Exploiting Auxiliary Caption for Video Grounding

Hongxiang Li[1], Meng Cao[2], Xuxin Cheng[1], Yaowei Li[1], Zhihong Zhu[1], Yuexian Zou[1†]

[1]School of Electronic and Computer Engineering, Peking University
[2]International Digital Economy Academy (IDEA)
{lihongxiang, chengxx, ywl, zhihongzhu}@stu.pku.edu.cn; {mengcao, zouyx}@pku.edu.cn

Abstract

Video grounding aims to locate a moment of interest matching the given query sentence from an untrimmed video. Previous works ignore the sparsity dilemma in video annotations, which fails to provide the context information between potential events and query sentences in the dataset. In this paper, we contend that exploiting easily available captions which describe general actions, i.e., auxiliary captions defined in our paper, will significantly boost the performance. To this end, we propose an Auxiliary Caption Network (ACNet) for video grounding. Specifically, we first introduce dense video captioning to generate dense captions and then obtain auxiliary captions by Non-Auxiliary Caption Suppression (NACS). To capture the potential information in auxiliary captions, we propose Caption Guided Attention (CGA) project the semantic relations between auxiliary captions and query sentences into temporal space and fuse them into visual representations. Considering the gap between auxiliary captions and ground truth, we propose Asymmetric Cross-modal Contrastive Learning (ACCL) for constructing more negative pairs to maximize cross-modal mutual information. Extensive experiments on three public datasets (i.e., ActivityNet Captions, TACoS and ActivityNet-CG) demonstrate that our method significantly outperforms state-of-the-art methods.

Introduction

Video grounding (Gao et al. 2017; Zhang et al. 2020; Wang et al. 2022b; Cao et al. 2022a; Mun, Cho, and Han 2020; Anne Hendricks et al. 2017; Cao et al. 2022b; Zhang et al. 2022; Cao et al. 2023; Li et al. 2023) aims to identify the timestamps semantically corresponding to the given query within the untrimmed videos. It remains a challenging task since it needs to not only model complex cross-modal interactions, but also capture comprehensive contextual information for semantic alignment.

Currently, due to the costly labeling process, the annotations of existing video grounding datasets (Krishna et al. 2017; Regneri et al. 2013) are *sparse*, *i.e.*, only a few actions are annotated despite the versatile actions within the video. For example in Figure 1, the video from ActivityNet Captions (Krishna et al. 2017) dataset lasts for 218 seconds

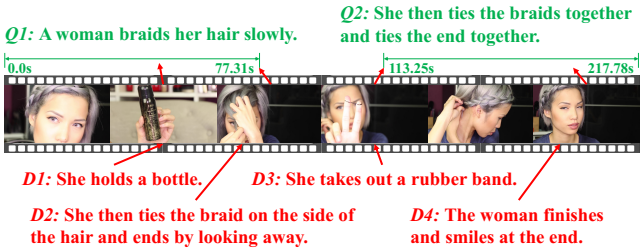

Figure 1: The sparse annotation dilemma in video grounding. The annotated captions (marked by green) in the dataset are sparse while there still exist many uncovered captions (marked by red). This 218-second video from ActivityNet Captions with 2 annotations.

and only 2 moment-sentence pairs (marked by green) are annotated. Previous methods ignore the presence of these unlabeled action instances (marked by red) associated with the query, which will facilitate the grounding. As shown in Figure 1, the missing D_3 contains the process of "`take out a rubber band`", which is preparatory for the action "`tie the braids`" in the queried sentence Q_2.

However, it is labor-intensive to manually annotate all actions in the video. Recently end-to-end dense video captioning (DVC) (Krishna et al. 2017; Li et al. 2018; Suin and Rajagopalan 2020; Wang et al. 2021), which combines event localization and video captioning together, has achieved satisfactory advances. A straightforward solution is to resort to dense video captioning for plausible caption generation. Intuitively, we can incorporate the DVC generated captions as a data augmentation (DA) strategy into the video grounding training. This simple solution, however, suffers from two inherent weaknesses: (1) The generated dense captions of timestamps and sentences may be rough and unreliable. (2) There may be overlaps between dense captions and ground truth. The incorrect caption of the ground truth moment will cause the model to learn incorrect information from training samples. Experimentally, we implement this data augmentation idea on two representative methods (*i.e.*, MMN (Wang et al. 2022b) and 2D-TAN (Zhang et al. 2020)). The experimental results on ActivityNet Captions dataset are shown in Figure 2. We have seen that directly using such data augmentation leads to performance degradation. For example,

† Corresponding author.

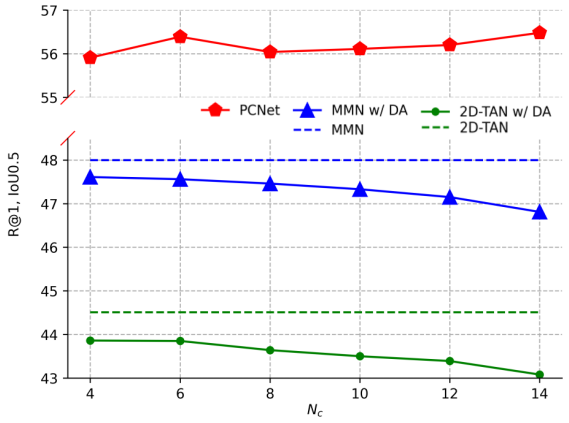

Figure 2: Performance comparison with ACNet and two representative models (Wang et al. 2022b; Zhang et al. 2020) with dense caption data augmentation (w/ DA) on ActivityNet Captions. l_c denotes the number of additional moment-sentence pairs per video.

when using 10 additional dense captions, the performance drops by 1.01% in 2D-TAN.

Despite this intuitive data augmentation does not achieve improvements, we still argue that these dense descriptions contain beneficial information for video grounding. In this paper, we first generate several dense captions from the input video using the off-the-shelf dense video captioning model. To improve the reliability of the generated captions, we propose Non-Auxiliary Caption Suppression (NACS), which selects high-quality and general moment-sentence pairs from the dense captions, defined as auxiliary captions. Unlike the intuitive data augmentation strategy, we propose a novel Auxiliary Caption Network (ACNet) to maximize the utilization of the generated captions rather than simply as extra training data as shown in Figure 3. Our ACNet exploits the potential information embedded in the auxiliary caption through the regression branch and contrastive learning branch.

In the regression branch, we propose Caption Guided Attention (CGA) to investigate the prior knowledge in the auxiliary caption. Our motivation lies in that the auxiliary caption is a well-established prior indication, *i.e.*, it provides an approximate temporal range for the action needed to be grounded. Specifically, we obtain the correlation information between the sentence of the auxiliary caption and the input query through the cross-attention mechanism. Then, we encode the timestamp of the auxiliary caption into a two-dimensional temporal map and linearly project semantic relations into the temporal map to obtain visual features with prior knowledge. In this manner, video clips related to the query semantics are assigned higher weights and unrelated ones are assigned lower weights, providing prompt information for the subsequent localization module.

In the contrastive learning branch, we introduce Asymmetric Cross-modal Contrastive Learning (ACCL) to capture more negative samples in the auxiliary caption. Traditional cross-modal contrastive learning treats all classes equally (Khosla et al. 2020; Wang et al. 2022a), which is exhibited in video grounding as matched fragment-sentence pairs are treated as positive pairs and mismatched fragment-sentence pairs are treated as negative pairs while pulling is applied within positive pairs and pushing among negative pairs. However, the generated auxiliary moment-sentence pairs are not as accurate as the manually annotated ones. Additionally, there exist conflicts between auxiliary caption and ground truth as they are independent of each other. Therefore, while pulling the ground truth pairs together, we push the auxiliary caption sentences away from the ground truth moments but do not push the auxiliary caption moments away from the ground truth sentences. Auxiliary captions provide more descriptions related to the video content, which are treated as hard negative pairs with the ground truth moments. Our ACCL mines more supervision signals from unannotated actions without compromising the original representation capability.

Our main contributions are summarized in three fields:

- We present the sparse annotation dilemma in video grounding and propose to extract information about potential actions from unannotated moments to mitigate it.
- We propose Caption Guided Attention (CGA) to fuse auxiliary captions with visual features to obtain prior knowledge for video grounding. Moreover, we propose Asymmetric Cross-modal Contrastive Learning (ACCL) to mine potential negative pairs.
- Extensive experiments on three public datasets demonstrate the effectiveness and generalizability of ACNet.

Related Work

Video Grounding. Video grounding also known as natural language video localization and video moment retrieval, was first proposed by (Gao et al. 2017; Anne Hendricks et al. 2017). Existing methods are primarily categorized into proposal-based methods and proposal-free methods. Proposal-based methods focus on the representation, ranking, quality and quantity of proposals. They perform various proposal generation methods such as sliding windows (Gao et al. 2017; Anne Hendricks et al. 2017; Ning et al. 2021), proposal networks (Xiao et al. 2021; Chen and Jiang 2019), anchor-based methods (Chen et al. 2018; Liu et al. 2020; Zhang et al. 2020) to extract candidate moments and then semantically match a given query sentence with each candidate through multi-modal fusion. The proposal-free method directly predicts video moments that match query sentences. Specifically, the regression-based method (Yuan, Mei, and Zhu 2019; Chen et al. 2020; Lu et al. 2019; Zeng et al. 2020) calculates the error of time pair with ground truth for model optimization. Span-based method (Zhao et al. 2021; Zhang et al. 2021a) predicts the probabilities of each video frame being the starting, ending and content location of the target moment. Existing methods ignore the annotation sparsity in video grounding, DRN (Zeng et al. 2020) is the pioneer to notice this issue which uses the distance between frames within the ground truth and the starting (ending) frame as dense supervision signals. However, DRN does not exploit

moment-sentence pairs of unannotated video frames. In this paper, we leverage potential information in them to significantly improve the grounding performance.

Dense Video Captioning. Dense video captioning (Krishna et al. 2017; Li et al. 2018; Suin and Rajagopalan 2020; Yang, Cao, and Zou 2023; Mao et al. 2023) techniques typically consist of event detection and caption generation. Most approaches enrich event representations through contextual modeling (Wang et al. 2018), event-level relationships (Wang et al. 2020), or multimodal fusion (Iashin and Rahtu 2020b,a). (Wang et al. 2021) proposed a simple yet effective framework for end-to-end dense video captioning with parallel decoding (PDVC). In practice, by stacking a newly proposed event counter on the top of a transformer decoder, the (Wang et al. 2021) precisely segments the video into several event pieces under the holistic understanding of the video content. In this work, we introduce PDVC (Wang et al. 2021) to generate dense video captions.

Method

Problem Formulation

Given an untrimmed video and a query sentence, we aim to retrieve a video moment that best matches the query sentence, *i.e.*, the start time t^s and end time t^e. We denote the video as $V = \{x_i\}_{i=1}^{T}$ frame by frame, where x_i is the i-th frame in the video and T is the total number of frames. We also represent the given sentence query as $Q = \{w_i\}_{i=1}^{N_q}$ word-by-word, where w_i is the i-th word and N_q is the total number of words.

Feature Encoder

Video encoder. We extract visual representations from the given video and encode them into a 2D temporal moment feature map following (Zhang et al. 2020; Wang et al. 2022b; Cao et al. 2022c). We first segment the input video into small video clips and then perform fixed interval sampling to obtain N_v video clips $V = \{v_i\}_{i=1}^{N_v}$. We extract visual features using a pre-trained CNN model (*e.g.*, C3D) and fed them into the convolutional neural network and average pooling to reduce their dimensions. Then, We construct 2D visual feature maps $F_v \in \mathbb{R}^{N_v \times N_v \times d_v}$ referring to previous works (Zhang et al. 2020; Wang et al. 2022b) based on visual features by max pooling and L layer convolution with kernel size K, where N_v and d_v are the numbers of sampled clips and feature dimension, respectively.

Language encoder. Most of the existing works employ glove embedding with manually designed LSTM as the language encoder (Gao et al. 2017; Zhang et al. 2020), instead of uniformly employing pre-trained models for encoding as in the case of video processing. For a given query sentence, we generate tokens for the words Q by the tokenizer and then feed them into pre-trained BERT (Kenton and Toutanova 2019) with text aggregation to get sentence embedding $F_q \in \mathbb{R}^{1 \times d_s}$, where d_s is the feature dimension.

Unified visual-language feature embedding. We apply two parallel convolutional or linear layers after the encoders to project F_v and F_q to the same feature dimension d_n and employ them for regression (V_r, Q_r) and cross-modal contrastive learning (V_c, Q_c), respectively.

Algorithm 1: Non-Auxiliary Caption Suppression (NACS)

Input: $\mathcal{E} = [e_1, ..., e_M]$, $e_i = (s_i, t_i)$,
$\mathcal{C} = [c_1, ..., c_M]$,
l_c, θ, $\mathcal{F}$
$\mathcal{E}$ is the set of generated moment-sentence pairs
$\mathcal{C}$ contains the corresponding confidence scores
l_c and θ are predefined values
$\mathcal{F}$ records the annotated video intervals
Output: $\mathcal{U} \leftarrow \{\}$
begin
  while $\mathcal{E} \neq$ *empty* and $\mathcal{U}.length < l_c$ **do**
    $m \leftarrow \text{argmax}\, \mathcal{C}$;
    $\mathcal{U} \leftarrow \mathcal{U} \cup e_m$; $\mathcal{E} \leftarrow \mathcal{E} - e_m$;
    $\mathcal{C} \leftarrow \mathcal{C} - c_m$; $\mathcal{F} \leftarrow \mathcal{F} \cup t_m$;
    for e_i *in* $\mathcal{E}$ **do**
      $c_i \leftarrow exp(-\frac{IoU(\mathcal{F}, t_i)^2}{\theta}) c_i, \forall t_i \notin \mathcal{U}$;
    end
  end
  return $\mathcal{U}$
end

Auxiliary Caption Generation

In general, queries in video grounding should be visually based on the temporal region, but the boundaries of the generated dense captions are rough. Moreover, due to the complexity of the video content, there are overlaps between the dense caption intervals and the ground truth intervals. The incorrect descriptions of ground truth are detrimental to the learning of the model. To solve the above issues, we propose to exploit a reliable moment-sentence pair from the generated dense caption, *i.e.*auxiliary caption.

Specifically, we first feed the input video into an off-the-shelf dense video captioning model (*i.e.*PDVC (Wang et al. 2021)) to generate the dense caption set $E = \{s_i, t_i, c_i^s, c_i^p\}_{i=1}^{M}$, where s_i and t_i are the generated descriptions and corresponding timestamps, respectively; c_i^s and c_i^p are the confidence scores of sentences and proposals, respectively; M is the pre-defined number of dense captions per video. Then, we propose Non-Auxiliary Caption Suppression (NACS) inspired by (Bodla et al. 2017) for set E. The computation process is shown in Algorithm 1. To minimize the interval overlap between auxiliary captions and between auxiliary captions and ground truth, we define $\mathcal{F}$ to record the intervals that the video is currently annotated with, which is initialized to all ground truth intervals. We calculate the confidence scores $\mathcal{C}$ and sort $\mathcal{E}$ in descending order accordingly by $\mathcal{C}$. For each e_i, its confidence score c_i is defined as follows:

$$c_i = (c_i^s + c_i^p)\frac{t_i^e - t_i^s}{d_i} \tag{1}$$

where t_i^s and t_i^e are the start and end timestamps, respec-

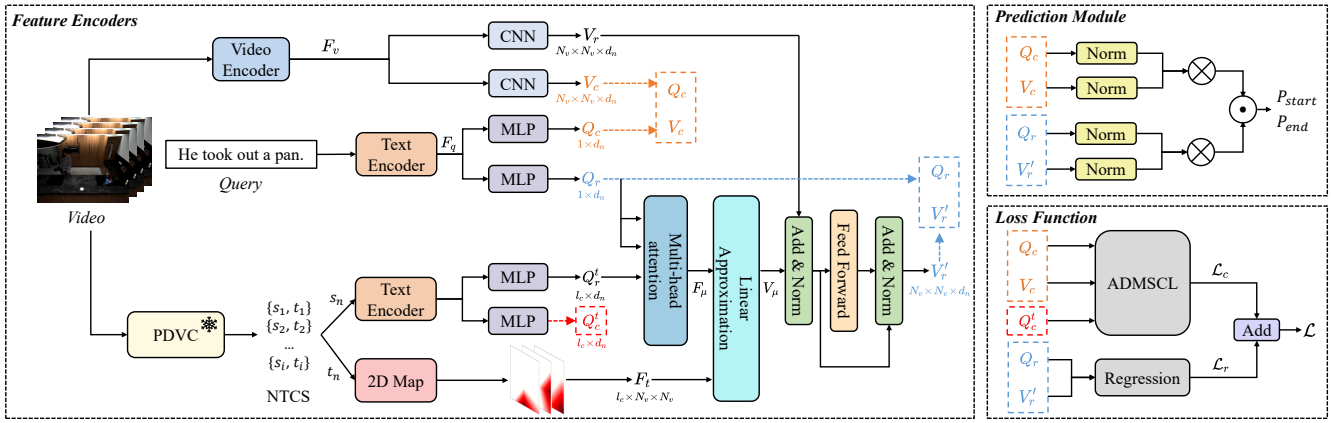

Figure 3: Overview of the proposed Auxiliary Caption Network (ACNet). Auxiliary Caption is filtered through our proposed Non-Auxiliary Caption Suppression algorithm (NACS) from PDVC (Wang et al. 2021) outputs. We convert the timestamp of the auxiliary caption to the 2D map form following (Zhang et al. 2020; Wang et al. 2022b). Then, video segments and query sentences are encoded by the respective feature encoders for regression learning and cross-modal contrastive learning. In the regression branch, Caption Guided Attention (CGA) calculates semantic relations between query features Q_r and auxiliary caption features Q_r^t. Then we project them to visual space to obtain visual representations V_r' with prior knowledge. V_r' and query features Q_r are used for prediction and loss computation. In the cross-modal learning branch, the encoded video features V_c and query features Q_c are directly fed into the prediction module and loss function. $\otimes$ and $\odot$ indicate matrix multiplication and Hadamard product, respectively.

tively, and d_i is the duration of the whole video. The action described by e_i is considered a general action if it has a long duration, and is given a higher score. Then, the e_i with the highest c_i is selected and the annotated video interval $\mathcal{F}$ is updated. Finally, the confidence scores c_i of other e_i are decayed with a Gaussian penalty function (Bodla et al. 2017) according to the degree of overlap with $\mathcal{F}$. The above operations are repeated until $\mathcal{E}$ is empty or the number of elements in $\mathcal{U}$ is equal to l_c. Finally, as with the query sentence, sentences of auxiliary captions are encoded as Q_c^t and Q_r^t for two branches, respectively. We refer to 2D-TAN (Zhang et al. 2020) to encode timestamps of auxiliary captions as two-dimensional temporal maps $F_t \in \mathbb{R}^{l_c \times N_v \times N_v}$, where l_c is the number of auxiliary captions. We provide details of the 2D temporal map in the supplementary material.

Caption Guided Attention (CGA)

The CGA is responsible for extracting the prior knowledge and coarse-grained estimation about the target moment from the auxiliary caption as shown in Figure 4. We first employ the co-attention mechanism to obtain the semantic relations F_μ between the sentence features Q_r^t of auxiliary caption and the query sentence features Q_r:

$$F_\mu = \mathrm{MHA}(Q_r^t, Q_r, Q_r) \tag{2}$$

where MHA stands for standard multi-head attention (Vaswani et al. 2017) which consists of m parallel heads and each head is defined as scaled dot-product

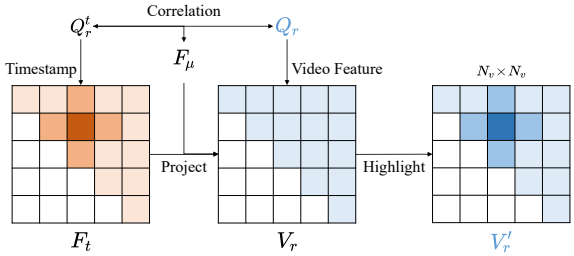

Figure 4: Illustration of our Caption Guided Attention (CGA).

attention:

$$\mathrm{Att}_i(X,Y) = \mathrm{softmax}\left(\frac{X\mathrm{W}_i^{\mathrm{Q}}\left(Y\mathrm{W}_i^{\mathrm{K}}\right)^T}{\sqrt{d_m}}\right)Y\mathrm{W}_i^{\mathrm{V}} \tag{3}$$

$$\mathrm{MHA}(X,Y) = [\mathrm{Att}_1(X,Y);\ldots;\mathrm{Att}_n(X,Y)]\mathrm{W}^{\mathrm{O}} \tag{4}$$

where $X \in \mathbb{R}^{l_x \times d}$ and $Y \in \mathbb{R}^{l_y \times d}$ denote the Query matrix and the Key/Value matrix, respectively; $\mathrm{W}_i^{\mathrm{Q}}$, $\mathrm{W}_i^{\mathrm{K}}$, $\mathrm{W}_i^{\mathrm{V}} \in \mathbb{R}^{d \times d_n}$ and $\mathrm{W}^{\mathrm{O}} \in \mathbb{R}^{d \times d}$ are learnable parameters, where $d_m = d/m$. $[\cdot;\cdot]$ stands for concatenation operation.

Then, we linearly project the semantic relation feature F_μ onto the two-dimensional temporal map F_t to obtain prior knowledge V_μ:

$$V_\mu = \mathrm{MLP}(F_\mu \otimes F_t) \tag{5}$$

where $\otimes$ represents the matrix multiplication. Note that the value in the temporal map F_t represents the intersection over union (IoU), *i.e.*temporal correlation, between the current

clip and the corresponding clip of Q_r. Finally, we obtain V_r' used to predict the target moment by integrating prior knowledge V_μ to visual features V_r by a fully connected feed-forward network:

$$V_r' = \max\left(0, (V_r + V_\mu)\mathrm{W_f} + \mathrm{b_f}\right)\mathrm{W_{ff}} + \mathrm{b_{ff}} \tag{6}$$

where $\max(0, *)$ represents the ReLU activation function; $\mathrm{W_f}$ and $\mathrm{W_{ff}}$ denote learnable matrices for linear transformation; $\mathrm{b_f}$ and $\mathrm{b_{ff}}$ represent the bias terms. In this way, we assign different weights to the video features according to the semantic and temporal position of the auxiliary caption,

Asymmetric Cross-modal Contrastive Learning (ACCL)

For traditional cross-modal contrastive learning, matched pairs are considered as positive pairs and mismatched pairs are considered as negative pairs. However, the temporal boundaries of auxiliary caption may be coarse and should not be pulled close to the corresponding sentences for the localization task. In addition, the intervals of auxiliary caption may overlap with ground truth so as to disagree on the same moment. Therefore, we propose asymmetric cross-modal contrastive learning (ACCL). We consider video grounding as a dual matching task, *i.e.*moment to text and text to moment. Figure 5 illustrates the core idea of ACCL: ACCL applies pulling and pushing in ground truth pairs, and applies pushing between ground truth moments and prompt sentences. We adopt the noise contrastive estimation (NCE) (Gutmann and Hyvärinen 2010) to calculate our ACCL loss, which is defined as:

$$\mathcal{L}_c = \lambda_v \mathcal{I}_{v\to s} + \lambda_s \mathcal{I}_{s\to v} \tag{7}$$

$$\mathcal{I}_{v\to s} = -\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \log \frac{\exp(f(v_i)^\top f(s_i)/\tau_v)}{\sum_{j\in\mathcal{A}_s}\exp(f(v_i)^\top f(s_j)/\tau_v)} \tag{8}$$

$$\mathcal{I}_{s\to v} = -\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \log \frac{\exp(f(s_i)^\top f(v_i)/\tau_s)}{\sum_{j\in\mathcal{A}_v}\exp(f(s_i)^\top f(v_j)/\tau_s)} \tag{9}$$

where i and j are indexes of video moment v or sentence s from V_c, Q_c and Q_c^t; λ_v and λ_s are hyperparameters to balance the contribution of contrastive loss for each direction; τ_v and τ_s are temperatures. At first glance, $\mathcal{I}_{v\to s}$ and $\mathcal{I}_{s\to v}$ *seem* identical to the vanilla cross-modal contrastive learning loss. However, the *key difference* lies in the definitions of $\mathcal{P}$, $\mathcal{A}_s$ and $\mathcal{A}_v$, as we detail below.

Asymmetry of Positive pairs and Negative pairs (APN). We do not pull moments of the auxiliary caption and sentences together, *i.e.*, $\mathcal{P} = \mathcal{G}$. This design is motivated by the fact that we cannot guarantee the accuracy of the auxiliary caption. The boundaries of the auxiliary caption moments are rough, while video grounding is an exact and frame-level matching task. If we construct them as positive pairs, which will hinder cross-modal learning for exact matching.

Asymmetry of Negative pairs in Dual Matching (ANDM). Moment-sentence pairs of auxiliary caption are only contained in $\mathcal{A}_s$ and not in $\mathcal{A}_v$, *i.e.*, $\mathcal{A}_s = \mathcal{G}_s \cup \mathcal{D}_s$, $\mathcal{A}_v = \mathcal{G}_v$. We only push target moments away from auxiliary caption sentences, but do not push query sentences away from auxiliary caption moments. Since auxiliary caption moments and target moments are independent of each other and they may refer to the same video moments, *i.e.*, it is possible for auxiliary caption moments to match with query sentences. On the other hand, the auxiliary caption sentences provide more descriptions of the video content, and we treat the moment-sentence pairs they form with the ground truth moments as hard negative pairs to enhance joint representation learning.

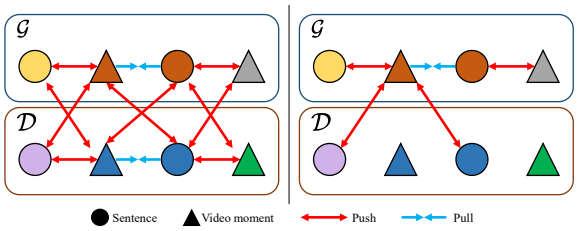

Figure 5: Illustration of our asymmetric push-and-pull strategy, in contrast to those in the original supervised contrastive learning, where elements with the same color mean they come from the same moment-sentence pair. $\mathcal{G}$ and $\mathcal{D}$ are the sets of moment-sentence pairs of ground truth and auxiliary caption, respectively.

Training and Inference

Training. In the regression branch, we employ cross-entropy loss to optimize the model:

$$\mathcal{L}_r = \frac{1}{C}\sum_{i=1}^{C} y_i \log p_i + (1-y_i)\log(1-p_i) \tag{10}$$

where p_i is the prediction score of a moment and C is the total number of valid candidates. Our contrastive loss relies on the binary supervision signal to learn cross-modal alignment and the regression loss counts on the IoU supervision signal for moment ranking. Finally, we employ these two complementary losses for training. The overall training loss $\mathcal{L}$ of our model is

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r \tag{11}$$

where λ_c and λ_b are hyperparameters to balance the contribution of each loss.

Inference. During inference, we calculate the cosine similarity of the video moments and queries as the prediction scores

$$\mathcal{S}_r = \sigma(f(Q_r)f(V_r')^\top),\ \mathcal{S}_c = f(Q_c)f(V_c)^\top \tag{12}$$

where σ is the sigmoid function.

Due to the difference in the value region of $\mathcal{S}_r$ and $\mathcal{S}_c$ (especially the negative regions), we fuse them after scaling to obtain the final prediction scores $\mathcal{S}$.

$$\mathcal{S} = \mathcal{S}_r \odot (\frac{\mathcal{S}_c + 1}{2})^\gamma \tag{13}$$

where $\odot$ denote the element-wise multiplication and γ is the hyperparameter. Finally, We rank all the moment proposals according to $\mathcal{S}$ followed by a non-maximum suppression (NMS) function.

Method	ActivityNet Captions						TACoS					
	R@1 IoU0.3	R@1 IoU0.5	R@1 IoU0.7	R@5 IoU0.3	R@5 IoU0.5	R@5 IoU0.7	R@1 IoU0.1	R@1 IoU0.3	R@1 IoU0.5	R@5 IoU0.1	R@5 IoU0.3	R@5 IoU0.5
CTRL	47.43	29.01	10.34	75.32	59.17	37.54	24.32	18.32	13.30	48.73	36.69	25.42
CBP	54.30	35.76	17.80	77.63	65.89	46.20	–	27.31	24.79	–	43.64	37.40
SCDM	54.80	36.75	19.86	77.29	64.99	41.53	–	26.11	21.17	–	40.16	32.18
2D-TAN	59.45	44.51	26.54	85.53	77.13	61.96	47.59	37.29	25.32	70.31	57.81	45.04
DRN	–	45.45	24.39	–	77.97	50.30	–	–	23.17	–	–	33.36
FVMR	60.63	45.00	26.85	86.11	77.42	61.04	53.12	41.48	29.12	78.12	64.53	50.00
RaNet	–	45.59	28.67	–	75.93	62.97	–	43.34	33.54	–	67.33	55.09
DPIN	–	47.27	28.31	–	77.45	60.03	–	46.74	32.92	–	62.16	50.26
MATN	–	48.02	31.78	–	78.02	63.18	–	48.79	37.57	–	67.63	57.91
CBLN	66.34	48.12	27.60	88.91	79.32	63.41	49.16	38.98	27.65	73.12	59.96	46.24
SMIN	–	48.46	30.34	–	81.16	62.11	–	48.01	35.24	–	65.18	53.36
GTR	–	50.57	29.11	–	80.43	65.14	–	40.39	30.22	–	61.94	47.73
MMN	65.05	48.59	29.26	87.25	79.50	64.76	51.39	39.24	26.17	78.03	62.03	47.39
SPL	–	52.89	32.04	–	82.65	67.21	–	42.73	32.58	–	64.30	50.17
G2L	–	51.68	33.35	–	81.32	67.60	–	42.74	30.95	–	65.83	49.86
ACNet	66.82	52.51	32.51	87.11	79.89	66.68	57.66	48.13	36.79	80.11	69.08	58.10
ACNet$^{\diamond}$	67.07	53.55	34.68	88.21	80.94	67.78	58.76	48.74	37.14	82.43	71.47	60.66
ACNet$^{\natural}$	**70.31**	**56.39**	**38.19**	**89.26**	**82.87**	**70.77**	**62.76**	**51.64**	**38.84**	**86.83**	**74.73**	**62.86**

Table 1: Performance comparisons on ActivityNet Captions and TACoS. $^{\diamond}$ denotes using the generated auxiliary captions and $^{\natural}$ denotes introducing manual annotations from other moments within the video as auxiliary captions during inference.

Experiments

Datasets and Evaluation

ActivityNet Captions. ActivityNet Captions (Krishna et al. 2017) contains 20,000 untrimmed videos and 100,000 descriptions from YouTube (Caba Heilbron et al. 2015), covering a wide range of complex human behavior. The average length of the videos is 2 minutes, while video clips with annotations have much larger variations, ranging from a few seconds to over 3 minutes. Following the public split, we use 37417, 17505 and 17031 sentence-video pairs for training, validation and testing, respectively.

TACoS. TACoS (Regneri et al. 2013) contains 127 videos from the cooking scenarios, with an average of around 7 minutes. We follow the standard split (Gao et al. 2017), which has 10146, 4589 and 4083 video query pairs for training, validation and testing, respectively.

ActivityNet-CG. ActivityNet-CG (Li et al. 2022) aims to evaluate how well a model can generalize to query sentences that contain novel compositions or novel words. It is a new split of ActivityNet Captions, which is re-split into four sets: training, novel-composition, novel-word, and test-trivial.

Evaluation. Following previous work (Gao et al. 2017; Zhang et al. 2020), we adopt "R@n, IoU=m" as the evaluation metric. It calculates the percentage of IoU greater than "m" between at least one of the top "n" video moments retrieved and the ground truth.

Implementation Details

Following (Zhang et al. 2020; Wang et al. 2022b), we employed a 2D feature map to generate moment proposals. For the input video, we used exactly the same settings as in the previous work (Wang et al. 2022b) for a fair comparison, including visual features (both C3D features), NMS thresholds (0.5, 0.4), number of sampled clips (64, 128), number of 2D convolution network layers (3, 4) and kernels (4, 2) for ActivityNet Captions and TACoS, respectively. For the query sentence, the pre-trained BERT (Kenton and Toutanova 2019) was employed for each word of the query. Specifically, the average pooling results of the last two layers are used to obtain the embedding of the whole sentence. During the training, we used AdamW (Loshchilov and Hutter 2018) optimizer to train our model with learning rate of 8×10^{-4}. The batch size B was set to 48 and 8 for ActivityNet Captions and TACoS, respectively. We employed the same settings as ActivityNet Captions on ActivityNet-CG.

Comparison with State-of-the-art Methods

Benchmark. We compare our ACNet with state-of-the-art methods in Table 1. ACNet achieves significant improvements compared with all other methods. Specifically, on ActivityNet Captions, our ACNet achieves performance improvements of up to 6% compared with the cutting edge method SPL (Liu and Hu 2022). SPL (Liu and Hu 2022) investigates the imbalance of positive and negative frames in video grounding and develops a coarse-grained and fine-grained two-step framework, but does not consider the relationship between potential actions and queries. In contrast, our method encodes the video feature under the guidance of the auxiliary caption with a stronger correlation to the query. For TACoS, our ACNet outperforms the strongest competitor MATN (Zhang et al. 2021b) by up to 7 points. MATN (Zhang et al. 2021b) proposes a multi-level aggregated transformer, but it can easily overfit to the point of confusing similar actions due to the neglect of the sparse annotation dilemma. Our ACNet mines more supervision sig-

Method	Test-Trivial		Novel-Comp	
	R@1 IoU0.5	R@1 IoU0.7	R@1 IoU0.5	R@1 IoU0.7
TMN	16.82	7.01	8.74	4.39
TSP-PRL	34.27	18.80	14.74	1.43
VSLNet	39.27	23.12	20.21	9.18
LGI	43.56	23.29	23.21	9.02
2D-TAN	44.50	26.03	22.80	9.95
VISA	47.13	29.64	31.51	16.73
ACNet	**51.81**	**33.52**	**33.30**	**17.09**
ACNet†	46.33	28.67	30.71	15.80

Table 2: Performance comparison on ActivityNet-CG. "†" denotes without NACS.

NACS	CGA	ACCL	Reg	R@1 IoU0.3	R@1 IoU0.5	R@1 IoU0.7
			✓	62.73	46.74	27.12
		✓		64.57	47.28	28.09
✓		✓		66.74	51.83	32.29
	✓		✓	67.70	52.08	32.02
		✓	✓	65.03	50.24	30.02
✓		✓	✓	68.83	54.85	36.48
	✓	✓	✓	68.36	55.27	36.91
✓	✓	✓	✓	**70.31**	**56.39**	**38.19**

Table 3: Component ablations on ActivityNet Captions.

nals from the unannotated moments and employs two complementary loss functions to improve the grounding quality. Notably, most methods cannot achieve the best performance on both datasets simultaneously due to the differences between the two datasets, but ACNet does, which demonstrates the superiority of our method.

Compositional Generalization. Table 2 shows the result comparison between state-of-the-art methods on ActivityNet-CG. Unlike ActivityNet Captions and TACoS, ActivityNet-CG focuses on verifying the generalizability of the model on novel compositions or novel words, proposed by VISA (Li et al. 2022). We observe that our ACNet brings performance improvement of up to 4% compared with VISA (Li et al. 2022), demonstrating the excellent compositional generalization of our model. Notably, our variant "†" model is weaker than VISA on all splits, indicating that auxiliary caption is crucial for generalizability.

Ablation Study

Main Ablation Study. In Table 3, we conduct a thorough ablation study on the proposed components to verify their effectiveness. The first two rows of Table 3 show our single-branch base model. Based on these, we add NACS and CGA respectively. it can be noticed that the performance is improved by about 4% and 5% respectively, as shown in the third and fourth rows of Table 3. Row 5 of Table 3 shows our two-branch base model, which improves "IoU=0.5" to 50.24. In rows 6 and 7 of Table 3, we add NACS and CGA, respectively, to the two-branch model and find that the performance improves again by about 5%. The last row of Table 3 shows the performance of our full model, which further improves the "IoU=0.5" to 56.39% and achieves the best performance among ablation variants.

Model	Training	Inference
2D-TAN (Zhang et al. 2020)	0.13s	32s
MMN (Wang et al. 2022b)	0.32s	37s
Base Model	0.39s	40s
ACNet	0.94s	53s

Table 4: Time consumption on ActivityNet-Captions.

Model	R@1 0.3	R@1 0.5	R@1 0.7
CL	66.25	48.59	30.34
w/o APN	67.43	53.79	37.68
w/o ANDM	67.54	53.58	37.70
Full ACCL	70.31	56.39	38.19

Table 5: Ablation studies of ACCL on ActivityNet Captions.

Comparisons of Time Consumption. In Table 4, we compute the average training time per iteration and total inference time. Our method requires more computational costs but these are worth compared to the significant performance improvements.

Effect of Asymmetric Components. To evaluate the detailed components in ACCL more deeply, we conduct an ablation study of APN and ANDM on ActivityNet Captions in Table 5. We observe that removing any of the components brings significant performance degradation, indicating that this asymmetric design is capable of mining more hard negative samples from the auxiliary caption and thus improving the representation learning.

Conclusion

In this paper, we propose an Auxiliary Caption Network (ACNet) for video grounding. Firstly, we propose Non-Auxiliary Caption Suppression (NACS) to obtain auxiliary captions from dense captions. Then, we design a simple but effective Caption Guided Attention (CGA) to extract prior knowledge from the auxiliary captions and approximately locate the target moment. Moreover, we propose Asymmetric Cross-modal Contrastive Learning (ACCL) to fully mine negative pairs and construct extra supervision signals from unannotated video clips. Extensive experiments have demonstrated that ACNet can achieve excellent performance and superior generalizability on public datasets.

Acknowledgments

This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No: GXWD20201231165807007-20200814115301001).

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, 5803–5812.

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, 5561–5569.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.

Cao, M.; Jiang, J.; Chen, L.; and Zou, Y. 2022a. Correspondence matters for video referring expression comprehension. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4967–4976.

Cao, M.; Wei, F.; Xu, C.; Geng, X.; Chen, L.; Zhang, C.; Zou, Y.; Shen, T.; and Jiang, D. 2023. Iterative Proposal Refinement for Weakly-Supervised Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6524–6534.

Cao, M.; Yang, T.; Weng, J.; Zhang, C.; Wang, J.; and Zou, Y. 2022b. Locvtp: Video-text pre-training for temporal localization. In *European Conference on Computer Vision*, 38–56. Springer.

Cao, M.; Zhang, C.; Chen, L.; Shou, M. Z.; and Zou, Y. 2022c. Deep motion prior for weakly-supervised temporal action localization. *IEEE Transactions on Image Processing*, 31: 5203–5213.

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 162–171.

Chen, L.; Lu, C.; Tang, S.; Xiao, J.; Zhang, D.; Tan, C.; and Li, X. 2020. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10551–10558.

Chen, S.; and Jiang, Y.-G. 2019. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8199–8206.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, 5267–5275.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304. JMLR Workshop and Conference Proceedings.

Iashin, V.; and Rahtu, E. 2020a. A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer. In *The 31st British Machine Vision Virtual Conference*. British Machine Vision Association, BMVA.

Iashin, V.; and Rahtu, E. 2020b. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 958–959.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33: 18661–18673.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Li, H.; Cao, M.; Cheng, X.; Li, Y.; Zhu, Z.; and Zou, Y. 2023. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12032–12042.

Li, J.; Xie, J.; Qian, L.; Zhu, L.; Tang, S.; Wu, F.; Yang, Y.; Zhuang, Y.; and Wang, X. E. 2022. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3032–3041.

Li, Y.; Yao, T.; Pan, Y.; Chao, H.; and Mei, T. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7492–7500.

Liu, D.; and Hu, W. 2022. Skimming, Locating, then Perusing: A Human-Like Framework for Natural Language Video Localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4536–4545.

Liu, D.; Qu, X.; Liu, X.-Y.; Dong, J.; Zhou, P.; and Xu, Z. 2020. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4070–4078.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5144–5153.

Mao, Y.; Xiao, J.; Zhang, D.; Cao, M.; Shao, J.; Zhuang, Y.; and Chen, L. 2023. Improving Reference-based Distinctive Image Captioning with Contrastive Rewards. *arXiv preprint arXiv:2306.14259*.

Mun, J.; Cho, M.; and Han, B. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10810–10819.

Ning, K.; Xie, L.; Liu, J.; Wu, F.; and Tian, Q. 2021. Interaction-integrated network for natural language moment localization. *IEEE Transactions on Image Processing*, 30: 2538–2548.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding Action Descriptions in Videos. *Transactions of the Association for Computational Linguistics*, 1: 25–36.

Suin, M.; and Rajagopalan, A. 2020. An efficient framework for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12039–12046.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Zhang, A.; Zhu, Y.; Zheng, S.; Li, M.; Smola, A. J.; and Wang, Z. 2022a. Partial and Asymmetric Contrastive Learning for Out-of-Distribution Detection in Long-Tailed Recognition. In *International Conference on Machine Learning*, 23446–23458. PMLR.

Wang, J.; Jiang, W.; Ma, L.; Liu, W.; and Xu, Y. 2018. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7190–7198.

Wang, T.; Zhang, R.; Lu, Z.; Zheng, F.; Cheng, R.; and Luo, P. 2021. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6847–6857.

Wang, T.; Zheng, H.; Yu, M.; Tian, Q.; and Hu, H. 2020. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5): 1890–1900.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022b. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2613–2623.

Xiao, S.; Chen, L.; Zhang, S.; Ji, W.; Shao, J.; Ye, L.; and Xiao, J. 2021. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2986–2994.

Yang, B.; Cao, M.; and Zou, Y. 2023. Concept-Aware Video Captioning: Describing Videos With Effective Prior Information. *IEEE Transactions on Image Processing*.

Yuan, Y.; Mei, T.; and Zhu, W. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9159–9166.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10287–10296.

Zhang, C.; Yang, T.; Weng, J.; Cao, M.; Wang, J.; and Zou, Y. 2022. Unsupervised pre-training for temporal action localization tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14031–14041.

Zhang, H.; Sun, A.; Jing, W.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021a. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence*.

Zhang, M.; Yang, Y.; Chen, X.; Ji, Y.; Xu, X.; Li, J.; and Shen, H. T. 2021b. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12669–12678.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12870–12877.

Zhao, Y.; Zhao, Z.; Zhang, Z.; and Lin, Z. 2021. Cascaded prediction network via segment tree for temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4197–4206.