多模态大语言模型中的上下文学习

杨旭 东南大学

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站	K+全球软件研发行业创新峰会	时间：2024.06.21-22
K+峰会 敦煌站	K+思考周®研习社	时间：2024.10.17-19
K+峰会 香港站	K+思考周®研习社	时间：2024.11.10-12

AiDD峰会

AiDD峰会 上海站	AI+研发数字峰会	时间：2024.05.17-18
AiDD峰会 北京站	AI+研发数字峰会	时间：2024.08.16-17
AiDD峰会 深圳站	AI+研发数字峰会	时间：2024.11.08-09

K+峰会详情

AiDD峰会详情

演讲嘉宾

杨旭

东南大学 副教授

杨旭博士2021年6月从南洋理工大学计算机科学与技术系获工学博士学位，导师为蔡剑飞，张含望教授。现为东南大学计算机科学与工程学院、软件学院、人工智能学院副教授、任东南大学新一代人工智能技术与交叉应用教育部重点实验室副主任。现主要从事视觉文本多模态大模型应用研究以及一种新的大模型训练-部署模式：学习基因的研究。

目录
CONTENTS

PART 01
Background

"Why do we need In Context Learning? "

The Development of GPT

GPT (2018)

GPT-2 (2019)

1.5B Parameters

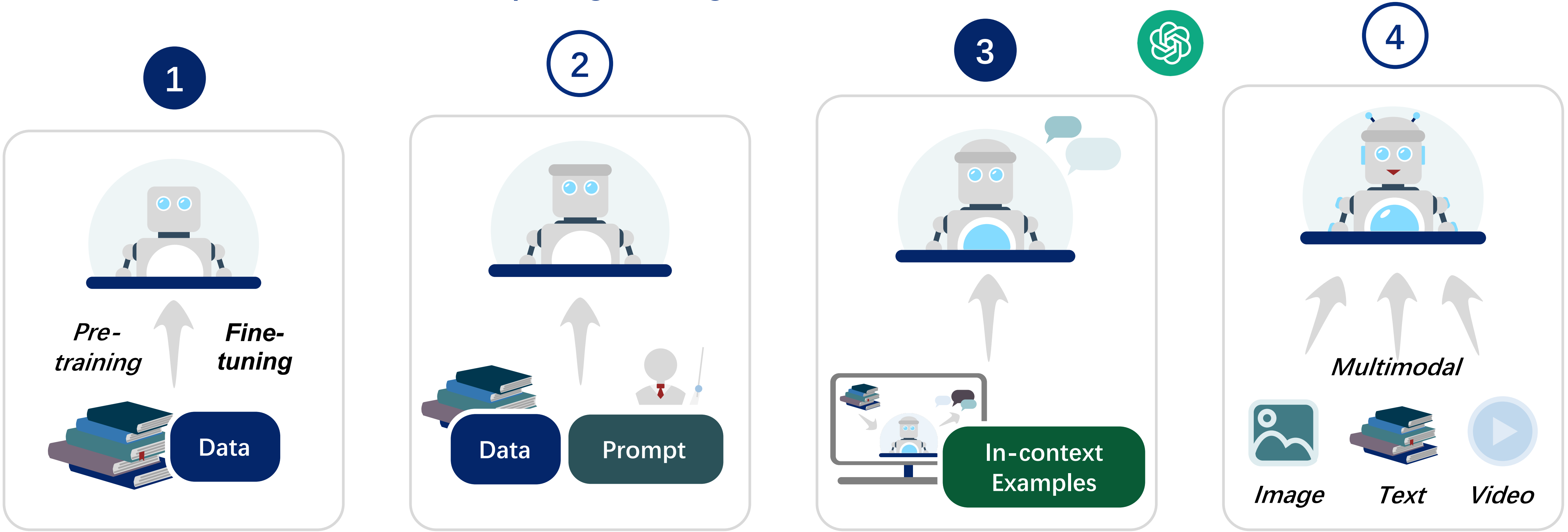
Prompt Engineering

GPT-3 (2020)

175B Parameters

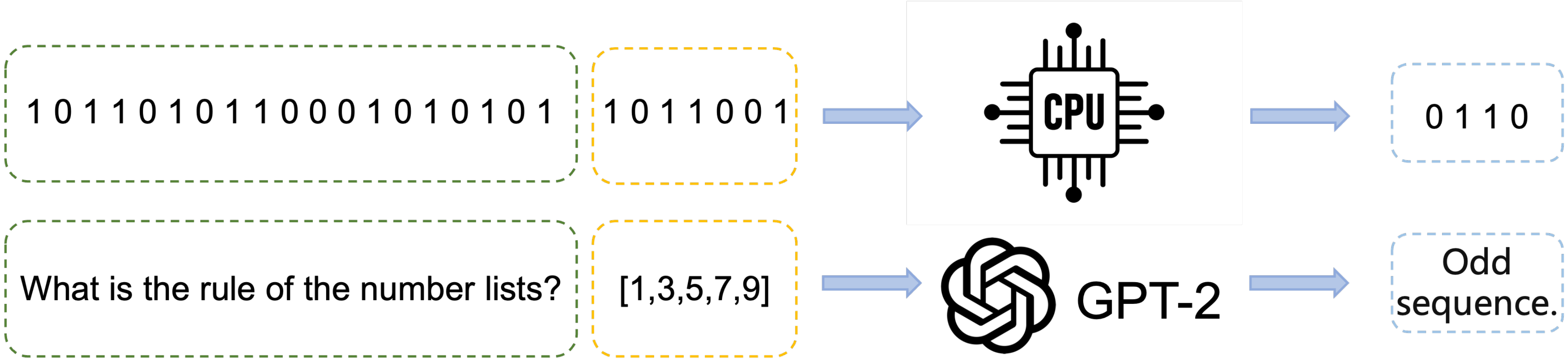
In-context Learning

GPT-4 (2023)



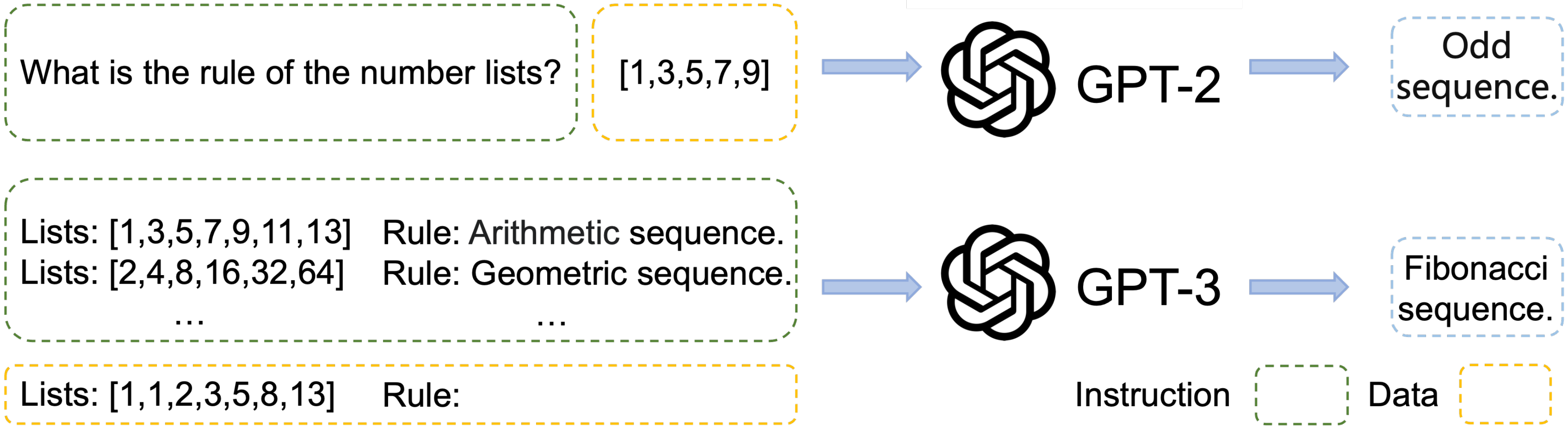
GPT-2's Capability of Prompt Engineering

- GPT-2 exhibits a distinctive feature known as **"prompt engineering"**.
- This can be compared to the architecture of modern computers, where both data and commands exist in the form of 0s and 1s encoding.



GPT-3’s Capability of In-Context Learning

- GPT-3 possesses a unique capability known as **“In-context learning”.**
- It will learn the representation of tasks from the provided in-context examples.



Why In-Context Learning?

Prompt Engineering —few shot→ In-Context Learning

Prompt Engineering

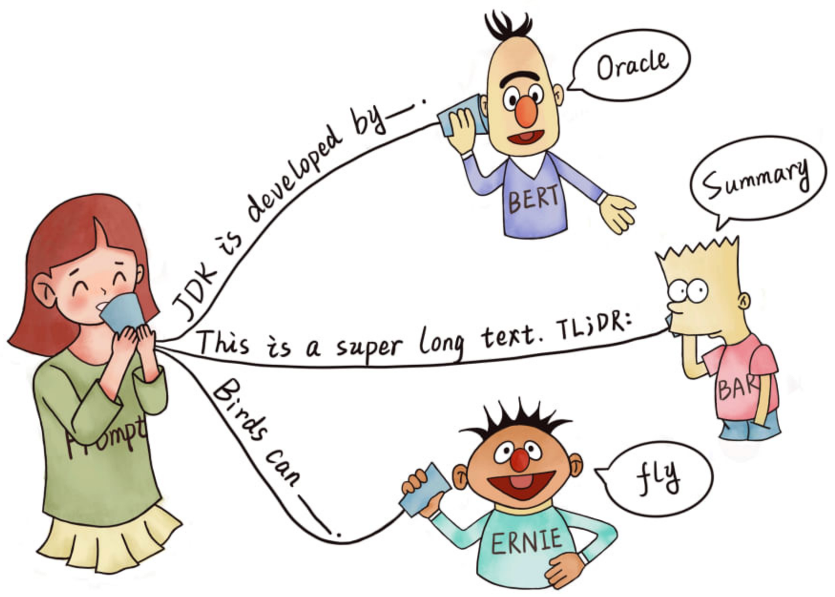
Yield precise responses

Unlock the potential of LLMs

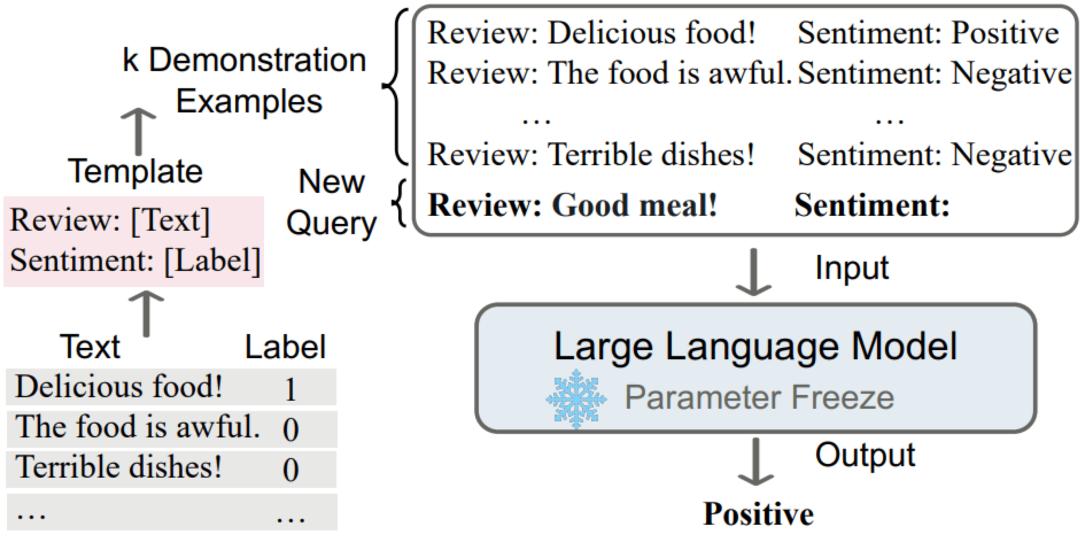
In-Context Learning

A specialized prompt engineering

Adapt to a task using a few examples



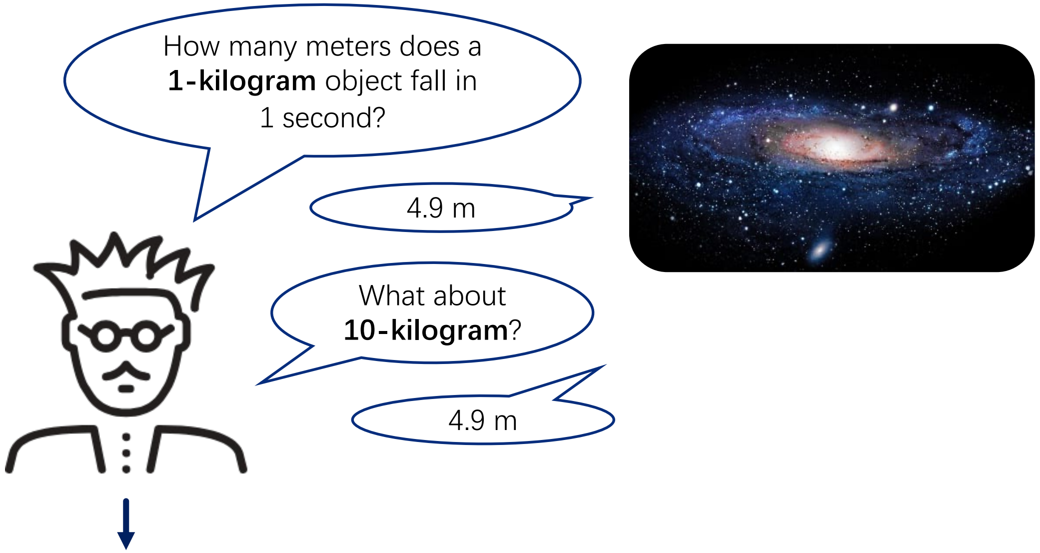
Liu, Pengfei, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

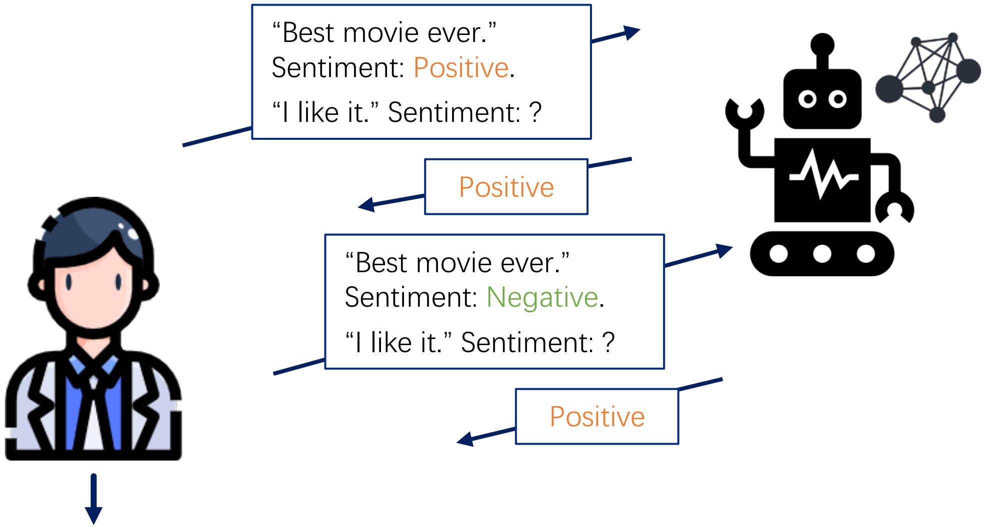


Dong, Qingxiu, et al. A survey for in-context learning.

Why In-Context Learning?

"outside-in" methodologies to unravel the inner properties of LLMs





Objects fall with a constant acceleration due to gravity, regardless of their mass.

Providing incorrect examples does not affect the LLM's ability to make correct judgments.

Pros of ICL

- **Flexible controllability**
- **Encapsulate more information**

GPT-4: Large Multimodal Model

What is LMM?

Process visual data & understand and generate natural language

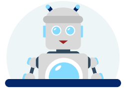
What color is the purse?

blue

Answer questions about the images

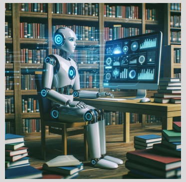
How does this food taste?

Delicious, especially the cake!



Refer to visual information in conversations

How about GPT-4?



These two images represent two different robots, respectively…

Excellent Multimodal capabilities

Incorporate the understanding of visual content

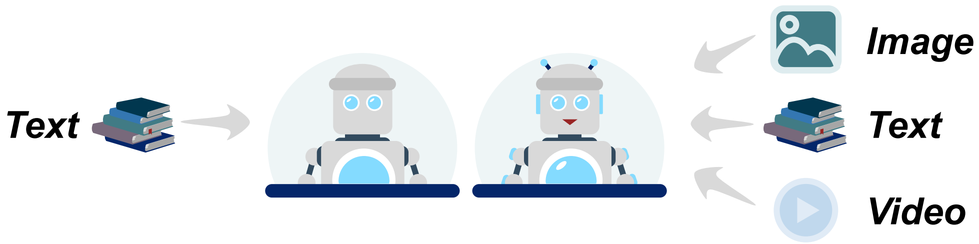
Not open-source

internal workings and training processes are opaque

Why Multimodal Model In-Context Learning?

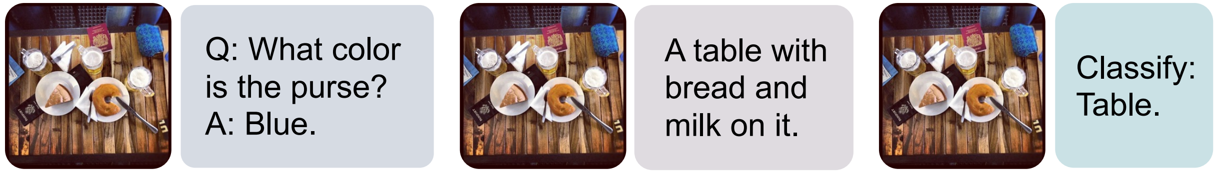
The development of large models from single-modal to multi-modal

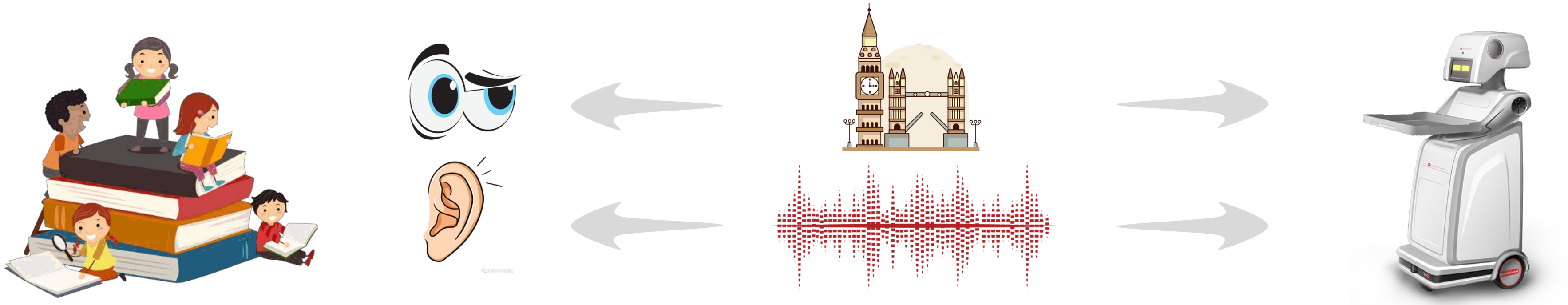
Expands the application scope of the model: various image/video understanding tasks.



Visual Question Answering

Image Caption



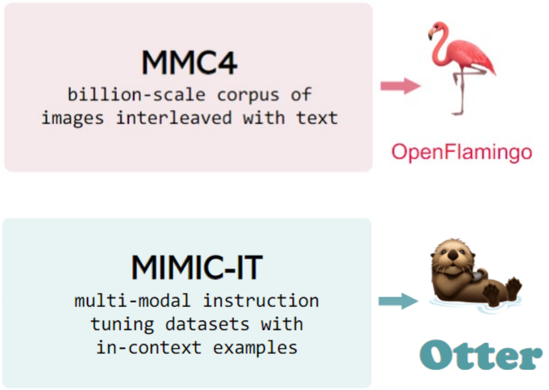


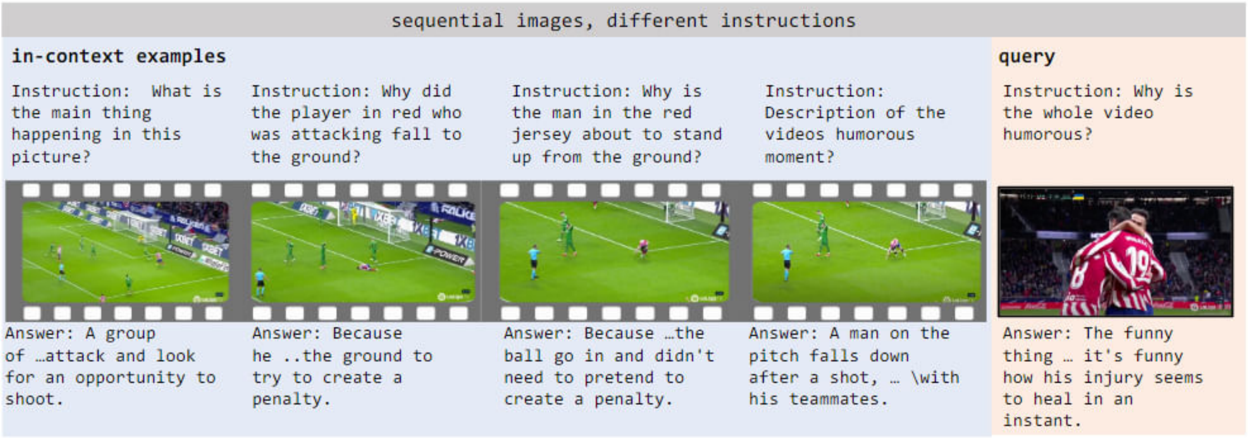
Imitate real humans and achieve multi-modal analogy capabilities

Why Multimodal Model In-Context Learning?

- **Less research** in the **Multimodal Model** In-Context Learning
- Most of the work only considers the field of Natural Language Processing

- Some large multimodal models are not well adapted to in-context learning, such as miniGPT-4, LLAVA, mPLUG owl, etc.
- Large multimodal model with good in-context learning: Flamingo, Otter, **IDEFICS**···





PART 02
Heuristic-based configuration strategies

"Take IC and VQA as examples"

Exploring Diverse In-Context Configurations for Image Captioning （NIPS 2023）

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

arXiv: https://arxiv.org/abs/2305.14800
code: https://github.com/yongliang-wu/ExploreCfg

Exploring Diverse In-Context Caption: Background and Motivation

- Transitioning from **single-modal** to **multi-modal** leads to increased complexity.
  - In image modality, which image optimizes testing?
  - In caption modality, what is the ideal choice for model generation?

Query

Review: Good meal!   Sentiment:

Retrieve

Candidates

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

Query

Retrieve

Candidates

Select

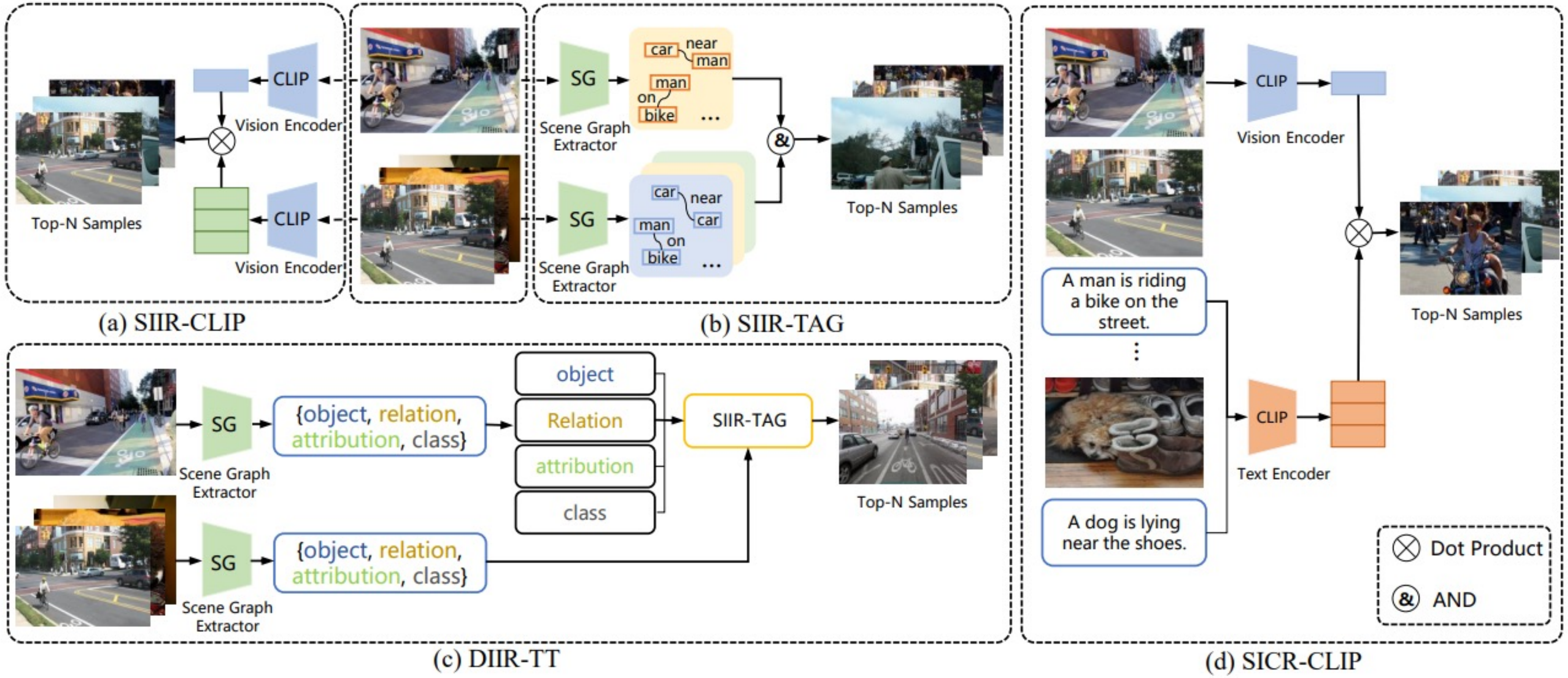
GTs

A bunch of bananas hanging from a wire.
A bunch of bananas stuck on a line above a kitchen.
...
Bushels of bananas hanging from a rope in a store.

Which pair is better?

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

- **Random Selection** (RS): Randomly select k examples for few-shot in-context learning.
- **Similarity**-based **Image-Image** Retrieval (SIIR)
- **Similarity**-based **Image-Caption** Retrieval (SICR)
- **Diversity**-based **Image-Image** Retrieval (DIIR)



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

- Ground Truth Caption (GTC)
- Model Generated Caption (MGC)
- Model Generated Caption as Anchor (MGCA)
- Iterative Prompting (IP)

Each image has five human-annotated captions.
Choose the first caption in our experiments

Use a VLM or an offline captioner to generate corresponding caption

Compute which GTC have higher CIDEr with the generated caption.

Generate captions and then using these captions paired with the images to iteratively prompt VLM for enhanced captions

Exploring Diverse In-Context Caption： Caption Assignment Strategies

- Model Generated Caption (MGC)
  - Given an image, we can use a VLM or an offline captioner to generate caption.
  - It might be helpful since the generated caption usually have the same pattern with the output.

Vision Language Model
Or
Offline Captioner

A little girl eating a piece of chocolate cake.

A vision language model or offline caption to generate caption as in-context examples.

Exploring Diverse In-Context Caption： Caption Assignment Strategies

- Model Generated Caption as Anchor (MGCA)
  - Once get the generated caption, We can compute CIDEr scores to find the best caption.
  - The selected one will have the advantages of both GTC and MGC, more precise expression and more consistent pattern.

Model Generated Caption

A little girl eating a piece of chocolate cake.

Select →

Ground truth Caption

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.

We can use the model-generated caption as anchor to select the best caption from human-annotated captions.

Exploring Diverse In-Context Caption: Conclusions

- Similar Images lead to **short-cut inference**.
  - (1) Same as test image (2) Similar images (3) Random images
  - Ensure the captions are irrelevant to the images to avoid biased inferences.

From top to bottom: The outputs start from imitation to inferencing from the vision cues.

Exploring Diverse In-Context Caption： Conclusions

- Simpler sentence patterns are more easily recognized by the VLM.
  - Ground truth captions use more diverse words and complex patterns
    Which have more precise expression
  - Model-generated captions have more salient objects and simple patterns
    Which have more consist patterns

The top: Model-generated captions. The bottom: Ground truth captions.

Exploring Diverse In-Context Caption： Conclusions

- There is a **synergy effect** between the two modalities.
  - When similar images are used, lower-quality captions can become toxic examples
  - When dissimilar images are used, the negative effects of these low-quality captions are diminished.

Image Similarity	Caption Quality	4-shot	8-shot	16-shot	32-ahot	mean
High	High	95.64	96.62	97.66	98.32	97.06
Low	High	72.35	70.10	72.73	77.76	73.23
High	Low	65.98	69.52	71.88	73.49	70.22
Low	Low	70.45	73.92	74.83	77.00	74.05

How to Configure Good In-Context Sequence for Visual Question Answering

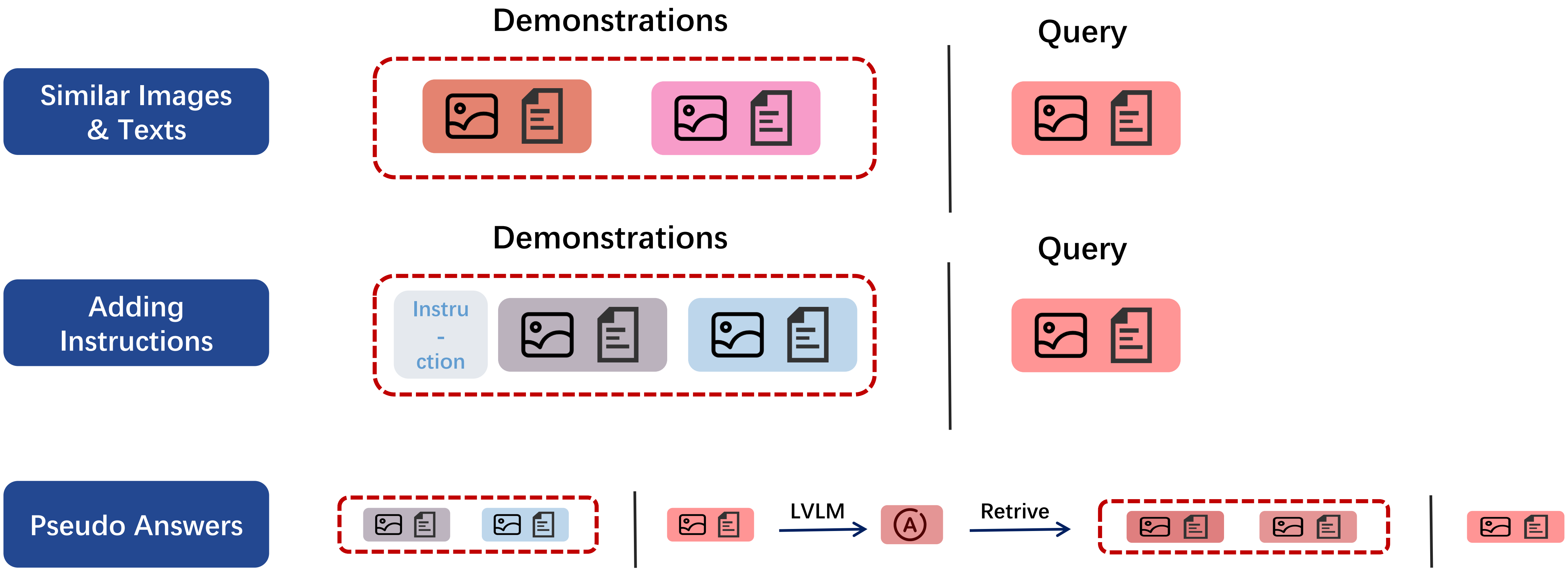
Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, Xu Yang

arXiv: https://arxiv.org/abs/2312.01571

code: https://github.com/GaryJiajia/OFv2_ICL_VQA

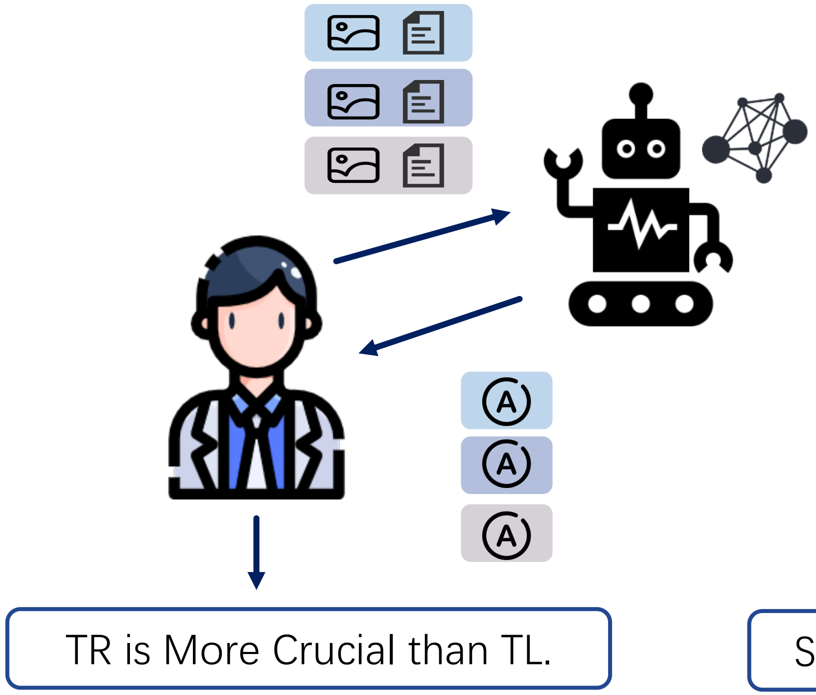
How to Configure Good In-Context Sequence for VQA: Background

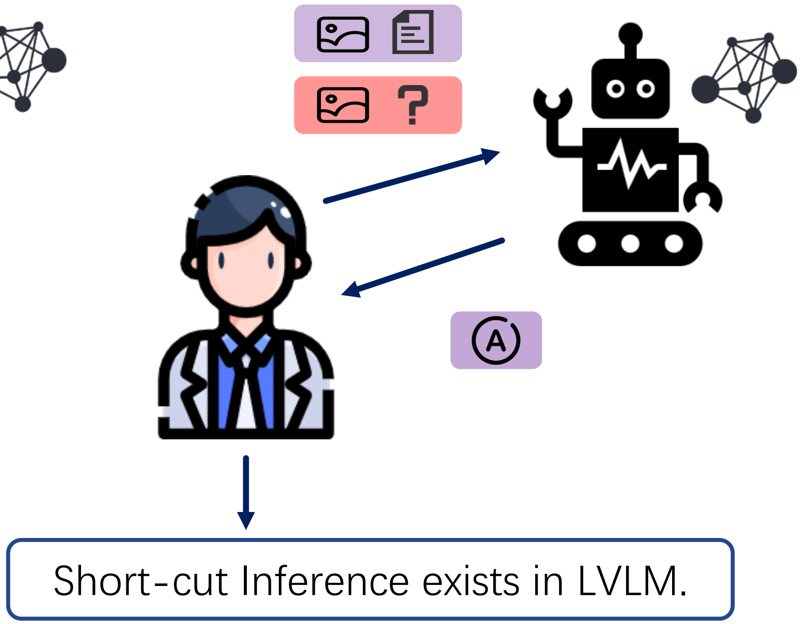
Explore effective In-context examples configuration strategies

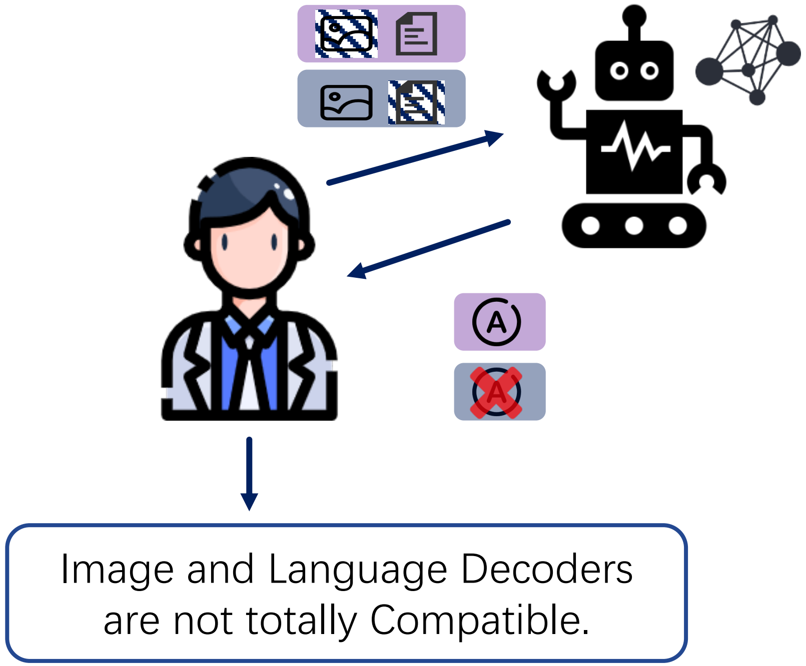


How to Configure Good In-Context Sequence for VQA: Background

Gain a better understanding of the inner properties of LVLM

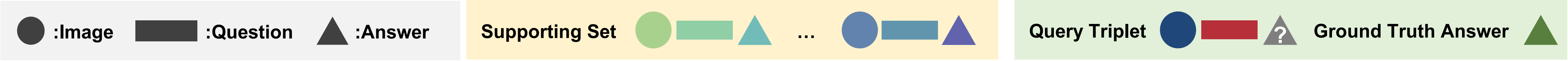




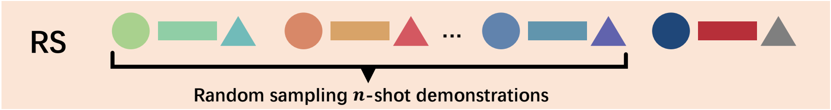


How to Configure Good In-Context Sequence for VQA: Approach

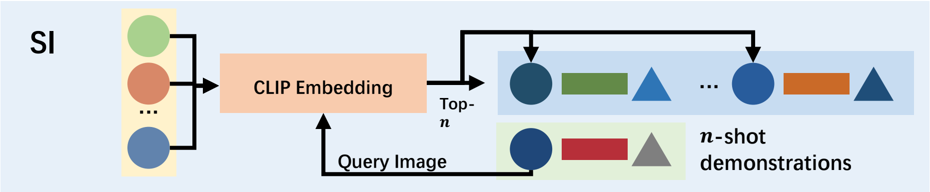
Retrieving In-context examples



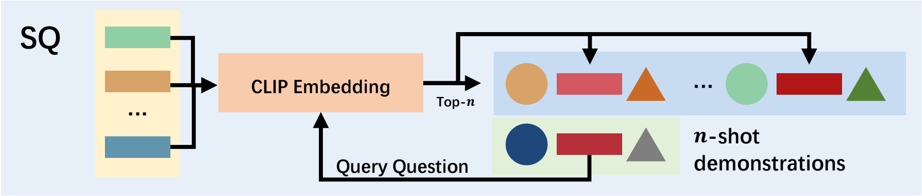
Random Sampling (RS)



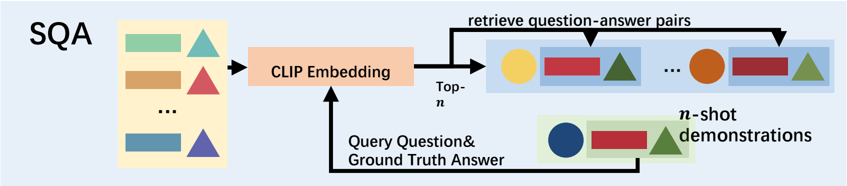
Retrieving via Similar Image (SI)



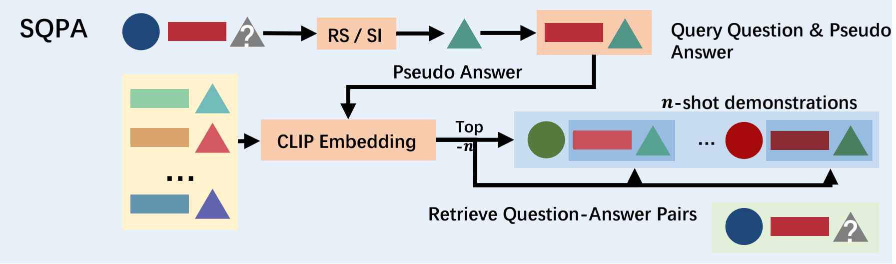
Retrieving via Similar Questions (SQ)



Retrieving via Similar Question&Answer (SQA)



Retrieving via Similar Question&Pseudo Answer(SQPA)

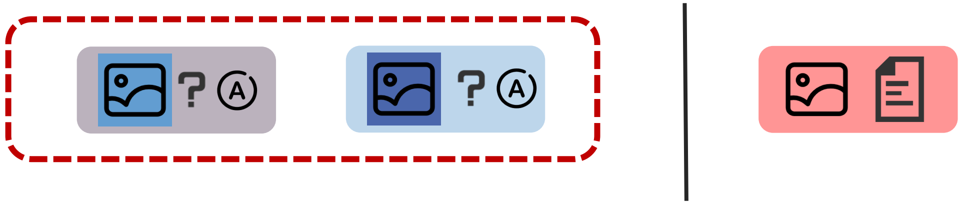


How to Configure Good In-Context Sequence for VQA: Approach

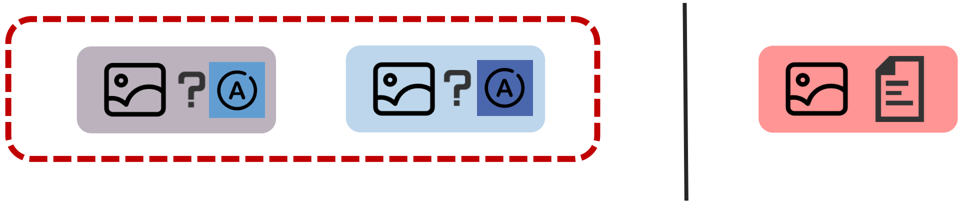
Manipulating examples

Mismatching the Triplet

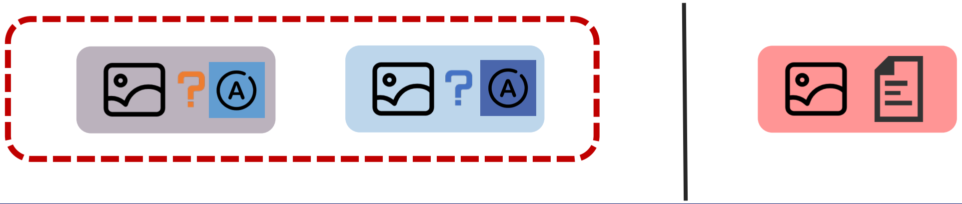
Mismatching Image (MI)



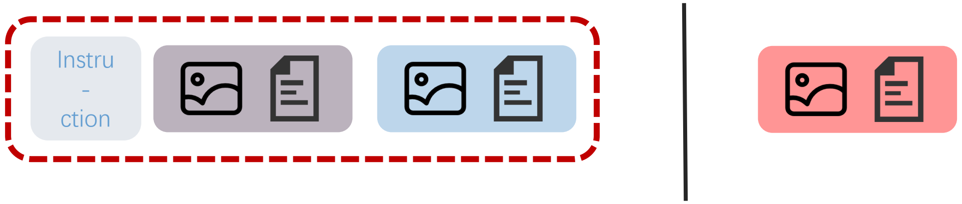
Mismatching Answer (MA)



Mismatching Question-Answer pair (MQA)



Using Instructions



e.g.

According to the previous question and answer pair, answer the final question.

<image>Question:What number is on the bus? Short Answer:284<|endofchunk|>

<image>Question:Where would a taxi park to wait for a customer? Short Answer:curb<|endofchunk|>

<image>Question:What is the man doing in the street? Short Answer:

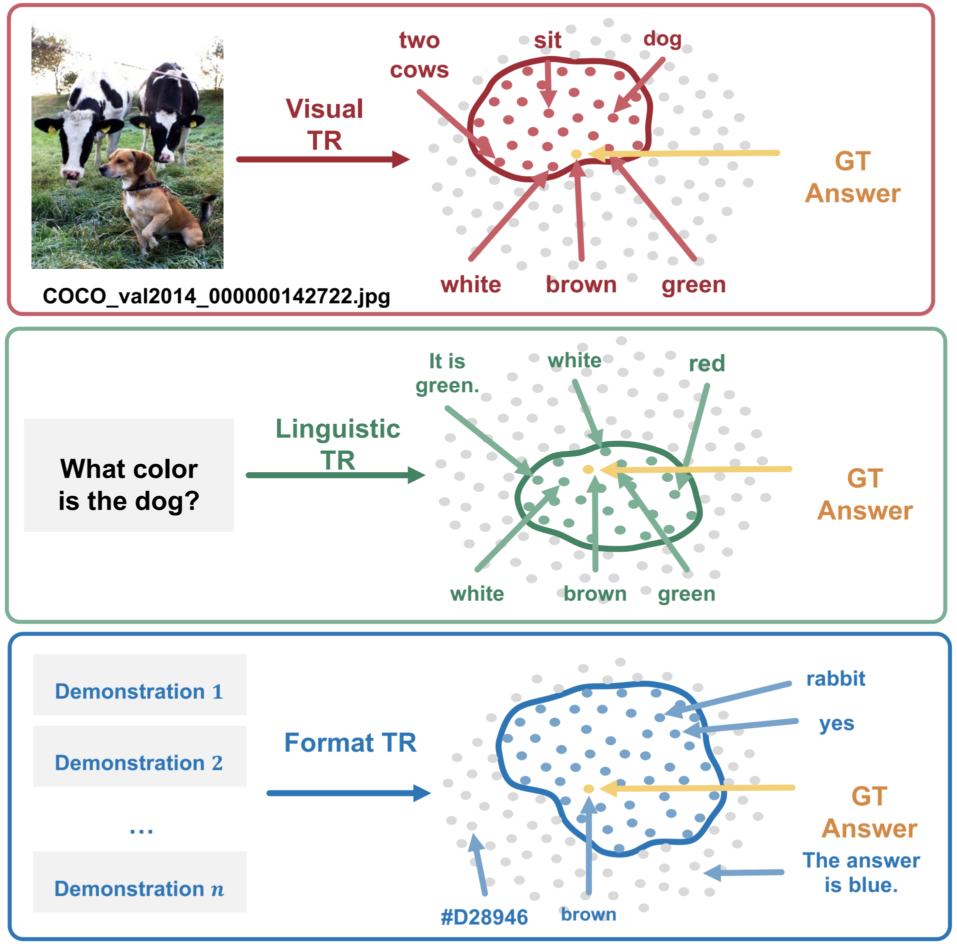
How to Configure Good In-Context Sequence for VQA: Approach

Extend TR and TL Hypothesis in the VL domain

Task Recognition

Recognizes the **distribution of the task**

Applying **pre-trained priors** of LLM



The recall of **pre-trained visual** / language knowledge

Identify:

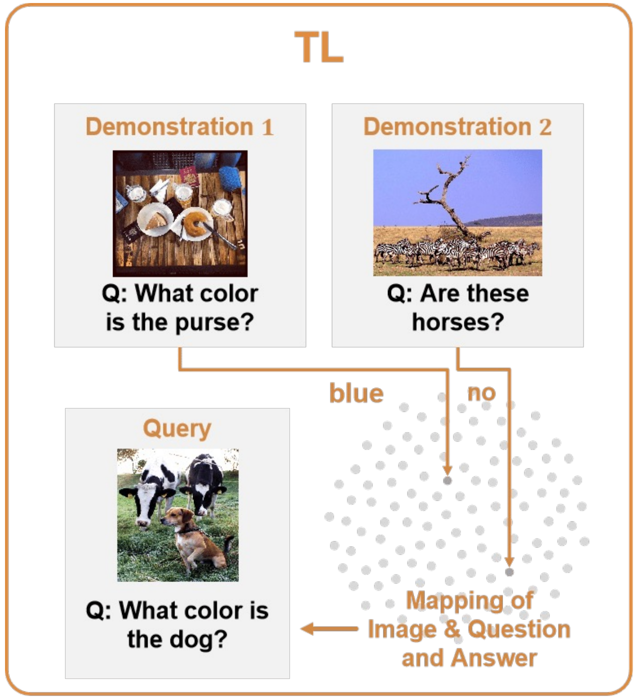
- task format,
- input distribution
- label space from demonstrations

How to Configure Good In-Context Sequence for VQA: Approach

Extend TR and TL Hypothesis in the VL domain

Task Learning

Learn the **mapping relationship** between QA pairs from the demonstrations

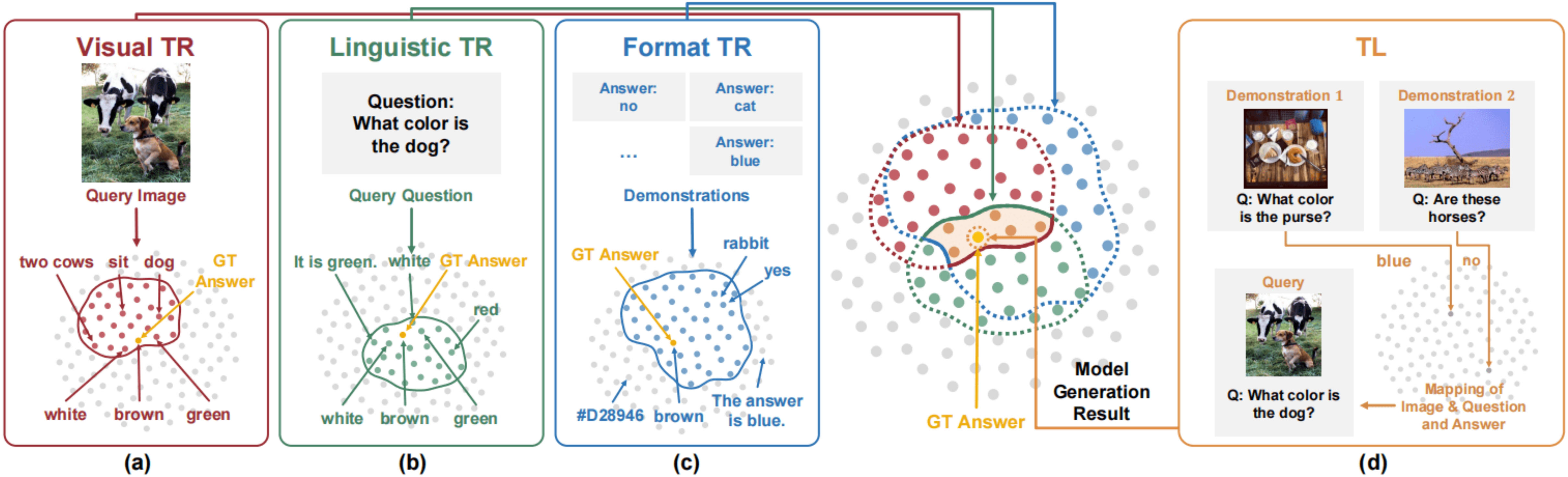


- Treats QAs from demonstrations as "**training samples**"
- **Implicit learning process** analogous to explicit fine-tuning

How to Configure Good In-Context Sequence for VQA: Approach

Extend TR and TL Hypothesis in the VL domain

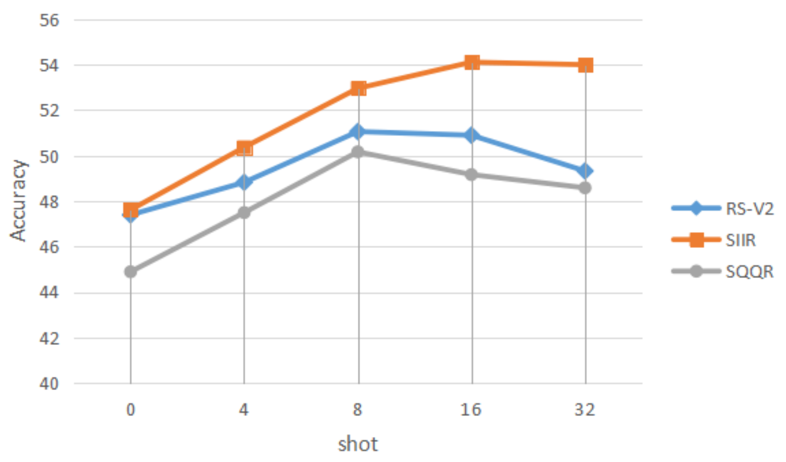
In ICL , TR and TL coexist simultaneously

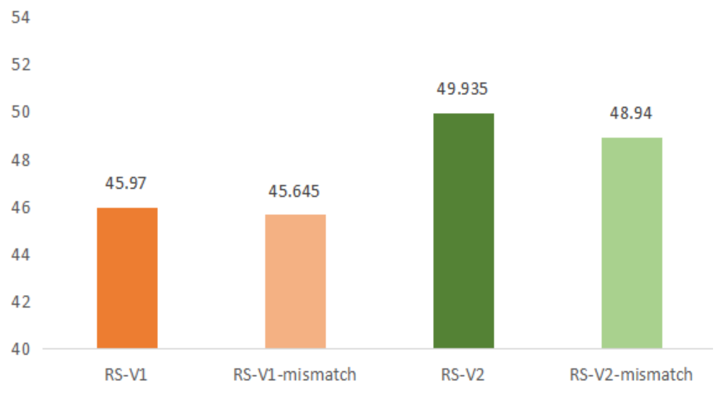


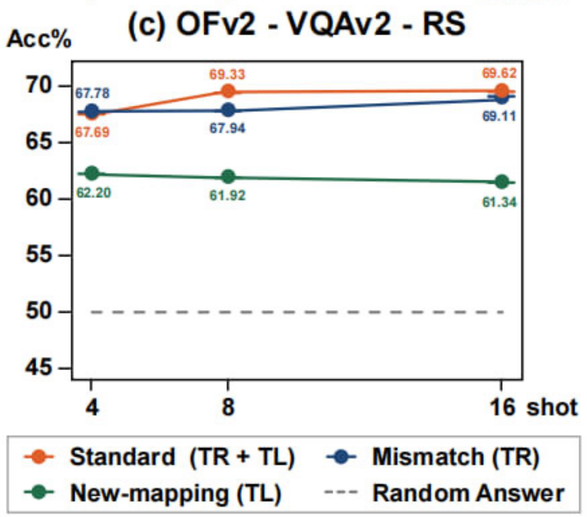
How to Configure Good In-Context Sequence for VQA: Analysis

Three important inner properties of LVLM during ICL

1. Limited TL capabilities







- As the **number** of **shots increases**, the improvement of the **model diminishes**
- **Replacing incorrect answers** in demonstrations did **not** significantly **impact** the model's performance.
- Disentangle TR and TL and find that the accuracy of **TR** is significantly **higher than TL**

How to Configure Good In-Context Sequence for VQA: Analysis

Three important inner properties of LVLM during ICL

2. The presence of a short-cut effect

SQ

Q: What is the design on the sheets?

A: alligators and bears

Q: What is the design of the bed cover?

A: alligators and bears

GT: zebra

SQ

Q: What is the scientific name of this leaf?

A: tulip

Q: What is the scientific name of this leaf?

A: tulip

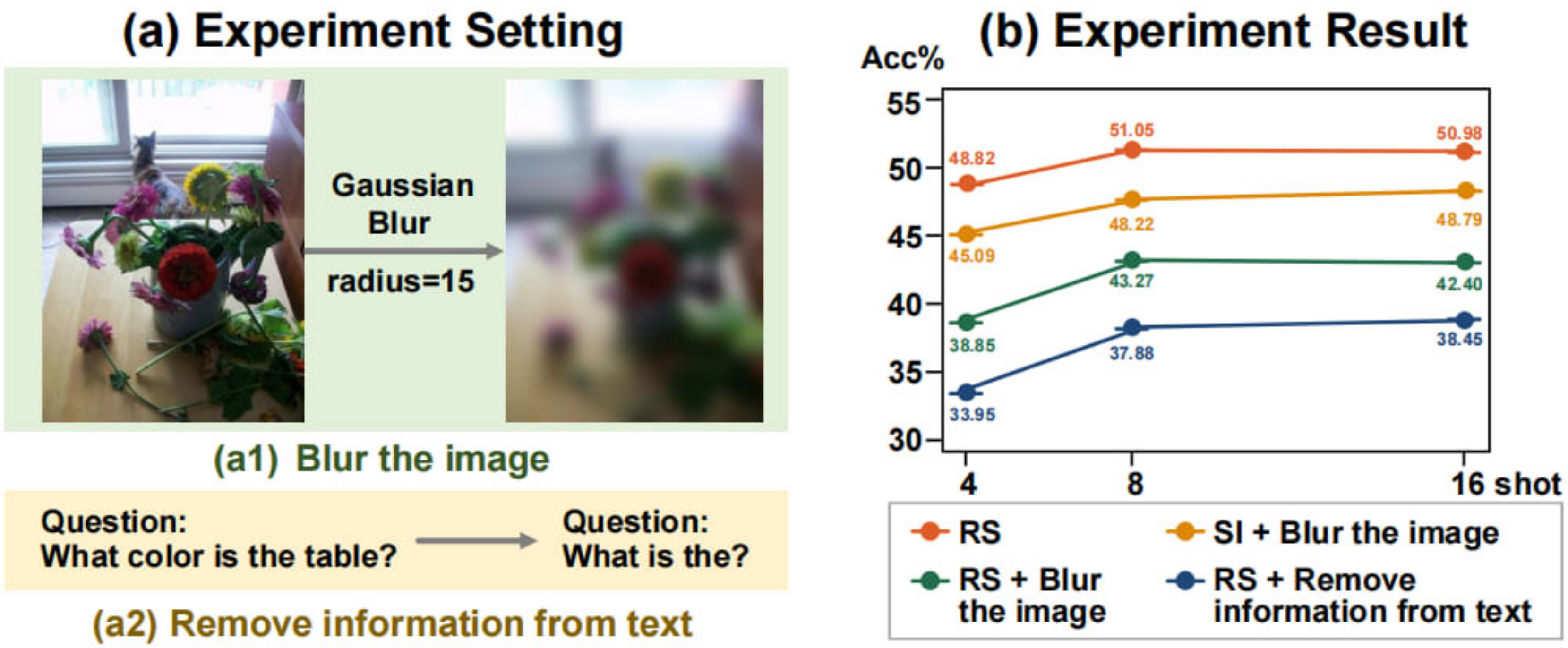
GT: camellia

Copy rate(%)	OFv1	OFv2
RS	43.64	37.34
SI	50.44	54.38
SQ	77.26	79.84
SQA	87.74	89.47
SQA(sole)	47.39	45.82
SQA(sole wrong)	37.07	45.71

How to Configure Good In-Context Sequence for VQA: Analysis

Three important inner properties of LVLM during ICL

3. Partial compatibility between vision and language modules



linguistic TR plays a more substantial role than visual TR

How to Configure Good In-Context Sequence for VQA: Analysis

Three important inner properties of LVLM during ICL

3. Partial compatibility between vision and language modules

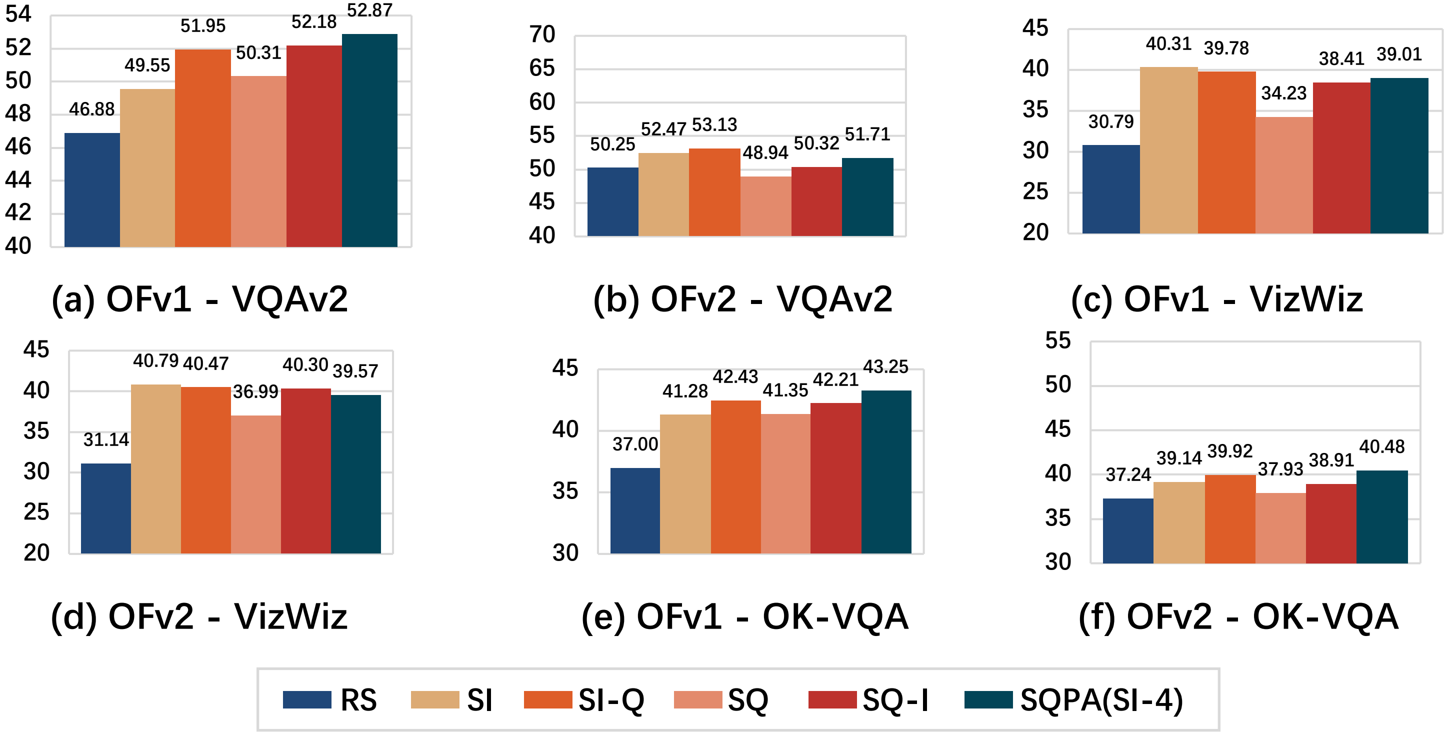
	Dataset	4-shot	8-shot	16-shot
RS(OFv1)	VQAv2	44.56	47.38	48.71
instruct1(OFV1)	VQAv2	43.75	46.91	48.67
RS(OFv2)	VQAv2	48.82	51.05	50.89
instruct1(OFv2)	VQAv2	**49.93**	**52.71**	**50.95**

Some language reasoning ability lose efficacy in the VL case

How to Configure Good In-Context Sequence for VQA: Analysis

Effective Configuration Strategies

- **Similar images and texts lead to better performance**
  - Similar images compensate visual information missed or incorrectly recognized
  - Similar texts brings unstable improvements due to the presence of the short-cut



How to Configure Good In-Context Sequence for VQA: Analysis

Effective Configuration Strategies

- **Instruction enhances the performance of linguistically advanced model**
  - increasing information density in demonstrations
  - do not yield significant improvements in inferior language encoder

Instruct1: According to the previous question and answer pair, answer the final question.

Instruct2: Consider the semantic relationship between the question and the image.

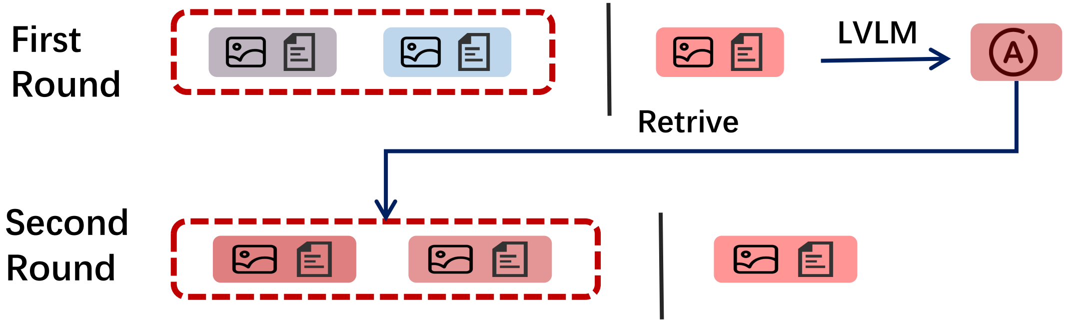
Instruct3: You will be engaged in a two-phase task. Phase 1: Absorb the information from a series of image-text pairs. Phase 2: Use that context, combined with an upcoming image and your own database of knowledge, to accurately answer a subsequent question.

	Dataset	4-shot	8-shot	16-shot
RS	VQAv2	48.82	51.05	50.89
instruct1	VQAv2	**49.93**	**52.71**	**50.95**
RS	OK-VQA	34.82	38.54	39.55
instruct1	OK-VQA	35.72	39.38	40.46
instruct2	OK-VQA	**36.45**	40.17	**41.11**
instruct3	OK-VQA	35.53	**40.19**	40.02

How to Configure Good In-Context Sequence for VQA: Analysis

Effective Configuration Strategies

- Pseudo answers have potential for expeditious enhancement of performance



	Dataset	4-shot
RS	VQAv2	48.82
SQPA(RS-4)	VQAv2	49.85
SI	VQAv2	50.36
SQPA(SI-4)	VQAv2	**50.57**
RS	VizWiz	22.07
SQPA(RS-4)	VizWiz	30.02
SI	VizWiz	36.30
SQPA(SI-4)	VizWiz	**38.37**
RS	OK-VQA	34.82
SQPA(RS-4)	OK-VQA	38.92
SI	OK-VQA	36.46
SQPA(SI-4)	OK-VQA	**39.34**

PART 03

Learning-based configuration strategies

"Take IC, and
VQA as examples"

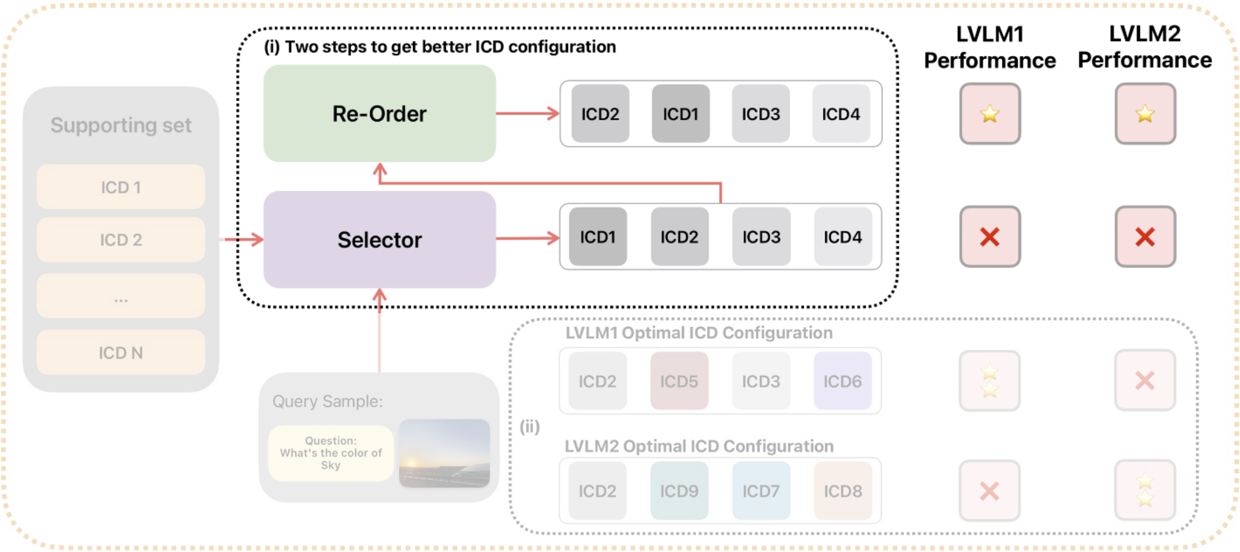
ICD-LM: Configuring Vision-Language In-Context Demonstrations by Language Modeling

Yingzhe Peng, Xu Yang, Haoxuan Ma, Shuo Xu, Chi Zhang, Yucheng Han, Hanwang Zhang

arXiv: https://arxiv.org/abs/2312.10104

code: https://github.com/ForJadeForest/ICD-LM

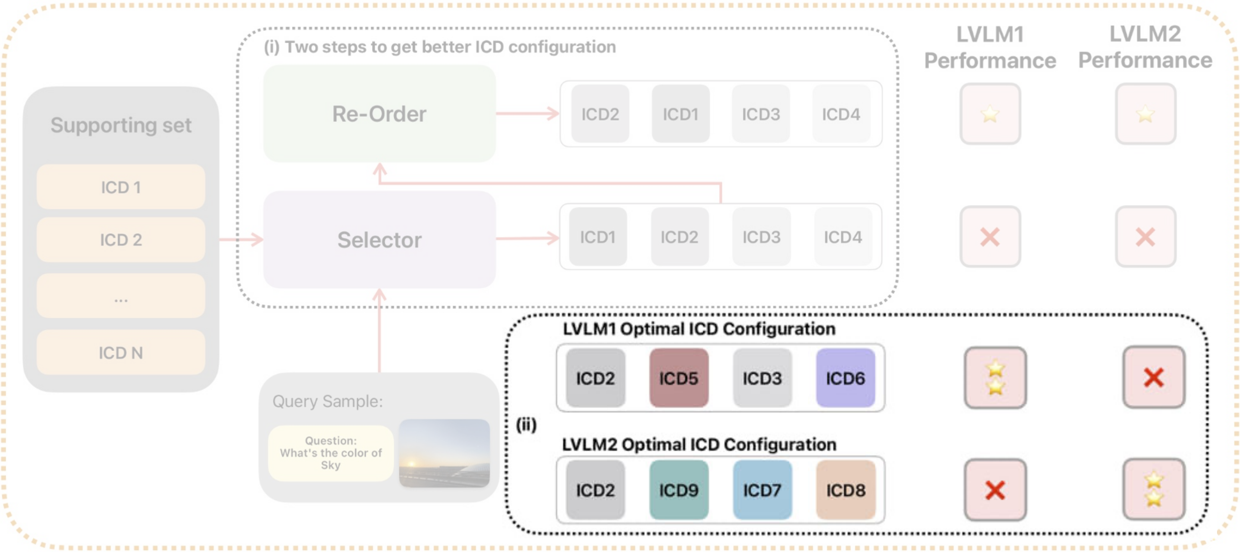
ICD-LM: Traditional Configure ICD Methods



(a) Traditional Two Steps method

- Require selecting and reordering ICD sequences.
- Different LVLMs have different optimal ICD sequence.

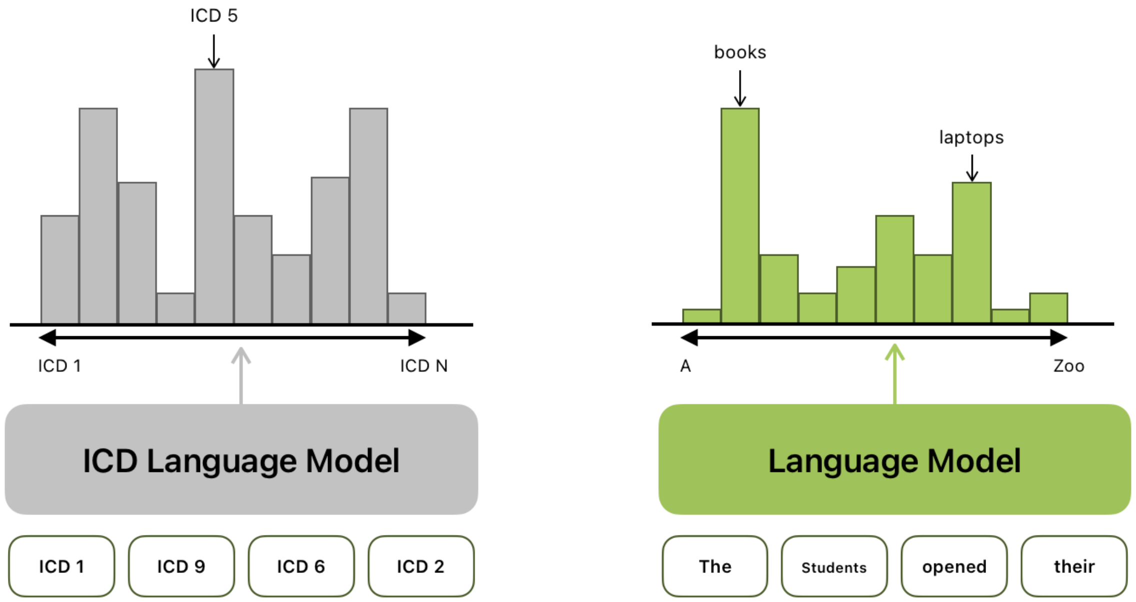
ICD-LM: Traditional Configure ICD Methods



(a) Traditional Two Steps method

- Require selecting and reordering ICD sequences.
- Different LVLMs have different optimal ICD sequence.

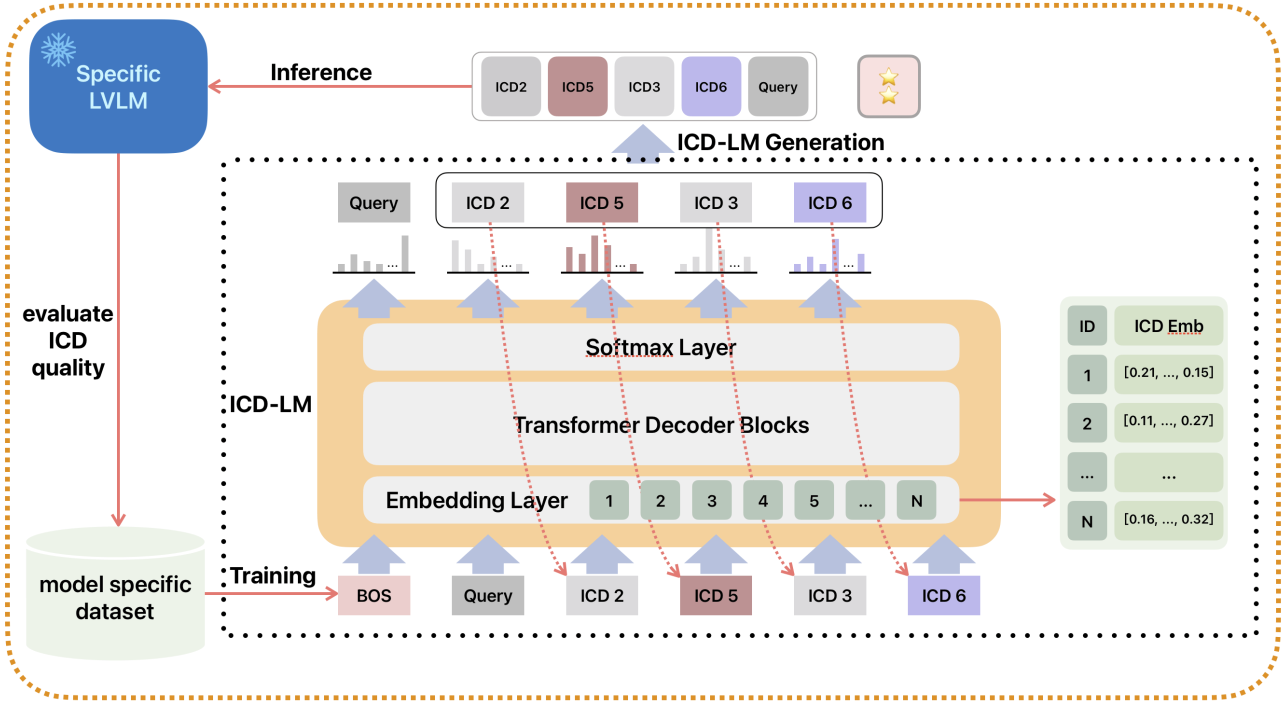
ICD-LM: ICD Language Model



Based on the following observation:

Obtaining an optimal ICD sequence can be likened to sentence generation in a language model.

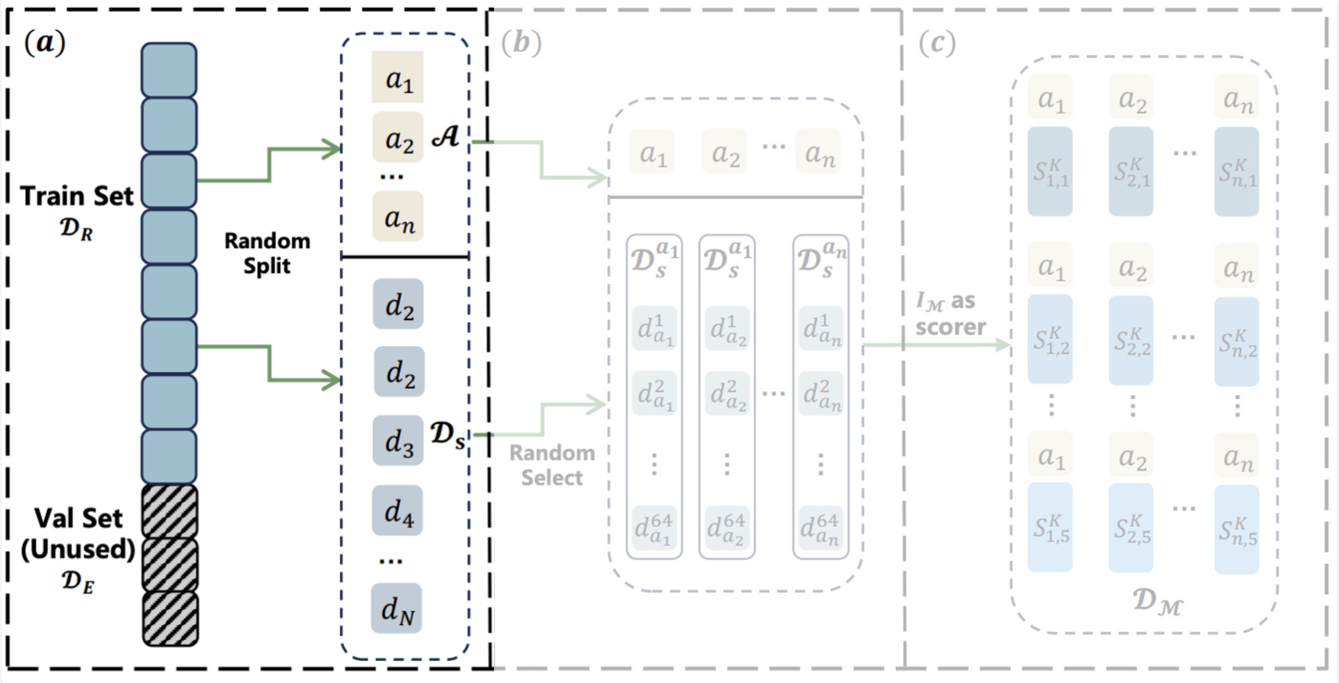
ICD-LM: ICD Language Model



(b) Our ICD-LM

- One selects the most fluent word (ICD) from a vocabulary (ICD set) one by one.
- Using a **language model** enables **learning** to select and arrange optimal ICDs.

ICD-LM: Dataset Construction

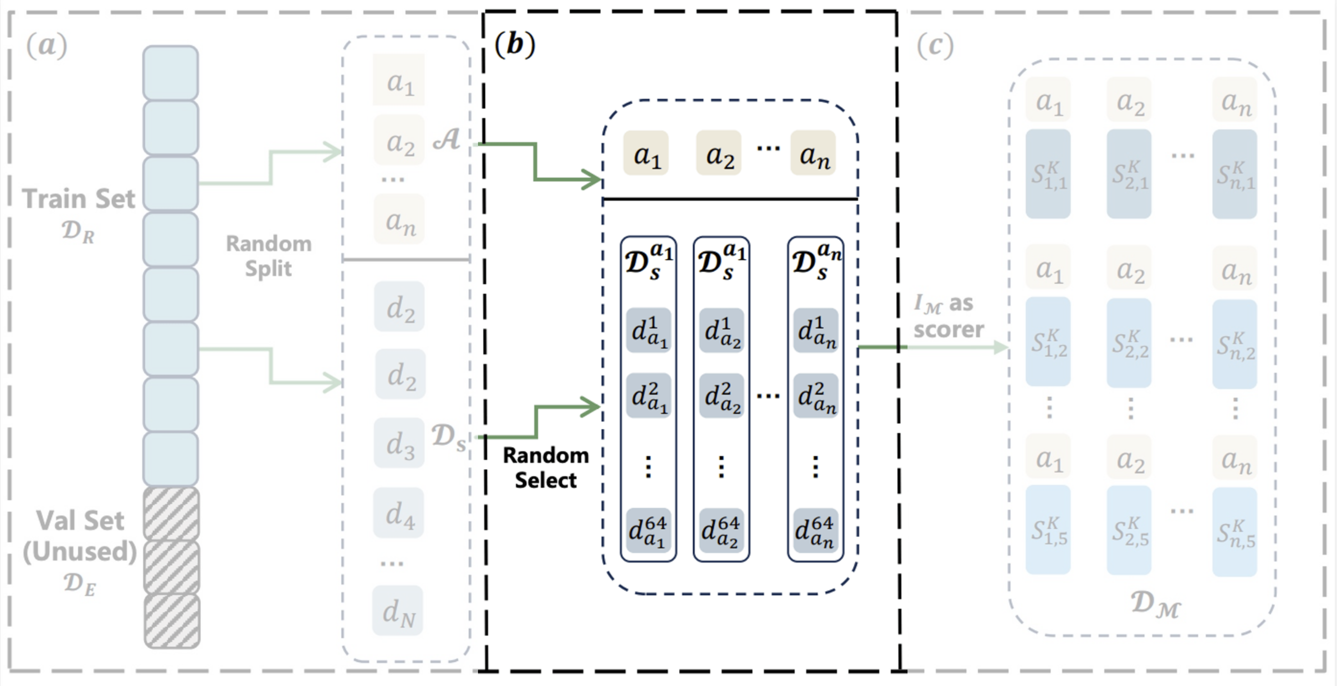


a) **Anchor set selection.**
- Anchor sample simulate a query sample during testing.
- Other train data samples will be used as supporting set.

b) **Sub-Supporting set sampling.**
- To reduce the time complexity.

c) **Use I_M to evaluate the ICD sequence.**
- Obtain the optimal ICD sequence using a greedy algorithm.

ICD-LM: Dataset Construction

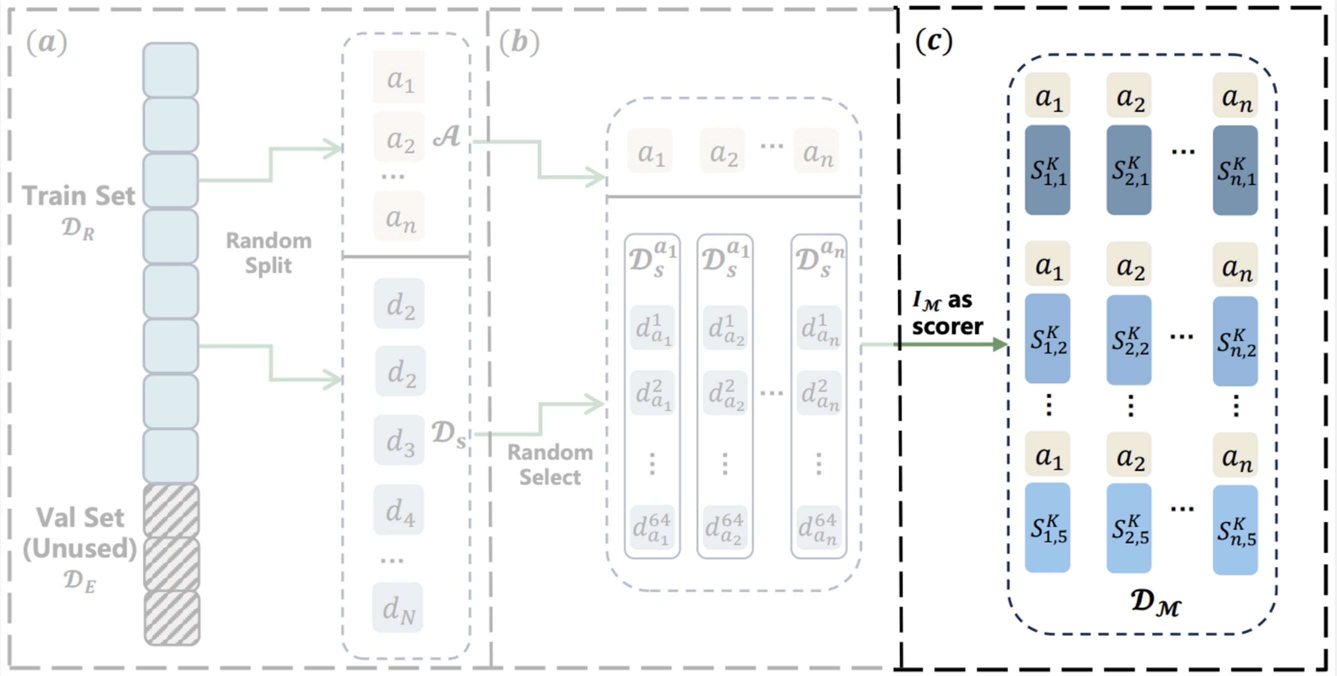


a) Anchor set selection.
- Anchor sample simulate a query sample during testing.
- Other train data samples will be used as supporting set.

b) Sub-Supporting set sampling.
- To reduce the time complexity.

c) Use I_M to evaluate the ICD sequence.
- Obtain the optimal ICD sequence using a greedy algorithm.

ICD-LM: Dataset Construction



a) **Anchor set selection.**
- Anchor sample simulate a query sample during testing.
- Other train data samples will be used as supporting set.

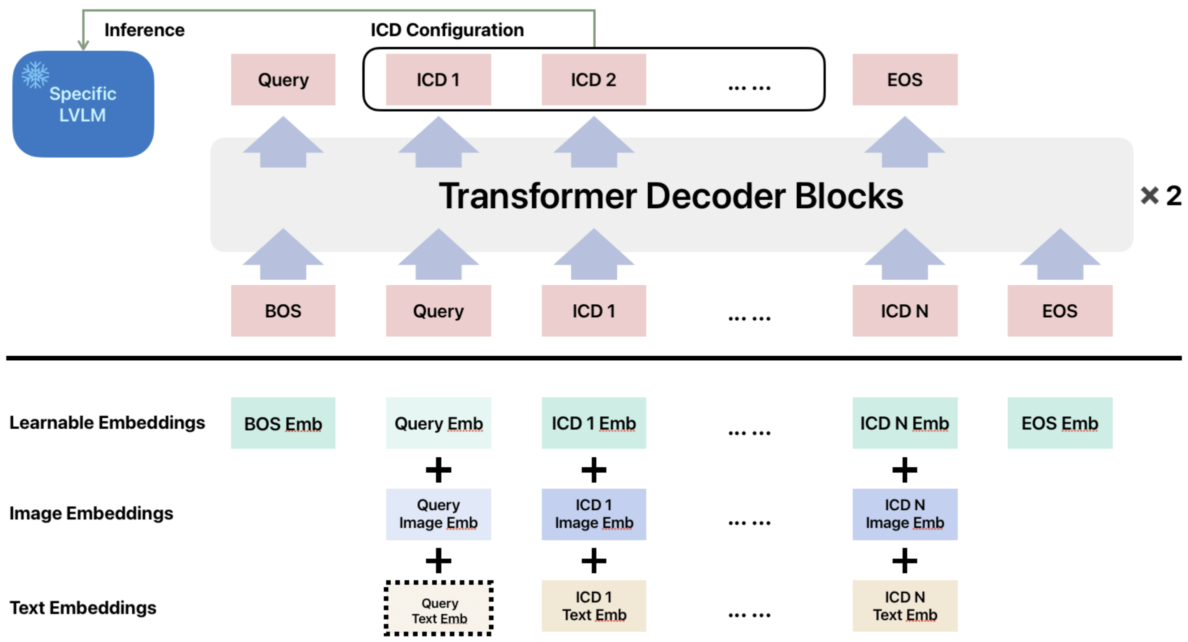
b) **Sub-Supporting set sampling.**
- To reduce the time complexity.

c) **Use I_M to evaluate the ICD sequence.**
- Obtain the optimal ICD sequence using a greedy algorithm.

$$I_{\mathcal{M}}(\mathcal{S}^K, \boldsymbol{a}) = P_{\mathcal{M}}(\boldsymbol{y}|\mathcal{S}^K, \boldsymbol{x}) = \prod_t P_{\mathcal{M}}(y^{(t)}|\mathcal{S}^K, \boldsymbol{x}, y^{(1:t-1)}).$$

$$\hat{\boldsymbol{d}}_k = \arg\max_{\boldsymbol{d}\in\mathcal{D}_S} I_{\mathcal{M}}(\{\boldsymbol{d}, \mathcal{S}^{k-1}\}, \boldsymbol{a}) - I_{\mathcal{M}}(\mathcal{S}^{k-1}, \boldsymbol{a})$$

ICD-LM: Training LM



We use CLIP to extract multimodal features as the embedding of LM.

The final Embedding is sum of:

a) Learnable Embedding: Randomly initialized

b) Image Embedding

c) Text Embedding

ICD-LM: Experiments Setting

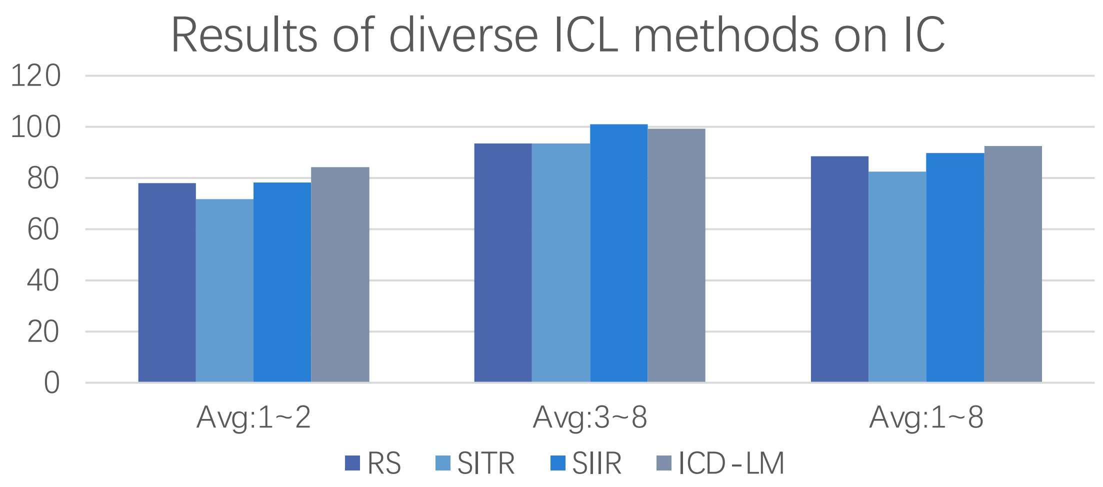
Compared Methods

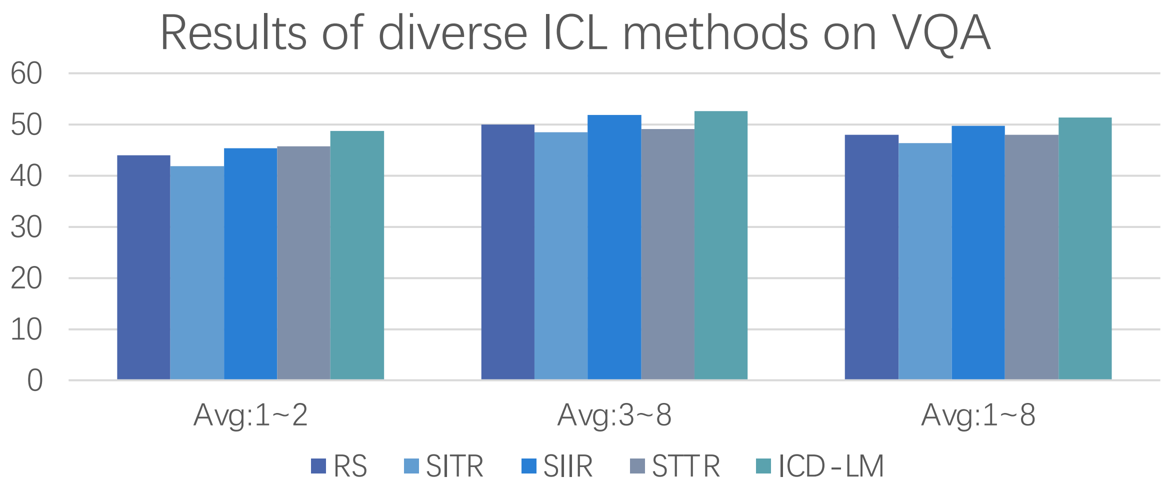
1. Random Sample (**RS**)
2. Similarity-based Retrieval methods:
    1. Similarity-based Image-Image Retrieval (**SIIR**)
    2. Similarity-based Text-Text Retrieval (**STTR**)
    3. Similarity-based Image-Text Retrieval (**SITR**)

ICD-LM: Main Result

We construct 2-shot ICD configurations dataset to train the ICD-LM.

- ICD-LM achieve **the best performance** compared with other methods.
- The trained ICD-LM excels in configuring 4-shot ICDs with **strong length extrapolation ability**.





ICD-LM: Ablation Result: Diverse configuration of dataset construction.

We select **three** factors for our ablation studies:

1. **Beam size b .**

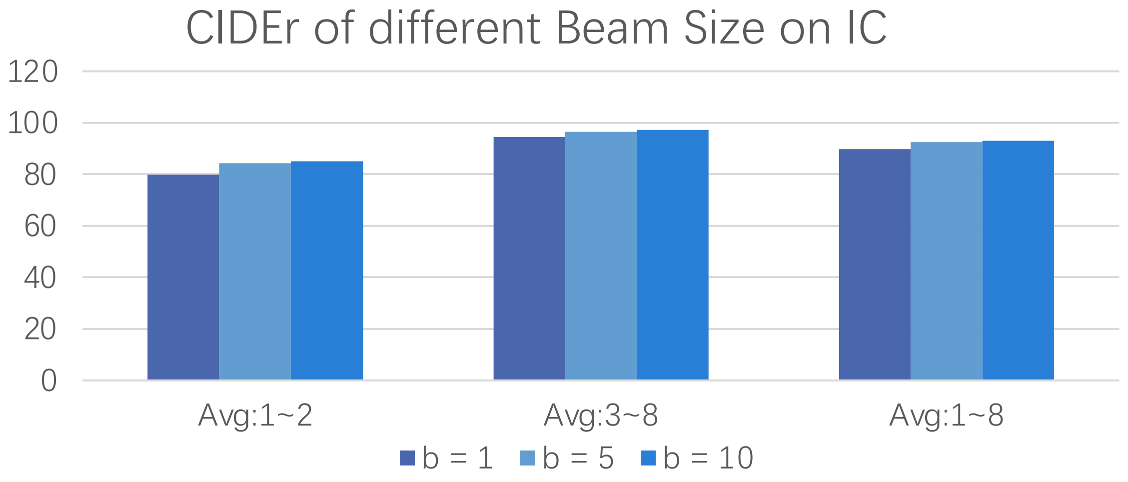
2. **The number n of samples in anchor set.**

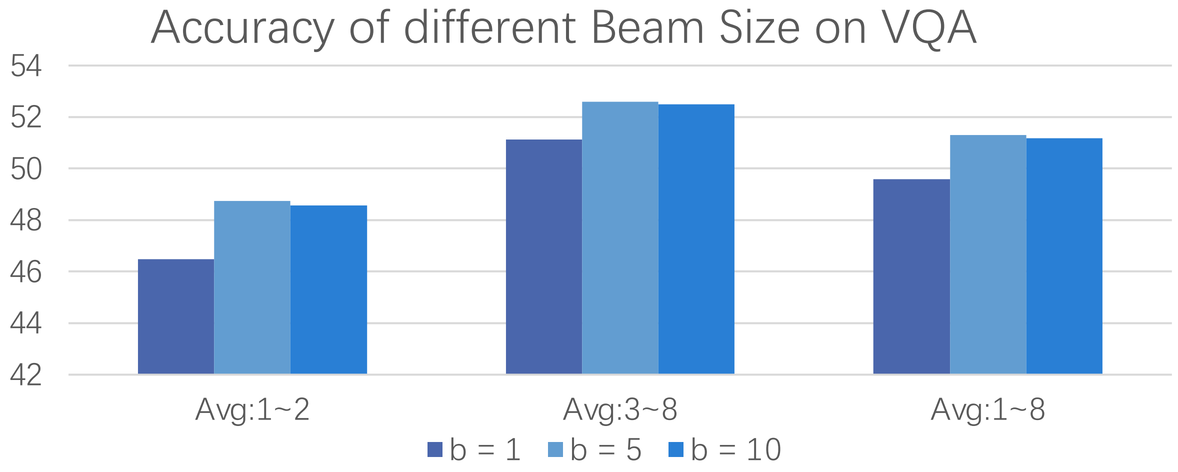
3. **The sampling method of sub-supporting set:**
    - Random: Selecting randomly from total supporting set.
    - Similar Text (Sim-T): Selecting the highest textual similarity sample with anchor sample a from total supporting set.
    - Similar Image (Sim-I): Selecting the highest visual similarity sample with anchor sample a from total supporting set.

ICD-LM: Ablation Result: Diverse configuration of dataset construction.

1. Beam size b .

- Increasing the beam size has a positive correlation with ICD-LM performance.
- An excessively large beam size can negatively impact performance.
    - The performance drop is due to lower-scoring ICD sequences introduced with a large beam size, misleading the ICD-LM during training.

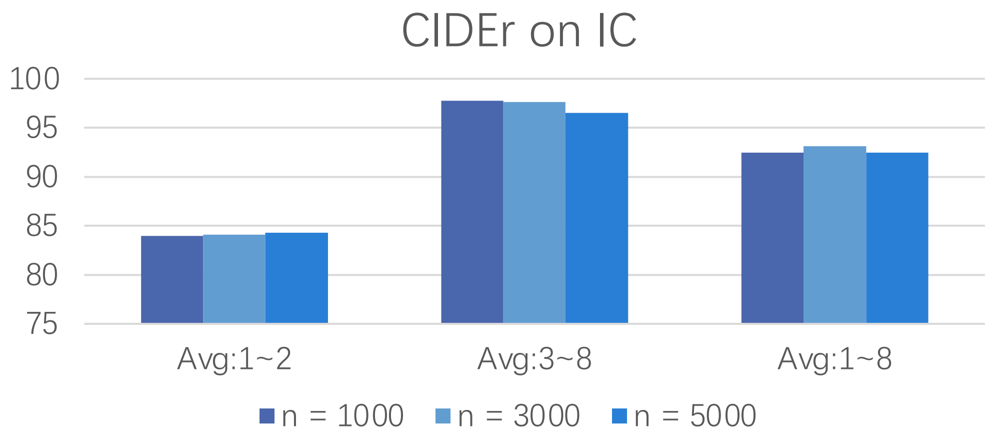


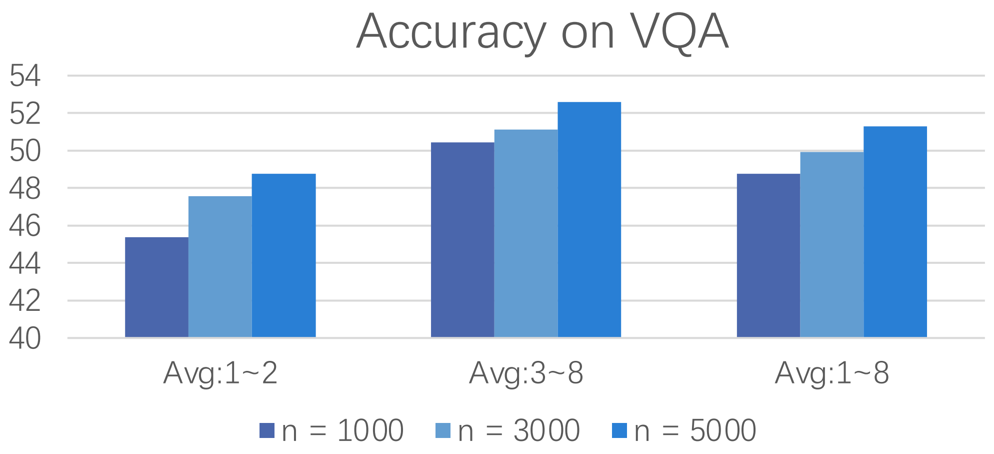


ICD-LM: Ablation Result: Diverse configuration of dataset construction.

2. The number n of samples in anchor set.

- Using more anchor samples can improve the interpolation performance in both IC and VQA
- However, on IC, the extrapolation performance decay when n changes from 3000 to 5000.

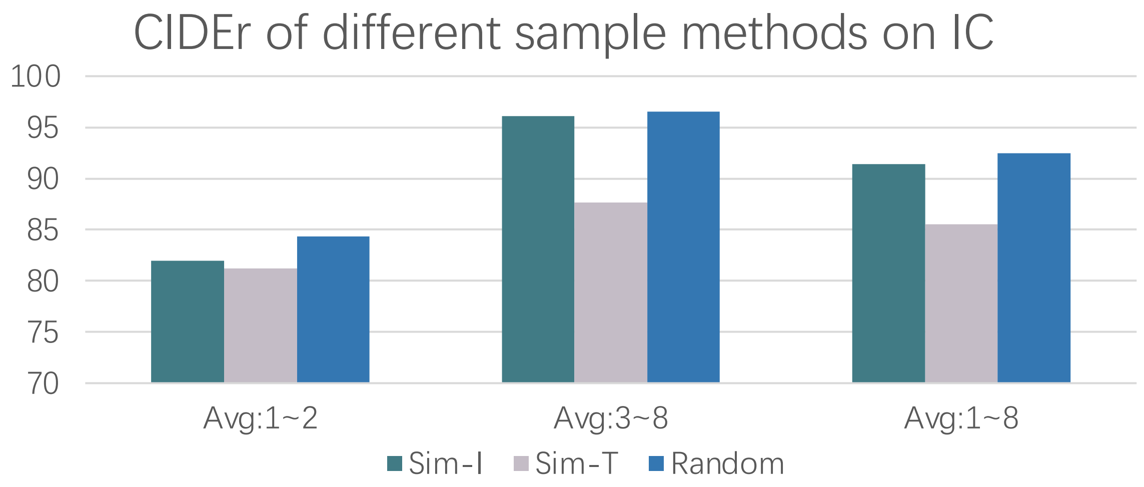


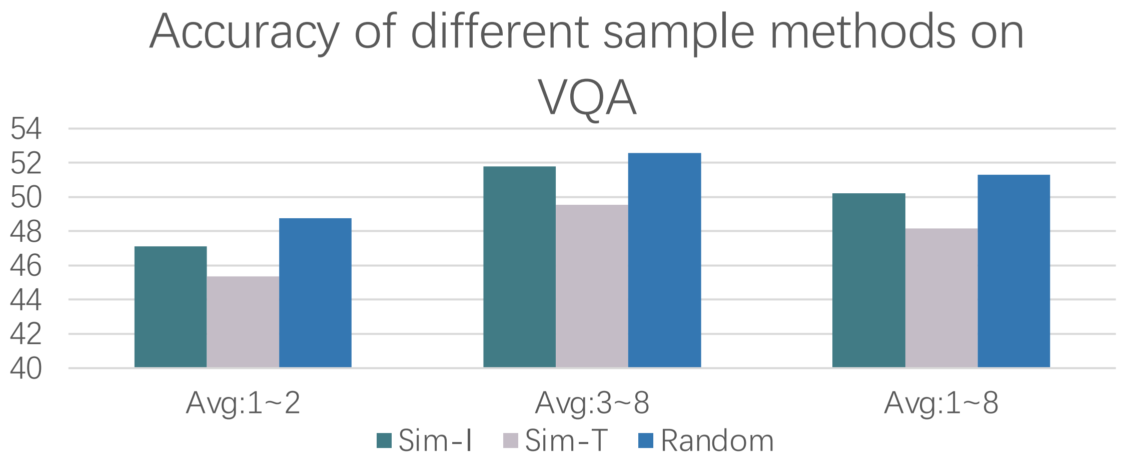


ICD-LM: Ablation Result: Diverse configuration of dataset construction.

3. The sampling method of sub-supporting set.

- We find *Random* is the best in both IC and VQA.
  - We suppose this is because selecting similar ICDs with the anchor sample will damage the diversity of ICD sequence.



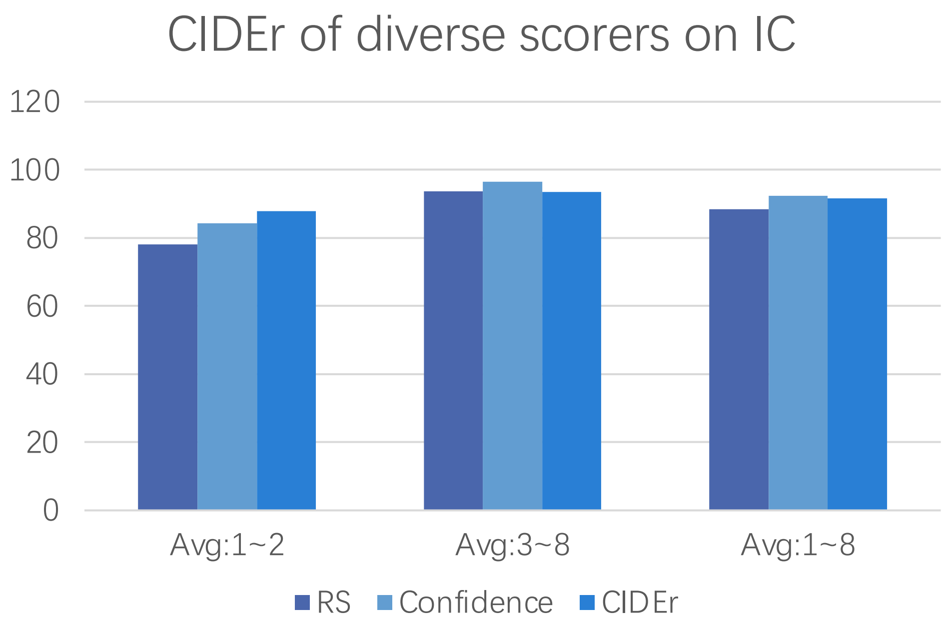


ICD-LM: Ablation Result: Diverse scorers structure

- Using **task-specifical scorers** will increase the interpolation performance.

$$\hat{\boldsymbol{d}}_k = \arg\max_{\boldsymbol{d}\in\mathcal{D}_S} I_{\mathcal{M}}(\{\boldsymbol{d},\mathcal{S}^{k-1}\},\boldsymbol{a}) - I_{\mathcal{M}}(\mathcal{S}^{k-1},\boldsymbol{a})$$

- Accuracy is not suitable for I_M
  - Binary Metric



科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINKing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

THANKS