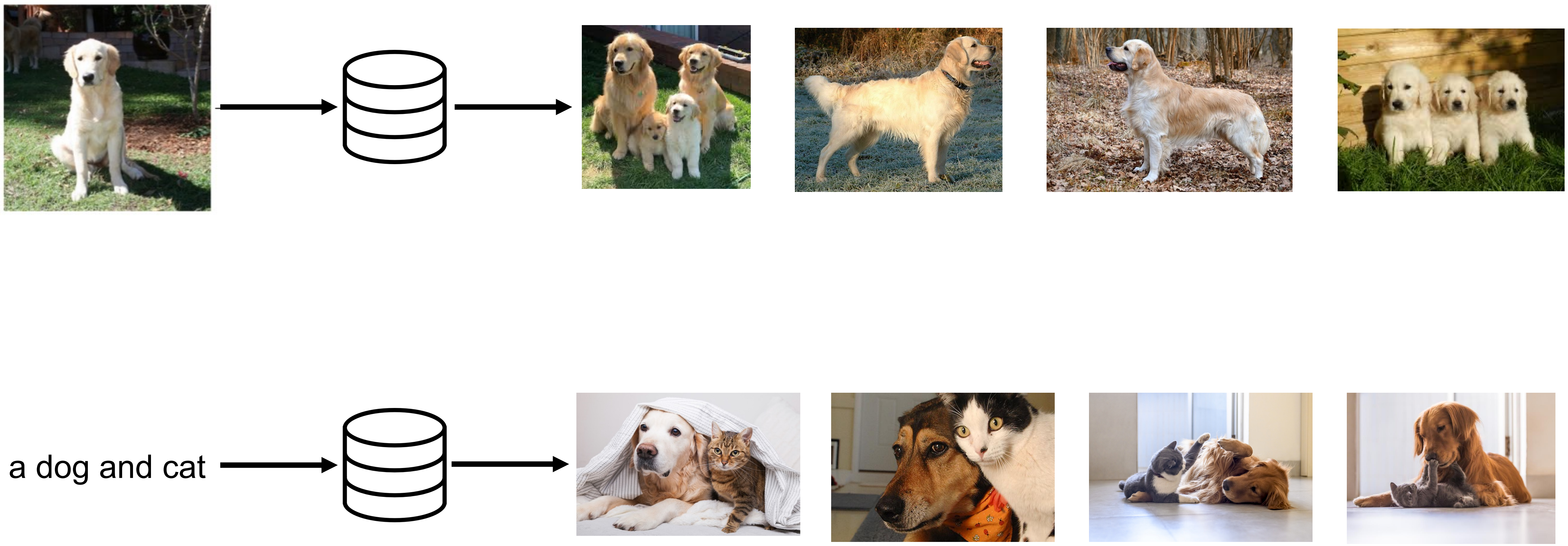
Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval

CVPR2023

Kuniaki Saito[1,2] , Kihyuk Sohn[3], Xiang Zhang[2] , Chun-Liang Li[2] ,
Chen-Yu Lee[2], Kate Saenko[1,4], Tomas Pfister[1]
[1]Boston University, [2]Google Cloud AI Research, [3]Google Research, [4]MIT-IBM Watson AI Lab

Image retrieval with image or text query



Query

Retrieved images

a dog and cat

Can we combine text and image query?

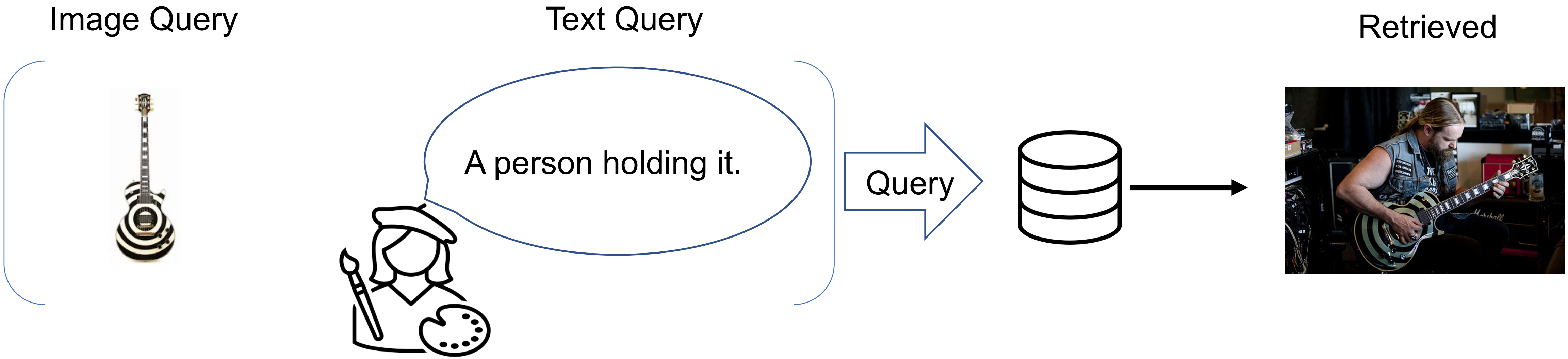


Image Query

Text Query

A person holding it.

Query

Retrieved

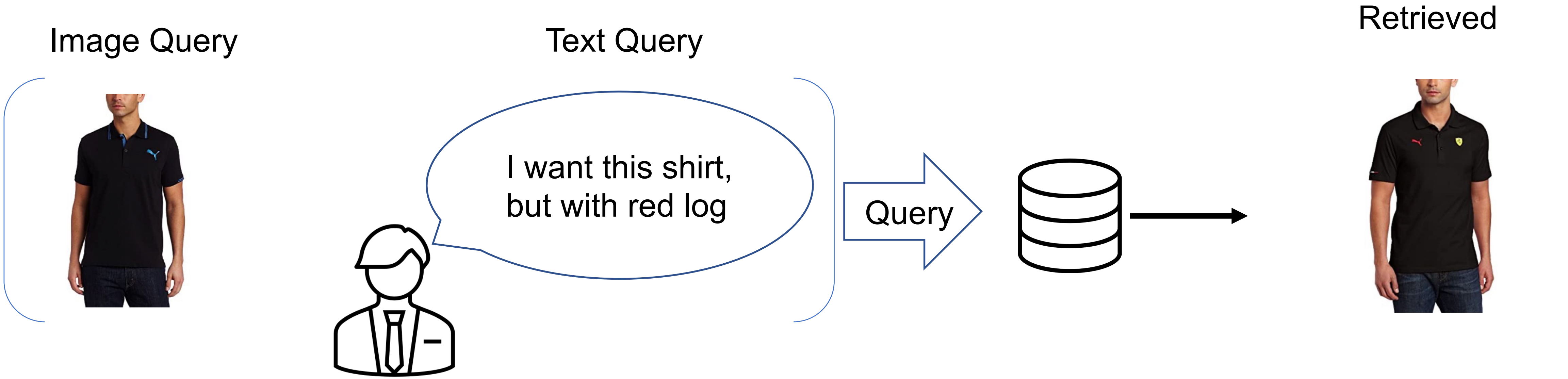


Image Query

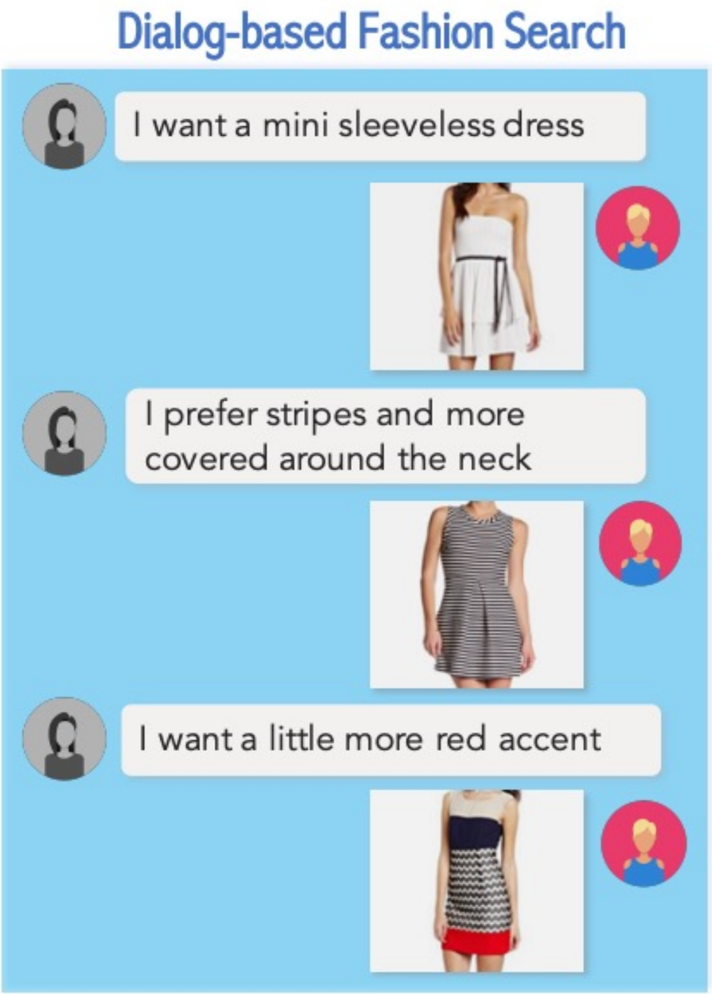
Text Query

I want this shirt,
but with red log

Query

Retrieved

Application



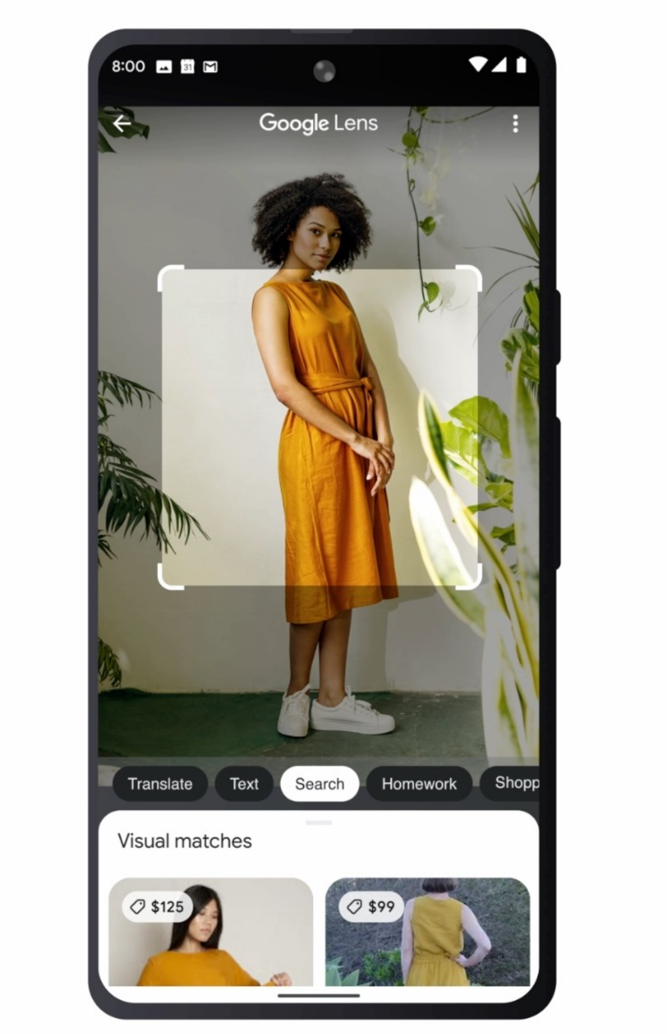
Dialog-based Fashion Search

I want a mini sleeveless dress

I prefer stripes and more covered around the neck

I want a little more red accent

https://arxiv.org/pdf/1905.12794.pdf



Multisearch allows people to search with both images and text at the same time.

https://blog.google/products/search/multisearch/

Composed Image Retrieval

- Goal: (Query Image, Query Text) => Target Image
- Query Text: Modification to query image

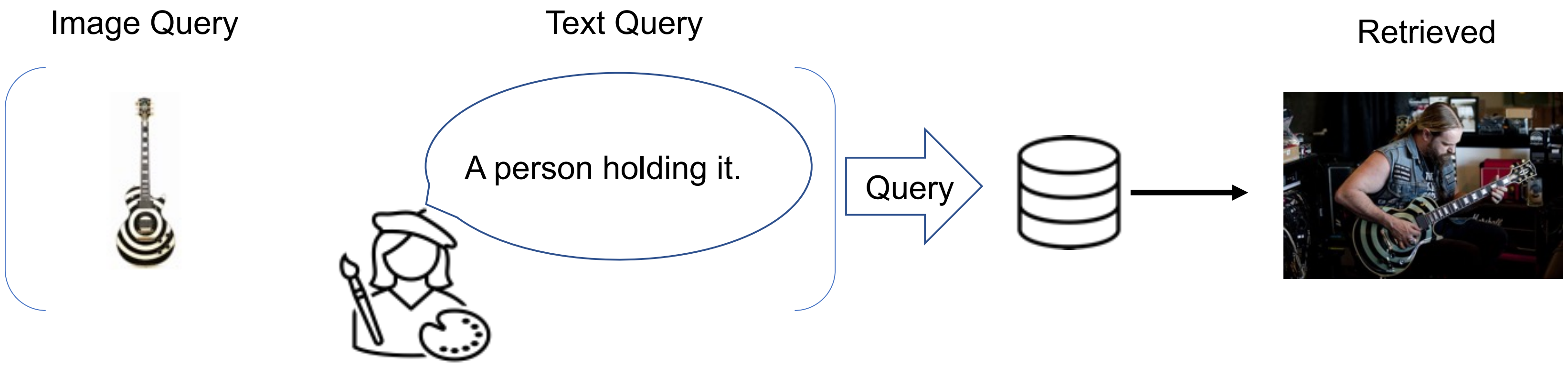


Image Query

Text Query

A person holding it.

Query

Retrieved

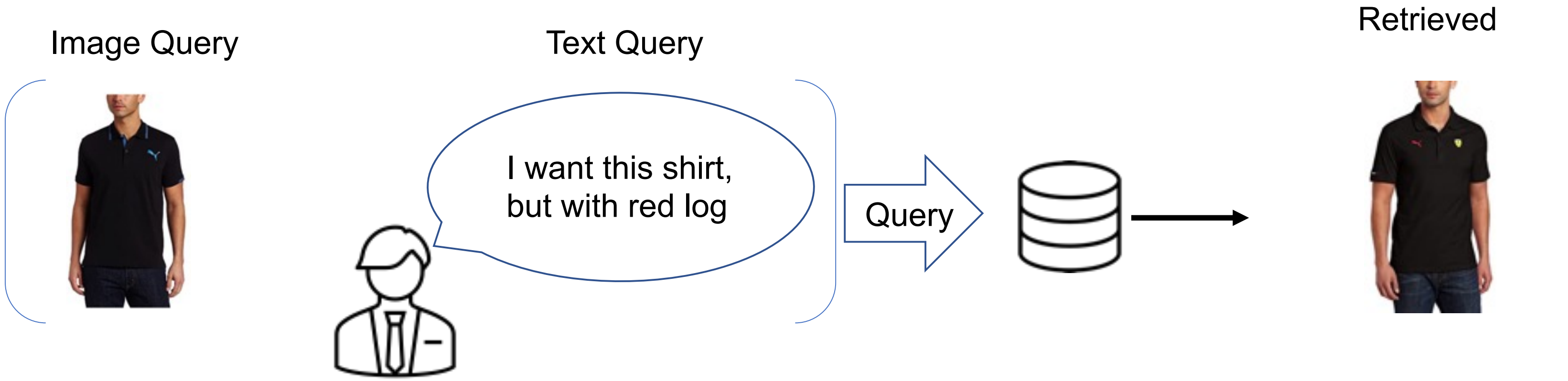


Image Query

Text Query

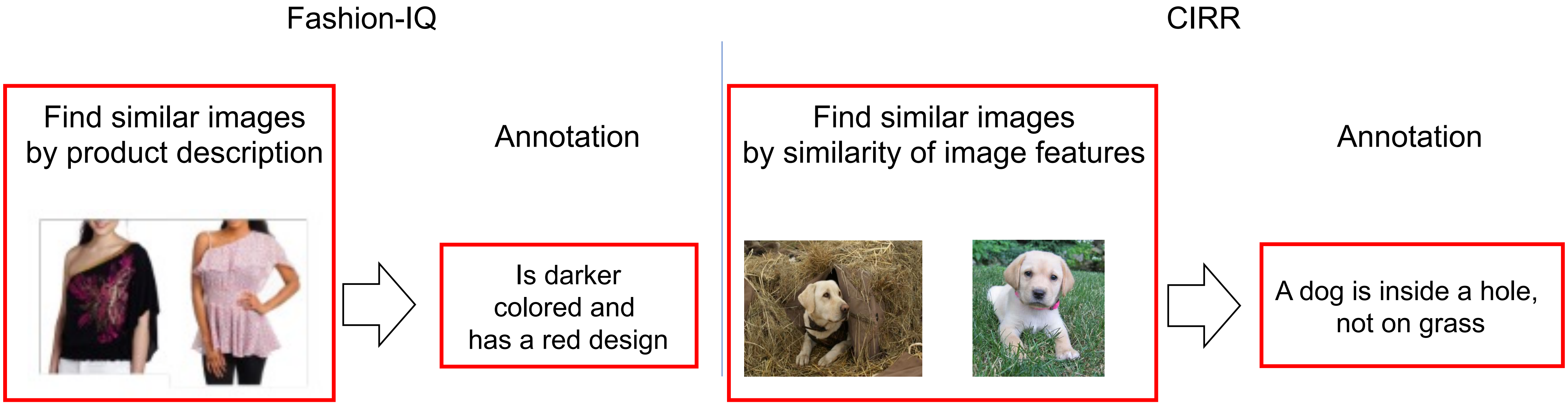
I want this shirt,
but with red log

Query

Retrieved

Data Collection in Composed Image Retrieval

- Triplet: (Query Image, Query Text, Target Image)



Fashion-IQ

Find similar images by product description

Annotation

Is darker colored and has a red design

CIRR

Find similar images by similarity of image features

Annotation

A dog is inside a hole, not on grass

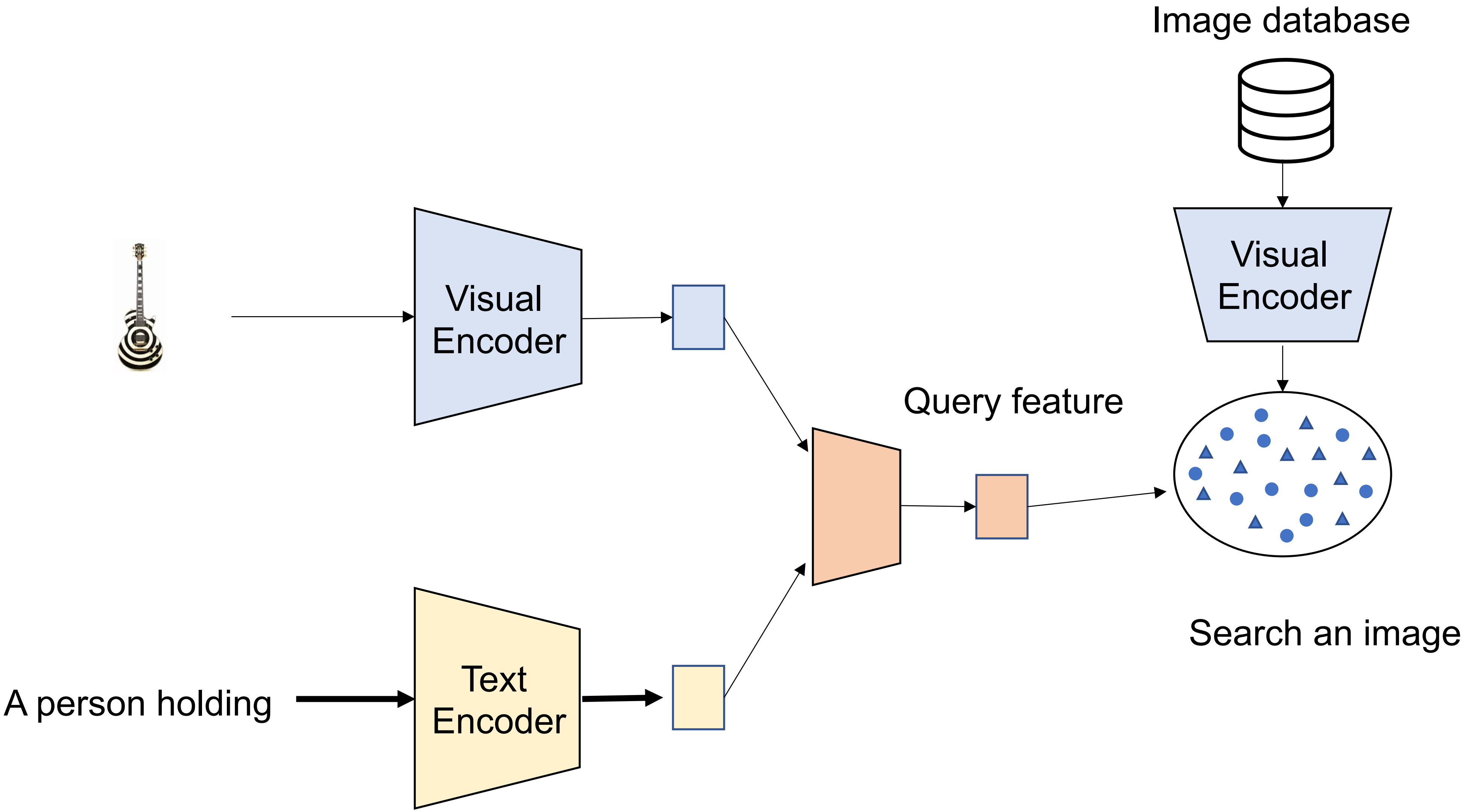
- Expensive!
- Dataset-specific bias

Zero-shot composed image retrieval

- Can we train composed image retrieval model without triplet?
  - One model applicable to diverse retrieval tasks.

- Vision-language model (CLIP) pre-trained with image-text pairs
  - Relationship between image and text

Limitation in supervised methods



[Baldrati CVPR'22]

Idea: Represent image as a word token

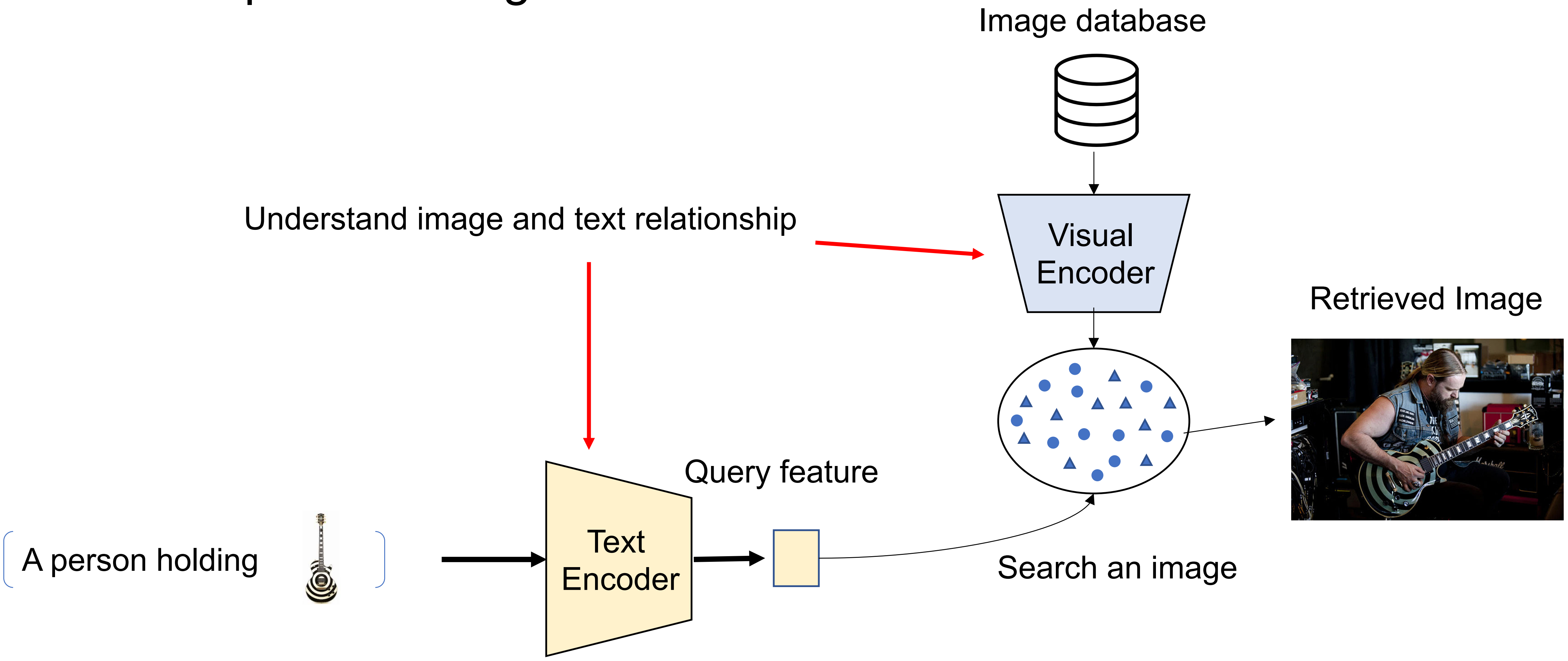


Image database

Understand image and text relationship

Visual Encoder

Retrieved Image

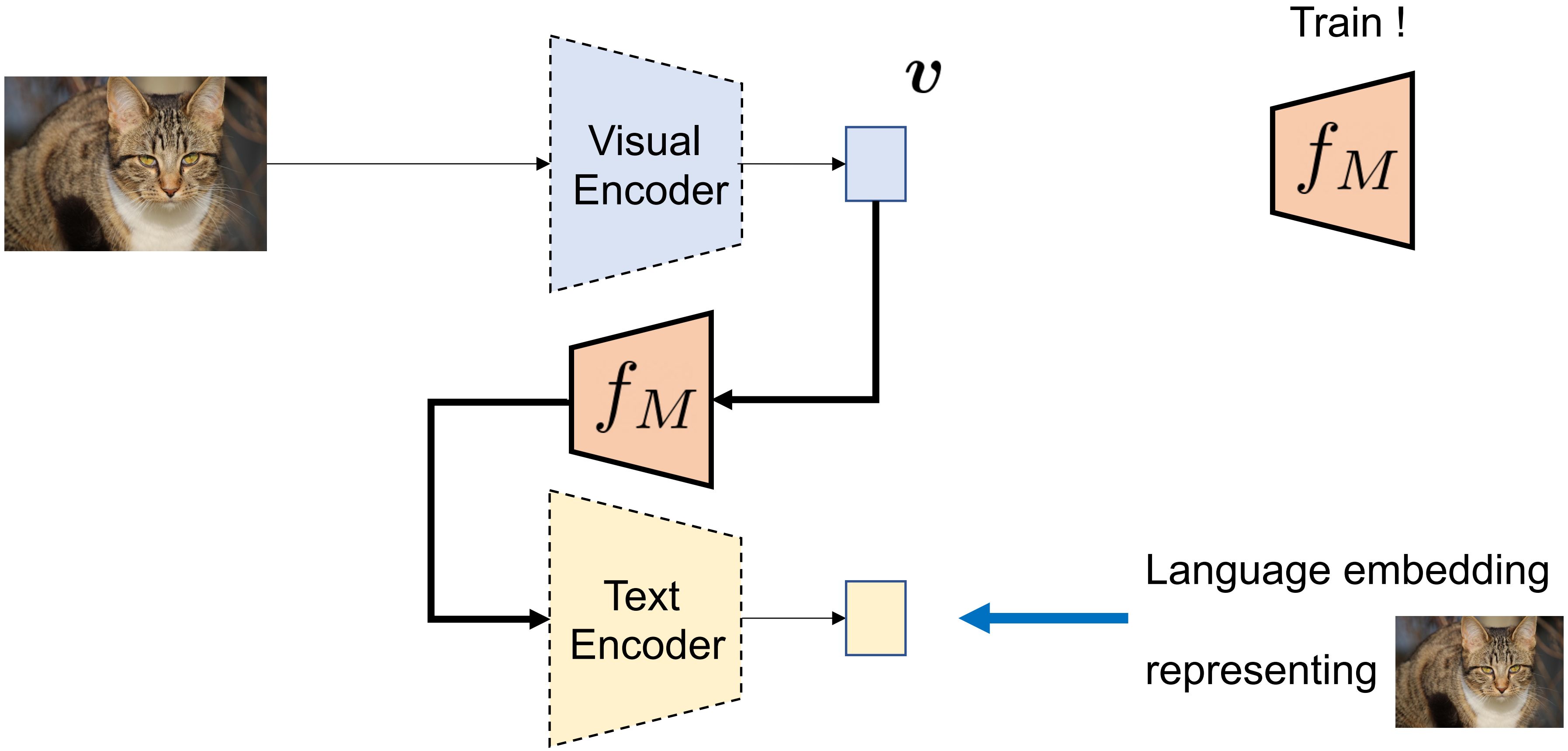
Query feature

A person holding

Text Encoder

Search an image

Mapping image into a word token



Train !

$\boldsymbol{v}$

Visual Encoder

f_M

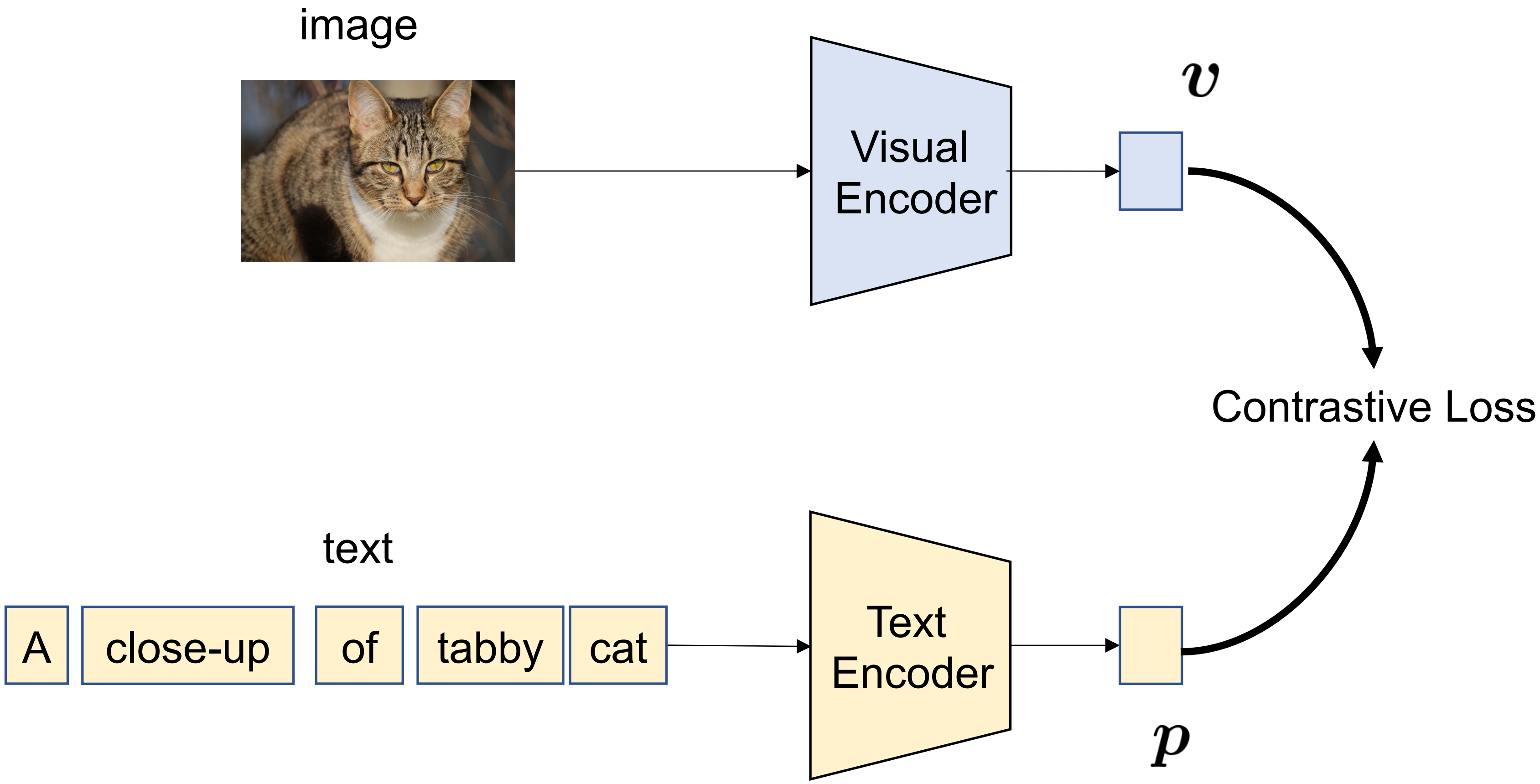
f_M

Text Encoder

Language embedding

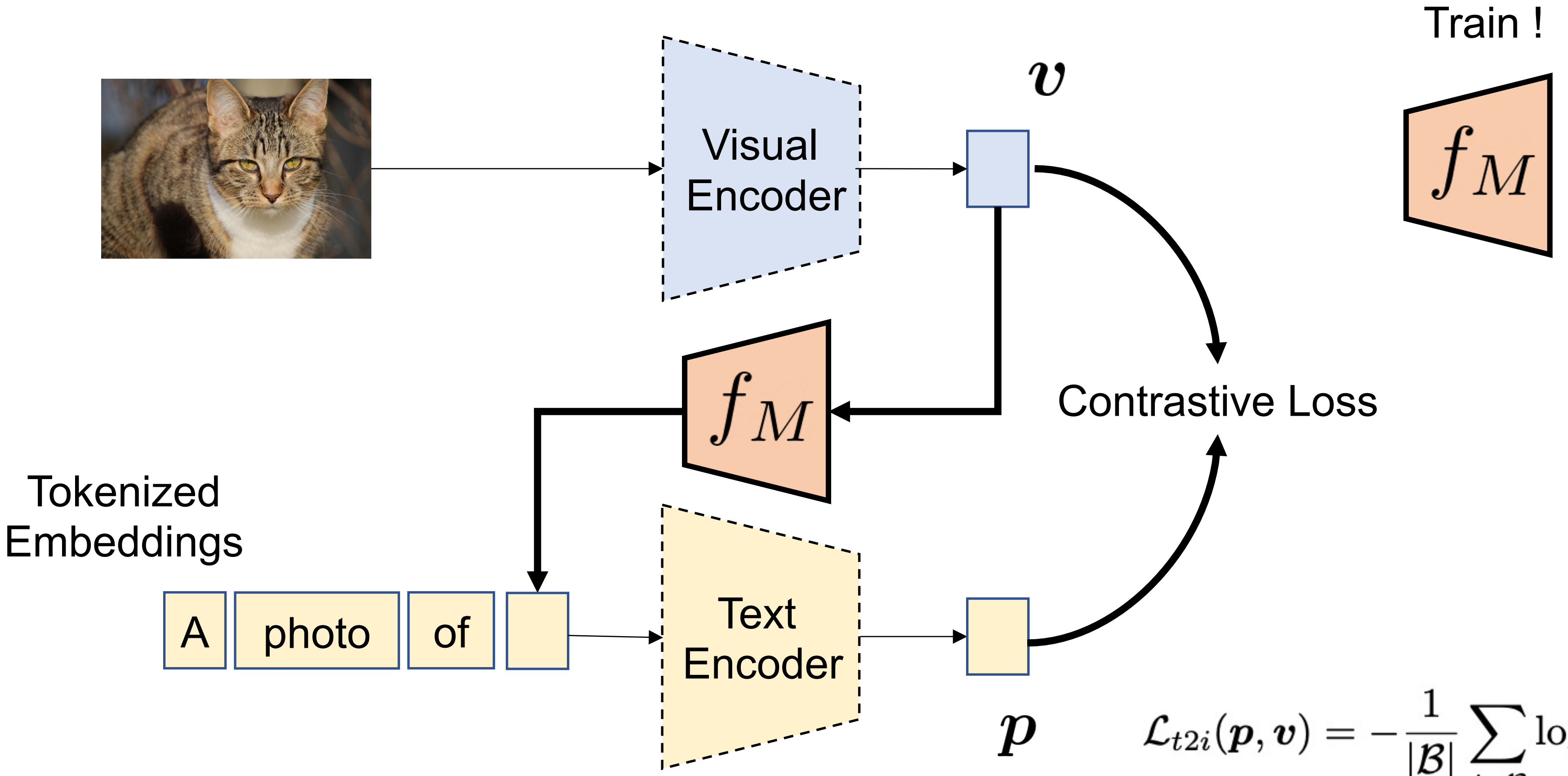
representing

How encoders are pre-trained?



Find mapping $\boldsymbol{v}$ to a word token so that contrastive loss is minimized!

Mapping network training



$$\mathcal{L}_{t2i}(\boldsymbol{p},\boldsymbol{v}) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \frac{\exp(\tau \boldsymbol{p}_i^T \boldsymbol{v}_i)}{\sum_{j\in\mathcal{B}} \exp(\tau \boldsymbol{p}_i^T \boldsymbol{v}_j)},$$

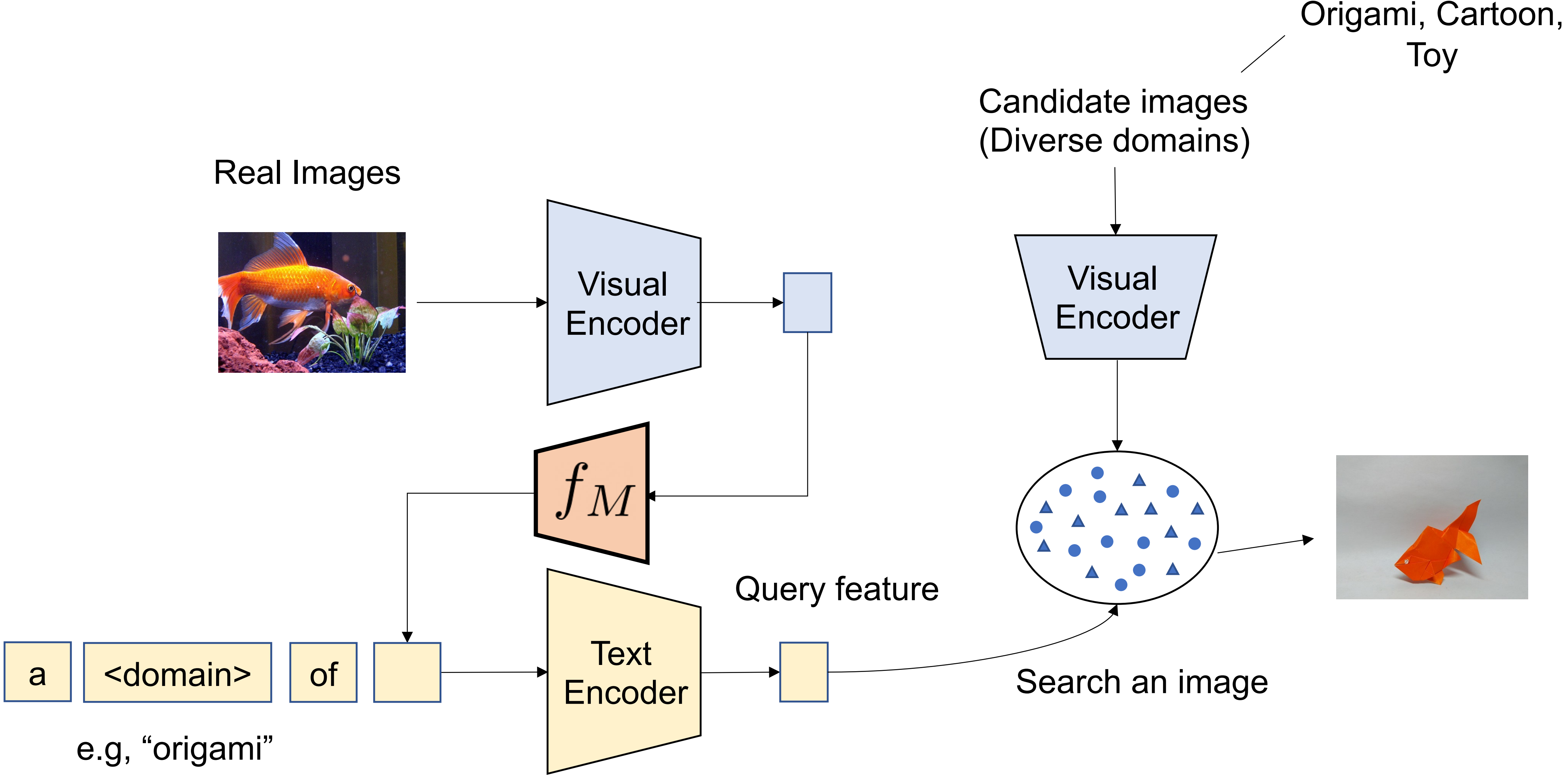
$$\mathcal{L}_{i2t}(\boldsymbol{p},\boldsymbol{v}) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \frac{\exp(\tau \boldsymbol{v}_i^T \boldsymbol{p}_i)}{\sum_{j\in\mathcal{B}} \exp(\tau \boldsymbol{v}_i^T \boldsymbol{p}_j)},$$

Training data : CC3M (Image only)

Experiments

- Converting domain (ImageNet)
- Fashion attribute manipulation (Fashion-IQ)
- Object composition manipulation (COCO)
- General composition manipulation (CIRR)

Composing with domain description



Real, Sculpture, Origami, Cartoon, Toy

Candidate images (Diverse domains)

Real Images

Visual Encoder

f_M

Visual Encoder

a

<domain>

of

e.g, “origami”

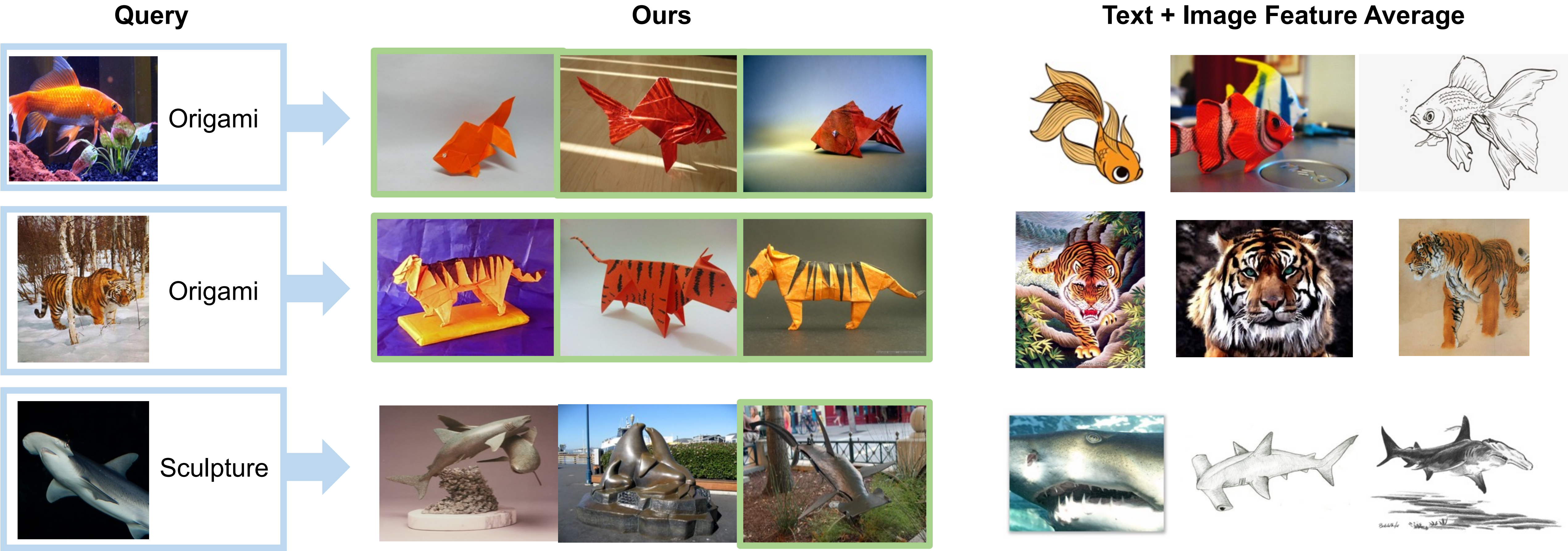
Text Encoder

Query feature

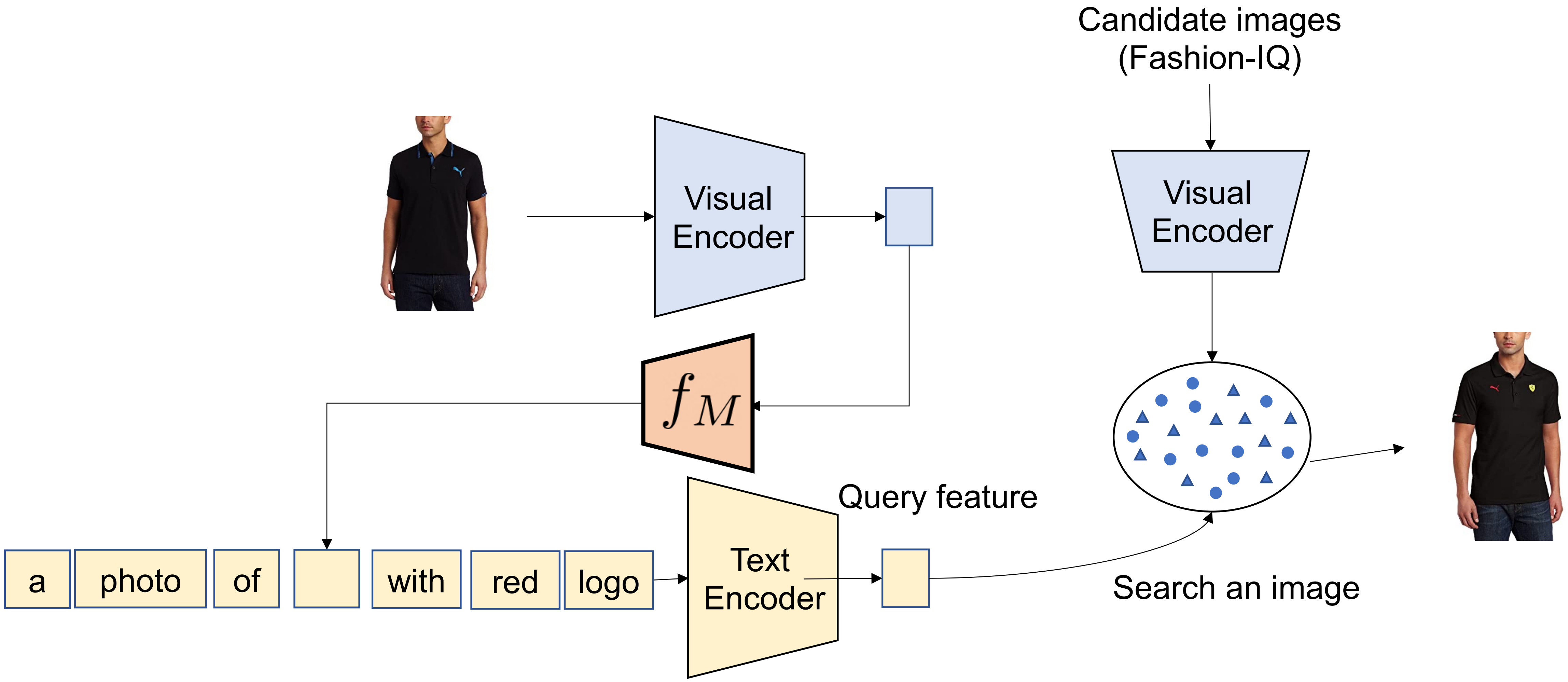
Search an image

Retrieve images with corresponding domain and class

R10, averaged over four test domains: Cartoon, Origami, Toy, Sculpture


Query
Ours
Text + Image Feature Average
Origami
Origami
Sculpture

Fashion attribute composition



Candidate images
(Fashion-IQ)

Visual Encoder

Visual Encoder

f_M

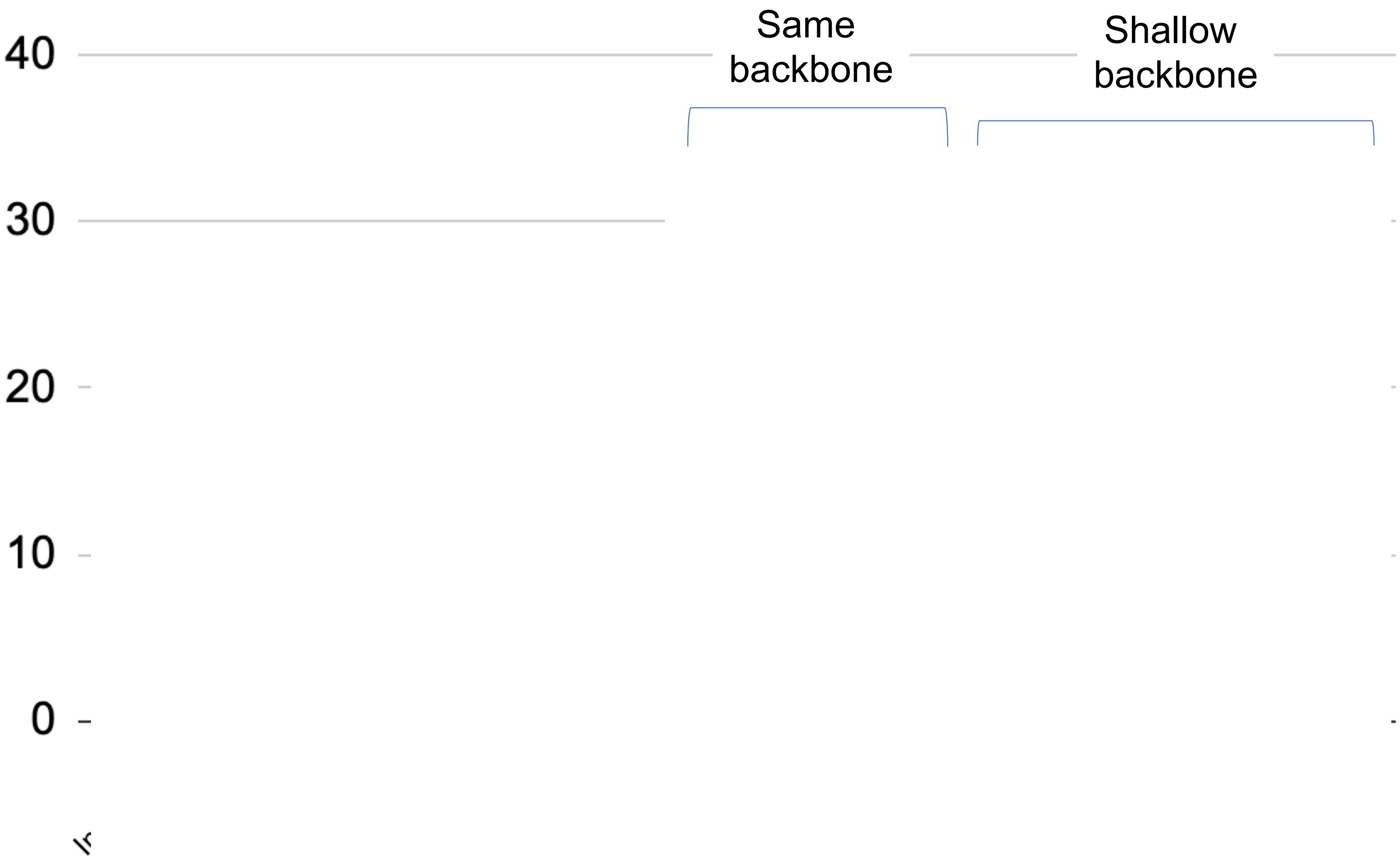
a photo of with red logo

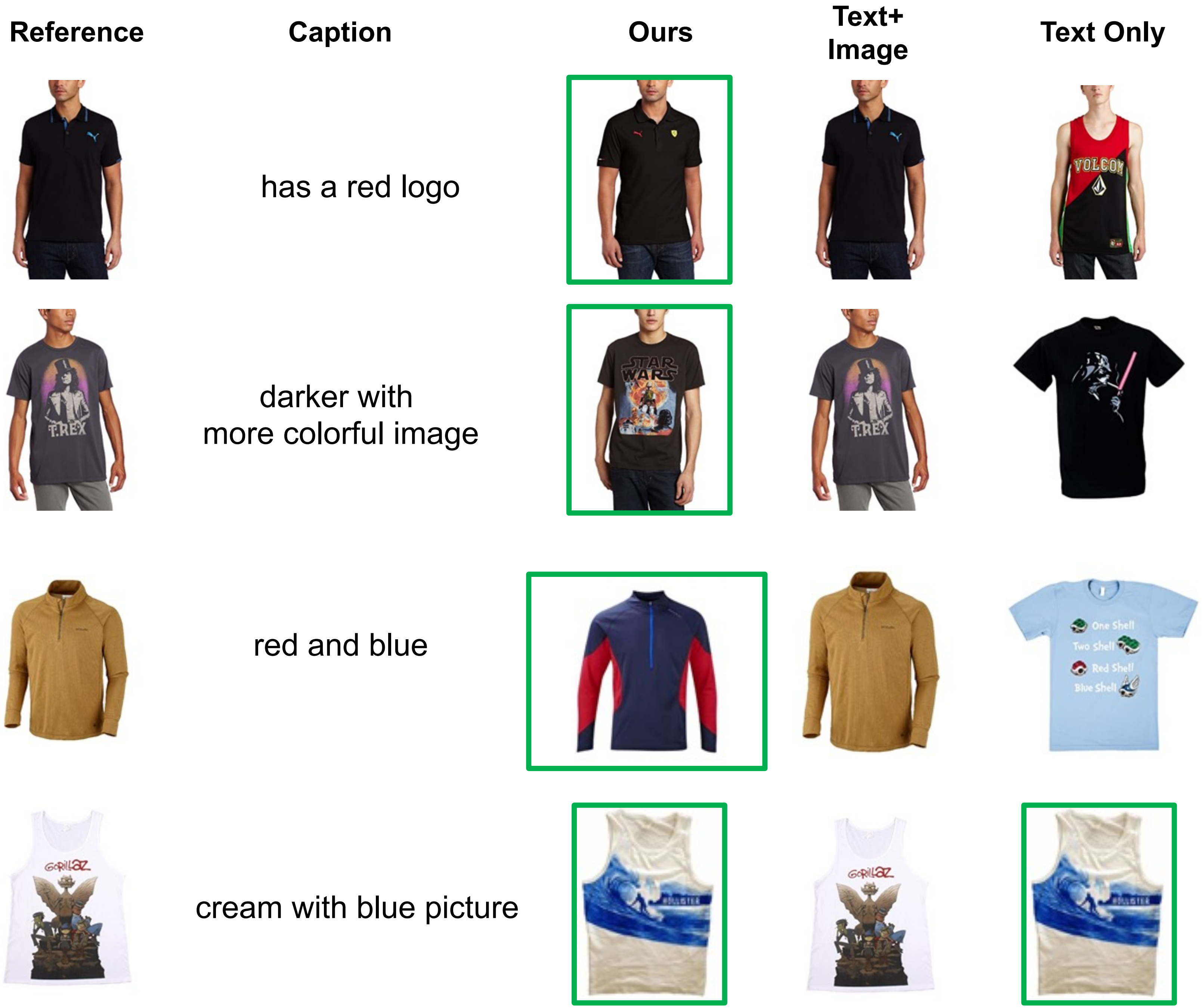
Text Encoder

Query feature

Search an image

R10, averaged over Dress, Shirt, TopTee





Reference

Caption

Ours

Text+ Image

Text Only

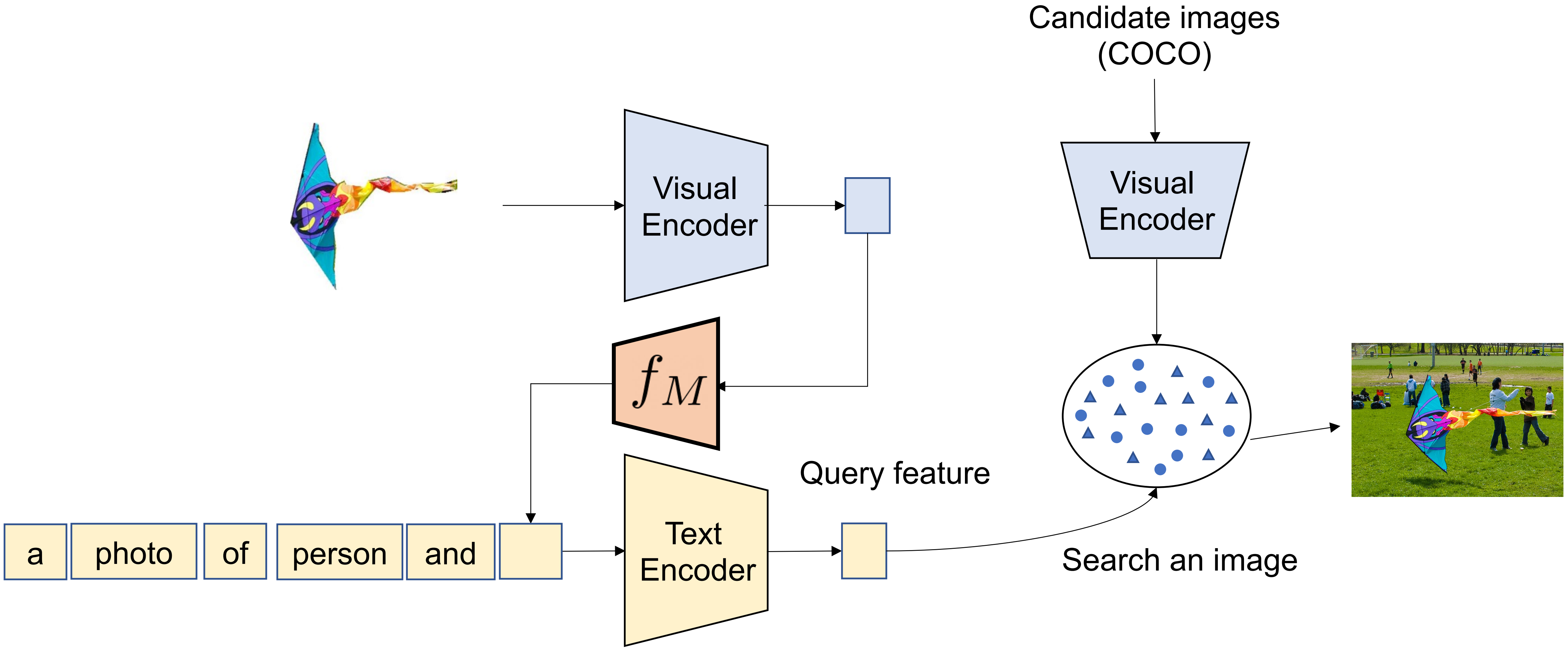
has a red logo

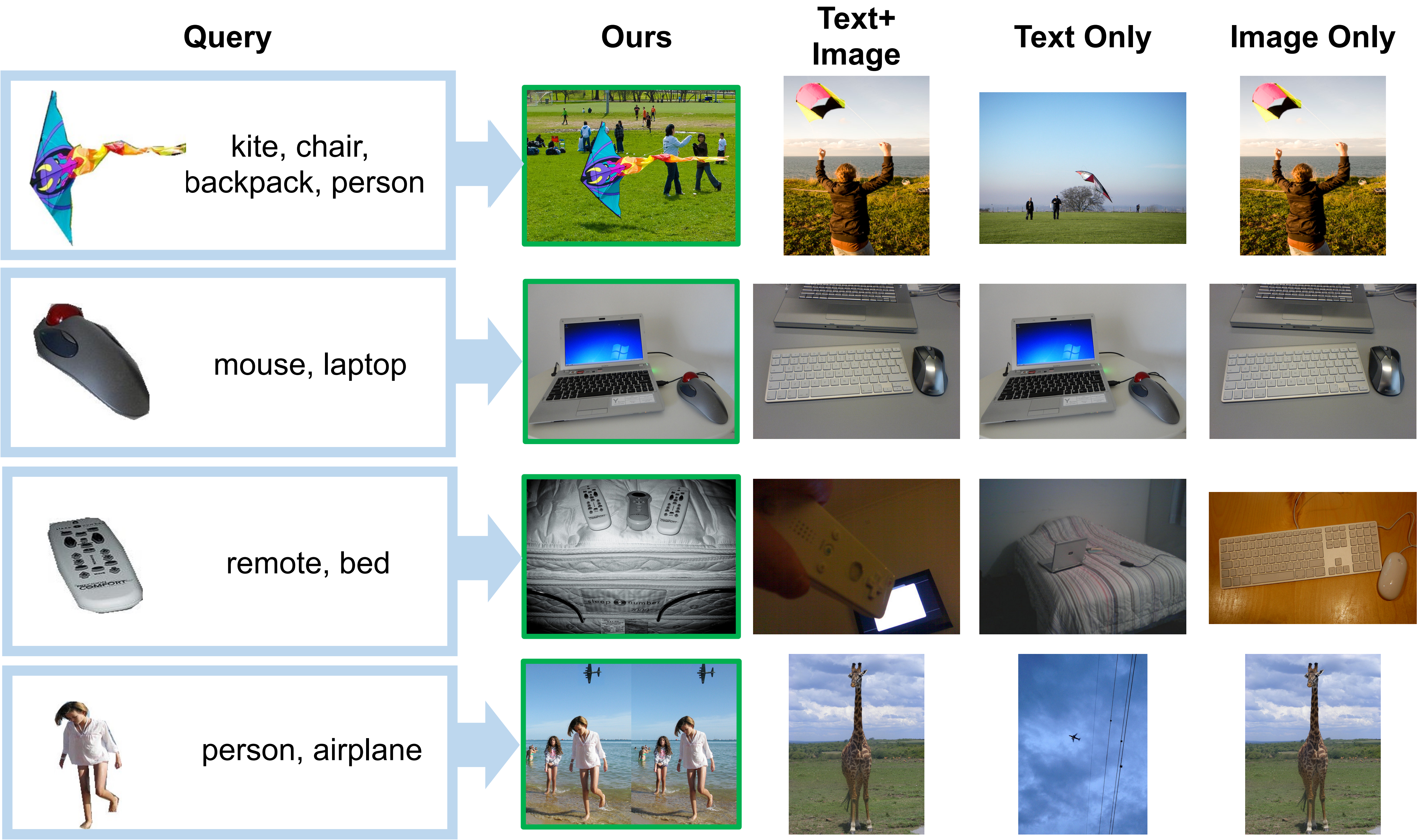
darker with more colorful image

red and blue

cream with blue picture

Composing with object classes





Query

Ours

Text+ Image

Text Only

Image Only

kite, chair, backpack, person

mouse, laptop

remote, bed

person, airplane

Ours

Text + Image feature averaging

Summary

- Zero-shot composed image retrieval
  - Image-Text pairs are enough for composed image retrieval
- Can we utilize image-text pair to training mapping network?