Article

Not peer-reviewed version

Bridging Vision and Texts: An External Graph Framework for Enhanced Language Comprehension

Martínez Pérez , Emily Marwood , Martina Fernández Gómez *

Posted Date: 13 March 2025

doi: 10.20944/preprints202503.1014.v1

Keywords: Multimodal Graphs; External Knowledge Integration; Language Comprehension; Multilingual NER; Visual Sense Disambiguation

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2025 doi:10.20944/preprints202503.1014.v1

Article

Bridging Vision and Texts: An External Graph Framework for Enhanced Language Comprehension

Martínez Pérez, Lobry Hsu and Martina Fernández Gómez *

Bond University, Robina, QLD 4226, Australia
* Correspondence: martinafg@bond.edu.au

Abstract: In this work, we introduce a novel framework that augments language understanding systems with external multimodal graph structures. Instead of increasing the internal capacity of language models by scaling parameters, our approach leverages a dedicated external repository—an enriched knowledge graph—to provide additional visual and textual cues during inference. Specifically, given multilingual inputs (for example, German sentences), our method retrieves corresponding entities from the graph and incorporates their multimodal embeddings to boost performance on various downstream tasks. Our framework, herein referred to as **AlphaKG**, integrates state-of-the-art tuple-based and graph-based learning strategies to generate representations for entities and their inter-relations. By fusing data from diverse modalities such as textual descriptions available in 14 languages and multiple visual samples per entity, we design a robust representation learning scheme that is predictive of the underlying graph structure. Experiments on multilingual named entity recognition (NER) and crosslingual visual verb sense disambiguation (VSD) show promising results, with improvements reaching up to 0.7% in F1 score for NER and up to 2.5% in accuracy for VSD. Additionally, we derive new equations to refine the integration process between the retrieved external features and the language model inputs, thereby offering a comprehensive solution to enhance parameter efficiency while maintaining competitive performance.

Keywords: multimodal graphs; external knowledge integration; language comprehension; multilingual NER; visual sense disambiguation

---

1. Introduction

Recent advances in natural language understanding (NLU) and natural language generation (NLG) have significantly transformed the landscape of artificial intelligence. State-of-the-art models have achieved remarkable performance across various benchmarks [18,40,41], yet this progress has come with an ever-increasing demand for computational resources and a rapid escalation in the number of model parameters [8,11,30]. This surge in complexity has led to substantial financial, computational, and environmental costs, which pose serious challenges for both academia and industry [34].

Traditional approaches aimed at improving efficiency, such as model distillation [32] or enforcing parameter sharing [21], primarily focus on compressing or reorganizing the internal structure of language models. However, these techniques still necessitate the storage of vast amounts of information within the model parameters, limiting their scalability and flexibility in dynamic environments. By contrast, our proposed method advocates for the externalization of knowledge, thereby relieving language models from the need to memorize extensive amounts of information.

This work introduces the idea of augmenting language models with a dedicated external repository that houses a rich, multimodal knowledge graph. Such an approach permits the retrieval of additional contextual cues, both textual and visual, during the inference process. Consequently, models can leverage up-to-date and diversified data without overburdening their internal architectures, leading to enhanced performance and improved parameter efficiency. The concept of retrieving external information to support language understanding has been explored in prior research [27,29]. Yet, most

 doi:10.20944/preprints202503.1014.v1

existing efforts focus solely on textual data and often neglect the substantial benefits provided by visual cues. In our framework, these visual elements are integrated alongside multilingual textual descriptions, offering a more comprehensive representation of entities. This multimodal strategy not only enriches the information available to the language model but also bridges the gap between purely text-based representations and the complex, real-world scenarios in which these models are deployed.

Moreover, our framework, designated as **AlphaKG**, is designed to interface seamlessly with contemporary language models, providing them with external representations that are both visually and linguistically grounded. This integration enables the model to dynamically access and utilize supplementary information, thereby supporting more nuanced and context-aware decision-making processes. The external repository can be updated independently of the main language model, which allows for continuous learning and rapid adaptation to new information. Beyond the technical advantages, the externalization of knowledge offers significant practical benefits. By decoupling the storage of extensive background information from the core model, our approach facilitates a modular design. Such modularity allows individual components to be refined or replaced without necessitating a complete overhaul of the system. This flexibility is especially critical in applications where the underlying data evolves rapidly or where frequent updates are required to maintain high performance. Another notable aspect of our method is its potential for scalability. As the external knowledge graph is updated with additional data—ranging from emerging visual trends to newly available multilingual text—the language model benefits from a continuously expanding repository of relevant information. This ensures that the model remains effective even as the scope of real-world data broadens, without the need to increase its intrinsic parameter count.

In addition, our approach mitigates the challenges associated with overfitting that are common in large-scale language models. By offloading a significant portion of the required knowledge to an external graph, the model can focus on learning how to effectively integrate and interpret this supplementary information. This decoupling of knowledge storage from inference processes encourages more robust generalization and better performance across diverse tasks. Furthermore, the use of an external multimodal knowledge graph provides an innovative pathway for integrating disparate sources of information. The synergy between textual and visual data enhances the overall representational capacity of the system, paving the way for breakthroughs in tasks such as multilingual named entity recognition (NER) and visual verb sense disambiguation (VSD). By combining these modalities, the system is better equipped to capture subtle contextual cues that are often missed by models relying solely on one type of data.

In summary, the externalization of knowledge through a multimodal graph framework represents a significant shift in the design of language understanding systems. The **AlphaKG** model exemplifies how decoupling knowledge storage from the internal parameters of language models can lead to enhanced efficiency, improved scalability, and greater adaptability. This paradigm not only alleviates the growing burden of model size but also opens new avenues for future research in integrating multimodal information into language processing pipelines.

2. Related Work

Retrieval Augmented Models

Another significant research direction involves retrieval augmented models, where external information is accessed by querying pre-indexed knowledge bases rather than relying solely on internal memory. In these approaches, the external repository is populated with data that extends far beyond the training corpus of the target task. For example, Lee et al. [22] introduced a framework for Open Retrieval Question Answering (ORQA) in which both the retrieval and answering components are jointly trained to leverage external data sources. Similarly, Karpukhin et al. [19] developed a dense passage retriever (DPR) that surpasses traditional sparse retrieval methods such as TF-IDF or BM25 by significantly enhancing retrieval quality, which in turn leads to improved performance on question answering tasks.

 doi:10.20944/preprints202503.1014.v1

Additional work, such as REALM [14], incorporates a dense Wikipedia index and fine-tunes both the index and the language model simultaneously to tackle open-domain QA problems. In parallel, Petroni et al. [28] examined the effect of feeding BERT with contexts retrieved or generated through different techniques, revealing that external information can substantially influence unsupervised QA performance. Moreover, Lewis et al. [23] integrated a retrieval module into an encoder-decoder architecture to condition the generation process on factual data extracted from Wikipedia. While these models predominantly focus on text-based retrieval, our proposed AlphaKG framework expands upon this paradigm by incorporating structured multimodal information. Unlike conventional retrieval systems that treat facts as unstructured text, AlphaKG is designed to retrieve and leverage both visual and textual features that are inherently organized according to a knowledge graph's structure.

Multimodal Pretraining

Pretraining methods that jointly model vision and language have recently emerged as a powerful trend, achieving state-of-the-art results on various multimodal reasoning tasks [24,37,47]. These approaches generally adopt masked multimodal modeling techniques over image-text pairs to learn rich, joint representations that capture the intricate interactions between visual content and linguistic cues. Unlike end-to-end models that rely solely on raw paired data, these multimodal pretraining frameworks harness the synergy between modalities to better capture context and semantic nuances.

While many existing methods focus on implicitly learning cross-modal connections through large-scale data, our approach explicitly incorporates structured external knowledge into the model. The AlphaKG framework leverages a well-organized knowledge graph that contains not only multilingual textual descriptions but also visual representations of entities. This explicit modeling of entity-centric relationships enables a more precise retrieval of multimodal information, which is crucial for tasks requiring fine-grained reasoning. Furthermore, by structuring the external information, AlphaKG facilitates interpretable alignments between visual cues and textual semantics, providing an additional layer of robustness and control.

The trend in recent research is evident: external and structured knowledge sources are increasingly recognized as valuable complements to internal model representations. Memory networks, retrieval augmented models, and multimodal pretraining techniques all contribute unique perspectives on how best to integrate external information. Our AlphaKG framework builds on these insights by uniting the strengths of dynamic memory access, sophisticated retrieval mechanisms, and multimodal pretraining. This integration is expected to yield significant improvements in tasks that demand a deep understanding of both visual and textual data.

In conclusion, the body of work encompassing external memory augmentation, retrieval-based approaches, and multimodal pretraining offers a comprehensive foundation for advancing language understanding systems. By synthesizing these diverse methodologies, our proposed AlphaKG framework presents a novel approach to incorporating structured, multimodal knowledge into neural models, thereby addressing critical challenges in scalability, efficiency, and interpretability.

Memory in Neural Networks

The idea of augmenting neural models with an external memory has a long-standing history in the literature. Early studies demonstrated that recurrent neural networks could be enriched with external memory mechanisms to capture context-free grammars and other complex structures [9,48]. More contemporary frameworks, such as memory networks [36,43] and neural Turing machines [13], further advanced this concept by enabling networks to dynamically read from and write to external storage. These architectures provide models with the ability to maintain long-term dependencies and to manipulate contextual information beyond the limitations of fixed internal representations.

In these systems, the memory access is typically managed via differentiable attention mechanisms. For example, a canonical approach to reading from memory involves computing a weighted sum over memory slots. Although such formulations are not the focus of our present work, they offer important insights into how external memory components can enhance neural computations. Recent research

 doi:10.20944/preprints202503.1014.v1

has also extended these techniques to multimodal contexts, where visual and textual data are stored in a unified memory system. For instance, Xiong et al. [45] adapted memory networks for both textual question answering and visual question answering by aligning visual features with textual queries, while Su et al. [35] and Wang et al. [42] demonstrated that incorporating a visual memory component improves performance in tasks like video captioning and visual QA. These advances underscore the importance of dynamic memory modules that can integrate heterogeneous information, a principle that underlies our proposed AlphaKG framework.

3. Methodology

3.1. Overview and Motivation

AlphaKG [1] represents a state-of-the-art multilingual and multimodal knowledge graph (KG) constructed by leveraging BabelNet v4.0 [26] and ImageNet [31]. Unlike many traditional KGs that focus solely on textual or structured data, AlphaKG integrates visual information by associating multiple images with each node. Each node corresponds to a *synset*—a set of synonymous terms that describe a specific concept—and is enriched with descriptions in several languages. For example, the synset representing the concept of *dog* may be accompanied by the gloss "The dog is a mammal in the order Carnivora," along with several illustrative images. This rich, multimodal integration makes AlphaKG particularly suitable for bridging the gap between vision and language tasks.

The design of AlphaKG was motivated by the need for high-quality, well-curated data that is directly applicable in modern neural pipelines for vision-and-language research. To ensure visual relevance, nodes are selected based on criteria that include both their linguistic descriptions and the presence of strong visual features. The knowledge graph covers a wide range of topics and includes 13 distinct relation types that emphasize visual components. These relation types include: *is-a*, *has-part*, *related-to*, *used-for*, *used-by*, *subject-of*, *receives-action*, *made-of*, *has-property*, *gloss-related*, *synonym*, *part-of*, and *located-at*. Such a diverse set of relations enables the KG to capture intricate semantic connections and nuanced visual relationships among concepts.

To our knowledge, AlphaKG is the only publicly available multimodal KG that has been specifically designed for seamless integration into neural model pipelines. Although our experiments focus on AlphaKG, the underlying framework we propose, AlphaKG, can be extended to any similar knowledge repository, thereby broadening its applicability across various research domains.

3.2. Graph Structure and Mathematical Notation

We formalize the AlphaKG KG as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes (or synsets) and $\mathcal{E}$ represents the set of directed edges corresponding to typed semantic relations between these nodes. Each edge $e_r \in \mathcal{E}$ is associated with a relation type r from the predefined set of 13 relation categories.

For any given node $v_i \in \mathcal{V}$, we define its local neighborhood $\mathcal{N}_i$ as:

$$\mathcal{N}_i = \{v_j \in \mathcal{V} \mid (v_i, e_r, v_j) \in \mathcal{E} \text{ for some } e_r\}.$$

This neighborhood function is instrumental in many graph-based learning algorithms where information from adjacent nodes is aggregated to learn robust representations.

We denote the set of all valid relational triples (or factual tuples) in the KG as

$$\mathcal{D} = \{(v_i, e_r, v_j) \mid v_i, v_j \in \mathcal{V}, e_r \in \mathcal{E}\}.$$

To facilitate training of embedding models, a set of corrupted triples $\mathcal{D}'$ is generated by randomly substituting either the head or tail node such that the corrupted tuple (v_i, e_r, v'_j) does not exist in $\mathcal{G}$. Such negative sampling is common in contrastive learning and ranking-based loss formulations.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2025 doi:10.20944/preprints202503.1014.v1

Our representation learning approach in the AlphaKG framework involves constructing two key embedding matrices:

$$TV \in \mathbb{R}^{|\mathcal{V}|\times d_n} \quad \text{and} \quad TE \in \mathbb{R}^{|\mathcal{E}|\times d_r},$$

where d_n and d_r are the dimensionalities of node and relation embeddings, respectively. The embedding for node v_i is denoted by the row vector $Tv_i = TV[i,:]$, and for a relation e_r, the embedding is given by $Te_r = TE[r,:]$. These embeddings are learned such that they preserve both the structural and semantic properties of the KG.

In addition, each node $v_i \in \mathcal{V}$ is augmented with two types of auxiliary data:

- **Multilingual Glosses:** A set of textual descriptions $\mathcal{T}_i$, where each gloss $t \in \mathcal{T}_i$ provides language-specific information about the concept.
- **Visual Images:** A collection $\mathcal{I}_i$ of images that visually depict the corresponding concept.

We adopt the notation $[Tx; Ty]$ to represent the concatenation of vectors Tx and Ty, and $Tx \odot Ty$ to denote their element-wise product. These operations play a vital role in our subsequent fusion and gating mechanisms.

Furthermore, to quantify the connectivity of the graph, we define the degree of a node v_i as:

$$\deg(v_i) = |\{v_j \in \mathcal{V} \mid (v_i, e_r, v_j) \in \mathcal{E} \text{ or } (v_j, e_r, v_i) \in \mathcal{E}\}|.$$

This measure is crucial for understanding the distribution of node connectivity and for designing neighborhood aggregation strategies in graph-based models.

3.3. Statistical Properties and Integration Details

The scale of AlphaKG is significant: it comprises over 100,000 nodes and nearly 2 million relations, along with more than 1.5 million images. Such a large-scale dataset offers a rich testbed for learning multimodal representations that can capture the interplay between visual and textual modalities. Although a detailed statistical summary (including comparisons with other multimodal KGs such as WN9-IMG and FB15-IMG) was provided in earlier studies [1], here we briefly summarize some key properties:

- **Node Count:** $|\mathcal{V}| \approx 10^5$
- **Relation Count:** The KG encompasses $|\mathcal{E}| \approx 1.9 \times 10^6$ edges, distributed across 13 distinct relation types.
- **Image Associations:** Each node is linked to multiple images, leading to an overall count of approximately 1.5×10^6 images.

These statistics underscore the comprehensive nature of AlphaKG and its suitability as a foundation for multimodal learning tasks. The intricate structure of AlphaKG is exploited by the AlphaKG framework to retrieve and integrate both textual and visual cues during model inference. In particular, the embedding matrices TV and TE are optimized not only to reconstruct the observed relational structure in $\mathcal{D}$ but also to effectively incorporate multimodal signals from $\mathcal{T}_i$ and $\mathcal{I}_i$.

To further elucidate the embedding learning process, consider a scoring function for a valid triple (v_i, e_r, v_j) given by

$$\phi(v_i, e_r, v_j) = f(Tv_i, Te_r, Tv_j),$$

where $f(\cdot)$ is a function that measures the compatibility of the node and relation embeddings. In many models, this function might be defined as a simple dot product, a bilinear form, or even a more complex neural network function. The training objective is to maximize the score for true triples while minimizing it for corrupted ones. This objective is often formalized using a margin-based ranking loss:

$$\mathcal{L} = \sum_{(v_i,e_r,v_j)\in\mathcal{D}} \sum_{(v_i,e_r,v_j')\in\mathcal{D}'} \max\Big(0, \gamma + \phi(v_i, e_r, v_j') - \phi(v_i, e_r, v_j)\Big),$$

 doi:10.20944/preprints202503.1014.v1

where γ is a margin hyperparameter. Although this specific loss function is common in knowledge graph embedding literature, our overall framework, AlphaKG, builds upon such principles while introducing novel multimodal integration strategies.

In summary, AlphaKG is not only a repository of extensive visual and textual information but also a well-structured graph that captures rich semantic relationships among concepts. Its integration into the AlphaKG framework provides a powerful external knowledge source that enhances the capabilities of neural models in processing multimodal information.

4. Experiments

In this section, we present a comprehensive evaluation of the proposed AlphaKG framework on the link prediction task and two downstream applications: named entity recognition (NER) and crosslingual visual verb sense disambiguation (VSD). We merge all experimental analyses into this single section, detailing our experimental setup, reporting extensive quantitative results, and providing thorough discussions on the impact of incorporating additional multimodal features. In our experiments, we compare several baseline models and our hybrid architectures based on graph neural networks augmented with a DistMult layer. We also explore the effect of adding multilingual gloss (text) and image features into node and edge representations via gating mechanisms.

4.1. Experimental Setup and Training Details

We evaluate all models on the link prediction task, i.e., to identify whether a given pair of *head* and *tail* nodes in the knowledge graph are connected by a *relation*. For each observed triplet (v_i, e_r, v_j) in the dataset, we generate k corrupted triplets by substituting the tail v_j (or, equivalently, the head v_i) with a random node such that the resulting triplet is not part of the original graph. We experiment with two settings for the number of corrupted examples, $k \in \{100, 1000\}$. Details on the architectures for the hybrid models (GraphSage+DistMult and GAT+DistMult) are provided in Appendix **??**.

Negative Sampling and Loss Function

All models are trained using negative sampling [25] with the goal of maximizing the probability of positive triplets while minimizing the probability of corrupted triplets. The overall loss function is given by:

$$\begin{aligned}\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{(v_i,e_r,v_j)\in\mathcal{D}} \Big[&-\log\sigma(\phi(v_i,e_r,v_j)) \\ &- \sum_{(v_i,e_r,v'_j)\in\mathcal{D}'} \log\sigma(-\phi(v_i,e_r,v'_j))\Big],\end{aligned} \tag{1}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function, $\mathcal{D}$ is the set of positive triplets, and $\mathcal{D}'$ denotes the set of corrupted (negative) triplets. For models such as TransE and DistMult, we employ both head-corrupted (v'_i, e_r, v_j) and tail-corrupted triplets (v_i, e_r, v'_j).

Scoring Function

Let $\phi(v_i, e_r, v_j)$ be the scoring function for a triplet. For graph-based models, we compute the score using a simple dot product between the final hidden states of the head and tail nodes, i.e.,

$$\phi(v_i, e_r, v_j) = Th_i^\top Th_j,$$

which does not involve any learned relation parameters. For hybrid models, however, the score is computed as:

$$\phi(v_i, e_r, v_j) = Th_i \odot Te_r \odot Th_j,$$

 doi:10.20944/preprints202503.1014.v1

where $\odot$ denotes element-wise multiplication and Te_r is the learned embedding for relation e_r contained in the matrix TE.

When multimodal features are incorporated, the input node embedding Tv_i is replaced by either Tv_i^t (using text features), Tv_i^m (using image features), or $Tv_i^{t,m}$ (using both modalities). In the hybrid models, the relation embedding is similarly updated to Te_r^t, Te_r^m, or $Te_r^{t,m}$ respectively.

Additional Training Details and Hyperparameters

We perform an extensive hyperparameter search for all models. In addition to the learning rate, batch size, and embedding dimensions, we also tune the dropout rate and the number of graph convolution layers. An additional regularization term is added in some experiments to constrain the norm of the node and relation embeddings:

$$\mathcal{L}_{reg} = \lambda \left(\sum_{v_i \in \mathcal{V}} \|Tv_i\|^2 + \sum_{e_r \in \mathcal{E}} \|Te_r\|^2 \right),$$

where λ is a regularization coefficient. All final results are averaged over 5 independent runs, and model selection is performed based on the best validation MRR.

Evaluation Metrics

For link prediction, we use standard metrics: Mean Reciprocal Rank (MRR) and Hits@$\{1,3,10\}$. MRR is computed as the mean of the reciprocal rank of the correct triplet, while Hits@k measures the proportion of correct triplets ranked within the top-k predictions. Increasing the number of negative examples k typically renders the task more challenging.

4.2. Results on Link Prediction Without Additional Features

Tuple-based Models

We first compare tuple-based models on the link prediction task using the full set of negative samples. Table 1 presents results for **TransE**, **DistMult**, and **TuckER** on the AlphaKG test set. As can be seen, TuckER significantly outperforms both TransE and DistMult. In particular, TuckER achieves an MRR of 6.1, Hits@1 of 3.4, Hits@3 of 6.3, and Hits@10 of 11.1, roughly twice the performance of the other two models. These findings are consistent with previous literature [4].

Table 1. Link prediction results on AlphaKG's test set using all negative samples.

	MRR	**Hits@1**	**Hits@3**	**Hits@10**
TransE	3.2	0.2	3.3	8.2
DistMult	3.6	1.9	3.5	7.6
TuckER	*T*6.1	*T*3.4	*T*6.3	*T*11.1

Graph-based vs. Hybrid Models

We now compare the performance of graph-based models and their hybrid counterparts that incorporate a DistMult layer to learn relation embeddings. We evaluate vanilla **GAT** and **GraphSage** as well as the hybrid models **GAT+DistMult** and **GraphSage+DistMult**. Results with 100 negative examples per positive triplet are summarized in Table 2. Note that the column labeled **R** indicates whether relation features are learned (3) or not (7). Although vanilla GAT and GraphSage perform relatively poorly compared to TuckER, their hybrid variants show marked improvements. In particular, GraphSage+DistMult attains an MRR of 78.4, Hits@1 of 56.8, and perfect scores (Hits@3 and Hits@10 at 100.0) under this setting, clearly outperforming TuckER.

 doi:10.20944/preprints202503.1014.v1

Table 2. Link prediction results on AlphaKG's test set using 100 negative samples. **R** denotes whether the model learns relation features.

	R	MRR	Hits@1	Hits@3	Hits@10
TuckER	3	19.0	12.3	17.7	30.0
GAT	7	10.0	3.8	12.6	29.7
+DistMult	3	34.8	13.6	54.4	69.3
GraphSage	7	8.6	2.3	6.4	18.0
+DistMult	3	T78.4	T56.8	T100.0	T100.0

4.3. Results on Link Prediction with Additional Multimodal Features

In this set of experiments, we study the impact of incorporating additional multimodal features from AlphaKG—specifically, textual features from multilingual glosses ($\mathcal{T}_i$) and visual features from images ($\mathcal{I}_i$)—on link prediction performance. These features are integrated into the model through node and edge gating modules.

Table 3 shows a comprehensive comparison of different feature combinations under two settings: using 100 negative examples and 1000 negative examples per positive triplet. For each model, we evaluate configurations with (i) no additional features, (ii) only visual features, (iii) only textual features, and (iv) both textual and visual features. In many cases, the hybrid models benefit considerably from the additional modalities. For instance, when using GraphSage+DistMult with both modalities and 1000 negatives, the best configuration achieves an MRR of 61.6 and Hits@1 of 50.6, outperforming models that use only one type of feature or none at all.

Table 3. Link prediction results on the AlphaKG test set with additional textual ($\mathcal{T}_i$) and visual features ($\mathcal{I}_i$). Best overall scores per metric are shown in bold, and the best scores across feature types for a given model are underlined.

	Features		100 Negative Examples				1000 Negative Examples			
	$\mathcal{T}_i$	$\mathcal{I}_i$	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
GAT +DistMult	7	7	34.8	13.6	54.4	69.3	4.4	0.0	0.0	4.7
	7	3	50.2	43.3	55.6	55.6	29.8	8.9	28.4	55.5
	3	7	69.4	57.2	81.2	81.2	24.3	7.4	26.4	71.2
	3	3	61.8	50.4	63.8	70.1	28.2	9.6	29.3	69.3
GraphSage +DistMult	7	7	78.4	56.8	T100.0	T100.0	38.0	13.4	48.6	T99.9
	7	3	80.7	61.5	T100.0	T100.0	46.9	31.9	47.2	98.3
	3	7	T84.7	T69.5	T100.0	T100.0	36.4	13.8	42.8	T99.9
	3	3	80.7	61.4	T100.0	T100.0	T61.6	T50.6	T63.6	97.2

Discussion on Multimodal Integration

Our results indicate that incorporating multimodal features via the gating mechanisms in AlphaKG generally leads to significant improvements in link prediction performance, particularly for the hybrid models. Although the gains are more pronounced in certain configurations (e.g., GraphSage+DistMult with both features at 1000 negatives), the overall trend is clear: both textual and visual cues contribute complementary information that enhances the learned representations. The improvement in MRR and Hits@k metrics suggests that the additional external features help the model better capture the semantic relationships and visual context embedded in AlphaKG.

4.4. Downstream Task Evaluation

To assess the practical utility of the representations learned via AlphaKG, we integrate them into two downstream tasks: Named Entity Recognition (NER) and Crosslingual Visual Verb Sense Disambiguation (VSD). In both cases, the pretrained AlphaKG node representations are used as external knowledge to augment the base models, and we explore different strategies for integrating these features.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2025 doi:10.20944/preprints202503.1014.v1

4.4.1. Named Entity Recognition (NER)

Datasets and Experimental Model

We evaluate on two NER datasets: **GermEval 2014** for German and **WNUT-17** for English. For WNUT-17, we use a pretrained English BERT model (`bert-large-cased`), while for GermEval 2014 we use a multilingual BERT model (`bert-base-multilingual-cased`). Our baseline NER system is a standard BERT-based classifier where the final token representations Tz_i are fed to a softmax layer:

$$T\hat{y}_i = \texttt{softmax}(TW^n Tz_i), \tag{2}$$

with TW^n as the classification weight matrix.

To incorporate external knowledge, we retrieve the top-k closest nodes from AlphaKG using its sentence retrieval model. Two strategies are investigated:

1. **Concatenation (concat):** The retrieved node representation Th_i^{RET} is concatenated with the token representation:

$$T\hat{y}_i = \texttt{softmax}(TW^n [Tz_i; TW^{RET} Th_i^{RET}]), \tag{3}$$

where TW^{RET} projects the retrieved node to the appropriate dimension.

2. **Attention (attend):** An attention mechanism is applied over the top-5 retrieved nodes, with Tz_i serving as the query:

$$Ta = \texttt{Attention}(Tz_i, \{Th_i^{RET}\}_{k=1}^{5}), \tag{4}$$

$$T\hat{y}_i = \texttt{softmax}(TW^n [Tz_i; TW^{RET} Ta]). \tag{5}$$

Quantitative Results

Table 4 reports the NER performance on the WNUT-17 (EN) and GermEval (DE) test sets. For English, the baseline achieves an F1 score of 47.4. With the addition of AlphaKG representations via the attention mechanism over the top-5 retrieved nodes (using node features without any additional multimodal data), the F1 score improves to 48.1, a 0.7% absolute gain. For German, the baseline already performs well with an F1 of 86.1, and integrating the AlphaKG representations via concatenation slightly boosts the score to 86.4. These improvements, though moderate, validate the benefit of enriching NER systems with structured external knowledge.

Table 4. NER results on the WNUT-17 (EN) and GermEval (DE) test sets. The incorporation of AlphaKG representations improves the F1 score by up to 0.7% over the baseline.

		Precision	Recall	F1 Score
EN	**Baseline**	58.4	39.9	47.4
	+concat $Th_i^{\mathbf{IMG}}$	57.1	39.1	46.4
	+attend $\{Th_i^{\mathbf{NODE}}\}_{k=1}^{5}$	T61.5	39.5	T48.1
DE	**Baseline**	86.0	86.2	86.1
	+concat $Th_i^{\mathbf{NODE}}$	T86.2	T86.6	T86.4
	+attend $\{Th_i^{\mathbf{TXT+IMG}}\}_{k=1}^{5}$	85.7	86.0	85.9

4.4.2. Crosslingual Visual Verb Sense Disambiguation (VSD)

Dataset and Task Description

We evaluate on the **MultiSense** dataset [12], which comprises 9,504 images associated with 55 English verbs and their corresponding translations in German (154 unique German verbs). Each sample in the dataset includes an ambiguous English verb, a textual context describing the verb, and an image that visually illustrates the action. The task is to disambiguate the correct translation in German based on both textual and visual context.

Baseline Model and Integration of AlphaKG

Our baseline model encodes the visual modality using a pretrained ResNet-152 [17], extracting the 2048-dimensional activation from the *pool5* layer as visual features Tz_i^m. Concurrently, the ambiguous English verb along with its context is encoded using a pretrained BERT model (`bert-large-cased`), with the resulting token embedding serving as the textual feature Tz_i^t. These features are projected to lower dimensions via learned projection matrices TW^m and TW^t, respectively, and then concatenated and passed through a hidden layer with ReLU activation:

$$Th_i = \texttt{ReLU}(TW^h[TW^m Tz_i^m; TW^t Tz_i^t]), \tag{6}$$

followed by a final projection to the output space:

$$T\hat{y}_i = \texttt{softmax}(TW^o Th_i). \tag{7}$$

To integrate external knowledge, we retrieve the top-1 nearest node representation Th_i^{RET} from the AlphaKG using a sentence retrieval model that processes the concatenation of the English verb and its textual context. The hidden layer is then redefined as:

$$Th_i = \texttt{ReLU}(TW^h[TW^m Tz_i^m; TW^t Tz_i^t; Th_i^{RET}]), \tag{8}$$

where the hidden layer size is adjusted to maintain a comparable number of parameters with and without the additional feature.

Quantitative Results and Analysis

Table 5 summarizes the accuracy on the MultiSense test set (German). Our baseline model attains an accuracy of 94.4%, significantly outperforming the earlier reported results of 55.6% in [12], which we attribute to improvements in model design and data preprocessing. When integrating AlphaKG representations, we observe that augmenting with node features (+$Th_i^{\textbf{NODE}}$) raises the accuracy to 96.8%, while incorporating image features (+$Th_i^{\textbf{IMG}}$) further boosts the accuracy to 97.2%. These gains indicate that the additional multimodal and structured information provided by AlphaKG can enhance crosslingual VSD performance, especially when the base model is already highly competitive.

Table 5. Accuracy on the MultiSense test set (German). The addition of AlphaKG representations leads to improvements over the strong baseline.

	Accuracy
[12]	55.6
Our Baseline	94.4
+$Th_i^{\textbf{NODE}}$	96.8
+$Th_i^{\textbf{IMG}}$	**97.2**

4.5. Summary of Experimental Findings

Across our experiments, the proposed AlphaKG framework consistently improves link prediction performance on AlphaKG as well as downstream task performance on both NER and crosslingual VSD. In link prediction, hybrid models that combine graph neural network architectures with a DistMult layer (notably GraphSage+DistMult) yield substantial gains when augmented with additional multimodal features. For downstream tasks, even modest improvements in F1 and accuracy metrics demonstrate the practical benefits of integrating structured external knowledge into state-of-the-art models.

Furthermore, the incorporation of additional textual and visual features through well-designed gating mechanisms enables the models to capture richer semantic and visual context, leading to better generalization and improved task performance. The experimental results indicate that leveraging a

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2025 doi:10.20944/preprints202503.1014.v1

multimodal KG such as AlphaKG within the AlphaKG framework is a promising avenue for enhancing various natural language processing and computer vision applications.

Overall, our comprehensive evaluation validates the effectiveness of AlphaKG in both intrinsic (link prediction) and extrinsic (NER, VSD) tasks, setting the stage for future research on integrating external multimodal knowledge into neural architectures.

5. Conclusions and Future Directions

In this work, we presented a systematic investigation comparing various tuple-based and graph-based architectures for learning robust multimodal representations for the AlphaKGknowledge graph. Our study revealed that integrating the rich visual information (illustrative images) and descriptive textual glosses available at each node significantly enhances the quality of node and entity embeddings, as measured on the link prediction task. In particular, our best-performing method—AlphaKG, a hybrid approach that merges the strengths of both tuple- and graph-based paradigms—demonstrated its efficacy by yielding substantial improvements in downstream applications. For example, on crosslingual visual verb sense disambiguation, AlphaKG improved accuracy by 2.5% compared to a strong baseline, while in multilingual named entity recognition, performance gains ranged from 0.3% to 0.7% in F1 score. These results were achieved using relatively simple downstream architectures, suggesting that further gains might be obtained by exploring more sophisticated integration strategies.

Beyond our empirical findings, we introduced an enhanced training objective that combines standard negative sampling with an auxiliary regularization term designed to encourage smoothness in the learned embedding spaces. This additional refinement underscores the potential of carefully designed loss functions to further improve the quality of multimodal representations in the AlphaKG framework.

Our findings motivate several promising avenues for future research. First, it would be valuable to extend our evaluation to a broader set of downstream tasks. For instance, integrating AlphaKG with vision-centric tasks—such as object detection or scene understanding—could reveal further benefits of leveraging structured multimodal knowledge. Additionally, challenging generative tasks like image captioning, where the fusion of visual and textual modalities is critical, may also benefit from the rich representations produced by AlphaKG.

Another promising direction involves applying our framework to other knowledge graphs that encode different types of information. For example, incorporating commonsense knowledge from resources such as ConceptNet could enable AlphaKG to handle even more diverse and complex reasoning scenarios. In this case, one could adapt our hybrid training objective to jointly optimize for multiple types of semantic relationships, thereby learning a unified representation that captures both factual and commonsense dimensions.

Furthermore, an exciting line of inquiry is the integration of structured knowledge graph representations within large-scale retrieval-based language models. By dynamically retrieving and incorporating external structured knowledge during inference, such models could achieve enhanced contextual understanding without the need to store all knowledge implicitly within their parameters. This could be realized through a modular approach where AlphaKG serves as an external memory component that interfaces with large pretrained models via attention-based mechanisms.

In summary, our work demonstrates that the AlphaKG framework effectively bridges visual and textual modalities within a structured knowledge graph, yielding improved performance across both intrinsic (link prediction) and extrinsic (NER, VSD) evaluation tasks. The encouraging results and the modular nature of our approach open up a wide spectrum of future work, ranging from the incorporation of additional data modalities and knowledge sources to the integration with large-scale language models. We anticipate that these directions will not only further enhance the efficiency and accuracy of multimodal representation learning but also contribute to the development of more robust and adaptable AI systems.

References

1. Houda Alberts, Ningyuan Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2021. VisualSem: a high-quality knowledge graph for vision and language. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 138–152, Punta Cana, Dominican Republic. Association for Computational Linguistics.
2. Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.
3. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointlylearning to align and translate. In *International Conference on Learning Representations (ICLR)*.
4. Ivana Balazevic, Carl Allen, and Timothy Hospedales. 2019. TuckER: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5185–5194, Hong Kong, China. Association for Computational Linguistics.
5. Peter Battaglia, Jessica Blake Chandler Hamrick, Victor Bapst, Alvaro Sanchez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andy Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Jayne Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv*.
6. Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).
7. Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2787–2795. Curran Associates, Inc.
8. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.
9. Sreerupa Das, C. Lee Giles, and Guo zheng Sun. 1992. Learning context-free grammars: Capabilities and limitations of a recurrent neural network with an external stack memory. In *CONFERENCE OF THE COGNITIVE SCIENCE SOCIETY*, pages 791–795. Morgan Kaufmann Publishers.
10. Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.
11. William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.
12. Spandana Gella, Desmond Elliott, and Frank Keller. 2019. Cross-lingual visual verb sense disambiguation.
13. Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines.
14. Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.
15. Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1024–1034. Curran Associates, Inc.
16. Xu Han, Shulin Cao, Lv Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. Openke: An open toolkit for knowledge embedding. In *Proceedings of EMNLP*.
17. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
18. Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.
19. Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

20. Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
21. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.
22. Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.
23. Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.
24. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.
25. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.
26. Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.
27. Aida Nematzadeh, Sebastian Ruder, and Dani Yogatama. 2020. On memory in human and artificial language processing systems. In *Proceedings of the Bridging AI and Cognitive Science Workshop at ICLR 2020*.
28. Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.
29. Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
30. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
31. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252.
32. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*.
33. Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling relational data with graph convolutional networks.
34. Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.
35. Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. 2018. Learning visual knowledge memory networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7736–7745.
36. Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.
37. Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers.
38. Ledyard Tucker. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.
39. Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.
40. Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3266–3280. Curran Associates, Inc.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2025 doi:10.20944/preprints202503.1014.v1

41. Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.
42. Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. 2018. M3: Multimodal memory modelling for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7512–7520.
43. Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.
44. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.
45. Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406.
46. Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
47. Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graph.
48. Zheng Zeng, R. M. Goodman, and P. Smyth. 1994. Discrete recurrent neural networks for grammatical inference. *IEEE Transactions on Neural Networks*, 5(2):320–330.
49. Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network. In *Proceedings of the Web Conference 2021*. 1673–1685.
50. Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management CIKM*. 729–738.
51. 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.
52. 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).
53. 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.
54. 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.
55. 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)*.
56. 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.
57. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
58. 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).

59. 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. 10.1007/s00530-010-0182-0.
60. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. 10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
61. 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/.
62. 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.
63. 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.
64. 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. 10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
65. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
66. 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.
67. 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.
68. A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.

80. 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.
81. 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.
82. 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.
83. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
84. 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.
85. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
86. 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.
87. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
88. 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.
89. 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. 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
90. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
91. 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.
92. 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.
93. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
94. 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.
95. 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.
96. 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.
97. 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.
98. 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.
99. 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.
100. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
101. 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.

102. 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.
103. 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.
104. 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.
105. 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.
106. 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.
107. 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.
108. 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.
109. 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.
110. 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.
111. 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.
112. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
113. 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.
114. 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.
115. 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.
116. 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.
117. 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.
118. 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.
119. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.