Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding

Zhenzhi Wang   Limin Wang[*]   **Tao Wu   Tianhao Li   Gangshan Wu**

State Key Laboratory for Novel Software Technology, Nanjing University, China
{zhenzhiwang,tianhaolee}@outlook.com, {lmwang,gswu}@nju.edu.cn, wt@smail.nju.edu.cn

Abstract

Temporal grounding aims to localize a video moment which is semantically aligned with a given natural language query. Existing methods typically apply a detection or regression pipeline on the fused representation with the research focus on designing complicated prediction heads or fusion strategies. Instead, from a perspective on temporal grounding as a metric-learning problem, we present a Mutual Matching Network (MMN), to directly model the similarity between language queries and video moments in a joint embedding space. This new metric-learning framework enables fully exploiting negative samples from two new aspects: constructing negative cross-modal pairs in a mutual matching scheme and mining negative pairs across different videos. These new negative samples could enhance the joint representation learning of two modalities via cross-modal mutual matching to maximize their mutual information. Experiments show that our MMN achieves highly competitive performance compared with the state-of-the-art methods on four video grounding benchmarks. Based on MMN, we present a winner solution for the HC-STVG challenge of the 3rd PIC workshop. This suggests that metric learning is still a promising method for temporal grounding via capturing the essential cross-modal correlation in a joint embedding space. Code is available at https://github.com/MCG-NJU/MMN.

1 Introduction

Video analysis is a fundamental problem in computer vision and has drawn increasing attention in recent years because of the potential applications in surveillance, robotics, and Internet videos. While considerable progress has been made in video classification (Wang et al. 2016, 2018; Feichtenhofer et al. 2019; Wang et al. 2021) and action localization (Zhao et al. 2017; Lin et al. 2019; Li et al. 2020), such tasks are still limited to recognizing a pre-defined list of activities, such as playing baseball or peeling potatoes. As videos often contain complex activities that may cause a combinatorial explosion if described by a list of actions and objects, the task of grounding language queries in videos (Hendricks et al. 2017; Gao et al. 2017a; Tang et al. 2020), as a generalization of the action localization task to overcome the constraints, has recently gained plenty of interest (He et al.
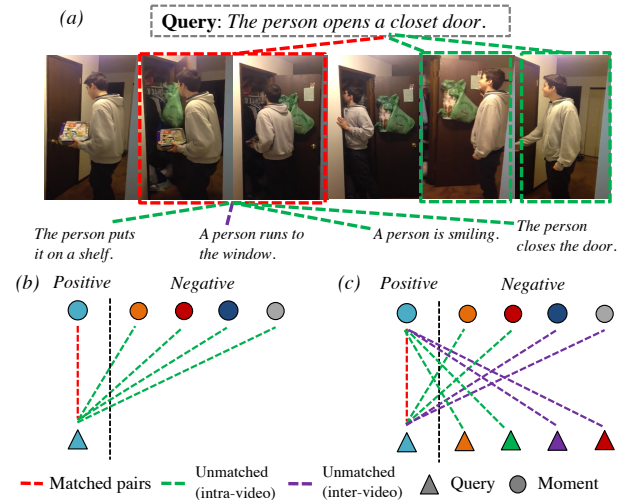
---
[*]Corresponding author.

Figure 1: (a) In addition to matching moments given a query, we propose a new auxiliary task named *mutual matching* to differentiate the matched query (red dash line) among unmatched queries intra- (green) or inter-video (purple) for the GT moment (red box). (b) Most previous methods only consider intra-video negative moments (green). (c) Our *mutual matching* uses negative pairs from both modalities intra- (green) or inter-video (purple). **Best view with colors**.

2019; Zhang et al. 2019; Zeng et al. 2020; Yuan et al. 2019; Zhang et al. 2020d; Tang et al. 2020) in both computer vision and language community. Formally, given a verbal description, the goal of temporal grounding is to determine the temporal moment (i.e., the start and end time) that semantically corresponds to the query best in a given video.

Although temporal grounding opens up great opportunities for detailed video perception by using the new language modality to capture the complex relations between sentences and videos, most of the previous approaches (Zhang et al. 2019; Yuan et al. 2019; Zhang et al. 2020d; Zeng et al. 2020; Tang et al. 2020) still tackle this problem in a *detection/regression* way with *early-fusion* designs, e.g., using the fused multi-modal features to predict the offset of action moments from anchors; or directly regressing the desired region on

a globally aggregated multi-modal feature. These indirect grounding methods typically ignore the essential relation between all cross-modal pairs (i.e., pairs of moment and language queries), and they only simply utilize the IoU-based scores between the given sentence and moments from the *same video* as supervision (Fig. 1(b)). However, we argue that the negative relations between the given moment and other *unmatched descriptions* are also important for learning a joint cross-modal embedding. Intuitively, the auxiliary task, given a video moment with visually informative actions, training the model to contrast the matched query and unmatched descriptions, (Fig. 1(a)) is beneficial for the temporal grounding task. We propose to use this auxiliary task for the first time in temporal grounding, which shows advantages in less extra computational cost compared with previous auxiliary tasks. In order to model such relations from mutual directions (Fig. 1(c)), we revisit the temporal grounding task from a *metric-learning perspective*, which allows us to directly model the essential similarity measurement in joint cross-modal embedding space, rather than designing sophisticated detection/regression strategies based on the fused representation as in many previous methods.

In our metric-learning view, sentences and moments play an equally important role in temporal grounding and both matching directions are adopted, i.e., select the right instance in a modality given the groundtruth of another modality. Thus, supervisions are constructed in a symmetric form for two modalities by our approach (Fig. 1(c)), creating more supervisions than previous methods. Our framework shows several advantages: Firstly, the metric-learning perspective enables us to mine negative samples from a *mutual matching scheme*. In this sense, an *unmatched relation* is also informative by implying that this moment and sentence should be pushed away in the joint space. To achieve this, we adopt a cross-modal mutual matching objective to contrast pos./neg. moment-sentence pairs inspired by a cross-modal pretraining method (Li and Wang 2020). Secondly, we exploit much more negative samples to improve our representation learning by adopting *inter-video* negative samples. In contrast, most of the previous methods only utilize *intra-video* moment-sentence pairs as supervision due to the early-fusion strategy. Finally, we model the video and language feature in a Siamese-alike network architecture and use a simple dot-product for cross-modal similarity computation. Thus, our framework has less computation cost by sharing moment features among sentences inside each video.

While using the negative pairs can lead to a better feature, the binary supervision signal itself (i.e., matched or unmatched) is still weak to precisely rank the moments inside a video. In addition, although the negative pairs used in our mutual matching scheme provide much more supervision signals, most of them are easy negatives (especially the inter-video ones). So we adopt another embedding space for learning the cross-modal similarity for precisely ranking moments based on the IoU supervision, which enables us to estimate finer relations (i.e., a scalar instead of a binary signal) and accurately differentiate the correct pair among hard negatives (e.g., negative moments with high IoU). These two complementary objectives share the same feature encoders

as backbone and have their own customized heads.

Our main contributions are three-fold: (1) We revisit the metric-learning perspective on the temporal grounding task in a *late-fusion* fashion and leverage **inter-video** negative visual-language pairs. Our framework shows advantages on both performance and training cost. (2) Our metric-learning view enables a new auxiliary task for temporal grounding, named **cross-modal mutual matching** to significantly add more supervision signals. Instead of heavy additional networks for previous auxiliary tasks, ours has a more direct idea and is also more effective. (3) Comprehensive quantitative and qualitative analyses are conducted on four benchmarks from both temporal and spatio-temporal video grounding tasks to show our method's generalization ability.

2 Related Works

2.1 Temporal Grounding

Temporal Grounding Methods. Previous methods can be mainly categorized into four groups: **(1) Regression based methods** directly predict the boundaries (i.e., start and end) of the target moment from the fused multi-modal features relying on either a clip-wise boundary classification in local regions (Ghosh et al. 2019; Opazo et al. 2020; Zhang et al. 2020b; Chen et al. 2020a) or a direct boundary regression on the aggregated global feature (Yuan, Mei, and Zhu 2019; Wang, Ma, and Jiang 2020; Mun, Cho, and Han 2020). Some methods also introduce some heuristics, e.g., aggregating clip-wise actionness scores modeled by compositional reasoning (Liu et al. 2018) or taking the expectation of probability for start/end (Ghosh et al. 2019). **(2) Detection based methods** often first generate candidate moments and then evaluate them on the fused multi-modal features (Gao et al. 2017a; Ge et al. 2019). Their evaluation on moments utilize various designs, e.g., LSTM (Chen et al. 2018; Xu et al. 2019), dynamic filtering (Zhang et al. 2019; Opazo et al. 2020) or modulation (Yuan et al. 2019), graph convolution (Zhang et al. 2019), anchor-free detectors (Lu et al. 2019; Zeng et al. 2020), and 2D moment map (Zhang et al. 2020d). **(3) Reinforcement Learning based methods** (He et al. 2019; Wang, Huang, and Wang 2019; Wu et al. 2020) localize the target moment iteratively by defining the *state* and *action* on the video and treating this task as a sequential decision-making process. **(4)** The only previous **metric learning based method** (Hendricks et al. 2017) uses a triplet loss with ℓ_2 distance as the similarity measurement to match the given sentence to the correct video moment. However, it lacks both the important supervision of negative sentence samples, i.e., our novel *mutual matching scheme* and the effective relation modeling of moments, thus achieves much worse results than ours.

Multi-modal Fusion. Many methods from the group 1) to 3) mainly adopt a *early-fusion* pipeline for cross-modal modeling, e.g., by concatenation (Chen et al. 2018; Lin et al. 2020; Wang, Ma, and Jiang 2020), dynamic convolution (Opazo et al. 2020; Zhang et al. 2019; Yuan et al. 2019), cross-attention (Lu et al. 2019; Zhang et al. 2020a), or hadamard product (Zeng et al. 2020; Mun, Cho, and Han 2020; Zhang et al. 2020d). On the contrary, we use a simple inner-product

in the joint visual-language space to measure the cross-modal similarity in a *late-fusion* manner. It not only enables our mutual matching scheme, but also shows advantages in computational cost during training by sharing the video feature among sentences in the same video.

Auxiliary Tasks in Temporal Grounding. Query reconstruction as an auxiliary task for temporal grounding was explored by (Lin et al. 2020; Xu et al. 2019) who added a video caption loss following image-based grounding methods (Ramanishka et al. 2017). Yet it introduces extra overhead of parameters and computational cost, e.g., LSTM (Xu et al. 2019) or Transformer (Lin et al. 2020). Our proposed auxiliary task of mutual matching avoid the usage of the heavy additional network and is also more direct and effective.

2.2 Spatio-Temporal Video Grounding

As the recent progress achieved in spatio-temporal action localization (Li et al. 2020), spatio-temporal video grounding (Tang et al. 2020) is also proposed as an extension of temporal grounding. By adapting our MMN on the linked human bounding boxes (Kalogeiton et al. 2017) for temporal trimming, significant performance gain is achieved over the previous transformer-based method (Tang et al. 2020).

2.3 Metric Learning

The family of metric learning loss has been explored to learn powerful representations with the supervised setting (Hadsell, Chopra, and LeCun 2006; Khosla et al. 2020), where the positive sample is chosen from the same class and the negative one from other classes; or self-supervised setting (Wu et al. 2018; Chen et al. 2020c; He et al. 2020), which select positive samples with data augmentation or co-occurrence. Different from both setting, we choose pos./neg. samples according to the groundtruth in a *supervised* way, yet we have *no pre-defined categories*. Furthermore, most aforementioned methods use single-modal samples, e.g., images, while we utilize cross-modal moment-sentence pairs. There are some video-language pre-training methods (Li and Wang 2020; Miech et al. 2020) similar to our setting, yet they aim to learn *video-level* representations in an *unsupervised* way while we want to enhance *proposal-level* features in a *supervised* way, where no pre-training dataset is used. We adopt cross-modal pair discrimination loss (Li and Wang 2020) for our mutual matching scheme, where each instance in both modalities defines an unique class and the binary classes for cross-modal pairs, i.e., matched or unmatched, provide us valuable supervisions. Cross-modal image/video retrieval methods (Wang, Li, and Lazebnik 2016; Gabeur et al. 2020) also utilize negative samples in a metric learning framework, yet they still treat the *image/video as a whole*. Thus we address a different problem from them.

3 Model

We propose to use negative sentence samples to construct a cross-modal mutual matching scheme for modeling bidirectional matching relations in visual and language modalities. To enable this auxiliary task, we adopt a late modality fusion from the metric-learning prospective with the advantage of less computational cost than the baseline (Zhang

et al. 2020d). Sec. 3.1 introduces the formulation of temporal grounding. Sec. 3.2 illustrates how to construct our MMN from the baseline. Then, we analyze the two complementary loss of IoU regression and mutual matching in Sec. 3.3.

3.1 Problem Formulation

Given an untrimmed video V and a natural language query S, the temporal grounding task aims to localize a temporal moment (x_s, x_e) that matches the query. We denote the video as a sequence of frames $V = \{x_i\}_{i=1}^{l_v}$, where x_i is a frame and l_v is the total number of frames; the query sentence as a sequence of words $S = \{s_i\}_{i=1}^{l_s}$, where s_i is a word and l_s is the total number of words. Ideally, the retrieved moment (x_s^*, x_e^*) should deliver the same semantic as the sentence S. We adopt the feature vectors in the joint space to represent sentence S and moments (x_i, x_j), therefore the inner product of visual and text features after ℓ_2-norm should be maximized.

3.2 Architecture

Our MMN adopts a Siamese-alike network architecture with a late modality fusion by a simple inner product in the joint visual-language space, as illustrated in Fig. 2.

Language Encoder. Previous works commonly utilize a LSTM (Hochreiter and Schmidhuber 1997) upon a sequence of word vectors embedded by GloVe (Pennington, Socher, and Manning 2014). Yet, some concerns about unfair comparisons in previous methods, e.g., GloVe models pre-trained on different corpora, motivate us to adopt a standard language encoder. We choose DistilBERT (Sanh et al. 2019) for its light-weighted model capacity. For each input sentence S, we first generate the tokens of words by the tokenizer and add a class embedding token '[CLS]' in the beginning. Then we feed the tokens into DistilBERT to get a feature sequence $\{\mathbf{f}_i^S\}_{i=1}^{l_s+1}$, where $\mathbf{f}_i^S \in \mathbb{R}^{d^S}$ and d^S=768 is the feature dimension. There are two commonly used aggregation approaches to get the whole sentence's embedding: 1) global average pooling over all tokens ; 2) the class embedding '[CLS]'. Our experiment shows that global average pooling has faster convergence and better performance in this task, so we take average pooling as default.

Video Encoder. We extract features of the input video and encodes them as a 2D temporal moment feature map following the baseline 2D-TAN (Zhang et al. 2020d). We segment the input video stream into small video clips $\{v_i\}_{i=1}^{l_v/t}$ with each clip v_i containing t frames, then extract clip-level features with an off-the-shelf pre-trained CNN model (e.g., C3D). We perform a fixed-length sampling to obtain N clip-level features for each video by an even stride $\frac{l_v}{t \cdot N}$ and pass the fixed-length features through an FC layer to reduce their dimension, denoted as $\{\mathbf{f}_i^V\}_{i=1}^N$, where $\mathbf{f}_i^V \in \mathbb{R}^{d^V}$. Then we build up the 2D feature map for candidate moments following the baseline (Zhang et al. 2020d) as $\mathbf{F}^M \in \mathbb{R}^{N \times N \times d^V}$, where we adopt the max-pooling as the moment-level feature aggregation strategy following. We also utilize the sparse sampling strategy which removes highly overlapped moments to reduce the number of candidate moments as
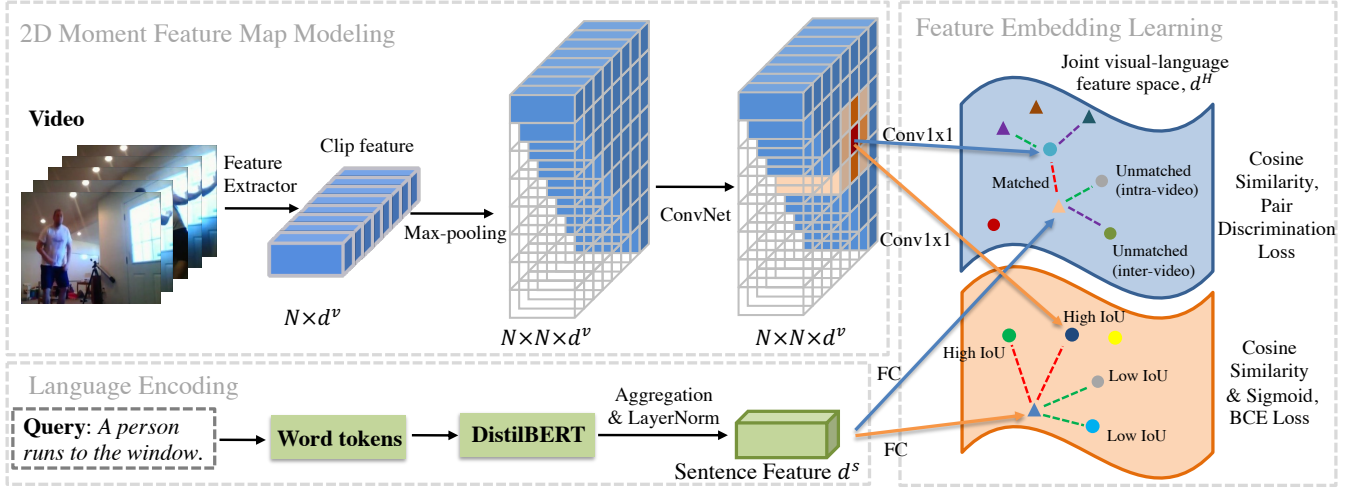
Figure 2: Overview of our framework. Different from the baseline, we adopt a late modality fusion strategy and learning the feature embedding for moments and sentences on two independent spaces with pair discrimination and BCE losses, respectively. The dots and triangles are the feature of moments and sentences. The red dash lines are matched moment-sentence pairs to be pulled in, while green/purple dash lines are negative samples intra/inter-video to be pushed away. **Best view with colors**.

well as computation cost following 2D-TAN. Different from 2D-TAN (Zhang et al. 2020d), we directly model the relations of moments $\mathbf{F} \in \mathbb{R}^{N \times N \times d^V}$ by L layers of 2D convolution with kernel size K only based on visual features. As discussed in (Otani et al. 2020), the performance of 2D-TAN has no significant drop when the order of input video clips is randomly permuted (i.e., the 'sanity check'), which is counter-intuitive. It indicates that to some extent 2D-TAN ignores the visual features and overfits the bias in the dataset, such as temporal distributions of actions conditioned on sentences. The design of our MMN forces the convolution filters to actually utilize the visual features.

Joint Visual-Language Embeddings. Finally, we estimate the matching quality of each moment based on the similarity of two modalities for both supervision signals (i.e., IoU regression and cross-modal mutual matching). We adopt a LayerNorm (Ba, Kiros, and Hinton 2016) for more stable convergence in language feature aggregation. We use a linear projection layer or a 1x1 convolution to project the language and visual features into the same dimension d^H respectively. The final representations of sentence feature are $\mathbf{f}_{mm}^S, \mathbf{f}_{iou}^S \in \mathbb{R}^{d^H}$ for cross-modal mutual matching (subscript mm) and IoU regression (subscript iou). Moment features are $\mathbf{F}_{mm}^V, \mathbf{F}_{iou}^V \in \mathbb{R}^{N \times N \times d^H}$.

$$\mathbf{f}^S = \text{LayerNorm}(\mathbf{f}_1^S) \text{ or } \text{LayerNorm}(\frac{1}{l_s} \sum_{i=2}^{l_s+1} \mathbf{f}_i^S) \quad (1)$$

$$\mathbf{f}_{mm}^S = \mathbf{W}_{mm}\mathbf{f}^S + \mathbf{b}_{mm}, \quad \mathbf{f}_{iou}^S = \mathbf{W}_{iou}\mathbf{f}^S + \mathbf{b}_{iou} \quad (2)$$

$$\mathbf{F}_{mm}^V = \text{conv}_{mm}(\mathbf{F}, 1, 1), \quad \mathbf{F}_{iou}^V = \text{conv}_{iou}(\mathbf{F}, 1, 1) \quad (3)$$

where $\mathbf{W}_\cdot$ and $\mathbf{b}_\cdot$ are learnable parameters, $\text{conv}_\cdot(\mathbf{x}, k, s)$ is 2D convolution with kernel size k and stride s for 2D feature map $\mathbf{x}$. The subscripts id or iou means the weights of

two branches are independent. Then we regard the cosine similarity as moments' estimation scores for both losses.

$$\forall \mathbf{f}_{mm}^V \in \mathbf{F}_{mm}^V, s^{mm} = \mathbf{f}_{mm}^{V\mathrm{T}}\mathbf{f}_{mm}^S$$
$$\forall \mathbf{f}_{iou}^V \in \mathbf{F}_{iou}^V, s^{iou} = \mathbf{f}_{iou}^{V\mathrm{T}}\mathbf{f}_{iou}^S \quad (4)$$

where we enforce the embedding $||\mathbf{f}^V||_2 = ||\mathbf{f}^S||_2 = 1$ via a ℓ_2-normalization layer.

3.3 Loss Functions

Our MMN integrates two complementary losses: a binary cross entropy loss for regressing the IoU and a pair discrimination loss for learning discriminative features.

Binary Cross Entropy. We follow 2D-TAN (Zhang et al. 2020d) to adopt the scaled IoU values y_i for each candidate moments as the supervision signal, as shown in Tab. 1 (IoU regression). The IoU values are linearly scaled from (t_{min}, t_{max}) to $(0,1)$ and truncate the values beyond $(0,1)$. We directly use the value of t_{min} and t_{max} reported in 2D-TAN for fair comparisons. We notice that the range of cosine similarity is $s^{iou} \in (-1,1)$, yet the IoU signal is $y_i \in (0,1)$. So we adopt a commonly-used sigmoid function σ to highlight the value change near the neutral region (e.g., y_i=0.5). We heuristically amplify the s^{iou} by a factor of 10 to make the range of final prediction of our model $p_i^{iou} = \sigma(10 \cdot s_i^{iou})$ cover most of the regions in $(0, 1)$. The regression branch of our MMN is trained by a BCE loss:

$$L_{bce} = -\frac{1}{C} \sum_{i=1}^{C} \left(y_i \log p_i^{iou} + (1 - y_i) \log(1 - p_i^{iou}) \right), \quad (5)$$

where p_i^{iou} is the final score of a moment and C is the total number of valid candidates.

Cross-modal Mutual Matching. As discussed in Sec. 1, our cross-modal mutual matching creates more supervision signals for temporal grounding. By contrasting the positive

Loss	Modality	IoU with GT (moment or its matched sent.)				
		1	(0.5, 1)	(0, 0.5)	0	**other video**
Mutual Matching	vid.	pos.	-	neg.	neg.	neg.
	sent.	pos.	-	neg.	neg.	neg.
IoU Regression	vid.	max	high	low	min	-

Table 1: Supervision signals of our MMN. We categorize moments/sentences as pos./neg. samples according to IoU of this moment (or this sentence's corresponding moment) with the GT moment. We introduce two novel types of supervisions: 1) negative sentence samples in mutual matching scheme (row 2), and 2) inter-video negatives (column 7).

moment-sentence pairs with the negative ones sampled from *both intra and inter videos*, encoders will learn a more discriminative features for both modalities *without any extra pre-training dataset*. As shown in Tab. 1, our mutual matching objective introduces two novel aspects of supervision signals: 1) pairs of the ground-truth moment and pos./neg. sentences (row 2); and 2) inter-video negative pairs (column 7). In contrast, previous detection/regression methods only sample pos./neg. samples based on IoU signals (row 3), and the previous metric learning method (Hendricks et al. 2017) only has single direction of matching (i.e., only row 1) yet lacks the important cross-modal mutual matching (i.e., both row 1 and row 2). Specifically, we adapt the cross-modal pair discrimination loss (Li and Wang 2020) from video-level to proposal-level to learn features for moments $\mathbf{f}_{mm}^V$ and sentences $\mathbf{f}_{mm}^S$, where these features should be similar if the moment-sentence pair is semantically matched and dissimilar if it is semantically unrelated. We adopts the following conditional distribution in a non-parametric softmax form:

$$p(i_s|v) = \frac{\exp\left((\mathbf{f}_i^{S\mathrm{T}}\mathbf{f}^V - m)/\tau_v\right)}{\exp\left((\mathbf{f}_i^{S\mathrm{T}}\mathbf{f}^V - m)/\tau_v\right) + \sum_{j\neq i}^{N_s}\exp\left(\mathbf{f}_j^{S\mathrm{T}}\mathbf{f}^V/\tau_v\right)}$$

$$p(i_v|s) = \frac{\exp\left((\mathbf{f}_i^{V\mathrm{T}}\mathbf{f}^S - m)/\tau_s\right)}{\exp\left((\mathbf{f}_i^{V\mathrm{T}}\mathbf{f}^S - m)/\tau_s\right) + \sum_{j\neq i}^{N_v}\exp\left(\mathbf{f}_j^{V\mathrm{T}}\mathbf{f}^S/\tau_s\right)} \quad (6)$$

where the ith sentence or moment define a instance-level class i_s or i_v, the feature embedding $\mathbf{f}^S$ and $\mathbf{f}^V$ are ℓ_2 normalized, τ_s and τ_v are temperatures and N_s and N_v are total numbers of sampled instances in the batch. Although the conditional distribution of video moments and sentences are similar, their differences are non-trivial due to the selection of negative samples, i.e., $N_v - 1$ negative video moments are sampled from low IoU moments inside videos or moments from other videos; and $N_s - 1$ negative sentences are sampled from other sentences in the video or from other videos. To enable a steady training process, we only adopt the moments whose IoU with ground-truth moment lower than a threshold as negative samples (e.g., ≤ 0.5). To further reduce the potential false negative signals, the sentences similar to the ground-truth sentence is automatically removed from the negative sample set, i.e., by computing their matched moments' IoU with ground-truth moment and removing them if their IoU ≥ 0.5. When we construct the negative sample set, we hold an assumption that only a small portion of inter-video negative samples (both sentences and moments) will semantically close to the positive sample. It is reasonable if

the size of the training video corpus is large enough.

The objective of our cross-modal mutual matching is to maximize the likelihood $\prod_{i=1}^N p(i_s|v_i)\prod_{i=1}^N p(i_v|s_i)$ where N is the total number of moment-sentence pairs for training. The cross-correlation enables the network to effectively capture the mutual information between modalities by guiding the learning process of feature representation learning with the binary pair supervision. The loss function is as follows

$$L_{mm} = -\left(\sum_{i=1}^N \log p(i_v|s_i) + \sum_{i=1}^N \log p(i_s|v_i)\right) \quad (7)$$

Our final loss function L is a linear combination of binary cross entropy loss and mutual matching loss, and the final prediction score s for candidate moments given the query sentence are these two scores' product.

$$L = L_{bce} + \lambda L_{mm}, \quad s = s^{iou} \cdot s^{mm}. \quad (8)$$

4 Experiments

4.1 Datasets

ActivityNet-Captions (Krishna et al. 2017) is built on ActivityNet v1.3 dataset (Heilbron et al. 2015), where videos cover a wide range of complex human actions. It is originally designed for video captioning, and recently introduced into temporal grounding. There are 37,417, 17,505, and 17,031 moment-sentence pairs for training, validation, and testing respectively. Following the setting of 2D-TAN (Zhang et al. 2020d), we report the evaluation result on val_2 set.

TACoS consists of 127 videos selected from the MPII-Cooking dataset (Rohrbach et al. 2012). It is comprised of 18,818 video-language pairs of different cooking activities in the kitchen annotated by (Regneri et al. 2013). A standard split (Gao et al. 2017a) consists of 10,146, 4,589, and 4,083 moment-sentence pairs for training, validation and testing, respectively. We report evaluation results on test set in our experiments.

Charades-STA (Gao et al. 2017a) is an extended version of action recognition and localization dataset Charades (Sigurdsson et al. 2016) by (Gao et al. 2017a) for temporal grounding. It contains 5,338 videos and 12,408 query-moment pairs in the training set, and 1,334 videos and 3,720 query-moment pairs in the test set.

HC-STVG dataset is introduced by (Tang et al. 2020) to spatio-temporally localize an action tubelet of the target person from an untrimmed video based on a given textual description. The dataset contains 5,660 videos selected from AVA (Gu et al. 2018) where each video has one video-sentence pair. A standard split consists of 4,500 pairs for training and 1,160 pairs for testing.

4.2 Experimental Settings

Evaluation Metrics. Following previous setting (Gao et al. 2017a) of temporal grounding, we evaluate our model by computing *Rank n@m*. It is defined as the percentage of sentence queries having at least one correctly localized moment, i.e. IoU $\geq m$, in the top-n retrieved moments. There are specific settings of n and m for different datasets. Specifically,

we report the results as $m \in \{0.5, 0.7\}$ for Charades-STA, $m \in \{0.1, 0.3, 0.5\}$ for TACoS and $m \in \{0.3, 0.5, 0.7\}$ for ActivityNet Captions dataset with $n \in \{1, 5\}$.

For spatio-temporal video grounding, we report the mean of vIoU and vIoU@$\{0.3, 0.5\}$ following HC-STVG benchmark (Tang et al. 2020), where vIoU is computed by the top-1 predicted tube with the GT tube.

Implementation Details. We adopt standard off-the-shelf video feature extractors without fine-tuning on each dataset. Our convolution network for 2D proposal feature modeling uses exactly the same settings with 2D-TAN (Zhang et al. 2020d) (max-pool version) for fair comparisons, including visual features (VGG feature for Charades and C3D feature for ActivityNet-Captions and TACoS), number of sampled clips N, number of 2D convolution network layers L, kernel size K and channels d^V, non maximum suppression (NMS) threshold, scaling thresholds t_{min} and t_{max}. We set the dimension of the joint feature space $d^H = 256$, and temperatures $\tau_s = \tau_v = 0.1$. We use the HuggingFace (Wolf et al. 2019) implementation of DistilBERT (Sanh et al. 2019) with pre-trained model 'distilbert-base-uncased' for better standardization in temporal grounding. In pair discrimination loss, we only sample negative moments lower than IoU=0.5 and set margin m as 0.1, 0.3 and 0.4 for TACoS, ActivityNet-Captions and Charades-STA respectively. We use AdamW (Loshchilov and Hutter 2019) optimizer with learning rate of 1×10^{-4} and batch size 48 for Charades, learning rate of 8×10^{-4} and batch size 48 for ActivityNet Captions and learning rate of 1.5×10^{-3} and batch size 4 for TACoS. We set λ of pair discrimination loss as 0.1 for ActivityNet Captions and 0.05 for Charades and TACoS following the principle that both loss should contribute equally weighted gradients. We early stop the pair discrimination loss when we observe the performance on validation set starts to drop. The learning rate of DistilBERT is always $1/10$ of our main model. Each mini-batch has B *videos* instead of B *moment-sentence pairs*, where B is batch size.

4.3 Ablation Study

Cross-modal Mutual Matching. We investigate the most important part in our method, i.e., 'cross-modal mutual matching', by ablating different types of negative samples on both Charades-STA and ActivityNet Captions datasets, as shown in Tab. 2. Row 1 is the baseline of our model for not using mutual matching. Row 2 uses the same information with previous works (e.g., intra-video moment negatives), which shows similar performance with row 1. By introducing sentence negatives, the notable performance gain is observed in row 3 to row 6 when compared to row 1 and row 2. It proves the effectiveness of our mutual matching scheme. Besides, we conclude some observations based on the differences inside row 3 to row 8: (1) Due to the larger number of negatives, only using inter-video negatives (row 4) tend to be more effective than only using intra-video negatives (row 3), especially on smaller datasets (e.g., Charades-STA); (2) Best results are often obtained by both intra and inter-video negatives (row 8), and sub-optimal results are often achieved in row 5 and 6 by discarding some specific type of inter-

BCE	intra-video vid.	intra-video sent.	inter-video vid.	inter-video sent.	R@1 IoU0.5	R@1 IoU0.7	R@5 IoU0.5	R@5 IoU0.7
✓					40.12	23.89	79.57	53.26
✓	✓				40.65	24.49	77.98	55.89
✓	✓	✓			43.36	26.48	81.85	57.24
✓			✓	✓	46.32	26.59	83.41	56.77
✓	✓	✓	✓		45.13	26.24	_83.69_	58.13
✓	✓	✓		✓	_46.91_	**27.46**	83.09	_58.36_
	✓	✓	✓	✓	44.70	24.92	79.76	53.71
✓	✓	✓	✓	✓	**47.31**	_27.28_	**83.74**	**58.41**

BCE	intra-video vid.	intra-video sent.	inter-video vid.	inter-video sent.	R@1 IoU0.3	R@1 IoU0.5	R@1 IoU0.7	R@5 IoU0.3	R@5 IoU0.5	R@5 IoU0.7
✓					62.73	46.74	27.12	85.67	77.26	61.64
✓	✓				62.27	46.43	27.75	85.80	77.63	61.68
✓	✓	✓			63.41	47.52	**29.69**	86.37	78.17	62.50
✓			✓	✓	_64.64_	48.26	28.75	86.52	78.82	_64.49_
✓	✓	✓	✓		64.01	47.47	28.89	**87.38**	_79.26_	64.05
✓	✓	✓		✓	64.25	48.24	28.96	86.65	78.17	62.62
	✓	✓	✓	✓	63.47	45.28	26.29	85.55	76.89	60.62
✓	✓	✓	✓	✓	**65.05**	**48.59**	_29.26_	_87.25_	**79.50**	**64.76**

Table 2: Ablation on negative samples in our mutual matching scheme. (top) Charades-STA; (bottom) ActivityNet Captions. Intra/inter-video means we use video moments (vid.) or sentences (sent.) inside or cross videos as negatives.

Method	R@1 IoU0.3	R@1 IoU0.5	R@1 IoU0.7	R@5 IoU0.3	R@5 IoU0.5	R@5 IoU0.7
Only Intra	63.41	47.52	**29.69**	86.37	78.17	62.50
#neg. = Intra	_64.18_	_47.55_	28.48	_86.74_	_79.01_	_64.15_
Intra + Inter	**65.05**	**48.59**	_29.26_	**87.25**	**79.50**	**64.76**

Table 3: Ablation on number of negatives (ActivityNet).

video negatives. It indicates hard negative mining of inter-video samples could be further investigated; (3) Only using our mutual matching scheme achieves similar or better performance than original BCE loss on two datasets (row 7 vs. row 1), demonstrating the powerful additional supervision signals introduced by our mutual matching scheme.

Number vs. Type of Negatives. We have shown inter-video negatives are beneficial by introducing much more negatives than previous methods in Tab. 2. However, whether the number or the type of negatives is more important in our method? To ablate this, we construct a baseline that randomly picks the same numbers of negatives with intra-video negatives among all intra and inter-video negatives (row 2). Therefore, it uses the **same** number with only intra-video negatives (row 1) *in each iteration* yet will eventually see more types of negatives in the training process. Tab. 3 shows that the performance of this baseline (row 2) is similar to only using intra-video negatives (row 1) in R@1 metrics yet is better in R@5 metrics, indicating the importance of more types of negatives introduced by our mutual matching scheme.

Previous Methods with DistilBERT. Two baselines (DRN (Zeng et al. 2020) and 2D-TAN (Zhang et al. 2020d)) with DistilBERT are constructed in Tab. 4, which shows that the usage of DistilBERT has very little impact to the final performance of both methods. Therefore, we believe comparisons of MMN with previous SOTA methods are fair.

Aggregation of Text Feature. In Tab. 5 (top), we compare two aggregation strategies for the sentence feature. While many methods in NLP including BERT (Devlin et al. 2019)

Charades-STA	Feature	R@1		R@5	
		IoU0.5	IoU0.7	IoU0.5	IoU0.7
DRN[†] (Zeng et al. 2020)	C3D	*43.26*	*23.95*	85.43	52.82
DRN w/ DistilBERT	C3D	42.62	23.50	*85.84*	*53.59*
2D-TAN (Zhang et al. 2020d)	VGG	39.70	23.31	80.32	51.26
2D-TAN[†]	VGG	40.03	23.15	<u>82.18</u>	49.89
2D-TAN w/ DistilBERT	VGG	<u>40.32</u>	<u>23.81</u>	79.49	<u>51.96</u>

Table 4: Previous methods with DistilBERT. [†] means our implementation.

Charades-STA	R@1		R@5	
	IoU0.5	IoU0.7	IoU0.5	IoU0.7
Class Embedding 'CLS'	45.11	27.25	82.37	58.09
Average Pooling	**47.31**	**27.28**	**83.74**	**58.41**
$m = 0.3$	45.30	27.13	83.33	**59.02**
$m = 0.4$	**47.31**	**27.28**	**83.74**	58.41
$m = 0.5$	46.32	26.82	83.09	58.10
$\lambda = 0.05$	**47.31**	**27.28**	**83.74**	58.41
$\lambda = 0.1$	45.63	27.13	83.34	57.33
$\lambda = 0.2$	44.91	26.95	82.97	57.69
$\tau_v = \tau_s = 0.1$	**47.31**	**27.28**	**83.74**	**58.41**
$\tau_v = \tau_s = 0.2$	43.88	25.81	82.23	57.55

Table 5: Ablation on (top) text feature aggregations and (bottom) hyper-parameters in the loss function.

Figure 3: Sanity Check on the Charades-STA dataset. [†]values are obtained from (Otani et al. 2020).

Figure 4: Performance comparisons decomposed by moment length on Charades-STA, R@1(IoU=0.5).

directly adopt the class embedding as the representation of the whole sentence, we find the average pooling over all words is more robust in our task. It indicates that the scale of video grounding datasets is limited in vocabulary size or annotated sentences, e.g., TACoS dataset only has a vocabulary size of 2255 words and a training set of 10146 sentences, while NLP pre-training datasets commonly have millions of sentences. We empirically use average pooling as default.

Hyper-parameters. In Tab. 5 (bottom), we ablate important hyper-parameters of pair discrimination loss like margin m, temperature τ_s, τ_v and weight λ. Due to the small size of Charades-STA, they do have some influences but no notable change of the gap between our MMN and baseline (Zhang et al. 2020d). We set them by just a few tries on each dataset.

Sanity Check on Visual Input. (Otani et al. 2020) observe that the blind baselines which only uses training set priors or conditional distributions of moments given queries already outperform many previous SOTA methods. So they define a test named 'sanity check' to examine the contribution of visual inputs by evaluating the performance gap between the ordered input clips and randomly shuffled ones. We follow this setting on Charades-STA in Fig. 3: Our model with randomized visual input predicts an even lower result (37.5 vs. 37.8) than Blind-TAN (Otani et al. 2020) which does not use any visual input. Our performance gap between ordered inputs and shuffled ones is also larger than 2D-TAN's, which means our MMN shows better result in this sanity check. It indicates our method better exploits visual information.

Performance Comparisons by Moment Length. To further discover our improvement over the baseline (Zhang et al. 2020d), we decompose the performance comparisons by the ratio of moment length over the video length in Fig. 4. The x-axis is grouped by the moment length ratio which stops at 0.6 because the maximum length of moments on
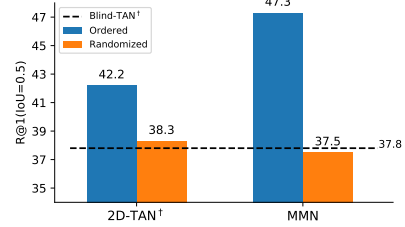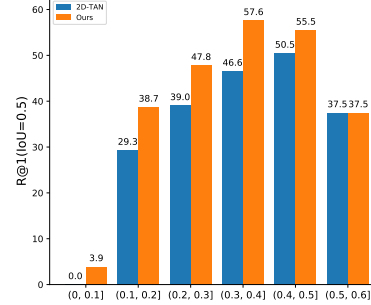
Charades-STA is less than 0.6. It is worth noticing that the number of moments in each ratio interval is not balanced, e.g., GT moments have 1020 instances in (0.1, 0.2] while only 16 in (0.5, 0.6]. Our MMN consistently outperforms 2D-TAN in various moment lengths.

4.4 Comparison with the State of the Art

We achieve highly competitive results on three datasets in temporal grounding: ActivityNet Captions, Charades-STA and TACoS, as reported in Tab. 7, Tab. 6, and Tab. 8. The top-2 performance values in tables are highlighted by **bold** or <u>underline</u>. Based on the results, we have several observations. **Firstly**, our proposed MMN outperforms the strong baseline (Zhang et al. 2020d) with a significant margin by leveraging our mutual matching scheme (i.e., using both video and sentence negative samples from both intra and inter videos). As a result, it outperforms or is on par with most of the previous or recently proposed SOTA methods. Specifically, we compare our MMN with several concurrent approaches which share the similar loss function or architecture with ours, i.e., IVG-DCL (Nan et al. 2021), SSCS (Ding et al. 2021) and FVMR (Gao and Xu 2021). IVG-DCL and SSCS use contrastive loss to align the sentences and video *clips* instead of the sentences and *moments* in MMN. Thus they lack our important auxiliary task 'mutual matching' and ignore the important sentence negative samples. They are also more complex than ours due to the additional IVG module (in IVG-DCL) or captioning/support-set module (in SSCS). As a result, they achieve similar or worse results than ours. They also do not address temporal grounding from a metric-learning perspective, thus their proposed mod-
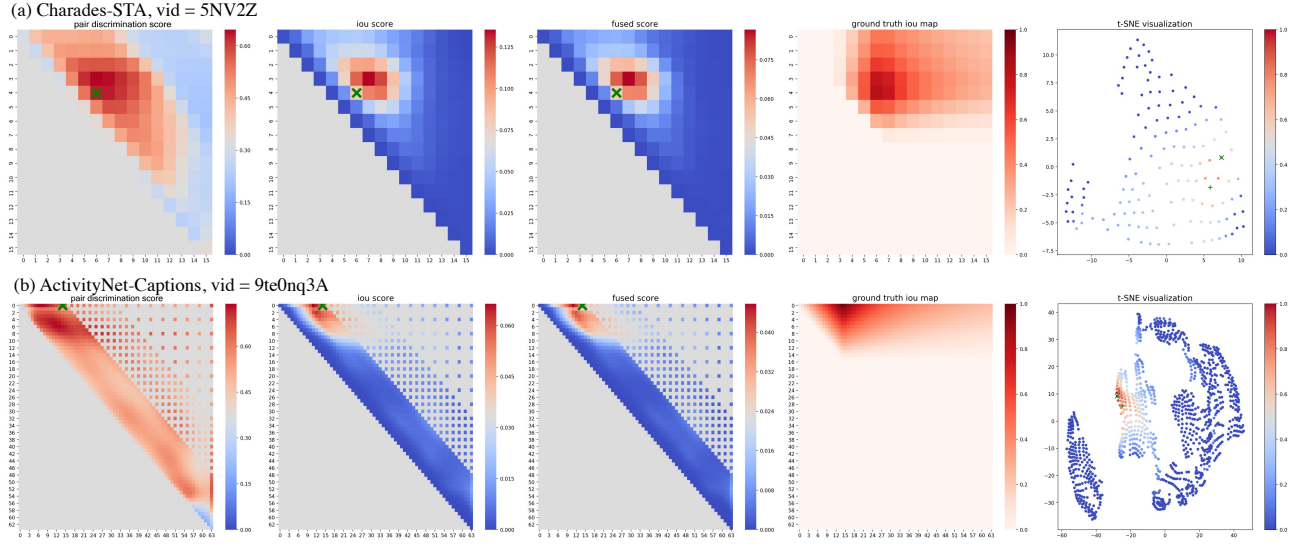
Figure 5: Visualizations on two datasets. Green 'x' in the first three maps shows location of GT moment. In the last map, the matched moment (i.e., has the highest IoU among candidate moments) is a green 'x' and the given sentence is a green '+'. The colors of unmatched moments shows their IoU with GT moment.

Method	R@1		R@5	
	IoU0.5	IoU0.7	IoU0.5	IoU0.7
MCN (Hendricks et al. 2017)	17.46	8.01	48.22	26.73
SAP (Chen and Jiang 2019)	27.42	13.36	66.37	38.15
MAN (Zhang et al. 2019)	41.21	20.54	83.21	51.85
DRN (Zeng et al. 2020)†	42.90	23.68	87.80	54.87
2D-TAN (Zhang et al. 2020d)	39.70	23.31	80.32	51.26
CBLN (Liu et al. 2021)	43.67	24.44	**88.39**	56.49
CPN (Zhao et al. 2021)	46.08	25.06	—	—
FVMR (Gao and Xu 2021)	42.36	24.14	83.97	50.15
SSCS (Ding et al. 2021)	43.15	25.54	84.26	54.17
Our MMN	**47.31**	**27.28**	83.74	**58.41**

Table 6: Performance comparison on Charades-STA (VGG feature). †our re-production with official code is much lower than the value reported in paper.

Method	R@1			R@5		
	IoU0.3	IoU0.5	IoU0.7	IoU0.3	IoU0.5	IoU0.7
MCN (Hendricks et al. 2017)	39.35	21.36	6.43	68.12	53.23	29.70
CTRL (Gao et al. 2017a)	47.43	29.01	10.34	75.32	59.17	37.54
QSPN (Xu et al. 2019)	52.13	33.26	13.43	77.72	62.39	40.78
SCDM (Yuan et al. 2019)	54.80	36.75	19.86	77.29	64.99	41.53
PMI (Chen et al. 2020b)	59.69	38.28	17.83	—	—	—
LGI (Mun, Cho, and Han 2020)	58.52	41.51	23.07	—	—	—
CMIN (Lin et al. 2020)	63.61	44.62	24.48	82.39	69.66	52.96
DRN (Zeng et al. 2020)	—	45.45	24.39	—	77.97	50.30
2D-TAN (Zhang et al. 2020d)	59.46	44.51	26.54	85.53	77.13	61.96
IVG-DCL (Nan et al. 2021)	63.22	43.84	27.10	—	—	—
CBLN (Liu et al. 2021)	**66.34**	48.12	27.60	**88.91**	79.32	63.41
CPN (Zhao et al. 2021)	62.81	45.10	28.10	—	—	—
FVMR (Gao and Xu 2021)	60.63	45.00	26.85	86.11	77.42	61.04
SSCS (Ding et al. 2021)	61.35	46.67	27.56	86.89	78.37	63.78
Our MMN	65.05	**48.59**	**29.26**	87.25	**79.50**	**64.76**

Table 7: Performance comparison on ActivityNet Captions (C3D feature).

ules will only add more computations to their baselines instead of reducing them, which is different from ours. FVMR also address temporal grounding from a metric-learning perspective, yet our supervision signals are totally different: 1) FVMR adopts a knowledge distillation loss where the teacher is the original hadamard product of cross-modal features for fusion with a MLP for prediction and the student is the inner-product in the common space. Our MMN adopts a contrastive loss and directly uses a inner-product in the common space without other fusion strategies as guidance. 2) FVMR also ignores the important sentence negative samples and lacks our proposed 'mutual matching scheme', thus it uses the same supervisions as 2D-TAN and fails to construct more supervisions like ours. Besides, FVMR utilizes a complex semantic role tree for extracting phrase-level fine-grained sentence feature, while we use a simple average pooling over all words to obtain our sentence features. We

achieve a better performance than FVMR on two datasets out of three, including the largest ActivityNet. **Secondly**, it is worth noting that our framework is independent to the moment-relation modeling network or language encoders, so our MMN could be implemented in any proposal-based temporal grounding methods and further boost the performances on strong baselines, e.g., MS-2D-TAN (Zhang et al. 2020c) with multi-scale modelling capacity. **Thirdly**, our computational cost is significantly reduced compared with 2D-TAN (Zhang et al. 2020d) due to shared video features between sentences inside each video. The training process of the baseline takes 36 GPU hours to converge on ActivityNet-Captions dataset while ours only takes 10 GPU hours on the same GPU server. The testing cost could also be reduced if there are multiple sentences in a video, e.g., there are hundreds of sentences per video in TACoS dataset.

Visualizations. In Fig. 5, we show the visualizations on test

Method	R@1			R@5		
	IoU0.1	IoU0.3	IoU0.5	IoU0.1	IoU0.3	IoU0.5
MCN (Hendricks et al. 2017)	14.42	–	5.58	37.35	–	10.33
CTRL (Gao et al. 2017b)	24.32	18.32	13.30	48.73	36.69	25.42
QSPN (Xu et al. 2019)	25.31	20.15	15.23	53.21	36.72	25.30
SCDM (Yuan et al. 2019)	–	26.11	21.17	–	43.35	28.53
CMIN (Lin et al. 2020)	36.68	27.33	19.57	64.93	40.16	32.18
DRN (Zeng et al. 2020)	–	–	23.17	–	–	33.36
2D-TAN (Zhang et al. 2020d)	47.59	37.29	25.32	70.31	57.81	45.04
IVG-DCL (Nan et al. 2021)	49.36	38.84	29.07	–	–	–
CBLN (Liu et al. 2021)	49.16	38.98	27.65	73.12	59.96	46.24
FVMR (Gao and Xu 2021)	**53.12**	**41.48**	29.12	**78.12**	**64.53**	**50.00**
SSCS (Ding et al. 2021)	50.78	41.33	**29.56**	72.53	60.65	48.01
Our MMN	51.39	39.24	26.17	78.03	62.03	47.39

Table 8: Performance comparison on TACoS (C3D feature).

Method	m_vIoU	vIoU@0.3	vIoU@0.5
2D-TAN + WSSTG[†]	15.43	19.83	6.81
STGVT (Tang et al. 2020)[†]	18.15	26.81	9.48
STVGBert (Su, Yu, and Xu 2021)	20.42	29.37	11.31
Yu *et al.* (Yu et al. 2021)	30.02	–	–
Tan *et al.* (Tan et al. 2021)[‡]	30.40	50.40	18.80
Our stage-1+2D-TAN	22.83	36.07	16.96
Our stage-1+MMN	**30.32**	**49.02**	**25.56**

Table 9: Performance comparison on HC-STVG val set. [†]use smaller train set, values reported in (Tang et al. 2020). [‡]uses an ensemble of 10 models.

set of two datasets. Invalid regions of the first three maps are manually set to be in color 'grey'. Due to highly unbalanced pos./neg. samples of BCE loss in 2D-TAN (Zhang et al. 2020d), the absolute value of top-1 result is lower than our expectation. Yet, the gap between top-1 prediction and backgrounds is still discriminative for localizing moments. We conclude that IoU scores tend to be sharp near the GT location yet could be wrong in many situations, while pair discrimination scores tend to cover the right location yet are not sharp among hard negatives. These two score maps are complementary to each other. In our MMN, IoU regression branch is already improved by sharing the same backbone enhanced by pair discrimination loss. The final performance could be further improved a bit (about 0.5%) by fusing their scores with a simple product in our experiments. In the last map, we use t-SNE (van der Maaten and Hinton 2008) to visualize features in the joint space (d^H=256) to a 2D image. Based on the distribution of moments, we believe our joint space has learnt an effective similarity measurement for both video and language modalities.

4.5 Spatio-Temporal Video Grounding

We also extend MMN to spatio-temporally localize action tubes given queries to further show its generalization ability. We find that MMN is even more effective on the HC-STVG dataset due to its smaller scale, where our additional supervisions are very important for limited annotations.

Adapting MMN to the STVG task. We adopt a two-stage method for this task: In stage 1, we first detect human bounding boxes by Faster R-CNN (Ren et al. 2015) and link them to be candidate tubes following ACT (Kalogeiton et al. 2017); then we use CSN (Tran et al. 2019) to extract 1024-d RoI features for each bounding box to form a feature sequence per candidate tube. In stage 2, we use our MMN to temporally refine the candidate tubes with proper adaptation. There are several candidate tubes (as feature sequences) generated in stage 1 in each video, while temporal grounding only have one feature sequence per video. Thus, we generate a 2D moment map for each candidate tube, yet there is only one positive moment per video selected by the highest vIoU globally and all the rest moments including other 2D moment maps are negatives. The IoU regression branch also uses the scaled vIoU as the groundtruth. As a result, vIoU-based supervisions will automatically guide our MMN to find the right tube and the appropriate temporal refinement

and we do not need to design complex evaluation module (e.g., transformer (Tang et al. 2020)) to select the right candidate tube to be refined.

Performance Comparisons. In Tab. 9, our MMN shows significant performance gain over some baselines or recently proposed transformer-based methods, e.g., 2D-TAN+WSSTG (Chen et al. 2019), STGVT (Tang et al. 2020) and STVGBert (Su, Yu, and Xu 2021). Our MMN also shows notable performance gain over 2D-TAN (Zhang et al. 2020d) on the same candidate tubes generated from our stage 1. Finally, our method outperforms some competitive opponents (Yu et al. 2021; Tan et al. 2021) which use strong multi-modal pre-training models (e.g., LXMERT (Tan and Bansal 2019) or MDETR (Kamath et al. 2021)) and ranks first in the HC-STVG challenge of the 3rd PIC workshop[1]. In contrast, cross-modal interaction modeling in our method is solely achieved by our MMN and no multi-modal pre-training model on extra large-scale video datasets is used.

5 Conclusion

In this paper, we propose the Mutual Matching Network (MMN) in the metric-learning prospective for temporal grounding. Particularly, we first propose to use the auxiliary task of *mutual matching* which asks the model to select the correct sentence in a constructed negative sentence set for video moments in addition to existing supervisions. By leveraging the powerful textual negatives, more discriminative features for both modalities are learned by cross-modal mutual matching. Moreover, inter-video negatives in temporal grounding are effectively exploited for the first time in our method. With this simple, effective and efficient framework, we achieve state-of-the-art performance on four challenging video grounding benchmarks: Charades-STA, TACoS, ActivityNet-Captions, and HC-STVG.

---

[1]http://www.picdataset.com/challenge/task/hcvg/

References

Ba, L. J.; Kiros, J. R.; and Hinton, G. E. 2016. Layer Normalization. *CoRR*, abs/1607.06450.

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T. 2018. Temporally Grounding Natural Sentence in Video. In *EMNLP*.

Chen, L.; Lu, C.; Tang, S.; Xiao, J.; Zhang, D.; Tan, C.; and Li, X. 2020a. Rethinking the Bottom-Up Framework for Query-Based Video Localization. In *AAAI*.

Chen, S.; Jiang, W.; Liu, W.; and Jiang, Y. 2020b. Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos. In *ECCV*.

Chen, S.; and Jiang, Y. 2019. Semantic Proposal for Activity Localization in Videos via Sentence Query. In *AAAI*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020c. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*.

Chen, Z.; Ma, L.; Luo, W.; and Wong, K. K. 2019. Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video. In *ACL (1)*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*.

Ding, X.; Wang, N.; Zhang, S.; Cheng, D.; Li, X.; Huang, Z.; Tang, M.; and Gao, X. 2021. Support-Set Based Cross-Supervision for Video Grounding. In *ICCV*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slow-Fast Networks for Video Recognition. In *ICCV*.

Gabeur, V.; Sun, C.; Alahari, K.; and Schmid, C. 2020. Multi-modal Transformer for Video Retrieval. In *ECCV (4)*.

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

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

Gao, J.; Yang, Z.; Sun, C.; Chen, K.; and Nevatia, R. 2017b. TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. In *ICCV*.

Ge, R.; Gao, J.; Chen, K.; and Nevatia, R. 2019. MAC: Mining Activity Concepts for Language-Based Temporal Localization. In *WACV*.

Ghosh, S.; Agarwal, A.; Parekh, Z.; and Hauptmann, A. G. 2019. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *NAACL-HLT (1)*.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; Schmid, C.; and Malik, J. 2018. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *CVPR*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *CVPR (2)*.

He, D.; Zhao, X.; Huang, J.; Li, F.; Liu, X.; and Wen, S. 2019. Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos. In *AAAI*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. C. 2017. Localizing Moments in Video with Natural Language. In *ICCV*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8).

Kalogeiton, V.; Weinzaepfel, P.; Ferrari, V.; and Schmid, C. 2017. Action Tubelet Detector for Spatio-Temporal Action Localization. In *ICCV*.

Kamath, A.; Singh, M.; LeCun, Y.; Misra, I.; Synnaeve, G.; and Carion, N. 2021. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. *CoRR*, abs/2104.12763.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *NeurIPS*.

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

Li, T.; and Wang, L. 2020. Learning Spatiotemporal Features via Video and Text Pair Discrimination. *CoRR*, abs/2001.05691.

Li, Y.; Wang, Z.; Wang, L.; and Wu, G. 2020. Actions as Moving Points. In *ECCV*.

Lin, T.; Liu, X.; Li, X.; Ding, E.; and Wen, S. 2019. BMN: Boundary-Matching Network for Temporal Action Proposal Generation. In *ICCV*.

Lin, Z.; Zhao, Z.; Zhang, Z.; Zhang, Z.; and Cai, D. 2020. Moment Retrieval via Cross-Modal Interaction Networks With Query Reconstruction. *IEEE Trans. Image Process.*, 29.

Liu, B.; Yeung, S.; Chou, E.; Huang, D.; Fei-Fei, L.; and Niebles, J. C. 2018. Temporal Modular Networks for Retrieving Complex Compositional Activities in Videos. In *ECCV*.

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

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, C.; Chen, L.; Tan, C.; Li, X.; and Xiao, J. 2019. DEBUG: A Dense Bottom-Up Grounding Approach for Natural Language Video Localization. In *EMNLP/IJCNLP (1)*.

Miech, A.; Alayrac, J.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-End Learning of Visual Representations From Uncurated Instructional Videos. In *CVPR*.

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

Nan, G.; Qiao, R.; Xiao, Y.; Liu, J.; Leng, S.; Zhang, H.; and Lu, W. 2021. Interventional Video Grounding With Dual Contrastive Learning. In *CVPR*.

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

Otani, M.; Nakashima, Y.; Rahtu, E.; and Heikkilä, J. 2020. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval. In *BMVC*.

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

Ramanishka, V.; Das, A.; Zhang, J.; and Saenko, K. 2017. Top-Down Visual Saliency Guided by Captions. In *CVPR*.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding Action Descriptions in Videos. *Trans. Assoc. Comput. Linguistics*, 1.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.

Rohrbach, M.; Regneri, M.; Andriluka, M.; Amin, S.; Pinkal, M.; and Schiele, B. 2012. Script Data for Attribute-Based Recognition of Composite Activities. In *ECCV*.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

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

Su, R.; Yu, Q.; and Xu, D. 2021. STVGBert: A Visual-Linguistic Transformer Based Framework for Spatio-Temporal Video Grounding. In *ICCV*.

Tan, C.; Lin, Z.; Hu, J.; Li, X.; and Zheng, W. 2021. Augmented 2D-TAN: A Two-stage Approach for Human-centric Spatio-Temporal Video Grounding. *CoRR*, abs/2106.10634.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP/IJCNLP (1)*.

Tang, Z.; Liao, Y.; Liu, S.; Li, G.; Jin, X.; Jiang, H.; Yu, Q.; and Xu, D. 2020. Human-centric Spatio-Temporal Video Grounding With Visual Transformers. *CoRR*, abs/2011.05049.

Tran, D.; Wang, H.; Feiszli, M.; and Torresani, L. 2019. Video Classification With Channel-Separated Convolutional Networks. In *ICCV*.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86).

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

Wang, L.; Li, W.; Li, W.; and Gool, L. V. 2018. Appearance-and-Relation Networks for Video Classification. In *CVPR*.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *CVPR*.

Wang, L.; Tong, Z.; Ji, B.; and Wu, G. 2021. TDN: Temporal Difference Networks for Efficient Action Recognition. In *CVPR*.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Gool, L. V. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV*.

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

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR*, abs/1910.03771.

Wu, J.; Li, G.; Liu, S.; and Lin, L. 2020. Tree-Structured Policy Based Progressive Reinforcement Learning for Temporally Language Grounding in Video. In *AAAI*.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *CVPR*.

Xu, H.; He, K.; Plummer, B. A.; Sigal, L.; Sclaroff, S.; and Saenko, K. 2019. Multilevel Language and Vision Integration for Text-to-Clip Retrieval. In *AAAI*.

Yu, Y.; Wang, X.; Hu, W.; Luo, X.; and Li, C. 2021. 2rd Place Solutions in the HC-STVG track of Person in Context Challenge 2021. *CoRR*, abs/2106.07166.

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *NeurIPS*.

Yuan, Y.; Mei, T.; and Zhu, W. 2019. To Find Where You Talk: Temporal Sentence Localization in Video with Attention Based Location Regression. In *AAAI*.

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

Zhang, B.; Li, Y.; Yuan, C.; Xu, D.; Jiang, P.; and Shan, Y. 2020a. A Simple Yet Effective Method for Video Temporal Grounding with Cross-Modality Attention. *CoRR*, abs/2009.11232.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.; and Davis, L. S. 2019. MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment. In *CVPR*.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020b. Span-based Localizing Network for Natural Language Video Localization. In *ACL*.

Zhang, S.; Peng, H.; Fu, J.; Lu, Y.; and Luo, J. 2020c. Multi-Scale 2D Temporal Adjacent Networks for Moment Localization with Natural Language. *CoRR*, abs/2012.02646.

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

Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; and Lin, D. 2017. Temporal Action Detection with Structured Segment Networks. In *ICCV*.

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