Exploring Diverse In-Context Configurations for Image Captioning

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, Xin Geng

Pattern Learning and Mining (PALM) Lab http://palm.seu.edu.cn/
School of Computer Science and Engineering, Southeast University, China

Outline

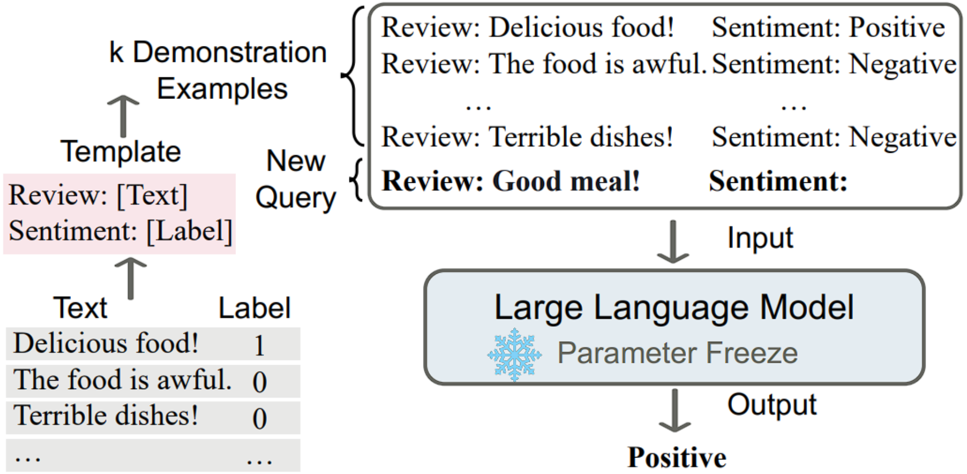
Background

Methods

Experiments

Takeaways

In-Context Learning:
Allows a model to adapt to a task using a few examples

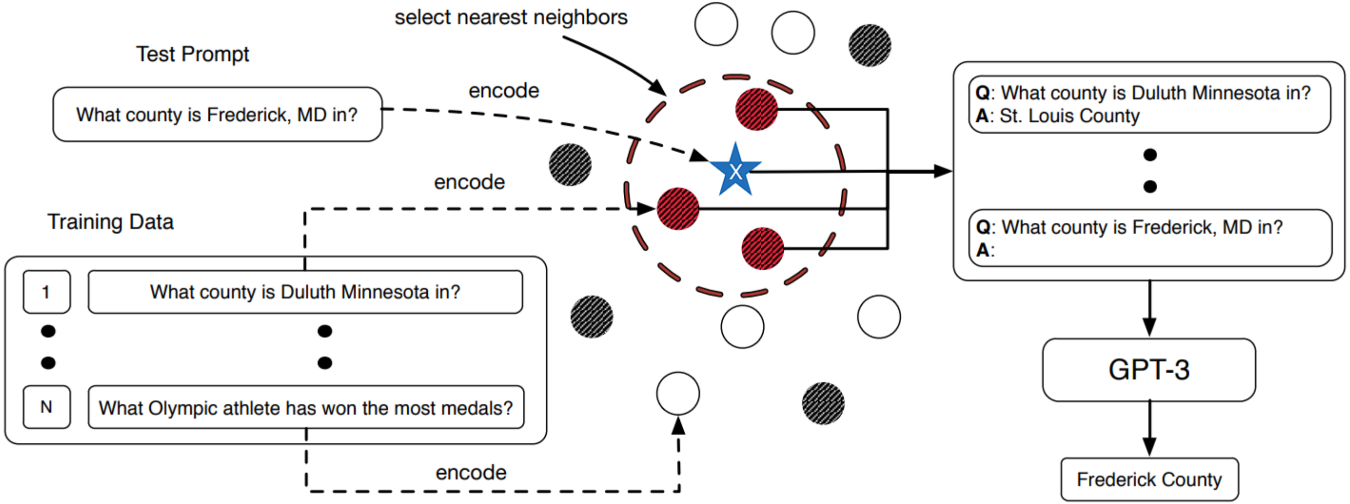


"We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-ofthe-art fine-tuning approaches." -- "Language Models are Few-Shot Learners" (GPT-3)

Image Source: Dong, Qingxiu, et al. "A survey for in-context learning." arXiv preprint arXiv:2301.00234 (2022).

Previous Study: Demonstration Selection

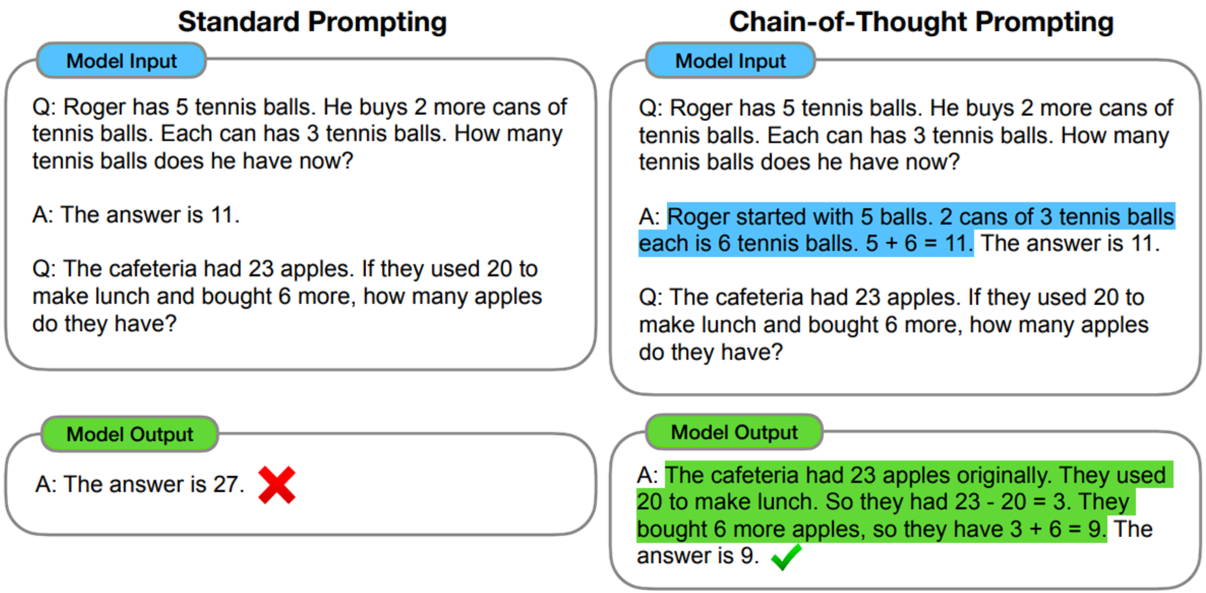
Liu et al.[1] suggest retrieving semantically-similar examples corresponding to a test sample



References
[1] Liu, Jiachang, et al. "What Makes Good In-Context Examples for GPT-3?." DeeLIO 2022

Previous Study: Demonstration Formatting

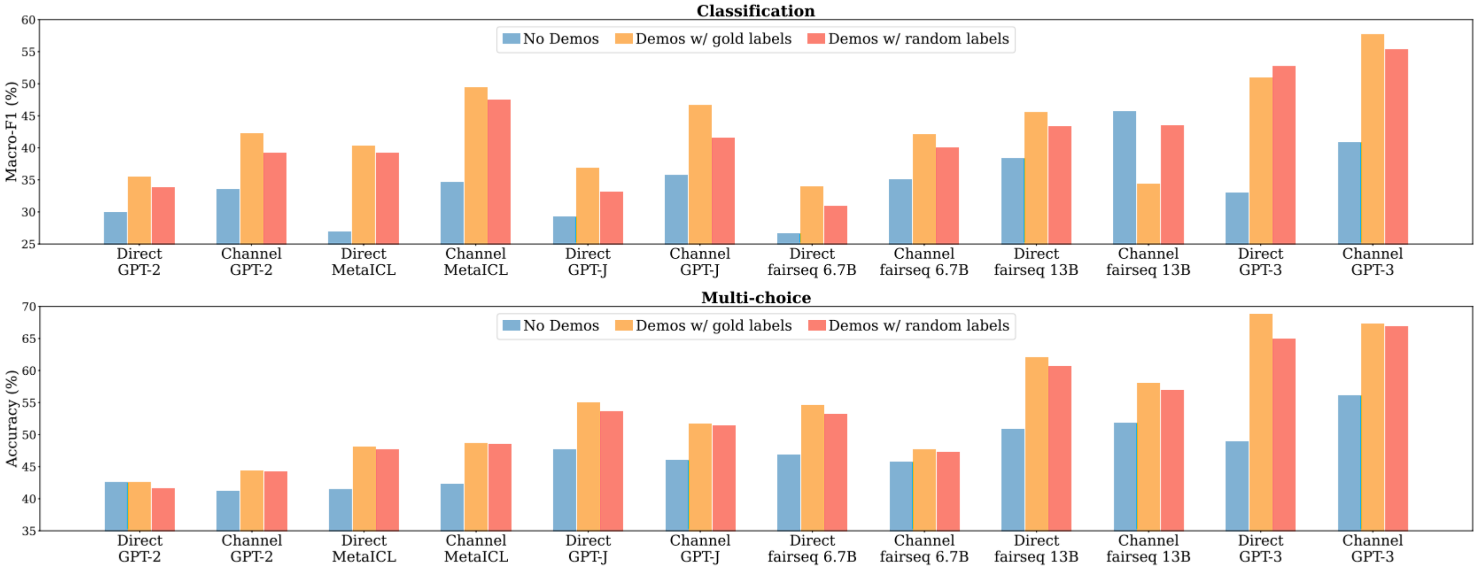
Wei et al.[1] adding intermediate reasoning steps, commonly known as "Chain of Thought."



References
[1] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." NeurIPS 2022

Previous Study: Mechanism Exploration

Min et al.[1] find that even random label replacements have minimal impact on performance.



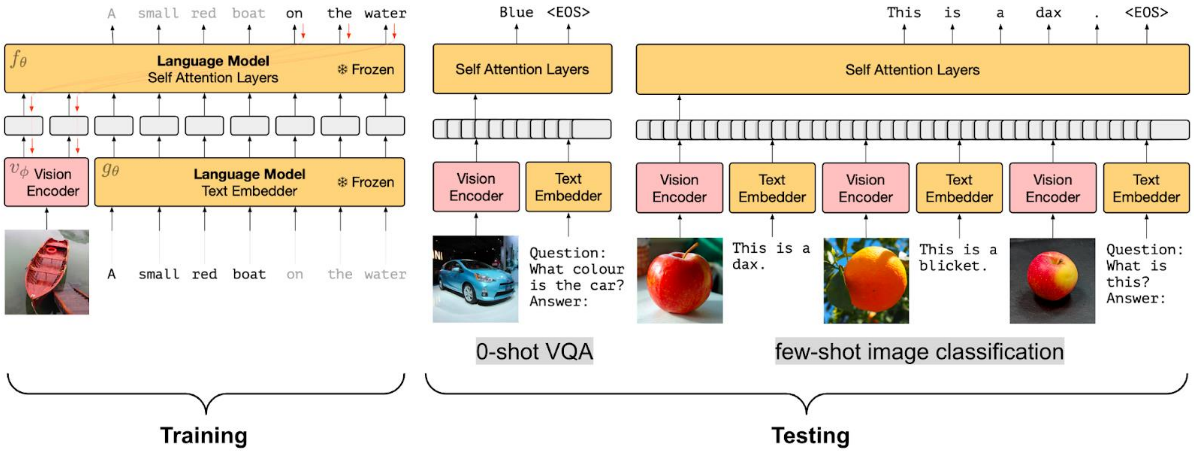
References

[1] Min, Sewon, et al. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?." EMNLP 2022.

Status Quo: From LLMs to VLMs

- Numerous Vision Language Models (VLMs), such as Flamingo[1] and MiniGPT-4[2] have emerged

- The exploration of in-context learning configurations on VLMs is still limited



References
[1] Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." NeurIPS 2022
[2] Zhu, Deyao, et al. "Minigpt-4: Enhancing vision-language understanding with advanced large language models."

Image Source: https://lilianweng.github.io/posts/2022-06-09-vlm/

Outline

Background

Methods

Experiments

Takeaways

From Single-Modal to Multi-Modal: More Complex

Query

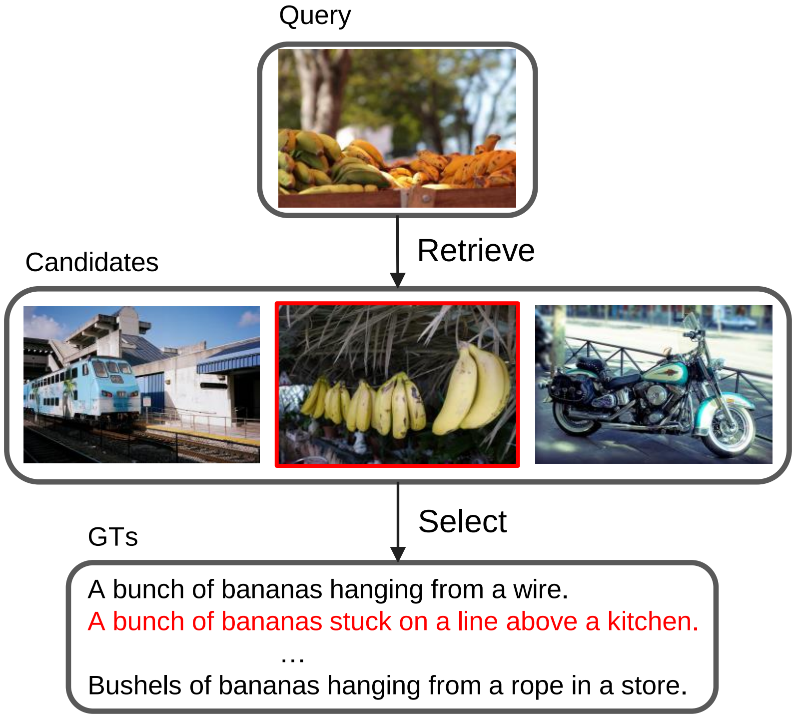
Review: Good meal! Sentiment:

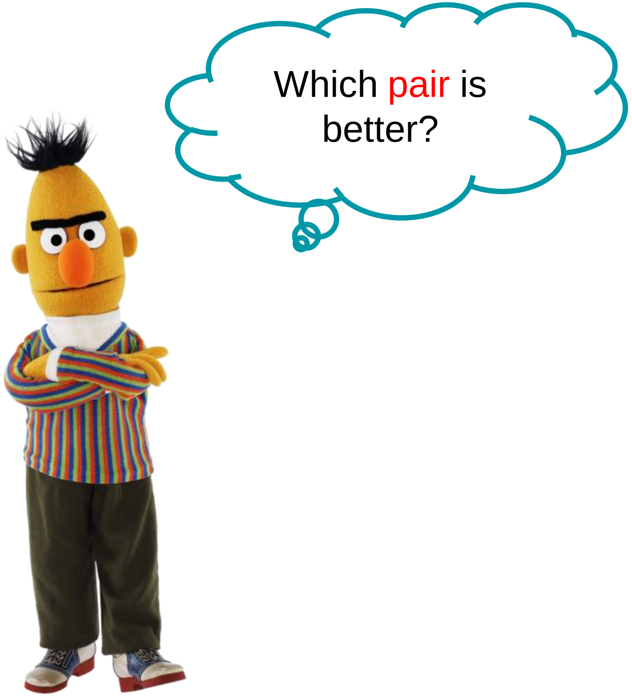
Retrieve

Candidates

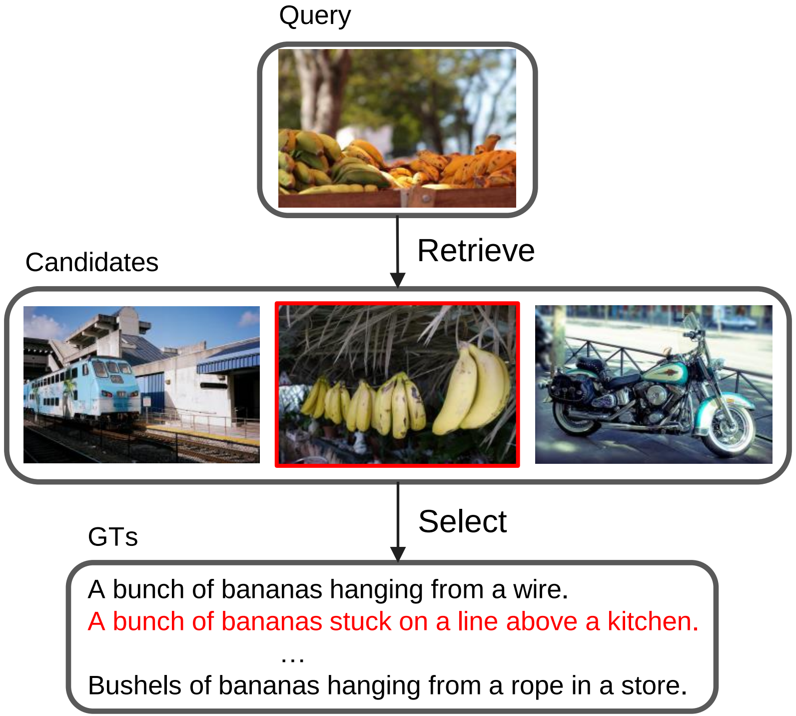
Review: Delicious food! Sentiment: Positive
Review: The movie is awful. Sentiment: Negative
… …
Review: Terrible dishes! Sentiment: Negative

From Single-Modal to Multi-Modal: More Complex





From Single-Modal to Multi-Modal: More Complex



Step1: Given a test image, how to select the proper image?

Step2: Given the selected image, how to choose the suitable caption?

Method: Image Selection Strategies

Step1: Given a test image, how to select the proper image?

1. Random Selection (RS)
2. Similarity-based Image-Image Retrieval (SIIR)
3. Similarity-based Image-Caption Retrieval (SICR)
4. Diversity-based Image-Image Retrieval (DIIR)

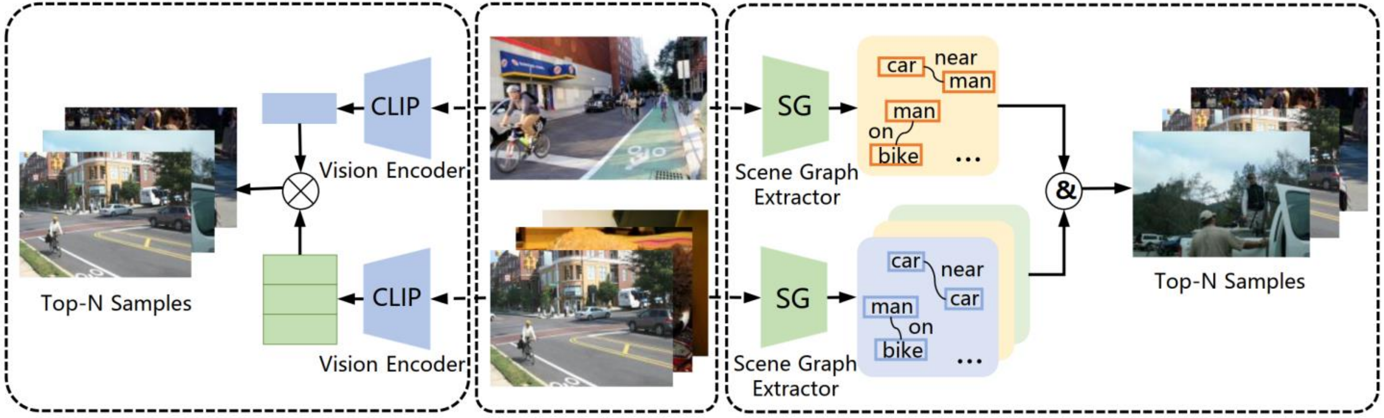
We randomly select k examples for few-shot in-context learning.

Method: Image Selection Strategies

Step1: Given a test image, how to select the proper image?

1. Random Selection (RS)
2. Similarity-based Image-Image Retrieval (SIIR)
3. Similarity-based Image-Caption Retrieval (SICR)
4. Diversity-based Image-Image Retrieval (DIIR)

We use some models to extract the representations from the images, such as CLIP or VinVL.
Then retrieve the top-k examples.



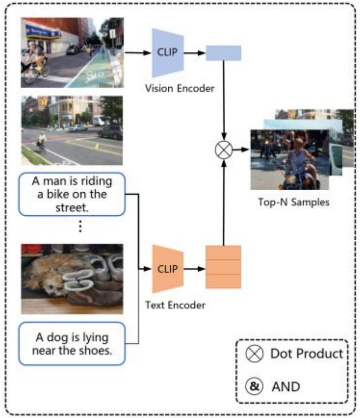
(a) SIIR-CLIP (b) SIIR-TAG

Method: Image Selection Strategies

Step1: Given a test image, how to select the proper image?

1. Random Selection (RS)
2. Similarity-based Image-Image Retrieval (SIIR)
3. Similarity-based Image-Caption Retrieval (SICR)
4. Diversity-based Image-Image Retrieval (DIIR)

As CLIP could embedding the image and text into same latent space.
We could to retrieve the proper caption for the test image then get the pairs.

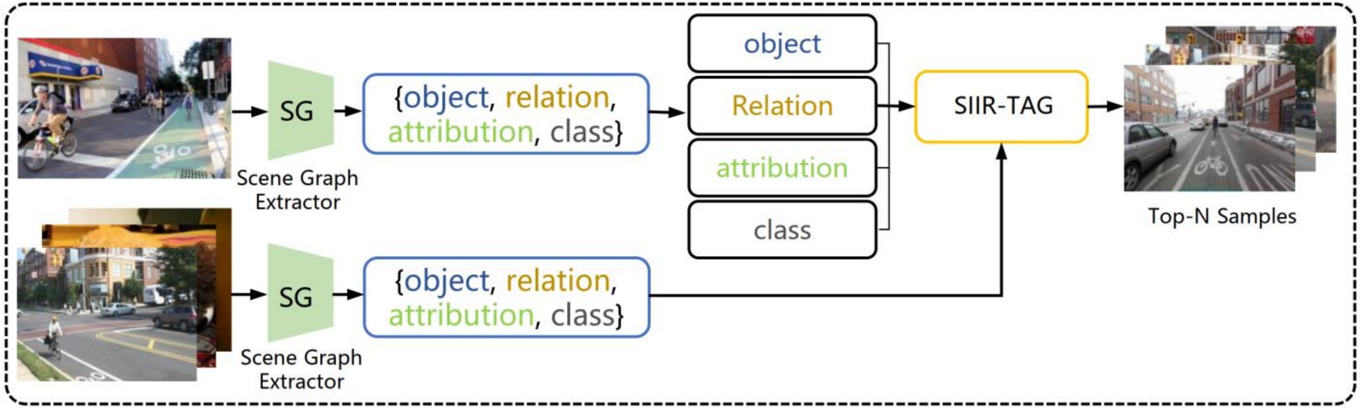


(d) SICR-CLIP

Method: Image Selection Strategies

Step1: Given a test image, how to select the proper image?

1. Random Selection (RS)
2. Similarity-based Image-Image Retrieval (SIIR)
3. Similarity-based Image-Caption Retrieval (SICR)
4. Diversity-based Image-Image Retrieval (DIIR)



(c) DIIR-TT

Method: Caption Assignment Strategies

Step2: Given the selected image, how to choose the suitable caption?

1. Ground Truth Caption (GTC)
2. Model Generated Caption (MGC)
3. Model Generated Caption as Anchor (MGCA)
4. Iterative Prompting (IP)

For MSCOCO Dataset, each image has five human-annotated captions. We choose the first caption in our experiments

① A close up of a young person at a table eating cake.
② A small girl takes a bite of chocolate cake.
③ A young girl eating a piece of chocolate cake.
④ A little girl taking a big bite out of chocolate cake.
⑤ A young child enjoying a serving of cake and ice cream.

Method: Caption Assignment Strategies

Step2: Given the selected image, how to choose the suitable caption?

1. Ground Truth Caption (GTC)
2. Model Generated Caption (MGC)
3. Model Generated Caption as Anchor (MGCA)
4. Iterative Prompting (IP)

Given a image, we can use a VLM or an offline captioner to generate corresponding caption

Vision Language Models
Or
Offline Captioner

A little girl eating a piece of chocolate cake.

Method: Caption Assignment Strategies

Step2: Given the selected image, how to choose the suitable caption?

1. Ground Truth Caption (GTC)
2. Model Generated Caption (MGC)
3. Model Generated Caption as Anchor (MGCA)
4. Iterative Prompting (IP)

Once get the generated caption, we could compute which GTC have higher CIDEr with it.

MGC

A little girl eating a piece of chocolate cake.

Select →

GTC

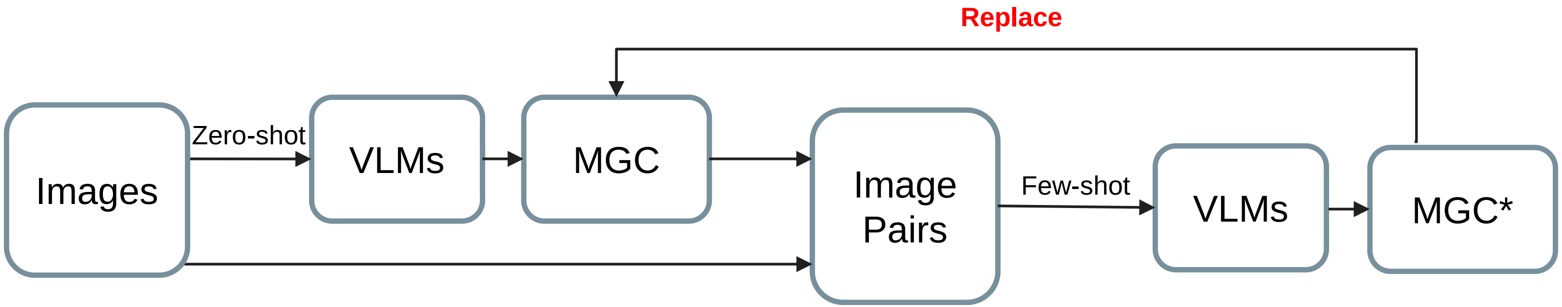
① A close up of a young person at a table eating cake.
② A small girl takes a bite of chocolate cake.
③ A young girl eating a piece of chocolate cake.
④ A little girl taking a big bite out of chocolate cake.
⑤ A young child enjoying a serving of cake and ice cream.

Method: Caption Assignment Strategies

Step2: Given the selected image, how to choose the suitable caption?

1. Ground Truth Caption (GTC)
2. Model Generated Caption (MGC)
3. Model Generated Caption as Anchor (MGCA)
4. Iterative Prompting (IP)

We generate captions using MGC-VLM and then using these captions paired with the images to iteratively prompt VLM for enhanced captions.



Outline

Background

Methods

Experiments

Takeaways

Effects of Caption Qualities

Conclusion 1: the performance typically improves with an increase in shot numbers. However, the rate of improvement varies, or even declines, depending on the quality of the captions.

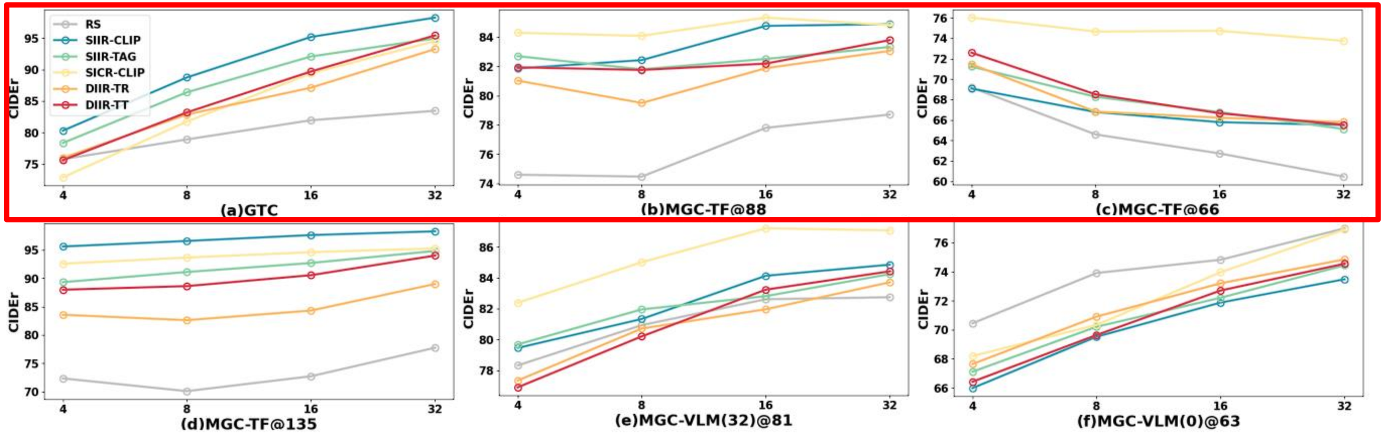


Figure 4: The line charts of various in-context images with diverse caption-assignment strategies.

Effects of Caption Qualities

Conclusion 2: up to a certain threshold of descriptiveness, the VLM more readily identifies simpler sentence structures, enhancing the generation of captions, particularly when the images supply sufficient visual cues.

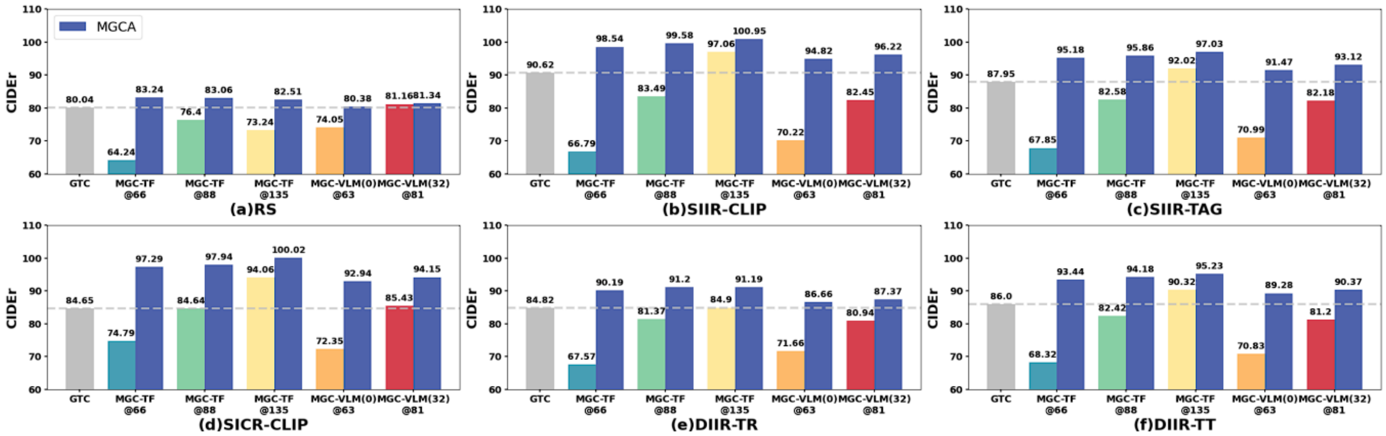
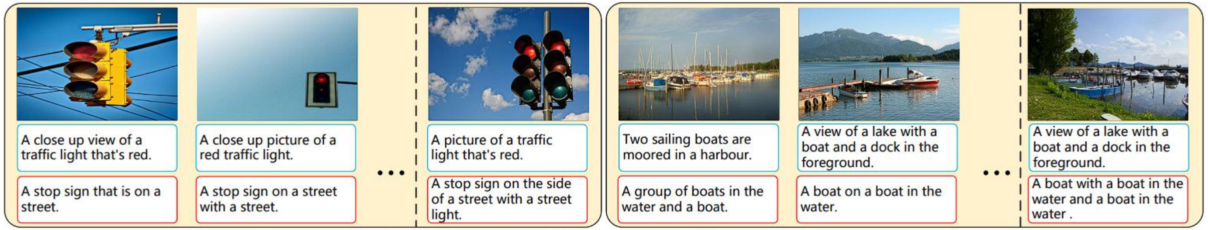


Figure 5: The histograms of various in-context captions with diverse image-selection strategies.

Effects of Caption Qualities

Conclusion 2: up to a certain threshold of descriptiveness, the VLM more readily identifies simpler sentence structures, enhancing the generation of captions, particularly when the images supply sufficient visual cues.

(a) MGC-TF@135(blue) vs. GTC(red)



(b) MGC-VLM(0)@63(blue) vs. MGC-TF@66(red)

Effects of Caption Qualities

Conclusion 3: MGCA consistently improve over GTCs, likely because they verbalize major image patterns, such as salient objects. This helps identify the GTC with the most detailed information about these patterns, which helps VLMs generate better captions.

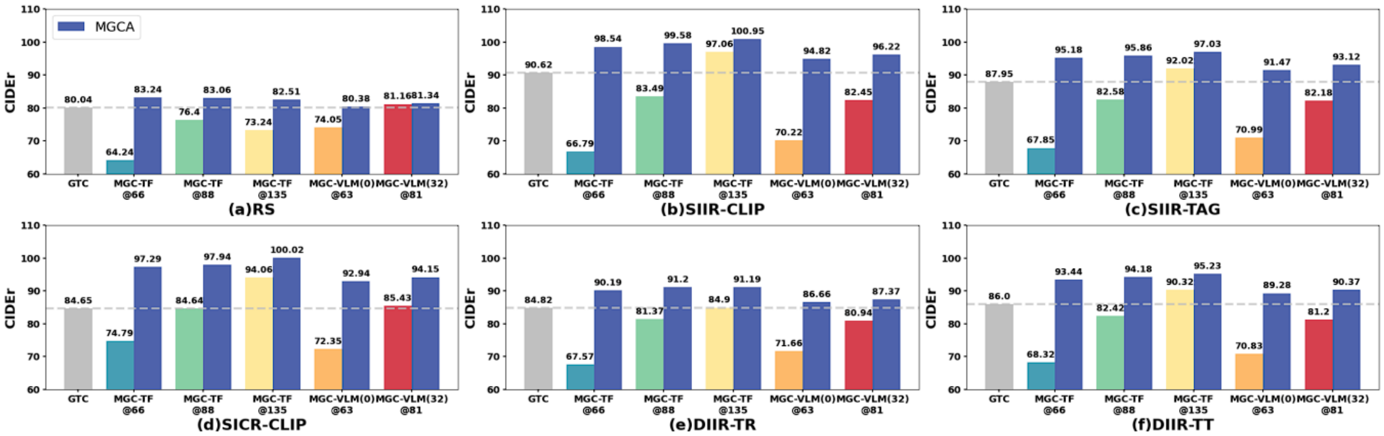


Figure 5: The histograms of various in-context captions with diverse image-selection strategies.

Effects of Caption Qualities

Conclusion 4: The results in the table suggest that extended VLM iterations are redundant, as both MGC-VLM(0) and MGC-VLM(32) stabilize after the third and second iterations, respectively. Even when limited to only 32-shot GTCs, two iterations of IP are sufficient to achieve performances comparable to those seen when all GTCs are used.

Iter	1	2	3	4	5
MGC-VLM(0)	63.0	74.1	79.9	79.3	77.3
MGC-VLM(32)	85.3	80.5	79.4	78.9	77.1

Table 1: The CIDEr scores of IP in different iterations.

Effects of Image Qualities

Conclusion 5: The effectiveness of using similar images is closely linked to the quality of the corresponding captions. We need to consider the synergy between modalities.

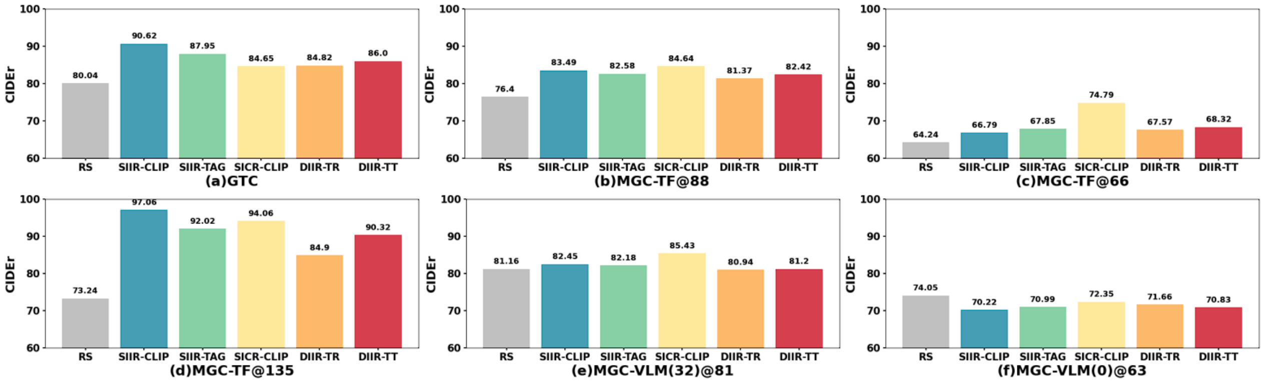
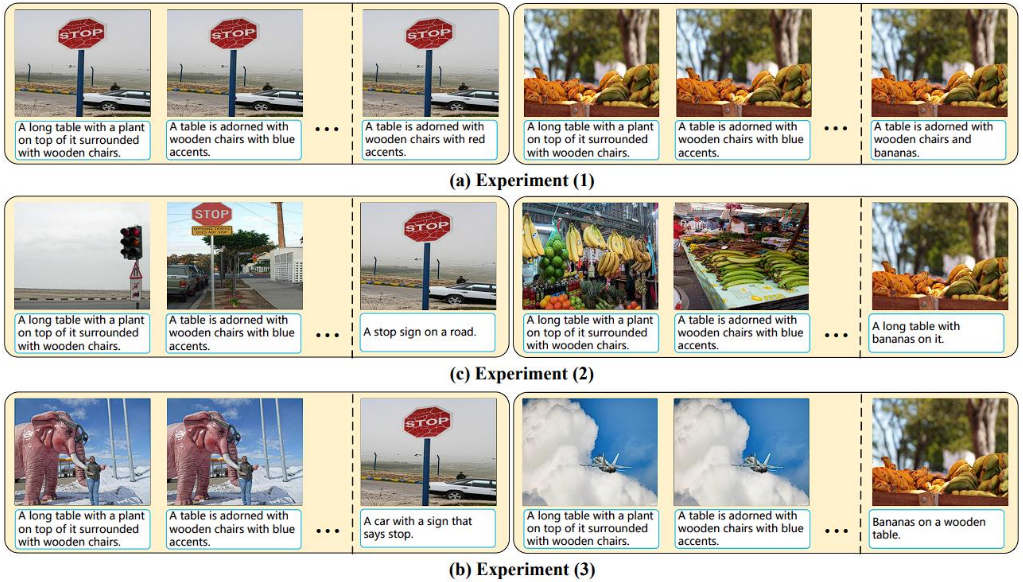


Figure 4: The histograms of various in-context images with diverse caption-assignment strategies.

Effects of Image Qualities

Conclusion 6: when in-context images are similar to the test image, VLM may take a short-cut by leveraging in-context captions to generate a new one, rather than learning how to caption from the in-context pairs



Outline

Background

Methods

Experiments

Takeaways

Takeaways

❖ **Objective**

- **Examine the impact of varied multi-modal in-context configurations using image captioning as a case study.**

❖ **Key Findings**

- **The VLM is better at identifying and generating captions for simpler sentence structures up to a certain level of descriptiveness. This is especially true when the images provide enough visual information.**
- **The VLM tends to take shortcuts when the test image is similar to in-context images. Instead of learning to caption from the in-context pairs, it reuses captions from similar in-context images to generate a new one.**

❖ **Limitations**

- **Relied on Open-Flamingo, which underperforms due to less training data compared to official Flamingo.**
- **The findings need to be validated on more models.**

Thanks!
Any questions?