Conditioned Visual Captioning with Spatially-Aware Multimodal Modeling

Poppy Marwood *, Lobry Hsu, Harry Taylor

Posted Date: 19 February 2025

doi: 10.20944/preprints202502.1538.v1

Keywords: Multimodal Captioning; Scene Text Recognition; Image Understanding

 doi:10.20944/preprints202502.1538.v1

Article

Conditioned Visual Captioning with Spatially-Aware Multimodal Modeling

Poppy Marwood [1,*], Lobry Hsu [2] and Harry Taylor [3]

[1] Bond University
[2] Bond University; lobryhsu@gmail.com
[3] Bond University; harry.taylor@bond.edu.au
* Correspondence: poppymarwood@bond.edu.au

Abstract: Understanding scene text in images is crucial for various real-world applications, especially for visually impaired individuals who rely on comprehensive and contextually relevant descriptions. Traditional text-aware image captioning systems, however, fail to generate personalized captions that cater to diverse user inquiries. To bridge this gap, we introduce a novel and challenging task called Question-driven Text-aware Image Captioning (Q-TAG), where captions are dynamically tailored based on specific user queries. Given an image embedded with multiple scene texts, the system must comprehend user-posed questions, extract relevant textual and visual features, and construct fluent, contextually enriched captions. To facilitate research in this domain, we construct benchmark datasets derived from existing text-aware captioning datasets through an automated data augmentation pipeline. These datasets provide comprehensive quadruples of <image, initial coarse caption, control questions, enriched captions>. We propose an advanced model, Q-TAG, which integrates a Spatially-aware Multimodal Encoder to fuse object-region and scene-text features while considering their geometric relationships. Additionally, a Question-driven Feature Selector filters the most relevant visual-textual elements based on user queries. Finally, a Multimodal Fusion Decoder synthesizes these components to generate highly informative captions. Experimental evaluations demonstrate that Q-TAG surpasses strong baselines in both captioning quality and question relevance, producing more diverse and context-sensitive descriptions than existing models.

Keywords: Multimodal Captioning; Scene Text Recognition,; Image Understanding

1. Introduction

Scene text is an essential source of information, often providing critical details such as product labels, book titles, and road signs. Recognizing and describing such textual content is crucial for a wide range of applications, particularly in assistive technologies for visually impaired users [2,10,14,18,23,28,32]. Recent advancements in text-aware image captioning [24,29,31,33,35] have significantly improved the integration of scene text into captions. However, these models typically generate static descriptions without adapting to user-specific information needs.

In scenarios where an image contains multiple scene texts, summarizing all textual content in a single caption can be cumbersome and inefficient. Prior research [19] indicates that visually impaired users prefer a progressive interaction paradigm, where they receive an initial high-level summary followed by detailed captions based on their specific queries. For example, when encountering an image of a book, users might ask targeted questions such as "Who is the author?" or "What is the title?" instead of receiving a verbose caption listing all textual elements indiscriminately.

To address this limitation, we introduce the Question-driven Text-aware Image Captioning (Q-TAG) task, wherein users pose scene-text-related questions following an initial general caption, prompting the model to generate more personalized and informative captions. This differs from the Text-based Visual Question Answering (TextVQA) [4,13,15,25], which typically produces brief textual answers to single questions without incorporating scene understanding. Instead, Q-TAG

 doi:10.20944/preprints202502.1538.v1

requires the model to generate coherent, natural language descriptions enriched with textual and visual context. Compared to prior controllable captioning methods [5,6,34], our approach is more intuitive and user-friendly, leveraging natural language queries instead of pre-selected object tokens or predefined visual regions.

Given the lack of existing datasets for this task, we develop an automated framework to generate Q-TAG datasets from TextCaps [24] and VizWiz-Captions [11]. Our pipeline extracts relevant scene text from images, constructs initial coarse descriptions, and generates diverse question-answer pairs to simulate real-world user interactions.

To tackle the Q-TAG task, we introduce a novel Question-driven Text-aware Captioning Generator (Q-TAG), which consists of:

- A Spatially-aware Multimodal Encoder, which integrates region-based object features and text-based scene representations while encoding spatial relationships using geometric modeling.
- A Question-driven Feature Selector, which dynamically attends to relevant scene-text-visual features based on user queries, filtering extraneous information.
- A Multimodal Fusion Decoder, which synthesizes the retrieved information to generate fluent and context-aware captions tailored to user queries.

Our extensive experiments on the datasets demonstrate that Q-TAG generates more diverse, informative, and user-centric captions than prior text-aware captioning models.

Key contributions of our work:

- We introduce the Q-TAG task, pioneering question-controlled text-aware image captioning to enhance accessibility for visually impaired individuals.
- We develop a novel model, Q-TAG, which integrates spatial-aware encoding, question-guided feature selection, and multimodal fusion for enhanced caption generation.
- Our model achieves superior performance against state-of-the-art baselines, demonstrating improved contextual awareness, linguistic diversity, and user adaptability.

By leveraging natural language interactions, our approach facilitates intuitive user engagement, enabling more effective comprehension of visually rich environments. Future research directions include incorporating reinforcement learning for adaptive caption refinement and exploring multimodal pretraining for further generalization.

2. Related Work

A plethora of deep learning-based frameworks [2,10,12,14,18,23,28,32] have been developed to address the challenge of general image captioning, aiming to automatically generate detailed and semantically rich textual descriptions of images. Among these, AoANet [14] achieves state-of-the-art performance by leveraging an advanced attention-on-attention mechanism, which significantly enhances context comprehension.

Fishch et al. [10] introduced an innovative training paradigm, wherein question-answering accuracy serves as an auxiliary reward to reinforce information richness in generated captions. However, despite their effectiveness, these methods primarily produce static captions, failing to dynamically adapt to user preferences or highlight scene texts in a manner that aligns with individual user needs.

Text-aware image captioning focuses on interpreting embedded scene text in images while integrating it with surrounding visual objects for coherent descriptions. Notable datasets supporting this research direction include TextCaps [24], derived from Open Images V3, and VizWiz-Captions [11], comprising images captured by visually impaired users. Approximately 63% of VizWiz images contain readable text; however, real-world capturing conditions introduce challenges such as poor lighting, blur, and overexposure, making the dataset highly representative of practical scenarios.

Leveraging these datasets, various models [29,31,33,35] have been proposed to refine text-aware captioning. Wang et al. [29] incorporated spatial reasoning into OCR token processing to better contextualize scene text. Zhu et al. [35] formulated a multi-attention-based strong baseline model. Wang et al. [31] enhanced text selection with confidence embeddings, filtering out irrelevant content,

 doi:10.20944/preprints202502.1538.v1

while Yang et al. [33] designed pre-training tasks specifically tailored for text-centric scene understanding. Although these methods contribute significantly, they do not address user-driven personalized captioning, which remains a crucial gap in this field.

TextVQA models [4,13,15,25] are designed to locate and interpret scene text to answer specific questions. The primary distinction between the TextVQA task and the proposed Q-TAG framework lies in two aspects. First, Q-TAG demands the simultaneous processing of multiple questions, necessitating advanced multi-query comprehension mechanisms. Second, rather than outputting fragmented one-word or short-phrase answers, Q-TAG systems must construct complete, naturally flowing captions that fluently incorporate the relevant scene text and associated visual elements, offering a richer, more holistic textual representation.

Controllable Image Captioning endeavors to generate diverse descriptions by emphasizing different aspects of an image [5,6,34]. Zheng et al. [34] introduced a method that guides captioning via object-focused tokens, allowing the system to tailor descriptions based on predefined object categories. Cornia et al. [6] proposed region-based control signals, enabling caption generation for specific localized areas within an image. Additionally, Chen et al. [5] advanced controllability by employing abstract scene graphs to encode user intentions at a finer granularity.

While these approaches offer precise user-driven control, they rely on explicit, structured signals, such as object selections or bounding-box specifications, which are impractical for visually impaired users. Unlike these rigid control methods, our Q-TAG model introduces a language-driven interactive framework, where users can simply pose natural language queries to obtain personalized captions. This method is more intuitive, removing the prerequisite of visual perception while maintaining high expressiveness, making it an ideal solution for assistive technologies catering to visually impaired individuals.

3. Our Methodology

In this section, we present the Q-TAG framework, a novel architecture designed for the Question-Controlled Text-Aware Captioning (Qc-TextCap) task. Our model consists of three major components: the Spatially-Aware Multimodal Encoder, the Question-Guided Feature Selector, and the Multimodal Fusion Decoder. These modules work in synergy to analyze image regions, process textual questions, and generate rich, personalized captions that integrate scene text information with object relationships.

3.1. Spatially-Aware Multimodal Encoder

To effectively capture the spatial relationships between objects and scene text, our Spatially-Aware Multimodal Encoder fuses object features with scene text features while leveraging geometric properties. Given an input image I, an initial caption C^{ini}, and a set of user queries $\mathcal{Q}$, we apply an object detection model [2] to identify bounding boxes B^{obj} and an OCR engine to extract scene text bounding boxes B^{ocr}. We then extract feature representations V^{obj} and V^{ocr} for these regions.

To model the spatial relationships, we define each bounding box b_i using its geometric attributes:

$$b_i = (c_i^x, c_i^y, w_i, h_i) \tag{1}$$

where c^x and c^y denote the center coordinates, while w and h represent the width and height. The spatial relationships between two regions i and j are encoded as:

$$s_{ij}^g = \log\left(\frac{|c_i^x - c_j^x|}{w_i}, \frac{|c_i^y - c_j^y|}{h_i}, \frac{w_j}{w_i}, \frac{h_j}{h_i}\right) W^g \tag{2}$$

 doi:10.20944/preprints202502.1538.v1

These spatial features are incorporated into the self-attention mechanism, ensuring that the encoder prioritizes regions based on spatial relevance. The attended visual feature is then computed as:

$$\hat{v}_i = \sum_j s_{ij} V_j \tag{3}$$

where s_{ij} is the spatially weighted attention score.

3.2. Question-Guided Feature Selector

To address multiple user queries in a single caption, the Question-Guided Feature Selector dynamically extracts relevant visual-textual information. Instead of treating questions holistically, we perform word-level alignment between question tokens and image features. Given the token embeddings T^{que}, we compute attention scores to identify the most relevant visual features for each word:

$$s_{ij}^q = t_i^{que} W_Q^q (\hat{v}_j W_K^q)^T \tag{4}$$

where W_Q^q, W_K^q are learned projection matrices. The refined feature representation for each question token is then computed as:

$$\hat{t}_i^{que} = t_i^{que} W_T^q + \sum_j s_{ij}^q \hat{v}_j W_V^q \tag{5}$$

This selective feature extraction ensures that each generated caption explicitly answers the posed questions while maintaining linguistic coherence.

3.3. Multimodal Fusion Decoder

The Multimodal Fusion Decoder synthesizes the extracted features into a fluent and meaningful caption. At each decoding step t, the decoder integrates multiple feature sources:

$$Y_t^{dec} = \text{Transformer}([\hat{V}^{obj}, \hat{V}^{ocr}, \hat{T}^{que}, T^{ini}, Y_{t-1}^{dec}]) \tag{6}$$

where $\hat{V}^{obj}, \hat{V}^{ocr}$ are the encoded object and text features, $\hat{T}^{que}$ are the question-guided features, and T^{ini} represents the initial caption.

To generate the next token y_t, we employ a pointer network that selects between vocabulary words and OCR tokens:

$$\hat{y}_t = \text{Softmax}(W^{voc} z_{t-1}^{dec} + W^{ocr} z^{ocr}) \tag{7}$$

where z_{t-1}^{dec} represents the decoder state from the previous step, and z^{ocr} denotes the OCR-based contextual embedding.

3.4. Training Strategy

To enhance the model's ability to incorporate scene text into captions, we experiment with different training objectives. The primary loss function is defined as:

$$\mathcal{L} = -\sum_{t=1}^{l} y_t \log(\hat{y}_t) \tag{8}$$

where y_t is the ground-truth token. We also explore a multi-task learning approach by incorporating a contrastive loss to refine text-image alignment.

3.5. Evaluation and Results

We evaluate Q-TAG on the ControlTextCaps and ControlVizWiz datasets, comparing it against state-of-the-art models. Table 1 presents our performance metrics, showing substantial improvements in BLEU, CIDEr, and SPICE scores.

 doi:10.20944/preprints202502.1538.v1

Table 1. Performance Comparison on Benchmark Datasets

Model	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	Answer Recall
M4C-Captioner	8.98	15.53	32.05	102.41	20.58	-
ControlM4CC	23.81	25.76	48.48	215.45	37.00	46.56
Q-TAG (Ours)	**26.52**	**27.31**	**51.24**	**234.89**	**40.17**	**52.86**

Our results indicate that Q-TAG significantly enhances caption quality, providing more informative, user-adaptive descriptions. Future work will explore reinforcement learning for iterative refinement and multimodal pretraining to improve generalization across diverse image-text scenarios.

4. Experiments

In this section, we conduct extensive evaluations of our proposed Q-TAG framework on the benchmark datasets. We analyze the performance of Q-TAG under various training strategies, compare it against state-of-the-art baselines, and assess its effectiveness in generating personalized, text-aware image captions.

4.1. Experimental Setup

Baseline Methods. To evaluate Q-TAG comprehensively, we compare it against the following models:

1) **M4C-Captioner** [24]: A state-of-the-art text-aware image captioning model that fuses object and scene text features using multi-layer transformers and a pointer network. 2) **ControlM4CC**: An extension of M4C-Captioner adapted for question-controlled captioning, where questions and initial captions are concatenated with visual features. 3) **Q-TAG w/o Spatial Encoder**: A variant of our model that omits the Spatially-Aware Multimodal Encoder to analyze its contribution to overall performance.

Evaluation Metrics. We utilize the standard captioning evaluation metrics: BLEU [20], METEOR [7], ROUGE-L [16], CIDEr [27], and SPICE [1]. Since CIDEr emphasizes rare words, it is particularly suitable for text-aware captioning tasks. Additionally, we introduce Answer Recall (AnsRecall) to measure how accurately the generated captions capture key scene text details in response to given questions.

Implementation Details. We set the maximum sequence lengths for initial captions and concatenated questions to 20 tokens, while generated captions are capped at 30 tokens. Each image is associated with up to 100 object bounding boxes and 50 scene text bounding boxes. During training, the batch size is set to 50, with training steps of 10,000 for ControlVizWiz and 16,000 for ControlTextCaps. Greedy decoding is used for inference unless otherwise specified.

4.2. Evaluation of Q-TAG on Qc-TextCap Task

Comparison of Non-Controllable vs. Controllable Models. Table 1 shows the performance of various models. The question-controlled models outperform M4C-Captioner, demonstrating the benefits of question-guided caption generation. Our Q-TAG model surpasses ControlM4CC in both captioning and question-answering capabilities, validating the effectiveness of the overall architecture.

Ablation studies reveal that removing the Spatially-Aware Multimodal Encoder significantly impacts CIDEr scores (+14.1/+13.8), highlighting its role in fusing spatially aligned visual-textual information. Additionally, Q-TAG achieves higher AnsRecall scores, confirming that spatial reasoning enhances the incorporation of scene text in generated captions.

Impact of Training Strategies. To investigate how different training approaches influence performance, we compare three strategies: 1) Auto: Using only automatically generated initial captions. 2) Pseudo: Using pseudo captions lacking scene text. 3) Rand(auto, pseudo): Randomly selecting one of the two strategies for each training instance.

As seen in Table 2, models trained with the 'rand(auto, pseudo)' strategy outperform those trained with a fixed approach. This suggests that varying training inputs improves generalization, enabling better adaptation to diverse user queries.

 doi:10.20944/preprints202502.1538.v1

Table 2. Comparison of Different Training Strategies

Training Strategy	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	AnsRecall
Auto	25.66	26.52	50.07	231.74	38.44	50.92
Pseudo	14.72	19.89	38.97	143.36	25.46	49.47
Rand(auto, pseudo)	**26.13**	**26.83**	**50.50**	**238.20**	**38.69**	**51.27**

Table 3. Diversity Evaluation Results

Model	Div-1	Div-2	SelfCIDEr
M4C-Captioner	7.44	21.11	62.58
Q-TAG	**14.72**	**38.00**	**78.32**

4.3. Diversity Evaluation

Q-TAG is designed to generate diverse captions tailored to different user queries. To quantify this, we compute Diversity-1 (Div-1), Diversity-2 (Div-2) [3,5], and SelfCIDEr [30]. Table 3 shows that Q-TAG significantly outperforms M4C-Captioner across all diversity metrics, demonstrating its ability to generate more varied and informative captions.

4.4. Qualitative Evaluation and Human Assessment

We conduct human evaluations to compare Q-TAG against M4C-Captioner in terms of scene text accuracy (ST Info) and overall caption quality. Six evaluators rated captions based on how well they incorporated scene text and their fluency.

As seen in Table 4, Q-TAG is preferred over M4C-Captioner in most cases, reinforcing its ability to deliver superior, user-personalized text-aware captions. Our experimental findings establish that Q-TAG significantly improves upon existing text-aware captioning models by incorporating question-controlled guidance. The model enhances caption diversity, ensures more accurate scene text representation, and provides greater user adaptability. Future work will focus on integrating reinforcement learning for iterative caption refinement and extending the framework to handle multi-turn question interactions.

5. Conclusions and Future Directions

In this work, we introduce Question-Guided Text-Aware Image Captioning (Q-TAG), a novel and challenging task aimed at generating personalized, text-aware image captions tailored to the needs of visually impaired users. Unlike conventional text-aware captioning models, Q-TAG employs user-posed questions as control signals to dynamically focus on relevant scene text and visual elements, thereby enhancing both informativeness and interactivity in generated captions.

The Q-TAG task demands that models possess the capability to comprehend user queries, identify relevant scene text regions, and seamlessly integrate extracted information with an initial coarse-grained caption to generate a final, contextually enriched description. To facilitate research in this domain, we construct two benchmark datasets—ControlTextCaps and ControlVizWiz—by augmenting existing text-aware captioning datasets with question-controlled annotations. These datasets will be released publicly to support further advancements in the field.

Table 4. Human Evaluation Results

Model Comparison	ST Info	Overall Quality
Q-TAG $>$ M4C-Captioner	43.48%	51.38%
Q-TAG $\simeq$ M4C-Captioner	42.29%	27.67%
Q-TAG $<$ M4C-Captioner	14.23%	20.95%

 doi:10.20944/preprints202502.1538.v1

To tackle this task, we propose the Q-TAG framework, which incorporates a Spatially-Aware Multimodal Encoder, a Question-Guided Feature Selector, and a Multimodal Fusion Decoder to progressively integrate relevant visual-textual features. Our extensive experimental results on both datasets demonstrate that Q-TAG significantly outperforms baseline methods in terms of both captioning quality and question-answering accuracy. The inclusion of explicit question-guided control signals enables the generation of more informative, diverse, and user-adaptive captions, making it a practical enhancement for assistive technologies.

Moving forward, several promising research directions emerge from our work:

- **Multi-turn Question-Controlled Captioning:** Extending Q-TAG to support interactive, multi-turn dialogues where users can refine or request additional details iteratively.
- **Incorporating Commonsense Knowledge:** Enhancing scene text interpretation by integrating external knowledge sources to infer implicit relationships and contextual meanings.
- **Leveraging Reinforcement Learning:** Employing reinforcement learning to optimize caption generation based on user engagement and feedback, ensuring continuously refined outputs.
- **Multimodal Pretraining Strategies:** Exploring large-scale multimodal pretraining approaches to improve generalization and robustness across diverse real-world datasets.
- **Adaptive Decoding Mechanisms:** Investigating more flexible decoding strategies, such as constrained beam search or neural-symbolic approaches, to ensure higher coherence in generated captions.

By advancing these directions, we envision Q-TAG evolving into a powerful, intelligent assistive system capable of delivering highly customized and meaningful textual descriptions for users with diverse accessibility needs.

References

1. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV (5) (Lecture Notes in Computer Science, Vol. 9909)*. Springer, 382–398.
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 *CVPR*. IEEE Computer Society, 6077–6086.
3. Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander G. Schwing. 2019. Sequential Latent Spaces for Modeling the Intention During Diverse Image Captioning. In *ICCV*. IEEE, 4260–4269.
4. Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Scene Text Visual Question Answering. In *ICCV*. IEEE, 4290–4300.
5. Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs. In *CVPR*. IEEE, 9959–9968.
6. Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *CVPR*. Computer Vision Foundation / IEEE, 8307–8316.
7. Michael J. Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *WMT@ACL*. The Association for Computer Linguistics, 376–380.
8. Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, and David A. Forsyth. 2019. Fast, Diverse and Accurate Image Captioning Guided by Part-Of-Speech. In *CVPR*. Computer Vision Foundation / IEEE, 10695–10704.
9. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.
10. Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan H. Clark, and Regina Barzilay. 2020. CapWAP: Image Captioning with a Purpose. In *EMNLP (1)*. Association for Computational Linguistics, 8755–8768.
11. Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning Images Taken by People Who Are Blind. In *ECCV (17) (Lecture Notes in Computer Science, Vol. 12362)*. Springer, 417–434.
12. Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image Captioning: Transforming Objects into Words. In *NeurIPS*. 11135–11145.

 doi:10.20944/preprints202502.1538.v1

13. Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative Answer Prediction With Pointer-Augmented Multimodal Transformers for TextVQA. In *CVPR*. IEEE, 9989–9999.
14. Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. Attention on Attention for Image Captioning. In *ICCV*. IEEE, 4633–4642.
15. Yash Kant, Dhruv Batra, Peter Anderson, Alexander G. Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. 2020. Spatially Aware Multimodal Transformers for TextVQA. In *ECCV (9) (Lecture Notes in Computer Science, Vol. 12354)*. Springer, 715–732.
16. Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
17. Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV (5) (Lecture Notes in Computer Science, Vol. 8693)*. Springer, 740–755.
18. Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In *CVPR*. IEEE Computer Society, 3242–3250.
19. Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. In *CHI*. ACM, 59.
20. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*. ACL, 311–318.
21. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67.
22. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *EMNLP*. The Association for Computational Linguistics, 2383–2392.
23. Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7008–7024.
24. Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *ECCV (2) (Lecture Notes in Computer Science, Vol. 12347)*. Springer, 742–758.
25. Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. In *CVPR*. Computer Vision Foundation / IEEE, 8317–8326.
26. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.
27. Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*. IEEE Computer Society, 4566–4575.
28. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*. IEEE Computer Society, 3156–3164.
29. Jing Wang, Jinhui Tang, and Jiebo Luo. 2020. Multimodal Attention with Image Text Spatial Relationship for OCR-Based Image Captioning. In *ACM Multimedia*. ACM, 4337–4345.
30. Qingzhong Wang and Antoni B. Chan. 2019. Describing Like Humans: On Diversity in Image Captioning. In *CVPR*. Computer Vision Foundation / IEEE, 4195–4203.
31. Zhaokai Wang, Renda Bao, Qi Wu, and Si Liu. 2021. Confidence-aware Non-repetitive Multimodal Transformers for TextCaps. In *AAAI*. AAAI Press, 2835–2843.
32. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML (JMLR Workshop and Conference Proceedings, Vol. 37)*. JMLR.org, 2048–2057.
33. Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei A. F. Florêncio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2020. TAP: Text-Aware Pre-training for Text-VQA and Text-Caption.. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8751–8761.
34. Yue Zheng, Yali Li, and Shengjin Wang. 2019. Intention Oriented Image Captions With Guiding Objects. In *CVPR*. Computer Vision Foundation / IEEE, 8395–8404.
35. Qi Zhu, Chenyu Gao, Peng Wang, and Qi Wu. 2021. Simple is not Easy: A Simple Strong Baseline for TextVQA and TextCaps. In *AAAI*. AAAI Press, 3608–3615.

 doi:10.20944/preprints202502.1538.v1

36. 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.
37. 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.
38. 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.
39. 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).
40. 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.
41. 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.
42. 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)*.
43. 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.
44. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
45. 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).
46. 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.
47. 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.
48. 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/.
49. 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.
50. 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.
51. 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.
52. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
53. 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.

 doi:10.20944/preprints202502.1538.v1

54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.
69. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
70. 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.
71. 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.
72. 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.
73. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
74. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.

 doi:10.20944/preprints202502.1538.v1

75. 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.
76. 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.
77. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.
83. 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.
84. 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.
85. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
86. 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.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. 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.
94. 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.

95. 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.
96. 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.
97. 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.
98. 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.
99. 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.
100. 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.