Tell as You Want: Customizing Image Narrative with Knowledge and Thoughts

Ziwei Yao[1,2], **Qian Wang**[1,2], **Ruiping Wang**[1,2*], **Xilin Chen**[1,2]

[1]Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
{ziwei.yao, qian.wang}@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn

Abstract

With the advancement of vision-language models, image captioning has made significant progress, leading to the generation of more accurate and detailed descriptions. Current image captioning primarily focuses on describing the apparent visual characteristics, which are easily observed by most humans, but less helpful in real-world scenarios. When users seek a deeper understanding of visual content, they may be concerned with fine-grained categories, function properties, and other background knowledge, rather than merely appearances. Additionally, as users' interests vary, there is a growing demand for customizable content generation. To address these challenges, we propose the task of image narrative generation, which aims to produce knowledge-rich natural language responses for input images, customized to the user preference. Furthermore, we propose T^4, an image narrative generation model progressing through cascade steps: **T**ailor, re**T**rieve, **T**hink, and **T**ell. Specifically, it takes the image and various types of prompts as input, and first refines or predicts potentially interesting queries that are tailored to the user expertise level. Subsequently, the model enriches contextual knowledge through retrieval-augmentation and employs chain-of-thoughts to decompose the generation process step by step, thereby telling an accurate and logically coherent image narrative. In addition, we construct the ImgNarr-23K dataset to support task training and evaluation. Experimental results demonstrate that the proposed approach generates image narratives that better satisfy user requirements, and achieves state-of-the-art performance in knowledge-based VQA tasks without additional finetuning. T^4 presents a promising solution for customized content generation in specialized domains.

1 Introduction

Image caption generation has made significant progress, owing to advances in multimodal learning (Vinyals et al. 2015; Anderson et al. 2018; Li et al. 2023). The generated captions have become increasingly accurate and detailed, leading to the development of diverse tasks such as dense captioning (Yang et al. 2017), detailed captioning (Urbanek et al. 2024), grounded captioning (Cornia, Baraldi, and Cucchiara 2019), controllable captioning (Wang et al. 2023), etc. Image captions are expected to be utilized in scenarios such as children education and exhibit introductions, to provide explanations of content and knowledge that may be unfamiliar to the users. However, most captioning models primarily concentrate on the visible visual elements that are easily recognizable to humans, such as objects, attributes and relationships, while neglecting deeper knowledge that users may be concerned about, which to some extent limits its application in real-world scenarios. In addition, as assistants for interpreting images, the outputs of current captioning models are typically fixed, lacking diversity and customization to accommodate various user preferences.

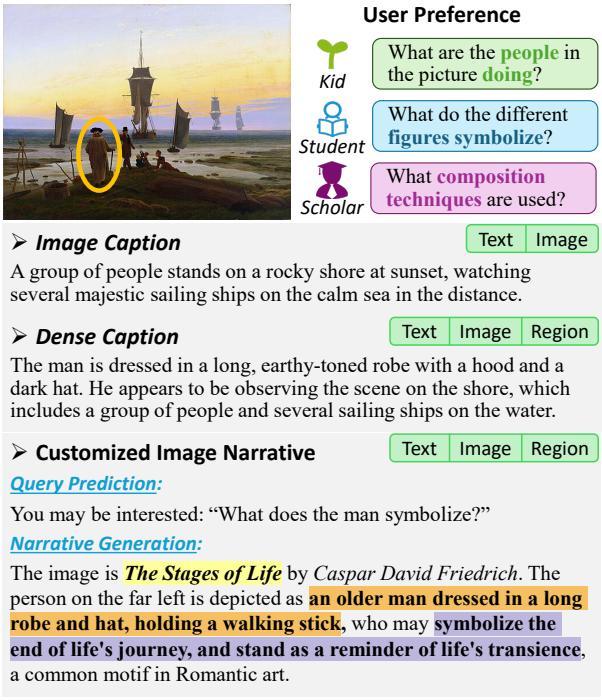

Figure 1: When attempting to understand the image content, users require not only the basic visual appearance but also personalized contextual knowledge. We propose the customized image narrative generation, which provides detailed and knowledge-rich natural language responses tailored to the user-specific interests.

*Corresponding author.

Specifically, as shown in Fig. 1, when aiming to gain a deep understanding of visual content, users often expect specific and in-depth knowledge—such as taxonomy, function, and historical development—rather than simply appearance description. Furthermore, users seek personalized responses tailored to their individual preferences. For instance, when presented with the same item, children generally ask for basic popular science information, while scholars prefer more professional explanations. This requires the model not only to possess specialized knowledge, but also to respond flexibly and controllably to customized needs. Although recent multi-modal large language models (Dai et al. 2023; Liu et al. 2023; Chen et al. 2024) enhance the flexibility of generative tasks through dialogue and demonstrate a fundamental understanding of general world knowledge and common sense, they still struggle with specialized domains that have limited data. In these contexts, MLLMs may misclassify categories, generate inaccurate or excessively general responses, and even produce hallucinated information. Such limitations significantly undermine users' trust and the practical value of the model.

Therefore, we introduce the task of **customized image narrative generation**, which generates a detailed natural language response regarding user-specific interests about the input image. This response uses fine-grained object categories as anchors, includes descriptions of image content relevant to the user's query, and provides a reply with a simple reasoning process. Compared with image captioning, which only provides appearance descriptions, image narrative offers more knowledge that users may be interested in. Moreover, the reasoning process can enhance interpretability and strengthen the user's trust in the generated content. To accomplish this task, we propose T^4, an image narrative generation model that progresses through cascade steps: **T**ailor, re**T**rieve, **T**hink, and **T**ell. T^4 takes images and flexible multimodal prompts as input, and refines the specific query according to user preference. Prompts may include textual instruction as well as region-specific inputs within the image, such as points or bounding boxes, allowing users to customize their queries. Then it employs a two-stage retrieval-augmented approach to fetch relevant documents and passages from the knowledge base, as the factual references for subsequent generation. Furthermore, by leveraging chain-of-thought reasoning, T^4 decomposes the query into structured reasoning steps and finally generates the narrative, which enhances its interpretability and logical rationality.

To support the model training and evaluation, we construct the **ImgNarr-23K** dataset, collecting image narrative annotations for images from InfoSeek (Chen et al. 2023) and Encyclopedic-VQA (Mensink et al. 2023) datasets. Experimental results demonstrate that T^4 generates more user-adaptive, accurate, and detailed image narratives than current open-source MLLMs. It also achieves SoTA performance in knowledge-based VQA benchmarks without additional finetuning, illustrating the generalization and robustness of our framework.

2 Related Works

2.1 Knowledge-Related Vision-Language Tasks

As vision-language research advances towards deeper semantic understanding, the integration of knowledge is attracting increasing attention, as reflected in tasks such as visual commonsense reasoning, visual question answering, and image captioning. Visual commonsense reasoning requires models to reason and rationalize about commonsense knowledge, such as complex human activities and social interactions, with the help of explicit knowledge base (Park et al. 2020) or implicit model knowledge (Zellers et al. 2019). KB-VQA aims to answer questions beyond what is visually observable using external knowledge (Marino et al. 2019; Schwenk et al. 2022; Shah et al. 2019), while recent encyclopedia VQA models (Mensink et al. 2023; Chen et al. 2023) focus more on entity recognition and specialized-domain knowledge learning of fine-grained objects. In image captioning, knowledge helps make descriptions more informative and specific. TextKG (Gu et al. 2023), EVCAP (Li et al. 2024b) and MeaCap (Zeng et al. 2024) enhance fine-grained category recognition by integrating external knowledge documents. This also has potential applications in specialized fields, such as medical image analysis (Yang et al. 2022; Bu et al. 2024), remote sensing image captioning (Li et al. 2024c), and autonomous driving (Hussien et al. 2025). The proposed image narrative generation not only provides deeper contextual knowledge in natural language descriptions but also offers customization capabilities tailored to user preferences.

2.2 Retrieval-Augmented Generation

Retrieval-augmented Generation (RAG) is a widely adopted approach for integrating external knowledge into models to enhance reliability and mitigate hallucinations. It has been applied to a range of tasks, including natural language processing (Karpukhin et al. 2020; Shuster et al. 2021) and multimodal applications. In multimodal tasks, RAG can be used to improve caption quality. EXTRA (Ramos, Elliott, and Martins 2023) retrieves relevant captions and encodes them together with the image. RAG is also used in knowledge-enhanced VQA to improve question answering accuracy. Wiki-LLAVA (Caffagni et al. 2024) retrieves relevant wiki passages and feeds them into the multimodal large language model. EchoSight (Yan and Xie 2024) retrieves relevant wiki passages and uses a Q-Former-based re-ranking module to sort them. ReflectiVA (Cocchi et al. 2025) utilizes reflective tokens to decide if external knowledge is needed and to predict the relevance of retrieved information. We employ retrieval-augmented methods to incorporate domain-specific knowledge into image narratives, thereby increasing the richness of knowledge and enhancing the accuracy of responses.

2.3 Chain-of-Thought Reasoning

Chain-of-Thought (CoT) reasoning improves large language models by encouraging them to solve problems step by step (Wei et al. 2022). Early CoT uses simple prompts to guide the reasoning process (Kojima et al. 2022; Shum,

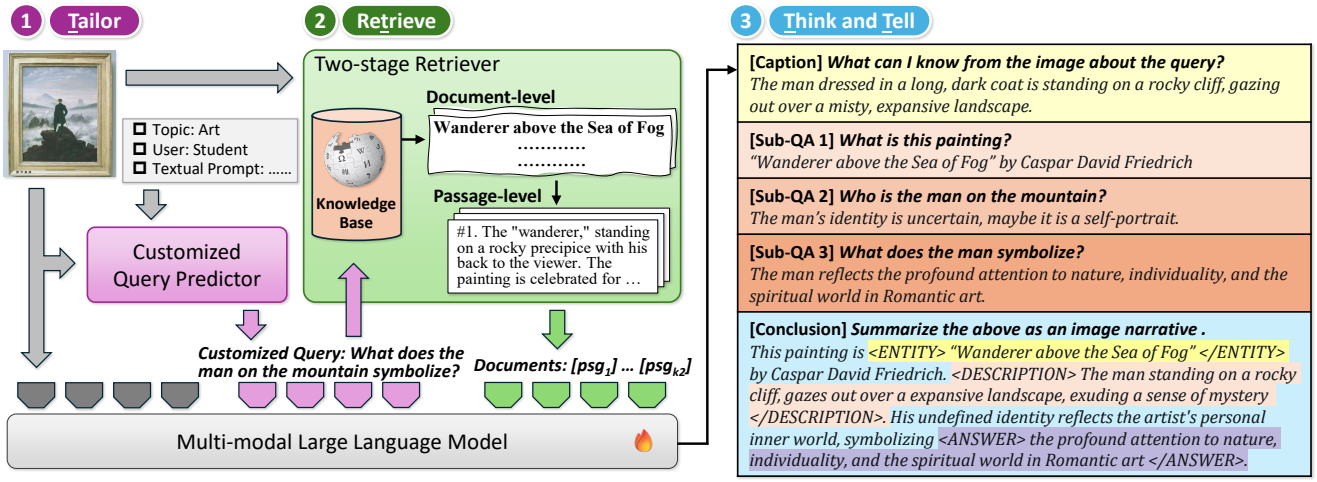

Figure 2: The overall framework of the proposed T^4. Given the image and user prompts, T^4 first predicts candidate queries according to user preference. Then, the model employs a two-stage retrieval-augmented approach to obtain relevant knowledge passages as references. Finally, we design a chain-of-thought-based reasoning process, enabling the model to generate image descriptions, answer decomposed sub-questions separately, and ultimately produce a fluent and informative narrative.

Diao, and Zhang 2023). Tree-of-Thought (ToT) extends CoT into a tree-based framework that enables optimized exploration of reasoning paths (Yao et al. 2023), while Graph-of-Thought (GoT) extends CoT to a graph-based paradigm that supports reflection and refinement through cyclic reasoning (Besta et al. 2024). CoT has also been extended to multimodal large language models. Zhang et al. (2023) and Mitra et al. (2024) divide the CoT process into first generating a rationale and then predicting the answer, enhancing the interpretability of results. Guo et al. (2025) and Xu et al. (2025) make the CoT longer and more complex, including problem decomposition, multi-step reasoning, and providing the final response, better mimicking human-like slow thinking. Applying CoT to the image description task, our T^4 proposes a three-stage reasoning process, which effectively constrains the thinking path and enhances the transparency and credibility of the generation.

3 Method

In this section, we introduce the task of customized image narrative generation. Then, we propose the T^4 framework, which generates narratives through a cascade of steps: **T**ailor, re**T**rieve, **T**hink and **T**ell, as shown in Fig. 2. T^4 first predicts the customized query that the user may be interested in. Then, the knowledge retrieval module extracts relevant document segments from the knowledge base as references. Finally, the query is decomposed and answered step by step through the chain-of-thought reasoning, leading to a structured image narrative output. In addition, we introduce the construction of the ImgNarr-23K dataset for both training and evaluation, as well as the training approach we employ.

3.1 Customized Image Narrative Generation

The customized image narrative task aims to generate a natural language response that addresses the user-specific concerns regarding a given input image. Given an image and the user prompts, the model should first predict candidate queries according to the user's preference, and then produce a natural language explanation in response to the query. As described in the Introduction section, to enable more natural interactions and enhance readability in practical scenarios, a narrative $\mathcal{N}$ contains the detailed description of the image content related to the query N_d, identification of relevant fine-grained entity categories N_e, and the reasoning process that leads to the final answer N_a. We design three pairs of special tags to mark the corresponding content `<DESCRIPTION></DESCRIPTION>`, `<ENTITY></ENTITY>`, and `<ANSWER></ANSWER>`. The latter two tags are used to mark specific entity names and answers, rather than the entire sentence, to achieve more precise localization.

A RAG-based MLLM framework to generate customized narratives can be formulated as:

$$\mathcal{N} = MLLM(I, Q, K), \tag{1}$$

where I denotes the input image, Q denotes the predicted query, and K denotes the related background knowledge to ensure correctness.

3.2 Tailor - Customized Query Prediction

Given an input image I, configuration information C (which includes the image topic and user type), and optional user prompts T and R, T^4 first predicts the query Q that the user is likely interested in, that is,

$$Q = QuePred(I, C, T, R), \tag{2}$$

where T denotes textual instruction and R denotes visual prompts, such as bounding boxes and points, which are used to specify the region of interest.

As shown in Fig. 3, we utilize an MLLM to perform detailed analysis of the input above, converting the image content, and user-interested regions into textual descriptions,

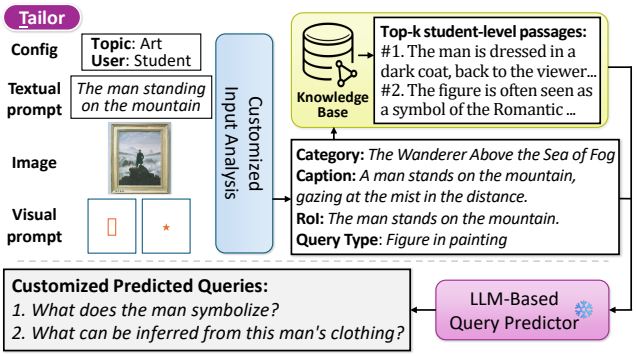

Figure 3: The customized query prediction module of T^4. Given an image and various types of prompts, the module first performs content analysis and generates detailed intermediate results. Then, it retrieves relevant knowledge entries that match the user's level of expertise from a knowledge base, and finally predicts customized queries.

and roughly predict query type. To reduce hallucinations when facing unfamiliar specialized domains, we employ a knowledge base as the foundational resource for query prediction. Inspired by Britannica Kids (Encyclopædia Britannica 2025), users are categorized into three levels based on their cognitive development: `Kids` (up to 5th grade), `Students` (grades 6 to 8), and `Scholars` (grades 9 and up). Knowledge documents aimed at specific groups, such as children's popular science books, can be directly utilized. For general knowledge resources like Wikipedia, we use GPT-4o (Hurst et al. 2024) to rewrite the content into knowledge documents tailored to the user's identity. Subsequently, the documents are segmented into passages, forming a knowledge base. Then, based on the analyzed image content and region of interest, the relevant passages are retrieved from the knowledge base. Finally, the passages are transformed into several predicted questions by the LLM. The generated queries are used as targets for the following image narrative generation. More details are available in the supplementary material.

3.3 Retrieve - Knowledge Retrieval Augmentation

We employ a two-stage retrieval approach that first conducts coarse document retrieval and then performs fine-grained passage retrieval, based on the input image I and query Q.

In the first stage of coarse document retrieval, the top k_1 most relevant documents are selected from the knowledge base based on the input image. Let e_I denote the embedding of the input image I, and e_i represent the embedding of document D_i in the knowledge base. Here, e_i is encoded from either the metadata or the image of each document, depending on the test setup. The EVA-CLIP model (Sun et al. 2024) is utilized to extract embeddings for both the input and the knowledge base items. The similarity score between the query image and knowledge documents to be retrieved is computed as

$$Sim(I_q, D_i) = \frac{e_I \cdot e_i}{\|e_I\|\|e_i\|}. \quad (3)$$

In the second stage, fine-grained passage retrieval is performed to further retrieve the passages most relevant to the question from the documents related to the image. The top k_1 relevant documents are divided into passages $P = \{P_1, P_2, \ldots, P_n\}$ according to paragraphs or length. Then, we employ the multi-modal retriever PreFLMR (Lin et al. 2024) to refine the retrieval, identifying the most relevant passages from P based on the input image I and query Q.

$$Sim(Q, P_i) = PreFLMR(I, Q, P_i) \quad (4)$$

We then select the top k_2 most relevant passages, $P' = \{P'_1, P'_2, \ldots, P'_{k_2}\}$, as knowledge references.

3.4 Think and Tell - Generation with Thoughts

Chain-of-thought has been demonstrated to effectively enhance the model's reasoning logic and its accuracy in answering questions. Inspired by this, we design a three-stage reasoning process to generate a narrative in response to a customized query, including the Captioning stage, Sub-QA stage, and Conclusion stage.

In the **Captioning stage**, the model needs to describe the content of the image, especially details relevant to the query, making the model's understanding of the image more specific and precise. In the **Sub-QA stage**, the model should break down the query into a series of sub-questions and address them one by one through step-by-step reasoning, ultimately arriving at the answer to the original query. It is recommended that the initial sub-question focuses on fine-grained entity recognition, and the reference document should be consulted when addressing these sub-questions. The final **Conclusion stage** is responsible for summarizing the above reasoning process and generating a fluent and natural narrative. The output of each stage is enclosed within pairs of special tags, `<CAPTION></CAPTION>`, `<SUB-QA></SUB-QA>`, and `<CONCLUSION></CONCLUSION>`, to standardize the model's reasoning steps.

3.5 Data Preparation and Model Training

Data Construction. To construct training data, we utilize Gemini-2.5-Flash (Comanici et al. 2025) to generate the thinking process and final narrative output. First, the model generates query-relevant captions for images. Then, based on ground-truth reference documents, it generates a series of sub-questions and sub-answers. The first sub-question typically identifies the category related to the query to assist subsequent question-answer reasoning. Finally, the caption and the entire sub-question and answer process are summarized to produce a narrative tailored to the query. Descriptions, entity categories, and answers within the narrative are each marked with their respective tags.

Training Approach. Following LLaVA-CoT (Xu et al. 2025), we perform full-parameter fine-tuning on the Llama-3.2-11B-Vision (Meta 2024). Using images, questions, and reference documents as inputs, we perform next-token prediction throughout the entire chain-of-thought process.

Model	Mode	InfoSeek						E-VQA					
		QA	Ent	Des_{IR}	Des_{QR}	OVA_R	OVA_G	QA	Ent	Des_{IR}	Des_{QR}	OVA_R	OVA_G
Closed-Source MLLMs													
GPT-4o-mini-0718 (Menick et al. 2024)	zero-shot	27.8	19.3	77.7	3.3	54.1	2.0	19.5	3.2	78.7	2.0	53.0	3.3
Gemini-2.5-Flash (Comanici et al. 2025)	zero-shot	42.7	40.3	78.5	2.7	51.6	4.0	28.8	18.0	79.6	2.4	51.0	3.9
Gemini-2.5-Flash (Comanici et al. 2025)	RAG	44.9	51.6	79.3	2.9	53.6	**4.2**	**39.4**	17.9	80.3	2.6	53.7	**3.9**
Open-Source MLLMs													
LLaVA-OneVision-7B (Li et al. 2024a)	zero-shot	10.5	4.6	76.0	2.5	51.3	2.9	14.3	1.3	78.3	2.6	51.5	3.1
Qwen2.5-VL-7B (Bai et al. 2025)	zero-shot	22.2	16.3	81.6	2.7	55.3	3.0	21.0	6.2	**82.0**	2.3	54.5	2.9
Qwen2.5-VL-7B (Bai et al. 2025)	RAG	38.7	36.6	**82.0**	3.2	52.9	3.8	38.6	11.7	81.9	2.6	52.7	3.4
Llama-3.2-11B-Vision (Meta 2024)	zero-shot	21.3	14.4	55.7	2.0	58.1	2.9	18.5	4.4	56.2	2.1	59.1	3.0
Llama-3.2-11B-Vision (Meta 2024)	RAG	31.7	20.7	64.9	2.4	56.6	3.2	30.4	6.9	72.1	2.1	56.9	3.0
Our T^4	RAG	**47.4**	**56.1**	79.6	**3.3**	**63.2**	3.9	39.3	**28.1**	80.0	**3.1**	**65.2**	3.8

Table 1: Performance comparison of narrative generation on ImgNarr-23K. "QA" indicates the answer accuracy, while "Ent" indicates entity recognition accuracy. "Des_{IR}" and "Des_{QR}" denotes the relevance of the description to the image and query. "OVA_R" and "OVA_G" denote overall quality by RefCLIPScore and GPT-4o. Higher values indicate better performance.

4 Experiments

To demonstrate the effectiveness of our proposed T^4 model, we conduct quantitative experiments on image narrative generation using the constructed ImgNarr-23K dataset. In addition, we extend the model to knowledge-based VQA without extra finetuning, evaluated on InfoSeek (Chen et al. 2023) and Encyclopedic-VQA (Mensink et al. 2023).

4.1 Experimental Setup

Datasets and External Knowledge Bases. InfoSeek and Encyclopedic-VQA are commonly utilized datasets in knowledge-based VQA and serve as essential data resources for the construction of image narratives.

InfoSeek comprises 1M images and 1.3M question-answering samples, split into 934k for training, 73k for validation, and 348k for testing. The VQA evaluation is conducted on the validation set, as test set groundtruth is unavailable. The validation set are categorized into Unseen-Question (UQ) and Unseen-Entity (UE) types, which do not appear in the training set. Following previous works (Caffagni et al. 2024; Cocchi et al. 2025), we utilize the WiKi-100K as the knowledge base .

Encyclopedic-VQA includes 514k images and 1.03M question-answering samples, with 1M in the training set, 13.6k in the validation set, and 5.8k in the test set. Encyclopedic-VQA comprises four question types: automatic, templated, multi_answer, and 2_hop. The first three types are all one-hop questions. Its corresponding knowledge base is derived from the WiKi-2M dataset.

The proposed **ImgNarr-23K dataset** collects image-question pairs from the above two datasets and constructs image narrative data, comprising a total of 23,747 samples. It covers various specific domains, including animals, plants, landmarks, transportation, etc. Specifically, 7,997 training samples are selected from InfoSeek by retaining only one image for each unique entity-question pair, and 5,000 test samples are randomly chosen from its validation set. From Encyclopedic-VQA, 5,000 training samples are randomly selected, while all 5,750 samples from the original test set are included.

Implementation Details. We use Llama-3.2-11B-Vision (Meta 2024) as the baseline model and perform full parameter finetuning separately on the InfoSeek and Encyclopedic-VQA. For the two-stage retriever, the top 5 documents are retrieved for InfoSeek and the top 50 for Encyclopedic-VQA, due to its larger corpus. The top 10 passages are then selected for both datasets. More details can be found in the supplementary material. Please refer to our project page[1] for the supplementary material, source code and dataset.

4.2 Experiments on Image Narrative Generation

In this section, we evaluate the performance of the proposed T^4 model in generating narratives and compare it with current MLLMs under zero-shot and RAG settings. For MLLM, using images and customized queries as inputs, we design prompts to guide the model to generate outputs in accordance with the standardized narrative format.

Evaluation Metrics. As defined in the Methods section, an image narrative includes the relevant content description, fine-grained entity category, and inferential answer to the query. We extract the corresponding content between special tags from the conclusion, and evaluate them with the following metrics:

(1) Question answering accuracy is calculated using relaxed string matching on InfoSeek and the BERT-based matching score on Encyclopedic-VQA. If there are no <ANSWER> tags in the output, we adopt the last sentence as the answer. **(2) Entity recognition accuracy** is calculatd through string matching with the ground truth categories. **(3) Description relevance**: We use CLIPScore (Hessel et al. 2021) and GPT-4o (Hurst et al. 2024) to respectively assess the relevance of the description with the image and the query. GPT scores range from 1 to 5. **(4) Overall** performance is evaluated by RefCLIPScore between the generated

[1]https://github.com/VIPL-VSU/T4

Model	LLM	Retriever	InfoSeek			E-VQA	
			UQ	UE	All	Single-Hop	All
Closed-Source MLLMs							
GPT-4V* (Achiam et al. 2024)	-	-	15.0	14.3	14.6	26.9	28.1
Gemini-2.5-Flash (Comanici et al. 2025)	-	-	34.4	35.7	35.0	25.0	25.3
Open-Source MLLMs							
LLaVA-OneVision-7B (Li et al. 2024a)	Qwen2	-	11.9	11.0	11.5	15.2	15.0
Qwen2.5-VL-7B (Bai et al. 2025)	Qwen2.5	-	19.1	18.7	18.9	19.1	18.8
Llama-3.2-11B-Vision (Meta 2024)	Llama-3.1-8B	-	17.7	16.8	17.2	19.0	19.1
Retrieval-Augmented Models							
Wiki-LLaVA* (Caffagni et al. 2024)	Llama-3.1-8B	CLIP ViT-L/14+Contriever	28.6	25.7	27.1	18.3	19.6
EchoSight*† (Yan and Xie 2024)	Mistral-7B/Llama-3-8B	EVA-CLIP-8B+Q-Former	-	-	31.3	**41.8**	-
ReflectiVA* (Cocchi et al. 2025)	Llama-3.1-8B	EVA-CLIP-8B	40.4	39.8	40.1	35.5	35.5
Llama-3.2-11B-Vision (Meta 2024)	Llama-3.1-8B	EVA-CLIP-8B+PreFLMR	28.5	27.2	27.8	30.8	29.8
Our T^4	Llama-3.1-8B	EVA-CLIP-8B+PreFLMR	**45.9**	**44.0**	**44.9**	41.2	**39.3**
Retrieval-Augmented Models with GT docs							
ReflectiVA* (Cocchi et al. 2025)	Llama-3.1-8B	EVA-CLIP-8B	57.8	57.4	57.6	**75.2**	-
Llama-3.2-11B-Vision (Meta 2024)	Llama-3.1-8B	EVA-CLIP-8B+PreFLMR	43.2	41.5	42.3	59.2	56.9
Our T^4	Llama-3.1-8B	EVA-CLIP-8B+PreFLMR	**58.4**	**58.4**	**58.4**	**75.2**	**73.4**

Table 2: Performance comparison of knowledge-based VQA on InfoSeek and Encyclopedic-VQA datasets. * indicates the results are reported. †indicates that the results are not directly comparable due to different knowledge bases and validation split. "with GT docs" indicates retrieving top passages from ground-truth documents.

text, image and the reference narrative. In addition, we use GPT-4o to evaluate narrative's overall quality in terms of fluency, clarity, and logical coherence, from 1 to 5. Evaluaiton with GPT are conducted on a randomly selected 10% subset of the test sets. Evaluation prompts are available in the supplementary material.

Results. As shown in Table 1, our fine-tuned T^4 model exhibits significant improvements over the baseline, validating our training strategy. Performance gains in QA and Entity indicate that it can utilize the retrieved knowledge more efficiently, while the overall improvement attests that chain-of-thought reasoning makes the generated paragraphs more accurate and coherent. It outperforms Qwen2.5-VL-7B (Bai et al. 2025) in most metrics, but is slightly lower on Des_{IR}, mainly because Qwen2.5-VL-7B has stronger overall image captioning capability, whereas our descriptions focus more on the local aspects relevant to the queries with higher Des_{QR}. Closed-source models achieve higher zero-shot QA and entity recognition accuracy than open-source models, indicating that they learn fine-grained categories and knowledge better during training. As a result, the performance gains brought by additional knowledge are relatively limited. T^4 performs better or on par with Gemini-2.5-Flash on most tests, achieving remarkable transfer performance through low-cost finetuning with less than 10K samples.

4.3 Experiments on Knowledge-Based VQA

By extracting the answer text between tags in the generated narrative, our method can accomplish the knowledge-based VQA task without additional finetuning. We compare it with MLLMs as well as SoTA retrieval-augmented VQA models on the InfoSeek and Encyclopedic-VQA datasets.

Row	Finetuned	CoT	RAG	InfoSeek		E-VQA	
				VQA	REC	VQA	REC
r1	w/o ft			21.3	14.4	18.5	4.4
r2			✓	31.7	20.7	30.4	6.9
r3			GT	48.6	38.2	59.6	30.6
r4	w/ ft			30.2	28.4	22.5	14.8
r5		✓		32.7	28.7	23.7	17.7
r6			✓	40.5	48.0	35.1	18.3
r7		✓	✓	47.4	56.1	39.3	28.1
r8		✓	GT	**61.3**	**85.2**	**73.4**	**66.2**

Table 3: Ablation study on module effectiveness. "w/o ft" and "w/o ft" indicate the baseline model and our finetuned T^4. "CoT" indicates adding the thinking stage during generation. "RAG" indicates using retrieved knowledge as references, and "GT" means using groundtruth documents.

Evaluation Metrics. We follow the evaluation metrics provided along with the InfoSeek and Encyclopedic-VQA datasets. Questions in InfoSeek are divided into string-based, time-related, and numerical types. For string-based questions, accuracy is measured by exact string matching, while for the latter two types, a relaxed accuracy is calculated. For Encyclopedic-VQA, a BERT-based matching score (BEM) is used for semantic matching between the predicted and groundtruth answers.

Results. As shown in Table 2, the VQA accuracy of both closed-source and open-source MLLMs without RAG is lower than that of current retrieval-augmented models, indicating that incorporating relevant external knowledge can effectively address issues such as wrong fine-grained recog-

Gemini-2.5-Flash: <DESCRIPTION> The image features a prominent street clock standing on a boardwalk beside a wide sandy beach stretching towards the ocean under a vast blue sky. </DESCRIPTION> This location is recognized as part of the <ENTITY> **Long Beach Boardwalk** </ENTITY>. The city is named for its <ANSWER> geographical feature of **a long, sandy beach** </ANSWER>.

User: *Student*
Topic: *General*
Question: *What is this park named after?*

Llama-3.2-11B-Vision: The image depicts a large, ornate clock tower situated on a concrete platform, overlooking a vast body of water. The surrounding environment suggests that it is located in a public park or plaza. The clock tower's design suggests that it may be a memorial or tribute to <ANSWER> **a notable figure or event** </ANSWER>.

Our T⁴:

Thinking…

[Caption stage] *What can I know from the image?*

- *The image captures a serene seaside scene at dusk, featuring a large, ornate clock tower with four faces. In the background, a vast ocean stretches towards a distant horizon, with a sandy beach and a boardwalk visible in the foreground.*

[Sub-QA stage] *How can I answer the questions step by step?*

- *Sub-Q1: What is the name of the park shown in the image?*
- *Sub-A1: The park shown in the image is Jacob Riis Park.*
- *Sub-Q2: What is Jacob Riis Park named after?*
- *Sub-A2: Jacob Riis*

[Conclusion stage] ***How to generate a user-focused narrative response to this visual issue?***

<DESCRIPTION>This image captures a serene seaside scene at dusk, featuring a large, ornate clock tower with four faces. In the background, a vast ocean stretches towards a distant horizon, with a sandy beach and a boardwalk visible in the foreground. </DESCRIPTION> This park is the <ENTITY> **Jacob Riis Park** </ENTITY>. It is named after <ANSWER> **Jacob Riis** </ANSWER>.

Gemini-2.5-Flash: <DESCRIPTION> The image displays a green plant characterized by its finely dissected, fern-like leaves with serrated edges, and small, compact clusters of whitish flowers. </DESCRIPTION> This plant is recognized as <ENTITY> **Yarrow** </ENTITY>.Yarrow is widely distributed across Europe, and the northernmost country in Europe is <ANSWER> **Norway** </ANSWER>.

User: *Scholar*
Topic: *Natural history*
Question: *What is the northernmost country in the continent where this plant is widely distributed?*

Llama-3.2-11B-Vision: The image depicts a plant with serrated leaves and a white flower, which is likely a type of **cannabis plant**. The plant is widely distributed in the Northern Hemisphere, particularly in subarctic regions. The northernmost country in the continent where this plant is widely distributed is <ANSWER> **Norway** </ANSWER>.

Our T⁴ :

Thinking…

[Caption stage] *What can I know from the image?*

- *The image displays a green plant with serrated leaves and a cluster of small, light-colored flowers at its center. The plant is surrounded by other green foliage*

[Sub-QA stage] *How can I answer the questions step by step?*

- *Sub-Q1: What is this plant in the image?*
- *Sub-A1: Urtica dioica.*
- *Sub-Q2: What continent is Urtica dioica widely distributed in?*
- *Sub-A2: North America*
- *Sub-Q3: What is the northernmost country in North America?*
- *Sub-A3: Greenland*

[Conclusion stage] ***How to generate a user-focused narrative response to this visual issue?***

<DESCRIPTION>The image displays a green plant with serrated leaves and a cluster of small, light-colored flowers at its center. </DESCRIPTION> This plant is <ENTITY> **Urtica dioica** </ENTITY>. It is widely distributed in North America, and the northernmost country on that continent is <ANSWER> **Greenland** </ANSWER>.

Figure 4: Visualization of image narrative generation examples, where we compare the results provided by T^4 with those from Gemini-2.5-Flash and Llama-3.2-11B-Vision.

nition and incomplete learning in specialized domain knowledge. Among the retrieval-augmented models, our T^4 uses less training data while outperforming existing SoTA methods. This demonstrates the efficiency of our framework. If the groundtruth document is provided to enhance knowledge relevance, the accuracy of T^4 is further improved, demonstrating its higher upper bound.

4.4 Ablation Study

Effectiveness of Different Modules. We validate the effectiveness of each module on the baseline and finetuned model, with results shown in Table 3. For r4 and r6 that do not contain CoT, we re-trained the model to directly generate narratives by removing the thinking stage. Compared to the zero-shot setting (r1, r4) where outputs are directly generated without CoT or RAG, CoT (r5) and RAG (r2, r6) can individually contribute to performance improvements. Results of their combination in r7 yield the best results. Comparing results under the same settings, such as r1 v.s. r4, and r2 v.s. r6, the finetuned model shows significant enhancement, validating the effectiveness of our training strategy.

Results with Oracle Documents. To evaluate performance upper bound, we provide groundtruth documents to obtain the most relevant knowledge entries. As shown in Table 3, comparing r2 with r3, and r7 with r8, the inclusion of groundtruth documents leads to significant improvements across all metrics. On the baseline model, there is an average increase of 21.8 points, and on the finetuned model, an average increase of 28.8 points. This indicates that our model can more effectively utilize groundtruth information. This advantage is particularly evident in applications such as exhibition introductions, where accurate categories of exhibits can be easily obtained, thereby better demonstrating the superiority of the proposed model.

4.5 Qualitative Results

Visualization of Narrative Generation. In Fig. 4, we present visual examples of image narrative generation on specialized domains like landmarks and plants, where our T^4 is shown to provide more accurate and fluent responses. Furthermore, T^4's transparent reasoning process enhances logical consistency and interpretability, thereby boosting user trust in the generated results. More results can be found in the supplementary material.

5 Conclusion

In this work, addressing the demands for in-depth knowledge in image content description, we introduce the customized image narrative generation, aiming to generate a natural language response to users' specific concerns regarding a given input image. Furthermore, we propose T^4, generating image narrative through cascade steps: **T**ailor, re**T**rieve, **T**hink, and **T**ell. Taking an image and various prompts as input, it first predicts queries based on the user's expertise. Next, it retrieves external knowledge as references, thereby enhancing the reliability of the output. It utilizes chain-of-thought reasoning to decompose the response process, and finally generates accurate, logically coherent customized image narratives. Experimental results demonstrate that our proposed approach excels at generating image narratives and achieves SoTA on the knowledge-based VQA without extra finetuning. T^4 shows great potential to generate customized content suited to specific domains.

Acknowledgements

This work is partially supported by Natural Science Foundation of China under contract No. U21B2025, and National Key R&D Program of China No. 2023YFF1105104.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 6077–6086.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17682–17690.

Bu, S.; Li, T.; Yang, Y.; and Dai, Z. 2024. Instance-level Expert Knowledge and Aggregate Discriminative Attention for Radiology Report Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 14194–14204.

Caffagni, D.; Cocchi, F.; Moratelli, N.; Sarto, S.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 1818–1826.

Chen, Y.; Hu, H.; Luan, Y.; Sun, H.; Changpinyo, S.; Ritter, A.; and Chang, M.-W. 2023. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 14948–14968.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 24185–24198.

Cocchi, F.; Moratelli, N.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2025. Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 9199–9209.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Cornia, M.; Baraldi, L.; and Cucchiara, R. 2019. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 8299–8308.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 49250–49267.

Encyclopædia Britannica. 2025. Britannica Kids. https://kids.britannica.com/. Accessed: 2025-08-01.

Gu, X.; Chen, G.; Wang, Y.; Zhang, L.; Luo, T.; and Wen, L. 2023. Text With Knowledge Graph Augmented Transformer for Video Captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 18941–18951.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7514–7528.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. GPT-4o System Card. arXiv:2410.21276.

Hussien, M. M.; Melo, A. N.; Ballardini, A. L.; Maldonado, C. S.; Izquierdo, R.; and Sotelo, M. A. 2025. RAG-based explainable prediction of road users behaviors for automated driving using knowledge graphs and large language models. *Expert Systems with Applications*, 265: 125914.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P. S.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 22199–22213.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 2024a. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 19730–19742.

Li, J.; Vo, D. M.; Sugimoto, A.; and Nakayama, H. 2024b. EVCap: Retrieval-Augmented Image Captioning with External Visual-Name Memory for Open-World Comprehension. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 13733–13742.

Li, Y.; Zhang, X.; Cheng, X.; Tang, X.; and Jiao, L. 2024c. Learning consensus-aware semantic knowledge for remote sensing image captioning. *Pattern Recognition*, 145: 109893.

Lin, W.; Mei, J.; Chen, J.; and Byrne, B. 2024. PreFLMR: Scaling Up Fine-Grained Late-Interaction Multi-modal Retrievers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5294–5316.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 19730–19742.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 3195–3204.

Menick, J.; Lu, K.; Zhao, S.; Wallace, E.; Ren, H.; Hu, H.; Stathas, N.; and Such, F. P. 2024. GPT-4o mini: advancing cost-efficient intelligence. *Open AI: San Francisco, CA, USA*.

Mensink, T.; Uijlings, J.; Castrejon, L.; Goel, A.; Cadar, F.; Zhou, H.; Sha, F.; Araujo, A.; and Ferrari, V. 2023. Encyclopedic VQA: Visual Questions About Detailed Properties of Fine-Grained Categories. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 3113–3124.

Meta, A. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20: 2024.

Mitra, C.; Huang, B.; Darrell, T.; and Herzig, R. 2024. Compositional Chain-of-Thought Prompting for Large Multimodal Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 14420–14431.

Park, J. S.; Bhagavatula, C.; Mottaghi, R.; Farhadi, A.; and Choi, Y. 2020. VisualCOMET: Reasoning about the Dynamic Context of a Still Image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 508–524.

Ramos, R.; Elliott, D.; and Martins, B. 2023. Retrieval-augmented Image Captioning. arXiv:2302.08268.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 146–162.

Shah, S.; Mishra, A.; Yadati, N.; and Talukdar, P. P. 2019. KVQA: Knowledge-Aware Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8876–8884.

Shum, K.; Diao, S.; and Zhang, T. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. In *In Findings of the Association for Computational Linguistics: EMNLP 2023*, 12113–12139.

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3784–3803.

Sun, Q.; Wang, J.; Yu, Q.; Cui, Y.; Zhang, F.; Zhang, X.; and Wang, X. 2024. EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters. arXiv:2402.04252.

Urbanek, J.; Bordes, F.; Astolfi, P.; Williamson, M.; Sharma, V.; and Romero-Soriano, A. 2024. A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 26700–26709.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 3156–3164.

Wang, N.; Xie, J.; Wu, J.; Jia, M.; and Li, L. 2023. Controllable image captioning via prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2617–2625.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 24824–24837.

Xu, G.; Jin, P.; Wu, Z.; Li, H.; Song, Y.; Sun, L.; and Yuan, L. 2025. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2087–2098.

Yan, Y.; and Xie, W. 2024. EchoSight: Advancing Visual-Language Models with Wiki Knowledge. In *In Findings of the Association for Computational Linguistics: EMNLP 2024*, 1538–1551.

Yang, L.; Tang, K.; Yang, J.; and Li, L. 2017. Dense Captioning With Joint Inference and Visual Context. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 1978–1987.

Yang, S.; Wu, X.; Ge, S.; Zhou, S. K.; and Xiao, L. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80: 102510.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 11809–11822.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 6713–6724.

Zeng, Z.; Xie, Y.; Zhang, H.; Chen, C.; Chen, B.; and Wang, Z. 2024. MeaCap: Memory-Augmented Zero-shot Image Captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 14100–14110.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal Chain-of-Thought Reasoning in Language Models. arXiv:2302.00923.