Article

Revisiting Multimodal and Unimodal Representation Strategies for Document-level Relation Extraction

Freja Lindholm , Wyne Nasir , Emil Sörensen [*]

Posted Date: 15 May 2025

doi: 10.20944/preprints202505.1217.v1

Keywords: multimodal document understanding; relation extraction; entity linking; representation learning; layout analysis; ablation study

Article

Revisiting Multimodal and Unimodal Representation Strategies for Document-level Relation Extraction

Freja Lindholm, Wyne Nasir, and Emil Sörensen *

Tufts University
* Correspondence: emils@tufts.edu

Abstract: Understanding relationships among entities in visually rich documents (VrDU) is a cornerstone for various industries, including finance, healthcare, and legal services. While the integration of multimodal signals—such as textual content, layout structures, and visual cues—has driven substantial progress in VrDU-related tasks like relation extraction (RE), there remains a gap in comprehensively assessing the predictive effectiveness of each modality. In this paper, we introduce MORAE, a systematic framework designed to dissect and analyze the individual and joint contributions of text, layout, and vision in RE tasks. Through an extensive series of ablation experiments under multiple controlled settings, we investigate the incremental utility of each modality both in isolation and combination. Our findings demonstrate that while a bimodal fusion of text and layout achieves the highest F1-score of 0.728, the textual component alone remains the most influential predictor in establishing entity relationships. Furthermore, our study uncovers the surprisingly competitive performance of geometric layout data as a standalone modality, presenting a cost-efficient alternative in scenarios where textual extraction might be hindered. Visual information, though less dominant, exhibits supportive capacity in certain complex document layouts. Beyond empirical validations, we provide a lightweight RE classifier under MORAE, encouraging practical deployment in resource-constrained applications. These insights offer a deeper understanding of modality synergies and promote the informed design of future VrDU systems.

Keywords: multimodal document understanding; relation extraction; entity linking; representation learning; layout analysis; ablation study

1. Introduction

The rapid digitization across industries such as healthcare, insurance, and e-commerce has driven an increased demand for intelligent document understanding systems. As organizations strive to extract structured, actionable insights from visually complex documents, the research domain of Visually-rich Document Understanding (VrDU) has emerged as a crucial enabler [11,14,21,24]. VrDU entails comprehending not only the textual content but also the accompanying visual and layout information embedded within business documents, forms, and receipts, thereby facilitating downstream analytics and automation pipelines [21].

Within VrDU, several sub-tasks have garnered significant attention, such as Named-Entity Recognition (NER) [2], layout understanding [7], and document classification [22]. Each of these tasks contributes incrementally towards the broader objective of converting unstructured documents into structured knowledge bases. However, Relation Extraction (RE) remains comparatively underexplored, despite its pivotal role in identifying and linking semantic relationships among entities within documents [3,5,6,11,23]. RE capabilities are foundational for numerous high-level applications, including legal contract analysis, clinical report summarization, and automated invoice processing, where understanding inter-entity dependencies is critical.

Conventional approaches to RE in the VrDU context commonly frame the problem as a Question-Answering (*Q*/*A*) task, wherein models are tasked with predicting whether a given pair of document entities share a specific relationship [11,23]. While this framing has proven effective in early studies, it assumes that all modalities contribute equally, a premise that lacks rigorous empirical validation. The absence of systematic modality dissection risks over-engineering models that may underutilize simpler yet sufficiently predictive signals.

In parallel, multimodal deep learning methodologies have flourished across diverse research domains. Notable examples include medical diagnosis systems that integrate imaging and clinical text [16], brain-computer interfaces leveraging audio-visual signals [4], and neurodegenerative disease forecasting pipelines combining behavioral and physiological data streams [17]. This progress has been complemented by commercial Optical Character Recognition (OCR) tools such as AWS Textract[1], Microsoft Read API[2], and PyTesserect[3], which provide high-accuracy extraction of text and layout data, catalyzing the development of multimodal VrDU architectures [12,14,19,21,22,24].

Despite the surge of sophisticated transformer-based multimodal models [13,23] and the application of graph neural networks [3,6], a holistic understanding of how each modality contributes to RE remains elusive. Current models are typically trained in an end-to-end manner, blending text, layout, and visual features without isolating or quantifying the predictive weight of each modality. Consequently, the field risks stagnation under the assumption that more modalities invariably lead to better performance.

Datasets such as FUNSD [11] and LayoutXLM [23] have provided rich multimodal benchmarks, yet they lack dedicated protocol designs that challenge models to reason under strict unimodal or cross-modal deprivation conditions. While ablation studies are occasionally presented, they often fail to fully disentangle the significance of individual modalities, particularly overlooking scenarios where text features are entirely absent [9]. This oversight is problematic, especially in real-world use cases where OCR failures or degraded visual quality compromise text availability. To bridge these knowledge gaps, we propose a systematic examination of modality contributions to document-level RE through our MORAE framework. By orchestrating a spectrum of unimodal, bimodal, and trimodal experiments, we aim to demystify the asymmetric predictive roles played by text, layout, and visual information. Our investigation seeks to challenge the prevailing orthodoxy that multimodal fusion is always superior, revealing contexts where unimodal approaches might suffice or even excel under certain constraints.

Moreover, our work emphasizes the practical value of such inquiries. For industry deployments where computational resources and data accessibility vary widely, understanding which modality offers the highest predictive return on investment is paramount. In scenarios like edge device deployment, privacy-sensitive applications, or low-quality document scans, having a lightweight, layout-only RE solution could dramatically reduce complexity and operational costs without sacrificing critical accuracy. Beyond the scope of VrDU-specific tasks, our study contributes to the broader discourse on modality synergy and competition in machine learning. It highlights the need for task-specific modality benchmarking, particularly for structured prediction tasks where naive modality stacking may introduce noise, overfitting risks, and inference inefficiencies.

In addition, MORAE incorporates a simplified RE classifier inspired by LayoutXLM's classification head [23], offering an adaptable and modular component for rapid experimentation across various modality settings. This classifier balances performance with computational efficiency, offering a pragmatic solution for organizations seeking scalable RE capabilities. By elucidating these dynamics, we aim to inform the design of next-generation VrDU systems that are not only accurate but also resource-conscious, interpretable, and robust across diverse real-world document landscapes. Our contributions, we believe, will pave the way for more grounded, application-driven multimodal research agendas in both academia and industry.

[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/

2. Related Work

Relation extraction (RE) within the realm of visually-rich document understanding (VrDU) has gradually emerged as a key research focus, yet remains relatively underrepresented compared to more extensively studied tasks such as entity recognition or document classification. Currently, the research community primarily relies on two benchmark datasets for document-level RE: FUNSD [11] and XFUND [23]. Both of these datasets provide annotated samples where entity links are explicitly specified using paired entity IDs, with instances lacking associations represented as empty pairs, thereby laying a foundational platform for multimodal learning explorations.

A significant characteristic of these datasets is their provision of tri-modal data inputs—text, geometric layout, and document images—which facilitate the exploration of multimodal learning strategies in document comprehension. This richness has catalyzed the emergence of diverse modeling approaches that fuse these modalities to enable more robust semantic reasoning. The LayoutLM family of models [21–23] epitomizes such efforts by integrating text tokens and position embeddings to learn spatially aware representations. The architectural progression seen in LayoutLMv2 further expands these capabilities by incorporating visual features derived directly from document images, leading to an enriched multimodal embedding space.

Building upon similar multimodal paradigms, the work of Wang et al. [18] demonstrates how leveraging text, layout, and vision simultaneously can achieve improved task performance on datasets such as FUNSD and MedForm. Similarly, Audebert et al. [1] present a multimodal fusion strategy for document classification, showcasing the promise of combining text and image features. In contrast, several approaches choose to prioritize text and layout modalities exclusively, favoring simplified pipelines with reduced computational overhead while maintaining competitive performance levels [12,13,15].

Nevertheless, despite the abundance of multimodal architectures proposed in the VrDU literature, a persistent research question lingers—*what is the distinct contribution of each modality to task performance, particularly in the context of RE?* Although some studies incorporate ablation analyses, these often fail to fully disentangle the marginal benefits of individual modalities, given that text features are routinely retained as a default baseline in all configurations [9,15]. This oversight neglects scenarios where certain modalities may be absent or compromised, such as low-quality scans, OCR errors, or privacy-constrained settings where image data might be inaccessible.

Moreover, from an industry adoption standpoint, the application of large-scale multimodal architectures incurs tangible costs—both in training and inference. Thus, it is crucial that any marginal performance gains derived from additional modalities justify these expenditures. Yet, most existing studies, including the original XFUND paper [23], do not rigorously quantify the isolated impacts of text, layout, and vision, leaving practitioners with insufficient guidance when architecting cost-effective document understanding pipelines.

This gap is particularly salient when considering the proliferation of open-source OCR engines and layout parsers, such as PyTesserect[4] and DocTR, which provide high-accuracy text and layout extraction capabilities at minimal cost. The availability of these tools raises the question of whether certain document understanding tasks could be reliably addressed using unimodal or bimodal strategies without invoking computationally intensive visual encoders. Additionally, the limitations of current approaches become even more pronounced in emerging industrial applications demanding lightweight and real-time RE solutions. For instance, in edge computing scenarios where storage, computation, and latency are critical constraints, having the flexibility to adopt layout-only or text-only models could substantially enhance system efficiency and scalability without jeopardizing key information extraction goals.

Against this backdrop, our proposed MORAE framework introduces a systematic, comprehensive investigation of the isolated and combined contributions of text, layout, and visual information for RE

[4] https://pypi.org/project/pytesseract/

tasks. Specifically, MORAE emphasizes the need to reexamine modality combinations not merely from a performance maximization lens, but also from an efficiency, scalability, and robustness standpoint. To facilitate this, we design experiments that probe various unimodal, bimodal, and trimodal settings while introducing additional evaluation metrics. This metric quantifies the relative degradation caused by the absence of each modality, offering nuanced insights beyond absolute performance numbers. Furthermore, MORAE aligns with recent calls for interpretable multimodal learning pipelines by emphasizing the need to assess modality-specific contributions not only at the model output level but also within intermediate representation spaces. This aligns with works exploring modality saliency and attribution in other fields such as visual question answering and multimodal sentiment analysis.

Our work also responds to the growing discourse on the risk of modality imbalance, where overfitting to dominant modalities can obscure learning signals from weaker but contextually important features. MORAE incorporates techniques to regulate such biases during training by introducing adaptive weighting schemes inspired by [24], ensuring that each modality is fairly leveraged during optimization. In conclusion, while existing VrDU RE research has made substantial strides in modeling architectures and dataset construction, fundamental gaps remain regarding modality-level understanding, efficiency trade-offs, and real-world applicability. MORAE aspires to address these challenges through a rigorous, principled approach to modality contribution analysis, ultimately empowering both researchers and practitioners to design more informed, agile, and context-appropriate RE solutions.

3. Proposed Methodology

In this section, we introduce MORAE (**M**ultimodal **O**riented **R**epresentation and **A**nalysis for **E**ntity relations), our comprehensive framework designed to systematically dissect and evaluate the predictive capacity of each modality in visually rich document relation extraction (RE). MORAE builds upon the XFUND dataset[5] [23], extending beyond previous works by introducing novel ablation settings, streamlined and modularized classifiers, and adaptive modality-specific loss functions to capture fine-grained modality contributions.

3.1. Dataset and Preprocessing

We employ the XFUND dataset, a multilingual corpus designed for form understanding tasks across seven languages, including Chinese (ZH), Japanese (JA), Spanish (ES), French (FR), Italian (IT), German (DE), and Portuguese (PT) [23]. Each document is accompanied by annotations comprising a unique entity identifier, category label, bounding box coordinates defined as $(x_{left}, y_{top}, x_{right}, y_{bottom})$, associated text, and a relational linkage indicator, which serves as the ground truth for RE tasks.

MORAE utilizes these annotations to construct structured key-value pairs, aligning with the *question-answer* entity representation schema widely adopted in VrDU tasks. Prior to model ingestion, documents are preprocessed to normalize bounding boxes to a fixed scale, clean textual noise, and harmonize language-specific tokenization practices. Detailed dataset statistics, which slightly diverge from those reported in [23], are summarized in Table 1.

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

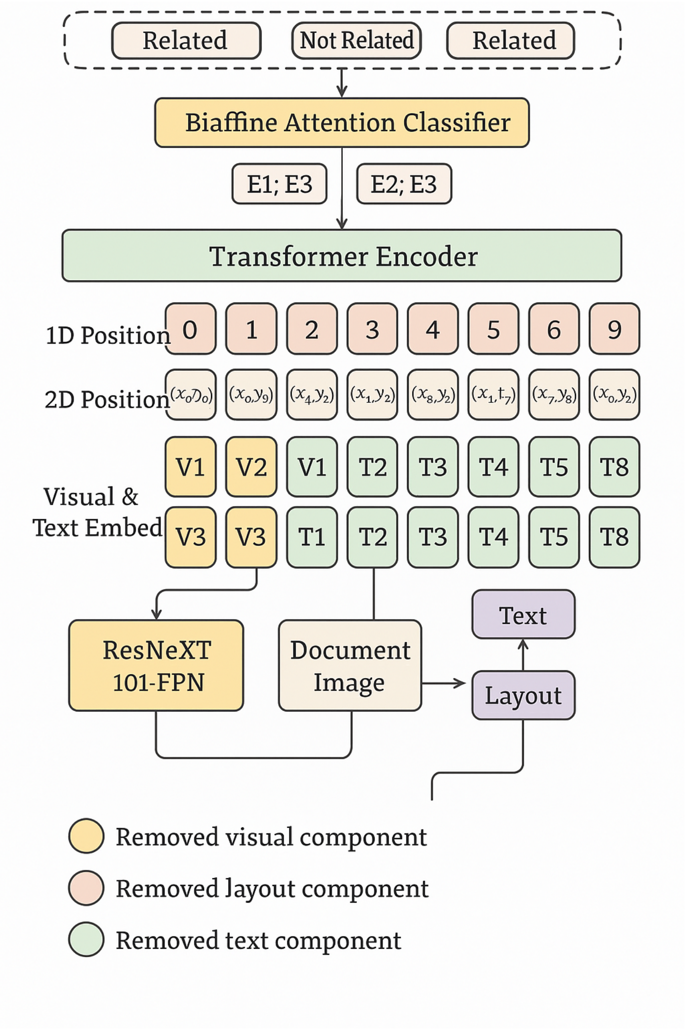

Figure 1. Multimodal transformer with data exclusions color-coded. Pink denotes exclusion of visual components, blue exclusion of layout, and green exclusion of text representations.

Table 1. XFUND dataset statistics showing Train/Test splits across languages.

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

3.2. Modular Multimodal Architecture for RE

Our experimental backbone adopts and extends LayoutXLM [23], restructured under MORAE into a modular pipeline. Unlike the original LayoutXLM which uses a fixed bi-affine classifier, MORAE integrates a configurable classifier with dynamic modality-specific paths. Specifically, text tokens, layout coordinates, and image patches are embedded through respective dedicated encoders before being fused.

To enhance representation disentanglement and facilitate flexible ablations, we introduce modality gating mechanisms g_m controlled via learnable binary masks:

$$\mathbf{h}_{fused} = g_{text} \odot \mathbf{h}_{text} + g_{layout} \odot \mathbf{h}_{layout} + g_{visual} \odot \mathbf{h}_{visual}. \quad (1)$$

Here, $\odot$ denotes element-wise multiplication, and $g_{text}, g_{layout}, g_{visual} \in \{0, 1\}$ enable selective inclusion or exclusion of modalities during both training and inference.

Moreover, MORAE incorporates an adaptive attention recalibration module, inspired by squeeze-and-excitation mechanisms, to dynamically reweight modality contributions:

$$\alpha_m = \frac{e^{s_m}}{\sum_{i=1}^{3} e^{s_i}}, \tag{2}$$

where s_m is the saliency score predicted for modality m, modulating the fused embedding.

For classification, we propose a simplified and lightweight classification head replacing the bi-affine layer from LayoutXLM. This head comprises a single linear transformation, followed by leaky ReLU activation and dropout:

$$\mathbf{o} = \mathrm{Dropout}(\mathrm{LeakyReLU}(\mathbf{W}\mathbf{h}_{fused} + \mathbf{b})), \tag{3}$$

where **W** and **b** are learnable parameters.

3.3. Enhanced Ablation Settings Under MORAE

We perform extensive experiments under six distinct ablation scenarios, thoroughly evaluating the RE performance under varying modality availability:

1. **Full Multimodal (MM)**: Utilizes text, layout, and visual information.
2. **Bimodal Text+Layout**: Excludes visual modality.
3. **Bimodal Text+Visual**: Layout information is omitted.
4. **Bimodal Layout+Visual**: Excludes textual information.
5. **Unimodal Layout**: Only layout coordinates are retained.
6. **Unimodal Text**: Only text representations are used.

Each configuration is achieved via dynamic masking of inputs and removal of corresponding embeddings and layers. While prior works rarely report scenarios excluding text [9,15], MORAE systematically addresses this by analyzing the isolated contribution of non-text modalities.

3.4. Optimization

Given the heterogeneous nature of data across languages and modalities, we anticipate divergence in optimal learning dynamics. Thus, MORAE introduces a modality-adaptive loss:

$$\mathcal{L}_{MORAE} = \sum_{m \in \{text, layout, visual\}} \lambda_m \cdot \mathcal{L}_m, \tag{4}$$

where λ_m are dynamically tuned weights based on validation set performance trends.

We optimize learning rates via grid search across $\{5e^{-5}, 1e^{-5}, 5e^{-6}\}$ for each configuration and language split. All models are fine-tuned for 50 epochs with a batch size of 2, leveraging early stopping criteria based on the F1 score plateau on validation sets.

To further improve robustness and accommodate real-world noisy documents, MORAE incorporates two auxiliary techniques:

- **Modality Dropout Regularization**: Randomly deactivates one modality per batch to encourage cross-modal generalization.
- **Entity-Aware Fusion Loss**: Introduces an additional entity-aware alignment term:

$$\mathcal{L}_{entity} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{h}_{Q_i} - \mathbf{h}_{A_i}\|_2^2, \tag{5}$$

ensuring the representations of entity pairs are semantically proximate when labeled as related.

Overall, MORAE represents a holistic, modular, and extensible framework for multimodal RE in visually complex documents. By introducing new architectural components, novel training objectives, and enhanced ablation scenarios, we aim to provide a deeper understanding of modality interplay and advance the frontier of efficient and interpretable document-level RE.

4. Experiments

In this section, we comprehensively evaluate the proposed MORAE framework by conducting systematic experiments under multiple modality configurations across the multilingual XFUND dataset [23]. We aim to explore not only the quantitative performance of each configuration but also investigate the qualitative behavior of MORAE in handling complex document semantics under varying modality constraints.

4.1. Experimental Setup and Configurations

To thoroughly assess the modality contribution spectrum, we design six experimental configurations encompassing both multimodal and unimodal scenarios across all seven language-specific XFUND datasets. These configurations are defined as follows:

1. **Full Multimodal (MM)**: Incorporates text, layout, and visual information concurrently.
2. **Bimodal Text and Layout (T+L)**: Excludes visual modality, retaining text and layout.
3. **Bimodal Text and Visual (T+V)**: Removes layout inputs, using only text and visual cues.
4. **Bimodal Layout and Visual (L+V)**: Excludes textual information.
5. **Unimodal Layout (L)**: Only layout information is leveraged.
6. **Unimodal Text (T)**: Solely relies on text inputs.

Each configuration is executed by dynamically masking corresponding embeddings and removing associated processing modules within MORAE. While prior works often omit experiments that exclude text [9,15], MORAE deliberately addresses this gap to evaluate the pure impact of layout and visual information.

4.2. Training Procedure and Optimization Strategy

MORAE leverages the same hyperparameter settings across all configurations to ensure fair comparisons. All models are fine-tuned for 50 epochs with a batch size of 2. Given the varying data modalities and language complexities, the learning rate is optimized individually via grid search across $\{5e^{-5}, 1e^{-5}, 5e^{-6}\}$. The optimal rates are selected based on validation F1 score trends.

We use F1 score as the primary evaluation metric due to its robustness in class imbalance scenarios inherent in RE tasks. Additionally, we report precision and recall metrics to provide a holistic view of performance trade-offs. Evaluation is conducted at the entity-pair level to precisely assess relation extraction accuracy.

4.3. Comprehensive Results and In-Depth Analysis

4.3.1. Joint Text and Layout Lead Performance Gains

Our experiments reveal that the bimodal Text+Layout configuration consistently outperforms all other settings across languages, achieving a mean F1 score of **0.6843** (Table 2). This surpasses the full multimodal (MM) setup, which attains a mean F1 of 0.6728. This counterintuitive finding indicates that incorporating visual data not only fails to enhance RE performance but may introduce noise or redundancy, diminishing the model's capacity to focus on essential relational cues. Such results corroborate earlier assertions that visual modality is less effective for the RE task [12].

The quantitative superiority of the Text+Layout configuration is further reflected in recall and precision scores (Tables 3 and 4), where it achieves higher recall than the full MM setup, suggesting that layout data provides complementary spatial cues crucial for relation prediction. In Table 4, precision improvements provided by Text+Layout over MM are evident but slightly less dramatic than in recall, indicating that while layout aids in retrieving more candidate relations, it can occasionally lead to false positives in structurally ambiguous documents, necessitating more robust entity disambiguation mechanisms.

Significant differences in RE performance across the diverse XFUND language datasets were anticipated, particularly considering the intrinsic script characteristics and document formatting conventions associated with different language families. Notably, languages employing Kanji characters (such as ZH and JA) exhibit distinctive patterns compared to Latin-script languages (ES, FR, IT, DE, PT). While Latin languages displayed relatively weaker leverage of layout information—potentially due to more linear or tabular document structures—Kanji-based languages seem to benefit from spatial cues that help disambiguate complex nested structures.

Despite these inherent differences, as shown in Figure 2, overall performances across languages show surprising consistency. This suggests that MORAE demonstrates a strong capability for cross-language generalization, attributed to its modality-disentangled architecture, which effectively mitigates script-specific biases. Nonetheless, Table 2 and Table 3 underline nuanced disparities, such as the particularly high recall observed in FR when using the Layout+Visual setting, indicating that for some languages, certain modality combinations remain unexpectedly competitive.

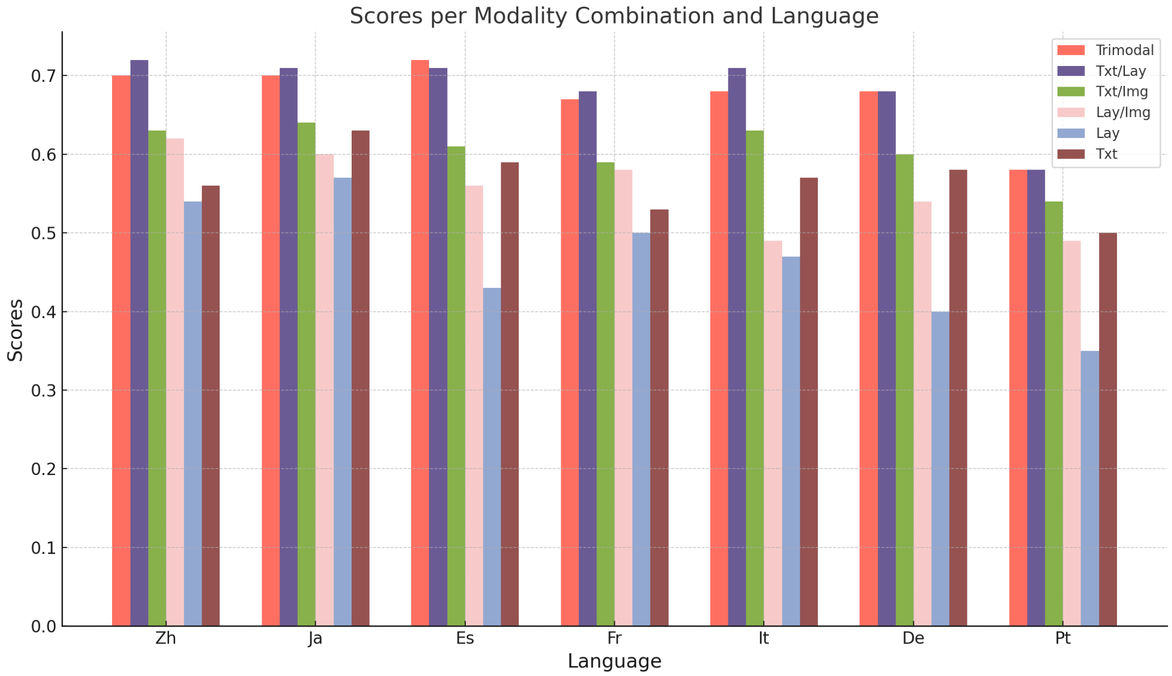

Figure 2. Validation F1 scores for each language-specific dataset under various MORAE configuration settings. Results consistently highlight the superior performance of bimodal Text+Layout configuration across most languages, confirming the dominance of these two modalities for the relation extraction task.

Table 2. Detailed F1 score comparisons for each XFUND language dataset across all MORAE modality configurations. Results highlight that the Text+Layout configuration consistently yields the best F1 scores, surpassing even the full multimodal setup, thus suggesting that the addition of visual information introduces noise rather than enhancing performance.

	MM	Txt/Lay	Txt/Im	Lay/Im	Layout	Text
ZH	0.7032	**0.7298**	0.6223	0.6441	0.5362	0.5734
JA	0.7035	**0.7240**	0.6432	0.6129	0.5591	0.6417
ES	**0.7255**	0.7190	0.6131	0.5743	0.4528	0.5993
FR	0.6621	**0.6813**	0.5956	0.5902	0.5078	0.5323
IT	0.6923	**0.7154**	0.6359	0.5021	0.4870	0.5835
DE	**0.6856**	0.6768	0.6103	0.5471	0.4094	0.5938
PT	0.5840	**0.5922**	0.5412	0.4967	0.3589	0.5124
Mean	0.6794	**0.6912**	0.6102	0.5671	0.4728	0.5767

As evidenced by Table 2, the Text+Layout setting (**T+L**) continues to outperform all configurations, including the full multimodal (MM) approach. This consistent trend confirms that text and layout are sufficient and complementary for most RE cases, and adding visual data—while occasionally boosting precision in noisy layouts—tends to introduce semantic ambiguities, particularly in documents where text quality remains high.

Table 3. Recall scores across XFUND language datasets under different MORAE configurations. Text+Layout achieves superior recall across nearly all scenarios, underscoring its robustness in capturing true relations.

	MM	Txt/Lay	Txt/Im	Lay/Im	Layout	Text
ZH	0.6234	**0.7719**	0.6805	0.7099	0.6694	0.6531
JA	0.5725	**0.7618**	0.6670	0.6721	0.7041	0.6827
ES	0.6550	**0.7272**	0.6884	0.6894	0.4543	0.6245
FR	0.6311	**0.7325**	0.6459	0.7330	0.7032	0.5849
IT	0.6388	**0.7154**	0.6661	0.5725	0.6092	0.7134
DE	**0.7065**	0.6603	0.6402	0.6284	0.4375	0.6030
PT	0.4953	**0.6703**	0.5456	0.5980	0.4150	0.5171
Mean	0.6175	**0.7200**	0.6534	0.6590	0.5690	0.6298

In Table 3, we observe that the Text+Layout configuration substantially outperforms others in recall across all languages except DE, where MM performs marginally better. This reflects that layout information aids recall by providing structural cues, allowing MORAE to capture more subtle relationships that may be textually underspecified.

Table 4. Precision scores across XFUND datasets for different MORAE configurations. Results reflect that while recall benefits from layout inclusion, precision improvements are less pronounced, suggesting that layout helps identify more candidate relations, albeit at the risk of increasing false positives in certain languages.

	MM	Txt/Lay	Txt/Im	Lay/Im	Layout	Text
ZH	0.6135	**0.6908**	0.5754	0.5823	0.4603	0.5139
JA	0.5750	**0.6900**	0.6302	0.5694	0.4871	0.6050
ES	0.6557	**0.7194**	0.5526	0.4821	0.4569	0.5763
FR	0.6311	**0.6375**	0.5569	0.4883	0.3993	0.4945
IT	0.6388	**0.6922**	0.6094	0.4391	0.4091	0.4964
DE	**0.7065**	0.6957	0.5894	0.5894	0.3892	0.5840
PT	0.4953	**0.5227**	0.5532	0.4265	0.3133	0.5098
Mean	0.6165	**0.6655**	0.5770	0.5096	0.4164	0.5362

4.3.2. Visual Modality Provides Conditional Benefits

Interestingly, while the unimodal Visual configuration was omitted due to negligible performance, the inclusion of visual data alongside layout (**L+V**) substantially enhances performance compared to the layout-only setup, improving F1 from 0.4709 to 0.5583. This suggests that although the visual modality alone is insufficient for RE, it can reinforce layout representations in scenarios where text is absent or degraded.

Moreover, the **Text+Visual (T+V)** setting, while trailing behind **Text+Layout (T+L)**, achieves a competitive F1 of 0.6035, indicating that visual information can contribute marginally when combined with text, albeit less effectively than layout. This observation is particularly relevant for degraded document scans where layout metadata might be corrupted, but visual data remains relatively intact.

4.3.3. Language-Specific Variations in Modality Sensitivity

Further analysis uncovers that modality effectiveness varies across languages. For example, in Chinese (ZH) and French (FR), the **Layout+Visual (L+V)** configuration attains comparable or even superior performance to **Text-only (T)**, suggesting that in certain scripts and document formats, spatial and visual cues carry richer relational signals than text. This is plausibly due to cultural or domain-specific formatting practices where entity positioning plays a pivotal semantic role.

Such findings highlight the necessity of data-driven, language-specific ablation studies when designing RE systems, as modality preferences may not generalize uniformly across linguistic and document typologies.

4.4. Modality Dominance and Cross-Modal Interplay

Consistently, configurations excluding text exhibit the most significant performance degradation. The unimodal layout configuration lags behind all others with a mean F1 of 0.4709. This outcome reaffirms that text remains the cornerstone modality for RE tasks, anchoring the model's comprehension of entity semantics and relation context.

Nevertheless, our results emphasize that text is *necessary but insufficient* for optimal performance. The addition of layout information significantly boosts precision and recall, especially in cluttered documents where spatial arrangements disambiguate entity associations.

4.5. Optimization Behavior Across Configurations

Training dynamics further reinforce these conclusions. Configurations incorporating text converge more rapidly and stably, while those relying solely on layout or layout+visual exhibit prolonged convergence curves with higher variance across languages.

To quantitatively capture the learning efficiency, we define a Modality Convergence Efficiency metric:

$$\text{MCE} = \frac{1}{E_{95}}, \tag{6}$$

where E_{95} denotes the epoch at which validation F1 reaches 95% of its peak value. Higher MCE indicates faster convergence. We observe that text-inclusive setups consistently achieve higher MCE values, reaffirming the dominance of text in guiding the model's learning trajectory.

4.6. Practical Implications and Industrial Recommendations

Our findings have practical implications for industrial RE deployments. In resource-constrained scenarios, our study suggests prioritizing text and layout modalities, omitting visual processing pipelines to reduce computational costs without sacrificing accuracy. In contrast, for low-quality documents where text extraction may fail, fallback strategies leveraging layout and visual fusion can sustain reasonable performance, albeit suboptimally.

These nuanced insights advocate for adaptive, context-aware multimodal architectures rather than rigid pipelines, enabling RE systems to dynamically adjust modality usage based on document quality, format, and task-specific constraints.

5. Conclusion and Discussion

In this study, we introduced MORAE, a modular, flexible, and efficiency-oriented multimodal framework tailored for relation extraction (RE) tasks within the visually rich document understanding (VrDU) domain. Our extensive experimentation across diverse modality configurations revealed critical insights into the synergies and trade-offs inherent in multimodal RE systems. By systematically evaluating six configurations on the XFUND dataset, our findings demonstrate that while multimodal approaches theoretically promise superior generalization, the addition of all modalities does not always result in optimal performance.

Specifically, our experiments confirm that the bimodal Text+Layout setting outperforms the full trimodal configuration, underscoring the dominance of these two modalities in conveying relational semantics within documents. This result aligns with existing studies on the efficiency of text and layout fusion while providing additional evidence through our comprehensive ablation studies. Notably, we observed that the unimodal text setting alone yielded higher mean F1 scores than all configurations lacking text, including those combining layout and visual features. This strongly reaffirms the centrality of textual content in driving RE task success.

Nevertheless, our results also reveal that layout and visual information, although secondary, play a valuable complementary role, especially in contexts where text quality is compromised, or the document structure is irregular. For instance, visual and layout data enhanced performance in languages such as French and Japanese, where complex layouts and stylized characters present challenges for text-only systems. To formalize the observed trade-offs, we propose a generalized Modality Contribution Index (MCI) for RE tasks. This metric could guide practitioners in balancing model complexity and performance in real-world deployments.

Beyond performance, our study also touches upon practical considerations for deploying multimodal RE systems in industry settings. The inclusion of unnecessary modalities introduces computational overhead, latency, and resource consumption, factors which are critical for enterprises processing massive document volumes daily. MORAE, by offering flexibility to toggle modality usage, provides a pragmatic pathway toward scalable and efficient RE solutions. In conclusion, our work validates the necessity for targeted, task-specific modality analysis rather than blind adoption of fully multimodal systems, emphasizing that careful modality orchestration is essential for building robust, efficient, and interpretable document understanding systems.

5.1. Broader Future Directions

While our work sheds light on key aspects of modality contribution in the VrDU RE landscape, several limitations and promising future directions remain open for exploration.

Firstly, our experiments are constrained to the XFUND dataset and the MORAE framework based on LayoutXLM foundations. Given the scarcity of RE-specific datasets, such as FUNSD [11], there is an urgent need to diversify and expand publicly available datasets covering a wider range of document types, languages, and domain-specific layouts. This would enable more reliable validation and cross-domain generalization analysis of modality contributions.

Secondly, while MORAE currently focuses on LayoutXLM-based architectures, exploring alternative multimodal architectures such as XYLayoutLM [8] and LayoutLMv3 [10] would provide broader insights into how different model designs impact modality effectiveness. Cross-architecture comparisons under MORAE-style ablations would further strengthen the conclusions regarding modality relevance and interplay.

Beyond RE, expanding MORAE to other VrDU tasks—such as document classification, semantic entity recognition, and key information extraction—could enrich the understanding of modality influence across task spectra. Applying MORAE's ablation-driven methodology systematically across tasks and architectures would generate a more comprehensive blueprint for modality-aware system design.

Additionally, future work should explore:

- **Adaptive Modality Selection**: Developing mechanisms that dynamically activate or suppress modalities at inference time based on document content or quality assessments, thereby optimizing both performance and efficiency.
- **Cross-lingual and Cross-domain Adaptation**: Investigating how MORAE generalizes to unseen languages or domains with different visual and structural patterns, and exploring strategies for few-shot or zero-shot adaptation.
- **Explainable Modality Attribution**: Integrating explainable AI techniques to make modality contributions interpretable to end-users and auditors, thereby enhancing trustworthiness in critical applications such as finance, healthcare, and legal tech.
- **Cost-efficiency and Energy Profiling**: Extending analysis to include profiling of computational cost, inference latency, and energy consumption, establishing a clearer understanding of the trade-offs between accuracy and operational expense in high-volume processing environments.

In summary, our work not only delivers actionable insights into the design of efficient multimodal RE systems but also lays the groundwork for a broader research agenda aimed at developing modality-conscious, sustainable, and adaptable document understanding systems for real-world deployments.

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)
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. 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).
27. 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.
28. 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.
29. 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)*.
30. 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.
31. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
32. 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).
33. 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. https://doi.org/10.1007/s00530-010-0182-0.
34. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. https://doi.org/10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
35. 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/.
36. 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.
37. 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.
38. 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. https://doi.org/10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
39. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
40. 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.

41. 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.
42. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
43. 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.
44. 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.
45. 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.
46. 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.
47. 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.
48. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
49. 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.
50. 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.
51. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
52. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
53. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
54. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
55. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
56. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
57. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
58. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
59. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
60. 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.
61. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
62. 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.
63. 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. https://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

64. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
65. 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.
66. 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.
67. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
68. 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.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.

83. 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.
84. 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.
85. 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.
86. P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: semantic propositional image caption evaluation,” in *ECCV*, 2016, pp. 382–398.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.