convolutions to capture long-span dependencies within videos (**?**) or as a form of cross-modal interaction (**??**). Moreover, there exist works that reframe NLVL into a generative task (**?**) or traditional NLP tasks such as multiple-choice reading comprehension (**?**) and dependency parsing (**?**).

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 (**????**). Many weakly-supervised approaches leverage contrastive learning to improve visual-textual alignment (**???**). Recent work employs graph-based methodologies to capture contextual relationships between frames (**?**) and iterative approaches for fine-grained alignment between individual query tokens and video frames (**?**).

Despite requiring fewer annotations, the effort involved in acquiring queries is still substantial. Unsupervised iterative approaches (**?**) and zero-shot NLVL (ZS-NLVL) (**?**) 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. **?** introduce pseudo-query generation for video localization, and subsequently, **?** for language grounding in images. **?** 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 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. **?** 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 complex and nuanced events in videos that require additional context and reasoning beyond visual features. 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.

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

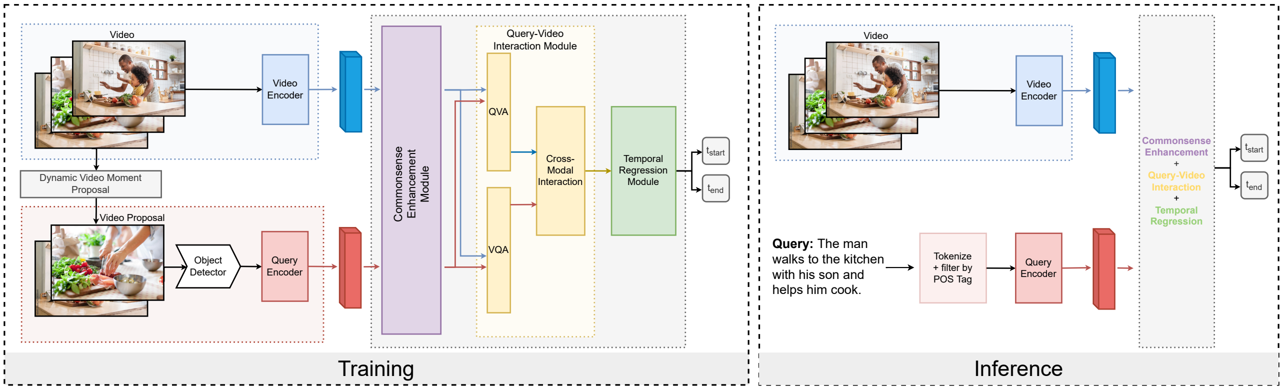

Video
Video Encoder
Dynamic Video Moment Proposal
Video Proposal
Object Detector
Query Encoder
Commonsense Enhancement Module
Query-Video Interaction Module
QVA
VQA
Cross-Modal Interaction
Temporal Regression Module
t_{start}
t_{end}
Training
Video
Video Encoder
Query: The man walks to the kitchen with his son and helps him cook.
Tokenize + filter by POS Tag
Query Encoder
Commonsense Enhancement + Query-Video Interaction + Temporal Regression
t_{start}
t_{end}
Inference

Figure 2: CORONET consists of a **Video Encoder** and a **Query Encoder**, the proposed **Commonsense Enhancement**, **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.

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 (**?**), Something Something (**?**), Violin (**?**), SUTD-TrafficQA (**?**), and VLEP (**?**) emphasize commonsense reasoning. Metrics have also been proposed to evaluate the commonsense reasoning abilities of video-language models (**??**). Commonsense has also been incorporated into tasks such as video captioning (**?**), video question answering (**?**), and visual story generation (**?**). Existing methods enhance query-based video retrieval using a co-occurrence graph of concepts mined from the target video moment (**??**). 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 (**?**) 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:

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

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

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

Figure **??** 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 **?**. 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

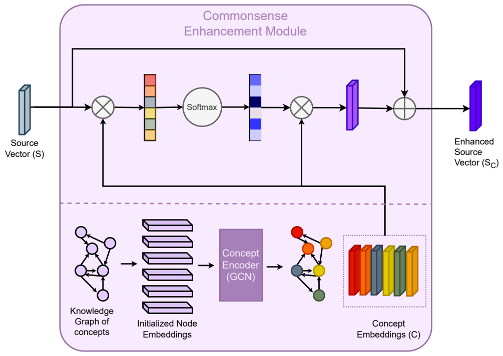


Figure 3: CORONET Commonsense Enhancement Module (CEM). CEM comprises a concept encoder and an enhancement mechanism that uses the previously encoded concept vectors to update a given input vector (video/query vectors). The concept encoder employs a Graph Convolution Network for encoding the nodes (concepts)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

is sampled uniformly at random. The resulting set of video moments corresponds to V_{span}.

Pseudo-query Generation ($f_{\mathbf{pq}}$). 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 (**?**)) 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 **??**. 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, 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 $c_i, cj \in C$ that start at c_i and end at c_j. To encode the commonsense information, we employ Graph Convolutional Networks (GCN) (**?**). 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 (**?**), 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 **??** 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

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.

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 (**?**) 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 based on the query features to obtain V_{ta}. Then, the span boundaries are localized using a regressor implemented as a Multi-Layer Perceptron (MLP).

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

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

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

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 encoded localized video features, V_{ta} represents the temporally attended video features, 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

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

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.

Experiments

Experimental Setup. Consistent with prior zero-shot NLVL research, we evaluate on Charades-STA (**?**) and ActivityNet-Captions (**??**). 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 (**??**), weakly supervised (**?????**) and fully supervised (**??**) 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*), that 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.

Experimental Results

Table **??** 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

		Charades-STA				ActivityNet-Captions			
Approach	**Supervision**	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**
CTRL (**?**)	Full	-	21.42	7.15	-	28.70	14.00	-	20.54
LGI (**?**)		72.96	59.46	35.48	51.38	58.52	41.51	23.07	41.13
TGA (**?**)	Weak	29.68	17.04	6.93	-	-	-	-	-
WSTG (**?**)		39.80	27.30	12.90	27.30	44.30	23.60	-	32.20
SCN (**?**)		42.96	23.58	9.97	-	47.23	29.22	-	-
WS-DEC (**?**)		-	-	-	-	41.98	23.34	-	28.23
WSLLN (**?**)		-	-	-	-	42.80	22.70	-	32.20
PSVL† (**?**)	None	46.63	30.84	13.57	31.09	43.03	25.14	10.96	30.77
LFVL† (**?**)		49.50	34.39	16.95	**33.19**	43.34	25.17	**13.10**	**31.67**
CORONET		49.21	**34.60**	**17.93**	32.73	**46.05**	28.19	12.84	31.11
CORONET$_{250}$		**50.98**	33.18	16.48	33.06	45.43	**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. The best-performing method is highlighted in bold and the second-best is underlined.

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 **??**, **F** without the spatial/temporal relations (**F-ST**) and **All** relation types.

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.

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 **??**) to accommodate domain invariance (**F**), 5) the relation set mentioned in the previous configuration excluding spatial and temporal relation types (**F-ST**), and 6) all relations in ConceptNet (**All**).

Table **??** 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 **??**, we showcase a few examples from the Charades-

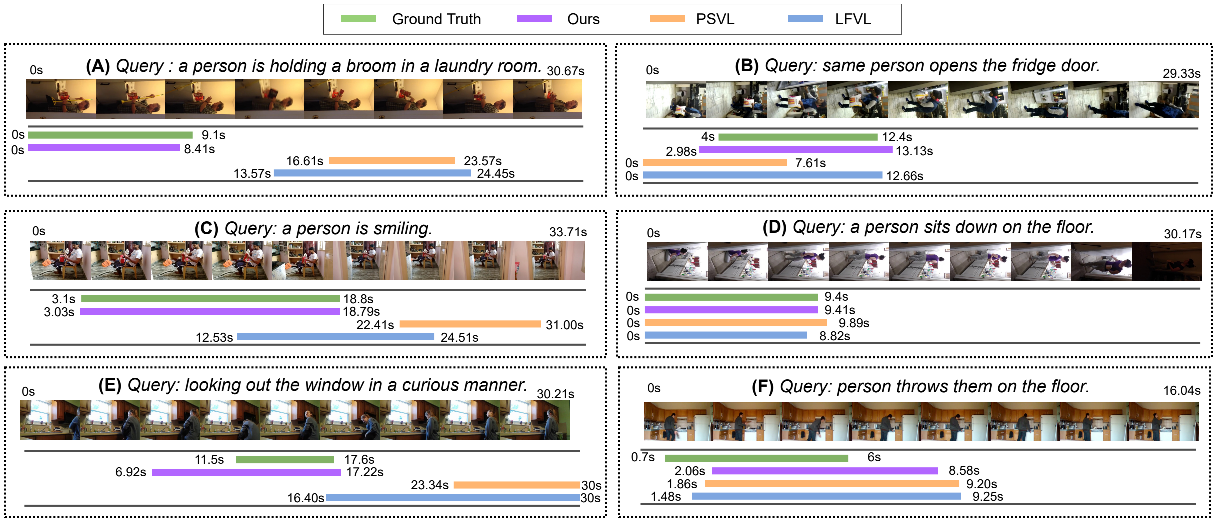


Figure 4: Qualitative inference results on examples from Charades-STA test data. Video span timestamps predicted by CORONET (**purple** lines), PSVL (**orange** lines), and LFVL (**blue** lines), juxtaposed with ground truth timestamps (**green** lines).

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 (**??**).

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, in an effort to reduce noise in pseudo-queries and enhance cross-modal grounding between video and query modalities. Our work highlights the benefits of commonsense knowledge in zero-shot natural language video localization. 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. Future research could explore more refined query representations, additional modalities such as audio or motion, and methods that capture richer query semantics.

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 and DARPA or the U.S. Government.

Implementation Details

We employ pre-trained I3D (**?**) and C3D (**?**) models to extract video frame features for Charades-STA and ActivityNet-Captions, respectively. We uniformly sample $T = 128$ features per video to ensure a fixed length. During the pseudo-query generation phase, we employ a Faster R-CNN object detector that is trained on objects enumerated in VisualGenome (**?**). We employ a top-k strategy to sample the most probable nouns found in the video segment. We choose a k value of 5 based on the experimental analysis by **?**. As for the CEM module, we rely on ConceptNet (**?**) for commonsense information and extract the English sub-graph for our experiments. Moreover, we prune the commonsense graph G_C by preserving edge types that convey relevant contextual information, as detailed in Table 1 in the main paper. We randomly initialize the weights for the GCN-based concept encoder. Experiments for the balancing hyperparameter λ, spanning a range of $\lambda \in \{0.75, 0.7, 0.3, 0.25\}$ show that performance is consistently high across all metrics for $\lambda = 0.7$, indicating the relative importance of temporal attention-guided loss over the overall localization regression loss.

Ablation Studies

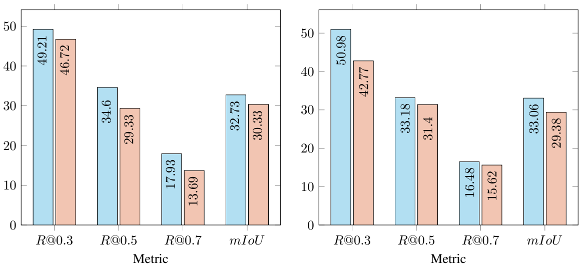
We further present ablation studies that focus on the Commonsense Enhancement Module (CEM) design and the overall efficacy of commonsense integration. Unless specified otherwise, we perform ablations on Charades-STA (**?**) and CORONET with 300 seed concepts.

How to Best Inject Commonsense?

We evaluate the effectiveness of the proposed enhancement mechanism by comparing it against an alternate configuration that omits the enhancement process and instead concatenates the encoded concept vectors with the text query vectors, treating the resultant feature set as the text query features. Figure **??** shows the relative performance of concatenation *vs.* enhancement across both CORONET configurations with 300 or 250 seed concepts. We observe that enhancement consistently outperforms concatenation, thereby reinforcing the effectiveness of the enhancement flow in injecting necessary commonsense information. Notably, the concatenation configuration for 300 seed concepts still outperforms the PSVL baseline at various recall thresholds *i.e.*, $k = \{0.3, 0.7\}$. This highlights the capacity of commonsense information to enhance localization performance even with a much simpler injection mechanism.

Which Modality to Enhance with Commonsense?

To analyze the importance of commonsense enhancement across modalities, we train CORONET with the following configurations: (1) Only query features Q are enhanced (**Q**), (2) Only video features V are enhanced (**V**). In addition, we employ two configurations for which both video and query are enhanced. CORONET's CEM makes use of the same concept vectors, but employs separate enhancement steps for video and text query, *i.e.*, we rely on two separate sub-modules $\phi_{C_{\text{vid}}}(V)$ and $\phi_{C_{\text{pq}}}(Q)$. This design choice stems from the hypothesis that the gap between video and query modalities is exacerbated when dealing with less sophisticated queries. Essentially, less sophisticated (or more general) queries may lack specificity or fail to capture the nuances of the desired information accurately. As a result, the gap between the information contained in the video and the intended query widens, making it more challenging to match the two modalities effectively. Having separate projection matrices for video and query allows differently enhancing V and Q with the same commonsense information (through the same concept vectors). To test this rationale, (3) we train CORONET with shared weights for V and Q, *i.e.*, $\phi_{C_{\text{vid}}}(V)$ and $\phi_{C_{\text{pq}}}(Q)$ are identical (**VQ**). Finally, (4) we represent the original setting of separate enhancement mechanisms for V and Q as **V+Q**.



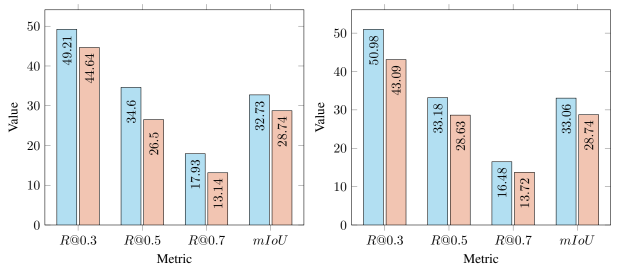
??

Figure 5: CORONET performance with enhancement *vs.* query-concepts concatenation for 300 (left) and 250 (right) seed concept sizes.

Method	R@0.3	R@0.5	R@0.7	mIoU
CORONET (**Q**)	39.72	22.16	8.10	26.06
CORONET (**V**)	<u>44.11</u>	31.64	15.20	<u>29.75</u>
CORONET (**VQ**)	40.21	<u>31.78</u>	**18.65**	28.45
CORONET (**V+Q**)	**49.21**	**34.60**	<u>17.93</u>	**32.73**

Table 4: CORONET performance with query enhancement only (**Q**), video enhancement only (**V**), shared video and query enhancement (**VQ**) and separate video and query enhancement (**V+Q**). The best and second-best scores are shown in **bold** and <u>underline</u>, respectively.

Table **??** presents results for the aforementioned configurations. We observe a significant drop in performance with **Q** across all metrics, which shows that the localization abilities are negatively impacted by omitting the video feature enhancement. To further support this observation, we see a consistent increase across all metrics for **V**, where only video features are enhanced and query feature enhancement is omitted. This highlights the positive impact of incorporating important commonsense information in the visual context for boosting model performance. Furthermore, we observe a consistent deterioration in model performance across all metrics in **VQ** except for R@0.7 when compared to **V+Q**. This could be attributed to the fact that a common enhancement flow for V and Q may potentially collapse diverging sources of information into one latent representation. Separating the enhancement for the two modalities allows disentangling the learned latent representations for video and



??

Figure 6: CORONET performance with pre-fusion enhancement *vs.* post-fusion enhancement for 300 (left) and 250 (right) seed concept sizes.

pseudo-query, thereby capturing different relationships, but with the same underlying commonsense knowledge. Finally, **V+Q** performing the best across all the aforementioned configurations validates our hypothesis of maintaining separate enhancement flows for video and text query features.

When to Perform Commonsense Enhancement?

CORONET separately enhances both video and query features prior to the cross-modal fusion step. However, an alternative option would be to perform commonsense enhancement on the unified video-query features after cross-modal fusion and interaction. Accordingly, we present results for pre-fusion as well as post-fusion enhancement. In Figure **??**, we observe that our approach of pre-fusion enhancement works significantly better than post-fusion enhancement across both 300 and 250 concept sizes. We believe the underlying reason for this observation is consistent with our previous prior findings, where employing separate enhancement modules for video and query features is best suited to inject necessary information and allowed CORONET to differently approach video and query enhancement.

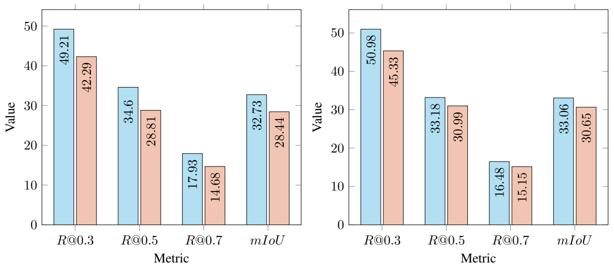
Does Retaining Relational Information Boost Localization?

CORONET employs a weighted directed graph as the concept graph G_C, where the edge weight between two nodes is the total number of relational edges from source to target nodes. In this ablation study, we analyze the impact of retaining multi-relational information in contrast to collapsing to a single weighted edge between two nodes. We replace the original G_C version with a multi-relational directed graph, where each edge belongs to one of the relation types in Table 1 in the main paper, and two nodes may be connected with multiple different edges. To this end, we employ Relational Graph Convolutional Networks (RGCNs) (**?**), where each graph convolution step is defined as

$$C^{(l+1)} = \sigma\left(\sum_{r\in R} A_r C^{(l)} W_r^{(l)}\right). \tag{10}$$

Model	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**
CORONET	**49.21**	**34.60**	**17.93**	**32.73**
CORONET-R	40.52	<u>27.92</u>	<u>13.85</u>	27.80
CORONET-R$_{post}$	<u>46.83</u>	25.57	12.45	<u>30.91</u>

Table 5: CORONET performance with multi-relational directed G_C with pre- (CORONET-R) and post-fusion (CORONET-R$_{post}$) enhancement. The best and second-best scores are shown in **bold** and <u>underline</u>, respectively.



??

Figure 7: We compare CORONET performance with 1-hop neighborhood graphs with their 0-hop neighborhood graph counterparts for 300 (left) and 250 (right) seed concepts.

Here, $W_r^{(l)}$ is the trainable weight matrix for layer $l \in [0, L]$ and A_r is the adjacency matrix for relation $r \in R$, where R denotes our relation set. We experiment with both pre- and post-fusion enhancement in this setup and respectively denote them by CORONET-R and CORONET-R$_{post}$.

Results in Table **??** show that contrary to one's intuition, having a higher-order contextual graph with more relational information does not help localization performance. A multi-relational adjacency matrix is much sparser than its weighted counterpart, where the adjacency matrix is aggregated along the relation dimension. Having a denser adjacency matrix could possibly enhance learning. Overall, this experiment showcases that the association between two given objects is more important in integrating commonsense, rather than the specific relation type.

Does Auxiliary Commonsense Information Boost Performance?

We analyze the impact of including auxiliary contextual information provided through G_C. We examine the performance of CORONET by replacing the proposed seed concept graph G_C with a bigger 1-hop neighborhood graph. Since including a 1-hop neighborhood leads to an exponential increase in the graph size, we limit G_C to include 1-hop neighborhood only with edge types that add valuable information to the localization setup. Specifically, we include edge types that may involve action information (*i.e.*, verbs) in relation to the objects observed in our video corpus (*e.g.*, $UsedFor$, $CapableOf$, $Causes$, *etc.*). Figure **??** shows the relative performance of this model variant in comparison to the original seed concept (0-hop) graph. Performance consistently worsens across all metrics for both 300 and 250 seed concept sizes, with more drastic drops for 300 concept sizes. We hypothesize that, despite the increased context via a larger graph, the additional information may prove to be noisy, thereby affecting localization accuracy as well as gen-

Encoder	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**	**time/epoch**
GRU	49.21	**34.60**	**17.93**	**32.73**	74.48s
Transformer	**53.57**	30.67	13.49	32.70	35.94s

Table 6: Performance with recurrent *vs.* Transformer-based encoders for video and query inputs. Time per epoch is measured in seconds. The best scores are presented in **bold**.

Model	**Enhancement**	**R@0.3**	**R@0.5**	**R@0.7**	**mIou**
LFVL (**?**)	None	49.50	**34.39**	**16.95**	33.19
+ CEM	Post	**54.39**	31.38	14.29	34.19
+ CEM$_{250}$		53.26	33.05	13.99	**34.30**
+ CEM	Pre	49.01	28.98	12.97	31.30
+ CEM$_{250}$		49.16	29.56	13.36	31.67

Table 7: Commonsense enhancement integrated to the LFVL (**?**) method. CEM and CEM$_{250}$ represent enhancement using our commonsense enhancement module with 300 and 250 seed concept graphs (G_C). Results for both post- and pre-fusion enhancement across CEM and CEM$_{250}$ are displayed. The best and second-best scores are shown in **bold** and underline, respectively.

eralization capabilities.

How to Best Encode Inputs?

We also investigate the impact of adopting a recurrent architecture (GRU/LSTM) *vs.* Transformers (**?**) for generating the video V and pseudo-query Q encodings. Table **??** quantitatively compares model performance under such encoding variants for CORONET. While Transformer-based methods are more than twice as fast as recurrent methods, they surprisingly impede model performance by large margins across most metrics.

Does Commonsense Help in Language-free Setups?

We also conduct an experiment to test the effectiveness of our CEM approach on a language-free NLVL (LFVL) setting (**?**). LFVL eliminates the need for query annotations by leveraging the cross-modal understanding of CLIP (**?**) to utilize visual features as textual information. We integrate our commonsense enhancement mechanism into the LFVL pipeline to analyze its impact in this setup. Table **??** compares model performances with two variants, CEM and CEM$_{250}$, which respectively contain 300 and 250 seed concepts. Furthermore, we examine the effectiveness of commonsense enhancement in a post- and pre-fusion setup. We find that there is a significant increase in the $mIoU$ and R@0.3 scores with both CEM and CEM$_{250}$ in post-fusion setup. This indicates that the integration of commonsense enhancement positively impacts the overall localization performance. Notably, the comparison between post- and pre-fusion enhancement reveals a striking difference in performance. These findings suggest that enhancing the fused video-query representation with commonsense information is more beneficial compared to enhancing a language-free query representation. The results imply that enhancing the fused representation allows for a more effective alignment between video and query, leading to improved localization performance.