Article

Not peer-reviewed version

Adaptive Semantic Fusion for Contextual Image Captioning

Ethan Cooper , Lobry Hsu , Sofia Ramirez *

Posted Date: 21 January 2025

doi: 10.20944/preprints202501.1491.v1

Keywords: Image Captioning; Visual Semantics; Language Modeling; Transformer; Semantic Refinement

 doi:10.20944/preprints202501.1491.v1

Article

Adaptive Semantic Fusion for Contextual Image Captioning

Ethan Cooper, Lobry Hsu and Sofia Ramirez *

Bond University
* Correspondence: sofia.ramirez@bond.edu.au

Abstract: The automatic generation of textual descriptions from visual data is a fundamental yet challenging task that requires the seamless integration of image understanding and sophisticated language modeling. It involves not only identifying and interpreting complex visual elements but also effectively mapping them to coherent and contextually relevant textual representations. In this paper, we propose a novel framework called the *Dynamic Iterative Refinement Model (DIRM)*, which addresses these challenges by dynamically adjusting the output vocabulary of the language decoder through decoder-driven visual semantics. By leveraging a dynamic gating mechanism and scatter-connected mappings, DIRM implicitly learns robust associations between visual tag words and corresponding image regions. This enables the model to generate captions that are semantically rich, contextually accurate, and capable of capturing fine-grained visual details. The proposed framework introduces a multi-step refinement strategy, wherein visual concepts are iteratively refined and integrated into the decoding process to enhance semantic alignment. Furthermore, DIRM incorporates a visual-concept vocabulary to guide the generation of descriptive keywords, effectively bridging the gap between high-level visual semantics and linguistic coherence. These innovations allow the model to adaptively focus on salient image features, reducing reliance on generic language patterns and promoting content-specific caption generation. Extensive experiments conducted on the MS-COCO dataset demonstrate the superiority of DIRM over existing visual-semantic-based approaches. The framework achieves state-of-the-art results across multiple evaluation metrics, including BLEU, CIDEr, and SPICE, reflecting its ability to generate captions with enhanced fluency, relevance, and descriptive depth. Additionally, qualitative analysis highlights the model's proficiency in capturing nuanced visual relationships and producing detailed captions that align closely with human annotations. Our work represents a significant advancement in image captioning, paving the way for future research in dynamic visual-linguistic integration and multimodal generation tasks.

Keywords: image captioning; visual semantics; language modeling; transformer; semantic refinement

---

1. Introduction

Image captioning, often referred to as image-to-text translation, is a fundamental task at the intersection of computer vision and natural language processing [2,4,7,12]. It aims to generate meaningful and descriptive textual captions for given images by integrating scene understanding and language generation.

Despite significant progress, many existing approaches tend to over-rely on frequently occurring n-grams in the training data, leading to captions that may not accurately reflect the content of the image [7]. To address this issue, visual concept prediction methods have been proposed to bridge the semantic gap by extracting meaningful tags from images and incorporating them into the caption generation process [7,8,26–28]. These methods typically predict the likelihood of semantic concepts derived from a predefined image-grounded vocabulary, enabling better alignment between visual content and generated captions.

 doi:10.20944/preprints202501.1491.v1

However, most prior approaches utilize Long Short-Term Memory (LSTM) networks [11] as language decoders, which, while effective, suffer from sequential processing limitations that hinder parallelization. In contrast, the Transformer architecture [22], originally developed for neural machine translation, has demonstrated exceptional parallelism and adaptability across various applications [6, 9,18,20,24]. By leveraging self-attention mechanisms, Transformers offer a powerful alternative for image captioning tasks.

Humans typically adopt a dual-cognitive process when describing images, alternating between "thinking" of words to construct coherent sentences and "looking" at image regions to capture specific content. This insight has inspired adaptive attention models [17], which observe that not all words in a caption are directly grounded in visual content. Functional words, such as determiners or conjunctions, can often be inferred through language modeling, while content words, such as nouns and verbs, require explicit visual grounding.

Building on this understanding, we propose a novel framework called *Dynamic Iterative Refinement Model (DIRM)*. This framework integrates a language-decoder-guided gating mechanism to dynamically modulate visual semantic vectors, ensuring that captions are both contextually coherent and semantically aligned with the image content. The key contributions of our work are summarized as follows:

- We introduce a novel dynamic gating mechanism that refines the language decoder's output using visual semantic vectors, enabling the generation of more descriptive and accurate captions.
- A scatter connection layer is proposed to effectively align visual-semantic features with the decoder's vocabulary, ensuring robust semantic representation.
- Extensive experiments on the MS-COCO dataset demonstrate the superiority of our approach, achieving state-of-the-art performance compared to existing visual-semantic-based captioning models.

2. Related Work

The task of image captioning has attracted substantial attention due to its potential applications in assistive technologies, content retrieval, and multimedia generation. This section reviews key advancements in image captioning methodologies, focusing on visual-semantic alignment, language modeling, and the integration of attention mechanisms.

2.1. Image Captioning with Visual-Semantic Alignment

Visual-semantic alignment is a foundational aspect of image captioning. Early approaches in this domain focused on encoding image features using convolutional neural networks (CNNs) and decoding captions with recurrent neural networks (RNNs). These models leveraged global image features, but their coarse-grained nature often resulted in generic captions. To address this limitation, researchers introduced object-level visual features, which involved detecting and encoding objects in an image to provide more fine-grained semantic representations [2]. The use of object detection models such as Faster R-CNN [21] enabled these systems to extract localized features, thereby improving caption quality. Concurrently, visual concept prediction methods [8,28] sought to predefine semantic tags associated with image content, which were then used as additional inputs to enhance caption generation. Despite these advancements, many models struggled to dynamically integrate semantic representations with language generation. This motivated the development of adaptive attention mechanisms [17], which allowed models to selectively focus on relevant image regions during different stages of caption generation. These approaches demonstrated that incorporating visual-semantic alignment at both global and local levels significantly enhances the quality of generated captions.

 doi:10.20944/preprints202501.1491.v1

2.2. *Transformer-Based Architectures for Captioning*

The advent of Transformer architectures [22] has revolutionized many areas of natural language processing and computer vision, including image captioning. Unlike RNNs, Transformers employ self-attention mechanisms to process entire sequences in parallel, making them more computationally efficient and scalable. Transformers have been successfully applied to image captioning in models such as Image Transformer [20] and Transformer-based encoder-decoder frameworks. These models leverage pre-trained vision encoders, such as Vision Transformers (ViT), to extract image embeddings, which are then processed by the Transformer decoder to generate captions. Recent works have focused on enhancing Transformer architectures for captioning by incorporating additional semantic cues. For instance, the OSCAR model uses object tags as auxiliary inputs to improve visual grounding. Similarly, M2 Transformer introduces multi-head cross-attention layers to better integrate visual and textual features. These methods have demonstrated state-of-the-art performance on benchmark datasets such as MS-COCO.

2.3. *Reinforcement Learning and Evaluation Metrics in Captioning*

Traditional supervised learning approaches for image captioning optimize cross-entropy loss, which does not directly align with evaluation metrics like BLEU, METEOR, ROUGE, and CIDEr. As a result, reinforcement learning (RL) techniques, particularly policy gradient methods, have been adopted to directly optimize these metrics. Self-critical sequence training (SCST) is a widely used RL-based approach in image captioning. It utilizes the model's own predictions as baselines to compute rewards, enabling it to generate captions that better align with human judgments. Additionally, RL has been used to optimize diverse objectives, such as diversity and coherence in caption generation. Despite the success of RL, challenges remain in balancing metric optimization with linguistic quality. Recent efforts have explored hybrid loss functions that combine RL with supervised learning to achieve a balance between human-like fluency and metric-based optimization.

2.4. *Limitations of Existing Approaches*

While significant progress has been made, existing image captioning models face notable limitations. Many approaches rely on static vocabularies, which restrict their ability to generalize to unseen visual concepts. Moreover, the focus on global evaluation metrics often overlooks the importance of fine-grained linguistic coherence and semantic richness. Addressing these challenges requires models that can dynamically refine semantic representations and adaptively align them with language generation. Our proposed framework builds upon these advancements by introducing a novel method for iterative refinement of visual semantics, enabling more descriptive and contextually accurate captions.

3. Methodology

This section details the proposed *Dynamic Iterative Refinement Model (DIRM)*, which integrates a Transformer-based relational encoder, a semantic-aware decoder, and a visual-concept refinement mechanism to generate high-quality captions. The model is designed to iteratively refine semantic representations, enabling better alignment between visual content and textual descriptions.

3.1. *Relational Encoding with Transformers*

Inspired by [2], we adopt Faster R-CNN [21] with ResNet-101 [10] as the base object detector. The detector generates M object proposals using a Region Proposal Network (RPN) and computes mean-pooled convolutional features for each proposal, resulting in a 2048-dimensional feature vector per object.

 doi:10.20944/preprints202501.1491.v1

Let the extracted features for all proposals be denoted as:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M]^\top \in \mathbb{R}^{M \times 2048}.$$

To reduce the feature dimensionality, we apply a fully connected layer:

$$\mathbf{X}_r = \mathbf{X}\mathbf{W}_\mathbf{r},$$

where $\mathbf{W}_\mathbf{r} \in \mathbb{R}^{2048 \times 512}$ is a learnable weight matrix, producing $\mathbf{X}_r \in \mathbb{R}^{M \times 512}$. Next, a Transformer encoder with N_{enc} layers processes the reduced features to capture object-object relationships [29]:

$$\mathbf{F} = \text{TransformerEncoder}(\mathbf{X}_r).$$

The output, $\mathbf{F} \in \mathbb{R}^{M \times 512}$, represents object-wise relational features that serve as input to the caption generation module.

3.2. Semantic-Aware Caption Decoder

The caption generator employs a Transformer decoder with N_{dec} layers. For a given caption sequence of length T, the decoder operates as follows: 1. Input words are first embedded via a word embedding layer and enriched with positional encodings:

$$\mathbf{E}_t = \mathbf{W}_e \mathbf{w}_t + \mathbf{P}_t, \quad t = 1, 2, \ldots, T,$$

where $\mathbf{W}_e \in \mathbb{R}^{|\mathcal{V}| \times 512}$ is the embedding matrix, and $\mathbf{P}_t$ encodes positional information. 2. A masked self-attention sublayer computes contextual relationships within the sequence, attending only to prior tokens. 3. A cross-attention sublayer integrates multi-modal information by aligning $\mathbf{E}_t$ with visual features $\mathbf{F}$:

$$\mathbf{C}_t = \text{softmax}\left(\frac{\mathbf{E}_t \mathbf{F}^\top}{\sqrt{d_k}}\right)\mathbf{F}.$$

4. Finally, a feed-forward network (FFN) processes the attention outputs:

$$\mathbf{H}_t = \text{FFN}(\mathbf{C}_t).$$

Each sublayer is wrapped with residual connections [10] and layer normalization [3], ensuring stable training and effective gradient flow.

3.3. Dynamic Visual-Concept Refinement

Visual Concept Layer

To capture semantic concepts, we construct an image-grounded vocabulary $\mathcal{V}_{\text{tag}}$ comprising the K most frequent nouns, verbs, and adjectives in the dataset [8], with $K = 1000$. For each image, visual features $\mathbf{F}$ are transformed into a K-dimensional concept vector:

$$\mathbf{v} = \sigma(\text{concat}(\mathbf{f}_1 \mathbf{W}_0, \mathbf{f}_2 \mathbf{W}_0, \ldots, \mathbf{f}_M \mathbf{W}_0)),$$

where $\mathbf{W}_0 \in \mathbb{R}^{512 \times (K/M)}$ and $\sigma(x) = 1/(1 + \exp(-x))$. The resulting $\mathbf{v} \in \mathbb{R}^K$ represents the likelihood of each concept in $\mathcal{V}_{\text{tag}}$.

Decoder-Guided Refinement

To dynamically modulate the visual-concept vector, we compute a decoder-guided gating mechanism:

$$\mathbf{g}_t = \sigma(\mathbf{W}_1 \mathbf{h}_t + \mathbf{W}_2 \mathbf{F}),$$

 doi:10.20944/preprints202501.1491.v1

where $\mathbf{h}_t \in \mathbb{R}^{512}$ is the t-th decoder hidden state, and $\mathbf{g}_t \in \mathbb{R}^K$ is a gating vector. The refined concept representation is:

$$\mathbf{o}_t = \mathbf{g}_t \odot \mathbf{v}.$$

Scatter-Connected Mapping

The refined concepts are integrated with the decoder vocabulary using scatter mapping. For each word $j \in \mathcal{V}_{\text{cap}}$:

$$\mathbf{h}_t[j] = \begin{cases} \mathbf{h}_t[j] + \mathbf{o}_t[k], & \text{if } \mathcal{V}_{\text{cap}}(j) = \mathcal{V}_{\text{tag}}(k), \\ \mathbf{h}_t[j], & \text{otherwise.} \end{cases}$$

The final vocabulary distribution is computed using a softmax function.

3.4. Training with Reinforcement Learning

The entire model is optimized using a hybrid loss function comprising cross-entropy and reinforcement learning (RL). For RL, we employ a policy gradient approach, where the reward is based on CIDEr [23] scores:

$$L_{\text{RL}} = -\mathbb{E}_{\pi_\theta}[R(\hat{y})],$$

where $R(\hat{y})$ is the CIDEr score for the generated caption $\hat{y}$. The total loss is:

$$L = L_{\text{XE}} + \lambda L_{\text{RL}},$$

where λ balances supervised and RL losses.

4. Experiments

This section presents a comprehensive evaluation of the proposed *Dynamic Iterative Refinement Model (DIRM)*. We detail the experimental setup, quantitative results, qualitative analysis, and additional insights through ablation studies and discussions.

4.1. Experimental Setup

Dataset and Evaluation

The MS-COCO dataset [16] is employed for evaluating DIRM. Following the *Karpathy* split, the dataset includes 113,287 images for training, 5,000 for validation, and 5,000 for testing, with each image paired with 5 human-generated captions. This diverse dataset provides a robust foundation for evaluating caption generation quality across varied scenes and contexts. We adopt the commonly used evaluation metrics BLEU [19], ROUGE-L [15], METEOR [5], CIDEr-D [23], and SPICE [1] to assess the generated captions. These metrics collectively evaluate fluency, semantic relevance, and adequacy.

Implementation Details

DIRM is implemented using PyTorch. The embedding dimension (D) is set to 512, with the encoder and decoder each comprising 3 Transformer layers for efficient yet effective learning. The batch size is set to 50, and the number of attention heads is 8. The feed-forward network (FFN) has a hidden size of 2048. To limit computational overhead, the maximum number of extracted object features per image is set to 50. Word embeddings are initialized randomly.

Optimization is performed using the Adam optimizer [13] with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The dropout rate is set to 0.1 to prevent overfitting, and early stopping is applied with a patience of 5 epochs. Beam search decoding is used with a beam width of 5. Prior to reinforcement learning (RL), the model is pretrained using supervised learning with cross-entropy loss. All experiments are conducted on a single NVIDIA Tesla V100 GPU.

 doi:10.20944/preprints202501.1491.v1

Training Workflow

The training consists of two stages: 1. **Supervised Pretraining**: The model is trained using cross-entropy loss to learn the mapping between images and captions. 2. **Reinforcement Learning Fine-tuning**: The pretrained model is further optimized using the CIDEr score as a reward signal. The REINFORCE algorithm [?] with a baseline strategy is used to stabilize training.

4.2. Quantitative Results

The quantitative evaluation results on the MS-COCO dataset are shown in Table 1. The proposed DIRM achieves superior performance across all evaluation metrics, demonstrating its ability to generate captions that are both semantically accurate and contextually appropriate. Compared to baseline models, DIRM exhibits significant improvements, particularly in CIDEr and SPICE, which reflect semantic richness and content relevance.

Table 1. Comparison of DIRM and state-of-the-art models on MS-COCO. Metrics include BLEU, METEOR, ROUGE-L, CIDEr, and SPICE. DIRM achieves notable improvements in CIDEr and SPICE.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
SemAttn [28]	0.709	0.537	0.402	0.304	0.243	-	-	-
Att-CNN+LSTM [26]	0.740	0.560	0.420	0.310	0.260	-	0.940	-
LSTM-C [27]	-	-	-	-	-	0.230	-	-
Skeleton Key [25]	0.673	0.489	0.355	0.259	0.247	0.489	0.966	0.196
SCN-LSTM [8]	0.728	0.566	0.433	0.330	0.257	-	1.041	-
Bridging [7]	-	-	-	0.330	0.264	**0.586**	1.066	-
DIRM (Ours)	**0.802**	**0.645**	**0.499**	**0.378**	**0.283**	0.580	**1.272**	**0.225**

Improved Semantic Understanding

The SPICE score of DIRM highlights its effectiveness in capturing semantic structures. This improvement stems from the visual-concept refinement module, which dynamically adjusts the semantic representations based on image content and linguistic context.

4.3. Ablation Study

To evaluate the contribution of key components in DIRM, we perform an ablation study. Table 2 presents the results of removing the visual-concept refinement module and scatter-connected mapping. Both components significantly impact performance, especially in CIDEr and SPICE scores, emphasizing their importance in enhancing semantic understanding.

Table 2. Ablation study of DIRM, showing the impact of removing key components.

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
DIRM (Full)	**0.802**	**0.378**	**0.283**	**0.580**	**1.272**	**0.225**
w/o Refinement	0.786	0.366	0.277	0.570	1.202	0.210
w/o Scatter-Connection	0.771	0.358	0.270	0.560	1.153	0.200

4.4. Qualitative Analysis

Contextual Accuracy

DIRM generates captions with high contextual relevance. For instance, in images containing multiple objects or complex scenes, the model accurately describes key interactions, avoiding common errors such as object misidentification or redundant phrases.

 doi:10.20944/preprints202501.1491.v1

Enhanced Semantic Detail

Compared to baseline methods, DIRM demonstrates a superior ability to capture fine-grained details. For example, it can correctly describe background elements or subtle visual cues, such as "a person holding an umbrella near a bustling market," which are often overlooked by simpler models.

Generalizability to Diverse Scenes

The implicit visual-concept modeling in DIRM enhances its generalization capability. It effectively generates accurate captions even for images containing rare objects or previously unseen combinations of objects.

4.5. Error Analysis and Future Directions

While DIRM achieves state-of-the-art results, there are some limitations. For example, the model occasionally generates captions with minor grammatical errors or overly verbose descriptions. Future work could explore incorporating syntactic constraints and more advanced language modeling techniques to address these challenges.

5. Conclusions and Future Directions

In this work, we introduced a novel image captioning framework, referred to as the *Dynamic Iterative Refinement Model (DIRM)*, designed to generate semantically rich and contextually appropriate descriptive sentences by dynamically integrating visual concepts and linguistic features. Our approach leverages a combination of visual-concept refinement and scatter-connected mappings, enabling more precise alignment between visual content and generated text. The experimental results on the MS-COCO dataset demonstrated the superiority of DIRM over existing state-of-the-art methods, achieving significant improvements across multiple evaluation metrics, including BLEU, CIDEr, and SPICE.

The key contributions of this work can be summarized as follows:

- We proposed a dynamic refinement mechanism that integrates visual concepts into the captioning process, allowing the model to focus on semantically important visual elements.
- A scatter-connected mapping strategy was introduced, effectively aligning the visual-concept vocabulary with the decoder's linguistic output, resulting in enhanced semantic accuracy.
- Extensive experiments validated the effectiveness of DIRM, highlighting its ability to generate more descriptive, accurate, and contextually relevant captions compared to baseline models.

5.1. Future Directions

Despite the promising results achieved by DIRM, there remain several areas for improvement and exploration. These include:

1. Enhancing Generalization to Diverse Datasets

While DIRM performs exceptionally well on the MS-COCO dataset, future research could explore its generalizability to other datasets with varying domain-specific challenges, such as Flickr 30k [**?**] or domain-specific datasets like VizWiz [**?**]. Adapting the model to diverse visual styles and less-structured annotations could provide further insights into its robustness.

2. Incorporating Multimodal Information

DIRM currently focuses on static image-to-text translation. Expanding this framework to handle multimodal inputs, such as video sequences or audio-visual data, could significantly broaden its applicability. Temporal dynamics and audio cues could enhance the descriptive quality of generated captions in complex scenarios.

3. Refining Linguistic Coherence and Style

While the scatter mapping mechanism effectively improves semantic alignment, future work could explore methods to further refine linguistic coherence and stylistic diversity in generated captions. Techniques such as controllable text generation or large-scale pretraining on diverse corpora may help achieve this goal.

4. Real-Time and Low-Resource Adaptation

Improving the computational efficiency of DIRM for real-time applications is a practical direction for future work. Additionally, investigating methods to reduce the reliance on large annotated datasets, such as semi-supervised or unsupervised learning approaches, could make the model more accessible for resource-constrained applications.

5. Explainable and Trustworthy Image Captioning

As the demand for transparent AI systems increases, integrating explainability into the DIRM framework could provide users with better insights into how captions are generated. Visualizing attention distributions or identifying key features responsible for specific caption elements could enhance trust and interpretability.

In conclusion, the proposed DIRM framework represents a significant step forward in bridging the gap between visual content and textual descriptions. Its dynamic refinement capabilities, coupled with its robust performance, pave the way for future advancements in image captioning and related multimodal tasks.

References

1. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.
2. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.
3. Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
4. Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667.
5. Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.
6. Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE.
7. Zhihao Fan, Zhongyu Wei, Siyuan Wang, and Xuan-Jing Huang. 2019. Bridging by word: Image grounded vocabulary construction for visual captioning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6514–6524.
8. Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639.
9. Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems*, pages 11179–11189.
10. 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.
11. Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

12. Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.
13. Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
14. Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8928–8937.
15. Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.
16. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
17. Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.
18. Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.
19. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
20. Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. *arXiv preprint arXiv:1802.05751*.
21. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.
22. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
23. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
24. Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, pages 1–5.
25. Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W Cottrell. 2017. Skeleton key: Image captioning by skeleton-attribute decomposition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7272–7281.
26. Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212.
27. Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6580–6588.
28. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.
29. Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, et al. 2018. Deep reinforcement learning with relational inductive biases. In *Proceedings of international conference on learning representations*.
30. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553): 436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
31. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
32. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.

33. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
34. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
35. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
36. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
37. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100: 101919, 2023.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
52. 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.

53. 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.
54. 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.
55. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
56. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
57. 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.
58. 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.
59. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
68. 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.
69. 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.
70. 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.
71. 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.

72. 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.
73. 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.
74. 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.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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.
82. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.
83. Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. 2024. LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding Reasoning and Planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26428–26438.