Preprints.org

Article

Not peer-reviewed version

Multi-modal Learning with Dynamic Integration for Video Temporal Localization

Mia Thompson , Lobry Hsu , Liam Harris *

Posted Date: 21 January 2025

doi: 10.20944/preprints202501.1492.v1

Keywords: Multi-modal learning; video temporal grounding; dynamic fusion; self-supervised learning

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Multi-Modal Learning with Dynamic Integration for Video Temporal Localization

Mia Thompson, Lobry Hsu and Liam Harris *

Bond University
* Correspondence: liam.harris@bond.edu.au

Abstract: We investigate the challenging task of text-guided video temporal grounding, which focuses on determining the precise temporal interval of a specific event in a video based on natural language descriptions. Unlike existing methods that primarily rely on RGB visual features, we introduce a novel multi-modal framework designed to extract and integrate complementary information from videos using multiple modalities. Our approach utilizes RGB frames to capture appearance, optical flow to highlight motion dynamics, and depth maps to discern scene structure. While RGB frames provide detailed visual context, they are susceptible to background noise and clutter. Therefore, optical flow is employed to focus on significant motion patterns, and depth maps are incorporated to understand spatial configurations, especially when actions involve objects identifiable by their shapes. To achieve effective integration and cross-modal learning, we propose a dynamic integration mechanism leveraging transformer-based architectures. This mechanism models interactions between modalities and dynamically assigns importance weights to each modality based on contextual relevance. Additionally, we incorporate intra-modal self-supervised learning to enhance feature representations within each modality, promoting robust learning across videos. Extensive experiments conducted on the Charades-STA and ActivityNet Captions datasets validate the superiority of our approach. Our framework, referred to as *MODIFUSE* (Multi-modal Dynamic Integration with Fused Understanding for Scene Events), significantly outperforms state-of-the-art methods, demonstrating its capability to handle complex scenarios with diverse visual and textual content.

Keywords: multi-modal learning; video temporal grounding; dynamic fusion; self-supervised learning

1. Introduction

The proliferation of video data across diverse domains such as entertainment, surveillance, and education has made video understanding an essential research area in computer vision. A particularly important problem in this space is temporal localization, which seeks to identify specific segments within a video based on high-level input, such as natural language descriptions. This capability is fundamental for enabling efficient video indexing, retrieval, and summarization, which are critical for managing the massive amounts of video content generated daily on platforms like YouTube, TikTok, and surveillance systems.

While video understanding is a well-explored field, the integration of multiple modalities, including vision and language, has gained increasing attention in recent years. Tasks like video captioning [17,23] and video question answering [16,18] exemplify the need to process video data in conjunction with textual information [22,26,33,34]. Among these tasks, text-guided video temporal localization is particularly challenging due to its requirement for precise spatiotemporal alignment between video content and natural language queries. This alignment must account for diverse scenes, varying actions, and subtle contextual cues, making it a cornerstone of advanced video analysis.

The primary objective of text-guided video temporal localization is to determine the start and end times of a video segment that corresponds to a given text description. For instance, given a query like "a person closing a door," the system must accurately identify the precise interval where this

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

action occurs. This task has wide-ranging applications, including automatic video editing, intelligent video retrieval, and enhanced human-computer interaction. In surveillance systems, for example, this capability could be used to locate security-critical events within hours of footage based on textual descriptions provided by human operators.

Despite its importance, existing methods face significant challenges due to their reliance on single-modality visual representations, typically RGB frames. While RGB features capture rich appearance information, they are often insufficient for complex real-world scenarios where motion and spatial context play critical roles. For instance, actions like "throwing a ball" may be visually ambiguous without motion information, while scenes involving objects with specific shapes, such as "sitting at a desk," benefit greatly from depth perception.

To address these limitations, multi-modal approaches have been proposed to incorporate additional modalities like optical flow and depth maps. Optical flow, which captures the motion dynamics between consecutive frames, is particularly effective for identifying large-scale actions and transient events. Depth maps, on the other hand, provide spatial and structural context, helping to disambiguate actions that involve subtle movements or distinct object shapes. Together, these modalities complement RGB features and offer a holistic understanding of video content.

However, integrating multiple modalities introduces its own challenges. A naive approach involves using separate streams to process each modality independently, followed by a late fusion of their outputs. While straightforward, this method often fails to capture the complex interdependencies and contextual relationships between modalities. For example, the importance of depth cues may vary depending on the presence of motion, and vice versa. Therefore, a more sophisticated integration strategy is required to dynamically weigh and combine information from different modalities based on the specific context of the query and the video.

In this work, we propose a novel multi-modal framework named *MODIFUSE* (Multi-modal Dynamic Integration with Fused Understanding for Scene Events). *MODIFUSE* is designed to address the shortcomings of existing approaches by introducing a dynamic integration mechanism based on transformer architectures. This mechanism allows for adaptive fusion of RGB, optical flow, and depth features, enabling the model to capture intricate cross-modal interactions and prioritize the most relevant modalities for each video-text pair. Additionally, we incorporate intra-modal self-supervised learning to enhance the consistency and robustness of feature representations within each modality, further improving the overall performance of our approach.

The contributions of this work are motivated by the need for scalable and effective solutions to text-guided video temporal localization in real-world applications. By leveraging the complementary strengths of multiple modalities and introducing novel integration techniques, we aim to push the boundaries of what is achievable in this domain. Our method is extensively validated on benchmark datasets, demonstrating state-of-the-art performance and highlighting its potential for a wide range of practical applications.

2. Related Work

2.1. Multimodal Modeling

Multi-modal learning has emerged as a powerful paradigm for understanding complex phenomena by integrating information from diverse sources. Actions and events are often described by signals spanning multiple modalities, such as vision, language, and audio. The ability to model cross-modal correlations is crucial for achieving a comprehensive understanding of the underlying tasks.

Recent advances in joint vision-language learning have garnered significant attention due to the natural alignment between visual content and textual descriptions [6,7,16,23]. Transformer-based models [8,24], such as BERT [8], have demonstrated exceptional performance in natural language tasks, inspiring their adaptation to multi-modal learning. For example, ViLBERT [20] employs co-attentional transformer layers to model interactions between image and text features, enabling the joint

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

representation of visual and linguistic information. Similarly, other frameworks [5,9,19,21,35] leverage large-scale pretraining on image-text pairs to learn transferable representations for downstream tasks.

Despite the success of these methods, multi-modal learning for video data remains underexplored, particularly in the context of temporal grounding tasks. Videos inherently combine spatial, temporal, and contextual information, making them more complex than static images. To address this, our framework, *MODIFUSE*, introduces a dynamic integration strategy for fusing RGB, optical flow, and depth features. By leveraging co-attentional transformer layers, our approach captures intricate cross-modal relationships and adapts to the specific requirements of text-guided video temporal grounding. Furthermore, we enhance feature learning through self-supervised contrastive techniques, ensuring robust and consistent representations across modalities and videos.

Through these innovations, *MODIFUSE* provides a comprehensive solution to the challenges of multi-modal video understanding, pushing the boundaries of what is achievable in video temporal localization tasks.

2.2. Video Temporal Grounding

Text-guided video temporal grounding is a critical task in video understanding, aiming to identify the temporal segment within a video that corresponds to a natural language query. This involves predicting the start and end times of the relevant video clip based on the semantic alignment between video content and textual descriptions. Existing methods addressing this problem can be broadly categorized into two paradigms: two-stage and one-stage approaches.

Two-stage methods employ a propose-and-rank framework, wherein candidate video segments are first generated as proposals and subsequently ranked based on their relevance to the query. Early methods, such as [10,14], relied on sliding window mechanisms to scan the entire video for generating segment proposals. However, the high computational cost and redundancy of such methods prompted the development of more efficient proposal-generation techniques. For instance, the TGN model [2] integrates frame-by-word interactions to localize proposals in a single pass. Similarly, query-guided proposal generation [31] and semantic activity proposals [3] were introduced to improve proposal relevance. Other models, such as MAN [34], leverage graph architectures to model temporal relationships among proposals, while reinforcement learning-based methods [13,30] aim to optimize video exploration strategies by intelligently skipping irrelevant frames. Despite their promising results, two-stage methods suffer from high computational overhead and dependency on proposal quality, limiting their scalability and performance.

In contrast, one-stage methods bypass the intermediate proposal-generation step and directly predict temporal boundaries by integrating video and query features. These methods predominantly focus on enhancing cross-modal interactions. For example, the ABLR model [32] employs co-attention mechanisms for location regression, while ExCL [11] leverages cross-modal interactions to refine predictions. Similarly, PfTML-GA [26] utilizes query-guided dynamic filters to enhance performance, and DRN [33] incorporates dense supervision to improve training efficiency. The LGI model [22] further advances one-stage methods by decomposing query sentences into semantic phrases and performing hierarchical video-text interactions.

Building on the strengths of LGI, our framework, *MODIFUSE*, incorporates RGB, optical flow, and depth features to capture complementary visual cues. Unlike LGI, which processes only RGB inputs, our approach introduces a novel inter-modality learning mechanism to enhance feature integration. Additionally, we apply contrastive learning to improve intra-modal feature consistency, further boosting the model's robustness and effectiveness.

3. Our MODIFUSE Framework

In this paper, we tackle the problem of text-guided video temporal grounding by introducing a novel multi-modal framework, *MODIFUSE*. Our approach is designed to leverage complementary information from RGB images, optical flow, and depth maps for a more comprehensive understanding of video content. The framework integrates inter-modal and intra-modal feature learning strategies to

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

enhance performance in identifying the temporal boundaries of events described by natural language queries.

Given an input video $V = \{V_t\}_{t=1}^{T}$ with T frames and a natural language query $Q = \{Q_i\}_{i=1}^{N}$ containing N words, the goal is to predict the starting and ending times, $[t_s, t_e]$, of the video segment that best corresponds to the query. To achieve this, our framework comprises the following components: 1. Feature extraction modules for RGB, optical flow, and depth modalities. 2. A textual encoder to represent the query. 3. Inter-modal feature fusion using co-attentional transformers. 4. Intra-modal self-supervised contrastive learning to enhance feature robustness. 5. A regression module to predict temporal boundaries.

3.1. Multi-Modal Feature Extraction

From the input video, we compute depth maps for each frame and optical flow between consecutive frames. Features are then extracted using specialized encoders for each modality:

$$\mathbf{F}_{rgb} = E_r(V), \quad \mathbf{F}_{flow} = E_f(O(V)), \quad \mathbf{F}_{depth} = E_d(D(V)),$$

where E_r, E_f, E_d denote the encoders for RGB, optical flow, and depth, respectively, and $O(\cdot)$ and $D(\cdot)$ represent optical flow and depth operations. The textual encoder E_t processes the query Q to produce $\mathbf{F}_t$, the textual feature representation.

The extracted features are then passed to the local-global interaction (LGI) modules, which incorporate the query information into each modality, generating $\mathbf{M}_{rgb}, \mathbf{M}_{flow}, \mathbf{M}_{depth}$. These serve as the input for inter-modal learning.

3.2. Inter-Modal Feature Fusion

Videos inherently contain spatial, temporal, and motion information distributed across different modalities. A naive approach would average the outputs of each modality, but this fails to account for the varying relevance of modalities across different scenarios. For instance, depth features might be more critical for static scenes, while optical flow is essential for dynamic actions. To address this, we employ a co-attentional transformer-based feature fusion mechanism.

Co-Attentional Transformer.

Inspired by [20], we utilize co-attentional transformer layers to enable cross-modal interactions. For each pair of modalities (e.g., RGB and depth), the features are processed as follows:

$$\mathbf{A}_{rgb,depth} = \text{Softmax}\left(\frac{\mathbf{Q}_{rgb}\mathbf{K}_{depth}^{\top}}{\sqrt{d}}\right)\mathbf{V}_{depth},$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ denote the query, key, and value matrices derived from the respective modality features, and d is the feature dimension. This operation ensures that each modality benefits from the contextual information provided by the others.

The outputs of these layers are dynamically weighted and combined:

$$\mathbf{F}_{fused} = \sum_{m} w_m \mathbf{A}_m,$$

where w_m are weights learned via a fully connected layer, normalized such that $\sum_m w_m = 1$.

3.3. Intra-Modal Contrastive Learning

To improve feature quality within each modality, we incorporate intra-modal contrastive learning. This ensures that features representing similar actions across different videos are closer in the feature space, while dissimilar actions are pushed apart.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

Given a set of positive pairs $(\mathbf{M}, \mathbf{M}_+)$ and negative pairs $(\mathbf{M}, \mathbf{M}_-)$, the contrastive loss is defined as:

$$L_{cl} = -\log \frac{\exp(\text{sim}(\mathbf{M}, \mathbf{M}_+)/\tau)}{\exp(\text{sim}(\mathbf{M}, \mathbf{M}_+)/\tau) + \sum_{-} \exp(\text{sim}(\mathbf{M}, \mathbf{M}_-)/\tau)},$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature parameter.

3.4. Temporal Regression Module

The fused multi-modal feature $\mathbf{F}_{fused}$ is fed into a regression module to predict the temporal boundaries $[t_s, t_e]$. The module includes: 1. Temporal attention to focus on relevant frames. 2. Boundary regression using a combination of linear and non-linear layers:

$$t_s, t_e = \text{REG}(\mathbf{F}_{fused}),$$

where $\text{REG}(\cdot)$ outputs the normalized start and end times.

3.5. Training Objective

The total loss is a combination of supervised and self-supervised components:

$$L = L_{grn} + L_{cl},$$

where L_{grn} comprises location regression, temporal attention guidance, and distinct query attention losses, as described in [22], and L_{cl} is the contrastive loss defined above.

3.6. Implementation Details

We generate optical flow and depth maps using RAFT [27] and MiDaS [25], respectively. The encoders E_r, E_f, E_d are based on I3D [1] for Charades-STA and C3D [28] for ActivityNet Captions. The textual encoder E_t is a bidirectional LSTM, with features obtained by concatenating the final hidden states. Training is performed using the Adam optimizer with a learning rate of 4×10^{-4}. All experiments are implemented in PyTorch, and the code is available at https://github.com/MODIFUSE.

4. Experiments

4.1. Settings

We evaluate the proposed *MODIFUSE* framework against state-of-the-art approaches on two benchmark datasets: Charades-STA [10] and ActivityNet Captions [17].

Charades-STA

This dataset, built upon the Charades dataset, is specifically designed to evaluate the video temporal grounding task. It contains 6,672 videos and 16,128 video-query pairs, split into 12,408 pairs for training and 3,720 pairs for testing. The videos have an average duration of 29.76 seconds, and each video includes 2.4 annotated moments, with an average duration of 8.2 seconds. These features make Charades-STA suitable for evaluating fine-grained temporal grounding capabilities.

ActivityNet Captions

Originally constructed for dense video captioning, the ActivityNet Captions dataset is repurposed for temporal grounding using captions as queries. It includes approximately 20,000 YouTube videos with an average duration of 120 seconds, covering 200 diverse activity categories. Each video contains 3.65 queries on average, with a mean query length of 13.48 words. The dataset is split into training, validation, and testing sets in a 2:1:1 ratio, resulting in 37,421, 17,505, and 17,031 video-query pairs, respectively. As the test set annotations are unavailable, evaluations follow the standard setup of using combined validation sets, denoted as val_1 and val_2.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

Evaluation Metrics

We use two widely adopted metrics to evaluate temporal grounding performance: 1. Recall at Intersection over Union (R@IoU): Measures the percentage of predictions achieving an IoU above a specified threshold with the ground truth. We report results for IoU thresholds {0.3, 0.5, 0.7}. 2. Mean IoU (mIoU): Calculates the average IoU across all predictions, providing a comprehensive performance indicator.

4.2. Overall Performance

Table 1 compares the performance of *MODIFUSE* with state-of-the-art methods, including two-stage approaches reliant on proposal-ranking [10,13,34] and one-stage models leveraging RGB features [12,22,26,33].

Table 1. Comparison with state-of-the-art methods. Results are shown for Charades-STA and ActivityNet Captions. For ActivityNet Captions, some methods report separate results on val_1 and val_2, while others report combined validation results. Single-modal results use only RGB inputs; two-modal results combine RGB and flow.

	Charades-STA				ActivityNet Captions			
Method	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
CTRL [10]	-	21.42	7.15	-	28.70	14.00	20.54	-
RWM [13]	-	36.70	-	-	-	36.90	-	-
MAN [34]	-	46.53	22.72	-	-	-	-	-
TripNet [12]	51.33	36.61	14.50	-	45.42	32.19	13.93	-
PfTML-GA [26]	67.53	52.02	33.74	-	51.28	33.04	19.26	37.78
DRN [33]	-	53.09	31.75	-	-	42.49/45.45	22.25/24.36	-
LGI [22]	72.96	59.46	35.48	51.38	58.52	41.51	23.07	41.13
Single-modal MODIFUSE	73.85	60.79	36.72	52.64	60.25	42.37	25.23	43.18
Two-modal MODIFUSE	74.26	61.93	38.69	53.92	61.80	43.71	26.43	44.82
Three-modal MODIFUSE	**76.68**	**63.03**	**40.15**	**54.89**	**62.91**	**45.72**	**27.79**	**45.86**

Performance Highlights

1. Compared to the baseline LGI [22], our method achieves consistent improvements across all metrics, demonstrating the effectiveness of inter-modal fusion and intra-modal contrastive learning. 2. Our single-modal model, which uses RGB features with contrastive learning, surpasses several prior methods. This indicates the robustness of our intra-modal learning strategy. 3. Incorporating additional modalities (optical flow and depth) into *MODIFUSE* leads to significant performance gains, validating the complementary nature of these modalities. For instance, the three-modal *MODIFUSE* achieves more than 5% improvement in all metrics across both datasets compared to single-modal baselines.

4.3. Ablation Study

We conduct extensive ablation experiments to analyze the contributions of individual components in *MODIFUSE*. Table 2 summarizes the results.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

Table 2. Ablation study on *MODIFUSE* components. Performance impact of removing inter-modal fusion, dynamic weights, and contrastive learning.

	Charades-STA				ActivityNet Captions			
Ablation	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
No Transformer	74.72	61.05	38.26	52.74	61.04	43.83	25.74	43.90
No Dynamic Weights	75.03	61.65	38.78	53.11	61.47	44.42	26.31	44.39
No Contrastive Loss	75.41	61.87	39.02	53.65	61.59	44.50	26.48	44.61
Full *MODIFUSE*	**76.68**	**63.03**	**40.15**	**54.89**	**62.91**	**45.72**	**27.79**	**45.86**

Inter-Modal Feature Fusion

The proposed inter-modal module, which includes co-attentional transformers and dynamic feature fusion, plays a critical role in the framework. Ablations reveal that: 1. Removing co-attentional transformers results in significant performance drops, highlighting their importance in capturing cross-modal interactions. 2. Dynamic fusion with learnable weights ensures that modality contributions are adapted to the input context. Removing this mechanism reduces performance significantly, as shown in Table 2.

Intra-Modal Contrastive Learning

Incorporating contrastive learning enhances feature representations within individual modalities. Removing this component leads to decreased performance, emphasizing its role in improving intra-modal robustness.

5. Conclusions and Future Directions

In this paper, we address the challenging task of text-guided video temporal grounding by introducing a novel multi-modal framework, *MODIFUSE*. Unlike traditional methods that rely solely on RGB features, our approach incorporates three complementary modalities: RGB, optical flow, and depth. This enables a more holistic understanding of video content, particularly in scenarios with cluttered backgrounds, subtle motions, or complex spatial structures.

RGB features, while rich in appearance information, often struggle to disambiguate actions in visually noisy environments. To address this limitation, we integrate optical flow, which captures dynamic motion patterns, and depth maps, which provide structural insights into the scene. These three modalities are effectively combined using our proposed inter-modal feature learning module. This module employs co-attentional transformers to facilitate cross-modal interactions and dynamically fuses the multi-modal features based on context, ensuring that the most relevant modality is prioritized for each query-video pair.

To further improve the training process, we introduce an intra-modal feature learning module, which enhances the robustness of individual modalities through self-supervised contrastive learning. By enforcing features of the same action across different videos to cluster closer in the feature space while pushing apart features from different actions, this module significantly improves intra-modal feature representation quality. The proposed contrastive loss promotes cross-video consistency and complements our inter-modal learning strategy.

We validate the effectiveness of *MODIFUSE* through comprehensive experiments on two benchmark datasets, Charades-STA and ActivityNet Captions. Our results demonstrate substantial improvements over state-of-the-art methods across multiple evaluation metrics. Specifically, we show that the inclusion of optical flow and depth features enhances the ability to recognize motion and structural cues, while the proposed inter- and intra-modal learning modules provide consistent performance gains. These results highlight the strength of our framework in addressing the complexities of text-guided video temporal grounding.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

Future Directions

The proposed *MODIFUSE* framework opens several promising avenues for future research: 1. Extension to Additional Modalities: While this work focuses on RGB, optical flow, and depth, incorporating other modalities such as audio or semantic segmentation could further enhance the understanding of complex videos. 2. Generalization to Longer and More Diverse Videos: Future work could investigate how to scale the framework to handle significantly longer videos with a wider variety of activities and queries, potentially leveraging hierarchical attention mechanisms. 3. Application to Real-Time Systems: Optimizing the computational efficiency of *MODIFUSE* could enable its deployment in real-time video analysis systems for applications like surveillance, sports analysis, and interactive media. 4. Unsupervised and Few-Shot Learning: Exploring the potential of unsupervised or few-shot learning paradigms could further reduce the dependency on large annotated datasets, making the framework more versatile for real-world scenarios. 5. Integration with Large Language Models (LLMs): With the growing capabilities of LLMs, integrating these models into *MODIFUSE* could provide deeper semantic understanding of queries and enhance cross-modal reasoning.

In conclusion, the proposed *MODIFUSE* framework represents a significant step forward in the field of video temporal grounding. By leveraging complementary modalities and advanced learning mechanisms, it sets a robust foundation for future advancements in multi-modal video understanding.

References

1. João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017.
2. Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, 2018.
3. Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, 2019.
4. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.
5. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.
6. Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. Referring expression object segmentation with caption-aware consistency. In *BMVC*, 2019.
7. Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Understanding synonymous referring expressions via contrastive features. *arXiv preprint arXiv:2104.10156*, 2021.
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.
9. Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020.
10. Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017.
11. Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander G. Hauptmann. ExCL: Extractive clip localization using natural language descriptions. In *NAACL*, 2019.
12. Meera Hahn, Asim Kadav, James M. Rehg, and Hans Peter Graf. Tripping through time: Efficient localization of activities in videos. In *BMVC*, 2020.
13. Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, 2019.
14. Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017.
15. Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *ICCV*, 2021.
16. Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D. Yoo. Modality shifting attention network for multi-modal video question answering. In *CVPR*, 2020.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

17. Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.
18. Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: localized, compositional video question answering. In *EMNLP*, 2018.
19. Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.
20. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.
21. Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.
22. Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020.
23. Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*, 2020.
24. Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018.
25. René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *PAMI*, 2020.
26. Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong LI, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *WACV*, 2020.
27. Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.
28. Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.
29. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.
30. Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, 2019.
31. Huijuan Xu, Kun He, Bryan A. Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019.
32. Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019.
33. Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, 2020.
34. Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2019.
35. Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020.
36. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
37. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
38. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
39. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
40. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. doi10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
41. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

42. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
43. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
44. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
45. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
46. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
47. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
48. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
49. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
50. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
51. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
52. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
53. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
54. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
55. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
56. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
57. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
58. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
59. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
60. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
61. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

62. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
63. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
64. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
65. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
66. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
67. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
68. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
69. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
70. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
71. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
72. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
73. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
74. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
75. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
76. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
77. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
78. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
79. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
80. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
81. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 January 2025 doi:10.20944/preprints202501.1492.v1

82. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
83. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
84. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
85. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
86. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
87. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
88. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.
89. Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. 2024. LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding Reasoning and Planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26428–26438.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.