Multi-Pair Temporal Sentence Grounding via Multi-Thread Knowledge Transfer Network

Xiang Fang[1,2], Wanlong Fang[3], Changshuo Wang[3*], Daizong Liu[4], Keke Tang[5], Jianfeng Dong[6], Pan Zhou[1*], Beibei Li[2*]

[1]Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology
[2]Key Laboratory of Data Protection and Intelligent Management (Sichuan University), Ministry of Education
[3]Nanyang Technological University, Singapore
[4]Peking University
[5]Guangzhou University
[6]Zhejiang Gongshang University
xfang9508@gmail.com, wanlongfang@gmail.com, wangchangshuo1@gmail.com, dzliu@stu.pku.edu.cn, tangbohutbh@gmail.com, dongjf24@gmail.com, panzhou@hust.edu.cn, libeibei@scu.edu.cn

Abstract

Given some video-query pairs with untrimmed videos and sentence queries, temporal sentence grounding (TSG) aims to locate query-relevant segments in these videos. Although previous respectable TSG methods have achieved remarkable success, they train each video-query pair separately and ignore the relationship between different pairs. To this end, in this paper, we pose a brand-new setting: Multi-Pair TSG, which aims to co-train these pairs. We propose a novel video-query co-training approach, Multi-Thread Knowledge Transfer Network, to locate a variety of video-query pairs effectively and efficiently. Firstly, we mine the spatial and temporal semantics across different queries to cooperate with each other. To learn intra- and inter-modal representations simultaneously, we design a cross-modal contrast module to explore the semantic consistency by a self-supervised strategy. To fully align visual and textual representations between different pairs, we design a prototype alignment strategy to 1) match object prototypes and phrase prototypes for spatial alignment, and 2) align activity prototypes and sentence prototypes for temporal alignment. Finally, we develop an adaptive negative selection module to adaptively generate a threshold for cross-modal matching. Extensive experiments show the effectiveness and efficiency of our proposed method.

Introduction

Temporal sentence grounding (TSG) (Gao et al. 2017; Li et al. 2023, 2024, 2022; Yu et al. 2024b; Ning et al. 2024, 2023a,b) is an important yet challenging multi-modal task, which has received increasing attention in recent years due to its wide potential applications, such as video understanding (Liu et al. 2024a; Wang et al. 2024c; Hu et al. 2023; Fei et al. 2024b,a; Wu et al. 2024; Wang et al. 2024b, 2023, 2022a; Zhang et al. 2024a,b; Yu et al. 2022) and human-computer interaction (Liu et al. 2024b, 2023c; Tang et al. 2022a, 2024a; Feng et al. 2023b,a; Zhao et al. 2024a,b; Wang et al. 2024a; Yu et al. 2024a, 2023). By complex multi-modal interactions and complicated context information, TSG targets the challenging problem of locating a variety of sentence queries about a video, which requires the designed models to understand both natural language and long video, including reasoning about activities, objects, sequence of events, and interactions within the video (Hu et al. 2022; Wei et al. 2023; Zhao et al. 2021, 2017, 2022, 2018; Jia et al. 2024; Gao et al. 2021, 2022, 2024a,b,c). Given an untrimmed video and a sentence query in Fig. 1(a), TSG aims to determine the segment boundaries that contain the query-relevant activity (Qu et al. 2020; Dong et al. 2022).

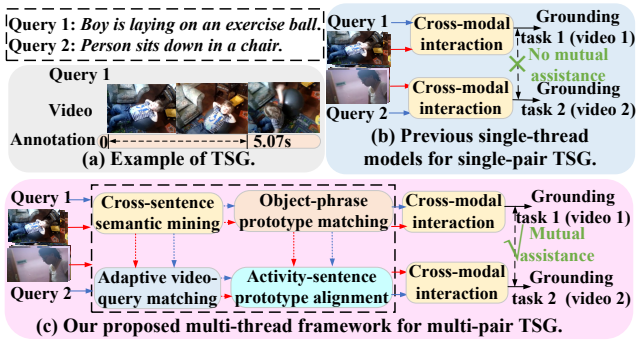

Figure 1: (a) Example of temporal sentence grounding (TSG). (b) Previous TSG models. (c) Our proposed model.

Most previous TSG works (Xiu et al. 2024; Liu et al. 2022; Ji et al. 2023b,a, 2024c,b,a) refer to a fully-supervised setting, where each frame is manually labeled as query-relevant or not. To avoid using such labor-intensive frame-level annotations, some recent works (Fang et al. 2025; Zhang et al. 2023d,c,b, 2022) explore a weakly-supervised setting with only the video-query correspondence to alleviate the reliance to a certain extent. Fully- and weakly-supervised methods only treat each video-query pair independently and ignore the semantic relationship between different video-query pairs in Fig. 1(b). Since a video often corresponds to multiple queries, their treatment will

*Corresponding Authors.

repeatedly extract video features and repetitively conduct complex multi-modal calculations, resulting in weak efficiency. Besides, ignoring the semantic relationship between different video-query pairs might miss important spatio-temporal semantic consistency (*e.g.*, common noun/appearance "woman" and temporal relationship "continues to" in Fig. 2), which limits their effectiveness. Thus, an effective and efficient model is expected to explore the latent semantics relationship between different video-query pairs.

Hence, we pose a novel task: *can we co-train multiple video-query pairs and transfer the grounding knowledge from a pair to another pair?* We show this brand-new task *"multi-pair TSG"* (MP-TSG) in Fig. 1(c). To the best of our knowledge, there is no such setting proposed in existing works. To address this brand-new and challenging setting, we propose a novel multi-thread framework to co-train different pairs. We notice that the semantic relationship between different video-query pairs includes four aspects: query-to-query relationship, video-to-query relationship, cross-modal spatial relationship (object-to-phrase) and cross-modal temporal relationship (activity-to-sentence). We first mine the shared spatial semantics and temporal relationships across different sentence queries to assist with each other in the TSG task. To mine the intra-modal information and obtain inter-modal representation simultaneously, we then design a cross-modal contrast module to explore the global-level semantic consistency between videos and queries by a self-supervised strategy. Moreover, we design an adaptive negative selection module to adaptively generate a dynamic threshold for cross-modal matching. To sufficiently align fine-grained visual information and fine-grained textual information from spatial and temporal perspectives, we design a prototype alignment strategy to 1) match the object prototypes and phrase prototypes to align appearance representations across modalities, and 2) align activity prototypes and sentence prototypes to integrate motion representations between different modalities.

Our main contributions include: 1) We pose and address a brand-new task: MP-TSG that aims to co-train multiple video-query pairs by exploring the semantic relationships between different pairs to assist with each other. We propose a novel multi-thread framework to co-train different pairs by mining the relationships in four aspects: query-to-query, video-to-query, object-to-phrase, activity-to-sentence. 2) We propose a novel cross-modal prototype alignment module to explore the semantic relationship between different queries/videos. To deeply explore cross-modal matching, we design an adaptive negative selection module to automatically generate a dynamic threshold for semantically matching video-query pairs. 3) Extensive experiments show both effectiveness and efficiency of our proposed method.

Related Works

Fully-supervised TSG. TSG (Gao et al. 2017; Ju et al. 2024, 2023; Liang et al. 2024a,b, 2023) aims at locating the most relevant segments from long videos corresponding to the given sentence descriptions. Traditional TSG methods (Gao et al. 2017; Rao et al. 2021b,a; Jiang et al. 2023; Zhang et al. 2023a; Wang et al. 2025b,a; Tang et al. 2024d, 2022b, 2021, 2022c, 2023, 2024b,c) typically utilize a *propose-and-rank* approach to make predictions based on interacted multi-modal features. Some *proposal-free* methods (Zhang et al. 2020a; Lin et al. 2024b,a,c, 2025; Wang et al. 2021a,b, 2020) are proposed to directly predict the temporal locations of the target segment without generating proposals.

Weakly-supervised TSG. The above fully-supervised methods heavily rely on the datasets that require numerous manually labeled annotations for training. To ease human labeling efforts, recent works (Mithun, Paul, and Roy-Chowdhury 2019; Liu et al. 2023a,b; Wen et al. 2023; Fang et al. 2023b,a, 2024) consider a weakly-supervised setting to only access the information of matched video-query pairs without accurate segment boundaries. However, their performance is less satisfactory with such weak supervision.

Many semantic relationships between different queries/videos are not explored in previous methods, leading to repeated training and much computational cost. Unlike them, we introduce a brand-new setting, MP-TSG, where different queries and videos can be co-trained to reduce the cross-modal gap between video and query.

Methodology

Problem definition. Given M_q video-query pairs $\{V_p, Q_p\}_{p=1}^{M_q}$, previous TSG methods aim to localize the query-described activity segment in the video for each video-query pair, where V_p and Q_p are the corresponding video and query, respectively. They independently regard each video-query pair with ignoring the semantic relationship between different queries and videos, and repeat the grounding process M_q times. Different from them, we pose a practical yet challenging setting, Multi-Pair TSG (MP-TSG), which aims to co-train multiple video-query pairs for effective and efficient grounding.

Pipeline. To tackle the MP-TSG task, we propose a novel framework in Fig. 2. The semantic relationship between different video-query pairs includes four aspects: query-to-query relationship, video-to-query relationship, cross-modal spatial relationship (object-to-phrase) and cross-modal temporal relationship (activity-to-sentence).

Video encoder. Given M_v videos $\{V_1, \cdots, V_{M_v}\}$, we first follow previous work (Gao et al. 2017) to extract its frame-wise features by a pre-trained 3D-CNN network (Tran et al. 2015), and then employ a multi-head self-attention (Vaswani et al. 2017) module to capture the long-range dependencies among video frames. For the a-th video with N_v frames, we denote the extracted video features as $V_a = \{v_g^i, v_{a_1}^i, \cdots, v_{a_C}^i\}_{i=1}^{N_v} \in \mathbb{R}^{N_v \times (C+1) \times d}$, where d is the feature dimension, C is the patch number, v_g^i is the global feature of the i-th frame.

Text encoder. Similarly, given M_q queries $\{Q_1, \ldots, Q_{M_q}\}$, by feeding any query Q_j to the pretained Glove network (Pennington, Socher, and Manning 2014), we can obtain the word-level features $Q_j = \{q_1^j, \cdots, q_{N_q}^j\} \in \mathbb{R}^{N_q \times d}$, where N_q is the word number. To extract the semantic of the whole sentence, the Skip-thought parser (Kiros et al. 2015) is employed to capture the query-level feature $q_e^j \in \mathbb{R}^d$.

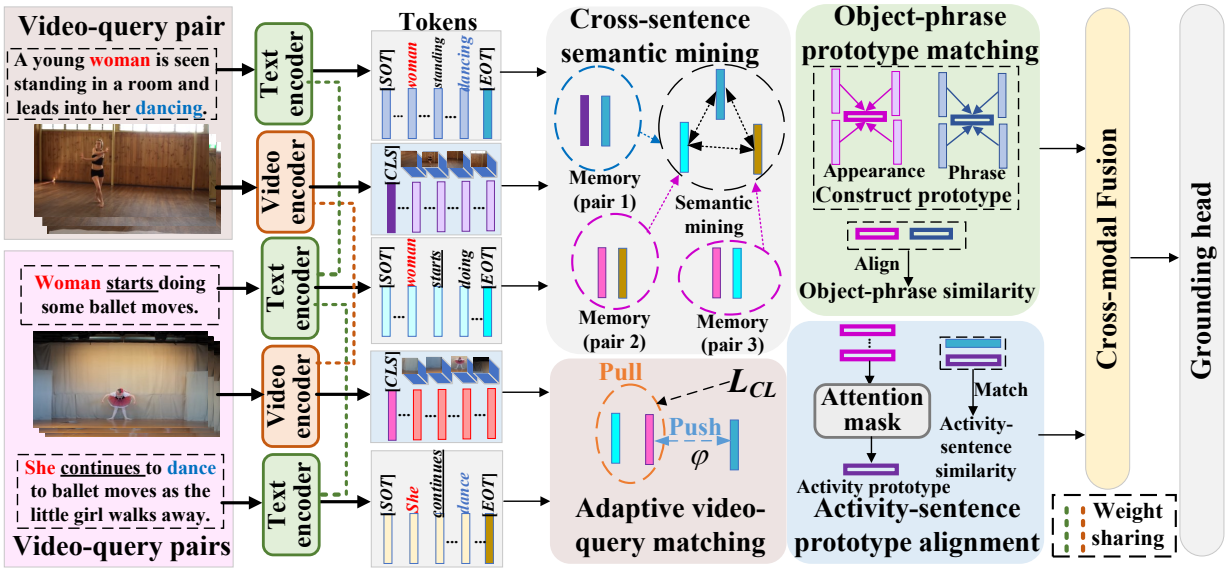

Figure 2: Overview of our proposed MKTN for the MP-TSG task. Best viewed in color.

Cross-Sentence Semantic Mining

Although previous TSG works (Anne Hendricks et al. 2017; Gao et al. 2017) try to fully understand textual information and visual information, they often ignore the semantic relationship between different sentences. The semantic relationship includes 1) the temporal information between different segments in the same video, and 2) the contextual information among different sentences. To sufficiently mine these query-to-query relationships, we extract query-level features $F_q = \{q_e^j\}_{j=1}^{M_q}$ from the multiple queries rather than learnable embeddings in previous works. Thus, we aim to model the query-level contexts, and explore the temporally and contextually related queries of each query. For example, in Fig. 2, "The woman starts doing some ballet moves." and "She continues to dance to ballet moves as the little girl walks away." share the same semantics ("woman" and "ballet moves"), and contains the temporal relationship ("start" and "continues to"). The shared semantics and temporal relationship will assist the grounding task of each sentence.

Given the query-level features F_q, we first encode temporal information by the position embedding layer, and then conduct the interactions among queries by the self-attention layers. After that, we extract a textual feature for each query to represent corresponding events from the multi-modal memory by cross-attention layers. Considering that some sentences might share the same particular words (*e.g.*, "woman" in Fig. 2), we extract the hierarchical textual features to conduct the cross-granularity interactions, which makes the decoder learn more contextual information. Finally, we compute the timestamps of each query-wise feature by a parallel regression layer. The above procedures are formulated as $T = MLP(\tau(F_q, F_{mem}))$, where $T = \{(t_s^j, t_e^j)\}_{j=1}^{M_q}$ denotes the ground-truth start and end timestamps (t_s^j, t_e^j) for M_q queries; τ denotes the decoder of the transformer to conduct the query-level position embedding; $F_{mem} = [F_v^j; F_w^j]_{j=1}^{M_q}$ denotes the multi-modal memory, where $F_v^j = \{v_{a_1}^j, \cdots, v_{a_C}^j\}$ is the frame-level feature of the video paired with query Q_j, $F_w^j = \{q_1^j, \cdots, q_{N_q}^j\}$ is the word-level feature, $[\cdot;\cdot]$ denotes the concatenating operation.

Adaptive Video-Query Matching

Adaptive negative selection. Common TSG datasets inherently treat a set of video-query pairs as positive matches. Negatives are assigned under the assumption that all non-corresponding pairs are semantically distinct. However, some queries labeled as negative may indeed partially or accurately align with a video, constituting false negatives. Therefore, we design a dynamic threshold-based negative selection strategy to adaptively select negatives. For a given query Q_j, the negative videos are selected as: $\mathcal{N}_q = \{V_i | S_{ij} < \phi\} \cap \mathcal{N}, S_{ij} = V_i (Q_j W_S)^\top \in \mathbb{R}^{N_v \times N_q}$. where S_{ij} denotes the similarity between video i and query j; $W_S \in \mathbb{R}^{d \times d}$ projects the query features into the same latent space as the video; $\mathcal{N}$ is the original negative set; ϕ is a dynamic threshold: $\phi = \phi_{final} - (\phi_{final} - \phi_{intial}) \cos(r\pi + 1)$, where ϕ_{intial} and ϕ_{final} are the thresholds at the start and end of training respectively, $r \in [0, 1]$ is the percentage of the training process. The design of the cosine annealing threshold where $\phi_{final} > \phi_{intial}$ is based on the intuition that the model has a higher confidence level in later training. In the early stages of training, negatives with relatively high similarity scores can be reliably regarded as false negatives.

Self-weighted cross-modal contrast. To mine the intra-modal information and obtain inter-modal representations simultaneously, we design a cross-modal contrast module to explore semantic consistency. As shown in Fig. 2, we map the word-level textual feature F_w and frame-level feature F_v into a shared subspace for semantic alignment. Especially, we introduce the transformer encoder ς to generate the transferred word-level textual feature F_w' and video feature F_v' as:

$F'_w = Norm(\varsigma(F_w)), F'_v = Norm(\varsigma(F_v))$.

We construct a triplet tuple (F'^+_v, F'^+_w, F'^-_w) to denote the pair relationship across queries, where (F'^+_v, F'^+_w) is a positive pair and (F'^+_v, F'^-_w) is a negative pair. To pull the positive pairs (F'^+_v, F'^+_w) together and push the negative pairs (F'^+_v, F'^-_w) away, we map the word-level textual feature F_w and frame-level feature F_v into a shared subspace for semantic alignment. Similarly, we can obtain the inter-video relationship: (F'^+_w, F'^+_v, F'^-_v), where (F'^+_w, F'^+_v) is a positive pair and (F'^+_w, F'^-_v) is a negative pair. Thus, we design the following contrastive loss for self-supervision:

$$\mathcal{L}_{CL} = \sum_{F'^+_v, F'^+_w} \{\theta \sum_{F'^-_w} \max[0, \phi - S_{(F'^+_v, F'^+_w)} + S_{(F'^+_v, F'^-_w)}] + (1-\theta) \sum_{F'^-_v} \max[0, \phi - S_{(F'^+_w, F'^+_v)} + S_{(F'^+_w, F'^-_v)}]\},$$

where ϕ is the dynamic threshold; the scoring function $S_{(\cdot,\cdot)}$ measures the similarity between the visual feature and textual feature in the joint space; $\theta \in (0,1)$ is a regularized parameter to balance the significance between negative videos and negative queries. We can obtain θ by: $\frac{\theta}{1-\theta} = \frac{(||F'^+_v - F'^-_w||_2 - ||F'^+_v - F'^+_w||_2)^2}{(||F'^+_v - F'^-_w||_2 - ||F'^+_v - F'^+_w||_2)^2}$.

Object-Phrase Prototype Matching

Constructing appearance prototype. For the given videos, we aggregate the patch features into object-level prototypes to represent fine-grained appearance representations, such as object instance, object part, and action region. During prototype construction, not all patch features are aggregated. Although the patch features contain important appearance representations, they also bring redundancy. For example, some background regions may interfere with cross-modal alignment. Hence, we filter out retrieval-superfluous information and generate object-level prototypes in a sparse aggregation manner. For convenience, three Fully Connected (FC) layers and ReLU function are utilized to predict sparse visual weights $W^i_a \in \mathbb{R}^{(C+1)\times N_a}$, where N_a denotes the number of object-level prototypes in the i-th video. Therefore, we prevent these object-level prototypes from being affected by redundant patches.

For each frame $F^i_v \in \mathbb{R}^{(C+1)\times d}$, its constructed appearance prototype is $P^i_a = W^{i^T}_a F^i_v \in \mathbb{R}^{N_a \times d}$. Ideally, each object prototype can adaptively aggregate the corresponding object-related or action-related patches. As for each phrase, we also apply a similar network structure to aggregate word features. Similarly, we use three FC layers and ReLU functions to obtain sparse textual weights $W^j_p \in \mathbb{R}^{(C+1)\times N_p}$, where N_p is the number of phrase prototypes in the j-th query. Thus, we can fully extract the significant appearance representations by the fine-grained patch features and word features. Also, the phrase prototypes $P^{i,j}_p = W^{j^T}_p F^i_v$ are optimized by spatial object-phrase prototype matching.

Object-phrase cross-modal alignment. We propose a prototype-wise query-video interaction from an appearance perspective. Specifically, we first compute the maximum similarity of object-phrase prototypes within each frame. This associates the phrase prototypes most similar to each object prototype, reflecting cross-modal fine-grained alignment. Then, for the multi-frame object similarity matrix, we find the largest similarity score across the frame sequences, which gives a more confident probability of object-phrase matching. Finally, the object-phrase matching scores are summed for the final similarity: $s_a = \frac{1}{N_a} \sum_{l=1}^{N_a} \max_{i=1}^{N_v} \max_{j=1}^{N_p} [P^{i,j}_p \times P^{l^T}_a]$.

Activity-Sentence Prototype Alignment

Building activity prototype. To fully understand the video activity, we aim to design diverse activity prototypes. A naive solution to obtain video-level features based on global frame features is by mean pooling, or by adding motion encoder layers. However, this leads to two issues: 1) failure to perceive local details and ignoring important objects will exacerbate the bias of video feature learning; 2) these strategies generate a single video-level feature, which can only quantify one-to-one relations. Therefore, we investigate how to incorporate key fine-grained objects and dynamic motion changes into diverse activity prototypes.

The core idea is to progressively aggregate spatial object prototypes into frame prototypes and then perform inter-frame interaction to generate various activity prototypes. A frame decoder is first designed to incorporate all object prototypes $P_a \in \mathbb{R}^{(N_v \times N_a)\times d}$ into frame-level prototypes $P_f \in \mathbb{R}^{N_a \times d}$, which implies fine-grained inter-object spatial relations. To learn frame-level object relations, we define the masked attention as: $P_f = Q_f + softmax(Q_f C^T_a + W_f)V_a$, where $Q_f \in \mathbb{R}^{N_a \times d}$ refers to frame queries (*i.e.*, a set of randomly initialized learnable features), V_a and C^T_a are the features after the linear transformation of object prototypes P_a. The attention mask $W_f \in \mathbb{R}^{N_v \times (N_v \times N_a)}$ is: $W_f(i,j) = 0$ if $N_a \cdot i \le J < N_a \cdot (i+1)$; otherwise $W_f(i,j) = \infty$. We add frame prototype p^i_f original global feature v^i_C of corresponding frames to enhance the robustness of the model: $p^i_f = (p^i_f + v^i_C)/2$.

Next, a dynamic activity decoder is developed to learn the inter-frame relationship in P_f, which can obtain different activity prototypes $P_e \in \mathbb{R}^{M_q \times d}$ to illustrate the rich information of videos. Our dynamic attention is formulated as: $P_e = Q_e + softmax(Q_e K^T_a)V_a$, where $Q_e = [q^1_e, q^2_e, \cdots, q^{M_q}_e] \in \mathbb{R}^{M_q \times d}$ refers to activity queries, V_f and K^T_f are the features after the linear transformation of object prototypes P_f. During training, each activity query learns how to adaptively focus on video frame prototypes, while multiple queries implicitly guarantee a certain activity diversity. Differently, since the same video often corresponds to multiple text semantic descriptions, we directly use the global text representation q_e as a sentence prototype to align with the activity prototypes P_e.

Activity-sentence alignment. The activity-sentence prototype alignment is expressed as: $s_{es} = \max_{i=1}^{M_q}(q_e, P_{e_i})$. By the similarity, we can find the closest activity prototype to the text representation for dynamic alignment.

Multi-Modal Fusion and Grounding Head

After obtaining the video feature V_i and query feature Q_j, we further utilize a co-attention mechanism (Lu et al. 2019)

to capture the cross-modal interactions between videos and queries. Specifically, we first calculate the similarity S_{ij} between V_i and Q_j. Then, we compute two attention weights as: $A = S_r(QW_S) \in \mathbb{R}^{N_v \times d}$ and $B = S_r S_c^{\mathrm{T}} V \in \mathbb{R}^{N_v \times d}$, where S_r and S_c are the row- and column-wise softmax results of S, respectively. We compose the final query-guided video representation by learning its sequential features: $F = \text{Bi-GRU}([V; A; V \odot A; V \odot B]) \in \mathbb{R}^{N_v \times d}$, where Bi-GRU$(\cdot)$ denotes the Bi-GRU layers, and $\odot$ is the element-wise multiplication. The output $F = \{f^i\}_{i=1}^{N_v}$ encodes visual features with query-guided attention, where $f^i \in \mathbb{R}^d$. In our model, we treat query-video pairs as positive examples, while considering all other pairwise combinations in the batch as negative examples. To fully leverage the query-video pair information, we introduce the query-to-video Robust InfoNCE (RINCE) loss as follows:

$$\mathcal{L}_{q\to v}(S) = 1/M_v \sum\nolimits_{i=1}^{M_v} [-\frac{\exp(S^{ii})}{\tau} + \frac{(\alpha \sum_{j=1}^{M_q} \exp(S^{ij}))^\tau}{\tau}],$$

where $\tau, \alpha \in (0, 1]$ are learnable parameters. Similarly, the video-to-query loss is:

$$\mathcal{L}_{v\to q}(S) = \frac{1}{M_q} \sum\nolimits_{j=1}^{M_q} [-\frac{\exp(S^{jj})}{\tau} + \frac{(\alpha \sum_{i=1}^{M_v} \exp(S^{ji}))^\tau}{\tau}].$$

Denoting activity-sentence and object-phrase prototype matching similarity matrices as S_{es} and S_{op} respectively, we design the alignment loss as follows:

$$\mathcal{L}_1 = \mathcal{L}_{v\to q}(S_{es}) + \mathcal{L}_{v\to q}(S_{op}) + \mathcal{L}_{q\to v}(S_{es}) + \mathcal{L}_{q\to v}(S_{op}).$$

To predict the segment start/end boundary quickly and accurately, we first introduce the span predictor (Zhang et al. 2020a) with two stacked transformer blocks and two feed-forward layers. Specifically, the multi-modal feature F is fed into the span predictor, followed by a softmax function, to obtain two probability scores of start and end boundaries. We denote them as $P_{s(e)} \in \mathbb{R}^{N_v}$. The rounded integer boundaries $\hat{t}_{s(e)}$ are used to generate one-hot label vectors $Y_{s(e)}$ as supervision.

$$\mathcal{L}_2 = L_{CE}(P_s, Y_s) + L_{CE}(P_e, Y_e), \tag{1}$$

where L_{CE} means the cross-entropy loss. The predicted boundary timestamps $\hat{t}'_{s(e)}$ are obtained by $P_{s(e)}$: $(\hat{t_s}', \hat{t_e}') = arg\max_{\hat{t}'_s, \hat{t}'_e} P_s(\hat{t_s}')P_e(\hat{t}'_e)$, where $0 \le \hat{t}'_s \le \hat{t}'_e \le N_v$.

Since the above span predictor can only predict coarse integer boundary values, we additionally design a parallel float predictor consisting of several feed-forward layers to provide fine-grained float boundaries by the following loss:

$$\mathcal{L}_3 = f_{\text{L1-smooth}}(\hat{t}'_s, t_s) + f_{\text{L1-smooth}}(\hat{t}'_e, t_e), \tag{2}$$

where $f_{\text{L1-smooth}}$ represents the smooth L1 loss and (t_s, t_e) is the ground-truth boundary. The predicted float boundaries O_s and O_e respectively represent the percentage of start and end boundary frames that are query-relevant. Therefore, the fine-grained boundary indexes $\tilde{t}'_{s(e)}$ are calculated by: $(\tilde{t}'_s, \tilde{t}'_e) = (\hat{t}'_s + 1 - O_s, \hat{t}'_e - 1 + O_e)$.

The multi-modal network is trained by minimizing the weighted sum of the above losses, denoted as $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{CL} + \lambda\mathcal{L}_1 + \gamma\mathcal{L}_2 + \mu\mathcal{L}_3, \tag{3}$$

where λ, γ and μ are parameters to weigh different losses.

ActivityNet Captions					
Method	Type	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
MRTNet	FS	42.02	24.25	-	-
2D-TAN	FS	44.51	26.54	77.13	61.96
MMN	FS	48.59	29.26	79.50	64.76
VDI	FS	48.09	28.76	79.69	64.88
ICVC	WS	29.52	-	66.61	-
VCA	WS	31.00	-	53.83	-
WSTAN	WS	30.01	-	63.42	-
CPL	WS	31.37	-	43.13	-
Ours	**MP**	**58.32**	**35.28**	**86.20**	**71.49**
Charades-STA					
Method	Type	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
VDI	FS	52.32	31.37	87.03	62.30
DRN	FS	53.09	31.75	89.06	60.05
MESM	FS	56.69	35.99	-	-
MRTNet	FS	62.50	43.63	-	-
ICVC	WS	31.02	16.53	77.53	41.91
VCA	WS	38.13	19.57	78.75	37.75
LCNet	WS	39.19	18.17	80.56	45.24
CPL	WS	49.24	22.39	84.71	52.37
MMDist	WS	54.72	26.00	-	-
Ours	**MP**	**69.93**	**46.27**	**97.16**	**69.13**
TACoS					
Method	Type	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
DRN	FS	-	23.17	-	33.36
2D-TAN	FS	37.29	25.32	57.81	45.04
MRTNet	FS	37.81	26.01	-	-
MMN	FS	39.24	26.17	62.03	47.39
MIGCN	FS	48.79	37.57	67.63	57.91
MESM	FS	52.69	39.52	-	-
Ours	**MP**	**53.38**	**42.62**	**73.54**	**62.24**

Table 1: Effectiveness comparison for TSG, where "FS" means "fully-supervised", "WS" means "weakly-supervised" and "MP" means "Multi-Pair".

Experiments

Datasets. For a fair comparison with previous works (Zhang et al. 2020b; Wang et al. 2022b; Zheng et al. 2022), we utilize the same ActivityNet Captions (Caba Heilbron et al. 2015), TACoS (Regneri et al. 2013), and Charades-STA (Sigurdsson et al. 2016) datasets for evaluation.

Evaluation metrics. Following (Gao et al. 2017), we evaluate 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.

Comparison With State-Of-The-Arts

Quantitative comparison. We compare our proposed method with other existing state-of-the-art approaches in Table 1 and Fig. 3(a). Obviously, our proposed method outperforms both fully-supervised and weakly-supervised methods by a large margin. The main reasons are as follows:

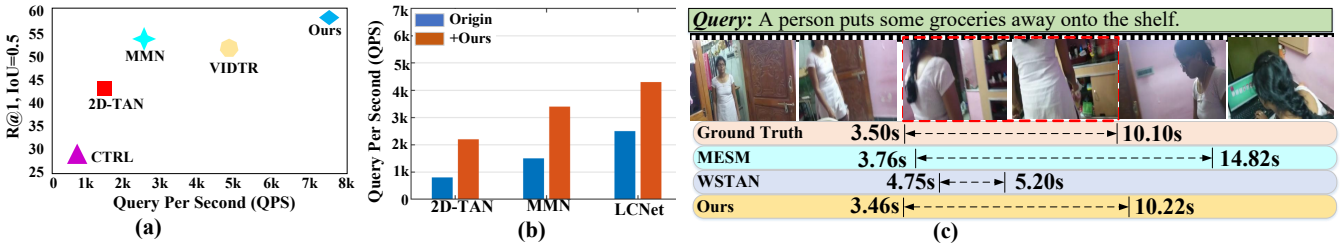

Figure 3: Performance comparison on Charades-STA. (a) compares the effectiveness (R@1, IoU=0.5) and the efficiency (QPS), (b) shows that our method can serve as a plug-and-play module for performance improvement, (c) is the qualitative results.

Methods	Variant	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	VPS
MMN	Origin	47.31	27.28	83.74	58.41	21.63
	+Ours	**49.36**	**31.03**	**88.75**	**62.39**	**48.38**
LCNet	Origin	39.19	18.17	80.56	45.24	23.88
	+Ours	**42.80**	**25.94**	**82.46**	**51.37**	**52.73**

Table 2: Our method serves as a plug-and-play module, where "VPS" is "video per second" during inference.

Model	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
w/o CSM	53.82	28.60	81.95	67.76
w/o OPM	55.40	30.33	83.95	70.02
w/o APA	56.27	30.86	80.21	68.49
w/o AVM	54.95	31.03	83.72	69.30
Full	**58.32**	**35.28**	**86.20**	**71.49**

Table 3: Main ablation study, where we remove each key individual component to investigate its effectiveness. "CSM" is "cross-sentence semantic mining", "OPM" is "object-phrase prototype matching", "APA" is "activity-sentence prototype alignment", "AVM" is "adaptive video-query matching", "Full" is our full model.

1) ActivityNet Captions: these sentence annotations share many nouns and sequencing words, and our method can mine the spatial and temporal relationship between different sentences by our cross-sentence semantic mining module. 2) Charades-STA: some video-query pairs share the same sentence queries. Thus, our method can co-train these pairs and transfer the knowledge from some easy pairs to difficult pairs. 3) TACoS: TACoS only contains activities of cooking scenarios, where these videos often share similar object/appearance information. Our method leverages the information to co-train different videos for better video understanding.

Plug-and-play. Our method can serve as a plug-and-play module for state-of-the-art models on Charades-STA. As shown in Table 2 and Fig. 3(b), our method can significantly improve their performance with higher efficiency.

Visualization comparison. To qualitatively investigate the effectiveness of our method, we report a representative example in Fig. 3(c), where our grounding result is closer to the ground truth than MESM and WSTAN.

Model	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
w/o TR	56.40	30.98	78.36	69.88
w/o CR	57.14	31.06	80.12	67.84
w/o SR	54.85	30.81	78.48	68.95
Full	**58.32**	**35.28**	**86.20**	**71.49**

Table 4: Ablation study on our CSM module, where "TR" is "Temporal relationship", "CR" is "Contextual relationship", "SR" is "Semantic relationship".

Cross-modal contrast via self-supervision					
Visual feature	Textual feature	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
✗	✓	53.03	29.96	80.31	68.69
✓	✗	52.84	30.12	79.68	67.45
✓	✓	**58.32**	**35.28**	**86.20**	**71.49**
Adaptive negative selection					
Fixed threshold		54.08	31.09	82.20	65.13
w/o $\cos(p\pi+1)$		55.15	31.26	82.39	68.04
w/ $\cos(p\pi+1)$		**58.32**	**35.28**	**86.20**	**71.49**

Table 5: Ablation study on our AVM module.

Ablation Study

Main ablation studies. To evaluate the contribution of each module, we perform the main ablation study in Table 3. All the modules contribute a lot to the final performances, demonstrating their effectiveness in exploring the intra- and inter-modal relationship in multiple video-query pairs.

Importance of CSM. About our CSM module, we compare different ablation models on ActivityNet Captions in Table 4, where we remove one relationship between different sentence queries in the first three ablation models. Obviously, our full model achieves the best performance because ActivityNet Captions contains a large number of semantically related queries, and our model can fully mine the semantic relationship between different queries for grounding.

Influence of cross-modal contrast (CC). About our CC module, we conduct corresponding experiments on ActivityNet Captions in Table 5. Both visual and textual features

CAP	OCA	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
✗	✓	56.80	31.03	83.16	70.93
✓	✗	57.03	30.72	84.13	69.48
✓	✓	**58.32**	**35.28**	**86.20**	**71.49**

Table 6: Ablation study on our OPM module, where "CAP" denotes "constructing appearance prototype" and "OCA" denotes "object-phrase cross-modal alignment".

BMP	AA	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
✗	✓	54.88	30.91	82.42	67.60
✓	✗	53.04	31.09	83.86	68.95
✓	✓	**58.32**	**35.28**	**86.20**	**71.49**

Table 7: Ablation study on activity-sentence prototype alignment, where "BMP" denotes "building motion prototype" and "AA" denotes "activity-sentence alignment".

contribute a lot to integrating different video-query pairs.

Effect of adaptive negative selection (ANS). About our ANS module, we change the threshold to obtain two ablation models in Table 5. Obviously, our full model obtains the best results since our ANS module can generate an adaptive threshold for negative query selection, which can fully match queries and relevant videos.

Significance of OPM. About our OPM module in integrating the cross-modal spatial information, an ablation experiment is conducted on ActivityNet Captions in Table 6. Our full model beats other ablation models by a large margin since appearance and phrase prototypes provide visual and textual spatial semantics for multi-modal fusion.

Analysis on APA. We further analyze the performance of our APA module for cross-modal temporal representations on ActivityNet Captions in Table 7. Obviously, both modules bring significant performance improvement since activity and sentence prototypes can understand the temporal semantics from visual and textual perspective respectively.

Conclusion

In this paper, we pose a brand-new and realistic setting: Multi-Pair TSG. For the challenging task, we propose a novel Multi-Thread Knowledge Transfer Network (MKTN) to deeply explore intra- and inter-modal relationships. Extensive experiments on multiple datasets show the effectiveness and efficiency of our proposed MKTN. Moreover, our MKTN can serve as a plug-and-play module for previous methods to enhance their effectiveness and efficiency.

Acknowledgments

The study was supported by the Key Laboratory of Data Protection and Intelligent Management, Ministry of Education, Sichuan University and also the Fundamental Research Funds for the Central Universities under Grant SCU2023D008 and the Sichuan Science and Technology Program under Grant No. 2024YFFK0443.

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *ICCV*.

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

Dong, J.; Wang, Y.; Chen, X.; Qu, X.; Li, X.; He, Y.; and Wang, X. 2022. Reading-strategy inspired visual representation learning for text-to-video retrieval. *TCSVT*.

Fang, X.; Easwaran, A.; Genest, B.; and Suganthan, P. N. 2024. Your Data Is Not Perfect: Towards Cross-Domain Out-of-Distribution Detection in Class-Imbalanced Data. *ESWA*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Cheng, Y.; Tang, K.; and Zou, K. 2023a. Annotations Are Not All You Need: A Cross-modal Knowledge Transfer Network for Unsupervised Temporal Sentence Grounding. In *Findings of EMNLP*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023b. You Can Ground Earlier than See: An Effective and Efficient Pipeline for Temporal Sentence Grounding in Compressed Videos. In *CVPR*.

Fang, X.; Xiong, Z.; Fang, W.; Qu, X.; Chen, C.; Dong, J.; Tang, K.; Zhou, P.; Cheng, Y.; and Liu, D. 2025. Rethinking weakly-supervised video temporal grounding from a game perspective. In *ECCV*.

Fei, H.; Wu, S.; Zhang, H.; Chua, T.-S.; and Yan, S. 2024a. VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. In *NeurIPS*.

Fei, H.; Wu, S.; Zhang, M.; Zhang, M.; Chua, T.-S.; and Yan, S. 2024b. Enhancing video-language representations with structural spatio-temporal alignment. *TPAMI*.

Feng, H.; Liu, Q.; Liu, H.; Zhou, W.; Li, H.; and Huang, C. 2023a. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv*.

Feng, H.; Wang, Z.; Tang, J.; Lu, J.; Zhou, W.; Li, H.; and Huang, C. 2023b. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv*.

Gao, C.; Jiang, Y.; Li, L.; Liu, D.; and Wu, F. 2024a. DMOFC: Discrimination Metric-Optimized Feature Compression. *arXiv*.

Gao, C.; Jiang, Y.; Wu, S.; Ma, Y.; Li, L.; and Liu, D. 2024b. IMOFC: Identity-Level Metric Optimized Feature Compression for Identification Tasks. *TCSVT*.

Gao, C.; Li, L.; Liu, D.; Chen, Z.; Li, W.; and Wu, F. 2022. Two-step fast mode decision for intra coding of screen content. *TCSVT*.

Gao, C.; Li, Z.; Li, L.; Liu, D.; and Wu, F. 2024c. Rethinking the Joint Optimization in Video Coding for Machines: A Case Study. In *DCC*.

Gao, C.; Liu, D.; Li, L.; and Wu, F. 2021. Towards task-generic image compression: A study of semantics-oriented metrics. *TMM*.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *ICCV*.

Hu, D.; Hou, X.; Wei, L.; Jiang, L.; and Mo, Y. 2022. MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations. In *ICASSP*.

Hu, J.; Wang, H.; Zheng, T.; Hu, J.; Chen, Z.; Jiang, H.; and Luo, J. 2023. Password-Stealing without Hacking: Wi-Fi Enabled Practical Keystroke Eavesdropping. In *CCS*.

Ji, W.; Li, L.; Lv, Z.; Zhang, W.; Li, M.; Wan, Z.; Lei, W.; and Zimmermann, R. 2024a. Backpropagation-Free Multi-modal On-Device Model Adaptation via Cloud-Device Collaboration. *TOMM*.

Ji, W.; Liang, R.; Liao, L.; Fei, H.; and Feng, F. 2023a. Partial annotation-based video moment retrieval via iterative learning. In *ACM MM*.

Ji, W.; Liang, R.; Zheng, Z.; Zhang, W.; Zhang, S.; Li, J.; Li, M.; and Chua, T.-s. 2023b. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In *CVPR*.

Ji, W.; Liu, X.; Sun, Y.; Deng, J.; Qin, Y.; Nuwanna, A.; Qiu, M.; Wei, L.; and Zimmermann, R. 2024b. Described Spatial-Temporal Video Detection. *arXiv*.

Ji, W.; Shi, R.; Wei, Y.; Zhao, S.; and Zimmermann, R. 2024c. Weakly Supervised Video Moment Retrieval via Location-irrelevant Proposal Learning. In *WWW*.

Jia, X.; Huang, Y.; Liu, Y.; Tan, P. Y.; Yau, W. K.; Mak, M.-T.; Sim, X. M.; Ng, W. S.; Ng, S. K.; Liu, H.; et al. 2024. Global Challenge for Safe and Secure LLMs Track 1. *arXiv*.

Jiang, L.; Wang, C.; Ning, X.; and Yu, Z. 2023. LTTPoint: A MLP-Based Point Cloud Classification Method with Local Topology Transformation Module. In *ACAIT*.

Ju, Y.; Lam, K.-M.; Xie, W.; Zhou, H.; Dong, J.; and Shi, B. 2024. Deep Learning Methods for Calibrated Photometric Stereo and Beyond. *TPAMI*.

Ju, Y.; Shi, B.; Chen, Y.; Zhou, H.; Dong, J.; and Lam, K.-M. 2023. GR-PSN: Learning to estimate surface normal and reconstruct photometric stereo images. *TVCG*.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. *NIPS*.

Li, L.; Zheng, M.; Liu, X.; Wu, W.; Liu, H.; El Naggar, M. H.; and Jiang, G. 2022. Numerical analysis of the cyclic loading behavior of monopile and hybrid pile foundation. *Computers and Geotechnics*.

Li, Q.; Guo, S.; Wu, J.; Li, J.; Sheng, J.; Peng, H.; and Wang, L. 2023. Event extraction by associating event types and argument roles. *TBD*.

Li, Q.; Li, J.; Wu, J.; Peng, X.; Ji, C.; Peng, H.; Wang, L.; and Philip, S. Y. 2024. Triplet-aware graph neural networks for factorized multi-modal knowledge graph entity alignment. *NN*.

Liang, K.; Liu, Y.; Zhou, S.; Tu, W.; Wen, Y.; Yang, X.; Dong, X.; and Liu, X. 2023. Knowledge graph contrastive learning based on relation-symmetrical structure. *TKDE*.

Liang, K.; Meng, L.; Liu, M.; Liu, Y.; Tu, W.; Wang, S.; Zhou, S.; Liu, X.; Sun, F.; and He, K. 2024a. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *TPAMI*.

Liang, K.; Meng, L.; Liu, Y.; Liu, M.; Wei, W.; Liu, S.; Tu, W.; Wang, S.; Zhou, S.; and Liu, X. 2024b. Simple Yet Effective: Structure Guided Pre-trained Transformer for Multi-modal Knowledge Graph Reasoning. In *ACM MM*.

Lin, M.; Chen, Z.; Liu, Y.; Zhao, X.; Wu, Z.; Wang, J.; Zhang, X.; Wang, S.; and Chen, H. 2024a. Decoding Time Series with LLMs: A Multi-Agent Framework for Cross-Domain Annotation. *arXiv*.

Lin, M.; Dai, E.; Xu, J.; Jia, J.; Zhang, X.; and Wang, S. 2025. Stealing Training Graphs from Graph Neural Networks. *KDD*.

Lin, M.; Xiao, T.; Dai, E.; Zhang, X.; and Wang, S. 2024b. Certifiably robust graph contrastive learning. *NeurIPS*.

Lin, M.; Zhang, Z.; Dai, E.; Wu, Z.; Wang, Y.; Zhang, X.; and Wang, S. 2024c. Trojan Prompt Attacks on Graph Neural Networks. *arXiv*.

Liu, C.; Wen, J.; Wu, Z.; Luo, X.; Huang, C.; and Xu, Y. 2023a. Information Recovery-Driven Deep Incomplete Multiview Clustering Network. *TNNLS*.

Liu, C.; Wu, Z.; Wen, J.; Xu, Y.; and Huang, C. 2023b. Localized Sparse Incomplete Multi-View Clustering. *TMM*.

Liu, D.; Liu, Y.; Huang, W.; and Hu, W. 2024a. A Survey on Text-guided 3D Visual Grounding: Elements, Recent Advances, and Future Directions. *arXiv*.

Liu, D.; Yang, M.; Qu, X.; Zhou, P.; Fang, X.; Tang, K.; Wan, Y.; and Sun, L. 2024b. Pandora's Box: Towards Building Universal Attackers against Real-World Large Vision-Language Models. In *NeurIPS*.

Liu, D.; Zhou, P.; Xu, Z.; Wang, H.; and Li, R. 2022. Few-Shot Temporal Sentence Grounding via Memory-Guided Semantic Learning. *TCSVT*.

Liu, Y.; Zhang, J.; Peng, D.; Huang, M.; Wang, X.; Tang, J.; Huang, C.; Lin, D.; Shen, C.; Bai, X.; et al. 2023c. Spts v2: single-point scene text spotting. *TPAMI*.

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

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly supervised video moment retrieval from text queries. In *CVPR*.

Ning, E.; Wang, C.; Zhang, H.; Ning, X.; and Tiwari, P. 2023a. Occluded person re-identification with deep learning: a survey and perspectives. *ESWA*.

Ning, E.; Wang, Y.; Wang, C.; Zhang, H.; and Ning, X. 2024. Enhancement, integration, expansion: Activating representation of detailed features for occluded person re-identification. *NN*.

Ning, E.; Zhang, C.; Wang, C.; Ning, X.; Chen, H.; and Bai, X. 2023b. Pedestrian Re-ID based on feature consistency and contrast enhancement. *Displays*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Qu, X.; Tang, P.; Zou, Z.; Cheng, Y.; Dong, J.; Zhou, P.; and Xu, Z. 2020. Fine-grained iterative attention network for temporal language localization in videos. In *ACM MM*.

Rao, H.; Hu, X.; Cheng, J.; and Hu, B. 2021a. SM-SGE: A self-supervised multi-scale skeleton graph encoding framework for person re-identification. In *ACM MM*.

Rao, H.; Xu, S.; Hu, X.; Cheng, J.; and Hu, B. 2021b. Multi-Level Graph Encoding with Structural-Collaborative Relation Learning for Skeleton-Based Person Re-Identification. In *IJCAI*.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *TACL*.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*.

Tang, J.; Lin, C.; Zhao, Z.; Wei, S.; Wu, B.; Liu, Q.; Feng, H.; Li, Y.; Wang, S.; Liao, L.; et al. 2024a. TextSquare: Scaling up Text-Centric Visual Instruction Tuning. *arXiv*.

Tang, J.; Qiao, S.; Cui, B.; Ma, Y.; Zhang, S.; and Kanoulas, D. 2022a. You can even annotate text with voice: Transcription-only-supervised text spotting. In *ACM MM*.

Tang, K.; Hou, C.; Peng, W.; Chen, R.; Zhu, P.; Wang, W.; and Tian, Z. 2024b. CORES: Convolutional Response-based Score for Out-of-distribution Detection. In *CVPR*.

Tang, K.; Ma, Y.; Miao, D.; Song, P.; Gu, Z.; Tian, Z.; and Wang, W. 2022b. Decision fusion networks for image classification. *TNNLS*.

Tang, K.; Miao, D.; Peng, W.; Wu, J.; Shi, Y.; Gu, Z.; Tian, Z.; and Wang, W. 2021. Codes: Chamfer out-of-distribution examples against overconfidence issue. In *CVPR*.

Tang, K.; Shi, Y.; Lou, T.; Peng, W.; He, X.; Zhu, P.; Gu, Z.; and Tian, Z. 2022c. Rethinking perturbation directions for imperceptible adversarial attacks on point clouds. *IoT-J*.

Tang, K.; Wang, Z.; Peng, W.; Huang, L.; Wang, L.; Zhu, P.; Wang, W.; and Tian, Z. 2024c. SymAttack: Symmetry-aware Imperceptible Adversarial Attacks on 3D Point Clouds. In *ACM MM*.

Tang, K.; Wu, J.; Peng, W.; Shi, Y.; Song, P.; Gu, Z.; Tian, Z.; and Wang, W. 2023. Deep manifold attack on point clouds via parameter plane stretching. In *AAAI*.

Tang, K.; Zhao, W.; Peng, W.; Fang, X.; Cui, X.; Zhu, P.; and Tian, Z. 2024d. Reparameterization Head for Efficient Multi-Input Networks. In *ICASSP*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Wang, A.-L.; Shan, B.; Shi, W.; Lin, K.-Y.; Fei, X.; Tang, G.; Liao, L.; Tang, J.; Huang, C.; and Zheng, W.-S. 2024a. ParGo: Bridging Vision-Language with Partial and Global Views. *arXiv*.

Wang, C.; He, S.; Fang, X.; Wu, M.; Lam, S. K.; and Tiwari, P. 2025a. Taylor Series-Inspired Local Structure Fitting Network for Few-shot Point Cloud Semantic Segmentation. In *AAAI*.

Wang, C.; Ning, X.; Li, W.; Bai, X.; and Gao, X. 2023. 3D person re-identification based on global semantic guidance and local feature aggregation. *TCSVT*.

Wang, C.; Wang, H.; Ning, X.; Shengwei, T.; and Li, W. 2022a. 3d point cloud classification method based on dynamic coverage of local area. *Journal of Software*.

Wang, C.; Wu, M.; Lam, S.-K.; Ning, X.; Yu, S.; Wang, R.; Li, W.; and Srikanthan, T. 2024b. GPSFormer: A Global Perception and Local Structure Fitting-based Transformer for Point Cloud Understanding. *arXiv*.

Wang, G.; Wu, X.; Liu, Z.; and Wang, H. 2021a. Hierarchical attention learning of scene flow in 3d point clouds. *TIP*.

Wang, G.; Wu, X.; Liu, Z.; and Wang, H. 2021b. Pwclo-net: Deep lidar odometry in 3d point clouds using hierarchical embedding mask optimization. In *CVPR*.

Wang, G.; Zhang, C.; Wang, H.; Wang, J.; Wang, Y.; and Wang, X. 2020. Unsupervised learning of depth, optical flow and pose with occlusion from 3d geometry. *TITS*.

Wang, H.; Hu, J.; Zheng, T.; Hu, J.; Chen, Z.; Jiang, H.; Zheng, Y.; and Luo, J. 2024c. MuKI-Fi: Multi-person Keystroke Inference with BFI-enabled Wi-Fi Sensing. *IEEE TMC*.

Wang, R.; Lam, S.-K.; Wu, M.; Hu, Z.; Wang, C.; and Wang, J. 2025b. Destination intention estimation-based convolutional encoder-decoder for pedestrian trajectory multimodality forecast. *Measurement*.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022b. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. In *AAAI*.

Wei, L.; Hu, D.; Zhou, W.; and Hu, S. 2023. Modeling Both Intra- and Inter-Modality Uncertainty for Multimodal Fake News Detection. *TMM*.

Wen, J.; Liu, C.; Deng, S.; Liu, Y.; Fei, L.; Yan, K.; and Xu, Y. 2023. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *TNNLS*.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2024. NExT-GPT: Any-to-Any Multimodal LLM. In *ICML*.

Xiu, J.; Li, M.; Ji, W.; Chen, J.; Zhao, H.; Satoh, S.; and Zimmermann, R. 2024. Hierarchical Debiasing and Noisy Correction for Cross-domain Video Tube Retrieval. In *ACM MM*.

Yu, S.; Sun, X.; Li, W.; Wen, C.; Yang, Y.; Si, B.; Hu, G.; and Wang, C. 2024a. NIDALoc: Neurobiologically Inspired Deep LiDAR Localization. *TITS*.

Yu, S.; Wang, C.; Lin, Y.; Wen, C.; Cheng, M.; and Hu, G. 2023. STCLoc: Deep LiDAR Localization With Spatio-Temporal Constraints. *TITS*.

Yu, S.; Wang, C.; Wen, C.; Cheng, M.; Liu, M.; Zhang, Z.; and Li, X. 2022. LiDAR-based localization using universal encoding and memory-aware regression. *PR*.

Yu, Z.; Li, L.; Xie, J.; Wang, C.; Li, W.; and Ning, X. 2024b. Pedestrian 3d shape understanding for person re-identification via multi-view learning. *TCSVT*.

Zhang, H.; Ning, X.; Wang, C.; Ning, E.; and Li, L. 2024a. Deformation depth decoupling network for point cloud domain adaptation. *NN*.

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

Zhang, H.; Wang, C.; Tian, S.; Lu, B.; Zhang, L.; Ning, X.; and Bai, X. 2023a. Deep learning-based 3D point cloud classification: A systematic survey and outlook. *Displays*.

Zhang, H.; Wang, C.; Yu, L.; Tian, S.; Ning, X.; and Rodrigues, J. 2024b. PointGT: A Method for Point-Cloud Classification and Segmentation Based on Local Geometric Transformation. *TMM*.

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

Zhang, Y.; Chen, Y.; Song, Z.; and King, I. 2023b. Contrastive cross-scale graph knowledge synergy. In *SIGKDD*.

Zhang, Y.; Zhu, H.; Song, Z.; Koniusz, P.; and King, I. 2022. COSTA: covariance-preserving feature augmentation for graph contrastive learning. In *SIGKDD*.

Zhang, Y.; Zhu, H.; Song, Z.; Koniusz, P.; and King, I. 2023c. Spectral feature augmentation for graph contrastive learning and beyond. In *AAAI*.

Zhang, Y.; Zhu, H.; Song, Z.; Koniusz, P.; King, I.; et al. 2023d. Mitigating the popularity bias of graph collaborative filtering: A dimensional collapse perspective. *NeurIPS*.

Zhao, X.; Chen, H.; Xing, Z.; and Miao, C. 2021. Brain-inspired search engine assistant based on knowledge graph. *TNNLS*.

Zhao, X.; Li, H.; Tang, Y.; Gao, D.; Bao, L.; and Lee, C.-H. 2018. A Smart Context-Aware Program Assistant Based on Dynamic Programming Event Modeling. In *ISSRE Workshops*.

Zhao, X.; Liu, Y.; Xu, Y.; Yang, Y.; Luo, X.; and Miao, C. 2022. Heterogeneous star graph attention network for product attributes prediction. *AEI*.

Zhao, X.; Xing, Z.; Kabir, M. A.; Sawada, N.; Li, J.; and Lin, S.-W. 2017. Hdskg: Harvesting domain specific knowledge graph from content of webpages. In *SANER*.

Zhao, Z.; Tang, J.; Lin, C.; Wu, B.; Huang, C.; Liu, H.; Tan, X.; Zhang, Z.; and Xie, Y. 2024a. Multi-modal In-Context Learning Makes an Ego-evolving Scene Text Recognizer. In *CVPR*.

Zhao, Z.; Tang, J.; Wu, B.; Lin, C.; Wei, S.; Liu, H.; Tan, X.; Zhang, Z.; Huang, C.; and Xie, Y. 2024b. Harmonizing Visual Text Comprehension and Generation. *arXiv*.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *CVPR*.