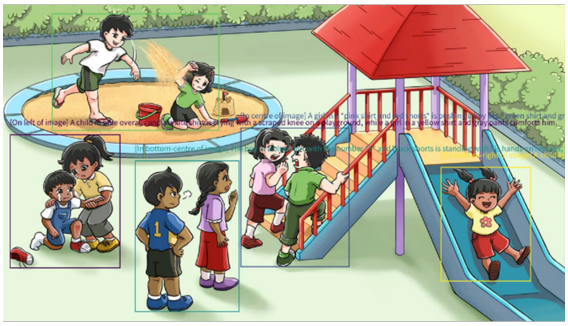
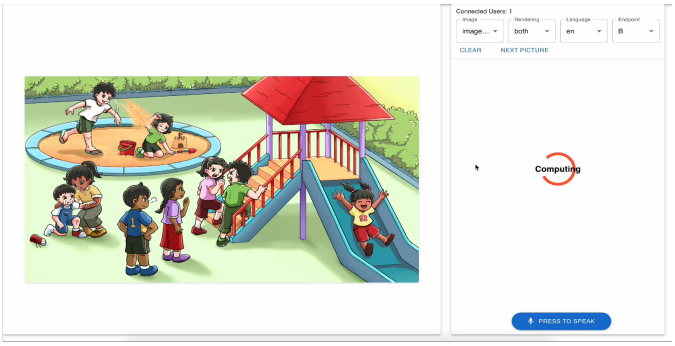
Contextualized Visual Storytelling for Conversational Chatbot in Education

Hui Li Tan, Ying Gu, Liyuan Li, Mei Chee Leong, Nancy F. Chen
{hltan,guy,lyli,leong_mei_chee,nfychen}@i2r.a-star.edu.sg
Institute for Research (I^{2}R), Agency for Science, Technology and Research (A*STAR)



Contextualized dense image captioning for visual storytelling, with **fine-grained understanding** (detailed description of the objects, characters, and activities), incorporating **cultural contextualization** (using localized references to learners' environment); **learner contextualization** (ensuring linguistic simplicity and age-appropriate phrasing); **pedagogical contextualization** (anchoring to curriculum-defined keywords).



Picture-guided conversational chatbot.

Figure 1: Contextualized dense image captioning for conversational chatbot in education.

Abstract

Interactive visual storytelling through conversational agents offers a means to enhance early childhood language learning. We developed a picture-guided conversational chatbot, driven by dense image captioning, for early childhood mother tongue language learning. However, state-of-the-art image captioning systems fall short in meeting the educational needs of young learners. They often lack cultural contextualization, use vocabulary that exceeds children's developmental level, and fail to align with curriculum-relevant learning goals. We investigated a contextualized dense image captioning framework, which augments dense image captioning with cultural and curriculum-aligned keyword retrieval through a Retrieval-Augmented Generation (RAG) module. This enables the generation of culturally appropriate, age-level suitable, and educationally anchored captions that enhance learner engagement and pedagogical relevance. We demonstrate that our approach outperforms existing captioning models in terms of linguistic appropriateness, and curriculum and cultural alignment. The contextualized dense image captioning framework supports the development of culturally grounded, education-oriented conversational agents for young learners.

This work is licensed under a Creative Commons Attribution 4.0 International License.
ICMI Companion '25, Canberra, ACT, Australia
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2076-5/2025/10
https://doi.org/10.1145/3747327.3764895

CCS Concepts

• **Human-centered computing. Human Computer Interaction (HCI). Interactive systems and tools.**; • **Social and professional topics. Professional topics. Computing education.**;

Keywords

dense image captioning, vision-language models, multi-cultural, human computer interaction, AI for education

ACM Reference Format:
Hui Li Tan, Ying Gu, Liyuan Li, Mei Chee Leong, Nancy F. Chen. 2025. Contextualized Visual Storytelling for Conversational Chatbot in Education. In *Companion Proceedings of the 27th International Conference on Multimodal Interaction (ICMI Companion '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3747327.3764895

1 Introduction

Mother tongue language education plays a critical role in early childhood cognitive development and cultural identity formation. In multi-lingual regions such as Southeast Asia, children often face challenges in acquiring mother tongue languages like Chinese due to limited immersive environments and misalignment between educational content and children's everyday experiences [10]. Visual storytelling, delivered through picture-guided conversational chatbot, offers an interactive solution to bridge the education gap. We developed a picture-guided conversational chatbot to support interactive mother tongue language learning for young learners. The

natural language descriptions associated with visual cues provide concrete semantic and vision concepts for language learning.

We propose a contextualized dense image captioning method to generate the visual storyline to guide the conversation flow. To support rich and engaging conversation, the dense image captioning needs to comprehensively identify the key events in the image, as well as generate detailed captions for each key event. Referring to Figure 1, the captions should include fine-grained description of the objects, characters, and activities while maintaining logical coherence with the overarching narrative. These ensure the generated content supports both local scene understanding and global narrative alignment as basis for interactive visual storytelling.

Additionally, to support the instructional goals of early childhood language education, the dense image captioning needs to be contextualized across multiple dimensions. (1) **Cultural context**, by using localized references that align with the learner's environment. Supported by *Situated Learning* [7], learning can be enhanced since learners relate better to materials embedded in their real-world context. For instance, the visual references and vocabulary should match the younger learners' cultural experience (e.g. local food, daily activities, familiar environments). (2) **Learner context**, by ensuring linguistic simplicity and age-appropriate phrasing. Supported by *Zone of Proximal Development* [15], the language should appropriately challenge but not overwhelm the learners. For instance, the word choice and sentence complexity should match the cognitive and linguistic level of the young learners. (3) **Pedagogical context**, by anchoring content generation to textbook vocabulary and grammar structures. Supported by *Intentional Vocabulary Instruction* [5], direct focus on high-utility words can enhance vocabulary acquisition. For instance, the target vocabulary should anchor to curriculum-defined wordlists (e.g. textbook). Together, these contextualization foster engagement, support comprehension and intentional vocabulary acquisition, making visual storytelling more effective and engaging for early childhood language education.

Dense image captioning generates descriptive sentences from an image, with each sentence corresponding to an image region spatially localized through a bounding box. Beyond traditional dense image captioning models, e.g. GRiT [16] and ControlCap [19], various multi-modal large-language models (MLLMs) can support dense image captioning without image region localization, e.g. InternVL [20], DeepSeek [17], Claude [2], ChatGPT[1]. Nonetheless, state-of-the-art image captioning methods are meant for general purpose, and typically generate generic descriptive sentences that lack the contextualization required for education needs. The image captions omit local cultural references, use advanced language beyond the young learners' proficiency, and fail to support the explicit acquisition of curriculum-defined keywords.

To address the limitations, we investigated a contextualized dense image captioning framework tailored for picture-guided conversational chatbot to support early childhood mother tongue language learning. The system aims to generate culturally-relevant, learner-appropriate, and curriculum-aligned dense image captions, by introducing Retrieval-Augmented Generation (RAG) for cultural and curriculum keyword anchoring. This research targets Chinese language spoken in Singapore.

2 Related Work

Traditional dense image captioning models, such as GRiT [16] and ControlCap [19], have been developed to generate multiple region-level captions from a single image. These models go beyond single image captioning models by providing fine-grained descriptions of different regions within a scene. Such detailed and localized captions can help scaffold vocabulary acquisition and support visual storytelling narratives. With the advancement of MLLMs, various open-source models, e.g. InternVL [20] and DeepSeek [17], and proprietary models, e.g. Claude [2] and ChatGPT [1], can generate dense image captions without localization bounding boxes.

There is limited work but increasingly interest in cultural diversity / inclusivity of MLLMs, for visual question answering [8, 9] visual grounding [6], visual captioning [18] etc. Cultural misalignment is particularly true for Southeast Asia, where under-resourced representations differ from norms embedded in most pretrained models, e.g. Singaporean/Malaysian Chinese differs from Mainland Chinese. Specific to image captioning, the authors in [18] generate cultural-aware image captions by introducing a visual question answering mechanism (VQA) to extract specific cultural elements, e.g. architecture, clothing, dance & music, region, food & drink, and then integrating the supplementary answers with the default captions for cultural-aware image captions. Our method provides a RAG-based solution on state-of-the-art VLM and MLLM.

Pedagogical alignment has been actively explored in the context of large-language models (LLMs) [3, 12], focusing on scaffolding problem-solving, tailoring feedback, and aligning generation with curriculum goals. However, such pedagogical alignment remains largely unaddressed in MLLMs, where visual captioning and storytelling are typically optimized for general comprehension than instructional intent.

3 Methodology

To leverage the common-sense knowledge and generalization capabilities of VLM and MLLM, while enhancing their effectiveness to produce contextualized storytelling captions on images for mother tongue language learning, we propose a RAG-based framework. The framework employs context-aware cultural and curriculum keyword anchoring to guide the generation of contextualized dense captions. The framework architecture is illustrated in Figure 2. It consists of two main modules, i.e., Context-aware Retrieval and Context-aware Captioning modules, as described in the following.

3.1 Vocabulary Database Creation

The framework begins with constructing a cultural and curriculum-aligned vocabulary database, which serves as the basis for the context-aware retrieval. To effectively teach learners storytelling expressions and skills grounded in visual content, the vocabulary used should represent clear semantic concepts that are directly associated with identifiable visual instances in the images. This promotes strong visual-linguistic alignment while minimizing ambiguity. Hence, we curated a set of visually groundable vocabulary terms, i.e., vocabulary terms which can be visually recognized from the images. We collected the groundable-person, groundable-object, and groundable-activity vocabulary terms. Examples of

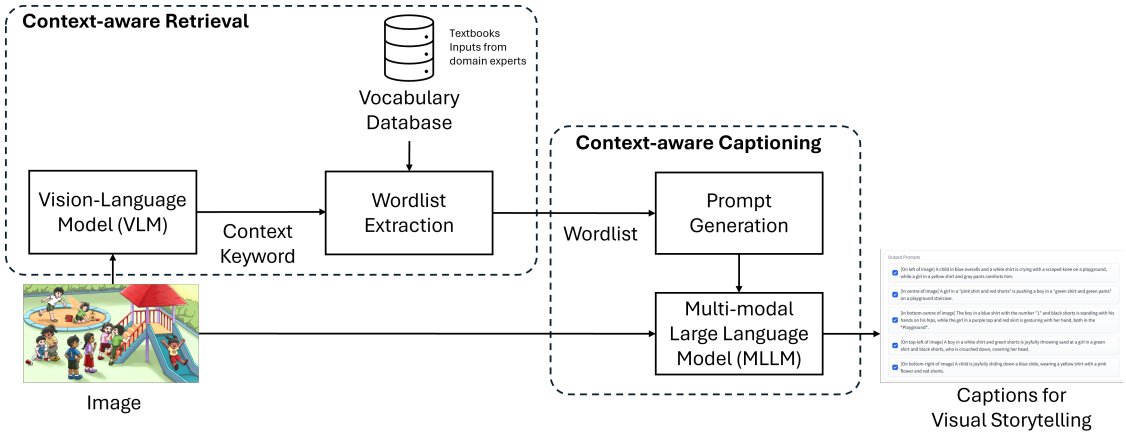


Figure 2: Contextualized dense image captioning by Retrieval-Augmented Generation (RAG) for cultural and curriculum keyword anchoring.

groundable-person, groundable-object, and groundable-activity vocabulary terms are 'sister', 'slide', and 'running' respectively. We exclude inferred vocabulary term which may not be visually recognized from the image, such as 'kind-hearted'.

The vocabulary terms are collected from all chapters of the primary one and two mother tongue textbooks. Additional vocabulary terms are added based on the education requirements and adaption to novel teaching materials. There are a total of 440 vocabulary terms, including 27 groundable-person, 295 groundable-object, and 118 groundable-activity vocabulary terms respectively, for the Chinese language. The vocabulary terms covers a wide range of cultural references (e.g. food or landmarks) and curriculum-defined words (e.g. household items, sports).

The vocabulary terms are organized into contexts, e.g. 'school', 'home', 'playground', 'shop', 'zoo' clusters, which are among the top topics in primary one and two mother tongue textbooks used in Singapore. The teacher can adjust this vocabulary base by adding new words and removing less popular words to produce new teaching materials for mother tongue language learning.

3.2 Context-aware Retrieval

A given image is classified into one of the contexts defined in the database. We used a state-of-the-art vision language model (VLM), EvaCLIP [13], for the zero-shot scene classification, which can be easily adapted to novel context to fit curriculum themes or learning goals, due to its generalization as per-trained on vast images.

Based on the identified context keyword, the respective vocabulary wordlist with cultural references and curriculum-defined words are retrieved. The retrieved vocabulary list is fed to the context-aware captioning module to guide caption generation.

3.3 Context-aware Captioning

Prompt-guided caption generation is used to produce the contextualized dense image captions for visual storytelling. To support the process, the retrieved vocabulary wordlist is incorporated for prompt refinement. The prompt is designed to elicit fine-grained description of each person, including the distinctive character traits and the corresponding activity. Additionally, the retrieved vocabulary wordlist serves as guiding words which should only be used when they are visible in the image. This constraint is needed to ensure semantic accuracy based on the visual content, mitigating hallucination where the MLLM generates imagined or inaccurate content solely based on the provided keywords. The refined prompt is then used to guide the MLLM in generating contextualized dense image captions. An example of the formatted prompt is:
'Describe the people in clockwise or anti-clockwise order, number each description and include the spatial position e.g. [top-right]. In the two-sentence description, include what they are wearing and doing. Use these helping words only if it is in the image. str(wordlist). No colon. Please reply in Chinese.'

4 Experimental Results

To evaluate the effectiveness of dense captioning on images for visual storytelling in mother tongue language learning, we conduct experiments following zero-shot setting on a set of novel pictorial images designed for visual storytelling exercises, where each image contain 5-10 distinctive activities. In this section, we present the details of task description, test data set collection, employed metrics, performance comparison, and visual examination on a representative example.

We focus on the teaching of Chinese language, one of the mother tongue languages in Singapore, for primary school students. In our experiments, we curated a set of 24 novel images[1,2] for storytelling skills for young learners. For each image, we collaborated with education professionals to develop age-appropriate captions containing curriculum-relevant keywords. These curated captions serve as ground-truth captions, for evaluating both linguistic appropriateness and instructional intent of the generated dense captions for early childhood Chinese learning.

The evaluation of dense image captioning remains an open-area of research, especially in visual storytelling. Typical automated evaluation metrics, e.g., BLEU [11], METEOR [4], CIDEr [14], which

[1] A*STAR Dataset for AI in Education Grand Challenge.
[2] AISG Dataset for AI in Education Grand Challenge.

Method	en2cn		cn	
	R	LA	R	LA
Generalized	0.299	0.925	0.284	0.919
Prompt-based anchoring	0.268	0.923	0.300	0.933
RAG-based anchoring (ours)	0.417	0.938	0.423	0.941

Table 1: Comparison of performance across the three methods. 'en' refers to English, 'cn' refers Chinese, 'R' refers to Recall, and 'LA' refers Linguistic Appropriateness.

relies on ground-truth captions, often fall short in capturing the full semantics and contexts of the image captions. For our task, we evaluated the dense image captions for curriculum fit and cultural resonance using Recall (R) against the ground-truth captions and Linguistic Appropriateness (LA) of the generated captions.

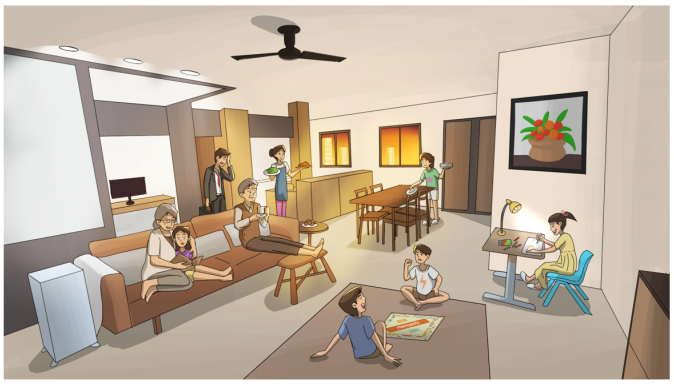
Recall (R): We employ ChatGPT as an evaluation tool to assess the quality of the generated dense image captions. Specifically, we use ChatGPT to extract relevant keywords from the generated captions and ground truth references. We then analyze if the generated captions contain the extracted keywords in comparison to the ground truth keywords to compute the recall. This approach enables quantitative measure of how well the generated captions align with the cultural context and curriculum defined learning goals.

Linguistic Appropriateness (LA): We prompt ChatGPT with a simulated perspective of a child aged 6–10, asking whether the generated caption is understandable and age-appropriate. This approach enables quantitative measure of how well the generated captions aligns with the target learner comprehension level.

The proposed RAG-based anchoring approach is compared against default off-the-shelf MLLM and prompt-based anchoring approach for the MLLM. The compared methods are Generalized, Prompt-based anchoring, RAG-based anchoring (ours). For prompt-based anchoring, we included additional prompt to instruct the MLLM that it is a Singapore primary school teacher, i.e. *'You are a Singapore primary school teacher.'* prepended to the first two lines of the formatted prompt given above. In this study, we design and evaluate our system using ChatGPT (GPT-4o), due to its advanced language generation capabilities and consistent performance. While we acknowledge the existence of other vision-language models (e.g., InternVL, Claude), our focus is on investigating the impact of contextual adaptation rather than benchmarking across model architectures. We experimented with using the MLLM in English and then translating the response (en2cn), as well as using the MLLM in Chinese directly (cn).

The performance comparisons are shown in Table 1. It is worth noting that the results of our method is obtained on the accuracy of context identification (zero-shot scene classification) of 92%. From the table, one can observe that our method improves recall (R) over 140% (0.417/0.299) w.r.t. two compared methods, and achieves best LA performance on both en2cn and cn tasks, which show the clear advantage of the effectiveness to produce contextualized dense captions on images for learning and exercise of visual storytelling skills in mother tongue language learning.

An example is shown in Figure 3. We observed that the RAG-based anchoring method outperforms the other two methods. The



Generalized (en2cn): 一位穿着黄色连衣裙的小女孩正坐在桌子旁，认真地低头*画画*，桌面上摆满了*彩笔*和*纸张*。
Generalized (cn): 穿着黄色连衣裙的小女孩坐在蓝色椅子上，正在灯光下认真地*绘画*。她面前摆放着多支*彩色笔*和*纸张*，透露出专注和创作的兴趣。
Prompt-based anchoring (en2cn): 一位穿着黄色连衣裙的小女孩坐在蓝色椅子上，正在桌前*画画*。她的面前有一盏亮着的台灯和一些*画笔*，看起来很专注。
Prompt-based anchoring (cn): 一位穿黄色连衣裙的女孩坐在蓝色的椅子上，认真地在书桌上*画画*。桌上有*画笔*和*纸张*，她看起来很专注于自己的创作。
RAG-based anchoring (en2cn):* 姐姐*穿着黄色连衣裙，坐在书桌旁*画画*，桌上有*彩色铅笔*和*活动纸*。
RAG-based anchoring (cn): 穿黄色连衣裙的*姐姐*坐在小椅子上，使用桌上的*彩色笔*在*活动纸*上认真地*画画*，桌上还有台灯。

Figure 3: Comparison results of six variants on example image[2]. We highlight curriculum-aligned keywords using **.

default MLLM uses language that is too difficult for the user. Prompt-based anchoring helps in simplifying the language to a limited extent. The RAG-based anchoring method not only helps in simplifying the language, but also uses the intended vocabulary. The default MLLM uses 女孩 (girl), while the RAG-based MLLM uses more relevant roles like 姐姐 (sister) in home setting. Additionally, the RAG-based anchoring method anchors the visual storytelling to curriculum keywords, e.g. 画画 (simple vocabulary for drawing) instead of 绘画 (advanced vocabulary for drawing). In other examples we observed the default MLLM typically uses 男士 (man) or 女士 (woman), while the RAG-based MLLM uses the appropriate roles like 爸爸 (father) and 妈妈 (mother) in home setting. Additionally, the default may be biased towards less local terms such as 西红柿 (tomato in Mainland Chinese) instead of 番茄 (tomato in Singapore/Malaysian Chinese).

5 Conclusion

We presented a contextualized dense image captioning framework to support interactive visual storytelling for early childhood language learning. By grounding dense image captioning in learning science principles, we aim to enhance learner engagement and support intentional vocabulary acquisition through cultural, learner and pedagogical contextualization. Through RAG-based cultural and curriculum-aligned keyword anchoring, the generated captions provide fine-grained visual understanding while maintaining culturally-relevant, age-appropriate, and curriculum-aligned. The approach bridges the gap between general-purpose models and the specific needs of early learners of mother tongue. Future work will explore extension to support multi-lingual and cross-cultural generalization (e.g., low-resource languages like Malay and Tamil), as well as more comprehensive evaluation.

6 Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-GC-2022-005), the A*STAR under its Japan-Singapore Joint Call: Japan Science and Technology Agency (JST) and A*STAR 2024 (R24I6IR136), and by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme (DesCartes).

References

[1] [n. d.]. ChatGPT. https://openai.com/index/chatgpt/. Accessed: 2025-07-02.
[2] [n. d.]. Claude. https://claude.ai/new. Accessed: 2025-07-02.
[3] Mohsen Balavar, Wenli Yang, David Herbert, and Soonja Yeom. 2025. Enhancing tutoring systems by leveraging tailored promptings and domain knowledge with Large Language Models. arXiv:2505.02849 [cs.CY] https://arxiv.org/abs/2505.02849
[4] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (Eds.). Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909/
[5] Isabel L Beck, Margaret G McKeown, and Linda Kucan. 2013. Bringing words to life: Robust vocabulary instruction. Guilford Press.
[6] Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. From Local Concepts to Universals: Evaluating the Multicultural Understanding of Vision-Language Models. arXiv:2407.00263 [cs.CL] https://arxiv.org/abs/2407.00263
[7] Jean Lave and Etienne Wenger. 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press.
[8] Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. CultureVLM: Characterizing and Improving Cultural Understanding of Vision-Language Models for over 100 Countries. arXiv:2501.01282 [cs.AI] https://arxiv.org/abs/2501.01282
[9] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking Vision Language Models for Cultural Understanding. arXiv:2407.10920 [cs.CV] https://arxiv.org/abs/2407.10920
[10] Teresa Wai See Ong and Quevada Hannah Rocafort. 2023. The Bilingual Dreams: Reasons to Learn Mother Tongues and Practical Tips. https://singteach.nie.edu.sg/2023/04/28/the-bilingual-dreams-reasons-to-learn-mother-tongues-and-practical-tips/
[11] 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 of the Association for Computational Linguistics, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135
[12] Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024. Pedagogical Alignment of Large Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2024, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 13641–13650. doi:10.18653/v1/2024.findings-emnlp.797
[13] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. arXiv:2303.15389 [cs.CV] https://arxiv.org/abs/2303.15389
[14] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. arXiv:1411.5726 [cs.CV] https://arxiv.org/abs/1411.5726
[15] L.S. Vygotsky. 1978. Mind in society: The development of higher psychological processes. Harvard University Press, Cambridge, MA.
[16] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2022. GRiT: A Generative Region-to-text Transformer for Object Understanding. arXiv:2212.00280 [cs.CV] https://arxiv.org/abs/2212.00280
[17] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. arXiv:2412.10302 [cs.CV] https://arxiv.org/abs/2412.10302
[18] Youngsik Yun and Jihie Kim. 2024. CIC: A Framework for Culturally-Aware Image Captioning. In Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 1625–1633. doi:10.24963/ijcai.2024/180
[19] Yuzhong Zhao, Yue Liu, Zonghao Guo, Weijia Wu, Chen Gong, Fang Wan, and Qixiang Ye. 2024. ControlCap: Controllable Region-level Captioning. arXiv:2401.17910 [cs.CV] https://arxiv.org/abs/2401.17910
[20] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479 [cs.CV] https://arxiv.org/abs/2504.10479