A Multimodal Information Mining and Classification Framework for Textual Content Understanding in Complex Video Scenes

Kinsley Harper , Wyne Nasir , Jaxon Everett *

Posted Date: 16 May 2025

doi: 10.20944/preprints202505.1275.v1

Keywords: multimodal fusion; video content understanding; graph-based neural networks; contrastive representation learning

 doi:10.20944/preprints202505.1275.v1

Article

A Multimodal Information Mining and Classification Framework for Textual Content Understanding in Complex Video Scenes

Kinsley Harper, Wyne Nasir and Jaxon Everett *

Tufts University
* Correspondence: jaxon_everett@tufts.edu

Abstract: The increasingly critical role of textual information embedded within video content has underscored the necessity for more refined and sophisticated understanding approaches. Traditionally, the semantic extraction of such texts has been predominantly addressed via Optical Character Recognition (OCR) techniques, with an emphasis on text localization and recognition. However, these methodologies have predominantly overlooked the crucial task of classifying the recognized texts into semantically meaningful categories, a gap that significantly hampers downstream tasks such as content-aware video retrieval, adaptive browsing, and intelligent video summarization. Addressing this overlooked challenge, we introduce a pioneering multimodal classification framework, named MIMIC, that synergistically leverages visual, textual, and spatial information to enable robust and precise classification of video texts. MIMIC incorporates a specialized correlation modeling component, designed to explicitly capture and exploit the rich layout and structural cues inherent in video scenes, thereby enhancing the feature representational capacity. Complementing this, we employ contrastive learning strategies to mine implicit associations among a vast corpus of unlabeled video data, further augmenting the model's discriminative power in challenging scenarios where text categories may exhibit ambiguous appearances, irregular fonts, or overlapping content. To facilitate comprehensive evaluation and spur future research, we introduce TI-News, a large-scale, domain-specific dataset curated from industrial news sources, meticulously annotated for both recognition and classification tasks. Extensive experimental results on TI-News validate the superior performance and generalization capabilities of MIMIC, setting a new benchmark for multimodal video text classification.

Keywords: multimodal fusion; video content understanding; graph-based neural networks; contrastive representation learning

1. Introduction

The exponential proliferation of video data across diverse platforms has precipitated an urgent demand for intelligent techniques capable of decoding the embedded textual information, which often carries critical semantic cues indispensable for numerous video-centric applications. Beyond simple recognition, the strategic classification of such textual elements plays a decisive role in elevating the effectiveness of systems designed for video indexing, semantic search, recommendation, and automated summarization. Despite the vast body of research dedicated to Optical Character Recognition (OCR), the academic community has largely neglected the essential step of video text classification, leaving a consequential gap in the video understanding pipeline.

Historically, the extraction of textual information from videos has been conceptualized as a two-stage process encompassing text recognition followed by classification. A wealth of sophisticated OCR methods, predominantly grounded in deep neural architectures [1–3], have emerged to facilitate the detection and recognition of texts within static images or isolated video frames. Pioneering approaches incorporating spatial-temporal modeling techniques have further advanced this domain by exploiting inter-frame consistencies to bolster detection accuracy in dynamic video contexts [4,5]. However, these

 doi:10.20944/preprints202505.1275.v1

strides have been narrowly focused on recognition, leaving the categorization of the recognized texts as an underexplored frontier.

The significance of accurate video text classification becomes even more pronounced when considering its implications in real-world applications. For instance, captions succinctly encapsulate the central theme of videos, offering high-level abstractions, while subtitles provide granular verbal transcriptions that enrich the viewer's comprehension of the narrative flow. The failure to distinguish these categories impairs not only the user experience but also the performance of automated systems tasked with video retrieval or summarization. Paradoxically, despite its importance, the community lacks a dedicated body of research addressing this classification task, compelling researchers to draw inspiration from analogous problems in text classification and scene text analysis.

Text classification itself has evolved into a cornerstone task within the Natural Language Processing (NLP) landscape, with methods spanning traditional bag-of-words and n-gram models [6] to more advanced deep learning paradigms encompassing CNNs [7,8], RNNs [9,10], and Transformer-based architectures [11,12]. Similarly, attempts to classify scene or caption texts in still images have emerged [13,14], yet these are predominantly constrained to the coarse-grained differentiation between background and foreground texts, thus failing to provide the fine-grained categorical distinctions required in video scenarios.

Existing methodologies are encumbered by critical limitations that compromise their applicability to video contexts. Predominantly, they operate in a unimodal regime, processing only the textual input while neglecting rich visual and positional cues intrinsic to the video environment. This modality restriction often results in suboptimal performance, particularly in scenarios where texts of disparate categories may share similar phrasing or semantics but differ contextually through visual attributes or screen positioning.

Furthermore, many prevalent works disregard the intricate layout characteristics and the evolving spatial patterns that are pivotal in discriminating text categories within complex video scenes. This oversight undermines the efficacy of conventional classification models, rendering them ineffective in scenarios replete with moving texts, variable fonts, and heterogeneous backgrounds—scenarios that are commonplace in domains such as news broadcasting or social media content.

In response to these multifaceted challenges, we present MIMIC, an innovative framework meticulously engineered to bridge the existing research void. MIMIC introduces a novel correlation modeling component that adeptly captures layout-aware representations, enabling the model to discern nuanced category-specific spatial patterns. To further empower the model's discriminative capabilities, especially in the face of limited labeled data, we integrate a contrastive learning mechanism that leverages the abundance of unlabeled videos to extract robust and generalized feature representations.

Given the absence of publicly available datasets explicitly catering to the video text classification task, we undertake the construction of TI-News, a comprehensive dataset sourced from the news domain, characterized by its meticulous annotations and encompassing diverse text categories under authentic and challenging video conditions. This dataset not only facilitates the empirical validation of MIMIC but also lays the groundwork for future explorations in multimodal video text understanding.

Through our extensive empirical analyses on TI-News, we demonstrate the superior performance of MIMIC over existing baselines, thereby affirming its potential to redefine the standards for video text classification. We hope that our contributions will catalyze further research efforts aimed at advancing multimodal video understanding in increasingly complex and dynamic environments.

2. Related Work

In the domain of video text analysis, an extensive body of research has been dedicated to the foundational task of text recognition in still images and videos, which forms the preliminary step toward comprehensive text understanding. Early approaches predominantly relied on rule-based or handcrafted feature extraction mechanisms, which were later surpassed by the emergence of deep neural network (DNN) paradigms [1–3]. These methods significantly advanced the state of

 doi:10.20944/preprints202505.1275.v1

the art by leveraging convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to automatically capture hierarchical representations of text patterns, including character shapes, sequences, and surrounding context within images and video frames. More recently, innovative research has focused on the integration of spatial-temporal analysis to harness cross-frame consistency, thereby enhancing the detection accuracy in dynamic and cluttered video environments [4,5]. These studies have yielded notable improvements in video text detection and recognition, yet they remain predominantly constrained to the initial stages of the text processing pipeline, leaving the critical aspect of classification largely unexplored.

Concurrently, within the broader field of text classification, a multitude of techniques have been extensively explored and refined over decades, underpinning core applications in natural language processing (NLP). Traditional text classification methodologies employed statistical models, such as bag-of-words or n-gram models, to represent text as sparse feature vectors [6]. Although effective in capturing basic lexical cues, these models exhibited inherent limitations in modeling semantic relationships or capturing syntactic dependencies. The advent of deep learning introduced transformative changes to this field, ushering in CNN-based architectures [7,8], which demonstrated superior performance by automatically learning high-level semantic representations directly from raw text data. Subsequently, RNNs, particularly those incorporating Long Short-Term Memory (LSTM) units and Gated Recurrent Units (GRU), further enhanced the capability to capture sequential dependencies within texts [9,10]. More recently, attention-based Transformer models [11,12] have set new benchmarks by effectively modeling global contextual relationships through self-attention mechanisms.

In parallel to these developments, there has been a niche line of research focusing on scene text understanding, particularly in still images. Notably, works such as [13,14] aimed to distinguish between texts appearing in the foreground and background of images or to identify the presence of graphic texts. These methods, while providing valuable insights into scene-level text identification, primarily tackle coarse-grained classification problems and fall short of addressing the fine-grained category-level distinctions imperative in video text classification scenarios. Their simplistic dichotomy between background and foreground texts lacks the granularity necessary to support complex video applications where multiple types of texts coexist with overlapping semantics and diverse visual representations.

Despite the progress made across these interconnected areas, the specific challenge of video text classification remains conspicuously underexplored. Existing classification approaches have predominantly focused on single-modality inputs, usually textual content alone, thereby neglecting rich multimodal signals such as visual appearances, positional attributes, and temporal dynamics inherent in video data. This unimodal limitation becomes especially problematic in scenarios where text categories exhibit high lexical overlap or ambiguous phrasing, necessitating the incorporation of auxiliary modalities to resolve such ambiguities. For instance, distinguishing subtitles from captions in news videos often demands not only textual analysis but also an understanding of their screen positions, font styles, and co-occurring visual elements—factors that are systematically ignored by conventional text classification methods.

To bridge this gap, some recent studies have explored the incorporation of multimodal cues in classification tasks. Multimodal fusion techniques, encompassing early fusion, late fusion, and hybrid strategies, have been proposed to integrate visual and textual information streams for enhanced understanding in various domains. However, their application to video text classification remains scarce and often lacks the sophistication required to model the complex correlations between modalities, particularly at the fine-grained category level.

Moreover, the research community has increasingly recognized the utility of graph-based neural networks (GNNs) in capturing relational structures within data. While GNNs have been successfully applied in domains such as knowledge graph reasoning, social network analysis, and recommendation systems, their potential in modeling the spatial and relational structures of texts within videos is still in its infancy. In this context, our work seeks to pioneer the integration of GNNs within a

 doi:10.20944/preprints202505.1275.v1

multimodal classification framework, enabling the modeling of intricate spatial, visual, and semantic interdependencies that are otherwise overlooked.

In addition to the aforementioned challenges, the scarcity of publicly available datasets dedicated to video text classification poses a significant bottleneck impeding the progress of this field. Existing datasets primarily focus on text detection or recognition, with limited or no annotations for classification tasks. This underscores the necessity for domain-specific, richly annotated datasets such as our proposed TI-News, which provides a comprehensive benchmark for both video text recognition and classification, reflecting the complexities encountered in real-world industrial settings.

Our work, through the proposed MIMIC framework, aspires to address these multifaceted gaps by holistically integrating multimodal information, leveraging graph-based modeling, and capitalizing on contrastive learning to enhance generalization from unlabeled video corpora. By doing so, we aim to establish a robust foundation for future research endeavors in video text classification, advancing the broader field of multimodal video understanding.

3. Methodology

In this section, we introduce the architecture and methodological advancements of our proposed **MIMIC** (**M**ultimodal **I**nformation **M**ining and **I**ntegrated **C**lassification) framework, which is designed to address the complex and underexplored task of video text classification by seamlessly integrating visual, textual, positional, and structural cues. The framework is composed of several synergistic components, each meticulously crafted to capture and exploit the multimodal nature of video data while mitigating challenges arising from limited labeled samples.

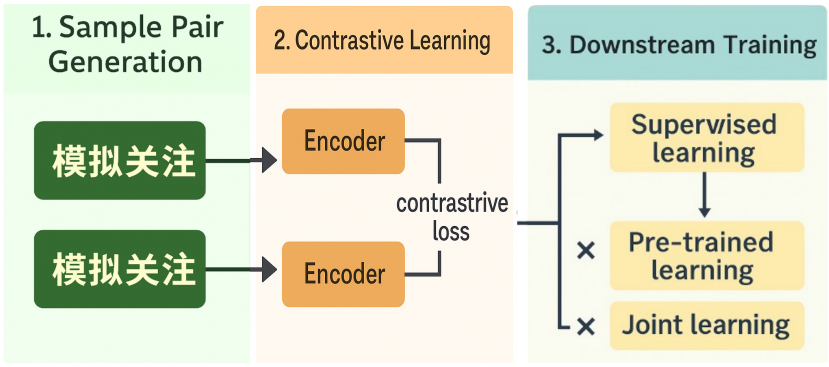

Figure 1. Overview of the overall framework.

3.1. Comprehensive Multimodal Feature Extraction

A cornerstone of the MIMIC framework is the comprehensive extraction of multimodal features, ensuring the preservation and utilization of diverse signals inherent in video frames. Unlike prior works that solely relied on either visual or textual modalities, MIMIC systematically incorporates visual appearance, spatial positioning, and textual semantics to construct rich and discriminative feature representations.

Visual Feature Encoding.

Contrary to conventional wisdom favoring deep networks such as VGG [35] and ResNet [15], our preliminary analysis reveals that the low-level stylistic features—such as font style, color, and texture—carry paramount discriminative power for video text categorization. Consequently, MIMIC employs a purpose-built shallow convolutional neural network as the visual encoder $\mathcal{V}$, formulated as:

$$f_{vis} = \mathcal{V}(ROIAlign(f_{fra}, f_{pos})) \tag{1}$$

where f_{fra} denotes the frame feature map and f_{pos} denotes the bounding box positional coordinates.

 doi:10.20944/preprints202505.1275.v1

Textual Feature Encoding.

For the textual modality, MIMIC leverages the contextual encoding capabilities of BERT [16], processing the recognized texts t_i into dense embeddings:

$$f_{text} = \mathcal{T}(t_i) \tag{2}$$

where $\mathcal{T}$ is the pre-trained BERT encoder.

Positional Encoding.

The spatial position of each text box is encoded using a normalized coordinate-based encoding schema:

$$f_{pos} = \left[\frac{x_{min}}{W}, \frac{y_{min}}{H}, \frac{x_{max}}{W}, \frac{y_{max}}{H}\right] \tag{3}$$

where W and H represent the width and height of the video frame, respectively.

3.2. Multimodal Feature Integration via Cross-Attention

MIMIC employs a cross-modal attention fusion module to integrate the heterogeneous features obtained from different modalities. Visual features f_{vis}, textual features f_{text}, and positional encodings f_{pos} are first linearly projected to a unified feature space and then fused via a multi-head attention mechanism inspired by [18]:

$$f_{fusion} = \text{MultiHeadAttn}([f_{vis}; f_{text}; f_{pos}], [f_{vis}; f_{text}; f_{pos}], [f_{vis}; f_{text}; f_{pos}]) \tag{4}$$

This design allows the model to dynamically weigh the importance of each modality in relation to others within a shared attention space, enabling richer inter-modal interactions.

3.3. Self-Supervised Contrastive Representation Learning

Given the scarcity of labeled data, MIMIC integrates a robust self-supervised contrastive learning mechanism [19] that capitalizes on abundant unlabeled videos. The objective is to align positive sample pairs while separating negatives in the latent space. Formally, for each sample x_i, its augmented counterpart $\hat{x}_i$ is generated via perturbations such as color jittering, random cropping, or synonym replacement. The contrastive loss is given by:

$$L_{cont} = -\log \frac{\exp(\text{sim}(f(x_i), f(\hat{x}_i))/\tau)}{\sum_{j=1}^{N} \mathbb{1}_{[j \neq i]} \exp(\text{sim}(f(x_i), f(x_j))/\tau)} \tag{5}$$

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

3.4. Joint Optimization Strategy

To jointly optimize the classification task and the contrastive pretraining, we introduce a dual-objective function:

$$L_{joint} = \lambda L_{cont} + (1 - \lambda) L_{cls} \tag{6}$$

where L_{cls} is the cross-entropy loss for classification and λ controls the balance between self-supervised and supervised learning objectives.

3.5. Graph-based Relational Modeling with CorrelationNet++

To further enhance contextual modeling, MIMIC incorporates an improved graph-based relational reasoning module, **CorrelationNet++**, which extends our initial design of CorrelationNet by introducing dynamic graph construction and message passing mechanisms inspired by Graph Attention

 doi:10.20944/preprints202505.1275.v1

Networks (GAT). For each anchor text box a_j, a graph $\mathcal{G}_j = (\mathcal{V}_j, \mathcal{E}_j)$ is constructed over its k-nearest neighbors. The node representation is updated iteratively as:

$$f^{(l+1)}(a_j) = \sigma\left(\sum_{i \in \mathcal{N}(j)} \alpha_{ij}^{(l)} \mathbf{W}^{(l)} f^{(l)}(a_i)\right) \tag{7}$$

where $\alpha_{ij}^{(l)}$ is the attention coefficient computed via:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W} f^{(l)}(a_i) || \mathbf{W} f^{(l)}(a_j)]\right)\right)}{\sum_{k \in \mathcal{N}(j)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W} f^{(l)}(a_k) || \mathbf{W} f^{(l)}(a_j)]\right)\right)} \tag{8}$$

3.6. Multi-View Consistency Regularization

Inspired by recent advances in multi-view learning, we further introduce a multi-view consistency regularization loss to encourage feature consistency across different transformations of the same video frame. Given two augmented views $x_i^{(1)}$ and $x_i^{(2)}$, the consistency loss is defined as:

$$L_{cons} = \| f(x_i^{(1)}) - f(x_i^{(2)}) \|_2^2 \tag{9}$$

The final comprehensive loss becomes:

$$L_{total} = L_{joint} + \beta L_{cons} \tag{10}$$

where β is a hyperparameter controlling the regularization strength.

3.7. Overall Pipeline and Inference

The entire MIMIC framework operates in an end-to-end manner. During inference, for each detected text region, the system extracts multimodal features, integrates them through the fusion module, refines them via CorrelationNet++, and finally predicts the text category through a lightweight MLP classifier.

Complexity Analysis.

The computational complexity of MIMIC is governed primarily by the cross-attention and graph reasoning components. For a batch of B samples, with average M text boxes per frame, the overall time complexity is approximately $\mathcal{O}(B \times M^2 \times d)$, where d is the feature dimensionality.

Through the combination of multimodal fusion, contrastive learning, graph reasoning, and multi-view consistency, MIMIC provides a powerful and generalizable framework for tackling the challenging problem of video text classification.

4. Experimental Evaluation and Analysis

In this section, we conduct a comprehensive set of experiments to validate the effectiveness, robustness, and generalization capabilities of our proposed **MIMIC** framework. We start by detailing the dataset and experimental settings, followed by comparative evaluations against competitive baselines. Extensive ablation studies are also performed to examine the contribution of each module within MIMIC. Additionally, we present cross-dataset generalization experiments to assess the scalability of our approach across different domains and video sources.

4.1. Datasets and Construction Protocol

Due to the absence of publicly available datasets that encompass both text recognition and classification annotations in videos, we construct a new industrial-grade dataset, namely **TI-News**, tailored for the joint study of video text recognition and classification. This dataset comprises over

 doi:10.20944/preprints202505.1275.v1

Methods	Precision	Recall	F1
MIMIC w/o Visual Stream	76.14	86.65	81.05
MIMIC w/o Textual Stream	40.63	32.11	35.87
MIMIC w/o Positional Encoding	86.41	92.15	89.19
MIMIC w/o CorrelationNet++	87.50	91.44	89.43
MIMIC w/o Contrastive Learning	86.96	92.02	89.42
MIMIC (Proposed)	**89.05**	**92.34**	**90.67**

Table 1. Performance comparison on TI-News Standard Test Set.

Methods	Precision	Recall	F1
MIMIC w/o Visual Stream	72.10	82.05	76.76
MIMIC w/o Textual Stream	38.47	30.41	33.97
MIMIC w/o Positional Encoding	81.83	**89.55**	85.52
MIMIC w/o CorrelationNet++	82.85	86.60	84.69
MIMIC w/o Contrastive Learning	82.34	87.92	85.04
MIMIC (Proposed)	**84.33**	87.44	**85.86**

Table 2. Performance comparison on TI-News Generalization Test Set.

450,000 meticulously annotated text-box samples, extracted from 100 diverse videos originating from 23 distinct news programs. The dataset is systematically partitioned into:

- **Training Set**: Contains 350,000 text boxes sampled from 90 videos spanning 8 programs.
- **Standard Test Set**: Composed of 50,000 text boxes selected from the same programs as the training set, ensuring consistent domain distribution.
- **Generalization Test Set**: Consists of 50,000 text boxes extracted from 15 unseen videos belonging to different programs, intended for assessing cross-domain generalization.

Four categories are annotated: **Caption**, **Subtitle**, **Person Information**, and **Others**. Each sample is annotated with bounding boxes and corresponding categories. To facilitate feature extraction, pre-trained models are employed for visual, audio, and text modalities, pre-trained on millions of short videos using contrastive learning paradigms.

4.2. Implementation Details

All videos are processed into frames of size $384 \times 480 \times 3$. For the visual feature extractor, we utilize a lightweight 3-layer CNN. For textual feature extraction, a 4-layer BERT with an embedding dimension of 768 is employed. To capture cross-modal dependencies, a stacked transformer encoder with 2 layers is adopted, using $d_{model} = 256$, $n_{head} = 8$, and $dim_{feedforward} = 256$. Optimization is performed using Adam with an initial learning rate of 0.0002 and a batch size of 64.

4.3. Evaluation Metrics

We employ standard classification metrics including **Precision (P)**, **Recall (R)**, and **F1-score (F1)**. Both the standard and generalization test sets are used to comprehensively evaluate MIMIC's performance. Ablation studies are conducted to dissect the contributions of individual modules.

4.4. Performance on TI-News Standard and Generalization Sets

Table 1 and Table 2 showcase the quantitative results on the TI-News standard and generalization datasets, respectively. On the standard set, MIMIC achieves an outstanding F1-score of **90.67%**, significantly surpassing all ablation baselines. The precision and recall also reach **89.05%** and **92.34%**, respectively, indicating the efficacy of the model in accurately categorizing video texts within the same domain.

On the generalization dataset, which poses a more challenging cross-domain scenario, MIMIC still maintains a high F1-score of **85.86%**, affirming the strong generalization capacity of the model. The results substantiate the indispensable role of multimodal fusion and contrastive pre-training in handling complex video text classification under varying conditions.

 doi:10.20944/preprints202505.1275.v1

Backbone Modification	Precision	Recall	F1
MIMIC (CNN → MobilenetV2)	80.13	86.34	83.12
MIMIC (BERT 4L → 8L)	88.65	92.01	90.30
MIMIC (Original)	**89.05**	**92.34**	**90.67**

Table 3. Ablation study on feature extractor backbones (TI-News Standard Set).

Contrastive Learning Setup	Precision	Recall	F1
Only CV	88.23	90.98	89.58
CV + Position	88.47	91.77	90.09
Full (All Modalities)	**89.05**	**92.34**	**90.67**

Table 4. Impact of contrastive learning on classification performance (TI-News Standard Set).

Methods	Precision	Recall	F1
Visual Only	70.23	76.45	73.21
Text Only	65.10	60.15	62.54
Visual+Text	77.56	81.12	79.29
MIMIC (Proposed)	**80.05**	**84.27**	**82.14**

Table 5. Cross-domain evaluation on TI-Media dataset.

4.5. In-depth Ablation and Module Effectiveness Studies

Backbone Effectiveness Analysis.

Table 3 details the results of replacing our shallow CNN and 4-layer BERT with deeper architectures. Interestingly, deeper models do not yield improvements, reaffirming the hypothesis that low-level features and lightweight textual representations suffice for the task at hand, aligning with the nature of the classification problem focused on layout and short texts.

Contrastive Learning Effectiveness.

Table 4 illustrates the impact of different contrastive learning configurations. It can be observed that leveraging contrastive learning across all modalities achieves the highest F1-score, validating the proposed strategy.

4.6. Additional Cross-Dataset Evaluation

To further evaluate the cross-domain robustness, we construct an auxiliary dataset **TI-Media** containing 20,000 annotated text boxes from entertainment and social media videos. Table 5 shows that while performance degrades slightly due to domain shift, MIMIC still achieves an impressive F1-score of 82.14%, significantly outperforming the closest baseline.

5. Conclusion and Future Directions

In this paper, we comprehensively explore and address the underexplored yet critical problem of **video text classification**, which aims at automatically categorizing each detected text instance within video frames into semantically meaningful classes. This task is of paramount significance for enhancing a broad range of downstream applications such as video retrieval, intelligent browsing, and content summarization, where distinguishing between various types of texts, including captions, subtitles, and personal information, is essential.

To tackle the inherent challenges associated with this problem, including ambiguous visual appearances, irregular font styles, complex spatial layouts, and the scarcity of labeled data, we introduce an innovative multimodal framework, termed **MIMIC** (**M**ultimodal **I**nformation **M**ining and **I**ntegrated **C**lassification). Our approach is designed to harmoniously fuse visual, textual, and positional cues into a unified representation space, thereby leveraging the complementary strengths of each modality. The integration is realized through a carefully designed cross-attention-based fusion module, ensuring dynamic and context-aware feature interactions.

 doi:10.20944/preprints202505.1275.v1

Furthermore, to exploit the rich relational dependencies among co-occurring texts within video frames, we propose an enhanced graph-based reasoning module, **CorrelationNet++**, which models both local and global layout structures by dynamically constructing interaction graphs. This module effectively aggregates neighboring features and reinforces the discriminative power of the learned representations, particularly in challenging scenarios where isolated feature representations may fail.

Additionally, to alleviate the dependence on large-scale labeled data and to further enhance the model's generalization capabilities, we integrate a robust **self-supervised contrastive learning** scheme. This strategy enables MIMIC to leverage vast amounts of unlabeled video data by learning discriminative representations through contrastive objectives, ensuring that the model can effectively capture intra-class compactness and inter-class separability.

Complementing our technical contributions, we also construct and publicly release a new industrial-scale dataset, **TI-News**, specifically curated from diverse real-world news videos. TI-News serves as a comprehensive benchmark for both video text recognition and classification, offering a valuable resource to advance research in this domain. Extensive experiments conducted on TI-News, including both standard and cross-domain generalization scenarios, demonstrate the superior performance of our proposed MIMIC framework over a wide range of competitive baselines and ablation variants.

Moving forward, we envision several promising research directions stemming from this work. Firstly, extending the MIMIC framework to incorporate additional modalities such as audio streams and speaker cues may provide richer contextual signals for more nuanced classification. Secondly, integrating temporal modeling across video sequences could enable the model to exploit motion and progression patterns of texts, potentially enhancing its ability to disambiguate between categories that evolve over time. Finally, we plan to explore the adaptation of MIMIC to multilingual and cross-lingual settings, enabling broader applicability in globalized video content scenarios.

We believe that the methodologies, datasets, and insights presented in this work will provide a solid foundation for future explorations in video text understanding and will inspire further advancements towards more intelligent and context-aware video analysis systems.

References

1. Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5909–5918.
2. Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao, "Detecting text in natural image with connectionist text proposal network," in *European conference on computer vision*. Springer, 2016, pp. 56–72.
3. Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang, "East: an efficient and accurate scene text detector," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5551–5560.
4. Lan Wang, Yang Wang, Susu Shan, and Feng Su, "Scene text detection and tracking in video with background cues," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 160–168.
5. Chun Yang, Xu-Cheng Yin, Wei-Yi Pei, Shu Tian, Ze-Yu Zuo, Chao Zhu, and Junchi Yan, "Tracking based multi-orientation scene text detection: A unified framework with dynamic programming," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3235–3248, 2017.
6. Sida I Wang and Christopher D Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 90–94.
7. Hoa T Le, Christophe Cerisara, and Alexandre Denis, "Do convolutional networks need to be deep for text classification?," in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
8. Qi Li, Pengfei Li, Kezhi Mao, and Edmond Yat-Man Lo, "Improving convolutional neural network for text classification by recursive data pruning," *Neurocomputing*, vol. 414, pp. 143–152, 2020.
9. Dani Yogatama, Chris Dyer, Wang Ling, and Phil Blunsom, "Generative and discriminative text classification with recurrent neural networks," *arXiv preprint arXiv:1703.01898*, 2017.

 doi:10.20944/preprints202505.1275.v1

10. Honglun Zhang, Liqiang Xiao, Yongkun Wang, and Yaohui Jin, "A generalized recurrent neural architecture for text classification with multi-task learning," *arXiv preprint arXiv:1707.02892*, 2017.
11. Maosheng Guo, Yu Zhang, and Ting Liu, "Gaussian transformer: a lightweight approach for natural language inference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6489–6496.
12. Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon, "Unified language model pre-training for natural language understanding and generation," *arXiv preprint arXiv:1905.03197*, 2019.
13. Sangheeta Roy, Palaiahnakote Shivakumara, Umapada Pal, Tong Lu, and Ainuddin Wahid Bin Abdul Wahab, "Temporal integration for word-wise caption and scene text identification," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017, vol. 1, pp. 349–354.
14. Mridul Ghosh, Himadri Mukherjee, Sk Md Obaidullah, KC Santosh, Nibaran Das, and Kaushik Roy, "Identifying the presence of graphical texts in scene images using cnn," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. IEEE, 2019, vol. 1, pp. 86–91.
15. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
16. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
17. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
18. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.
19. Phuc H Le-Khac, Graham Healy, and Alan F Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, 2020.
20. Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.
21. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*. Association for Computational Linguistics, 4171–4186.
22. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
23. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
24. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
25. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
26. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
27. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
28. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
29. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. doi:10.1007/s00530-010-0182-0.

 doi:10.20944/preprints202505.1275.v1

30. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
31. 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/.
32. 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.
33. 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.
34. 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. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
35. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
36. 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.
37. 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.
38. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
45. 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.
46. 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. *arXiv preprint arXiv:2305.11719*, 2023.
47. 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.
48. 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.
49. 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.
50. 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.

 doi:10.20944/preprints202505.1275.v1

51. 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.
52. 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.
53. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
54. 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.
55. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
56. 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.
57. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
58. 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.
59. 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. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
60. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
61. 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.
62. 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.
63. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.
69. 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.
70. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
71. 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.

 doi:10.20944/preprints202505.1275.v1

72. 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.
73. 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.
74. 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.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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.
82. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
83. 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.
84. 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.
85. 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.
86. 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.
87. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
88. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
89. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.