

SUMMARY

ACORN: Performant and Predicate-Agnostic Search Over Vector Embeddings and Structured Data

LIANA PATEL, Stanford University, USA

PETER KRAFT, DBOS, Inc., USA

CARLOS GUESTRIN, Stanford University, USA

MATEI ZAHARIA, UC Berkeley, USA



Presenter: Joohyung Yun

2024. 12. 11

Motivation: Hybrid Query

Approximate nearest neighbor search with predicates

Application

- E-commerce site: Search for t-shirts images similar to a reference image, filtering on price
- Literature search engine: Search for papers similar to natural language queries, filtering on publication date

Problems of Then-existing Techniques

Pre-filtering

1. Finds all records in the dataset X that pass the query predicate p , Xp
2. Linearly scans Xp , computing distances over each point
- **Limitations:** Cannot use the pre-built graph index structure, and thus we have to perform a linear search over the filtered candidates

Post-filtering

1. Performs ANN Search over X
2. Expand the search scope to find K vectors that pass the query predicate p
- **Limitations:** Have to build a candidate of $K_1 (\gg K)$ vectors to prepare for filtering that will be performed afterwards.

ACORN Overview: An Extended Pre-filtering Approach

- Construct a denser HNSW index by expanding and compressing neighbor lists to handle predicates efficiently.
- Traverse the predicate-specific subgraph using a filter method to ensure only neighbors satisfying the query predicate are considered.
- Incorporate a heuristic for expanding search neighborhoods dynamically during execution to optimize hybrid search performance.

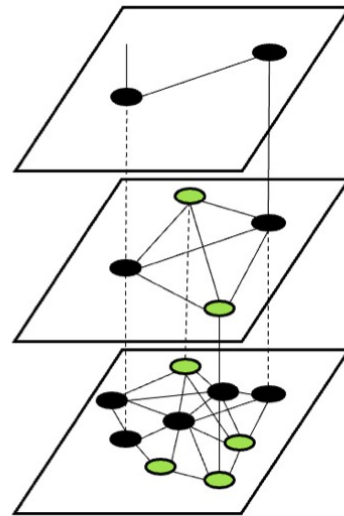
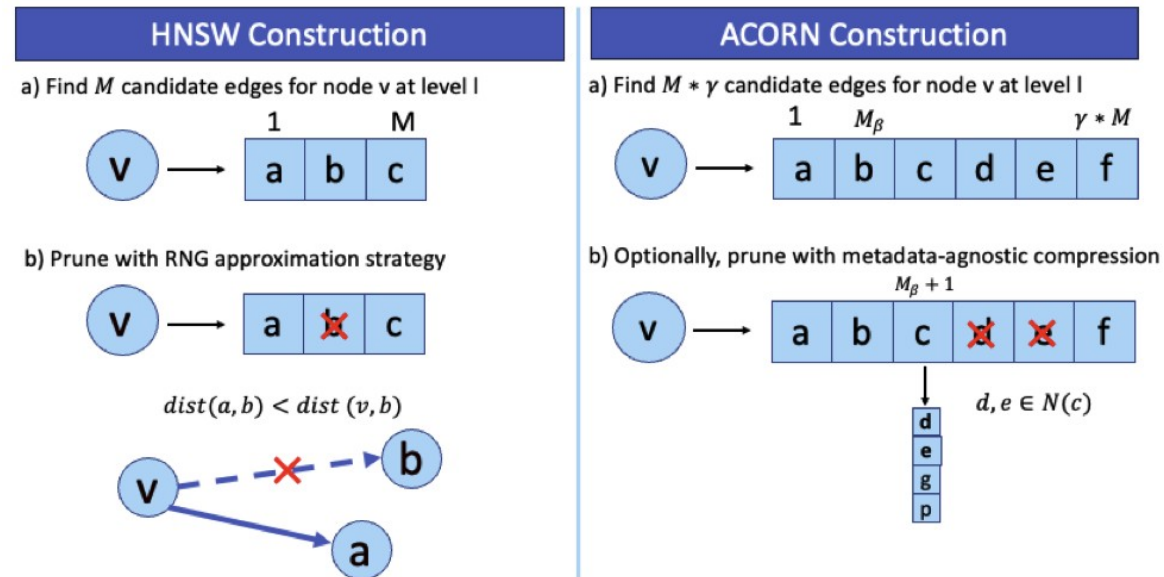


Fig. 3. An illustration of the predicate subgraph, shown by the green nodes. ACORN searches over the predicate subgraph to emulate search over an oracle partition index.

Creating a Denser Graph Index

- Expand neighbor lists by collecting a larger set of candidate edges (scaled by a factor γ) for each node during graph construction.
- Apply predicate-agnostic pruning by retaining a limited number of top candidates and pruning based on their connectivity within two-hop neighborhoods.
- Adjust the expansion factor γ dynamically based on the minimum predicate selectivity for robustness across queries.



Experiment

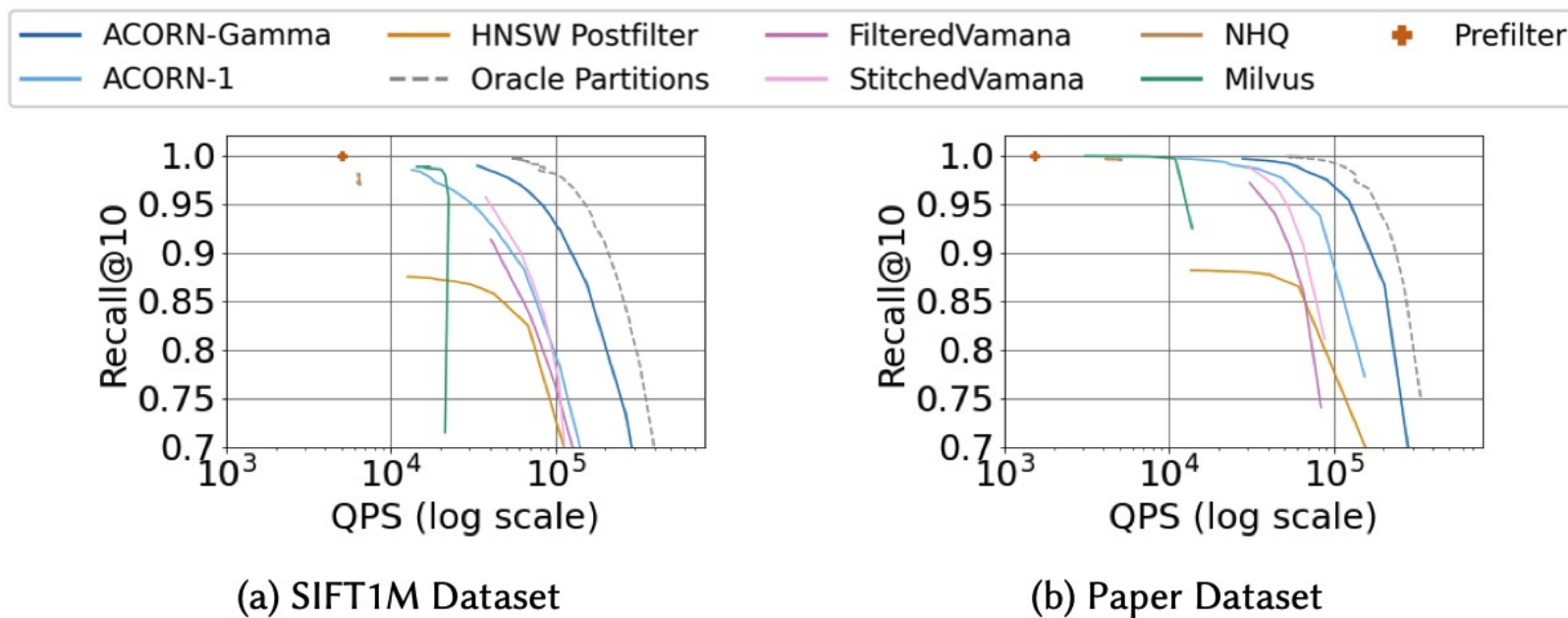


Fig. 7. Recall@10 vs QPS on SIFT1M and Paper

MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, William W. Cohen
Google Research

Presenter: Joohyung Yun

2024. 12. 11

Recap

Retrieval Augmented Generation

Motivation: Traditional language models rely only on *parametric knowledges* which are

- hard to fine-tune
- hard to explain (e.g., why is this output generated?)

Idea: *Retrieve, then generate*

- A *retriever* retrieves query-relevant documents from a corpus (i.e., database containing text)
- A *reader* generates the output based on the query and the retrieved results

In this class, we have focused on **two aspects of RAG**

- How can we retrieve query-relevant documents? How can we use them to generate an answer?
- What are **different modalities** that we can retrieve from? How should we **modify our retrieval and reading method** depending on the modality?

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 images

Barack Obama 248 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

"Barack" and "Obama" redirect here. For other uses, see Barack (disambiguation), Obama (disambiguation), and Barack Obama (disambiguation).

Barack Hussein Obama II^(b) (born August 4, 1961) is an American lawyer and politician who served as the 44th [president of the United States](#) from 2009 to 2017. A member of the Democratic Party, he was the first [African-American](#) president in U.S. history. Obama previously served as a [U.S. senator](#) representing [Illinois](#) from 2005 to 2008 and as an [Illinois state senator](#) from 1997 to 2004.

Obama was born in [Honolulu](#), Hawaii. He graduated from [Columbia University](#) in 1983 with a [Bachelor of Arts](#) degree in political science and later worked as a community organizer in Chicago. In 1988, Obama enrolled in [Harvard Law School](#), where he was the first black president of the *Harvard Law Review*. He became a civil rights attorney and an academic, teaching constitutional law at the [University of Chicago Law School](#) from 1992 to 2004. In 1996, Obama was elected to represent the 13th district in the [Illinois Senate](#), a position he held until 2004, when he successfully ran for the U.S. Senate. In the 2008 presidential election, after a close primary campaign against Hillary Clinton,

Barack Obama

Official portrait, 2012

44th President of the United States

In office
January 20, 2009 – January 20, 2017

Vice President Joe Biden

Preceded by George W. Bush

Succeeded by Donald Trump

United States Senator from Illinois

In office
January 3, 2005 – November 16, 2008

Preceded by Peter Fitzgerald

Succeeded by Roland Burris

Member of the Illinois Senate from the 13th district

In office

A example Wikipedia page

Lightning Deals Top 100 Today's big deals Toys & Games Trending Fashion Pet

Filtered by Toys & Games

Clear Filters

Department Any Category > Toys & Games

All

Arts & Crafts

Baby & Toddler Toys

Building Toys

Collectible Toys

See more

Brands

Buffalo Games

Ceaco

Disguise

Educational Insights

See more

Customer Reviews

All

5 stars & up

4 stars & up

3 stars & up

2 stars & up

1 star & up

Price

All

Under \$25

Under \$50

Under \$100

Under \$200

\$200 & Above

Discount

All

10% off or more

25% off or more

50% off or more

70% off or more

Prime Programs

Lightning Deals

Top 100

Today's big deals

Toys & Games

Trending

Fashion

Pet

360°

Cyber Monday Deal

SAHJUNG Galaxy Tab A9+ Plus Kids Edition 11" 64G...
+5 colors/patterns

Cyber Monday Deal

Tomb Raider Audio Player Starter Set with Playtime...
+5 colors/patterns

Cyber Monday Deal

Barbie Camper Play DreamCamper Toy
89% claimed

Cyber Monday Deal

Drone for Kids and Beginners, SYMA X440 RC...
24% claimed

Cyber Monday Deal

National Geographic Marble Run with Motorized Elevat...
2% claimed

Cyber Monday Deal

SEREED Baby Balance Bike for 1 Year Old Boys Girls 1...
+6

Cyber Monday Deal

Goopow Kids Camera Toys for 3-8 Year Old Girls Boys...
Save 10% with coupon

Cyber Monday Deal

GlowCity Light Up Football - Light U...
Save \$10.00 with coupon/Prime St...

An example Amazon page

Toys & Games > Tricycles, Scooters & Wagons > Balance Bikes

SEREED Baby Balance Bike for 1 Year Old Boys Girls 12-24 Month Toddler Balance Bike, 4 Wheels Toddler First Bike, First Birthday Gifts

Visit the SEREED Store

4.8 ★★★★★ 9,817 ratings | Search this page

Amazon's Choice

1 sustainability feature

30K+ bought in past month

Cyber Monday Deal

-26% \$37.98

List Price: \$50.99

No Import Charges & \$12.50 Shipping to Republic of Korea Details

Available at a lower price from other sellers that may not offer free Prime shipping.

Color: Green

\$39.99 \$42.49 \$42.49 \$37.98

\$39.99 \$39.99 \$42.49 \$42.49

Size Chart

Bike Type Balance Bike

Age Range (Description) Toddler

Brand SEREED

Number of Speeds 1

Color Green

Wheel Size 6.7 Inches

Frame Material Aluminum

Suspension Type Rigid

Special Feature Lightweight

Included Components Balance Bike

See less

Roll over image to zoom in

An example reddit page

Notation Definition

Multimodal Data Source

We typically deal with a data source generated from Wikipedia.

A Wikipedia webpage is simply a sequence of

- passages
- (image, caption) pairs

Definition)

$$D = \{P_1, \dots, P_n\} \cup \{M_1, \dots, M_m\}$$

- P : a natural language passage
- $M = (I, C)$: an (image, caption) pair
 - I : An image
 - C : A natural language caption



Barack Obama

248 languages

Article Talk

Read View source View history Tools

From Wikipedia, the free encyclopedia

"Barack" and "Obama" redirect here. For other uses, see [Barack \(disambiguation\)](#), [Obama \(disambiguation\)](#), and [Barack Obama \(disambiguation\)](#).

Barack Hussein Obama II^[a] (born August 4, 1961) is an American lawyer and politician who served as the 44th [president of the United States](#) from 2009 to 2017. A member of the [Democratic Party](#), he was the first [African-American](#) president in U.S. history. Obama previously served as a [U.S. senator](#) representing [Illinois](#) from 2005 to 2008 and as an [Illinois state senator](#) from 1997 to 2004.

Obama was born in [Honolulu](#), Hawaii. He graduated from [Columbia University](#) in 1983 with a [Bachelor of Arts](#) degree in political science and later worked as a [community organizer](#) in [Chicago](#). In 1988, Obama enrolled in [Harvard Law School](#), where he was the first black president of the [Harvard Law Review](#). He became a civil rights attorney and an academic, teaching [constitutional law](#) at the [University of Chicago Law School](#) from 1992 to 2004. In 1996, Obama was elected to [represent the 13th district in the Illinois Senate](#), a position he held until 2004, when he [successfully ran for the U.S. Senate](#). In the 2008 presidential election, after a close [primary campaign](#) against [Hillary Clinton](#),

Barack Obama



Official portrait, 2012

44th President of the United States

In office

January 20, 2009 – January 20, 2017

Vice President Joe Biden

Preceded by George W. Bush

Succeeded by Donald Trump

United States Senator from Illinois

In office

January 3, 2005 – November 16, 2008

Preceded by Peter Fitzgerald

Succeeded by Roland Burris

Member of the Illinois Senate from the 13th district

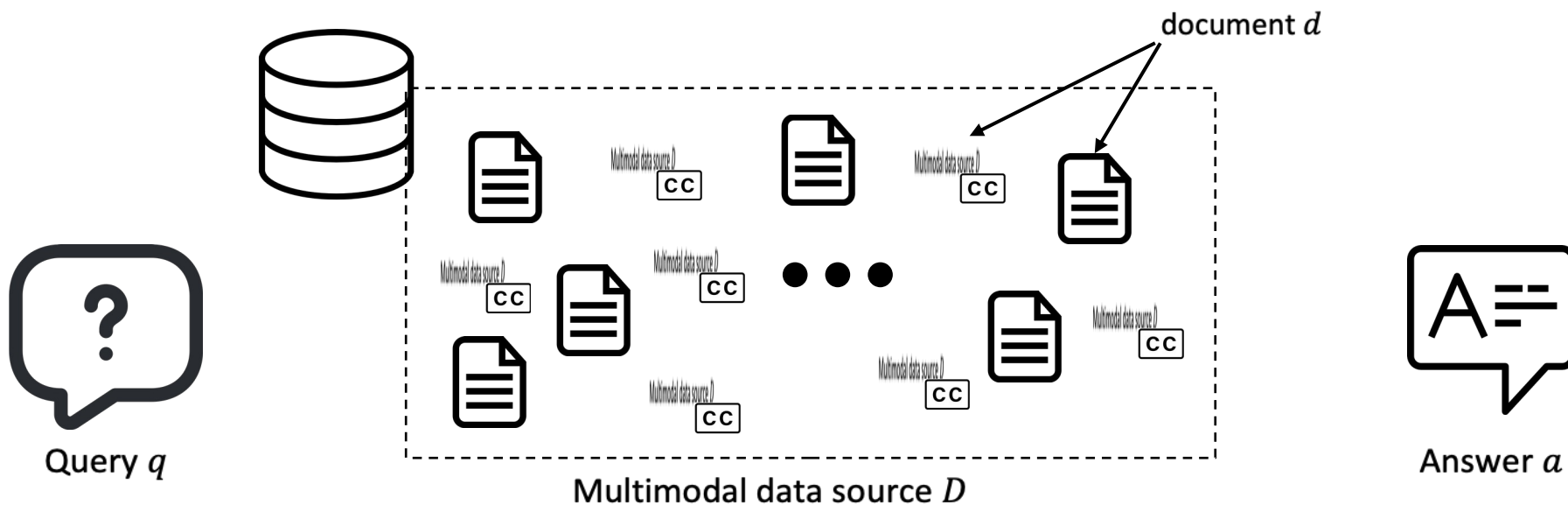
In office

Problem Description

Multimodal Question Answering

Given a query q and a multimodal data source D , generate the correct answer a .

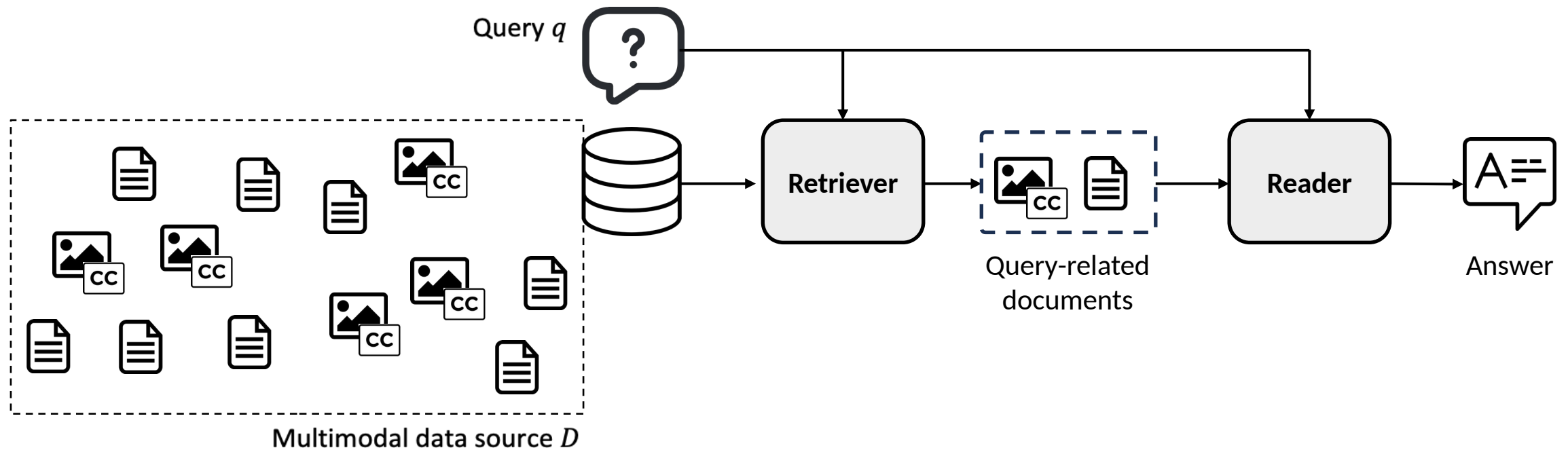
- Query $q = [q_1, \dots, q_t]$: A sequence of tokens
- Multimodal data source D : (Defined in the earlier page)
- Answer $a = [a_1, \dots, a_r]$: A sequence of tokens



MultiModal Retrieval Augmented Generation

A multimodal retrieval augmented generation (MMRAG) model is composed of

- Retriever: Retrieves query-relevant subset of documents from the multimodal data source
- Reader: Generates the predicted answer using the query and the retrieved documents

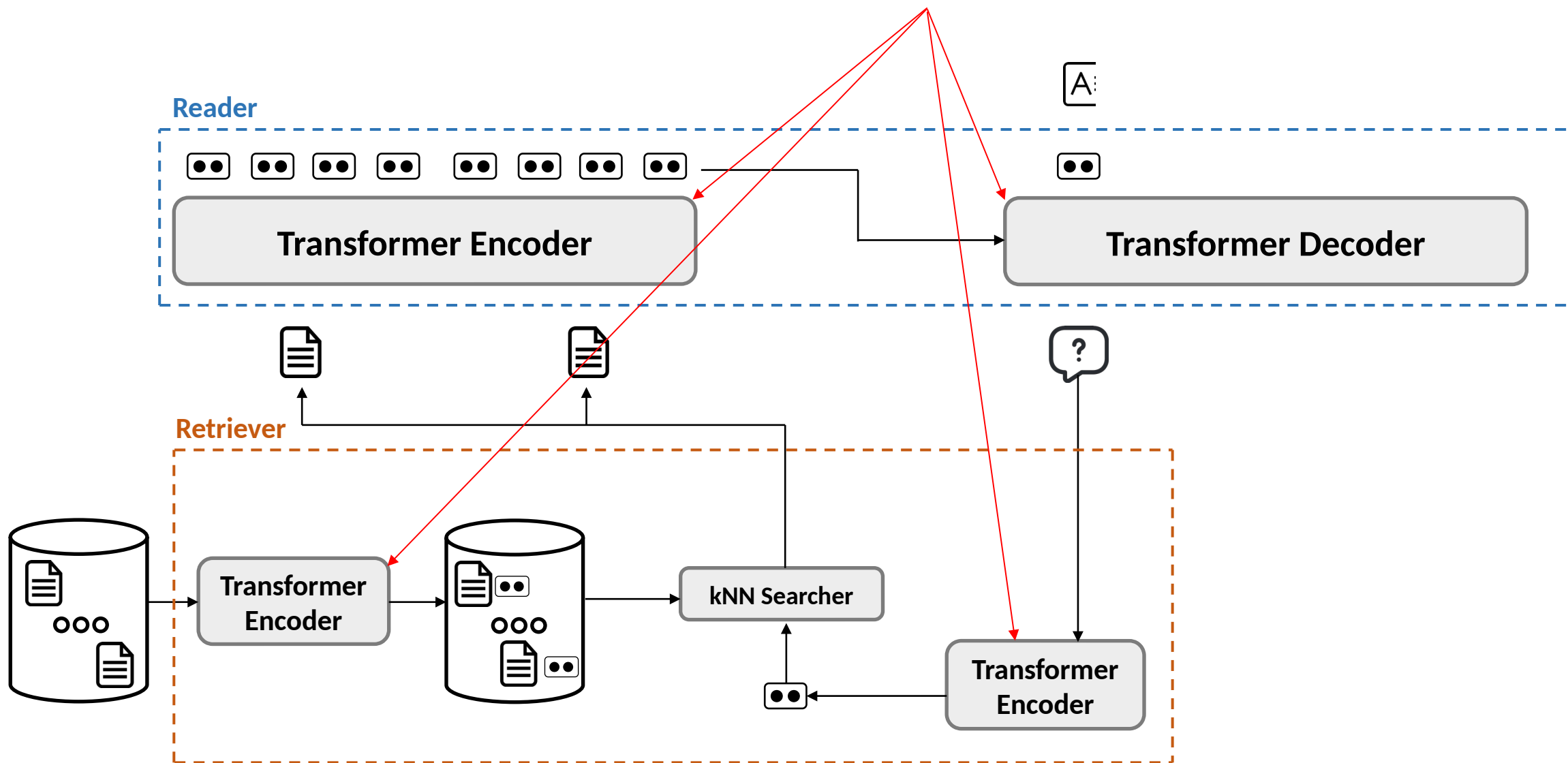


Recap

Two Challenges of MMRAG

Challenges:

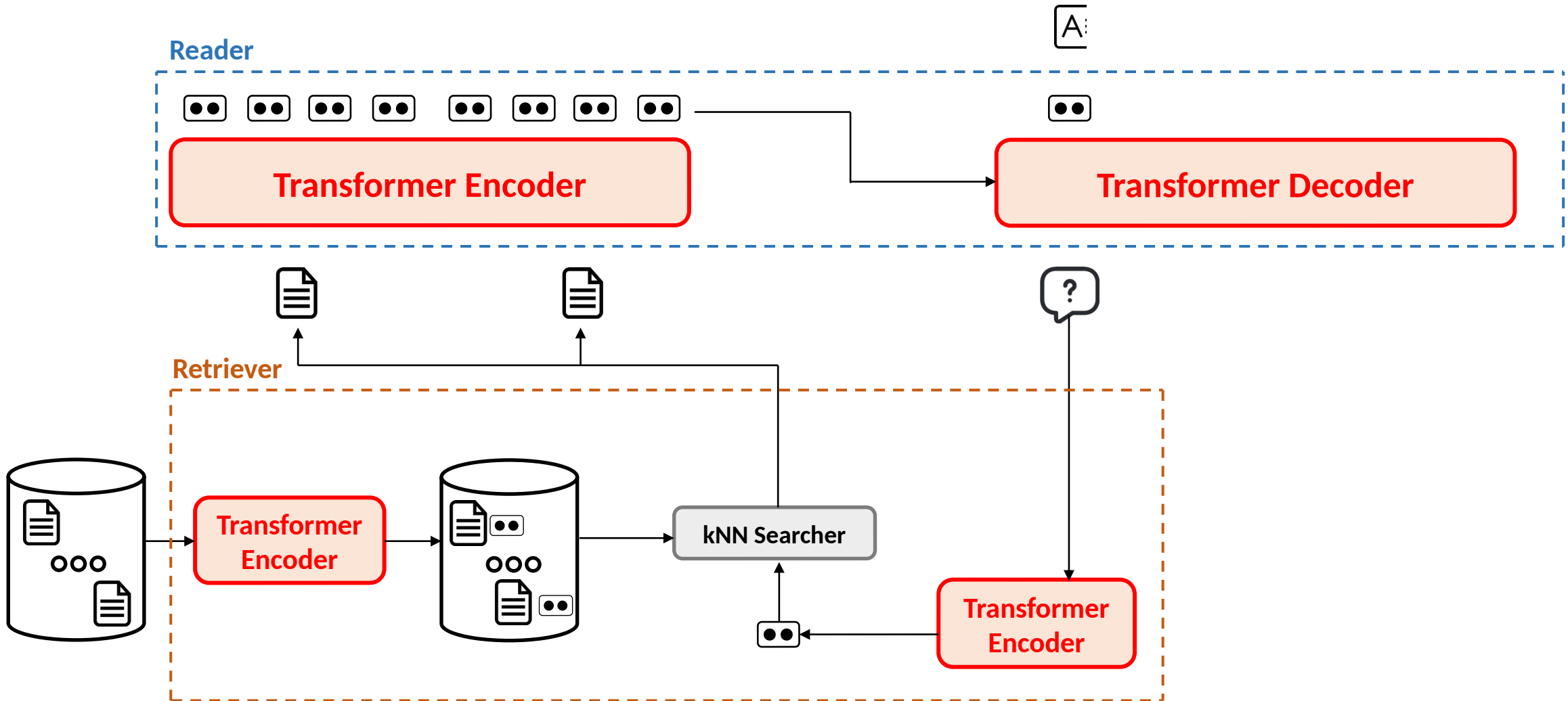
- How can we make the encoders and decoders handle multimodal data?
- How can we train such encoders and decoders?



Challenge #1

Designing Encoders and Decoders that Handle Multimodal Data

How can we make the encoders and decoders handle multimodal data?

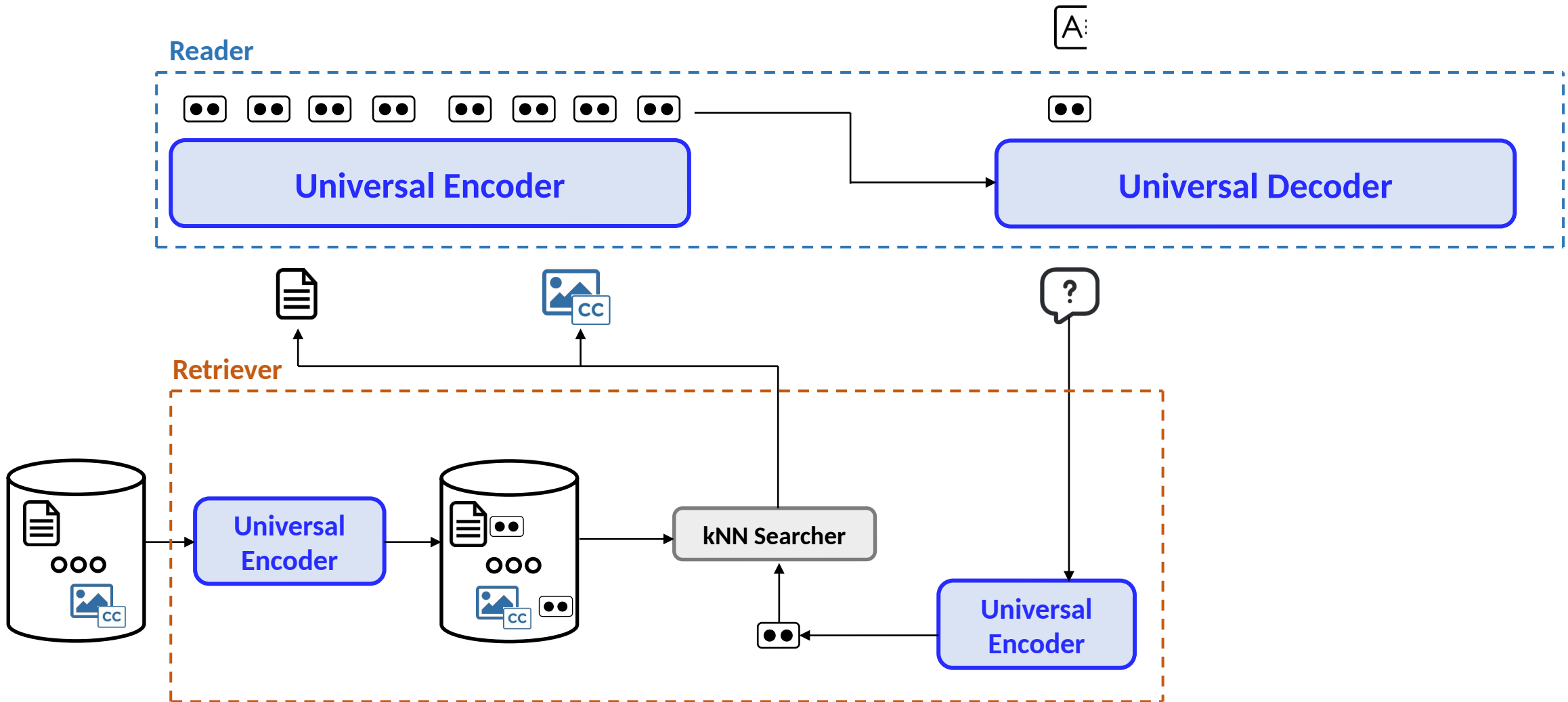


Solution #1

Universal Encoder & Decoder

Develop

- a universal encoder that can encode text, image, or even both.
- a universal decoder that can generate text using multimodal encodings.

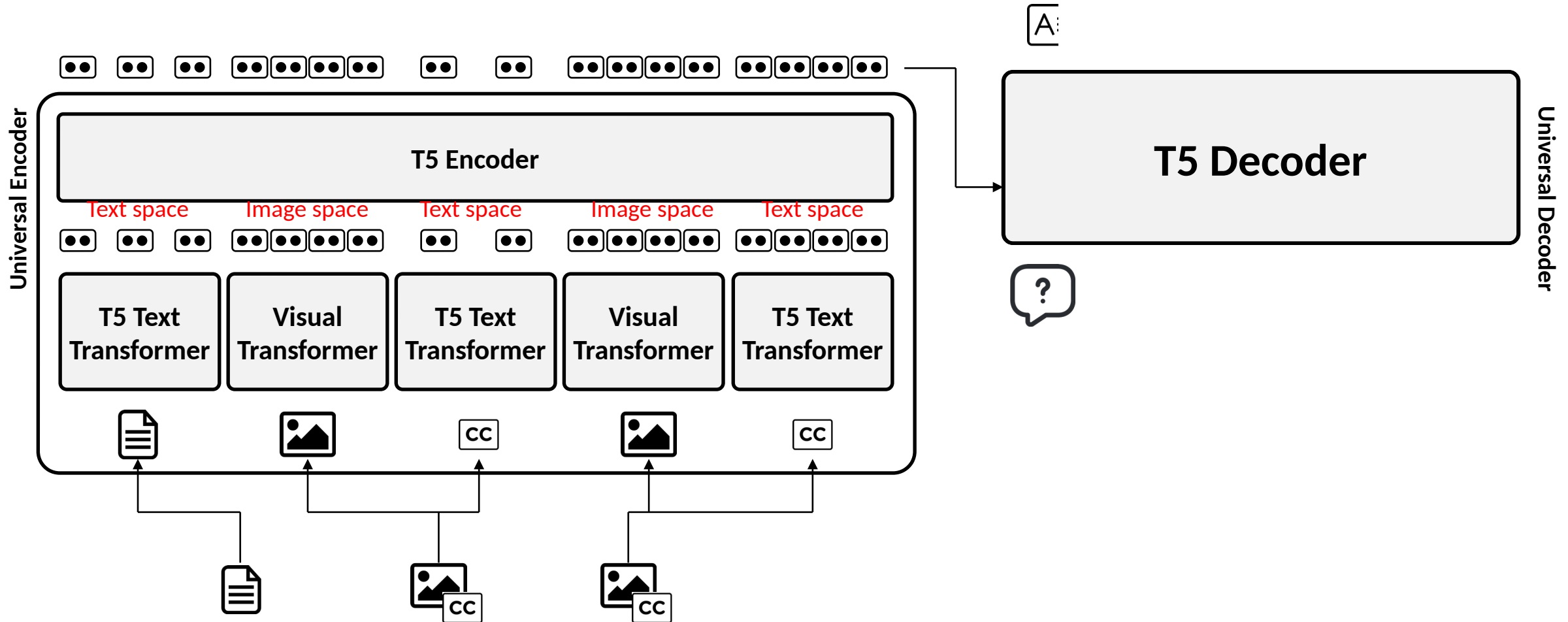


Solution #1

Inference Using Universal Encoder & Decoder

Problem: Different embeddings of two spaces!

Solution: Add an embedding layer to map them into a universal space



Challenge #2

Training Encoders and Decoders that Handle Multimodal Data

The encoders and decoders have to be retrained from the ground, as a new modality of image is introduced.

We have solve the following questions for two phases of training

- **Pretraining:**
 - What are some big datasets we can use for pretraining?
 - How should we set the loss functions?
 - How can we efficiently perform pretraining?
- **Fine-tuning:**
 - How should we set the loss functions?
 - How can we efficiently perform pretraining?

Datasets Used for Pretraining

Used 4 visual question answer benchmarks to perform pretraining

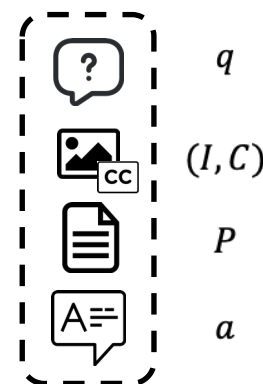
- CC
- LAION
- PAQ
- VQA

Dataset	#Size
CC	15M
LAION	200M
PAQ	65M
VQA	400K

Table 1: Pre-training Dataset Statistics

General configuration of the four datasets:

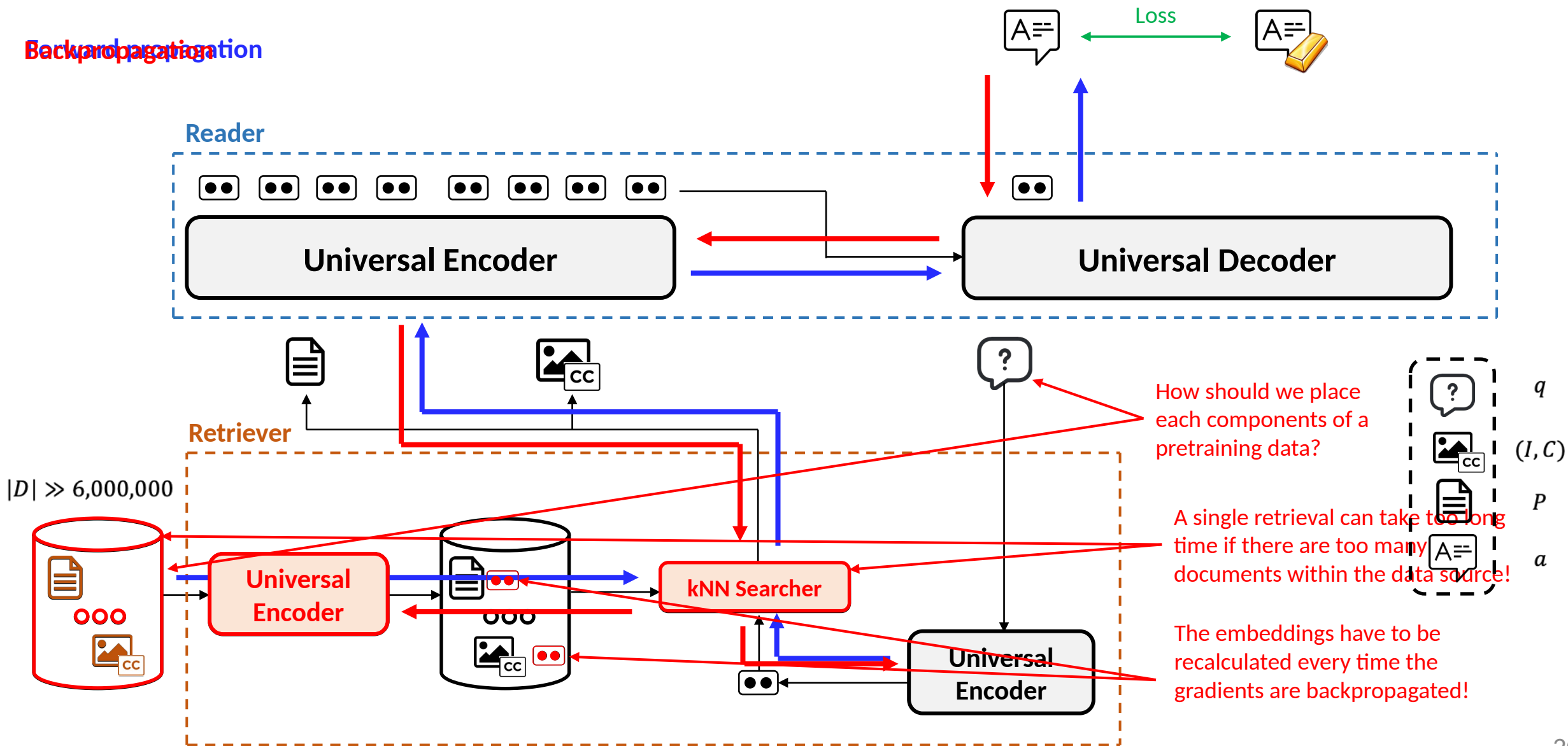
- **Input:** query
- **Input:** query-related (image, caption) pair
- **Input:** a passage describing the query-related image
- **Output:** answer



A pretraining data

Problems of Naïve Pretraining

Backpropagation



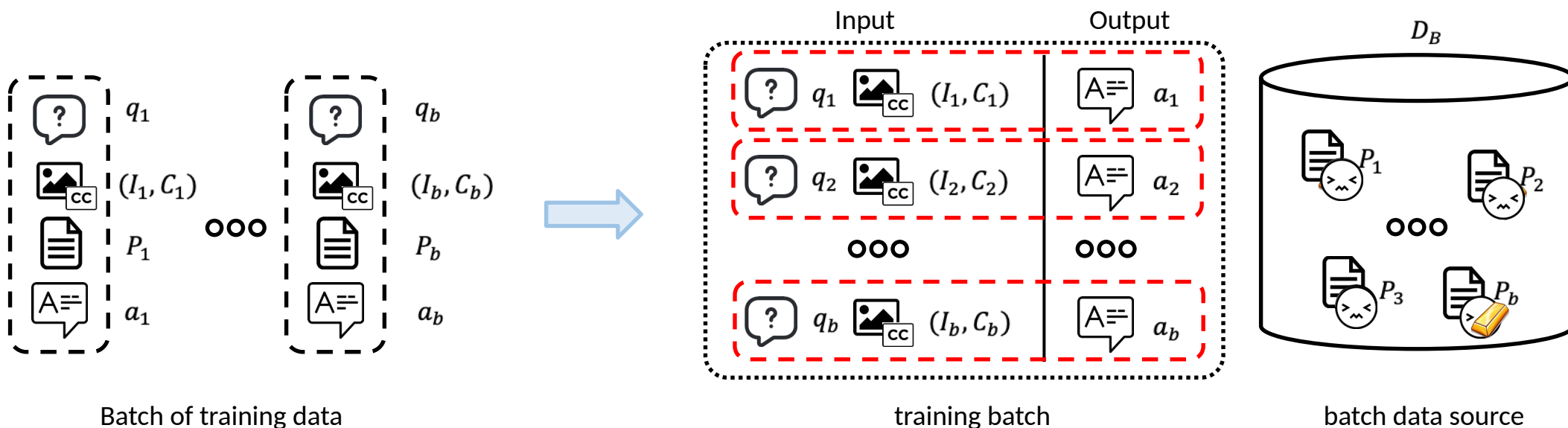
Solution to the Problems: In-Batch Pretraining

Generate a batch of multiple ($b = 4096$) pretraining data

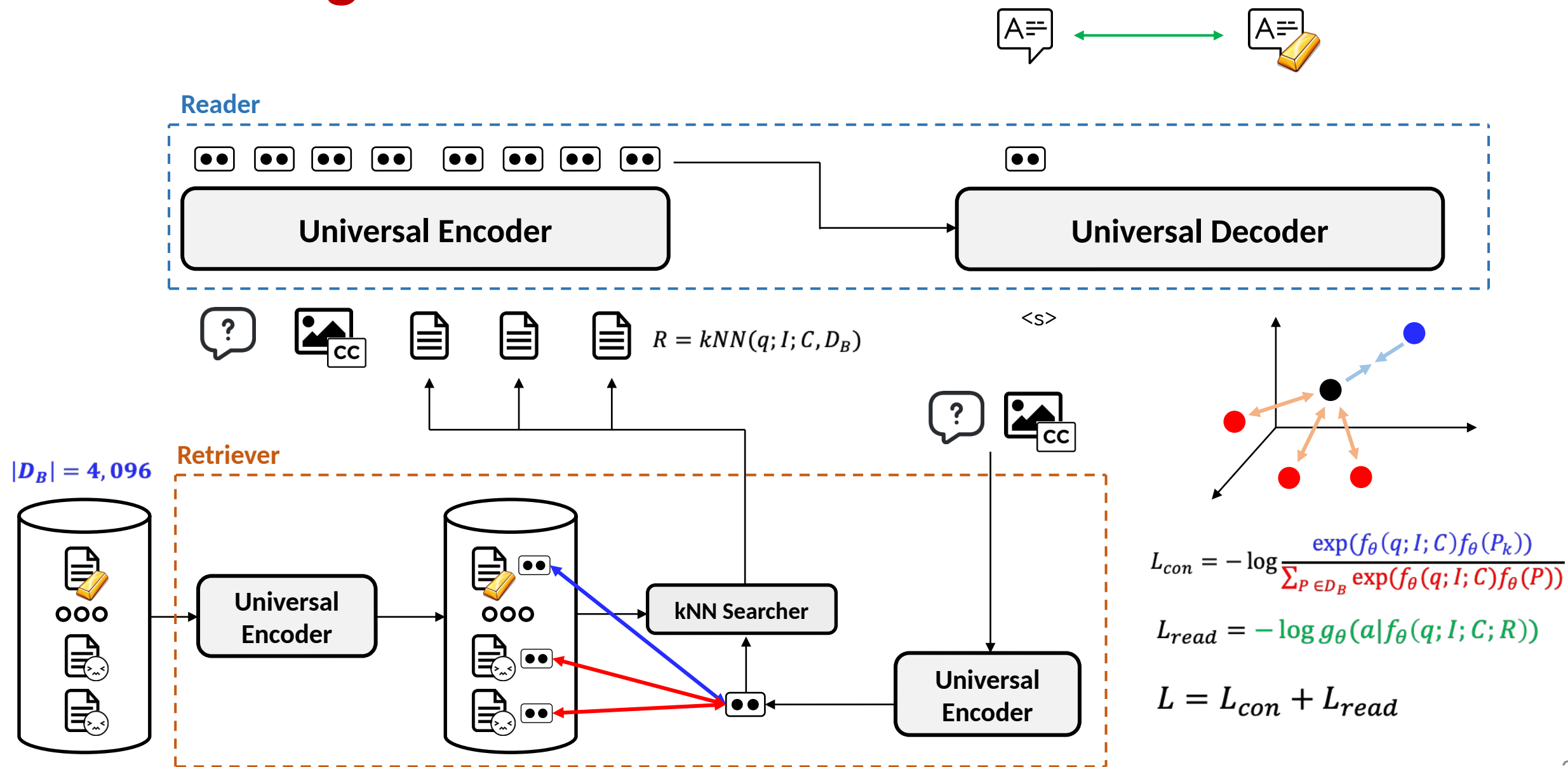
- Solution to the training configuration problem:**

- Input: $q_k; I_k; C_k$
- Data source: $D_B = \{P_1, P_2, \dots, P_k, \dots, P_b\}$
- Gold passage: P_k
- Answer: a_k

- Solution to data source size problem:** construct a data source composed of only passages from the batch.
- Solution to embedding recalculation problem:** update the weight parameters for the embedder in a batch-wise manner.



Pretraining Overview

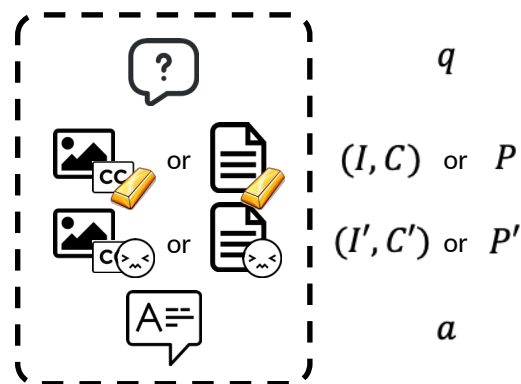


Goals and Problems of the Fine-tuning Phase

Goal: Fine-tune the universal embedder and decoder according to our downstream task

Configurations of the MMRAG task (revisited):

- **Input:** query
- **Label:** a gold document
- **Label:** a hard negative document
- **Output:** answer



A fine-tuning data

Problems of the fine-tuning phase:

- Embedding recalculation

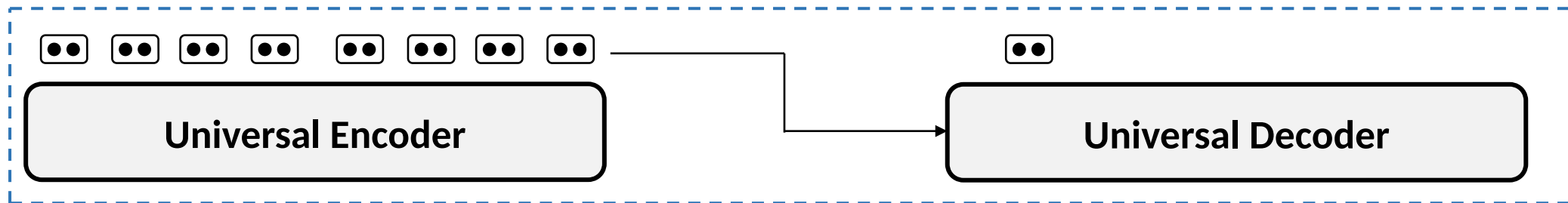
Two-stage Fine-tuning

In-batch Fine-tuning

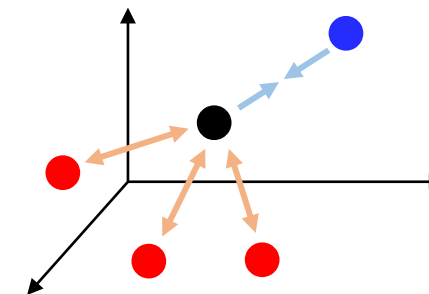
Resembles in-batch pretraining



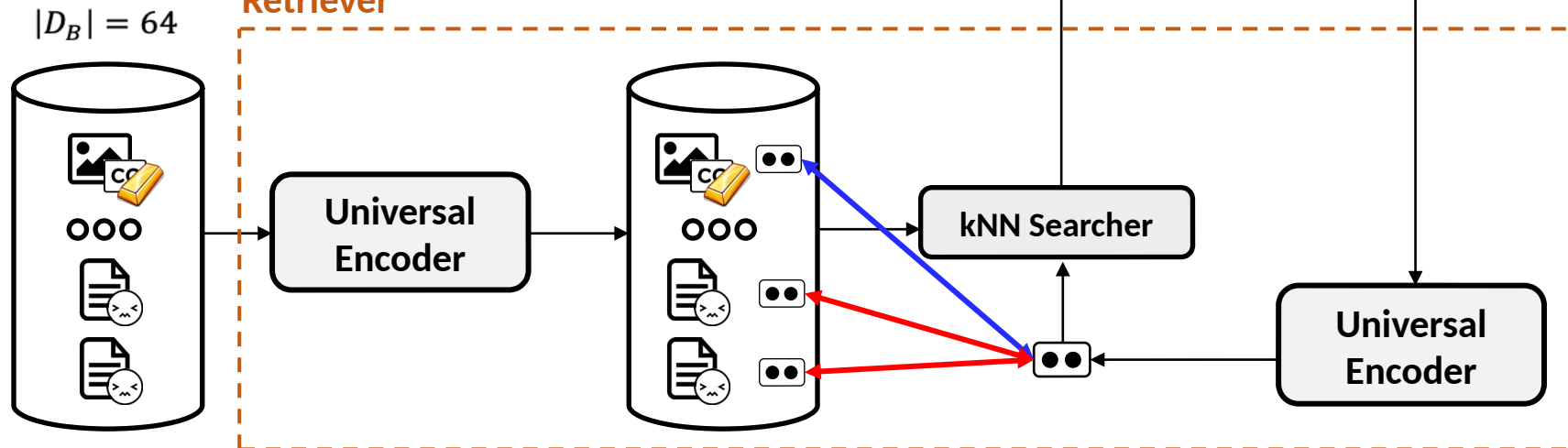
Reader



$\langle s \rangle$



Retriever



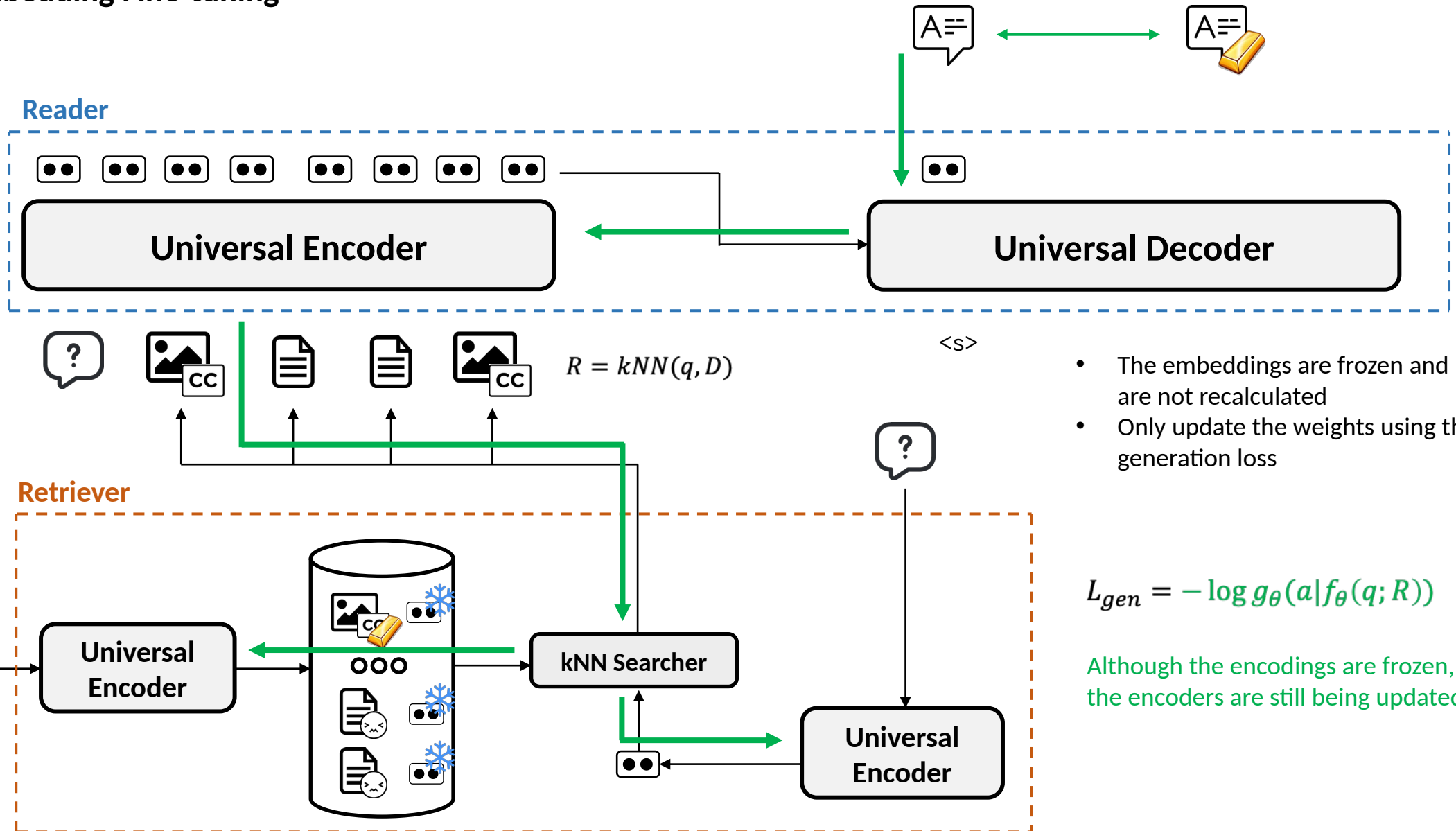
$$L_{con} = -\log \frac{\exp(f_{\theta}(q)f_{\theta}(d_k))}{\sum_{d \in D_B} \exp(f_{\theta}(q)f_{\theta}(d))}$$

$$L_{gen} = -\log g_{\theta}(a|f_{\theta}(q; R))$$

$$L = L_{con} + L_{gen}$$

Two-stage Fine-tuning

Fixed-Embedding Fine-tuning



- The embeddings are frozen and are not recalculated
- Only update the weights using the generation loss

Experiments

Goals of Experiment

1. **Accuracy comparison:** Does MuRAG accurately answer the multimodal questions?
2. **Ablation study:** How does each of MuRAG's training method affect the end-to-end accuracy?

Experiment Settings

Dataset

- MultimodalQA
- WebQA

Dataset	Train Image/Text	Dev Image/Text	Test Image/Text
WebQA	18K/17K	2.5K/2.4K	3.4K/4K
MultimodalQA	2.1K/7.4K	230/721	-

Table 3: Overall Statistics of downstream dataset.

Metrics

- MultimodalQA
 - Exact Match
 - F1
- WebQA
 - Retrieval F1
 - Generation Fluency
 - Generation Recall
 - Overall Generation

Competitors

- No retrieval (Question-Only)
- Retriever
 - CLIP (K): CLIP + kNN
- MultimodalQA Reader
 - AutoRouting
- WebQA Reader
 - VLP

Accuracy Comparison

MultimodalQA					
Evaluation	Distractor				
Metrics	Text		Image		All EM
	EM	F1	EM	F1	
Question-Only	15.4	18.4	11.0	15.6	13.8
AutoRouting	49.5	56.9	37.8	37.8	46.6
MuRAG	60.8	67.5	58.2	58.2	60.2

Full-Wiki					
Evaluation	Full-Wiki				
Metrics	Text		Image		All EM
	EM	F1	EM	F1	
CLIP (10) + AutoRouting	35.6	40.2	32.5	32.5	34.7
MuRAG	49.7	56.1	56.5	56.5	51.4

Table 5: Multimodal dev-set results on the subset.

WebQA				
Evaluation Metrics	Retr	Distractor		Overall
		FL	Accuracy	
Question-Only	-	34.9	22.2	13.4
VLP (Oscar)	68.9	42.6	36.7	22.6
VLP + ResNeXt	69.0	43.0	37.0	23.0
VLP + VinVL	70.9	44.2	38.9	24.1
MuRAG	74.6	55.7	54.6	36.1

Full-Wiki				
Evaluation	Retr	FL	Accuracy	Overall
CLIP (2) + VLP	11.9	34.2	24.1	14.6
CLIP (20) + VLP	24.0	36.1	27.2	16.1
MuRAG	39.7	50.7	47.8	31.5

Table 4: WebQA official test-set results indicated on leaderboard³ as of May 2022. Retr denotes the retrieval-F1 score. FL refers to fluency metric BARTScore, and Accuracy refers to keyword matching F1 score, they are combined as Overall.

MultimodalQA

- **Text question EM:** 10+%p increase
- **Image question EM:** 20+%p increase

WebQA

- **Retrieval performance:** 15%p increase in in full-wiki setting
- **Generation fluency:** 12%p increase under distractor setting
- **Generation accuracy:** 16%p increase under full-wiki setting

Ablation Study: Pre-training Data Source

Pre-train Dataset	FL	Accuracy	Overall
None	42.5	36.1	23.5
CC	46.4	41.3	25.6
LAION	47.8	44.8	28.3
VQA	47.0	44.4	27.4
PAQ	46.8	42.8	27.0
LAION+CC	49.5	47.4	30.7
LAION+CC+PAQ	53.7	51.8	34.4
LAION+CC+PAQ+VQA	55.7	54.6	36.1

Table 6: Ablation Study for different pre-training corpus, score under distractor setting.

Examined the contributions of different pre-training data sources

Without any pre-training: overall score 23.5, underperforms all baselines

Pretraining with only one data source: Best accuracy with LAION dataset

Ablation Study: Two-stage Fine-tuning

Model	WebQA	Multimodal
MuRAG (Only In-Batch)	29.4	49.6
MuRAG (Only Fixed-Retrieval)	25.8	40.7
MuRAG (Two Stage)	31.5	51.4

Table 7: Ablation Study for different fine-tuning stages to see their contributions. WebQA uses the overall score, and MultimodalQA refers to EM-all score.

Examined the contribution of each stage in two-stage fine-tuning

- Only in-batch fine-tuning: accuracy drop by 2%p in both datasets
 - Only fixed-retrieval fine-tuning: accuracy drop by about 6%p (WebQA) and 11%p (MultimodalQA)
- > Necessary to adapt MuRAG's pre-trained retriever to the downstream datasets.

APPENDIX

Human Analysis: Error Categorization

Evaluation	Model	Correct	Wrong
Distractor	MuRAG (Text)	80%	20%
	MuRAG (Image)	64%	36%
Full-Wiki	MuRAG (Text)	72%	28%
	MuRAG (Image)	54%	46%

Table 8: The human evaluation results on WebQA dataset separately for image/text queries.

Manually studied 200 model outputs and classified them into 3 categories

- Image queries are much harder than text queries
- Full-wiki setting is harder than the distractor setting
 - Distractor setting: 64%
 - Full-wiki setting: 54%

Qualitative Analysis: Error Categorization

Some qualitative reasons for these errors

- Some questions are related to infrequent objects
- The image scenes are highly complex with a large number of objects
- The questions require optical character recognition ability from images

Category	Count	Object	Color	Shape	Gender
Ratio	52%	29.4%	5.8%	5.8%	5.8%

Table 9: Error categorization and their ratios on sampled WebQA-dev image queries.

Waterpark in Ayia Napa ... Facade of Grand Palais, Paris 9 ...



Q: At the waterpark in Ayia Napa, are there waterslides that are taller than trees?

P: Yes, at the waterpark in Ayia Napa there are waterslides that are taller than trees.

R: Yes, there are waterslides taller than trees.

Q: How many spaces for banner flags to be hung are there between the columns on the Grand Palais?

P: There are two spaces for banner flags to be hung.

R: There are three spaces for banner flags to be hung.

Museum exhibits IMNH exhibit in located in Pocatello Idaho.



Museum of international folk art display Display at the ...



Q: What kind of lighting is on the ceiling at the Museum of International Folk Art in Santa Fe and the IMNH exhibit in Pocatello Idaho?

P: Track lighting is on the ceiling at the Museum of International Folk Art in Santa Fe and the IMNH exhibit in Pocatello Idaho.

R: Spot lights are on the ceiling.

Vision-Language Pretraining Model Overview

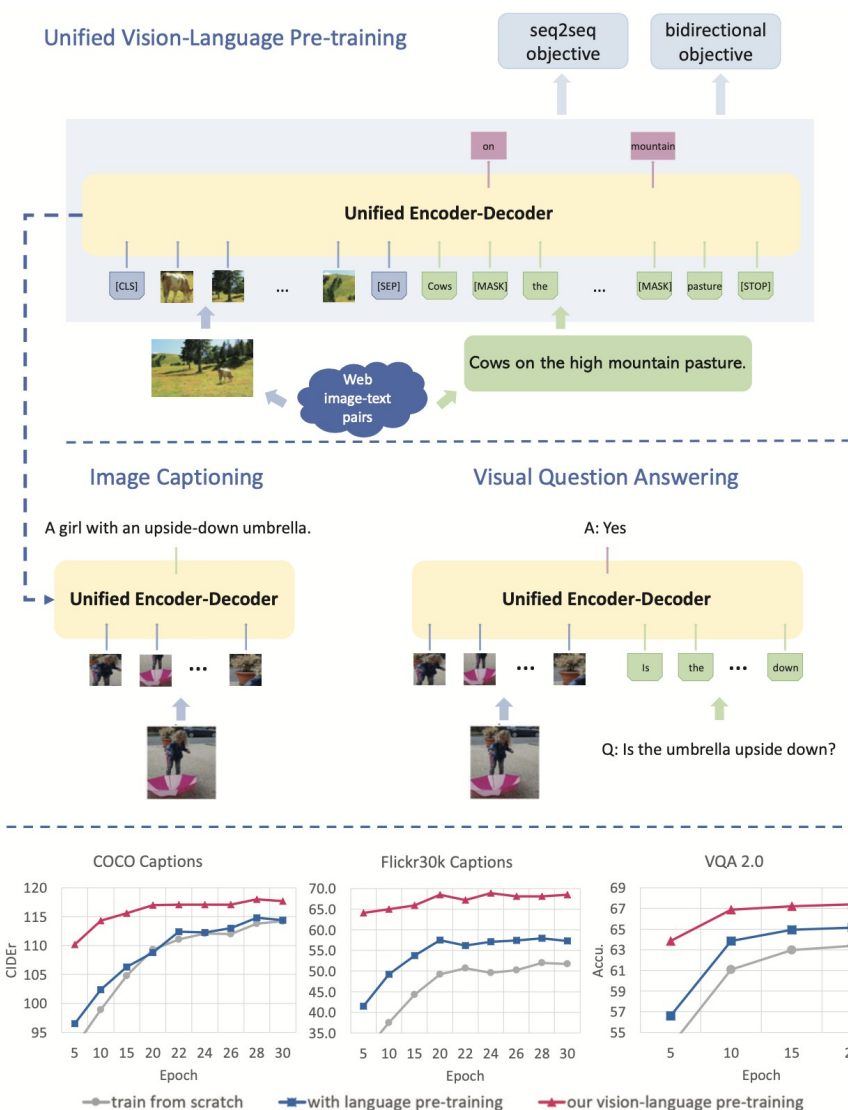
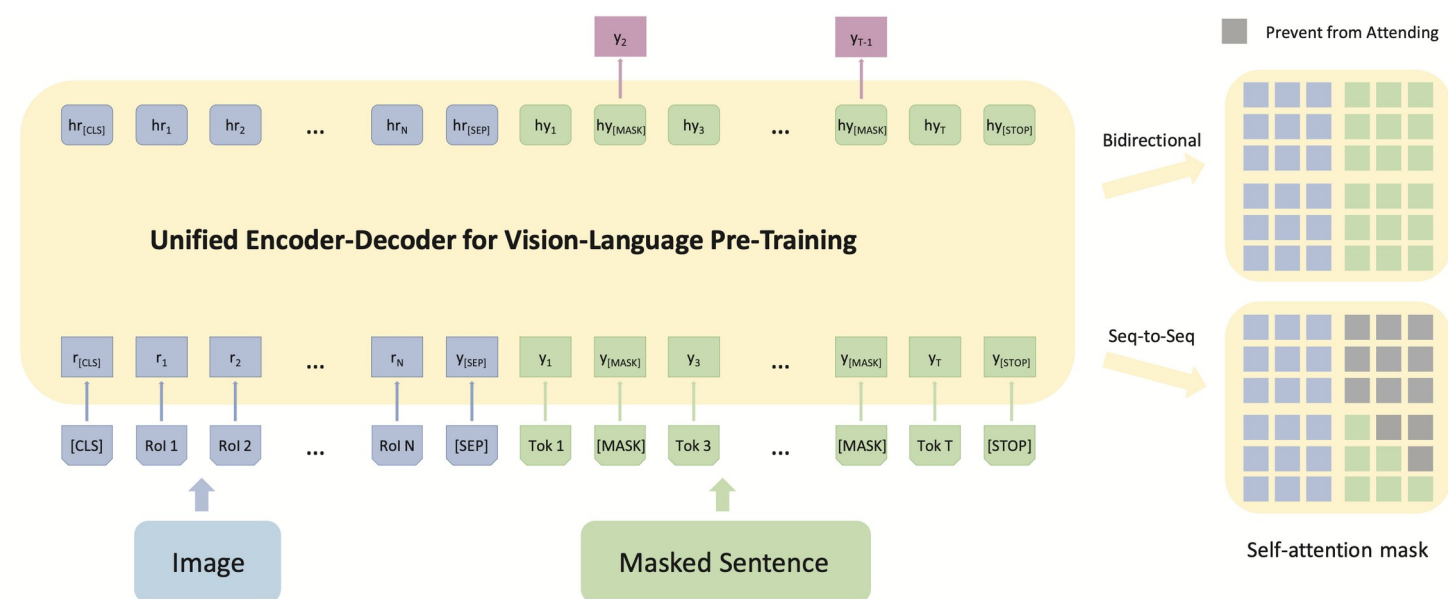


Figure 2: Model architecture for pre-training. The input comprises of image input, sentence input, and three special tokens ([CLS], [SEP], [STOP]). The image is processed as N Region of Interests (RoIs) and region features are extracted according to Eq. 1. The sentence is tokenized and masked with [MASK] tokens for the later masked language modeling task. Our Unified Encoder-Decoder consists of 12 layers of Transformer blocks, each having a masked self-attention layer and feed-forward module, where the self-attention mask controls what input context the prediction conditions on. We implemented two self-attention masks depending on whether the objective is bidirectional or seq2seq. Better viewed in color.