Localizing Moments in Video with Temporal Language

Lisa Anne Hendricks[1]*, **Oliver Wang**[2], **Eli Shechtman**[2],
Josef Sivic[2,3]* , **Trevor Darrell**[1], **Bryan Russell**[2]
[1] UC Berkeley, [2] Adobe Research, [3] INRIA

Abstract

Localizing moments in a longer video via natural language queries is a new, challenging task at the intersection of language and video understanding. Though moment localization with natural language is similar to other language and vision tasks like natural language object retrieval in images, moment localization offers an interesting opportunity to model temporal dependencies and reasoning in text. We propose a new model that explicitly reasons about different temporal segments in a video, and shows that temporal context is important for localizing phrases which include temporal language. To benchmark whether our model, and other recent video localization models, can effectively reason about temporal language, we collect the novel TEMPOral reasoning in video and language (TEMPO) dataset. Our dataset consists of two parts: a dataset with real videos and template sentences (TEMPO - Template Language) which allows for controlled studies on temporal language, and a human language dataset which consists of temporal sentences annotated by humans (TEMPO - Human Language).

1 Introduction

Consider the video and natural language query in Figure 1 where we seek to localize the desired moment in the video specified by the query. Queries like "the girl bends down" require understanding objects and actions, but do not require reasoning about different video moments. In contrast, queries like "the little girl talks after bending down" require reasoning about the *temporal relationship* between different actions ("talk" and "bend down"). Localizing natural language queries in video is an important challenge, recently studied in Hendricks et al. (2017) and Gao et al. (2017) with applications in areas such as video search and retrieval. We argue that to properly localize queries with temporal language, models must understand and reason about intra-video context.

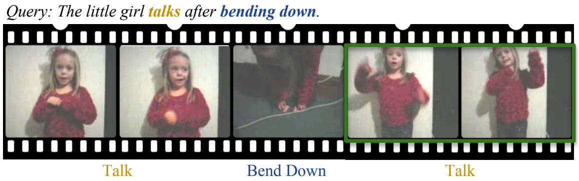

Figure 1: We consider localizing video moments which include *temporal language*. To properly localize "The little girl talks after bending down" localization models must understand how the action "talks" relates to the action "bend down."

Reasoning about intra-video context is difficult as we do not know a priori which moments should be involved in the contextual reasoning and different queries may require reasoning about different contextual moments. For example, in "the little girl talks after bending down", the relevant contextual moment "bending down" occurs just before the target moment "the little girl talks". This is in contrast to the query "the little girl talks *before* bending down" where the relevant contextual moment occurs just after. A limitation of current moment-localization models (Hendricks et al., 2017; Gao et al., 2017) is they consider query-independent video context when localizing moments. For example, when determining whether a proposed temporal region matches a natural language query, Gao et al. (2017) considers the proposed temporal region, as well as video regions just before and after the proposed region. Similarly, Hendricks et al. (2017) considers video context in the form of a global-context feature which represents the entire video. While both may implicitly include the appropriate contextual moment in their context feature, they do not explicitly determine the relevant context for the query.

To address this difficulty, we propose Moment

[*] Work done at Adobe during LAH's summer internship.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1380–1390
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

Localization with Latent Context (MLLC) which models video context as a *latent variable*. The latent variable enables the model to attend to different video contexts conditioned on the specific query/video pair, offering flexibility in the location and length of the contextual moment and overcoming the limitation of query-independent contextual reasoning. We validate the importance of latent context by showing that our model performs well both on simple queries without temporal words and more complex queries requiring temporal reasoning. Moreover, our formulation is generic and unifies approaches in Hendricks et al. (2017) and Gao et al. (2017), allowing us to ablate model component choices, as well as which kind of video context is best for localizing moments described with temporal language.

Though datasets used for moment localization in video (Hendricks et al., 2017; Regneri et al., 2013; Sigurdsson et al., 2016) include temporal language, as we will show, there is not enough temporal language to effectively train and evaluate models. We seek to extensively study this aspect, particularly with respect to temporal prepositions (Pratt-Hartmann, 2004). Thus, we collect the TEMPOral reasoning in video and language (TEMPO) dataset which builds off the recently collected DiDeMo dataset (Hendricks et al., 2017). The dataset consists of two parts: a dataset with real videos and sentences created with a template model (TEMPO - Template Language (TL)), and a dataset with real videos and newly collected user-provided temporal annotations (TEMPO - Human Language (HL)). Considering template sentences allows us to create a large dataset of sentences quickly for study of temporal language in a controlled setting. The human language data then allows us to see these trends transfer to more complex human-language queries. For data collection, we focus on the most common temporal referring words naturally occurring in language-and-video datasets.

Our contributions are twofold. (i) We are the first to study models for temporal language in video moment retrieval with natural language queries. To this end, we introduce TEMPO which includes examples of how humans use temporal language to refer to video moments. (ii) We propose MLLC for moment localization which treats video context as a latent variable and unifies prior approaches for moment localization. Our model outperforms prior work on TEMPO-TL and TEMPO-HL as well as the original DiDeMo dataset.

2 Related Work

Localizing Video Segments with Natural Language. Prior work has considered aligning natural language with video, e.g., instructional videos with transcribed text (Kiddon et al., 2015; Huang et al., 2017; Malmaud et al., 2014, 2015). Our work is most related to recent work in video moment retrieval with natural language (Gao et al., 2017; Hendricks et al., 2017). Both works take a natural language query and candidate video segment as input, and output a score for how well the natural language phrase aligns with the video segment. Gao et al. (2017) includes an additional loss to regress to start and end-points, whereas Hendricks et al. (2017) simplifies the problem by choosing from a discrete set of video segments. Importantly, to represent a proposed video segment, both models consider context features around a moment: Hendricks et al. (2017) uses global context by averaging features over an entire input video, and Gao et al. (2017) incorporates features adjacent to the proposed video segment. We argue that to do proper temporal reasoning, pre-determined, query independent context features may not cover all possible temporal relations. Thus, we propose to model the context as a *latent variable*, allowing our method to learn which context moments to consider as a function of the video and importantly, the query.

Both Gao et al. (2017) and Hendricks et al. (2017) collect data to test their models; Gao et al. (2017) considers the Charades (Sigurdsson et al., 2016) and TACoS (Regneri et al., 2013) datasets. While TACoS includes localized sentences, Charades only has sentences and activity detection localizations, so a semi-automatic method is used to align action detection annotations to visual descriptions in Charades. Hendricks et al. (2017) collected the Distinct Describable Moment (DiDeMo) dataset, which consists of Flickr (Thomee et al., 2016) videos with localized referring expressions. Both Charades and DiDeMo contain a large set of diverse videos (approximately 10,000 videos each). We chose to base TEMPO on DiDeMo because it contains more clip/sentence pairs (40,000 vs. 13,000), and is focused on general videos which we believe is an interesting and useful scenario, rather than being restricted to indoor activities.

Temporal Language. Prior work on temporal language processing has considered building explicit logical frameworks to process temporal prepositions like "during" or "until" (Pratt-Hartmann (2004), Konur (2008)). We do not derive a particular temporal logic, but rather learn to understand temporal language in a data driven fashion. Furthermore, we specifically consider how to understand temporal words commonly used when referring to video content. Other work has modeled dynamics for words which represent a change of state (e.g., "pick up") (Siskind (2001), Yu et al. (2015)) in limited environments. Though we limit the selection of temporal words in our study, the natural language in our data is open-world describing diverse events and how they relate to each other in video. Interpretation of temporal expressions in text ("The game happened on the 19^{th}") is a widely studied task (Angeli et al. (2012), Zhong et al. (2017)). Our work is distinctly different from this line of work as we specifically study temporal prepositions and how they refer to video.

Modeling Visual Relationships. A variety of papers have considered modeling spatial relationships in natural images (Dai et al., 2017; Hu et al., 2017; Peyre et al., 2017; Plummer et al., 2017). Our approach is analogous to this in the temporal domain; we hope to localize moments in videos. CLEVR, a synthetic visual question answering (VQA) dataset (Johnson et al., 2016), was created to allow researchers to systematically study the ability of models to perform complex reasoning. Our dataset is partially motivated by the success of CLEVR to enable researchers to study reasoning abilities of different models in a controlled setting. In contrast to CLEVR we consider a more diverse visual input in the form of real videos.

In the video domain, the TGIF-QA (Jang et al., 2017) and Mario-QA (Mun et al., 2016) datasets provide opportunities to study temporal reasoning for the task of VQA. The TGIF-QA dataset considers three types of temporal questions: before/after questions, repetition count, and determining a repeating action. Each question is accompanied by multiple choice answers. Videos we consider are much longer (25-30s as opposed to an average of 3.1s) which makes the use of temporal reasoning much more important. The MarioQA dataset is an additional VQA dataset designed to gauge temporal reasoning of VQA systems. Both TGIF-QA and MarioQA datasets include template-based natural language queries. In this paper, we consider synthetic queries similar to TGIF-QA and MarioQA, but also include human language queries. In addition, unlike the MarioQA dataset, that consists of synthetic data constructed from gameplay videos, our dataset consists of real visual inputs, and includes temporal grounding of natural language phrases. Finally, neither TGIF-QA nor MarioQA include temporal localization.

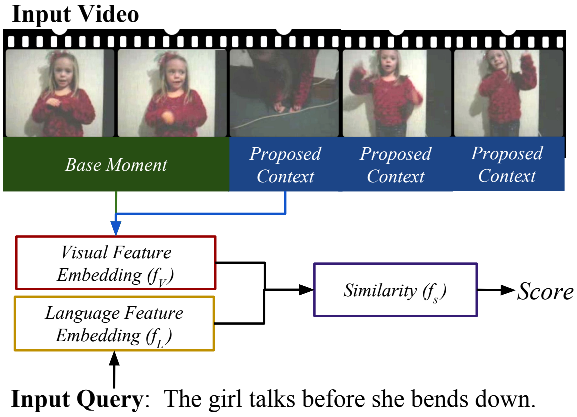
Input Video
Base Moment
Proposed Context
Proposed Context
Proposed Context
Visual Feature Embedding (f_V)
Language Feature Embedding (f_L)
Similarity (f_s)
Score
Input Query: The girl talks before she bends down.

Figure 2: Our model, Moment Localization with Latent Context (MLLC), takes a video and a text query as input and outputs the moment in the video corresponding to the query. MLLC considers many different *context* moments (blue) for a specific *base* moment (green).

3 Moment Localization with Latent Context

Given a video v and natural-language query q describing a moment in the video, our goal is to output the moment $\tau = \left(\tau^{(s)}, \tau^{(e)}\right)$ where $\tau^{(s)}$ and $\tau^{(e)}$ are temporal start and end points in the video, respectively. In the following, we formulate a generic, unified model which encompasses prior approaches (Hendricks et al., 2017; Gao et al., 2017). This allows us to explore and evaluate trade offs for different model components and extensions which then leads to higher performance. Unlike prior work, we consider a latent context variable which enables our model to better reason about temporal language.

Let the moment τ corresponding to the text query be the *base* moment and the set of other video moments T_τ be possible *context* moments for τ. We define a scoring function between the video moment and natural-language query by maximizing over all possible context moments

$\tau' \in \mathrm{T}_\tau$,

$$s_\phi(v, q, \tau) = \max_{\tau' \in \mathrm{T}_\tau} f_\mathcal{S}\left(f_\mathcal{V}\left(v, \tau, \tau'\right), f_\mathcal{L}(q)\right), \tag{1}$$

where $f_\mathcal{V}$ and $f_\mathcal{L}$ are functions computing features over the video and language query, $f_\mathcal{S}$ is a similarity function, and ϕ are model parameters. This formulation is generic and trivially encompasses the MCN and TALL formulations by letting the set of possible context moments T_τ be their respective single-context moment. Figure 2 shows the generic structure of our model.

With this formulation, we seek to answer the following questions: (i) Which combination of model components performs best for the moment-retrieval task? Though our primary goal is localizing moments with temporal language, we believe a good base moment retrieval model is important for localizing moments with temporal language. (ii) How best to incorporate context for moment retrieval with temporal language? We first detail the different terms and outline different model design choices, where design choices marked with ***bold-italic font*** is ablated in Section 5. Components which are used in our final proposed Moment Localization with Latent Context (MLLC) model and prior models are summarized in Table 3.

Video feature $f_\mathcal{V}$. The video feature $f_\mathcal{V} = (g(v, \tau), g(v, \tau'), f_\mathcal{T}(\tau, \tau'))$ is a concatenation of visual features for the base $g(v, \tau)$ and context $g(v, \tau')$ moments and endpoint features $f_\mathcal{T}(\tau, \tau')$. To compute visual features g for a temporal region τ, per-frame features are averaged over the temporal region. Note that if the context moment consists of more than one contiguous temporal region, then the visual features are computed over each contiguous temporal region and then concatenated (c.f., before/after context in TALL, explained below). There are many choices for visual features. TALL (Gao et al., 2017) compares average fc_7 features (extracted from (Simonyan and Zisserman, 2014)) to features extracted with C3D (Tran et al., 2015) and LSTM features (Donahue et al., 2015). Surprisingly, C3D features only outperform average fc_7 features by a small margin. We use the visual features used in the MCN model (Hendricks et al., 2017), which are similar to the fc_7 features from (Gao et al., 2017), but included motion features as well, computed from optical flow (extracted with (Wang et al., 2016)). We then pass the extracted visual features through a MLP. Note that we learn separate embedding functions for RGB and optical flow inputs and combine scores from different input modalities using a late-fusion approach (Hendricks et al., 2017).

Endpoint feature $f_\mathcal{T}$. Modeling temporal context requires understanding how different temporal segments relate in time. Hendricks et al. (2017) suggest including temporal endpoint features (***TEF***) $f_\mathcal{T} = \left(\tau^{(s)}, \tau^{(e)}\right)$ for the base moment which encode when the moment starts and ends to better localize sentences which include words like “first” and “last”. Note that TALL (Gao et al., 2017) does not incorporate TEFs. In order to understand temporal relationships, it is important that models also include features which indicate when a context moment occurs. In addition to providing TEFs for base moments, we also experiment with concatenating TEFs for context moments (***conTEF***) $f_\mathcal{T} = \left(\tau^{(s)}, \tau^{(e)}, \tau'^{(s)}, \tau'^{(e)}\right)$.

Language feature $f_\mathcal{L}$. Text queries are transformed into a fixed-length vector with an LSTM (Hochreiter and Schmidhuber, 1997). Before inputting words into the LSTM, they are embedded in the Glove (Pennington et al., 2014) embedding space. The final layer of the LSTM is projected into the shared video-language embedding space with a fully connected layer. Gao et al. (2017) considers LSTM language features and Skip-thought encoders. Our main goal is to study how context impacts moment localization with temporal language, so we use the LSTM features used on the original DiDeMo dataset.

Similarity $f_\mathcal{S}$. Given video $f_\mathcal{V}$ and language $f_\mathcal{L}$ features, we consider three ways to encode similarity between the features. Like Hendricks et al. (2017), we consider a ***distance-based*** similarity $f_\mathcal{S} = \left(|f_\mathcal{V} - f_\mathcal{L}|^2\right)$. Second, we consider a fused-feature similarity (***mult***) where the Hadamard product $f_\mathcal{V} \odot f_\mathcal{L}$ between the two features are passed to a MLP. We also explore unit normalizing features before the Hadamard product (***normalized mult***). Finally, we consider the similarity (***TALL similarity***) which consists of the concatenation $(f_\mathcal{V}, f_\mathcal{L}, f_\mathcal{V} \odot f_\mathcal{L}, f_\mathcal{V} + f_\mathcal{L})$ and then passed to a MLP.

Context moments T_τ. We consider three sets of context moments. First, we consider the entire video as the context moment (***global***) following Hendricks et al. (2017). Second, we consider using the moments just before and after the base moment (***before/after***). Finally, we consider using the set of all possible moments (***latent*** context) which offers greatest flexibility in contextual reasoning.

Training loss. We consider two training losses. The first loss is the MCN ***ranking loss*** which encourages positive moment/query pairs to have a smaller distance in a shared embedding space than negative moment/query pairs. To sample negative moment/sentence pairs, they consider negative moments *within* a specific video (called intra-video negative moments) and negative moments in different videos (called inter-video negative moments). This sampling strategy leads to a small improvement in performance (approximately one point on all metrics) when compared to just using intra-video negative moments. We also consider the alignment loss used in TALL (***TALL loss***) which is the sum of two log-logistic functions over positive and negative training query/moment pairs (intra-video negatives are used).

Supervising context moments. For the temporal sentences in our newly collected dataset (Section 4), we have access to the ground-truth context moment during training. Thus, we can contrast a ***weakly supervised*** setting in which we optimize over the unknown latent context moments during learning and inference to a ***strongly supervised setting***.

Implementation details. Candidate base and context moments coincide to the pre-segmented five-second segments used when annotating DiDeMo. Moments may consist of any contiguous set of five-second segments. For a 30-second video partitioned into six five-second segments, there are 21 possible moments. All models were implemented in Caffe (Jia et al., 2014) and optimized with SGD. Models were trained for $\sim$ 90 epochs with an initial learning rate of 0.05, which decreases every 30 epochs. Code is publicly released[*].

4 The TEMPO Dataset

We collect the TEMPOral reasoning in video and language (TEMPO) dataset based off the recently released DiDeMo dataset. Our dataset consists of two parts: TEMPO - Template Language (TL) and TEMPO - Human Language (HL). We create TEMPO - TL using language templates to augment the original sentences in DiDeMo with temporal words. The template allows us to generate a large number of sentences with known ground truth base and context moments. However, template language lacks the complexity of human language, so we then collect an additional fully user-constructed dataset, TEMPO - HL, consisting of sentences that contain specific temporal words.

Temporal Words in Current Datasets. We first analyze temporal words which occur in current natural language moment retrieval datasets. We consider temporal adjectives, adverbs, and prepositions found both by closely analyzing moment-localization datasets and consulting lists containing words which belong to different parts of speech. In particular, we rely on the preposition project (Litkowski and Hargraves, 2005)[†] to scrape relevant temporal words. Table 2 shows example temporal words and the number of times they occur in each dataset (TACoS (Regneri et al., 2013), Charades (Gao et al., 2017), DiDeMo (Hendricks et al., 2017)). Though all moment localization datasets use temporal words, they do not contain enough examples to reliably train and evaluate current models. Additionally, we observe that temporal words which are frequently used when describing video segments are different than those commonly used in text without video grounding. For example, in Pratt-Hartmann (2004), "during" is a common example, but we observe that "during" is infrequently used when describing video. Of temporal words, we focus on the four most common words, "before", "after", "then", and "while" when creating our dataset.

TEMPO - Template Language. To construct sentences in TEMPO-TL, we find adjacent moments in the DiDeMo dataset and fill in template sentences for "before", "after", and "then" temporal words. For "before", we use two templates: "X before Y" and "Before Y, X", where X and Y are sentences from the original DiDeMo dataset. Likewise for "after", we consider the templates "X after Y" and "After Y, X". For "then" we only consider one template, "X then Y."

TEMPO - Human Language. Though the template dataset is an interesting testbed for understanding temporal language, it is difficult to replicate the interesting complexities in human language. For example, when writing long sen-

[*] https://people.eecs.berkeley.edu/~lisa_anne/tempo.html

[†] http://www.clres.com/prepositions.html

	Endpoint Feature	Similarity	Context	Training Loss	Supervised Temp. Context
TALL (Gao et al., 2017)	None	TALL sim.	Before/After	TALL loss	None
MCN (Hendricks et al., 2017)	TEF	Distance-based	Global	Ranking	None
MLLC (ours)	**conTEF**	Normalized mult	**Latent**	Ranking	**Strongly sup.**

Table 1: Comparison of models. Bolded entries show our additions for localizing temporal language.

Dataset	Before	After	Then	While	Yet	During	Until
TACoS	50	62	731	82	23	0	4
Charades	281	27	1873	1165	0	3	1
DiDeMo	198	119	1021	266	16	21	22
TEMPO - TL	23,842	23842	11921	-	-	-	-
TEMPO - HL	6610	5495	5478	5425	-	-	-

Table 2: Word frequency of temporal words in natural language moment localization datasets.

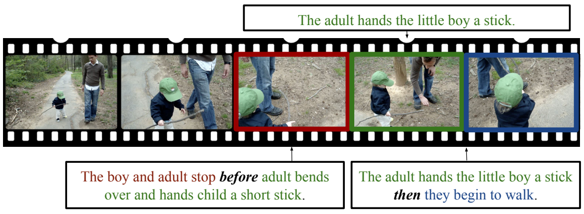
The adult hands the little boy a stick.
The boy and adult stop ***before*** adult bends over and hands child a short stick.
The adult hands the little boy a stick ***then*** they begin to walk.

Figure 3: Example sentences in TEMPO - HL. The top sentence corresponds to the reference moment (shown in green). The bottom sentences are newly collected sentences which use temporal language.

tences with temporal prepositions, humans frequently make use of language structure such as coreference to form more cohesive statements.

To collect annotations, we follow the protocol in Hendricks et al. (2017) and segment videos into 5-second temporal segments. After collecting descriptions, we ensure descriptions are localizable by asking other workers to localize each moment. To collect data for "before", "after", and "then", we ask annotators to describe a segment in *relation* to a "reference" moment from the DiDeMo dataset. For example, if the DiDeMo dataset includes a localized phrase like "the cat jumps", annotators write a sentence which refers to the segment "the cat jumps" using a specific temporal word. We provide both the phrase ("the cat jumps") and the reference moment to annotators, and the annotators provide a sentence describing a new moment which references the reference moment.

TEMPO-HL includes unique properties which are hard to replicate with template data. Figure 3 depicts the base moment provided to workers, as well as descriptions from TEMPO-HL. In Figure 3, the description "The adult hands the little boy the stick then they walk away" includes an example of visual coreference ("they"). We note that use of pronouns is much more prevalent in TEMPO-HL, with 28.1% of sentences in TEMPO-HL including pronouns ("he", "she", "it") in contrast to 10.3% of sentences in the original DiDeMo dataset. Additionally, annotators will refer to the base moment with different language than originally used in the base moment (e.g., "the girl waves at the camera" versus the base moment "the girl looks at the camera and waves") in order to make their sentences more fluent.

5 Experiments

Evaluation Method. We follow the evaluation protocol defined for the DiDeMo dataset (Hendricks et al., 2017) over all possible combinations of the five-second video segments. We report rank at one (R@1), rank at five (R@5), and mean intersection over union (mIOU) using their aggregator over three out of the four human annotators. We compare our models on TEMPO-TL, TEMPO-HL, and the DiDeMo dataset. When training our models, we combine the DiDeMo dataset with TEMPO-TL or TEMPO-HL. This enables our model to concurrently learn to localize the simpler DiDeMo sentences with more complex TEMPO sentences.

Baselines. We compare to the two recently proposed approaches for video moment localization: MCN (Hendricks et al., 2017) and TALL (Gao et al., 2017). We adapt the implementation of TALL (Gao et al., 2017) to the DiDeMo dataset in three ways. First, we do not include the temporal localization loss required to regress to specific start and end points as DiDeMo, and thus also TEMPO, is pre-segmented, so the model does not need to compute exact start and end points. Second, the original TALL model uses C3D features.

For a fair comparison we train both models with the same RGB and flow features extracted as was done for the original MCN model. Finally, the MCN model proposes temporal endpoint features (TEF) to indicate when a proposed moment occurs within a video. We train TALL with and without the TEF and show that TEF improves performance on the original DiDeMo dataset.

Ablations. To ablate our proposed latent context, we compare to other models which share the same MLLC base network. We consider the MLLC model with global context and before/after context. We also train a model with weakly supervised (WS) latent context and strongly supervised (SS) latent context. We also train models both with and without context TEF (conTEF).

The MLLC Base Model. We first ablate our MLLC base model (Table 3). We train our models on TEMPO-TL and DiDeMo and evaluate on the original DiDeMo dataset. All models are trained with global context. We find that the ranking loss is preferable on the DiDeMo dataset (compare lines 1 and 2) and that TALL-similarity performs better than the distance based similarity of the MCN model (compare lines 1 and 5). A simpler version of the TALL-similarity, in which the concatenated element wise multiplication, element wise sum, and concatenation is replaced by a single normalized elementwise multiplication, increases R@1 by almost one point and increases mIoU by over two points (compare lines 5-7). We call our best model the MLLC-Base model (line 7). Our MLLC-Base model performs better than previous models (MCN line 1 and TALL line 3).

	Model	Similarity	Training Loss	R@1	R@5	mIoU
1	MCN	Dist.-based	Ranking	26.63	73.38	41.14
2	MCN	Dist.-based	TALL	23.89	76.54	35.69
3	TALL	TALL-sim.	TALL	8.04	36.32	22.68
4	TALL w/TEF	TALL-sim.	TALL	23.56	72.74	35.58
5	MCN	TALL-sim	Ranking	27.52	**79.07**	41.48
6	MCN	Mult	Ranking	28.19	78.97	43.21
7	MLLC-Base	Norm. Mult	Ranking	**28.37**	78.64	**43.65**

Table 3: To select our base network, we consider different variants on the two previously proposed moment retrieval methods, TALL (Gao et al., 2017) and MCN (Hendricks et al., 2017). Results reported on val.

Results: TEMPO - TL. We first compare different moment localization models on TEMPO - TL (Table 4). In particular, our model performs well on "before" and "after" words. Additionally, our MLLC model with global context outperforms both the MCN model (Hendricks et al., 2017) and the TALL (Gao et al., 2017) model when considering all sentence types, verifying the strength of our base MLLC model.

Comparing MLLC with global context and MLLC with before/after context (compare row 4 and 5), we note that before/after context is important for localizing "before" and "after" moments. However, our model with strong supervision (row 9) outperforms the model trained with before and after context, suggesting that learning to reason about which context moment is correct (as opposed to being explicitly provided with the context before and after the moment) is beneficial. We note that strong supervision (SS) outperforms weak supervision (WS) (compare rows 7 and 9) and that the context TEF is important for best performance (compare rows 8 and 9).

We note that though the MLLC-global model outperforms our full model for "then" on TEMPO-TL, our full model performs better on then for the TEMPO-HL (Table 6). One possibility is that the "then" moments in TEMPO-TL do not require context to properly localize the moment. Because TEMPO-TL is constructed from DiDeMo sentences, constituent sentence parts are *referring*. For example, given an example sentence from TEMPO-TL (e.g., "The cross is seen for the first time *then* window is first seen in room"), the model does not need to reason about the ordering of "cross seen for the first time" and "window is seen for the first time" because both moments only happen once in the video. In contrast, when considering the sentence "The adult hands the little boy a stick *then* they begin to walk" (from Figure 3), "begin to walk" could refer to multiple video moments. Consequently, our model must reason about the temporal ordering of reference moments to properly localize the video moment.

On TEMPO - TL, sentences differ from original DiDeMo sentences solely because of the use of temporal words. Thus, we can do a controlled study of how well models understand temporal words. If a model has good temporal reasoning, then if it can localize a reference moment "the dog jumps" it should be easier for the model to localize the moment "the dog sits after the dog jumps". To test whether models are capable of this, we look at only sentences in TEMPO - TL where the model has correctly localized the cor-

		TEMPO - Template Language (TL)										
		DiDeMo		Before		After		Then		**Average**		
		R@1	mIoU	R@1	mIoU	R@1	mIoU	R@1	mIoU	R@1	R@5	mIoU
1	Frequency Prior	10.71	20.67	17.85	24.22	22.42	25.76	0.00	24.73	12.74	52.58	23.84
2	MCN	24.85	37.92	32.28	38.67	26.08	35.44	25.07	53.94	27.07	73.36	41.49
3	TALL	20.95	32.09	27.13	32.41	26.30	34.27	4.84	36.75	19.80	64.66	33.88
4	MLLC- Global	26.32	40.37	31.92	38.26	25.37	35.59	**27.53**	**57.08**	27.78	74.14	42.82
5	MLLC B/A	26.04	39.60	34.04	40.46	28.50	38.18	25.60	54.37	28.54	74.92	43.15
6	MLLC (WS)	26.57	40.99	30.56	37.64	24.76	35.10	26.95	56.49	26.95	74.18	42.55
7	MLLC (WS + conTEF)	25.87	40.37	32.01	39.51	24.31	33.94	24.98	55.22	26.79	74.04	42.27
8	MLLC (SS)	26.09	40.12	28.45	34.38	23.79	33.92	24.27	55.00	25.65	73.60	40.86
9	MLLC (SS + conTEF)	**27.46**	**41.20**	**35.31**	**41.81**	**29.38**	**38.90**	26.83	54.97	**29.74**	**76.76**	**44.22**

Table 4: Comparison of different model performance for different temporal words on TEMPO - TL on our test set. We report scores for the three temporal words in TEMPO - TL as well as on the original DiDeMo dataset. We find that our model performs best when considering all sentence types. B/A indicated before/after context, WS indicates weak context supervision, and SS indicates strong context supervision.

	Before		After		Then	
Context	R@1	mIoU	R@1	mIoU	R@1	mIoU
Global	-1.07	-2.72	-7.59	-6.75	43.30	31.57
Before/After	2.77	2.03	11.47	12.08	42.92	29.09
Latent	7.78	37.55	8.58	10.39	50.09	33.64

Table 5: Difference between performance on full dataset and set on which reference moments are localized properly for different methods on TEMPO-TL.

responding context moment in DiDeMo (Table 5). We report the *difference* in performance when considering only sentences in which temporal context was properly localized and all sentences. On our model, performance on all three temporal word types increases when the context moment can be properly localized. When considering global context, performance on "before" and "after" actually decreases, suggesting global context does not understand temporal reasoning well. Finally, even when the context is correctly localized, there is still ample room for improvement on all three sentence types motivating future work on temporal reasoning for moment retrieval.

Results: TEMPO - HL. Table 6 compares performance on TEMPO - HL. We compare our best-performing model from training on the TEMPO-TL (strongly supervised MLLC and conTEF) to prior work (MCN and TALL) and to MLLC with global and before/after context. Performance on TEMPO-HL is considerably lower than TEMPO-TL suggesting that TEMPO-HL is harder than TEMPO-TL.

On TEMPO - HL, we observe similar trends as on TEMPO-TL. When considering all sentence types, MLLC has the best performance across all metrics. In particular, our model has the strongest performance for all sentence types considering the mIoU metric. In addition to performing better on temporal words, our model also performs better on the original DiDeMo dataset. As was seen in TEMPO-TL, including before/after context performs better than our model trained with global context for both "before" and "after" words.

The final row of Table 6 shows an upper bound in which the ground truth context is used at test time instead of the latent context. We note that results improve for "before", "after", and "then", suggesting that learning to better localize context will improve results for these sentence types.

Localizing Context Fragments. TEMPO-HL sentences can be broken into two parts: a base-sentence fragment (which refers to the base moment), and a context-sentence fragment (which refers to the context moment). For example, for the sentence "The girl holds the ball before throwing it,", "the girl holds the ball" is the base fragment and "throwing it" is the context fragment. A majority of the "before" and "after" sentences in TEMPO-HL are of the form "X before (or after) Y", so we can determine a list of sentence fragments by splitting sentences based on the temporal word. Given "before" and "after" sentences, we determine the ground truth context fragment by considering which reference moment was given to annotators. We can then measure how well models localize context fragments. Table 7 compares two approaches to localizing context fragments: inputting just the context fragment into MLLC

TEMPO - Human Language (HL)

	DiDeMo		Before		After		Then		While		**Average**		
	R@1	mIoU	R@1	mIoU	R@1	mIoU	R@1	mIoU	R@1	mIoU	R@1	R@5	mIoU
Frequeny Prior	19.43	25.44	29.31	51.92	0.00	0.00	0.00	7.84	4.74	12.27	10.69	37.56	19.50
MCN	26.07	39.92	26.79	51.40	**14.93**	34.28	18.55	47.92	10.70	35.47	19.4	70.88	41.80
TALL + TEF	21.79	33.55	25.91	49.26	14.43	32.62	2.52	31.13	8.1	28.14	14.55	60.69	34.94
MLLC - Global	27.01	41.72	27.42	52.22	14.10	34.33	18.40	49.17	10.86	35.36	19.56	71.23	42.56
MLLC - B/A	26.47	40.39	31.95	55.89	**14.93**	34.78	17.36	47.52	**11.32**	35.52	20.40	70.97	42.82
MLLC (Ours)	**27.38**	**42.45**	**32.33**	**56.91**	14.43	**37.33**	**19.58**	**50.39**	10.39	**35.95**	**20.82**	**71.68**	**44.57**
MLLC (Ours) Context Sup. Test	27.39	42.25	52.58	80.37	36.48	75.79	36.05	70.51	10.39	35.87	32.58	79.86	60.96

Table 6: Comparison of different model performance on TEMPO - HL on the test set. “MLLC - Global” indicates our model with global context and “MLLC - B/A” indicated MLLC with before/after context.

	Before		After	
	R@1	mIoU	R@1	mIoU
Context Fragment	25.16	32.94	23.05	27.64
Full Sentence	**27.55**	**35.70**	**32.67**	**40.39**

Table 7: Comparison of different methods to localize context fragments (e.g., the text “she bends down” in the sentence “the girl talks after she bends down”). We compare localizing fragments with the MLLC model to localizing fragments with the latent context considered when localizing the whole query.

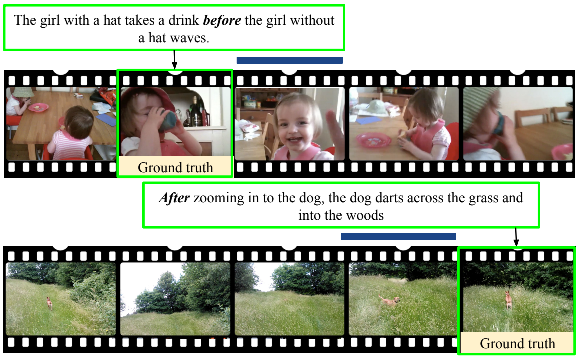
The girl with a hat takes a drink *before* the girl without a hat waves.
Ground truth
After zooming in to the dog, the dog darts across the grass and into the woods
Ground truth

Figure 4: Moment localization predictions on TEMPO - HL using our model. In addition to the localized query, we show the selected context segment (blue line) that our model considers when localizing the query.

and reporting the context used by MLLC when inputting the entire query into our model. We find that our model reliably selects the correct context fragments, most likely because it can properly exploit temporal understanding of how the context fragment relates to the base fragment.

Visualizing Context. In addition to a localized query, we can also visualize which context moment the temporal query refers to. Figure 4 shows predicted moments and their corresponding context moments. For the query “The girl with a hat takes a drink before the girl without a hat waves”, the little girl in the hat drinks twice, but our model correctly localizes the time she drinks *before* the other girl waves. Likewise, for the moment “After zooming in to the dog, the dog darts across the grass and into the woods”, the dog darts towards the woods twice (at the beginning of the video and at the end). Our model properly localizes the moment when the dog runs towards the forest the second time as well as the context fragment “zooming in on dog” when localizing the moment.

Discussion. We show promising results on both TEMPO-TL and TEMPO-HL, but there is potential improvement for building better frameworks for understanding temporal language. In Table 6, strongly supervising context at test time improves overall results, suggesting that models which can better localize context text will outperform our current model. Though TEMPO and DiDeMo have over 60,000 sentences combined, visual content is quite diverse. Integrating outside data sources (e.g., image retrieval and captioning) could possibly improve results on moment localization, both with and without temporal language queries. Additionally, in Table 5, even when the MLLC model can properly localize context, it does not always properly localize temporal sentences indicating that improved temporal reasoning can also improve our results. We believe our dataset, analysis, and method are an important step towards better moment retrieval models that effectively reason about temporal language.

Acknowledgements

We thank Anna Rohrbach for helpful feedback.

References

Gabor Angeli, Christopher D Manning, and Daniel Jurafsky. 2012. Parsing time: Learning to interpret time expressions. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 446–455. Association for Computational Linguistics.

Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. *arXiv preprint arXiv:1704.03114*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

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

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. *ICCV 2017*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Computer Vision and Pattern Recognition*.

De-An Huang, Joseph J. Lim, Li Fei-Fei, and Juan Carlos Niebles. 2017. Unsupervised visual-linguistic reference resolution in instructional videos. In *Computer Vision and Pattern Recognition*.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *arXiv preprint arXiv:1704.04497*.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*.

C. Kiddon, G. T. Ponnuraj, L. Zettlemoyer, and Y. Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes. In *Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP)*.

Savas Konur. 2008. An interval logic for natural language semantics. *Advances in Modal Logic*, 7:177–191.

Kenneth C Litkowski and Orin Hargraves. 2005. The preposition project. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171–179.

Jon Malmaud, Earl J. Wagner, Nancy Chang, and Kevin Murphy. 2014. Cooking with semantics. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 33–38.

Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's cookin'? Interpreting cooking videos using text, speech and vision. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 143–152.

Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. 2016. Marioqa: Answering questions by watching gameplay videos. *arXiv preprint arXiv:1612.01669*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. 2017. Weakly-supervised learning of visual relations. *arXiv preprint arXiv:1707.09472*.

Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1928–1937.

Ian Pratt-Hartmann. 2004. Temporal prepositions and their logic. In *Temporal Representation and Reasoning, 2004. TIME 2004. Proceedings. 11th International Symposium on*, pages 7–8. IEEE.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Jeffrey Mark Siskind. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of artificial intelligence research*, 15:31–90.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer.

Haonan Yu, N Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. 2015. A compositional framework for grounding language inference, generation, and acquisition in video. *Journal of Artificial Intelligence Research*, 52:601–713.

Xiaoshi Zhong, Aixin Sun, and Erik Cambria. 2017. Time expression analysis and recognition using syntactic token types and general heuristic rules. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 420–429.