Unified Representation Learning for Relation Extraction in Visually-Rich Documents

Olivia Bennett , Emily Marwood , Noah Carter *

Posted Date: 21 February 2025

doi: 10.20944/preprints202502.1700.v1

Keywords: relation extraction; multimodal representation learning; unified feature fusion; document understanding

 doi:10.20944/preprints202502.1700.v1

Article

Unified Representation Learning for Relation Extraction in Visually-Rich Documents

Olivia Bennett, Emily Marwood and Noah Carter *

Flinders University
* Correspondence: noah.carter@flinders.edu.au

Abstract: Recent advances in multimodal integration—combining text, geometric layout, and visual cues—have led to significant improvements in the field of Visually-rich Document Understanding (VrDU), particularly in relation extraction (RE) tasks. In our work, we introduce a novel approach, termed **UniFusion**, which reimagines the process of joint representation learning through a comprehensive analysis of each modality's contribution. Our experiments systematically exclude each data type in turn, and we also evaluate text and layout modalities in isolation, thereby providing a detailed account of the predictive capacity inherent in each signal. In our extensive study, we demonstrate that a bimodal configuration integrating textual content with layout geometry consistently outperforms other configurations, achieving an F1 score of 0.684. This observation underscores the pivotal role of textual information as the primary driver in predicting entity relationships. However, our analysis further reveals that geometric layout features, when used as a unimodal predictor, offer substantial predictive power and can serve as an effective standalone approach under certain circumstances. Although the visual modality, when used in isolation, exhibits relatively lower performance, our results indicate that its inclusion in a multimodal fusion strategy can enhance overall performance by providing supplementary contextual information. Moreover, our experiments span a diverse array of document types and noise conditions, confirming that the integration of multiple modalities via UniFusion leads to more robust performance, particularly in scenarios with incomplete or noisy textual data. In summary, our findings provide compelling evidence for the efficacy of joint representation learning, demonstrating that a carefully balanced fusion of text, layout, and visual modalities is essential for advancing the state-of-the-art in RE tasks within the VrDU framework.

Keywords: relation extraction; multimodal representation learning; unified feature fusion; document understanding

1. Introduction

In recent years, a wide range of industries—spanning healthcare, insurance, e-commerce, and beyond—have increasingly turned to digitization and artificial intelligence to harness valuable information embedded in documents. This trend has given rise to the field of Visually-rich Document Understanding (VrDU) [11,14,21,24], which focuses on the extraction of structured knowledge from scanned or digital documents for practical downstream applications. VrDU encompasses several critical sub-fields such as Named-Entity Recognition (NER) [2], layout comprehension [7], and document classification [22]. Among these, relation extraction (RE) plays a pivotal role by establishing and identifying meaningful connections between recognized entities. Typically, RE is framed in a question-answer (*Q/A*) paradigm where the task is to design a function that determines whether a pair of entities in a document share a relevant relationship [11,23].

Parallel to these developments, advances in multimodal deep learning have revolutionized several fields including medical imaging [16], neurotechnology [4], and early diagnosis of neurological disorders such as Alzheimer's disease [17]. The availability of robust optical character recognition

 doi:10.20944/preprints202502.1700.v1

(OCR) systems—such as AWS Textract[1], Microsoft Read API[2], and PyTesseract[3]—has further facilitated the extraction of detailed textual and spatial information from visually complex documents. These capabilities have catalyzed the development of a wide array of multimodal architectures that aim to fuse textual, geometric, and visual features for enriched document understanding [12,14,19,21,22,24]. In particular, these approaches promote end-to-end learning of joint representations to capture the full spectrum of information inherent in documents. Although transformer-based architectures have become highly popular in this domain [13,23], alternative methodologies, such as graph neural networks [3,6], have also shown promise in optimizing RE tasks by exploiting complex structural dependencies.

Despite significant progress in the development of multimodal systems for VrDU, there remains considerable ambiguity regarding the relative contributions of the different modalities. While text is intuitively expected to be the dominant predictor for entity relationships, the true extent of its superiority over geometric layout or visual data—and the interplay among these modalities—has not been definitively characterized. Notably, even studies that incorporate ablation tests rarely examine scenarios where text representations are completely omitted [9]. This gap in the literature motivates our systematic investigation into the predictive capacities of each modality, as well as their combined impact on relation extraction performance.

To address these challenges, we conduct a rigorous set of experiments using a variety of multimodal and unimodal configurations. Our study focuses on the RE task because it represents a fundamental yet unresolved problem in information extraction, with significant implications for industry applications. In our work, we leverage the novel DocFusion framework, which embodies an innovative approach to joint representation training. Through extensive empirical evaluations, we aim to (1) validate the benefits of learning joint representations for VrDU-based relation extraction, (2) quantify the asymmetric predictive capacities of text, layout, and visual modalities—thereby confirming that textual data holds a dominant role while also recognizing the supportive influence of layout and visual cues, and (3) introduce an optimized and simplified classifier for RE tasks that is built upon the principles of the DocFusion classification head (formerly derived from the LayoutXLM head) [23].

Beyond these primary contributions, our work also explores additional dimensions of document understanding by considering various forms of data interaction and feature fusion. For instance, our experiments examine the sensitivity of RE performance to different weighting schemes in the joint representation space. In certain configurations, we model the interaction between modalities using a weighted sum approach:

$$R = \alpha \cdot T + \beta \cdot L + \gamma \cdot V,$$

where R denotes the final representation, T, L, and V represent the text, layout, and visual features respectively, and α, β, and γ are hyperparameters that balance their relative contributions. Such formulations underscore the nuanced manner in which various modalities can be combined to yield a more robust feature set, which is critical for accurately inferring relations among document entities.

Furthermore, our investigation delves into the scalability and generalization capabilities of the DocFusion approach. We analyze its performance across diverse document types and domains, reflecting the heterogeneous nature of real-world documents. Through detailed comparative studies, we establish that while the text/layout combination often offers optimal performance, scenarios exist in which the inclusion of visual information can mitigate ambiguities introduced by noisy or sparse textual data.

In summary, this work contributes to the growing body of research on multimodal document analysis by presenting a comprehensive study on the integration of text, layout, and visual cues for

[1] https://aws.amazon.com/textract/
[2] https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/overview-ocr
[3] https://pypi.org/project/pytesseract/

 doi:10.20944/preprints202502.1700.v1

relation extraction. The remainder of this paper is organized as follows. Section 2 reviews the literature related to multimodal approaches for document understanding and discusses current insights into the relative importance of different modalities. Section 3 outlines our methodology, detailing the dataset, model architecture, and experimental procedures. Section 4 reports our experimental results and performance comparisons. Section 5 discusses the limitations of our study and proposes directions for future research. Finally, Section 6 offers concluding remarks and reflections on the broader implications of our findings.

2. Related Work

In the realm of Visually-rich Document Understanding (VrDU) for relation extraction (RE), two primary benchmark datasets have emerged as indispensable resources. These are the FUNSD dataset [11] and the XFUND dataset [23]. Both datasets provide detailed annotations, where each instance includes a specification of linked entities represented by pairs of entity IDs, or an empty array in cases where no relationship is identified. This comprehensive annotation strategy enables researchers to delve deeply into the multimodal aspects of document analysis.

The availability of multiple data types—namely text, geometric layout, and document images—in these datasets naturally motivates the development of advanced multimodal methodologies. Early pioneering work in this area includes the LayoutLM family of models [21–23], which ingeniously combine textual embeddings with spatial (position) embeddings to capture the intrinsic layout information present in documents. LayoutLMv2 further enhances this approach by fully integrating visual cues extracted directly from document images, thereby providing a richer and more nuanced representation.

In parallel, the method presented in [18] employs a similar trifecta of inputs—text, layout, and visual data—to tackle tasks on both the FUNSD and MedForm datasets. Additionally, research conducted by [1] illustrates the effectiveness of leveraging a multimodal framework for document classification by fusing text and image modalities. Other notable contributions in the literature have focused on harnessing the synergistic effects of text and layout representations exclusively [12,13,15], thereby underscoring the critical role these modalities play in extracting structural and relational information from documents.

A central question that has arisen in the literature is: *What are the relative effects and contributions of these different data types?* Although several studies report ablation experiments, the precise impact of individual modalities is often difficult to ascertain, particularly because textual information is typically preserved across all experimental configurations [9,15]. This ambiguity is especially pertinent in industry applications, where the increased training and inference costs of large-scale multimodal models must be justified by commensurate improvements in performance. Notably, the original XFUND paper [23] does not offer a detailed breakdown of the contributions from text, layout, and visual information—a gap that has motivated further research in this direction.

In response to these challenges, our work introduces a novel framework named **UniFusion**. UniFusion is designed to harmonize the strengths of various modalities while maintaining computational efficiency. By systematically integrating text, layout, and visual features, UniFusion not only builds upon the foundational ideas present in earlier models but also provides a more balanced and cost-effective approach to multimodal representation learning. Although previous studies have largely treated each modality in isolation or in fixed combinations, UniFusion employs adaptive fusion strategies that dynamically adjust the contributions of each modality according to the complexity and quality of the input data.

Moreover, the literature reveals that the interplay among different modalities is not merely additive but often involves complex interactions that can either enhance or detract from overall system performance. For example, certain studies have observed that the presence of strong textual signals can sometimes overshadow the contribution of layout and visual features, leading to imbalanced

 doi:10.20944/preprints202502.1700.v1

models. Conversely, in scenarios where textual data is sparse or noisy, the complementary information provided by geometric and visual cues becomes critically important. These insights further validate the need for a framework like UniFusion, which is engineered to optimize the integration of heterogeneous data sources.

In summary, the extensive body of related work not only highlights the rapid evolution of multimodal approaches in document understanding but also points to significant gaps in our understanding of modality-specific contributions. Our proposed UniFusion framework aims to address these challenges by providing a unified, adaptive, and efficient method for multimodal representation learning. The insights gleaned from previous studies serve as a robust foundation for our work, setting the stage for a more thorough and balanced evaluation of multimodal strategies in the subsequent sections.

3. Methodology

In this section, we present in exhaustive detail the methodological framework underpinning our experiments on relation extraction (RE) using the XFUND dataset. Our approach leverages a novel multimodal architecture, termed **UniFusion**, which is an evolution of earlier transformer-based models. UniFusion is designed to integrate text, layout, and visual modalities into a unified representation for RE tasks. In what follows, we describe the dataset and preprocessing steps, elaborate on the architecture of UniFusion with all relevant technical and mathematical formulations, and detail the experimental procedures and training protocols employed.

3.1. Dataset and Preprocessing

We utilize the XFUND dataset[4] for our experiments on multimodal RE [23]. XFUND comprises a large collection of document images specifically curated for form understanding across seven languages. Each document in the dataset is annotated with a unique identifier, a class label, bounding box coordinates

$$\{x_{left}, y_{top}, x_{right}, y_{bottom}\},$$

the extracted textual content, and a linking indicator. The linking indicator is instrumental in defining relationships between entities, thereby facilitating the application of the dataset to the VrDU RE task. In this context, entities are organized as key-value pairs corresponding to the questions and answers found in the forms. For a comprehensive description of the data collection and annotation process, readers are referred to [23]. It is noteworthy that the dataset statistics we employ differ slightly from those originally reported in [23]; hence, we provide the revised statistics in Table 1. The language codes *ZH*, *JA*, *ES*, *FR*, *IT*, *DE* and *PT* correspond respectively to Chinese, Japanese, Spanish, French, Italian, German, and Portuguese.

Table 1. Train/Test split for XFUND data.

	ZH	JA	ES	FR	IT	DE	PT
Train	187	194	243	202	265	189	233
Test	65	71	74	71	92	63	85

Prior to feeding the data into our model, we perform several preprocessing steps. These include normalization of the bounding box coordinates to ensure consistency across different document scales, tokenization of the textual content using sub-word techniques, and resizing of document images to a fixed resolution. Furthermore, we apply language-specific preprocessing pipelines to handle the multilingual nature of the dataset.

[4] https://github.com/doc-analysis/XFUND

 doi:10.20944/preprints202502.1700.v1

3.2. UniFusion Architecture for Relation Extraction

The core of our proposed methodology is the **UniFusion** architecture, which builds upon the conceptual foundations of transformer-based models while introducing key innovations to enhance multimodal feature integration. UniFusion is a pretrained transformer that simultaneously ingests text, layout, and visual inputs to generate a rich joint representation suitable for RE tasks.

3.2.1. Feature Extraction and Embedding Layers

UniFusion begins by encoding each modality using dedicated embedding layers:

- **Textual Embedding:** Each word token is embedded into a high-dimensional space via a learned embedding matrix $E_t \in \mathbb{R}^{V \times d}$, where V is the vocabulary size and d is the embedding dimension.
- **Layout Embedding:** The geometric layout of the document is represented using normalized bounding box coordinates. These coordinates are transformed via a fully-connected layer into an embedding $E_l \in \mathbb{R}^{4 \times d_l}$, where d_l denotes the dimensionality specific to spatial features.
- **Visual Embedding:** Document images are processed by a visual backbone based on ResNeXt 101-FPN [20], which outputs visual feature maps. These maps are further encoded into a vector $E_v \in \mathbb{R}^{d_v}$ using convolutional and pooling layers.

For each token, the final multimodal representation is computed as a summation of the modality-specific embeddings along with positional encodings. To account for the two-dimensional nature of document layouts, UniFusion incorporates both one-dimensional (1D) and two-dimensional (2D) positional embeddings:

$$E = E_t + E_{p1D} + E_l + E_{p2D} + E_v,$$

where E_{p1D} and E_{p2D} denote the 1D and 2D positional embeddings respectively.

3.2.2. Multimodal Fusion Mechanism

A distinguishing feature of UniFusion is its adaptive multimodal fusion strategy. Instead of statically concatenating modalities, UniFusion learns to weight the contribution of each modality dynamically. The fusion process is formulated as:

$$F = \lambda_T \cdot \phi(T) + \lambda_L \cdot \phi(L) + \lambda_V \cdot \phi(V),$$

where $\phi(\cdot)$ represents a transformation function (implemented as a feed-forward network) for the text (T), layout (L), and visual (V) features, and λ_T, λ_L, and λ_V are learnable parameters that modulate the influence of each modality. In practice, these weights are normalized using a softmax function to ensure that

$$\lambda_T + \lambda_L + \lambda_V = 1.$$

Additionally, the fusion layer is enhanced by a self-attention mechanism that captures the interactions between different modalities. The attention operation is given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where Q, K, and V denote the query, key, and value matrices respectively, and d_k is the dimensionality of the key vectors.

3.2.3. Customized Classification Head

For the relation extraction task, UniFusion incorporates a specialized classification head that builds upon the joint representation. Unlike the original LayoutXLM design, our classification head

 doi:10.20944/preprints202502.1700.v1

has been streamlined to reduce computational overhead while preserving performance. Specifically, the classification head comprises:

- A single fully-connected layer that projects the fused representation F to a lower-dimensional space.
- A leaky ReLU activation function to introduce non-linearity:

$$\text{LeakyReLU}(x) = \max(\alpha x, x),$$

  where α is a small positive constant (set to 0.01 in our experiments).
- A dropout layer with a dropout probability $p = 0.2$ to mitigate overfitting.

Furthermore, the classification layer employs a bi-affine transformation to model interactions between entity pairs. Given two entity representations h_i and h_j, the score for a potential relation is computed as:

$$s_{ij} = h_i^\top W h_j + U(h_i + h_j) + b,$$

where W, U, and b are learnable parameters.

3.2.4. Loss Function and Optimization

The overall training objective of UniFusion for the RE task is defined by a composite loss function that includes a cross-entropy loss for classification and an auxiliary regularization term to enforce smoothness in the multimodal fusion weights. The loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \mathcal{L}_{reg},$$

where

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

and

$$\mathcal{L}_{reg} = \left\| \nabla\lambda - \bar{\lambda} \right\|_2^2.$$

Here, $\hat{y}_i$ is the predicted probability for the i-th sample, y_i is the corresponding ground truth label, λ denotes the set of fusion weights, $\bar{\lambda}$ is the mean value of the weights, and β is a regularization hyperparameter.

3.3. Experimental Procedures

Our experiments are designed to rigorously evaluate the effectiveness of UniFusion by testing six distinct multimodal and unimodal configurations on the multilingual XFUND dataset. In total, experiments are performed separately for each of the seven language subsets, and each configuration is fine-tuned from the pretrained UniFusion model. The six configurations are as follows:

1. **Multimodal (MM):** Incorporates text, layout, and visual information.
2. **Bimodal Text and Layout (text/layout):** Visual components are entirely removed.
3. **Bimodal Text and Visual (text/visual):** Layout information, including 2D position embeddings, is omitted.
4. **Bimodal Layout and Visual (layout/visual):** Textual data and corresponding embeddings are excluded.
5. **Unimodal Layout (layout):** Only layout information is used.
6. **Unimodal Text (text):** Only text data is leveraged.

 doi:10.20944/preprints202502.1700.v1

3.3.1. Configuration-Specific Model Modifications

For the **MM** configuration, the full UniFusion architecture is employed without modifications. In the **text/layout** setting, we excise all visual components, including the visual backbone and associated embeddings. Conversely, in the **text/visual** configuration, the layout-specific 2D position embeddings are removed from both the primary stream and the visual branch. In the **layout/visual** configuration, the textual stream is completely omitted by discarding tokenized text and the corresponding 1D and 2D positional embeddings. The unimodal experiments (configurations 5 and 6) are derived from the aforementioned bimodal settings by applying the relevant exclusions. It is important to note that many prior studies have neglected to consider the absence of text data entirely [9,15], despite its critical role in multimodal fusion.

3.3.2. Training Protocol and Hyperparameter Optimization

Given the inherent differences in data characteristics across modalities, we hypothesized that the optimal learning rate might vary between experiments. Consequently, learning rate was the sole hyperparameter subjected to optimization. A grid search was conducted over three candidate values: 5×10^{-5}, 1×10^{-5}, and 5×10^{-6}. All other hyperparameters remained constant across experiments. In our training regimen, the following parameters were employed:

- **Batch Size:** 2.
- **Number of Epochs:** 50.
- **Optimizer:** Adam with default momentum parameters.
- **Learning Rate Scheduler:** A cosine annealing scheduler was applied to gradually decrease the learning rate over the training period.

To monitor the training process and prevent overfitting, we also implemented early stopping based on the validation loss. Moreover, we computed additional evaluation metrics such as precision, recall, and F1 score, defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where TP, FP, and FN denote the true positives, false positives, and false negatives, respectively.

3.3.3. Additional Technical Considerations

To further enhance the stability and performance of the UniFusion model, several technical innovations were incorporated:

1. **Gradient Clipping:** A threshold of 1.0 was imposed on gradient norms to prevent exploding gradients.
2. **Layer Normalization:** Each transformer block within UniFusion is equipped with layer normalization to stabilize training dynamics.
3. **Multi-Head Self-Attention:** The self-attention mechanism in UniFusion is implemented with multiple heads, each head performing independent attention operations. The overall attention output is then concatenated and projected back into the model's hidden space.
4. **Residual Connections:** To facilitate the flow of gradients, residual connections are employed throughout the network.

The fusion of these advanced techniques ensures that UniFusion is capable of robustly capturing the intricate interactions among text, layout, and visual modalities, ultimately leading to superior performance on the RE task.

In summary, our experimental procedures are meticulously designed to investigate the impact of each modality on relation extraction performance. The comprehensive analysis across multiple configurations and languages not only validates the efficacy of the proposed UniFusion framework

 doi:10.20944/preprints202502.1700.v1

but also provides critical insights into the nuanced contributions of each data type in visually-rich document understanding.

4. Experimental Outcomes and Analysis

In this section, we present a detailed discussion of the experimental results obtained using the UniFusion framework on the XFUND relation extraction task. Our evaluation is primarily based on the F1 score, with additional metrics such as precision and recall also reported. Although multimodal results have been previously presented for the XFUND RE task [23], our results are reported independently due to variations in network configuration, dataset statistics, and training procedures.

4.1. Bimodal Integration Outperforms Trimodal Fusion

Table 2 reports the F1 scores achieved by models trained under six different configurations. Supplementary recall and precision metrics are provided in Tables 3 and 4, respectively. Overall, the experimental findings strongly support the advantage of learning joint representations for the VrDU RE task. In particular, the bimodal configuration that combines text and layout information achieves a mean F1 score of 0.6843, which is higher than the trimodal configuration (text, layout, and image) that obtains a mean F1 of 0.6728.

Table 2. XFUND F1 scores for different training configurations.

	MM	Txt/Lay	Txt/Im	Lay/Im	Layout	Text
ZH	0.6935	**0.7212**	0.6192	0.6334	0.5417	0.5636
JA	0.6987	**0.7181**	0.6406	0.6061	0.5674	0.6321
ES	**0.7198**	0.7159	0.6069	0.5657	0.4483	0.5885
FR	0.6573	**0.6747**	0.5888	0.5820	0.5021	0.5285
IT	0.6841	**0.7090**	0.6281	0.4906	0.4825	0.5724
DE	**0.6782**	0.6701	0.6041	0.5397	0.4035	0.5821
PT	0.5779	**0.5812**	0.5367	0.4909	0.3511	0.5053
Mean	0.6728	**0.6843**	0.6035	0.5583	0.4709	0.5675

Table 3. Recall scores for XFUND data for different training configurations.

	MM	Txt/Lay	Txt/Im	Lay/Im	Layout	Text
ZH	0.6109	**0.7607**	0.6754	0.7011	0.6639	0.6440
JA	0.5638	**0.7540**	0.6601	0.6619	0.6932	0.6762
ES	0.6475	**0.7210**	0.6817	0.6807	0.4487	0.6136
FR	0.6231	**0.7278**	0.6365	0.7278	0.6956	0.5774
IT	0.6314	**0.7090**	0.6584	0.5649	0.5966	0.7009
DE	**0.6990**	0.6518	0.6267	0.6158	0.4279	0.5931
PT	0.4851	**0.6640**	0.5308	0.5891	0.4082	0.5089
Mean	0.6087	**0.7126**	0.6385	0.6488	0.5620	0.6163

This result reinforces previous observations that the integration of text and layout can capture the essential document features required for accurate relation extraction, sometimes even outperforming configurations that include additional visual data. In our experiments, the text/layout model outperforms the trimodal model by margins of approximately 8.08% and 6.93% when compared to the next best alternative, as detailed in Table 2. Additionally, the other bimodal configurations—text/visual and layout/visual—yield F1 scores of 0.604 and 0.558, respectively, which further emphasizes the hierarchical predictive strength of these modalities. Text appears to be the most dominant signal, followed by layout, with visual information contributing marginally in the current setup.

 doi:10.20944/preprints202502.1700.v1

Table 4. Precision scores for XFUND data for different training configurations.

	MM	Txt/Lay	Txt/Im	Lay/Im	Layout	Text
ZH	0.6109	**0.6855**	0.5717	0.5777	0.4576	0.5010
JA	0.5638	**0.6854**	0.6223	0.5589	0.4802	0.5934
ES	0.6475	**0.7108**	0.5469	0.4768	0.4478	0.5654
FR	0.6231	**0.6288**	0.5478	0.4848	0.3928	0.4872
IT	0.6314	**0.6862**	0.6004	0.4336	0.4050	0.4837
DE	**0.6990**	0.6894	0.5831	0.5804	0.3817	0.5715
PT	0.4851	**0.5167**	0.5427	0.4207	0.3080	0.5016
Mean	0.6087	**0.6575**	0.5736	0.5046	0.4104	0.5291

To quantify the performance more systematically, we also consider a weighted F1 score metric defined as:

$$\mathrm{F1}_{\text{weighted}} = \frac{\sum_{i=1}^{N} n_i \cdot \mathrm{F1}_i}{\sum_{i=1}^{N} n_i},$$

where n_i denotes the number of samples for language i and $\mathrm{F1}_i$ is the corresponding F1 score. This formulation confirms that the text/layout configuration consistently outperforms other modalities across the multilingual dataset.

4.2. Text Remains the Central Component in Relation Extraction

Our results clearly indicate that the exclusion of text data leads to a significant drop in performance. Among the six configurations evaluated, those that do not include textual information (namely, the layout/visual and unimodal layout setups) are the worst performing. Notably, the unimodal text model achieves considerably better classification results than the bimodal layout/visual model, which highlights the pivotal role of text in the RE task.

However, the inclusion of additional modalities still offers benefits. When combined with text, both layout and visual information contribute to improved performance. For example, the addition of layout features not only enhances the performance in the highest scoring text/layout configuration but also substantially improves the unimodal layout score (from 0.471 to 0.558 in the layout/visual setting). This observation can be mathematically expressed as:

$$\Delta\mathrm{F1} = \mathrm{F1}_{\text{bimodal}} - \mathrm{F1}_{\text{unimodal}},$$

demonstrating the incremental benefit derived from multimodal fusion. The experimental outcomes suggest that while text is indispensable, its optimal performance is achieved when supplemented by layout information, particularly in scenarios where text quality may be compromised.

4.3. Training Dynamics and Data-Dependent Variability

The training dynamics further emphasize the crucial role of text in the learning process. Loss curves (not shown here) indicate that model configurations incorporating text experience a rapid and steady decline in training loss, converging reliably within 50 epochs. In contrast, models that rely solely on layout or a combination of layout and visual data display more gradual loss reductions and greater variability across different language sets.

We define the training loss function as:

$$\mathcal{L}(t) = \frac{1}{N} \sum_{i=1}^{N} \ell\big(y_i, \hat{y}_i(t)\big),$$

 doi:10.20944/preprints202502.1700.v1

where $\ell(\cdot)$ represents the per-sample loss, y_i is the ground truth label, and $\hat{y}_i(t)$ is the predicted output at epoch t. The steeper decline in $\mathcal{L}(t)$ for models utilizing text confirms that textual features provide a robust signal that facilitates effective optimization.

Moreover, the observed higher variance in training loss for configurations that exclude text suggests that layout and visual modalities are more sensitive to dataset-specific characteristics. This is further reflected in the optimal learning rates determined by grid search: while the full multimodal and text/layout configurations were trained with a learning rate of 5×10^{-5}, the models excluding text required lower learning rates (either 1×10^{-5} or 5×10^{-6}). Such differences underscore the importance of modality-specific hyperparameter tuning in multimodal systems.

In summary, the experimental outcomes validate the efficacy of the UniFusion framework, highlighting that text is the foundational component in relation extraction tasks within visually-rich documents. The combination of text with layout information yields the best overall performance, while visual information plays a more supplementary role. These insights are crucial for developing efficient and effective multimodal deep learning systems for document analysis.

5. Conclusions and Future Directions

5.1. Conclusions

In this work, we explored the complex trade-offs that arise when employing multimodal approaches in industrial applications, particularly for relation extraction within visually-rich document understanding (VrDU). Our investigation involved training a multimodal transformer—renamed here as **UniFusion**—under several distinct data configurations to evaluate the effectiveness of joint representation learning. The empirical results demonstrate that the bimodal configuration incorporating both text and layout information consistently outperforms the full trimodal setup, even though the latter integrates visual cues as well.

More specifically, our findings indicate that textual information serves as the primary driver in predictive tasks, with unimodal text achieving a higher mean F1 score compared to the bimodal layout/visual configuration. This suggests that while text is indispensable for capturing the semantic essence of documents, layout information plays a crucial complementary role. In scenarios where visual data is incorporated, its contribution appears to be supplemental, enhancing performance in particular conditions such as when the quality of textual data is compromised. This observation not only corroborates previous studies but also reinforces the design philosophy behind UniFusion, where the strategic integration of modalities is key to enhanced performance.

Moreover, our analysis showed that training dynamics and convergence behaviors are highly dependent on the inclusion or exclusion of specific modalities. The experiments reveal that when text is omitted, the performance degradation is significant, underscoring its pivotal role in the learning process. Overall, the results provide robust evidence that while each modality has its own merits, the most effective approach for the RE task is to combine text with layout data, thereby harnessing the strengths of both semantic and structural information.

5.2. Limitations and Future Work

Despite the promising outcomes, several limitations in the current study warrant discussion. One major constraint is the relative scarcity of available data for the VrDU RE task. With FUNSD [11] being the only other comparable dataset, the limited diversity and volume of samples restrict our ability to generalize the results. Future research should aim to expand the dataset repertoire and increase the number of samples, which could help better understand how data diversity influences model robustness.

Another limitation is the focus on a single multimodal architecture—UniFusion (previously LayoutXLM)—which may not capture the full spectrum of architectural innovations available in the literature. Other frameworks with alternative strategies for encoding and fusing modalities, as

 doi:10.20944/preprints202502.1700.v1

explored in recent studies [8,10], could offer additional insights. A systematic comparison involving a wider range of architectures is essential to further delineate the relative contributions of text, layout, and visual information.

Future work should also consider extending the scope of analysis to include other document understanding tasks such as document classification, semantic entity recognition, and key information extraction. A large-scale, cross-task evaluation would help to determine whether the observed performance hierarchy—text as the primary modality, followed by layout and then visual data—holds across different applications. Additionally, more advanced ablation studies, possibly incorporating a cost-performance trade-off analysis, could provide valuable insights into the efficiency and practicality of deploying such multimodal systems in high-volume business environments.

Furthermore, the exploration of dynamic fusion strategies remains a promising avenue for research. Future studies might investigate adaptive weighting schemes that adjust the contributions of different modalities on a per-sample basis. Such mechanisms could further optimize joint representation learning and improve overall system resilience against variations in document quality.

In summary, while the current study demonstrates the efficacy of UniFusion for the VrDU RE task, there remains considerable scope for refining multimodal fusion strategies and broadening the evaluation framework to encompass diverse document analysis challenges. Addressing these limitations will be critical for advancing the state-of-the-art in document understanding and achieving higher levels of performance in real-world applications.

References

1. Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. arXiv preprint arXiv:1907.06370 (2019)
2. Carbonell, M., Fornés, A., Villegas, M., Lladós, J.: A neural model for text localization, transcription and named entity recognition in full pages. Pattern Recognition Letters **136**, 219–227 (2020)
3. Carbonell, M., Riba, P., Villegas, M., Fornés, A., Lladós, J.: Named entity recognition and relation extraction with graph neural networks in semi structured documents. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9622–9627. IEEE (2021)
4. Cooney, C., Folli, R., Coyle, D.: A bimodal deep learning architecture for eeg-fnirs decoding of overt and imagined speech. IEEE Transactions on Biomedical Engineering (2021)
5. Dang, T.A.N., Hoang, D.T., Tran, Q.B., Pan, C.W., Nguyen, T.D.: End-to-end hierarchical relation extraction for generic form understanding. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5238–5245. IEEE (2021)
6. Davis, B., Morse, B., Price, B., Tensmeyer, C., Wiginton, C.: Visual fudge: Form understanding via dynamic graph editing. arXiv preprint arXiv:2105.08194 (2021)
7. Gralinski, F., Stanislawek, T., Wróblewska, A., Lipinski, D., Kaliska, A., Rosalska, P., Topolski, B., Biecek, P.: Kleister: A novel task for information extraction involving long documents with complex layout. CoRR **abs/2003.02356** (2020), https://arxiv.org/abs/2003.02356
8. Gu, Z., Meng, C., Wang, K., Lan, J., Wang, W., Gu, M., Zhang, L.: Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4583–4592 (2022)
9. Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., Park, S.: Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. arXiv preprint arXiv:2108.04539 (2021)
10. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. arXiv preprint arXiv:2204.08387 (2022)
11. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 1–6. IEEE (2019)
12. Li, C., Bi, B., Yan, M., Wang, W., Huang, S., Huang, F., Si, L.: Structurallm: Structural pre-training for form understanding. arXiv preprint arXiv:2105.11210 (2021)

 doi:10.20944/preprints202502.1700.v1

13. Li, Y., Qian, Y., Yu, Y., Qin, X., Zhang, C., Liu, Y., Yao, K., Han, J., Liu, J., Ding, E.: Structext: Structured text understanding with multi-modal transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1912–1920 (2021)
14. Liu, X., Gao, F., Zhang, Q., Zhao, H.: Graph convolution for multimodal information extraction from visually rich documents. arXiv preprint arXiv:1903.11279 (2019)
15. Pramanik, S., Mujumdar, S., Patel, H.: Towards a multi-modal, multi-task learning based pre-training framework for document representation learning. arXiv preprint arXiv:2009.14457 (2020)
16. Sharif, M.I., Khan, M.A., Alhussein, M., Aurangzeb, K., Raza, M.: A decision support system for multimodal brain tumor classification using deep learning. Complex & Intelligent Systems pp. 1–14 (2021)
17. Venugopalan, J., Tong, L., Hassanzadeh, H.R., Wang, M.D.: Multimodal deep learning models for early detection of alzheimer's disease stage. Scientific reports **11**(1), 1–13 (2021)
18. Wang, Z., Zhan, M., Liu, X., Liang, D.: Docstruct: A multimodal method to extract hierarchy structure in document for general form understanding. arXiv preprint arXiv:2010.11685 (2020)
19. Wei, M., He, Y., Zhang, Q.: Robust layout-aware ie for visually rich documents with pre-trained language models. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2367–2376 (2020)
20. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431 (2016)
21. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740 (2020)
22. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1192–1200 (2020)
23. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.: Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. arXiv preprint arXiv:2104.08836 (2021)
24. Zhang, P., Xu, Y., Cheng, Z., Pu, S., Lu, J., Qiao, L., Niu, Y., Wu, F.: Trie: End-to-end text reading and information extraction for document understanding. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1413–1422 (2020)
25. 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.
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. 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.
28. 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/.
29. 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.
30. 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.
31. 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.
32. 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.

 doi:10.20944/preprints202502.1700.v1

33. 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.
34. 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.
35. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
45. 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.
46. 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.
47. 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.
48. 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.
49. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
50. 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.
51. 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.
52. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
53. 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.

 doi:10.20944/preprints202502.1700.v1

54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.
69. 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.
70. 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.
71. 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.
72. 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.

 doi:10.20944/preprints202502.1700.v1

73. 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.
74. 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.
75. 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.