CDTR: Semantic Alignment for Video Moment Retrieval Using Concept Decomposition Transformer

Ran Ran[1], **Jiwei Wei**[1*], **Xiangyi Cai**[1], **Xiang Guan**[1], **Jie Zou**[1], **Yang Yang**[1], **Heng Tao Shen**[1,2]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2] School of Computer Science and Technology, Tongji University
ranran@std.uestc.edu.cn, mathematic6@gmail.com, caixiangyi@std.uestc.edu.cn, duochuan.gx@gmail.com, {jie.zou, yang.yang}@uestc.edu.cn, shenhengtao@hotmail.com

Abstract

Video Moment Retrieval (VMR) involves locating specific moments within a video based on natural language queries. However, existing VMR methods that employ various strategies for cross-modal alignment still face challenges such as limited understanding of fine-grained semantics, semantic overlap, and sparse constraints. To address these limitations, we propose a novel Concept Decomposition Transformer (CDTR) model for VMR. CDTR introduces a semantic concept decomposition module that disentangles video moments and sentence queries into concept representations, reflecting the relevance between various concepts and capturing fine-grained semantics which is crucial for cross-modal matching. These decomposed concept representations are then used as pseudo-labels, determined as positive or negative samples by adaptive concept-specific thresholds. Subsequently, fine-grained concept alignment is performed in video intra-modal and textual-visual cross-modal, aligning different conceptual components within features, enhancing the model's ability to distinguish fine-grained semantics, and alleviating issues related to semantic overlap and sparse constraints. Comprehensive experiments demonstrate the effectiveness of the CDTR, outperforming state-of-the-art methods on three widely used datasets: QVHighlights, Charades-STA, and TACoS.

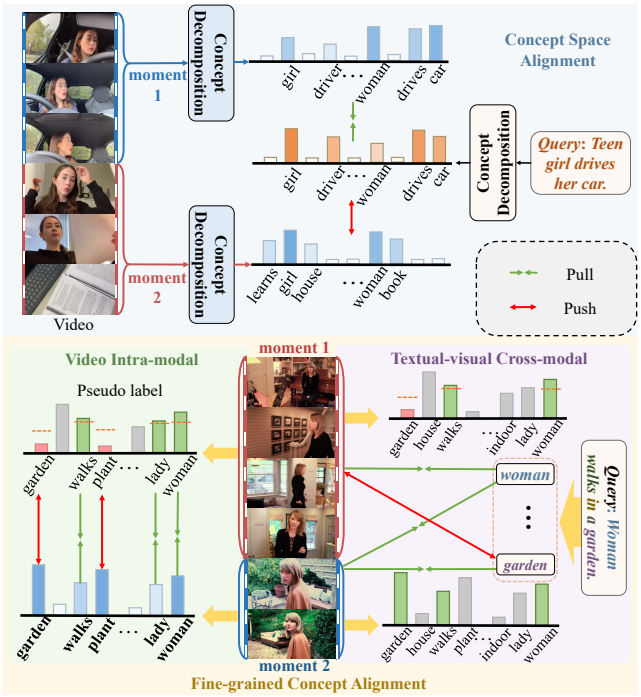

Figure 1: Video and query are decomposed into concept representations, indicating concept relevance. Annotated data facilitates concept space alignment. By using concept representations as pseudo-labels, fine-grained concept alignment is achieved by aligning the conceptual components within the video intra-modal and textual-visual cross-modal spaces.

Introduction

Video Moment Retrieval (VMR) refers to the task of locating specific moments within a video based on natural language queries (Gao et al. 2017; Jiang et al. 2022). Specifically, given a query describing a moment within an untrimmed video, VMR aims to determine the start and end timestamps of relevant video moments (Qu et al. 2020; Li et al. 2024a; Moon et al. 2023b). This challenging and meaningful task in the field of video understanding requires accurate comprehension of both the video content and the language query, as well as the alignment of their representations in a cross-modal space (Wang et al. 2022; Liu et al. 2024; Zhu et al. 2022).

Existing VMR methods employ various strategies for cross-modal alignment (Wei et al. 2023; Liu et al. 2022a; Qi et al. 2024). Some approaches suggest video moments and extract features at the moment-level, aligning them with sentence-level query features and using ranking structures to identify the best-matched moments (Wang et al. 2022; Li et al. 2023; Ning et al. 2021). Other approaches use clip-level video features and word-level query features for cross-modal interaction, followed by predicting the video moments through prediction heads (Mun, Cho, and Han 2020; Liu, Qu, and Hu 2022) or transformer decoder based on DETR (Jang et al. 2023; Lin et al. 2023; Sun et al. 2024).

Despite significant advancements, these methods still face several challenges. Firstly, due to the complexity of cross-

*Corresponding Author

modal semantics, current methods have a limited understanding of visual and linguistic elements, failing to account for various independent concepts within the semantics (Li et al. 2024a; Moon et al. 2023b). They cannot comprehend and recognize fine-grained semantics of concepts (Qu et al. 2020; Fang et al. 2023; Qi et al. 2024). Additionally, previous methods typically use contrastive learning to pull semantically similar positive samples closer while pushing apart semantically dissimilar negative samples. However, misaligned negative samples often exhibit substantial semantic overlap, resulting in contradictory feature representations that hinder accurate alignment (Liu, Qu, and Hu 2022; Li et al. 2024a; Zhu et al. 2022). Furthermore, the VMR encounters the issue of sparse constraints, where only a small portion of moments are annotated, leaving most moments as unconstrained negative samples (Li et al. 2023; Jung et al. 2023; Wang et al. 2022).

To address these issues, we propose a novel Concept Decomposition Transformer (CDTR) model for VMR. CDTR explicitly disentangles video moments and queries into semantic concepts through concept decomposition. By mapping complex cross-modal data to semantic concepts in natural language, we effectively capture intrinsic fine-grained semantics for cross-modal matching. Specifically, the concept decomposition module quantifies the semantic relevance between all semantic concepts from the input data. As shown in the upper part of Fig. 1, after being processed by the concept decomposition module, moments and query are disentangled into relevant vectors of specific concepts, like *teen*, *girl*, and *car*. Next, the decomposed concepts are used as pseudo-labels for fine-grained concept alignment, discerning relevant semantic concepts within different video moments. By aligning the various conceptual components, the issue of semantic overlap is effectively mitigated. Moreover, the use of pseudo-labels allows to exploit information from unlabeled moments, alleviating the sparse constraints.

Fine-grained concept alignment encompasses textual-visual cross-modal alignment and video intra-modal alignment, as shown in the lower part of Fig. 1. For textual-visual cross-modal alignment, we use the decomposed moment concepts as pseudo-labels and match them with the textual concepts, *i.e.* words in the query. Positive and negative samples for both tasks are selected based on adaptive concept-specific thresholds. Subsequently, through similarity evaluation that reflects the relevance of specific concepts by concept extraction and contrastive learning, conceptual components are aligned. Similarly, for video intra-modal alignment, pseudo-labels are used to align the conceptual components in other moments with the salient concepts in the annotated video moments.

In summary, the contributions of this paper are as follows:

- We propose a novel CDTR model for VMR that utilizes semantic concept decomposition to explicitly disentangle video moments and queries into independent concepts, effectively capturing intrinsic fine-grained semantics.
- By using the decomposed concepts as pseudo-labels for fine-grained concept alignment, the model aligns visual and textual conceptual components, enhancing its ability to distinguish subtle semantic features.
- We conduct comprehensive experimental analysis on three widely used datasets, QVHighlight, Charades-STA, and TACoS, to validate the effectiveness of CDTR.

Related Works

Video Moment Retrieval. Video Moment Retrieval (VMR) involves identifying the most relevant video moment based on a given query, a research area rooted in studies by (Gao et al. 2017; Anne Hendricks et al. 2017; Zhang et al. 2023). Existing VMR methods are generally classified into two categories: candidate-based and direct-prediction approaches (Lan et al. 2023; Chen et al. 2023). Candidate-based methods typically generate multiple potential moments and evaluate their relevance to the query using multimodal feature fusion, selecting the candidate with the highest relevance score. Some techniques use multiscale sliding windows to generate candidates and predict scores through operations on a two-dimensional matrix (Liu et al. 2021a), with enhancements achieved through metrics learning to improve alignment capabilities (Wang et al. 2022; Li et al. 2023). Recently, transformer decoder-based methods have been explored, offering end-to-end solutions that directly output predictions with notable success (Cao et al. 2021; Lei, Berg, and Bansal 2021; Moon et al. 2023b,a). Direct-prediction methods avoid generating explicit candidate segments and instead focus on predicting the target moments directly by learning interactions between video and sentence features (Zhang et al. 2020a; Mun, Cho, and Han 2020). Some methods improve VMR performance by designing attention mechanisms (Zeng et al. 2021), analyzing salient feature (Liu et al. 2022a; Liu, Qu, and Hu 2022), and enhancing model generalization (Liu et al. 2024; Li et al. 2022; Mun, Cho, and Han 2020).

Despite advancements in VMR, challenges like ineffective fine-grained semantic understanding, semantic overlap, and sparse constraints remain. To address these, we propose a model that uses semantic concept decomposition to map complex multi-modal data into clear semantic concepts. By leveraging pseudo-labels and adaptive thresholds, we aim to improve the alignment of visual and linguistic elements, handle unlabeled moments, and enhance VMR performance.

Cross-modal Alignment. Recently, contrastive learning has achieved significant success in cross-modal alignment tasks (Khosla et al. 2020; Wei et al. 2024a), such as text-image (Luo et al. 2023; Wei et al. 2020, 2024b) and text-video retrieval (Wang, Zhu, and Yang 2021; Ma et al. 2022; Wei et al. 2021a), by selecting positive and negative samples or using data augmentation to learn joint representations of visual and textual modalities and maximizing mutual information between representations (Hjelm et al. 2018). Some approaches have adopted hierarchical representations to learn the alignment between video and text (Jin et al. 2023; Chen et al. 2020), but they mainly rely on analyzing the query. In the VMR task, some methods improve accuracy by introducing causal reasoning (Nan et al. 2021), geodesic distance (Li et al. 2023), or considering the context of moments (Jung et al. 2023) in contrastive learning.

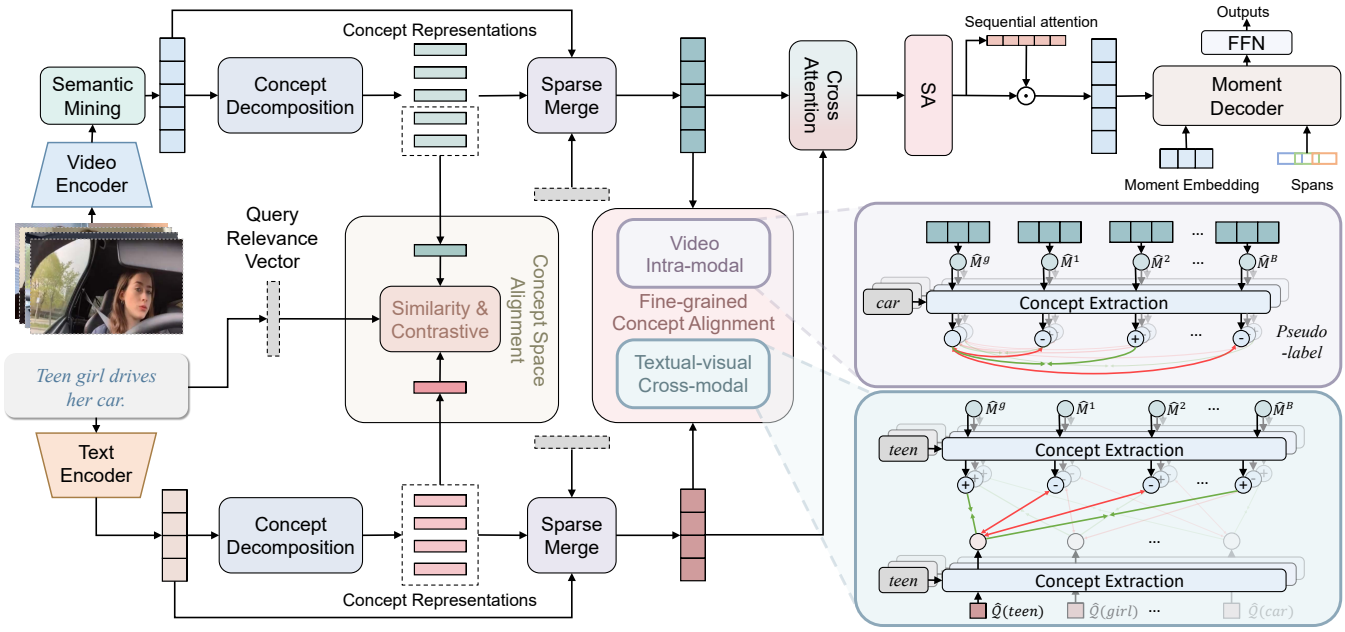

Figure 2: Overview of the proposed Concept Decomposition Transformer (CDTR). The model begins by extracting clip-level and word-level concept representations through the concept decomposition module, while generating query relevance to sparsify these concept representations. Annotated data is then used to supervise concept space alignment at the moment-query level. Subsequently, a sparse merge module aggregates features that are focused on the relevant concepts with query relevance. The concept representations are employed as pseudo-labels for all moments, aiding in the identification of positive and negative samples for specific concepts. CDTR enables fine-grained concept alignment within both video intra-modal and textual-visual cross-modal contexts through contrastive learning of the conceptual components obtained via concept extraction.

Methodology

Problem Formulation

Given an untrimmed video, denoted as V, and a sentence query Q. We represent the video as $V = \{f_i\}_{i=1}^{N_f}$, where N_f denotes the frame count. The sentence query is expressed as $Q = \{w_i\}_{i=1}^{N_w}$, with N_w representing the word count. The goal of VMR is to localize the target video moment in V that semantically aligns with the query Q. This process entails predicting the start and end timestamps (t_s, t_e).

Overview

Fig. 2 illustrates the architecture of CDTR. First, clip-level and word-level feature representations are extracted through text and video encoders. Next, concept representations are generated through concept decomposition, and query relevance vectors are produced to sparsify the concept representations, focusing on the most relevant concept dimensions. Annotated information is used for sentence-moment level concept space alignment, disentangling the fine-grained semantic concepts within the data. The sparse merge module then aggregates the concept representations, and video intra-modal and text-video cross-modal fine-grained alignment are employed to comprehend the fine-grained concepts within moments and queries. Specifically, the concept representations of video moments are used as pseudo-labels to select positive and negative samples, while concept extraction isolates the specific conceptual components of feature representations. Contrastive learning is employed to bring semantically similar concepts in videos and text closer together. Cross-modal interaction is achieved through cross-attention, and the sequential attention mechanism focuses on more significant features after the interaction, finally predicting the target moments through a moment decoder.

Feature Extractors

According to most previous VMR methods (Liu et al. 2024; Moon et al. 2023b), we employ fixed feature extractors to obtain pre-obtained features from the raw data to capture the semantics of the video V and the query Q, respectively. Generally, feature encoders are divided into video encoder and text encoder. We use MLPs to map the extracted video features and query features to a common space. The obtained video features can be represented as $\mathcal{F}_v \in \mathbb{R}^{N_t \times D}$, where N_t represents the number of video clips, D denotes the dimension of the features. The query features can be represented as $\mathcal{Q} = \{q_n\}_{n=1}^{N_w} \in \mathbb{R}^{N_w \times D}$.

Concept Decomposition

We transform the input into concept space representations through concept decomposition. For the concepts utilized, we employ the vocabulary from the BERT tokenizer (Kenton and Toutanova 2019) to represent different concepts, and

perform spelling checks using the SpellChecker library. The final concept library is comprised of 17,533 usable words, denoted as $C = \{c_i\}_{i=1}^{N_s}$, where N_s represents $||C||$, which is 17,533. Subsequently, we derive high-dimensional vectors for each word in the concept library using the CLIP text encoder. The final set of concept vectors is represented as $\{S(c_i)\}_{i=1}^{N_s}$, where $S(c_i)$ denotes a E-dimensional concept vector for query relevance and concept alignment.

Specifically, we disentangle the video and text features into the concept space. For video features, before concept decomposition, a semantic mining module containing multiple self-attentions is used to extract semantic information from the input video $\mathcal{F}_v$, obtaining $\mathcal{V} = \{v_t\}_{t=1}^{N_t} \in \mathbb{R}^{N_t \times D}$. Each clip feature of the video is then decomposed into concept representations of dimension N_s using a concept decomposition, where each value in the concept representations quantifies its relevance to the corresponding word concept. Similarly, the query features are decomposed into text concept representations using a concept decomposition.

A concept decomposition module includes normalization, mapping, and activation layers. For video clips or query features, it can be represented as:

$$\mathcal{V}_d = \psi(\text{norm}(\mathcal{V}) \cdot W_v), \mathcal{Q}_d = \psi(\text{norm}(\mathcal{Q}) \cdot W_q), \quad (1)$$

where $\mathcal{V}_d \in \mathbb{R}^{N_t \times N_s}$, $\mathcal{Q}_d \in \mathbb{R}^{N_w \times N_s}$, $\text{norm}(\cdot)$ denotes layer normalization, $W_v, W_q \in \mathbb{R}^{D \times N_s}$ are learnable mapping matrixes, and $\psi(\cdot)$ represents the softplus function.

Query Relevance Vector. Typically, the semantics of concepts within the concept set are too rich for queries and visual features. Therefore, we need to sparsify the obtained concept representations to determine which dimensions should contribute more to evaluate the similarity between features. Specifically, we generate the query relevance R based on the query. The query relevance $R(Q) \in \mathbb{R}^{N_s}$ represents the correlation between the query Q and the corresponding words in the concept representation. Since the words in the query are limited and cannot elicit more related concepts, we construct a concept set by identifying the most relevant concepts to the words in the query:

$$R(\mathcal{Q})[i] = \begin{cases} \max\limits_{c_n \in Q} cos(S(c_n), S(c_i)) & \text{if } c_i \in top_L(c_n), \forall c_n \in Q \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where $cos(\cdot)$ denotes cosine similarity, $top_L(c_n)$ denotes the top-L concepts most relevant to c_n based on $cos(\cdot)$.

Concept Space Alignment. We denote all moments (i.e., GT moments and non-GT moments in $\mathcal{V}_d$) containing clip-level concept representations from a batch as $\{\mathcal{M}^b\}_{b=1}^{B}$, and $\mathcal{M}^g$ denotes the concept representation of the GT moment corresponding to the query Q. Using max pooling, we aggregate the concept representations of moments and queries into N_s-dimensional global concept representations. The similarity between moment-level and query-level concept representations in relevant semantics is computed as:

$$sim(\mathcal{M}, \mathcal{Q}) = \pi(\mathcal{M}) \cdot \pi(\mathcal{Q}) \cdot R(Q), \quad (3)$$

where $\mathcal{M}$ and $\mathcal{Q}$ represent the concept representation of moment and query, and π denotes the max pooling operation. To align the fine-grained concept representations across modalities, we supervise the model with a contrastive loss based on the annotated moments:

$$\mathcal{L}_{csa} = -\sum_{\mathcal{Q} \in \mathbb{Q}} \log p(\mathcal{M}^g|\mathcal{Q}) - \sum_{\mathcal{M} \in \mathbb{M}^g} \log p(\mathcal{Q}^g|\mathcal{M}), \quad (4)$$

$$p(\mathcal{M}^g|\mathcal{Q}) = \frac{\exp(sim(\mathcal{M}^g, \mathcal{Q})/\tau)}{\sum_{\mathcal{M}^b \in \mathbb{M}} \exp(sim(\mathcal{M}^b, Q)/\tau)}, \quad (5)$$

$$p(\mathcal{Q}^g|\mathcal{M}) = \frac{\exp(sim(\mathcal{M}, \mathcal{Q}^g)/\tau)}{\sum_{\mathcal{Q}^b \in \mathbb{Q}} \exp(sim(\mathcal{M}, Q^b)/\tau)}, \quad (6)$$

where $\mathbb{M}$ and $\mathbb{Q}$ are concept representations of video moments and sentence queries in a training batch, respectively. $\mathbb{M}^g$ denotes moments from the ground truth set. $\mathcal{M}^g$ and $\mathcal{Q}^g$ correspond to the video moments and sentence queries matched to the current $\mathcal{Q}$ and $\mathcal{M}$. Moreover, inactive dimensions are sequentially activated to ensure proper supervision.

Fine-grained Concept Alignment

We aggregate the sparse concept representations through a sparse merge module as follows:

$$\widehat{\mathcal{V}} = \text{MLP}(\mathcal{V}_d \cdot R(Q)) + \mathcal{V}, \quad \widehat{\mathcal{Q}} = \text{MLP}(\mathcal{Q}_d \cdot R(Q)) + \mathcal{Q}. \quad (7)$$

To fully exploit the fine-grained semantic content in videos, we use the obtained fine-grained concept representations as pseudo-labels to align individual semantic concept components of the features $\widehat{\mathcal{V}}$ and $\widehat{\mathcal{Q}}$. This fine-grained concept alignment is divided into two parts: video intra-modal alignment and textual-visual cross-modal alignment.

Video Intra-modal Alignment. We first select the top-K highest value semantic concepts from $\pi(\mathcal{M}^g)$, denoted as $C^g = \{c_k^g\}_{k=1}^{K}$. We use the global concept representations of other video moments (non-GT and other video moments) as pseudo-labels to select positive and negative samples for the specific concept c_k^g. We adopt adaptive concept-specific thresholds: if the corresponding video moment exceeds the threshold $\rho^+(c_k^g)$, the moment is a positive sample for the semantic c_k^g; if it falls below the threshold $\rho^-(c_k^g)$, it is a negative sample. Adaptive concept-specific thresholds are set by analyzing global concept representations of all video moments in each epoch. For each concept, samples greater than the mean plus one standard deviation are positive samples, while those lower than the mean minus one standard deviation are negative samples. Ambiguous pseudo-labels due to classification difficulties can lead to semantic confusion, making it necessary to disregard the intermediate part.

Next, we extract the clip features corresponding to all moments from $\widehat{\mathcal{V}}$, and apply linear mapping followed by pooling to obtain the E-dimensional global representation of all moments, denoted as $\{\widehat{M}^b\}_{b=1}^{B}$. We then calculate the contrastive loss of these features concerning the concepts in C^g to achieve fine-grained concept alignment, expressed as:

$$\mathcal{L}_{via} = -\sum_{c \in C^g} \log\left(\frac{\sum_{\widehat{M}^+ \in \mathbb{P}_c} \exp(\text{csim}(\widehat{M}^g, \widehat{M}^+, S(c))/\tau)}{\sum_{\widehat{M}^b \in \mathbb{P}_c + \mathbb{N}_c} \exp(\text{csim}(\widehat{M}^g, \widehat{M}^b, S(c))/\tau)}\right), \quad (8)$$

$$\text{csim}(\widehat{M}^g, \widehat{M}^b, S(c)) = \frac{(\widehat{M}^g \odot S(c)) \cdot (\widehat{M}^b \odot S(c))}{\|\widehat{M}^g \odot S(c)\|_2 \cdot \|\widehat{M}^b \odot S(c)\|_2}, \quad (9)$$

where csim$(\cdot)$ denotes the similarity based on concept extraction, and $\odot$ represents the Hadamard product for concept extraction (Wei et al. 2021b; Ge et al. 2021), which is used to evaluate the similarity of two input features concerning specific semantics. $\mathbb{P}_c$ and $\mathbb{N}_c$ represent the positive and negative moments for concept c, respectively. $S(c)$ refers to the concept vectors mentioned earlier. The τ denotes the temperature parameter of contrastive loss.

Textual-visual Cross-modal Alignment. We align the fine-grained concepts of video moments with sentence queries. Since the words in the query are sparse and explicit, all words in the query $Q = \{w_i\}_{i=1}^{N_w}$ are directly used as semantic concepts to be aligned. We determine the positive and negative samples of moments by adaptive concept-specific threshold (the GT moment corresponding to the query is the positive sample for all concepts) and calculate the cross-modal contrastive loss:

$$\mathcal{L}_{tca} = -\sum_{w \in Q} \log \left(\frac{\sum_{\widehat{M}^+ \in \mathbb{P}_w} \exp(\text{csim}(\widehat{Q}(w), \widehat{M}^+, S(w))/\tau)}{\sum_{\widehat{M}^b \in \mathbb{P}_w + \mathbb{N}_w} \exp(\text{csim}(\widehat{Q}(w), \widehat{M}^b, S(w))/\tau)} \right), \quad (10)$$

where $\widehat{Q}(w)$ represents the textual feature at the corresponding position of the concept w in the query.

Fusion and Prediction

The video features $\widehat{\mathcal{V}}$ and text features $\widehat{\mathcal{Q}}$ are fused through a cross-attention mechanism, resulting in aggregated features $\mathcal{F} \in \mathbb{R}^{N_t \times D}$. A self-attention is applied to obtain features $\mathcal{F}_a$, and a sequential attention module is used to focus more on query-relevant features, as represented by:

$$\begin{aligned} \alpha &= \sigma(\text{MLP}(\mathcal{F}_a)) \in \mathbb{R}^{N_t}, \\ \mathcal{F}_{ex} &= \alpha \cdot \mathcal{F}_a, \end{aligned} \quad (11)$$

where $\sigma(\cdot)$ represents the sigmoid function. $\mathcal{F}_{ex} \in \mathbb{R}^{N_t \times D}$ is the temporal feature after adding sequential attention. We use binary cross-entropy to constrain the attention α to focus more on query-aligned content:

$$\mathcal{L}_{sa} = -\frac{1}{N_t} \sum_{i=1}^{N_t} (\bar{\alpha}_i \log \alpha_i + (1 - \bar{\alpha}_i) \log(1 - \alpha_i)), \quad (12)$$

where $\bar{\alpha}_i$ is set to 1 if the clip is in the GT moment corresponding to the query, otherwise it is set to 0.

Finally, the features $\mathcal{F}_{ex}$ are fed into a moment decoder for video moment prediction. We introduce learnable spans (Liu et al. 2021b) to effectively use multimodal features during moment prediction. The moment decoder directly uses moment embedding as queries, effectively utilizing $\mathcal{F}_{ex}$ to predict target moments by updating spans. The moment retrieval loss can be represented as:

$$\mathcal{L}_{mr} = \lambda_{iou}\mathcal{L}_{iou}(m, \bar{m}) + \lambda_{L1}\|m - \bar{m}\|_1 + \lambda_{ce}\mathcal{L}_{ce}(y, \bar{y}), \quad (13)$$

where m and $\bar{m}$ represent the predicted and ground truth moments, respectively. $\mathcal{L}_{iou}$ uses the generalized IoU loss (Rezatofighi et al. 2019), and y and $\hat{y}$ denote the confidence scores for the foreground and background of the predicted and ground truth moments (Carion et al. 2020), respectively. The λ_{iou}, λ_{L1}, and λ_{ce} are the balancing hyper-parameters.

Method	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
Results on **Test** Split					
MCN	11.41	2.72	24.94	8.22	10.67
CAL	25.49	11.54	23.40	7.65	9.89
XML	41.83	30.35	44.63	31.73	32.14
XML+	46.69	33.46	47.89	34.67	34.90
M-DETR	52.89	33.02	54.82	29.40	30.73
UMT	56.23	41.18	53.83	37.01	36.12
QD-DETR	62.40	44.98	62.52	39.88	39.86
UniVTG	58.86	40.86	57.60	35.59	35.47
MomentDiff	57.42	39.66	54.02	35.73	35.95
BM-DETR	60.12	43.05	63.08	40.18	40.08
MESM	62.78	45.20	62.64	41.45	40.68
UVCOM	63.55	47.47	63.37	42.67	43.18
CDTR(Ours)	**65.79**	**49.60**	**66.44**	**45.96**	**44.37**
Results on **Val** Split					
M-DETR	53.94	34.84	-	-	32.20
UMT	60.26	44.26	-	-	38.59
QD-DETR	62.68	46.66	62.23	41.82	41.22
EaTR	61.36	45.79	61.86	41.91	41.74
UniVTG	59.74	-	-	-	36.13
UVCOM	65.10	51.81	-	-	45.79
TaskWeave	64.26	50.06	65.39	46.47	45.38
CDTR(Ours)	**68.03**	**52.68**	**66.62**	**46.97**	**45.85**

Table 1: Performance comparison on QVHighlights test split and val split with C+SF features.

Therefore, the final loss of the CDTR can be expressed as:

$$\mathcal{L} = \mathcal{L}_{mr} + \lambda_{hl}\mathcal{L}_{hl} + \lambda_{csa}\mathcal{L}_{csa} + \lambda_{fa}(\mathcal{L}_{via} + \mathcal{L}_{tca}) + \lambda_{sa}\mathcal{L}_{sa}, \quad (14)$$

where λ_{csa},λ_{hl}, λ_{fa}, and λ_{sa} are hyper-parameters.

Experiments

Datasets and Evaluation

Datasets. We evaluate the proposed method on three widely used datasets: QVHighlights (Lei, Berg, and Bansal 2021), Charades-STA (Sigurdsson et al. 2016), and TACoS (Regneri et al. 2013). The QVHighlights dataset consists of 10,148 content-rich YouTube videos, each accompanied by at least one manually annotated text query that highlights specific moments. For a fair evaluation, the test set annotations are inaccessible and predictions must be uploaded to the QVHighlights CodaLab competition platform for assessment. The Charades-STA dataset, derived from the Charades dataset, contains 9,848 videos capturing daily indoor activities along with 16,128 human-tagged query texts. We use 12,408 samples for training and 3,720 for testing. The TACoS dataset includes long-term videos of cooking activities, providing a diverse range of scenarios for evaluation.

Evaluation Metrics. Our evaluation uses Recall@1 (IoU $\in \{0.5, 0.7\}$), mean Average Precision (mAP) at different thresholds, and mean Intersection over Union (mIoU). Recall@1 measures the percentage of top-1 predicted moments

Method	R1 @0.3	R1 @0.5	R1 @0.7	mIoU
2D-TAN	58.76	46.02	27.50	41.25
VSLNet	60.30	42.69	24.14	41.58
M-DETR	65.83	52.07	30.59	45.54
MomentDiff	-	55.57	32.42	-
QD-DETR	-	57.31	32.55	-
UniVTG	70.81	58.01	35.65	50.10
TR-DETR	-	57.61	33.52	-
UVCOM	-	59.25	36.64	-
CDTR(Ours)	**71.16**	**60.39**	**37.24**	**50.65**

Table 2: Comparison on Charades-STA with C+SF features.

Method	Feat	R1 @0.5	R1 @0.7
2D-TAN	VGG	40.94	22.85
DRN	VGG	42.90	23.68
CBLN	VGG	47.94	28.22
FVMR	VGG	42.36	24.14
SSRN	VGG	46.72	27.98
DCM	VGG	47.80	28.00
UMT	VGG	48.31	29.25
QD-DETR	VGG	52.77	31.13
BM-DETR	VGG	56.91	**36.24**
TR-DETR	VGG	53.47	30.81
CDTR(Ours)	VGG	**56.93**	36.20

Table 3: Comparison on Charades-STA with VGG features.

Method	R1 @0.3	R1 @0.5	R1 @0.7	mIoU
2D-TAN	40.01	27.99	12.92	27.22
VSLNet	35.54	23.54	13.15	24.99
M-DETR	37.97	24.67	11.97	25.49
MomentDiff	44.78	33.68	-	-
UniVTG	51.44	34.97	17.35	33.60
UVCOM	-	36.39	23.32	-
CDTR(Ours)	**53.41**	**40.26**	**23.43**	**37.28**

Table 4: Comparison on TACoS with C+SF features.

CD	VIA	TCA	SAT	R1 @0.5	R1 @0.7	mAP
				59.35	44.28	38.05
✓				62.19	47.22	41.06
✓	✓			62.74	48.77	42.51
✓		✓		63.99	49.19	43.02
✓	✓	✓		64.72	50.59	43.57
✓			✓	63.98	48.94	42.65
✓	✓		✓	64.49	50.27	43.01
✓		✓	✓	65.24	50.39	44.52
✓	✓	✓	✓	**68.03**	**52.68**	**45.85**

Table 5: Ablation study of concept decomposition (CD), video intra-modal alignment (VIA), textual-visual cross-modal alignment (TCA), and sequential attention (SAT).

with IoU above a threshold. mAP computes the mean precision across IoU thresholds, and mIoU calculates the average IoU with ground-truth annotations over all test samples.

Implementation Details. Experiments use CLIP + SlowFast (C+SF) (Radford et al. 2021; Feichtenhofer et al. 2019), VGG (Simonyan and Zisserman 2014) for video features, and CLIP (Radford et al. 2021) for query features. The hidden dimension is set to 256. For QVHighlights and Charades-STA, we use a batch size of 32, 300 training epochs, and a learning rate of 1e-4. For TACoS, the batch size is 32, epochs are 300, and the learning rate is 2e-4. Parameters L and K are set to 10 and 5. Add fine-grained loss after training for 60 epochs. The Adam optimizer with weight decay of 1e-4 is used.

Comparison with the State-of-the-Art Methods

The state-of-the-art methods include: 2D-TAN (Zhang et al. 2020b), VSLNet (Zhang et al. 2020a), MCN (Anne Hendricks et al. 2017), CAL (Victor et al. 2019), XML (Lei et al. 2020), XML+ (Lei et al. 2020), DRN (Zeng et al. 2020), CBLN (Liu et al. 2021a), FVMR (Gao and Xu 2021), SSRN (Zhu et al. 2022), DCM (Yang et al. 2021), M-DETR (Lei, Berg, and Bansal 2021), UMT (Liu et al. 2022b), QD-DETR (Moon et al. 2023b), UniVTG (Lin et al. 2023), MomentDiff (Li et al. 2024b), BM-DETR (Jung et al. 2023), MESM (Liu et al. 2024), UVCOM (Xiao et al. 2024), TaskWeave (Yang et al. 2024), and TR-DETR (Sun et al. 2024).

QVHighlights. Our method's performance on the QVHighlights dataset is shown in Table 1. Thanks to a comprehensive understanding of fine-grained concepts in the video, our CDTR model achieves new state-of-the-art performance, showing significant advantages across all metrics. Specifically, on the test split, CDTR surpasses UVCOM by an average of 2.38% across all metrics. On the val split, CDTR also outperforms the TaskWeave in all metrics, particularly exceeding by 3.77% in R1@0.5 and 2.62% in R1@0.7.

Charades-STA. We report the results on the Charades-STA benchmark, as shown in Table 2. CDTR surpasses the state-of-the-art methods using CLIP + SlowFast features across all metrics. Notably, it exceeds UVCOM by 0.80% in R1@0.7. Additionally, we evaluate the performance of our model using VGG features, as shown in Table 3. CDTR achieves state-of-the-art results in R1@0.5 and is the second best in R1@0.7, slightly behind BM-DETR.

TACoS. The performance of the CDTR on the TACoS dataset using CLIP + SlowFast features is shown in Table 4. CDTR achieves significant improvements across all metrics, particularly leading the UVCOM model by 3.87% in R1@0.5. CDTR outperforms existing methods across all metrics, demonstrating its effectiveness in fine-grained alignment and cross-modal matching.

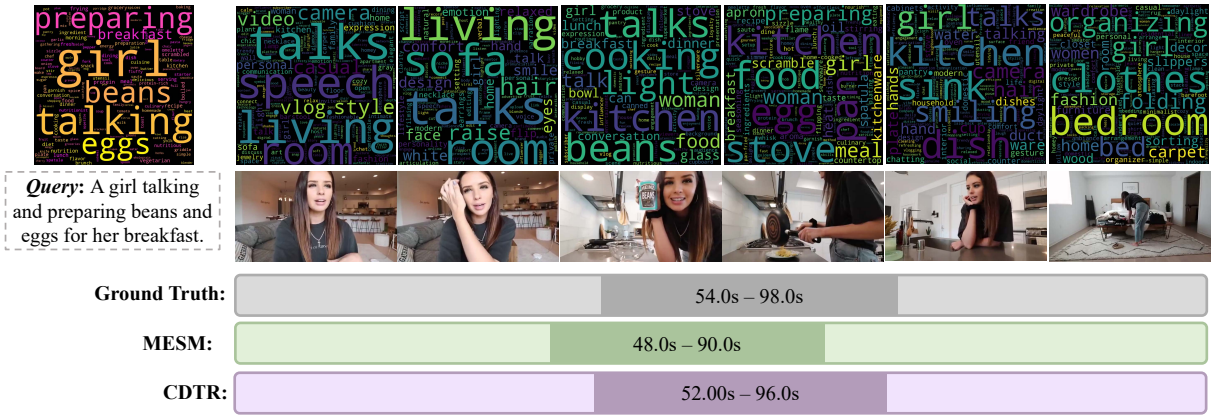

Figure 3: Visualization of results and word clouds for concept representations of query and clips.

Concept Extraction	R1 @0.5	R1 @0.7	mAP
No Extraction	61.98	43.42	39.63
Vector Projection	65.31	51.71	42.40
Hadamard Product	**68.03**	**52.68**	**45.85**

Table 6: Ablation study on concept extraction.

Method	R1 @0.5	R1 @0.7	mAP
w/o Query Relevance	65.12	50.43	43.71
w/o Semantic Mining	66.85	51.78	44.05
Full Model	**68.03**	**52.68**	**45.85**

Table 7: The ablation study of other components.

Ablation Study

To validate the effectiveness of each component, we conducted an ablation study on the QVHighlights val split.

Main Ablation. We constructed several baseline models incorporating different components of our model, with the results presented in Table 5. Specifically, the use of concept decomposition and its associated loss functions resulted in a significant improvement, enhancing the R1@0.5 by 2.84% compared to the baseline. For fine-grained concept alignments such as VIA and TCA, their inclusion also led to substantial performance gains. Notably, cross-modal TCA contributed more to the improvement; for instance, comparing rows six, seven, and eight for R1@0.5, the addition of TCA yielded a 1.26% increase, while VIA added 0.51%. This suggests that fine-grained information in cross-modal contexts is more critical. Sequential attention enhanced the model's focus on relevant content, improving performance. The best results were achieved when all modules were incorporated.

Concept Extraction. In fine-grained concept alignment, we employ the Hadamard product for feature extraction. Table 6 compares methods, including dot product projection and no processing. The Hadamard product enhances semantic concept representation, achieving superior performance. Dot product lacks multi-dimensional representation, leading to weaker results, while no processing causes semantic confusion and significant performance drops.

Other Components. Table 7 reports the impact of query relevance vectors and semantic mining after the video encoder on performance. Clearly, adding query relevance vectors during merging and similarity computation improves accuracy by focusing on relevant semantic concepts. Additionally, semantic mining after the video encoder captures complex temporal semantics in videos, enhancing the accuracy of concept decomposition and overall performance.

Qualitative Analysis

In Fig. 3, we visualize a result and word clouds for concept representations of queries and clips. The proposed model effectively focuses on relevant concepts within both the video and the query. Additionally, the ground truth clips contain more concepts related to the query, demonstrating that the proposed concept decomposition and fine-grained matching effectively decouple complex video semantics. For performance, CDTR aligns closer to the ground truth than MESM.

Conclusion

This paper proposes a novel Concept Decomposition Transformer (CDTR) model for VMR. The CDTR model addresses several critical challenges in existing VMR methods by introducing a semantic concept decomposition module that decouples video moments and sentence queries into independent concept representations. This approach enables the capture of fine-grained semantics, which is crucial for accurate cross-modal alignment. Using these decomposed concept representations as pseudo-labels, our model effectively performs fine-grained concept alignment within the video modality and across modalities, thereby mitigating issues related to semantic overlap and sparse constraints. Through comprehensive experimental analysis of the QVHighlight, Charades-STA, and TACoS datasets, we demonstrated the effectiveness of CDTR in performance.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under grant 62220106008 and 62306067, Sichuan Science and Technology Program under grand 2024NSFSC1463, Sichuan Province Innovative Talent Funding Project for Postdoctoral Fellows with Project BX202311, and China Postdoctoral Science Foundation with Project 2022M720660.

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*, 5803–5812.

Cao, M.; Chen, L.; Shou, M. Z.; Zhang, C.; and Zou, Y. 2021. On Pursuit of Designing Multi-modal Transformer for Video Grounding. In *EMNLP*, 9810–9823.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229.

Chen, S.; Zhao, Y.; Jin, Q.; and Wu, Q. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, 10638–10647.

Chen, Z.; Jiang, X.; Xu, X.; Cao, Z.; Mo, Y.; and Shen, H. T. 2023. Joint Searching and Grounding: Multi-Granularity Video Content Retrieval. In *ACM MM*, 975–983.

Fang, X.; Liu, D.; Zhou, P.; Xu, Z.; and Li, R. 2023. Hierarchical local-global transformer for temporal sentence grounding. *IEEE TMM*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *ICCV*, 6202–6211.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *ICCV*, 5267–5275.

Gao, J.; and Xu, C. 2021. Fast video moment retrieval. In *ICCV*, 1523–1532.

Ge, J.; Xie, H.; Min, S.; and Zhang, Y. 2021. Semantic-guided reinforced region embedding for generalized zero-shot learning. In *AAAI*, volume 35, 1406–1414.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. In *ICLR*.

Jang, J.; Park, J.; Kim, J.; Kwon, H.; and Sohn, K. 2023. Knowing Where to Focus: Event-aware Transformer for Video Grounding. In *ICCV*, 13846–13856.

Jiang, X.; Xu, X.; Zhang, J.; Shen, F.; Cao, Z.; and Shen, H. T. 2022. SDN: Semantic Decoupling Network for Temporal Language Grounding. *IEEE TNNLS*.

Jin, P.; Huang, J.; Xiong, P.; Tian, S.; Liu, C.; Ji, X.; Yuan, L.; and Chen, J. 2023. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *CVPR*, 2472–2482.

Jung, M.; Jang, Y.; Choi, S.; Kim, J.; Kim, J.-H.; and Zhang, B.-T. 2023. Overcoming Weak Visual-Textual Alignment for Video Moment Retrieval. *arXiv preprint arXiv:2306.02728*.

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

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

Lan, X.; Yuan, Y.; Wang, X.; Wang, Z.; and Zhu, W. 2023. A survey on temporal sentence grounding in videos. *ACM Trans. Multimed.*, 19(2): 1–33.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting moments and highlights in videos via natural language queries. *NeurIPS*, 34: 11846–11858.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 447–463.

Li, H.; Cao, M.; Cheng, X.; Li, Y.; Zhu, Z.; and Zou, Y. 2023. G2L: Semantically Aligned and Uniform Video Grounding via Geodesic and Game Theory. In *ICCV*, 12032–12042.

Li, H.; Cao, M.; Cheng, X.; Li, Y.; Zhu, Z.; and Zou, Y. 2024a. Exploiting Auxiliary Caption for Video Grounding. In *AAAI*, volume 38, 18508–18516.

Li, J.; Xie, J.; Qian, L.; Zhu, L.; Tang, S.; Wu, F.; Yang, Y.; Zhuang, Y.; and Wang, X. E. 2022. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *CVPR*, 3032–3041.

Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2024b. Momentdiff: Generative video moment retrieval from random to real. *NeurIPS*, 36.

Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023. Univtg: Towards unified video-language temporal grounding. In *ICCV*, 2794–2804.

Liu, D.; Qu, X.; Di, X.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022a. Memory-guided semantic learning network for temporal sentence grounding. In *AAAI*, volume 36, 1665–1673.

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021a. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*, 11235–11244.

Liu, D.; Qu, X.; and Hu, W. 2022. Reducing the vision and language bias for temporal sentence grounding. In *ACM MM*, 4092–4101.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2021b. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *ICLR*.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022b. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, 3042–3051.

Liu, Z.; Li, J.; Xie, H.; Li, P.; Ge, J.; Liu, S.-A.; and Jin, G. 2024. Towards Balanced Alignment: Modal-Enhanced Semantic Modeling for Video Moment Retrieval. In *AAAI*, volume 38, 3855–3863.

Luo, Z.; Zhao, P.; Xu, C.; Geng, X.; Shen, T.; Tao, C.; Ma, J.; Lin, Q.; and Jiang, D. 2023. Lexlip: Lexicon-bottlenecked language-image pre-training for large-scale image-text sparse retrieval. In *ICCV*, 11206–11217.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, 638–647.

Moon, W.; Hyun, S.; Lee, S.; and Heo, J.-P. 2023a. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *arXiv preprint arXiv:2311.08835*.

Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023b. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*, 23023–23033.

Mun, J.; Cho, M.; and Han, B. 2020. Local-global video-text interactions for temporal grounding. In *CVPR*, 10810–10819.

Nan, G.; Qiao, R.; Xiao, Y.; Liu, J.; Leng, S.; Zhang, H.; and Lu, W. 2021. Interventional video grounding with dual contrastive learning. In *CVPR*, 2765–2775.

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

Qi, Z.; Yuan, Y.; Ruan, X.; Wang, S.; Zhang, W.; and Huang, Q. 2024. Bias-Conflict Sample Synthesis and Adversarial Removal Debias Strategy for Temporal Sentence Grounding in Video. In *AAAI*, volume 38, 4533–4541.

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*, 4280–4288.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *Trans. Assoc. Comput.*, 1: 25–36.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 658–666.

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*, 510–526.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, H.; Zhou, M.; Chen, W.; and Xie, W. 2024. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *AAAI*, volume 38, 4998–5007.

Victor, E.; Mattia, S.; Josef, S.; Bernard, G.; and Bryan, R. 2019. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*.

Wang, X.; Zhu, L.; and Yang, Y. 2021. T2vlad: global-local sequence alignment for text-video retrieval. In *CVPR*, 5079–5088.

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

Wei, J.; Pan, C.; He, S.; Wang, G.; Yang, Y.; and Shen, H. T. 2024a. Towards Robust Person Re-Identification by Adversarial Training with Dynamic Attack Strategy. *IEEE Transactions on Multimedia*.

Wei, J.; Xu, X.; Yang, Y.; Ji, Y.; Wang, Z.; and Shen, H. T. 2020. Universal weighting metric learning for cross-modal matching. In *CVPR*, 13005–13014.

Wei, J.; Yang, Y.; Guan, X.; Xu, X.; Wang, G.; and Shen, H. T. 2024b. Runge-Kutta Guided Feature Augmentation for Few-Sample Learning. *IEEE Transactions on Multimedia*.

Wei, J.; Yang, Y.; Xu, X.; Song, J.; Wang, G.; and Shen, H. T. 2023. Less is better: Exponential loss for cross-modal matching. *IEEE TCSVT*, 33(9): 5271–5280.

Wei, J.; Yang, Y.; Xu, X.; Zhu, X.; and Shen, H. T. 2021a. Universal weighting metric learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6534–6545.

Wei, X.-S.; Shen, Y.; Sun, X.; Ye, H.-J.; and Yang, J. 2021b. A ^{2}-Net: Learning attribute-aware hash codes for large-scale fine-grained image retrieval. *NeurIPS*, 34: 5720–5730.

Xiao, Y.; Luo, Z.; Liu, Y.; Ma, Y.; Bian, H.; Ji, Y.; Yang, Y.; and Li, X. 2024. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *CVPR*, 18709–18719.

Yang, J.; Wei, P.; Li, H.; and Ren, Z. 2024. Task-Driven Exploration: Decoupling and Inter-Task Feedback for Joint Moment Retrieval and Highlight Detection. In *CVPR*, 18308–18318.

Yang, X.; Feng, F.; Ji, W.; Wang, M.; and Chua, T.-S. 2021. Deconfounded video moment retrieval with causal intervention. In *ACM SIGIR*, 1–10.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense regression network for video grounding. In *CVPR*, 10287–10296.

Zeng, Y.; Cao, D.; Wei, X.; Liu, M.; Zhao, Z.; and Qin, Z. 2021. Multi-modal relational graph for cross-modal video moment retrieval. In *CVPR*, 2215–2224.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2023. Temporal sentence grounding in videos: A survey and future directions. *IEEE TPAMI*, 45(8): 10443–10465.

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

Zhu, J.; Liu, D.; Zhou, P.; Di, X.; Cheng, Y.; Yang, S.; Xu, W.; Xu, Z.; Wan, Y.; Sun, L.; et al. 2022. Rethinking the Video Sampling and Reasoning Strategies for Temporal Sentence Grounding. In *EMNLP*, 590–600.