You Can Ground Earlier than See: An Effective and Efficient Pipeline for Temporal Sentence Grounding in Compressed Videos

Xiang Fang[1*]   Daizong Liu[2*]   Pan Zhou[1†]   Guoshun Nan[3]

[1]Hubei Key Laboratory of Distributed System Security, Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology
[2]Peking University   [3]Beijing University of Posts and Telecommunications

xfang9508@gmail.com   dzliu@stu.pku.edu.cn   panzhou@hust.edu.cn   nanguo2021@bupt.edu.cn

Abstract

Given an untrimmed video, temporal sentence grounding (TSG) aims to locate a target moment semantically according to a sentence query. Although previous respectable works have made decent success, they only focus on high-level visual features extracted from the consecutive decoded frames and fail to handle the compressed videos for query modelling, suffering from insufficient representation capability and significant computational complexity during training and testing. In this paper, we pose a new setting, compressed-domain TSG, which directly utilizes compressed videos rather than fully-decompressed frames as the visual input. To handle the raw video bit-stream input, we propose a novel Three-branch Compressed-domain Spatial-temporal Fusion (TCSF) framework, which extracts and aggregates three kinds of low-level visual features (I-frame, motion vector and residual features) for effective and efficient grounding. Particularly, instead of encoding the whole decoded frames like previous works, we capture the appearance representation by only learning the I-frame feature to reduce delay or latency. Besides, we explore the motion information not only by learning the motion vector feature, but also by exploring the relations of neighboring frames via the residual feature. In this way, a three-branch spatial-temporal attention layer with an adaptive motion-appearance fusion module is further designed to extract and aggregate both appearance and motion information for the final grounding. Experiments on three challenging datasets shows that our TCSF achieves better performance than other state-of-the-art methods with lower complexity.

1. Introduction

As a significant yet challenging computer vision task, temporal sentence grounding (TSG) has drawn increasing

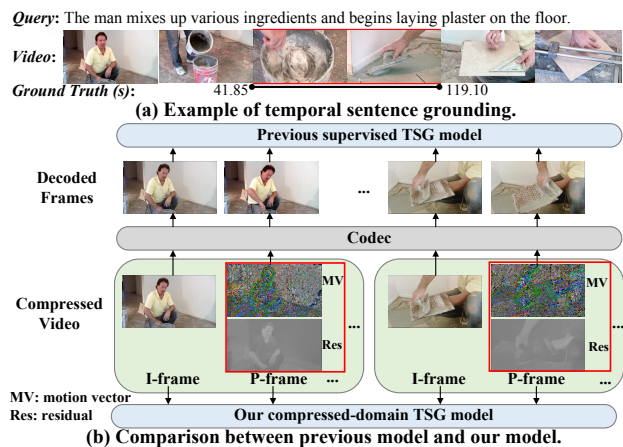---

*Equal contributions.   †Corresponding author.

Figure 1. (a) Example of the temporal sentence grounding (TSG). (b) Comparison between previous supervised TSG models and our compressed-domain TSG model. Previous models first decode the video into consecutive frames and then feed them into their networks, while our compressed-domain model directly leverages the compressed video as the visual input.

attention due to its various applications, such as video understanding [12–17, 25, 44, 76, 77, 79–81, 84] and temporal action localization [63, 71]. Given a long untrimmed video, the TSG task aims to locate the specific start and end timestamps of a video segment with an activity that semantically corresponds to a given sentence query. As shown in Figure 1(a), most of video contents are query-irrelevant, where only a short video segment matches the query. It is substantially more challenging since a well-designed method needs to not only model the complex multi-modal interaction among video and query, but also capture complicated context information for cross-modal semantics alignment.

By treating a video as a sequence of independent frames, most TSG methods [3, 17, 29, 34, 36–41, 43, 46, 65, 90, 93, 98] refer to the fully-supervised setting, where each frame is firstly fully decompressed from a video bit-stream and then manually annotated as query-relevant or query-irrelevant.

Despite the decent progress on the grounding performance, these data-hungry methods severely rely on the fully decompression and numerous annotations, which are significantly labor-intensive and time-consuming to obtain from real-word applications. To alleviate this dense reliance to a certain extent, some weakly-supervised works [8, 11, 16, 33, 35, 49, 51, 62, 64, 96, 97] are proposed to only leverage the coarse-grained video-query annotations instead of the fine-grained frame-query annotations. Unfortunately, this weak supervision still requires the fully-decompressed video for visual feature extraction.

Based on the above observation, in this paper, we make the first attempt to explore if an effective and efficient TSG model can be learned without the limitation of the fully decompressed video input. Considering that the real-world video always stored and transmitted in a compressed data format, we explore a more practical but challenging task: compressed-domain TSG, which directly leverages the compressed video instead of obtaining consecutive decoded frames as visual input for grounding. As shown in the Figure 1(b), a compressed video is generally parsed by a stream of Group of successive Pictures (GOPs) and each GOP starts with one intra-frame (I-frame) followed by a variable number of predictive frames (P-frames) [30, 74]. Specifically, the I-frame contains complete RGB information of a video frame, while each P-frame contains a motion vector and a residual. The motion vectors store 2D displacements between I-frame and its neighbor frames, and the residuals store the RGB differences between I-frame and its reconstructed frame calculated by Motion Vectors in the P-frames after motion compensation. The I-frame can be decoded itself, while these P-frames only store the changes from the previous I-frame by motion vectors and residuals.

Given the compressed video, our main challenge is how to effectively and efficiently extract contextual visual features from the above three low-level visual information for query alignment. Existing TSG works [3, 29, 34, 39, 65, 90, 93, 98] cannot be applied directly to the compressed video because their video features (*e.g.,* C3D and I3D) can only be extracted if all complete video frames are available after decompression. Moreover, decompressing all the frames will significantly increase computational complexity for feature extraction, leading to extra latency and extensive storage.

To address this challenging task, we propose the first and novel approach for compressed-domain TSG, called Three-branch Compressed-domain Spatial-temporal Fusion (TCSF). Given a group of successive picture (GOP) in a compressed video, we first extract the visual features from each I-frame to represent the appearance at its timestamp, and then extract the features of its P-frames to capture the motion information near the I-frame. In this way, we can model the activity content with above simple I-frame and P-frames instead of using their corresponding consecutive de-

coded frames. Specifically, we design a spatial attention and a temporal attention to integrate the appearance and motion features for activity modelling. To adaptively handle different fast-motion (P-frame guided) or slow-motion (I-frame guided) cases, we further design an adaptive appearance and motion fusion module to integrate the appearance and motion information by learning a balanced weight through a residual module. Finally, a query-guided multi-modal fusion is exploited to integrate the visual and textual features for final grounding.

Our contributions are summarized as follows:

- We propose a brand-new and challenging task: compressed-domain TSG, which aims to directly leverage the compressed video for TSG. To our best knowledge, we make the first attempt to locate the target segment in the compressed video.

- We present a novel pipeline for compressed-domain TSG, which can efficiently and effectively integrate both appearance and motion information from the low-level visual information in the compressed video.

- Extensive experiments on three challenging datasets (ActivityNet Captions, Charades-STA and TACoS) validate the effectiveness and efficiency of our TCSF.

2. Related Works

Temporal sentence grounding. Most existing TSG methods are under the fully-supervised setting, where all video-query pairs and precise segment boundaries are manually annotated based on the fully-decompressed video. These methods can be divided into two categories: 1) Proposal-based methods [1, 5, 45, 87, 94, 95]: They first pre-define multiple segment proposals and then align these proposals with the query for cross-modal semantic matching based on the similarity. Finally, the best proposal with the highest similarity score is selected as the predicted segment. Although achieving decent results, these proposal-based methods severely rely on the quality of the segment proposals and are time-consuming. 2) Proposal-free methods [7, 42, 53, 88, 92]: They directly regress the start and end boundary frames of the target segment or predict boundary probabilities frame-wisely. Compared with the proposal-based methods, proposal-free methods are more efficient. To alleviate the reliance to a certain extent, some state-of-the-art turn to the weakly-supervised setting [8, 11, 33, 49, 51, 62, 64, 96, 97], where only video-query pairs are annotated without precise segment boundaries in the fully-decompressed video.

In real-world computer vision tasks, we always collect the compressed video, rather than decompressed consecutive frames. In this paper, we present a brand-new practical yet challenging setting for TSG task, called compressed-domain TSL, with merely compressed video rather than a decompressed frame sequence.

Video compression. As a fundamental computer vision task, video compression [26, 27, 32, 48, 57, 72, 75] divides a video into a group of pictures (GOP), where each frame is coded as an I-, P-, and B- frame. An I-frame is the first frame of the GOP to maintain full RGB pixels as an anchor. The subsequent P-and B-frames are then coded using a block-based motion vector with temporal prediction. The prediction is conducted by searching the closest matching block of a previously coded frame as a reference frame. A vector of the current block to the reference block is determined as a motion vector. Since the current block and the matching block are often different, the transformed residual is used to denote the difference.

Compared with other deep features (*e.g.,* optical flow [24]) widely used in the TSG task, the compressed-domain features (MVs and residual) have the following advantages: 1) Lower computational costs. The compressed-domain features can be obtained during decoding, while other deep features need to decompress the compressed video and encode the video by a pretrained heavy-weight model (C3D [66] or I3D [4]). The compressed-domain features only even require partial-frame reconstruction by entropy decoding [100], inverse transform and quantization [28], and motion-compensation [10]. In entropy decoding, the most time-consuming process is skipping the motion-compensation [58], whose computational complexity is much smaller than that of other deep features. 2) No delay or dependency. The compressed-domain features can be instantly obtained. When we large-scale datasets, the advantages are more obvious.

3. Proposed Method

3.1. Overview

Problem statement. Given a video bit-stream $\mathcal{V}$ with T frames, the temporal sentence grounding (TSG) task aims to localize the precise boundary (τ_s, τ_e) of a specific segment semantically corresponding to a given query $\mathcal{Q} = \{q_j\}_{j=1}^M$, where q_j denotes the j-th word, M denotes the word number, τ_s and τ_e denote the start and end timestamps of the specific segment. In our compressed-domain TSG setting, we do not feed the decompressed frames video as input. Instead, we partially decode the video bit-stream at a low cost to extract the compressed video, which includes N group of pictures (GoPs). Each GoP G_i contains one reference I-frame $I_i \in \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times 3}$ followed by L number of P-frames $\{P_i^l\}_{l=1}^L$. Each P_i^l consists of a motion vector $M_i^l \in \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times 2}$ and a residual $R_i^l \in \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times 3}$, which can be extracted nearly cost-free from $\mathcal{V}$. For convenience, we assume that all GOPs contain the same number of P-frames. Thus, $T = N \times (L + 1)$. The video bit-stream can be represented as $\mathcal{V} = \{I_i, P_i^1, P_i^2, \cdots, P_i^L\}_{i=1}^N$, where i denotes the i-th GOP. Here, the I-frame contains complete RGB information of a video frame and can be decoded itself, while these P-frames only store the changes from the previous I-frame by motion vectors and residuals. The motion vectors store 2D displacements of the most similar patches between I-frame and the target frame, and the residuals store pixel-wise differences to correct motion compensation errors. We use above three low-level information contained in compressed videos as our visual input.

Pipeline. Our pipeline is summarized in Figure 2. Given a video bit-stream, we first utilize the entropy decoding approach [68, 73] to generate a group of successive pictures (GOP), which consists of several I-frames with their related P-frames. Then, we extract the visual appearance features from I-frames by a pre-trained ResNet-50 network, while a light-weight ResNet-18 network is used to extract the motion vector and residual features from P-frames. After that, we enrich these partial appearance and motion information with pseudo features to make the complete comprehension of the full video. A spatial-temporal attention module is further introduced to better model the activity content based on the motion-appearance contexts. Next, we design an adaptive appearance and motion fusion module to selectively integrate the attentive appearance and motion information guided by the residual information. Finally, we design a query-guided multi-modal fusion module to integrate the visual and textual features for final grounding.

3.2. Multi-Modal Encoding

Query encoder. Following [19], we first employ the Glove network [55] to embed each word into a dense vector. Then, a Bi-GRU network [9] and a multi-head self-attention module [67] are used to further integrate the sequential textual representations. Thus, final word-level features is denote as $Q = \{q_j\}_{j=1}^M \in \mathbb{R}^{M \times d}$, where d is the feature dimension. By concatenating the outputs of the last hidden unit in Bi-GRU with a further linear projection, we can obtain the sentence-level feature as $q_{global} \in \mathbb{R}^d$.

I-frame encoder. Following [31, 50], if the $\{t\}_{t=1^T}$-th frame is I-frame, we use a pretrained ResNet-50 model [22] to extract its appearance feature $a^t \in \mathbb{R}^{H \times W \times C}$, where H, W and C denotes dimensions of height, width, and channel.

P-frame encoder. Following [59, 73], if the $\{t\}_{t=1^T}$-th frame is P-frame containing a motion vector M^t and a residual R^t, we utilize a ResNet-18 network [22,78,79,82,83,85] to extract the motion vector feature $m^t \in \mathbb{R}^{H \times W \times C}$ and the residual feature $r^t \in \mathbb{R}^{H \times W \times C}$.

Pseudo feature generation. Since our compressed-domain TSG needs to locate the specific start and end frames of the target segment, we need to obtain the precise motion, compensation and appearance information of each frame for more accurate grounding. However, in the compressed video, we only have partially N-number I-frames of appearance and $(N \times L)$-number P-frames of motion and compensation, lacking enough full-frames (*i.e.*, T-number frames)
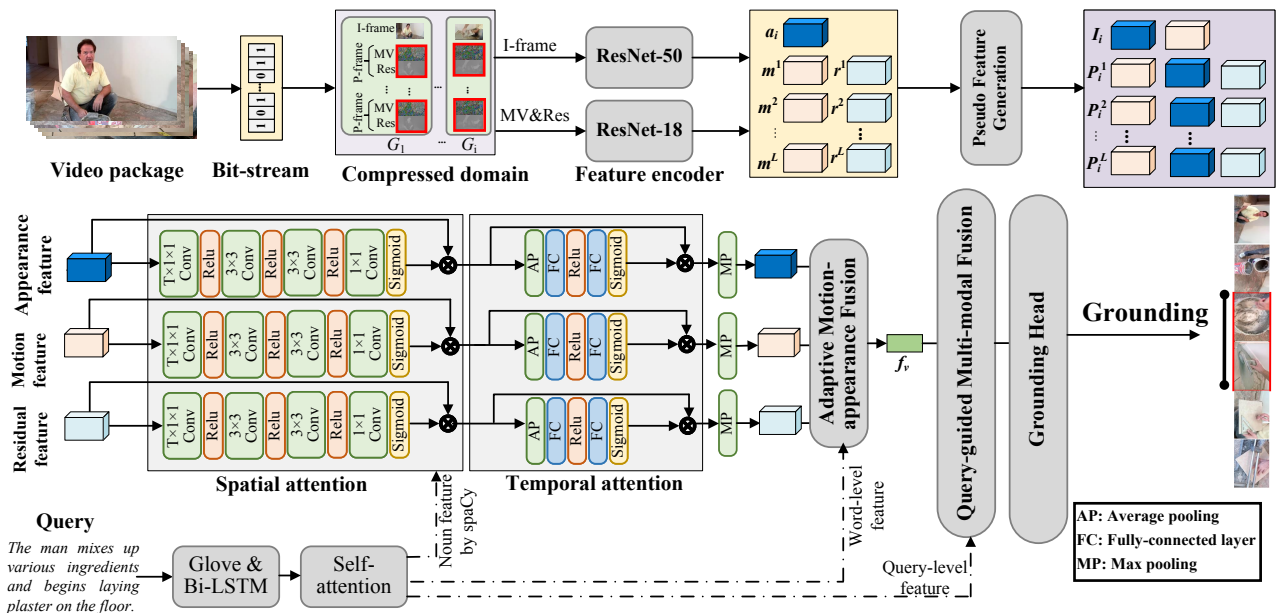
Figure 2. Overview of the proposed architecture. Firstly, we leverage the entropy decoding approach to obtain the compressed video, *i.e.*, I-frames and P-frames (containing motion vectors and residuals). Then, we enrich their information with pseudo features, and develop a three-branch spatial-temporal attention to model the query-related activity content. After that, we fuse the appearance and motion contexts, and integrate them with the query features for learning the joint multi-modal representations. At last, we feed the multi-modal features into the grounding head to predict the segment.

knowledge of the complete appearance-motion information. Thus, we tend to generate complementary pseudo features for the unseen frames of the video. For example, to warp the appearance feature from the current I-frame, we can use M^t to estimate the pseudo appearance feature a^{t+1} in its adjacent frame (its next frame). We can find that the pseudo feature generation approach exempts reconstructing each adjacent frame for feature extraction individually. We assume that the t-frame is I-frame. For constructing the pseudo appearance features of its n-th adjacent P-frame, we utilize a block-based motion estimation as:

$$a^{n+t}(s) = a^{n+t-1}(\delta M^{n+t-1}(s\delta) + s), \tag{1}$$

where a^{n+t} denotes the appearance feature of the $n + t$-th P-frame, s is a spatial coordinate of features, and δ is used as a scaling factor. By Eq. (1), we can obtain the appearance information of each P-frame based on off-the-shelf I-frames.

Similarly, we will generate the motion information of each I-frame based on P-frames. Following [18], we combine the temporal movement information of appearance features in these adjacent frames. In the channel axis, we concatenate consecutive n frames $[a^t; \cdots ; a^{n+t}]$ as $V^t \in \mathbb{R}^{H \times W \times C \times n}$. Setting $V_*^t = conv_{1\times1}(V^t)$, we can get

$$m^t = ReLU(V_*^t), \tag{2}$$

where m^t is the motion feature of t-th frame, ReLU is the ReLU function, and $conv_{1\times1}$ means 1×1 convolution layer with stride 1, producing a channel dimension of feature $C \times$

n to C. Thus, for the t-th frame, its appearance and motion features are a^t and m^t, respectively.

3.3. Three-branch Spatial-temporal Attention

In the TSG task, most of regions within a frame are query-irrelevant, where only a few regions are query-relevant. To automatically learn the discriminative regions relevant to the query, we need to obtain the fine-grained local spatial context. Besides, the temporal context is also important since we can correlate the region-attentive spatial information in time series for precisely modelling the activity. Therefore, we exploit previous-encoded three low-level features (appearance, motion and residual features) to obtain such query-relevant temporal-spatial information by designing a three-branch temporal and spatial attention.

Spatial attention. We propose the spatial attention to guide the model put more focus on the query-related region of the low-level features. Specifically, in the TSG task, most spatial visual information is noun-relevant. We first utilize the NLP tool spaCy [23] to parse nouns from the given query. Then, we exploit these nouns to enhance three visual features (appearance, motion and residual features) via an attention mechanism for helping the model learn to pay more attention on the spatial information precisely. The details of spatial attention is shown in Figure 2, where we leverage the combination of two 2D convolutional layers with kernel size of 3×3, two RelUs and a 2D convolutional layers with kernel size of 1×1 to obtain the spatial attention
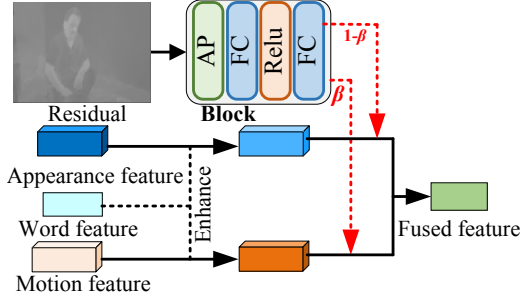
Figure 3. Framework of adaptive motion-appearance fusion.

map. Therefore, we can enhance the region-attentive appearance features a^t into a_*^t. Similarly, we can also obtain the region-attentive motion feature m_*^t and region-attentive residual features r_*^t.

Temporal attention. After learning the region-aware spatial information, we further learn to capture their temporal relation to better model the query-relevant activity. Specifically, we choose K consecutive frames (starting at the t-th frame) for extracting their temporal information via a newly proposed temporal attention. Here, we take the temporal attention on appearance features for example. For the appearance features, we first concatenate them as $\mathcal{A} = [a_*^t; \cdots; a'^{t+K-1}]$. To yield the temporal weights $w = [w^1, w^2, \cdots, w^K] \in \mathbb{R}^K$ on these consecutive frames, we first leverage a global average pooling along three dimensions $H \times W \times C$ to generate a temporal-wise statistics $S = [s^1, s^2, \cdots, s^K] \in \mathbb{R}^K$, where s^t represents the whole temporal information of w^t. Then, we utilize the temporal attention module shown in Figure 2 to generate the temporal weights w^t as:

$$w^t = \sigma(W_{F1} \circ ReLU(W_{F2} \circ S + b_2) + b_1), \qquad (3)$$

where W_{FC_1} and W_{FC_2} are the weights of two FC layers; $b_1 \in \mathbb{R}^K$ and $b_2 \in \mathbb{R}^K$ are the biases of two FC layers; $\circ$ denotes the convolution operation. Therefore, the final output of the appearance branch is:

$$f_{\hat{a}}^t = w_c a_*^t. \qquad (4)$$

Similarly, we can obtain the final outputs of the MV and residual branches as: $f_{\hat{m}}^t$ and $f_{\hat{r}}^t$.

3.4. Adaptive Motion-Appearance Fusion

After obtaining the attentive motion and appearance information, we tend to aggregate them to infer the activity content. Considering different videos may contain different abrupt temporal changes, we cannot equally fuse both the motion and appearance. Specifically, in the TSG task, a video with more abrupt temporal changes often corresponds to a related word. For example, a video corresponding to "run" often have more temporal changes than another video corresponding to "walk". Therefore, we propose an adaptive strategy to fuse motion and appearance reasonably. Specifically, we first enhance the appearance and motion

features based on the query features. Then, we leverage the residual information to balance the enhanced appearance features and the enhanced motion features adaptively.

Query-guided feature enhancement. We first utilize an attention mechanism to aggregate the word-level query features $\{q_j\}_{j=1}^M$ for each appearance feature $f_{\hat{a}}^t$ as:

$$X^t = W^\top \tanh(\boldsymbol{W}_1 f_{\hat{a}}^t + \boldsymbol{W}_2 \boldsymbol{q}_j + \boldsymbol{b}_0), \qquad (5)$$

$$A_a^t = \frac{X^t}{\sum_s X^t(s)}, \qquad (6)$$

where A_a^t is the attention, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are projection matrices, $\boldsymbol{b}_0$ is the bias vector, and the $W^\top$ is the row vector as in [95]. Based on Eq. (5), we can obtain the query-enhanced appearance feature $f_a^t = f_{\hat{a}}^t \odot A_a^t$, where $\odot$ denotes the operation of element-wise product. Similarly, we can obtain the query-enhanced motion feature f_m^t, which also semantically corresponds to the query.

Residual-guided feature fusion. The residual features not only represent the temporal changes (*i.e.*, motion context) among adjacent frames, but also denote the changes occur in RGB pixels (*i.e.*, appearance context). Thus, we utilize the residual feature as guidance to synchronize motion and appearance features by a learnable Block (shown Figure 3):

$$\beta^t = Block(f_{\hat{r}}^t), \qquad (7)$$

where $\beta^t \in [0, 1]$ is a learnable balance, the block contains an average pooling, two fully-connected layers and a RelU network. If there are many abrupt temporal changes between different scenes, β^t will approach 1. On the contrary, when there are few abrupt temporal changes, β^t goes nearly to 0. At last, we fuse the motion and appearance information with this balanced weight as:

$$f_v^t = \beta W_3 f_a^t + (1 - \beta) W_4 f_m^t, \qquad (8)$$

where matrices W_3 and W_4 are learnable parameters.

3.5. Multi-modal Fusion and Grounding Head

After obtaining the motion-appearance enhanced visual feature, we further integrate it with the textual features as:

$$o = W_5 \sum_{t=1}^T f_v^t + W_6 \sum_{j=1}^M q_j + W_7 q_{global}. \qquad (9)$$

where o is the fused feature, and W_5, W_6 and W_7 are learnable weight matrices.

Based on the multi-modal features o, we utilize two Multilayer Perceptron (MLP) layers to predict the start and end scores (p_s^t, p_e^t) on each video clip as

$$\gamma = \text{softmax}(\text{MLP}_1(o)), \; (p_s^t, p_e^t) = \text{MLP}_{\text{reg}}(\gamma^t o^t), \qquad (10)$$

where $\gamma \in \mathbb{R}^T$ is the attention weights for segments. Following [40, 92], we introduce the regression loss $\mathcal{L}_{\text{reg}}$ to learn the timestamp scoring prediction as follows:

$$\mathcal{L}_{\text{reg}} = \frac{1}{2T} \sum_{t=1}^T [\mathcal{S}(\hat{p}_s^t - p_s^t) + \mathcal{S}(\hat{p}_e^t - p_e^t)], \qquad (11)$$

Table 1. Effectiveness comparison for temporal sentence grounding on ActivityNet Captions dataset under official train/test splits.

Method	Type	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
CTRL [19]	FS	-	29.01	-	59.17
2D-TAN [94]	FS	59.45	44.51	85.53	77.13
DRN [89]	FS	-	45.45	-	77.97
RaNet [20]	FS	-	45.59	-	75.93
MIGCN [93]	FS	-	48.02	-	78.02
MMN [71]	FS	65.05	48.59	87.25	79.50
ICVC [6]	WS	46.62	29.52	80.92	66.61
LCNet [86]	WS	48.49	26.33	82.51	62.66
VCA [70]	WS	50.45	31.00	71.79	53.83
WSTAN [69]	WS	52.45	30.01	79.38	63.42
CNM [99]	WS	55.68	33.33	-	-
Our TCSF	**CD**	**66.87**	**48.38**	**88.75**	**80.24**

where $(\hat{p}_s^t, \hat{p}_e^t) \in [0, 1]$ are the ground-truth labels, $\mathcal{S}(x)$ is the cross-entropy loss. We also introduce a confident loss $\mathcal{L}_{\text{guide}}$ to guide timestamp prediction:

$$\mathcal{L}_{\text{guide}} = -\frac{\sum_{t=1}^{T} \hat{\gamma}^t \log(\gamma^t)}{\sum_{t=1}^{T} \hat{\gamma}^t}, \qquad (12)$$

where $\hat{\gamma}^t = 1$ if the t-th segment is located within the ground-truth boundary and $\hat{\gamma}^t = 0$ otherwise. By Eq. (12), we can obtain higher attention weights for the segments semantically relevant to the text query.

Therefore, the final loss function is formulated as:

$$\mathcal{L}_{final} = \mathcal{L}_{\text{reg}} + \alpha \mathcal{L}_{\text{guide}}, \qquad (13)$$

where α is a hyper-parameter.

Inference. Given a video bit-stream and a language query, we first feed them into our TCSF to obtain the fused cross-modal feature o in Eq. (9). Then, we predict the start and end boundary scores (p_s^t, p_e^t) by o in Eq. (10) and the confidence score in Eq. (12). Based on the predicted scores of the start/end timestamps and confidence scores, we generate several candidate moments, "Top-n (R@n)" candidates will be selected with non-maximum suppression.

4. Experiment

4.1. Datasets

ActivityNet Captions. Built from ActivityNet v1.3 dataset [2] for the dense video captioning task, ActivityNet Captions contains 20k YouTube videos and 100k language queries. On average, a video are 2 minutes and a query has about 13.5 words. Following the public split [19], we use 37421, 17505, and 17031 video-query pairs for training, validation and testing.

Charades-STA. Built upon the Charades dataset [19, 60], Charades-STA contains 16128 video-sentence pairs. Folowing [19], we utilize 12408 pairs for training and the others for testing. The average video length is 0.5 minutes.

Table 2. Performance comparison for temporal sentence grounding on Charades-STA dataset under official train/test splits.

Method	Type	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
CTRL [19]	FS	23.62	8.89	58.92	29.52
MMN [71]	FS	47.31	27.28	83.74	58.41
2D-TAN [94]	FS	39.81	23.25	79.33	52.15
RaNet [20]	FS	43.87	26.83	86.67	54.22
DRN [89]	FS	45.40	26.40	88.01	55.38
WSTAN [69]	WS	29.35	12.28	76.13	41.53
ICVC [6]	WS	31.02	16.53	77.53	41.91
CNM [99]	WS	35.15	14.95	-	-
VCA [70]	WS	38.13	19.57	78.75	37.75
LCNet [86]	WS	39.19	18.17	80.56	45.24
Our TCSF	**CD**	**53.85**	**37.20**	**90.86**	**58.95**

Table 3. Performance comparison for temporal sentence grounding on TACoS dataset under official train/test splits.

Method	Type	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
CTRL [19]	FS	18.32	13.30	36.69	25.42
ACRN [47]	FS	19.52	14.62	34.97	24.88
CMIN [95]	FS	24.64	18.05	38.46	27.02
SCDM [87]	FS	26.11	21.17	40.16	32.18
DRN [89]	FS	-	23.17	-	33.36
2D-TAN [94]	FS	37.29	25.32	57.81	45.04
MMN [71]	FS	39.24	26.17	62.03	47.39
FVMR [21]	FS	41.48	29.12	64.53	50.00
RaNet [20]	FS	43.34	33.54	67.33	55.09
MIGCN [93]	FS	48.79	37.57	67.63	57.91
Our TCSF	**CD**	**49.82**	**38.53**	**68.60**	**59.89**

The language annotations are generated by sentence decomposition and keyword matching with manual check.

TACoS. Collected from the cooking scene by [56], TACoS is employed for the video grounding and dense video captioning tasks. The dataset consists of 127 videos, whose average length is 4.8 minutes. Following the same split of [19], we leverage 10146, 4589, and 4083 video-query pairs for training, validation, and testing respectively.

4.2. Experimental Settings

Evaluation metric. Following [19, 47, 92], we evaluate the grounding performance by "R@n, IoU=m", which means the percentage of queries having at least one result whose Intersection over Union (IoU) with ground truth is larger than m. In our experiments, we use $n \in \{1, 5\}$ for all datasets, $m \in \{0.5, 0.7\}$ for ActivityNet Captions and Charades-STA, $m \in \{0.3, 0.5\}$ for TACoS.

Implementation details. All the experiments are implemented by PyTorch with an NVIDIA Quadro RTX 6000. For entropy decoding, following [68, 73], we use an MPEG-4 decoder [61] to decompress video bit-stream for obtaining I-frame and P-frame. As for query encoding, we embed each word to 300-dimension features by the Glove

Table 4. Time complexity (s) of 100 videos on ActivityNet Captions dataset. The total time T_{total} comprises the measurement time of decompressing video frames (T_{dec}), extracting the corresponding features (T_{ext}), and executing the network models (T_{exe}), where "Other" means the feature encoder (*e.g.,* C3D/I3D).

Model	T_{dec}	T_{ext}				T_{exe}	T_{total}
		I-frame	MV	Residual	Other		
CTRL [19]	50.72	-	-	-	30.36	372.74	453.82
RaNet [20]	50.72	-	-	-	30.36	406.30	487.38
2D-TAN [94]	50.72	-	-	-	30.36	434.91	515.99
MIGCN [93]	50.72	-	-	-	30.36	529.27	610.35
MMN [71]	50.72	-	-	-	30.36	556.43	637.51
DRN [89]	50.72	-	-	-	30.36	585.72	666.80
TAG [52]	50.72	-	-	-	30.36	162.28	243.36
WSTAN [69]	50.72	-	-	-	30.36	183.86	264.94
CNM [99]	50.72	-	-	-	43.86	175.37	269.95
Our TCSF	**12.67**	**1.84**	**0.61**	**0.28**	-	**30.76**	**46.16**

model [55]. Besides, we set the head size of multi-head self-attention to 8, and the hidden dimension of Bi-GRU to 512, respectively. During training, we optimize parameter by Adam optimizer with learning rate 4×10^{-4} and linear learning rate decay of 10 for each 40 epochs. The batch size is 16 and the maximum training epoch is 100. We set $\alpha = 0.8$ and $K = 7$ in this paper.

4.3. Comparison with State-of-the-Arts

We conduct performance comparison on three datasets. To evaluate efficiency, we only choose the open-source compared methods that are grouped into two categories: (i) Fully-supervised (FS) setting [19–21,47,71,87,89,93–95]; (ii) Weakly-supervised (WS) setting [6,69,70,86,97,99]. For convenience, we denote "compressed-domain setting" as "CD". Following [54,91], we directly cite the results of compared methods from corresponding works. Note that no weakly-supervised method reports its results on TACoS. The best results are **bold**. From Tables 1, 2 and 3, we can find that our TCSF outperforms all compared methods by a large margin. It demonstrates that our model can achieve effective performance in more challenging compressed-domain setting.

Efficiency comparison. To fairly evaluate the efficiency of our TCSF, we conduct comparison on ActivityNet Captions dataset with some state-of-the-art methods whose source codes are available. Table 4 reports the results, and we consider the decompressing time T_{dec}, the feature extracting time T_{ext}, the network executing time T_{exe}, where the time is measured via an average on the whole videos. As depicted in Table 4, we have the following observations: (i) Our model takes 12.67s to decompress GOPs in each video bit-streams and 1.84s, 0.61s, 0.28s to extract their three features, which is much efficient than previous works. The main reason is that previous works need to decompress full frames of the video and rely on the heavy-weight 3D encoder like C3D/I3D to extract the features. Instead, we need

Table 5. Main ablation study on ActivityNet Captions dataset, where we remove each key individual component to investigate its effectiveness. "PFG" denotes "pseudo feature generation", "TTA" denotes "Three-branch spatial-temporal attention", "AMF" denotes "adaptive motion-appearance fusion".

PFG	TTA	AMF	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
✗	✗	✗	50.59	32.84	76.12	68.33
✓	✗	✗	60.25	41.82	79.10	72.08
✗	✓	✗	62.79	45.87	79.45	76.13
✗	✗	✓	63.74	45.39	80.16	76.05
✓	✓	✗	64.19	47.56	83.77	76.90
✓	✓	✓	**66.87**	**48.38**	**88.75**	**80.24**

Table 6. Ablation study on pseudo feature generation.

Appearance feature	Motion feature	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
✗	✓	64.73	47.51	88.09	78.10
✓	✗	65.85	48.02	87.80	79.03
✓	✓	**66.87**	**48.38**	**88.75**	**80.24**

Table 7. Ablation study on three-branch spatial-temporal attention.

Spatial attention	Temporal attention	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
✗	✓	64.56	43.82	84.13	77.50
✓	✗	65.31	43.20	83.72	76.81
✓	✓	**66.87**	**48.38**	**88.75**	**80.24**

Table 8. Ablation study on adaptive motion-appearance fusion.

Query-guided enhancement	Residual-guided fusion	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
✗	✓	65.94	47.46	86.52	78.88
✓	✗	65.80	47.92	87.63	79.15
✓	✓	**66.87**	**48.38**	**88.75**	**80.24**

less frame-level context with much light-weight encoder. (ii) Our network executing T_{exe} also has less parameters to learn than previous work, thus achieving faster speed. Overall, experimental results demonstrate the time-efficiency of our method.

4.4. Ablation study

To validate the effectiveness of each component in our TCSF, we conduct extensive ablation studies on the most challenging ActivityNet Captions dataset.

Main ablation studies. To analyze how each component contributes to the challenging task, we perform main ablation study as shown in Table 5. Firstly, we set a baseline model that does not utilize pseudo feature, three-branch spatial-temporal attention module and adaptive motion-appearance fusion strategy to address the compressed-domain TSG. Similar to previous supervised methods, the baseline model directly generates multiple coarse segment proposals and then utilizes the rank loss for training. We
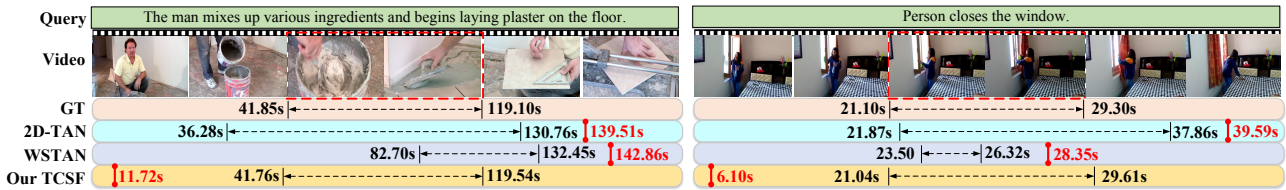
Figure 4. Qualitative prediction examples, where we complete the prediction at the red time. We find that our TCSF can ground earlier than the ground-truth start timestamp, while other methods ground later than the end timestamp.

Table 9. Effect of different low-level features.

I-frame	MV	Residual	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
✔	✗	✗	64.27	46.83	85.75	76.54
✔	✗	✔	65.66	47.82	86.36	78.59
✔	✔	✗	66.03	47.94	88.03	79.28
✔	✔	✔	**66.87**	**48.38**	**88.75**	**80.24**

Table 10. Effect of the nouns-formed query in spatial attention.

Changes	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
w/o query	64.26	47.82	88.17	77.54
w/ query	**66.87**	**48.38**	**88.75**	**80.24**

Table 11. Effect of different hyper-parameters.

Module	Changes	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
Temporal attention	$K = 6$	66.28	47.95	87.34	80.17
	$K = 7$	**66.87**	48.38	**88.75**	**80.24**
	$K = 8$	65.92	**48.53**	86.11	79.30
Grounding head	$\alpha = 0.7$	66.02	46.98	87.94	**80.31**
	$\alpha = 0.8$	**66.87**	**48.38**	**88.75**	80.24
	$\alpha = 0.9$	65.93	47.06	87.29	79.23

can find that this baseline performs worse than most state-of-the-art methods in Table 1. Secondly, by designing the pseudo feature generation (PFG) module, we can effectively improve the performance since it enriches the full-frame context of the video. Table 6 further analyzes the effective of both pseudo appearance and motion features. Thirdly, applying three-branch spatial-temporal attention (TTA) module also brings the large improvement since our well-designed spatial-temporal attention extracts the more fine-grained region-attentive temporal-spatial information for modelling more accurate activity content. As shown in Table 7, we further illustrate the effectiveness of spatial and temporal attentions separately. Besides, the adaptive motion-appearance fusion (AMF) strategy also boost the performance a lot because it can balance the importance between appearance and motion features. Table 8 illustrates the contributions of the query-guided feature enhancement and residual-guided fusion in AMF module. Overall, each component brings the performance improvement, and the full TCSF achieves the best results.

Effect of different low-level features. To analyze the contribution of different low-level features, we conduct the ablation study as shown in Table 9. Both MV and residual can significantly improve the performance. The improvement shows the effectiveness of MV and residual.

Effect of the nouns-formed query. In our spatial attention module, we utilize the noun feature to help us extract the spatial information. As shown in Table 10, we analyze the effect of the specific nouns-formed query. Based on the query, our TCSF improves the performance by 2.61% in

"R@1, IoU=0.3". This is because the nouns-formed query can locate the specific region for each frame, which reduces the distraction of background information in the video.

Analysis on the hyper-parameters. Moreover, we investigate the robustness of the proposed model to different hyper-parameters in Table 11. In the temporal attention module, we choose consecutive K frame to extract the temporal information. We find we can obtain the best performance when $K = 7$. In the grounding head module, we leverage α to balance the two losses. When $\alpha = 0.8$, our TCSF obtains the best performance.

4.5. Qualitative Results

As shown in Figure 4, we report the representative visualization of the grounding performance. Our TCSF can ground more accurate query-related segment boundaries than 2D-TAN and WSTAN with faster grounding.

5. Conclusion

In this paper, we introduce a brand-new compressed-domain setting into the temporal sentence grounding task to directly utilize the compressed video rather than decompressed frames. To handle the challenging setting, we propose a novel Three-branch Compressed-domain Spatial-temporal Fusion (TCSF) framework to extract and aggregate three kinds of low-level visual features for grounding. Experimental results on three challenging datasets (ActivityNet Captions, Charades-STA and TACoS) demonstrate that our TCSF significantly outperforms existing fully- and weakly-supervised methods.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 2

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 6

[3] Da Cao, Yawen Zeng, Meng Liu, Xiangnan He, Meng Wang, and Zheng Qin. Strong: Spatio-temporal reinforcement learning for cross-modal video moment localization. In *ACM MM*, pages 4162–4170, 2020. 1, 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 3

[5] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, pages 162–171, 2018. 2

[6] Jiaming Chen, Weixin Luo, Wei Zhang, and Lin Ma. Explore inter-contrast between videos via composition for weakly supervised temporal sentence grounding. *AAAI*, 2022. 6, 7

[7] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *AAAI*, 2020. 2

[8] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *ACL*, 2019. 2

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS*, 2014. 3

[10] Ajay Divakaran, Anthony Vetro, Kohtaro Asai, and Hirofumi Nishikawa. Video browsing system based on compressed domain feature extraction. *IEEE TCE*, 46(3):637–644, 2000. 3

[11] Xuguang Duan, Wen-bing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *NeurIPS*, 2018. 2

[12] Xiang Fang and Yuchong Hu. Double self-weighted multi-view clustering via adaptive view fusion. *arXiv preprint arXiv:2011.10396*, 2020. 1

[13] Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Wu. Animc: A soft approach for autoweighted noisy and incomplete multiview clustering. *IEEE TAI*, 3(2):192–206, 2021. 1

[14] Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. V 3 h: View variation and view heredity for incomplete multiview clustering. *IEEE TAI*, 1(3):233–247, 2020. 1

[15] Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. *IEEE TETCI*, 2021. 1

[16] Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE Transactions on Multimedia*, 2022. 1, 2

[17] Xiang Fang, Daizong Liu, Pan Zhou, Zichuan Xu, and Ruixuan Li. Hierarchical local-global transformer for temporal sentence grounding. *arXiv preprint arXiv:2208.14882*, 2022. 1

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *CVPR*, pages 6202–6211, 2019. 4

[19] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 3, 6, 7

[20] Jialin Gao, Xin Sun, Mengmeng Xu, Xi Zhou, and Bernard Ghanem. Relation-aware video reading comprehension for temporal language grounding. In *EMNLP*, pages 3978–3988, 2021. 6, 7

[21] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *ICCV*, pages 1523–1532, 2021. 6, 7

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[23] M Honnibal and I Montani. Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *Unpublished software application. https://spacy. io*, 2017. 4

[24] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 2462–2470, 2017. 3

[25] Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. Semi-supervised video paragraph grounding with contrastive encoder. In *CVPR*, pages 2466–2475, 2022. 1

[26] Daniel Lauzon, André Vincent, and Limin Wang. Performance evaluation of mpeg-2 video coding for

hdtv. *IEEE transactions on Broadcasting*, 42(2):88–94, 1996. 3

[27] Jeehong Lee, IlHong Shin, and HyunWook Park. Adaptive intra-frame assignment and bit-rate estimation for variable gop length in h. 264. *IEEE TCSVT*, 16(10):1271–1279, 2006. 3

[28] Se-Ho Lee, Je-Won Kang, and Chang-Su Kim. Compressed domain video saliency detection using global and local spatiotemporal features. *JVCIR*, 35:169–183, 2016. 3

[29] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 1, 2

[30] Congcong Li, Xinyao Wang, Longyin Wen, Dexiang Hong, Tiejian Luo, and Libo Zhang. End-to-end compressed video representation learning for generic event boundary detection. In *CVPR*, pages 13967–13976, 2022. 2

[31] Mu Li, Kede Ma, Jane You, David Zhang, and Wangmeng Zuo. Efficient and effective context-based convolutional entropy modeling for image compression. *IEEE TIP*, 29:5900–5911, 2020. 3

[32] Ting-Lan Lin, Sandeep Kanumuri, Yuan Zhi, David Poole, Pamela C Cosman, and Amy R Reibman. A versatile model for packet loss visibility and its application to packet prioritization. *IEEE TIP*, 19(3):722–735, 2009. 3

[33] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, 2020. 2

[34] Daizong Liu, Xiang Fang, Wei Hu, and Pan Zhou. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *arXiv preprint arXiv:2203.02966*, 2022. 1, 2

[35] Daizong Liu, Xiang Fang, Pan Zhou, Xing Di, Weining Lu, and Yu Cheng. Hypotheses tree building for one-shot temporal sentence localization. *arXiv preprint arXiv:2301.01871*, 2023. 2

[36] Daizong Liu and Wei Hu. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4536–4545, 2022. 1

[37] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *AAAI*, 2022. 1

[38] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for

temporal sentence localization in videos. In *EMNLP*, pages 9292–9301, 2021. 1

[39] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*, pages 11235–11244, June 2021. 1, 2

[40] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *ACM MM*, pages 4092–4101, 2022. 1, 5

[41] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *ACM MM*, pages 4070–4078, 2020. 1

[42] Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In *AAAI*, 2022. 2

[43] Daizong Liu, Xiaoye Qu, and Pan Zhou. Progressively guide to attend: An iterative alignment framework for temporal sentence grounding. In *EMNLP*, pages 9302–9311, 2021. 1

[44] Daizong Liu, Shuangjie Xu, Xiao-Yang Liu, Zichuan Xu, Wei Wei, and Pan Zhou. Spatiotemporal graph neural network based mask reconstruction for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2100–2108, 2021. 1

[45] Daizong Liu and Pan Zhou. Jointly visual-and semantic-aware graph memory networks for temporal sentence localization in videos. *arXiv preprint arXiv:2303.01046*, 2023. 2

[46] Daizong Liu, Pan Zhou, Zichuan Xu, Haozhao Wang, and Ruixuan Li. Few-shot temporal sentence grounding via memory-guided semantic learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1

[47] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *SIGIR*, pages 15–24, 2018. 6, 7

[48] Guo Lu, Tianxiong Zhong, Jing Geng, Qiang Hu, and Dong Xu. Learning based multi-modality image and video compression. In *CVPR*, pages 6083–6092, 2022. 3

[49] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D. Yoo. Vlanet:

Video-language alignment network for weakly-supervised video moment retrieval. In *ECCV*, pages 156–171, 2020. 2

[50] Detlev Marpe, Heiko Schwarz, Sebastian Bosse, Benjamin Bross, Philipp Helle, Tobias Hinz, Heiner Kirchhoffer, Haricharan Lakshman, Tung Nguyen, Simon Oudin, et al. Video compression using nested quadtree structures, leaf merging, and improved techniques for motion representation and entropy coding. *IEEE TCSVT*, 20(12):1676–1687, 2010. 3

[51] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, 2019. 2

[52] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, pages 11592–11601, 2019. 7

[53] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020. 2

[54] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, pages 122–132, 2020. 7

[55] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 3, 7

[56] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *TACL*, 1:25–36, 2013. 6

[57] Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. Overview of the scalable video coding extension of the h. 264/avc standard. *IEEE TCSVT*, 17(9):1103–1120, 2007. 3

[58] Bo Shen. Submacroblock motion compensation for fast down-scale transcoding of compressed video. *IEEE TCSVT*, 15(10):1291–1302, 2005. 3

[59] Zheng Shou, Xudong Lin, Yannis Kalantidis, Laura Sevilla-Lara, Marcus Rohrbach, Shih-Fu Chang, and Zhicheng Yan. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. In *CVPR*, 2019. 3

[60] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016. 6

[61] Thomas Sikora. The mpeg-4 video standard verification model. *IEEE TCSVT*, 7(1):19–31, 1997. 6

[62] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020. 2

[63] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1533–1542, 2021. 1

[64] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A. Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *WACV*, 2021. 2

[65] Haoyu Tang, Jihua Zhu, Meng Liu, Zan Gao, and Zhiyong Cheng. Frame-wise cross-modal matching for video moment retrieval. *IEEE TMM*, 24:1338–1349, 2021. 1, 2

[66] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 3

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 3

[68] Shiyao Wang, Hongchao Lu, and Zhidong Deng. Fast object detection in compressed video. In *ICCV*, pages 7104–7113, 2019. 3, 6

[69] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. Weakly supervised temporal adjacent network for language grounding. *IEEE TMM*, 2021. 6, 7

[70] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *ACM MM*, pages 1459–1468, 2021. 6, 7

[71] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, volume 36, pages 2613–2623, 2022. 1, 6, 7

[72] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *IEEE TCSVT*, 13(7):560–576, 2003. 3

[73] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. Compressed video action recognition. In *CVPR*, pages 6026–6035, 2018. 3, 6

[74] Kai Xu and Angela Yao. Accelerating video object segmentation with compressed video. In *CVPR*, pages 1342–1351, 2022. 2

[75] Qiang Xu, Xinghao Jiang, Tanfeng Sun, and Alex C Kot. Detection of hevc double compression with non-aligned gop structures via inter-frame quality degradation analysis. *Neurocomputing*, 452:99–113, 2021. 3

[76] Shuangjie Xu, Daizong Liu, Linchao Bao, Wei Liu, and Pan Zhou. Mhp-vos: Multiple hypotheses propagation for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 314–323, 2019. 1

[77] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, pages 16442–16453, 2022. 1

[78] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *ECCV*, pages 178–196. Springer, 2022. 3

[79] Jingkang Yang, Weirong Chen, Litong Feng, Xiaopeng Yan, Huabin Zheng, and Wayne Zhang. Webly supervised image classification with metadata: Automatic noisy label correction via visual-semantic graph. In *ACM MM*, pages 83–91, 2020. 1, 3

[80] Jingkang Yang and Santiago Segarra. Enhancing geometric deep learning via graph filter deconvolution. In *GlobalSIP*, pages 758–762. IEEE, 2018. 1

[81] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *ICCV*, pages 8301–8309, 2021. 1

[82] Jingkang Yang, Haohan Wang, Jun Zhu, and Eric P Xing. Sedmid for confusion detection: uncovering mind state from time series brain wave data. *arXiv preprint arXiv:1611.10252*, 2016. 3

[83] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 3

[84] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1

[85] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *arXiv preprint arXiv:2204.05306*, 2022. 3

[86] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE TIP*, 30:3252–3262, 2021. 6, 7

[87] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *NeurIPS*, 2019. 2, 6, 7

[88] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019. 2

[89] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, pages 10287–10296, 2020. 6, 7

[90] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. Multi-modal relational graph for cross-modal video moment retrieval. In *CVPR*, pages 2215–2224, 2021. 1, 2

[91] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering framework. *IEEE TPAMI*, 2021. 7

[92] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, pages 6543–6554, 2020. 2, 5, 6

[93] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *CVPR*, pages 12669–12678, 2021. 1, 2, 6, 7

[94] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020. 2, 6, 7

[95] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*, pages 655–664, 2019. 2, 5, 6, 7

[96] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACM MM*, pages 4098–4106, 2020. 2

[97] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *NeurIPS*, 2020. 2, 7

[98] Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. Cascaded prediction network via segment tree for

temporal video grounding. In *CVPR*, pages 4197–4206, 2021. 1, 2

[99] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *CVPR*, pages 15555–15564, 2022. 6, 7

[100] Jia Zhu and Ye Wang. Pop music beat detection in the huffman coded domain. In *ICME*, pages 60–63, 2007. 3