DATABASE LAB @POSTECH

UniIR: Training and Benchmarking Universal Multimodal Information Retrievers

Cong Wei[†], **Yang Chen**[‡], **Haonan Chen**[†], **Hexiang Hu**[°], **Ge Zhang**[†], **Jie Fu**[§], **Alan Ritter**[‡], **Wenhu Chen**[†]

[†]University of Waterloo [‡]Georgia Institute of Technology [§]Hong Kong University of Science and Technology [°]Google DeepMind

Presenter: Joohyung Yun

2025.3.21

ECCV Milan 2024

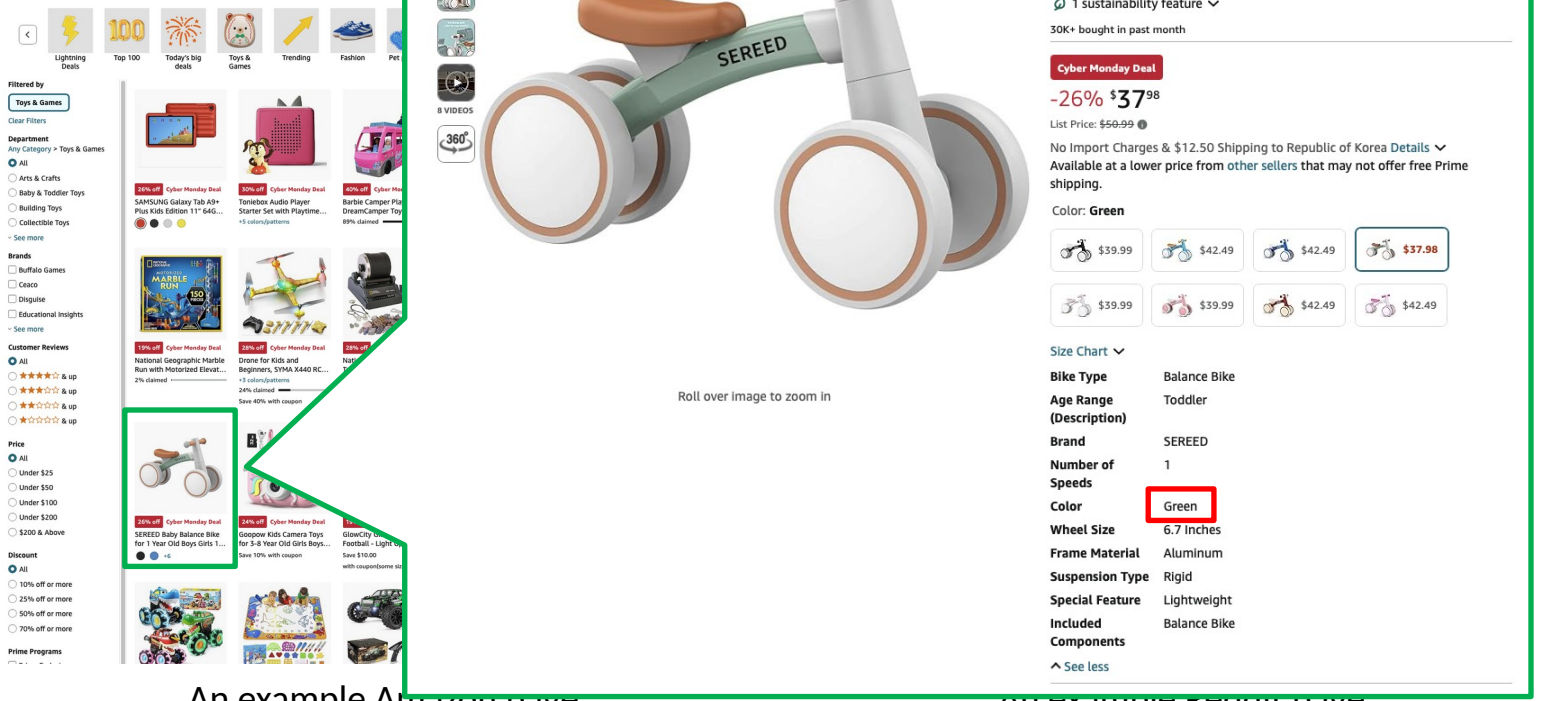
Introduction

Multimodal (Text & Image) Data Source

- Most of real-world web data are comprised of text and images.
- Images can play an irreplaceable role in answering some questions.
  - e.g., “Give me a recommendation of a white baby bike!”
- However, most of prior works do not consider ir



A example Wikipedia page



An example Amazon page

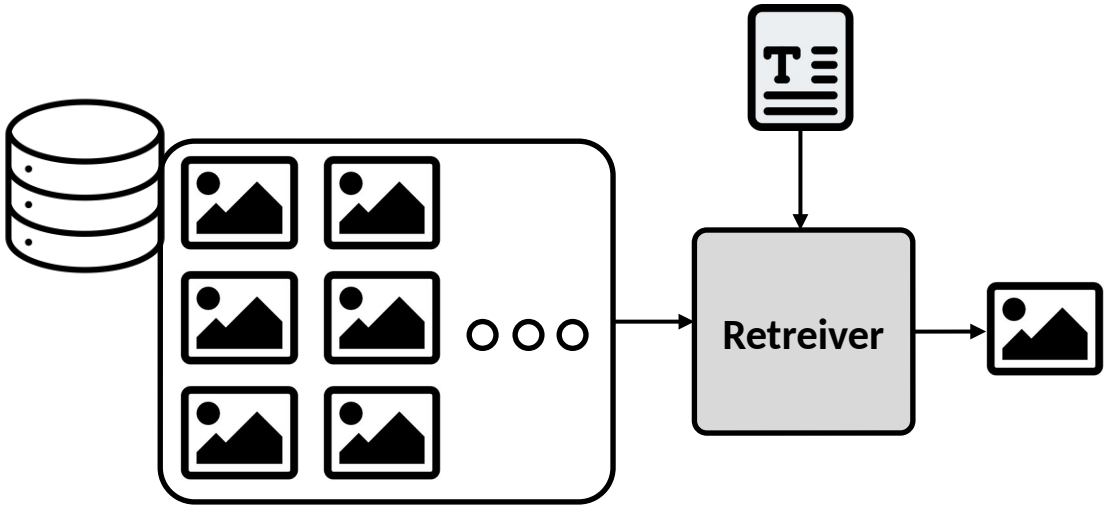
An example Reddit page

Problems of Existing Work

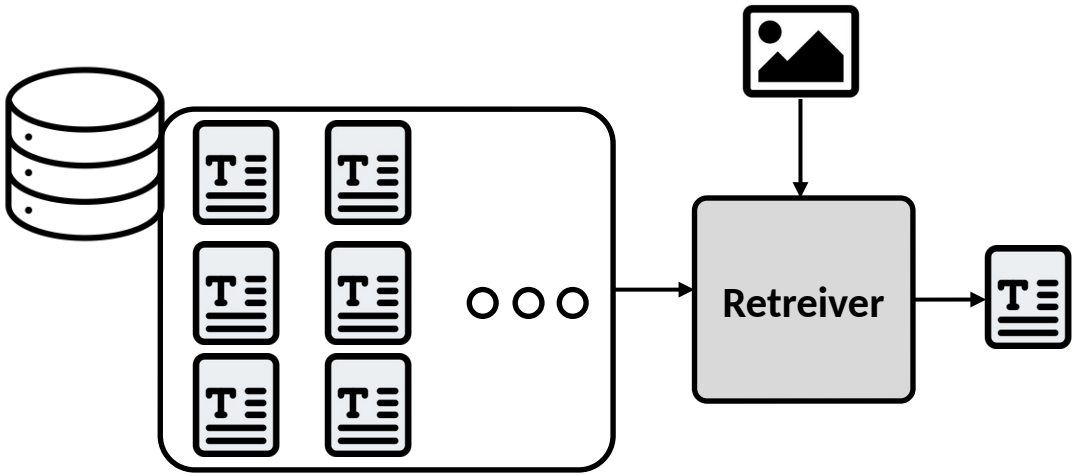
Existing studies focus on multimodal retrieval in limited scenarios.

For example, the MSCOCO benchmark only focuses on 2 kinds of retrieval scenarios.

- Input: caption, output: image
- Input: image, output: caption



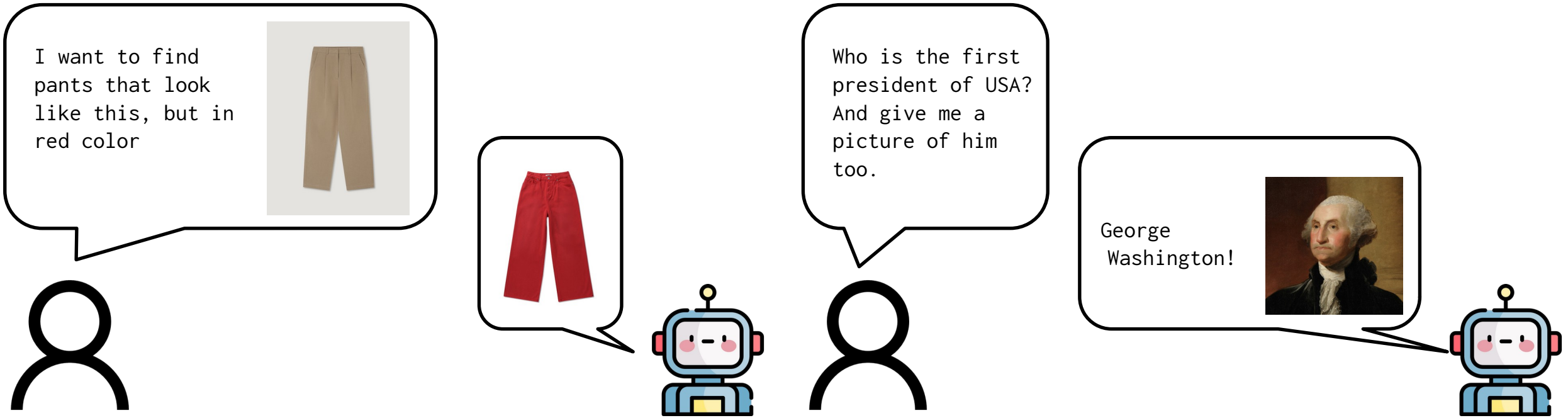
MSCOCO task 1 example



MSCOCO task 2 example

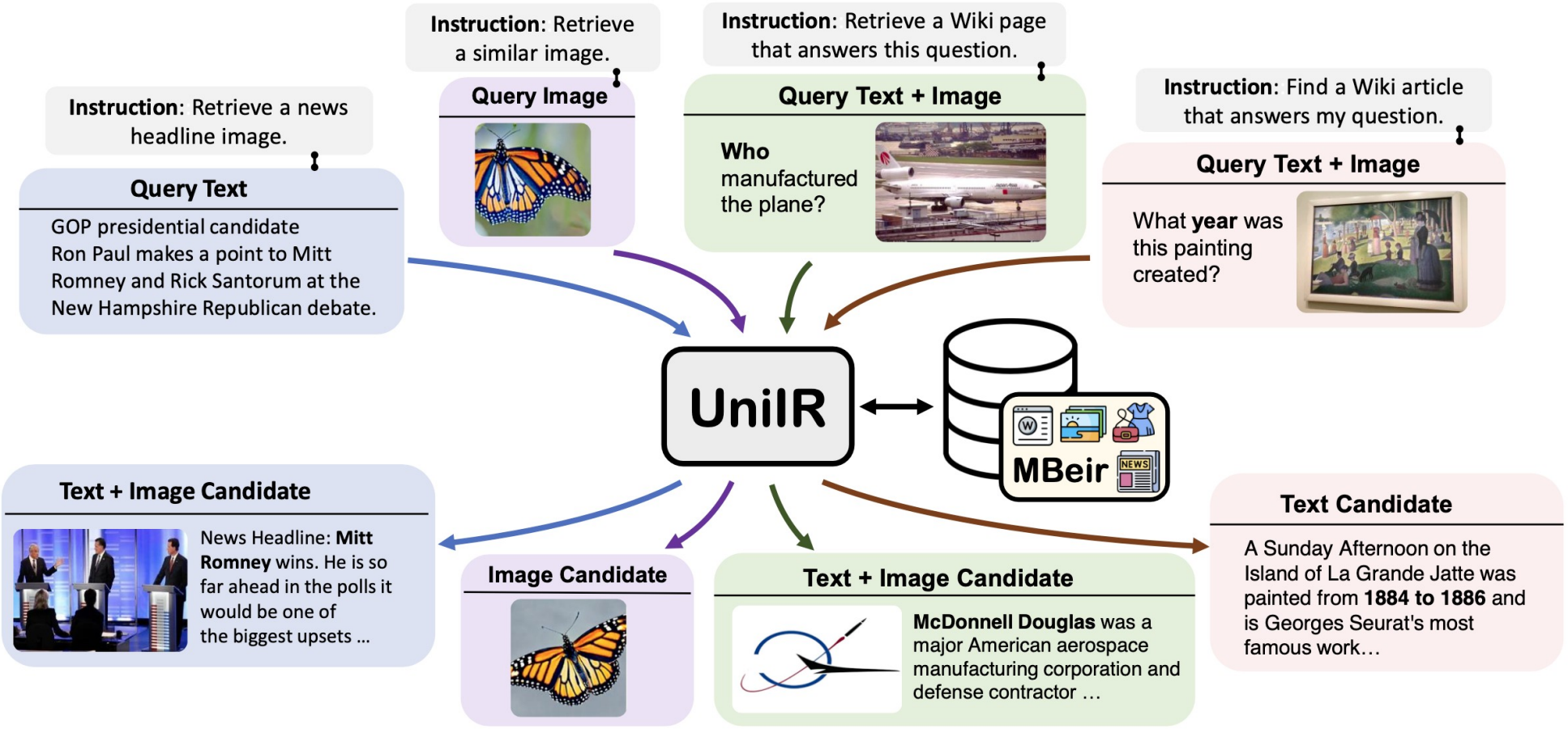
More Diverse Scenarios for Multimodal Retrieval

- Necessary to consider a wider range of domains and modalities in retrieval scenarios.
- A user conducts searches in various ways, including
  - the target (i.e., modality) that users want to search for
  - the methods of performing searching



More Diverse Scenarios for Multimodal Retrieval

This paper aims to introduce a broader range of scenarios (along with a benchmark) for multimodal retrieval and propose a framework for training multimodal retrievers capable of handling these tasks.



Contributions

- **UniIR framework**
  - A universal multimodal information retrieval framework designed to integrate various multi-modal retrieval tasks into a cohesive system.
- **M-BEIR benchmark**
  - A large-scale multimodal retrieval benchmark that assembles 10 diverse datasets from multiple domains, that has 8 distinct multimodal retrieval tasks.
- **UniIR models adopting UniIR framework & trained on M-BEIR benchmark**

Problem Definition

- Given a query $\boldsymbol{q}$ and a data source $\boldsymbol{c} \in C$, correctly retrieve a data $\boldsymbol{c}^*$ conforming to the instruction q_{inst}.
- Query $\boldsymbol{q}$ can be one of
  - A text q_t
  - An image q_i
  - A combination (q_t, q_i)
- A retrieval candidate $\boldsymbol{c}$ also be one of
  - A text c_t
  - An image c_i
  - A combination (c_t, c_i)
- An instruction q_{inst} is a natural language sentence that specifies the inputs and output modalities
  - ex) `Retrieve a Wiki image-description pair that answers this question.`
  - ex) `Find a new headline image that matches the given caption.`

$$\boldsymbol{c}^* = \underset{\{\boldsymbol{c}\} \in C}{\operatorname{argmax}} f(\boldsymbol{q}, q_{inst})^T f(\boldsymbol{c})$$

Problem Definition (Cont.)

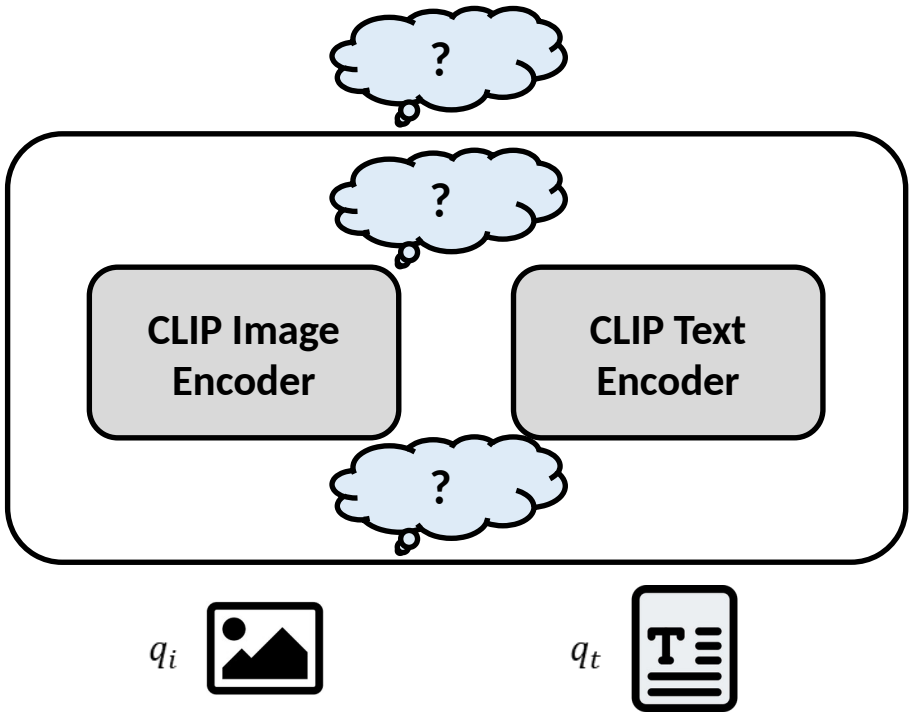
The authors define 8 retrieval scenarios total.

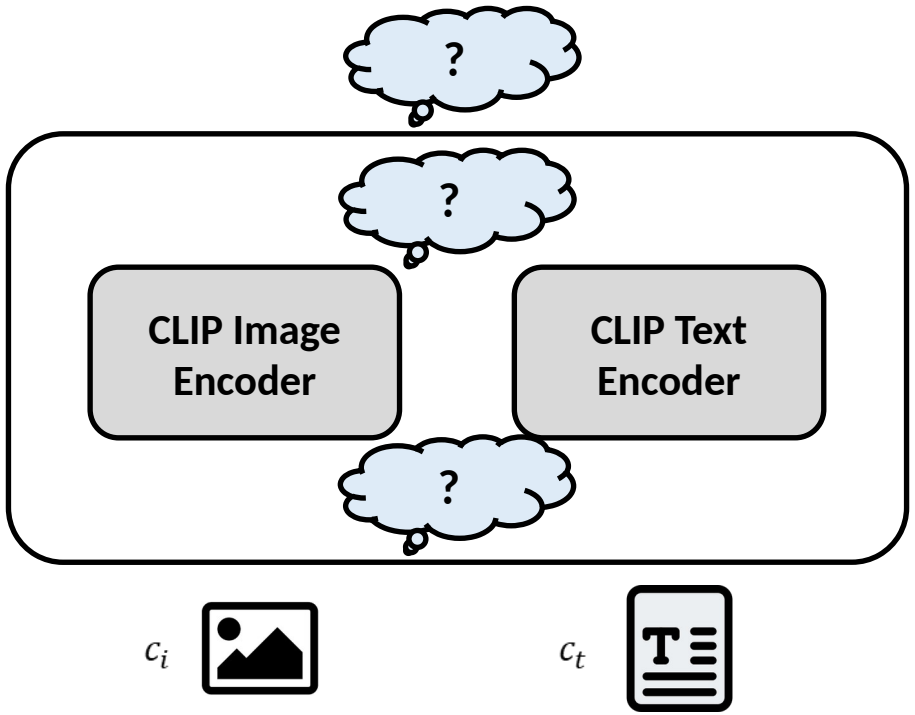
Task (query → candidate)	**Dataset**	**Instruction** (shown 1 out of 4)	**Domain**	**Train**	**Dev**	**Test**	**Pool**
1. $q_t \rightarrow c_i$	VisualNews [37]	Identify news-related image match with the description	News	99K	20K	20K	542K
	MSCOCO [35]	Find an everyday image match with caption	Misc.	100K	24.8K	24.8K	5K
	Fashion200K [18]	Based on fashion description, retrieve matched image	Fashion	15K	1.7K	1.7K	201K
2. $q_t \rightarrow c_t$	WebQA [6]	Find an paragraph from Wikipedia to answer the question	Wiki	16K	1.7K	2.4K	544K
3. $q_t \rightarrow (c_i, c_t)$	EDIS [39]	Find a news image matching with the caption	News	26K	3.2K	3.2K	1M
	WebQA [6]	Find a Wiki image that answer the question	Wiki	17K	1.7K	2.5K	403K
4. $q_i \rightarrow c_t$	VisualNews [37]	Provide a news-related caption for the displayed image	News	100K	20K	20K	537K
	MSCOCO [35]	Find a caption describe the an image	Misc.	113K	5K	5K	25K
	Fashion200K [18]	Find a description for the fashion item in the image	Fashion	15K	4.8K	4.8K	61K
5. $q_i \rightarrow c_i$	NIGHTS [15]	Find an image that is identical to the given image	Misc.	16K	2K	2K	40K
6. $(q_i, q_t) \rightarrow c_t$	OVEN [19]	Retrieve a Wiki text that answer the given query about the image	Wiki	150K	50K	50K	676K
	InfoSeek [10]	Find an article that answers the given question about the image	Wiki	141K	11K	11K	611K
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ [56]	Find an image to match the fashion image and style note	Fashion	16K	2K	6K	74K
	CIRR [40]	I'm looking for a similar everyday image with the described changes	Misc.	26K	2K	4K	21K
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN [19]	Find a Wiki image-text pair to answer a question regarding an image	Wiki	157K	14.7K	14.7K	335K
	InfoSeek [10]	Find a Wiki image-text pair to answers my question about this image	Wiki	143K	17.6K	17.6K	481K
	10 datasets	64 instructions	4 domains	1.1M	182K	190K	5.6M

UniIR &
M-BEIR Benchmark

UniIR Overview

- UniIR aims to design an embedder capable of embedding both images and text simultaneously.
- The UniIR framework utilizes pretrained CLIP and BLIP models.
- It is categorized into two types based on the method of their integration.
  - Score-level Fusion (SF)
  - Feature-level Fusion (FF)





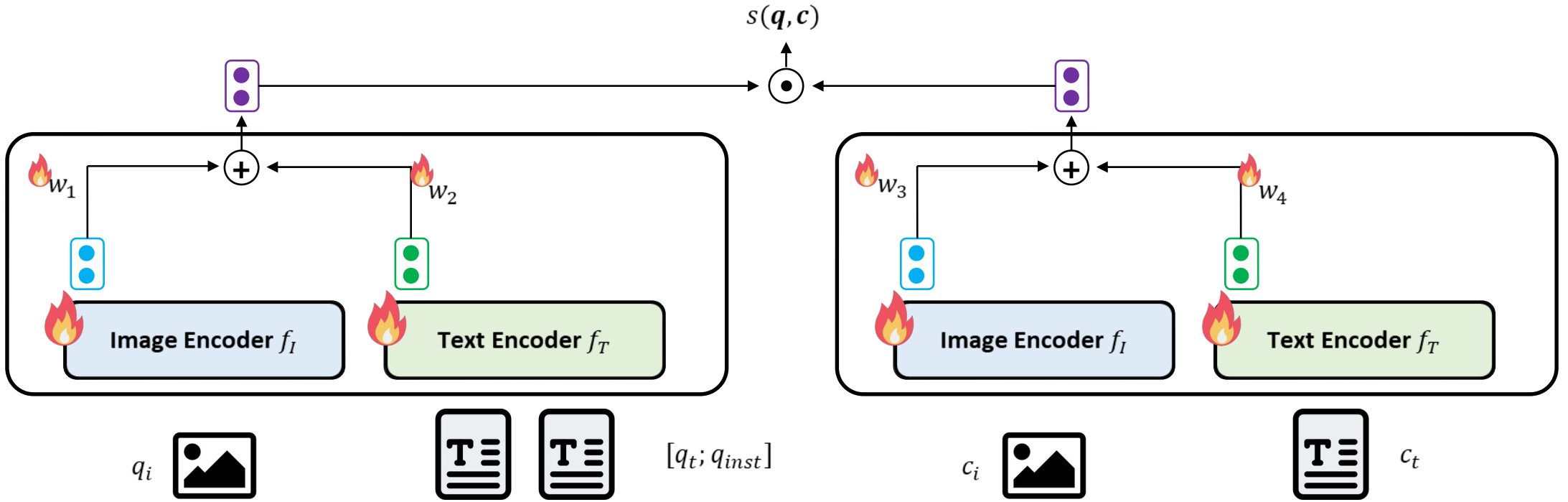
Design Choice 1: Score-level Fusion

1. Image and text inputs are processed into two individual vectors.
2. The two vectors undergo a (learnable) weighted sum to form a single representation vector.

$$f(q_i, q_t, q_{inst}) = w_1 f_I(q_i) + w_2 f_T(q_t, q_{inst})$$

$$f(c_i, c_t) = w_3 f_I(c_i) + w_2 f_T(c_t)$$

$$s(\boldsymbol{q}, \boldsymbol{c}) = f(q_i, q_t, q_{inst})^T f(c_i, c_t)$$



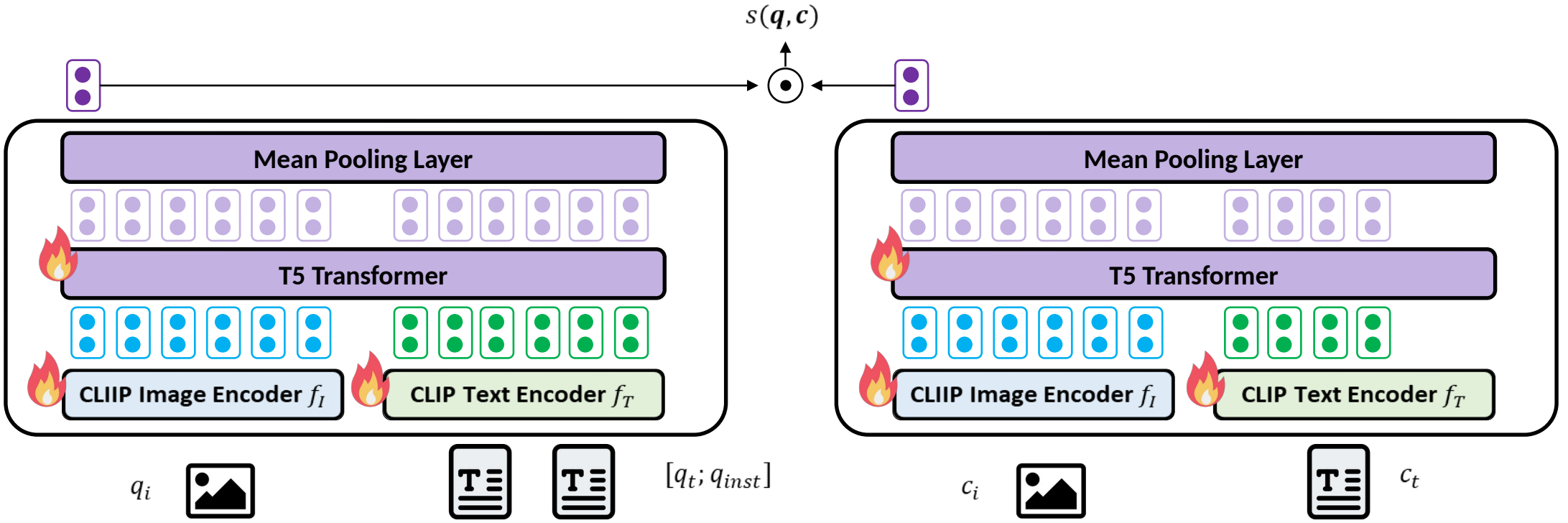
Design Choice 2: Feature-level Fusion

Feature-level fusion employs mixed-modality attention layers to integrate features from both modalities.

- Specific implementation varies between CLIP and BLIP for this fusion approach.

For CLIP

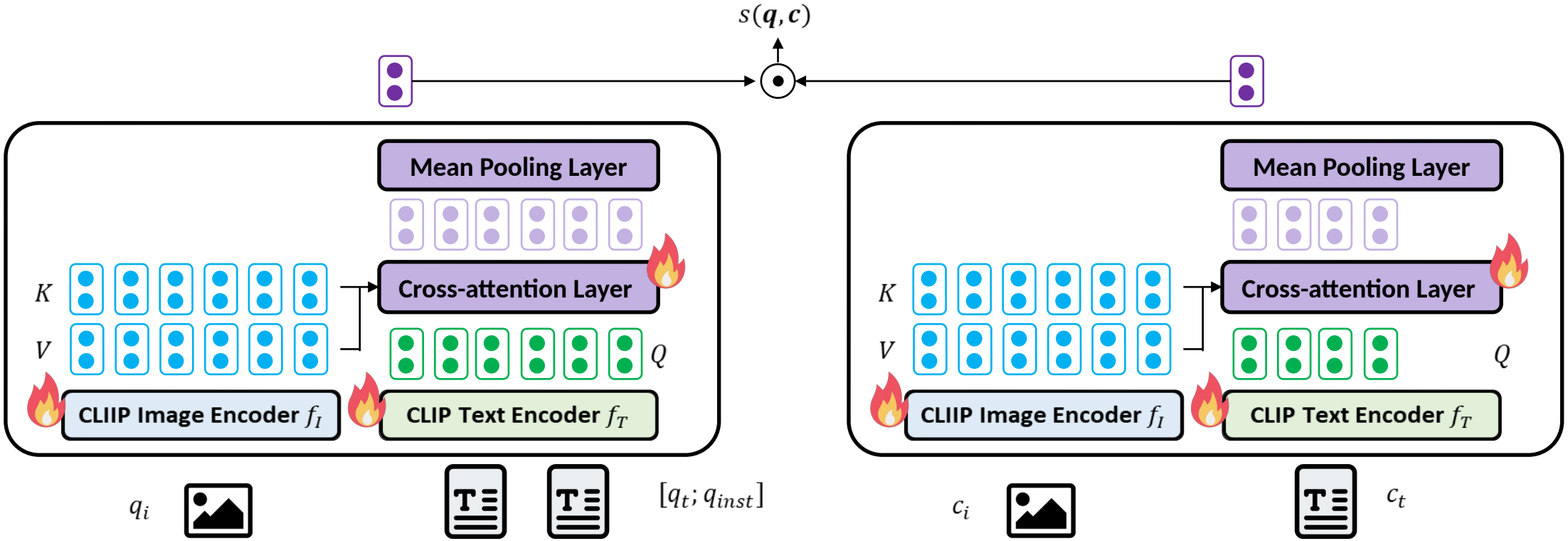
1. Image and text inputs are each processed into a sequence of encoded vectors.
2. Two sequences of vectors are put into a learnable T5 Transformer.
3. Perform mean pooling.



Design Choice 2: Feature-level Fusion

For BLIP

1. Image input is encoded into a sequence of key/value vectors.
2. Text input is encoded into a sequence of query vectors.
3. A cross-attention is performed on the query/key/value vectors.
4. Perform mean pooling.



M-BEIR Overview

A large-scale retrieval benchmark constructed to train and evaluate unified multimodal retrieval models.

- Composed of 10 different datasets, which have 8 different retrieval tasks.
- Four instructions q_{inst} are given along the task.

M-BEIR Components for a single task.

- $Q = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \dots, \boldsymbol{q_n}\}$
- $C = \{\boldsymbol{c_1}, \boldsymbol{c_2}, \dots, \boldsymbol{c_m}\}$
- q_{inst}: human-authored instruction to specify the intent of the retrieval task

For a single QA pair,

- $\boldsymbol{q}$: question
- $\boldsymbol{c}^+$: positive data
- $\{\boldsymbol{c_1^-}, \boldsymbol{c_2^-}, \dots, \boldsymbol{c_k^-}\}$: a set of negative data

M-BEIR: Dataset Collection

The M-BEIR dataset is composed of 8 tasks from 10 distinct benchmarks

Task (query → candidate)	**Dataset**	**Instruction** (shown 1 out of 4)	**Domain**	**Train**	**Dev**	**Test**	**Pool**
1. $q_t \rightarrow c_i$	VisualNews [37]	Identify news-related image match with the description	News	99K	20K	20K	542K
	MSCOCO [35]	Find an everyday image match with caption	Misc.	100K	24.8K	24.8K	5K
	Fashion200K [18]	Based on fashion description, retrieve matched image	Fashion	15K	1.7K	1.7K	201K
2. $q_t \rightarrow c_t$	WebQA [6]	Find an paragraph from Wikipedia to answer the question	Wiki	16K	1.7K	2.4K	544K
3. $q_t \rightarrow (c_i, c_t)$	EDIS [39]	Find a news image matching with the caption	News	26K	3.2K	3.2K	1M
	WebQA [6]	Find a Wiki image that answer the question	Wiki	17K	1.7K	2.5K	403K
4. $q_i \rightarrow c_t$	VisualNews [37]	Provide a news-related caption for the displayed image	News	100K	20K	20K	537K
	MSCOCO [35]	Find a caption describe the an image	Misc.	113K	5K	5K	25K
	Fashion200K [18]	Find a description for the fashion item in the image	Fashion	15K	4.8K	4.8K	61K
5. $q_i \rightarrow c_i$	NIGHTS [15]	Find an image that is identical to the given image	Misc.	16K	2K	2K	40K
6. $(q_i, q_t) \rightarrow c_t$	OVEN [19]	Retrieve a Wiki text that answer the given query about the image	Wiki	150K	50K	50K	676K
	InfoSeek [10]	Find an article that answers the given question about the image	Wiki	141K	11K	11K	611K
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ [56]	Find an image to match the fashion image and style note	Fashion	16K	2K	6K	74K
	CIRR [40]	I'm looking for a similar everyday image with the described changes	Misc.	26K	2K	4K	21K
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN [19]	Find a Wiki image-text pair to answer a question regarding an image	Wiki	157K	14.7K	14.7K	335K
	InfoSeek [10]	Find a Wiki image-text pair to answers my question about this image	Wiki	143K	17.6K	17.6K	481K
	10 datasets	64 instructions	4 domains	1.1M	182K	190K	5.6M

M-BEIR: Instructions

- The authors added four instructions q_{inst} for each retrieval task.
  - q_{inst} is a natural language sentence that describes a multimodal retrieval task by intent, domain, query modality, and target candidate modality.

Two settings for M-BEIR

- **M-BEIR 5.6M**: retrieve from the M-BEIR 5.6 million candidate pool $C_{5.6M}$, which consists of the retrieval corpus from all tasks
- **M-BEIR Local**: retrieve from a task-specific pool provided by the original benchmark

M-BEIR: Examples

Below shows 6 different benchmarks, with 4 different tasks.

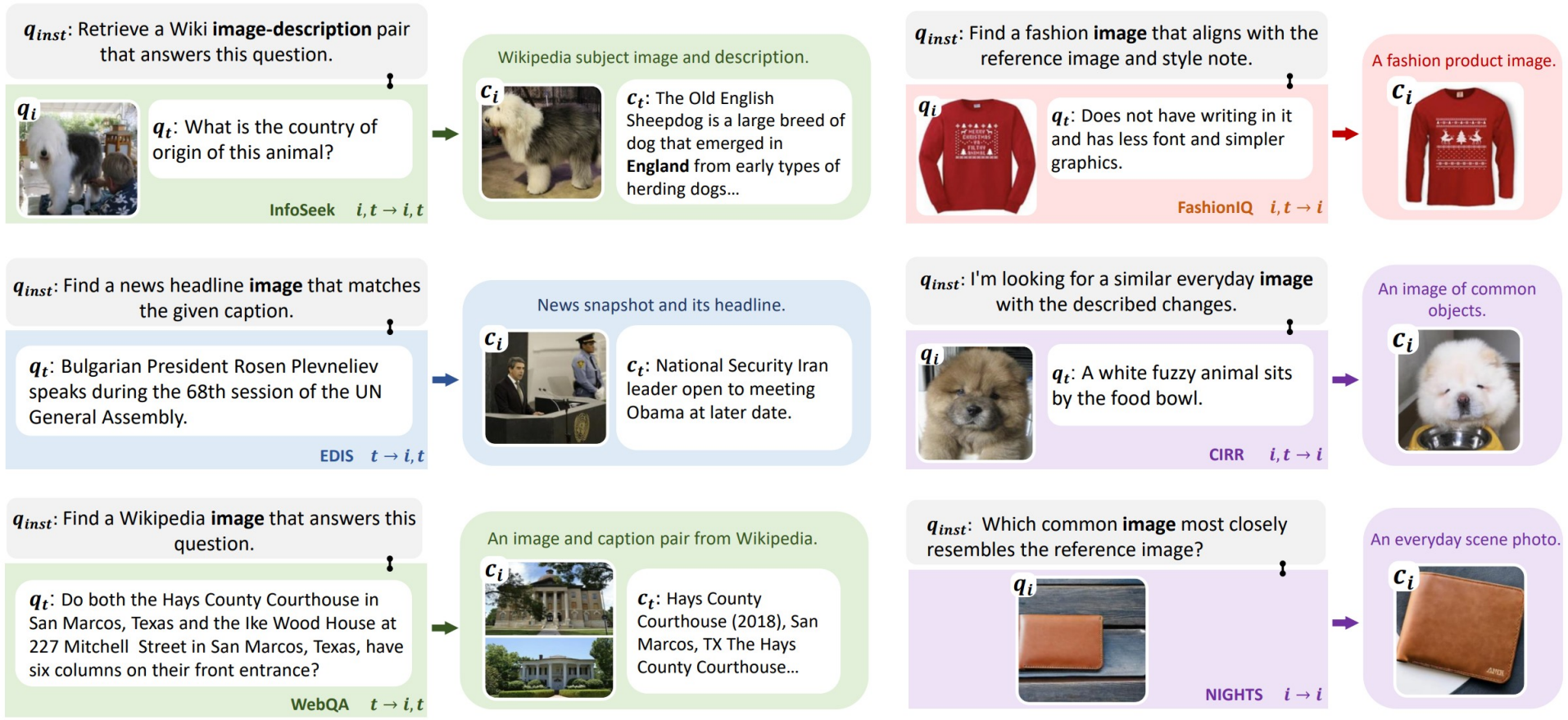


Figure 3. Examples of six query instances in the M-BEIR dataset. Each example query instance includes a query $\mathbf{q}$, a human-annotated natural language instruction q_{inst}, and a positive(relevant) candidate $\mathbf{c}^+$.

Experiments

Experiment Goals

- **UniIR effectivity:** How do UniIR models perform on the M-BEIR benchmark?
- **Instruction validity:** How do the instructions affect the retrieval accuracy, if we fine-tune UniIR using the instructions?
- **Model design choice superiority:** What is the most effective fusion strategy (score-level vs. feature-level) and base model (CLIP vs BLIP) for UniIR?

Experiment Settings

Benchmark:

- M-BEIR 5.6M
- M-BEIR Local

Metrics: adheres to the default metric of each benchmark

- Recall@10: Fashion200K, FashionIQ
- Recall@5: rest of the datasets

Models:

- `Zero-shot`: Vanilla pre-trained CLIP/BLIP
- `Single-task`: UniIR-framework models fine-tuned using M-BEIR Local's corresponding task (excluding the instructions)
- `Multi-task`: UniIR-framework models fine-tuned using M-BEIR 5.6M (excluding the instructions)
- `UniIR`: UniIR-framework models fine-tuned using M-BEIR 5.6M (including the instructions)

Main Experiment: Retrieval Accuracy

- Zero-shot retriever fails to comprehend retrieval intention.
- Instruction-tuning greatly improves retrieval accuracy on M-BEIR.

		Zero-shot	Multi-task (✗ instruction)					UniIR(✓instruction)				
Task	**Dataset**	BLIP2	CLIP_{SF}	CLIP_{FF}	BLIP_{SF}	BLIP_{FF}	$\text{BLIP}_{\text{FF, 384}}$	CLIP_{SF} (Δ)	CLIP_{FF} (Δ)	BLIP_{SF} (Δ)	BLIP_{FF} (Δ)	$\text{BLIP}_{\text{FF, 384}}$ (Δ)
1. $q_t \to c_i$	VisualNews	0.0	12.7	8.8	5.0	8.3	10.5	**42.6** (+29.9)	28.8 (+20.0)	20.9 (+15.8)	23.0 (+14.8)	26.5 (+16.0)
	MSCOCO	0.0	27.3	24.6	22.9	27.7	35.2	**77.9** (+50.6)	74.7 (+50.1)	71.6 (+48.7)	75.6 (+47.8)	75.7 (+40.4)
	Fashion200K	0.0	5.9	5.9	5.7	9.0	13.1	17.8 (+11.9)	15.5 (+9.7)	24.3 (+18.6)	25.4 (+16.4)	**26.7** (+13.6)
2. $q_t \to c_t$	WebQA	35.2	82.3	67.9	74.4	76.1	76.9	**84.7** (+2.5)	78.4 (+10.6)	78.9 (+4.4)	79.5 (+3.4)	79.2 (+2.4)
3. $q_t \to (c_i, c_t)$	EDIS	0.0	41.1	38.3	33.6	36.0	38.5	**59.4** (+18.3)	50.0 (+11.7)	47.2 (+13.6)	50.3 (+14.4)	51.4 (+12.9)
	WebQA	0.0	68.2	62.5	73.2	74.7	75.2	78.8 (+10.6)	75.3 (+12.8)	76.8 (+3.6)	**79.7** (+5.0)	79.4 (+4.2)
4. $q_i \to c_t$	VisualNews	0.0	12.1	8.2	4.8	4.9	6.0	**42.8** (+30.7)	28.6 (+20.4)	19.4 (+14.6)	21.1 (+16.3)	22.9 (+16.9)
	MSCOCO	0.0	84.6	80.8	74.9	76.9	81.4	**92.3** (+7.8)	89.0 (+8.2)	88.2 (+13.4)	88.8 (+11.9)	90.1 (+8.7)
	Fashion200K	0.0	1.2	1.3	2.6	3.6	4.0	17.9 (+16.7)	13.7 (+12.4)	24.3 (+21.7)	27.6 (+24.1)	**28.4** (+24.4)
5. $q_i \to c_i$	NIGHTS	24.0	31.0	30.8	32.9	31.3	32.5	32.0 (+1.0)	31.9 (+1.2)	33.4 (+0.4)	33.0 (+1.6)	**33.7** (+1.3)
6. $(q_i, q_t) \to c_t$	OVEN	0.0	36.8	31.6	33.2	37.7	39.2	39.2 (+2.4)	34.7 (+3.1)	35.2 (+2.0)	38.7 (+1.0)	**40.7** (+1.5)
	InfoSeek	0.0	18.3	15.4	11.9	17.8	17.1	**24.0** (+5.8)	17.5 (+2.1)	16.7 (+4.8)	19.7 (+1.9)	19.2 (+2.0)
7. $(q_i, q_t) \to c_i$	FashionIQ	3.9	22.8	19.7	26.1	28.1	29.0	24.3 (+1.5)	20.5 (+0.9)	26.2 (+0.1)	28.5 (+0.5)	**29.8** (+0.9)
	CIRR	6.2	32.0	32.7	36.7	45.1	47.4	43.9 (+11.9)	40.9 (+8.2)	43.0 (+6.3)	**51.4** (+6.3)	51.1 (+3.8)
8. $(q_i, q_t) \to (c_i, c_t)$	OVEN	13.8	58.7	50.1	51.0	51.6	53.1	**60.2** (+1.5)	55.8 (+5.7)	51.8 (+0.8)	57.8 (+6.2)	59.5 (+6.4)
	InfoSeek	11.4	42.3	31.5	23.0	25.4	25.2	**44.6** (+2.4)	36.8 (+5.3)	25.4 (+2.5)	27.7 (+2.3)	31.1 (+5.9)
	Average	5.9	36.1	31.9	32.0	34.6	36.5	**48.9** (+12.8)	43.3 (+11.4)	42.7 (+10.7)	45.5 (+10.9)	46.6 (+10.1)

Table 2. Benchmarking universal information retrieval on M-BEIR (5.6M candidates) with Recall@5 (except using Recall@10 for Fashion200K, FashionIQ). Δ: UniIR - Multi-task. 384 refers to image resolution. **Bold**: top-1 performance. Underline: top-2.

Validity of Instructions

conducted an analysis of the retrieval error for `Multi-task` and `UniIR`

Three kinds of errors

- `X modality`: Retrieving from modality differing to the target modality
- `X domain`: Retrieving from incorrect domain (i.e., other benchmark's data)
- Other

`X modality` errors decrease greatly from `Multi-task` to `UniIR`

- UniIR models were able to successfully learn to retrieve intended modalities

	Multi-task		UniIR	
Error Types	CLIP_{SF}	BLIP_{FF}	CLIP_{SF}	BLIP_{FF}
✗ modality	58.8%	50.9%	2.7%	15.2%
✗ domain	0.3%	0.5%	0.1%	0.0%
Other	40.9%	48.6%	97.2%	84.8%

Table 4. Error analysis on M-BEIR.

Model design choice superiority

1. Fine-tuned UniIR framework models outperform vanilla pre-trained models
2. For CLIP, Score-level Fusion design outperforms its counterpart.
3. For BLIP, Feature-level Fusion design outperforms its counterpart.
4. Overall, $CLIP_{SF}$ shows the highest retrieval accuracy.

- Reasoning or in-depth analysis for points 2, 3, 4 were not written in the paper.

	Zero-shot		UniIR			
Task	CLIP	BLIP	$CLIP_{FF}$	$CLIP_{SF}$ (Δ)	$BLIP_{SF}$	$BLIP_{FF}$ (Δ)
1.	0.0	0.0	39.7	46.1 (+6.4)	38.9	41.3 (+2.4)
2.	32.1	38.1	78.4	84.7 (+6.3)	78.9	79.5 (+0.7)
3.	6.1	0.0	62.7	69.1 (+6.4)	62.0	65.0 (+3.0)
4.	0.0	0.0	43.8	51.0 (+7.3)	44.0	45.9 (+1.9)
5.	25.3	25.1	31.9	32.0 (+0.1)	33.4	33.0 (–0.4)
6.	0.0	0.0	26.1	31.6 (+5.5)	26.0	29.2 (+3.2)
7.	4.9	4.8	30.7	34.1 (+3.4)	34.6	40.0 (+5.4)
8.	23.3	9.0	46.3	52.4 (+6.1)	38.6	42.7 (+4.1)
Avg	7.9	5.7	43.3	48.9 (+5.7)	42.7	45.5 (+2.8)

Table 5. **Experiments of multimodal feature fusion architecture design**. Retrieve from the M-BEIR (Recall@5). Δ: difference between two model architectures.

Comparison with Existing Methods

UniIR models outperform SOTA models for each task in average.

		SoTA Zero-Shot				ST	MT	UniIR	ST	MT	UniIR
Task	**Dataset**	CLIP	SigLIP	BLIP	BLIP2	CLIP_{SF}	CLIP_{SF}	CLIP_{SF} (Δ_s)	BLIP_{FF}	BLIP_{FF}	BLIP_{FF} (Δ_s)
1. $q_t \rightarrow c_i$	VisualNews	43.3	30.1	16.4	16.7	43.5	40.6	42.6 (–0.9)	20.0	22.8	23.4 (+3.4)
	MSCOCO	61.1	75.7	74.4	63.8	80.4	79.9	81.1 (+0.7)	77.3	78.3	79.7 (+2.3)
	Fashion200K	6.6	36.5	15.9	14.0	10.7	16.8	18.0 (+7.4)	17.1	25.8	26.1 (+9.0)
2. $q_t \rightarrow c_t$	WebQA	36.2	39.8	44.9	38.6	81.7	83.7	84.7 (+3.1)	67.5	77.9	80.0 (+12.5)
3. $q_t \rightarrow (c_i, c_t)$	EDIS	43.3	27.0	26.8	26.9	58.8	57.4	59.4 (+0.6)	38.2	51.2	50.9 (+12.7)
	WebQA	45.1	43.5	20.3	24.5	76.3	76.7	78.7 (+2.5)	67.8	79.2	79.8 (+11.9)
4. $q_i \rightarrow c_t$	VisualNews	41.3	30.8	17.2	15.0	42.7	40.0	43.1 (+0.4)	22.4	20.9	22.8 (+0.3)
	MSCOCO	79.0	88.2	83.2	80.0	89.8	90.3	92.3 (+2.6)	86.0	85.8	89.9 (+3.9)
	Fashion200K	7.7	34.2	19.9	14.2	12.0	18.4	18.3 (+6.3)	15.6	27.4	28.9 (+13.3)
5. $q_i \rightarrow c_t$	NIGHTS	26.1	28.9	27.4	25.4	33.5	31.1	32.0 (–1.5)	30.4	31.5	33.0 (+2.6)
6. $(q_i, q_t) \rightarrow c_t$	OVEN	24.2	29.7	16.1	12.2	45.4	46.6	45.5 (+0.1)	33.8	42.8	41.0 (+7.2)
	InfoSeek	20.5	25.1	10.2	5.5	23.5	28.3	27.9 (+4.4)	18.5	23.9	22.4 (+3.9)
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ	7.0	14.4	2.3	4.4	25.9	23.2	24.4 (–1.5)	3.0	28.4	29.2 (+26.2)
	CIRR	13.2	22.7	10.6	11.8	52.0	38.7	44.6 (–7.3)	13.9	48.6	52.2 (+38.2)
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN	38.8	41.7	27.4	27.3	66.2	69.0	67.6 (+1.4)	49.9	56.3	55.8 (+5.9)
	InfoSeek	26.4	27.4	16.6	15.8	47.4	49.2	48.9 (+1.5)	32.3	32.9	33.0 (+0.7)
-	Average	32.5	37.2	26.8	24.8	49.4	49.4	50.6 (+1.2)	37.1	45.8	46.8 (+9.7)

Table 6. Multi-task (MT) instruction-tuning experiments on $\text{M-BEIR}_{\text{local}}$. We report Recall@5 results of zero-shot retrieval, single-task (ST) fine-tuning, and UniIR (with or without instructions) on $\text{M-BEIR}_{\text{local}}$ except for Fashion200K and FashionIQ where we report Recall@10. Retrieval is conducted from $\text{M-BEIR}_{\text{local}}$ (single) candidate pools. Δ_s: absolute difference to single task fine-tuning.