NAVER LABS
Europe

ARTEMIS

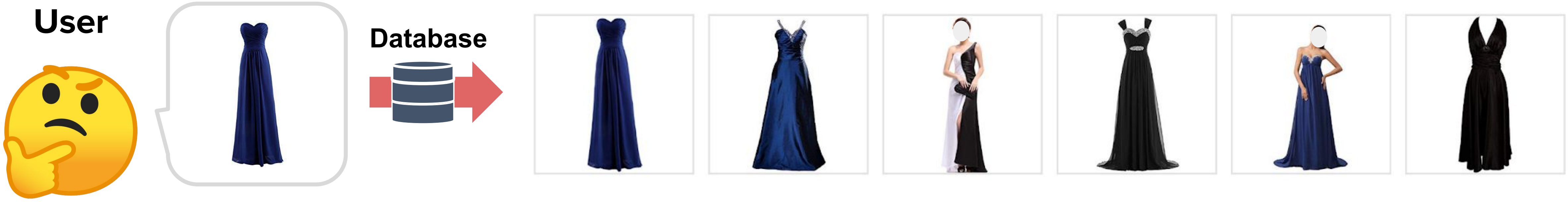
Attention-based **R**etrieval with **T**ext-**E**xplicit **M**atching and **I**mplicit **S**imilarity

Ginger Delmas, Rafael S. Rezende, Gabriela Csurka, Diane Larlus

ICLR

How to query for an image in a search engine?

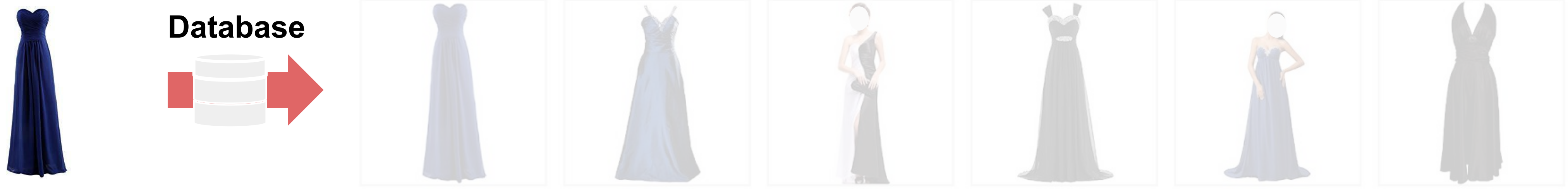
Query with an image? Visual image retrieval problem



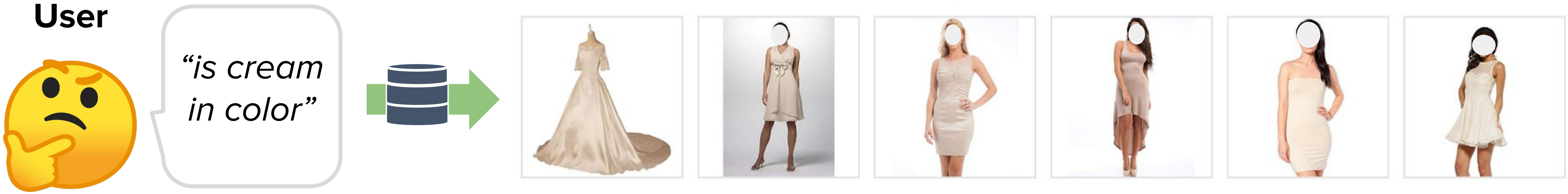
How to query for an image in a search engine?

Query with an image? Visual image retrieval problem

Query with text? Cross-modal image retrieval problem



Database



User

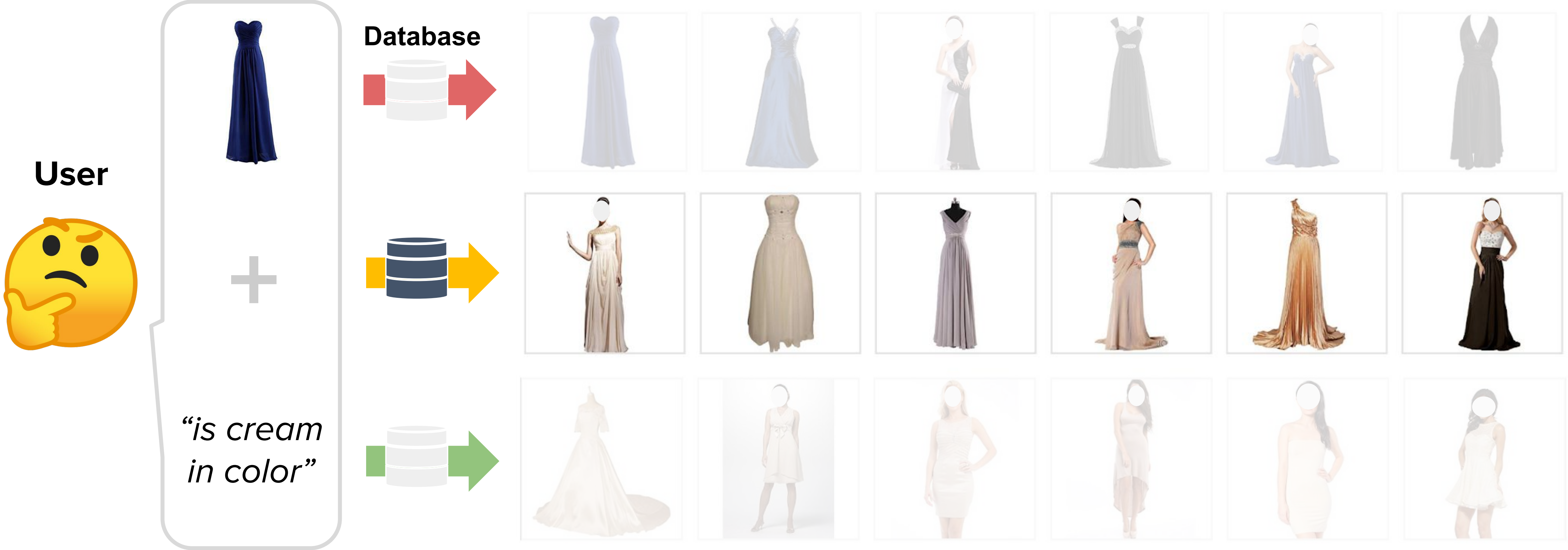
“is cream in color”

How to query for an image in a search engine?

Query with an image? Visual image retrieval problem

Query with text? Cross-modal image retrieval problem

Query with an image and text? Image search with text modifiers



User

Database

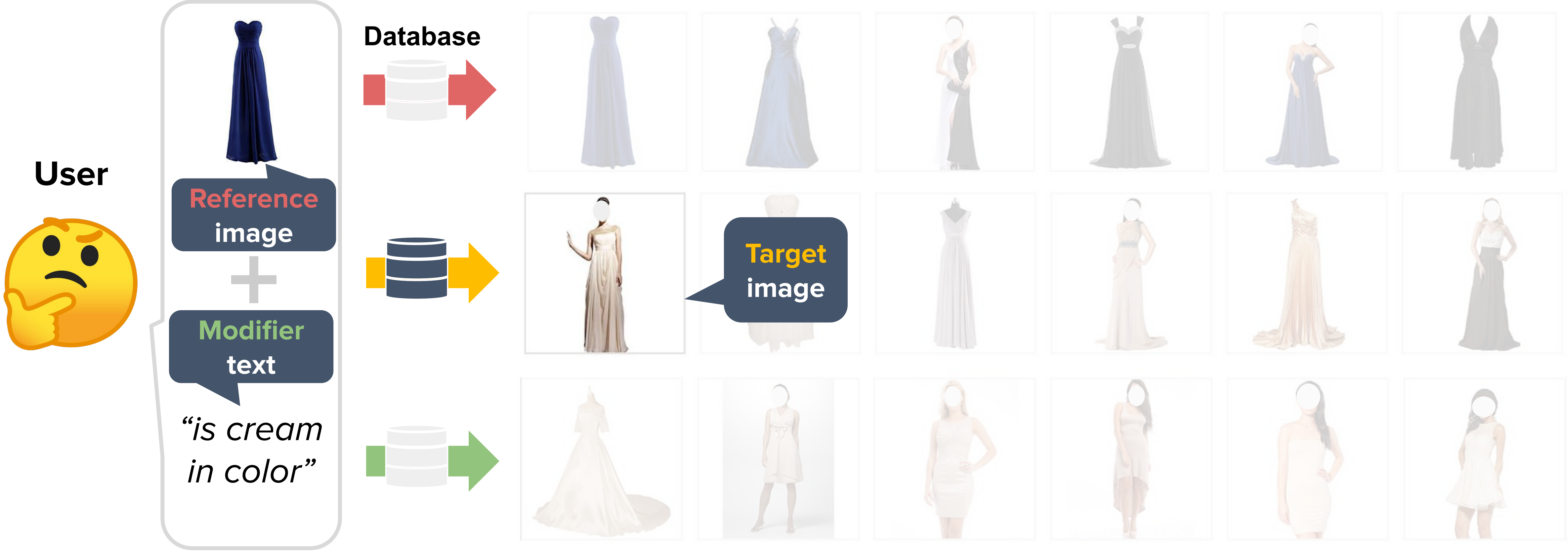
“is cream in color”

How to query for an image in a search engine?

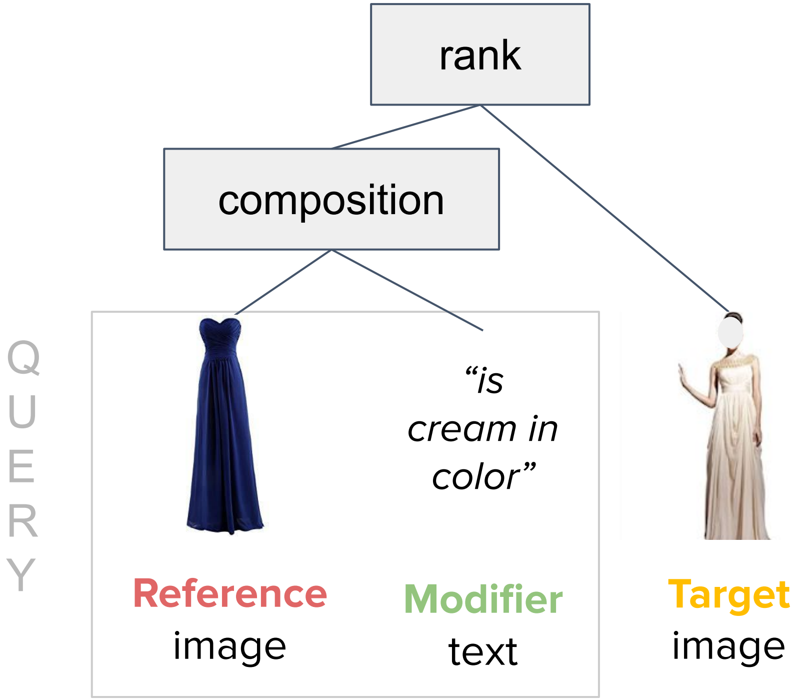
Query with an image? Visual image retrieval problem

Query with text? Cross-modal image retrieval problem

Query with an image and text? Image search with text modifiers.



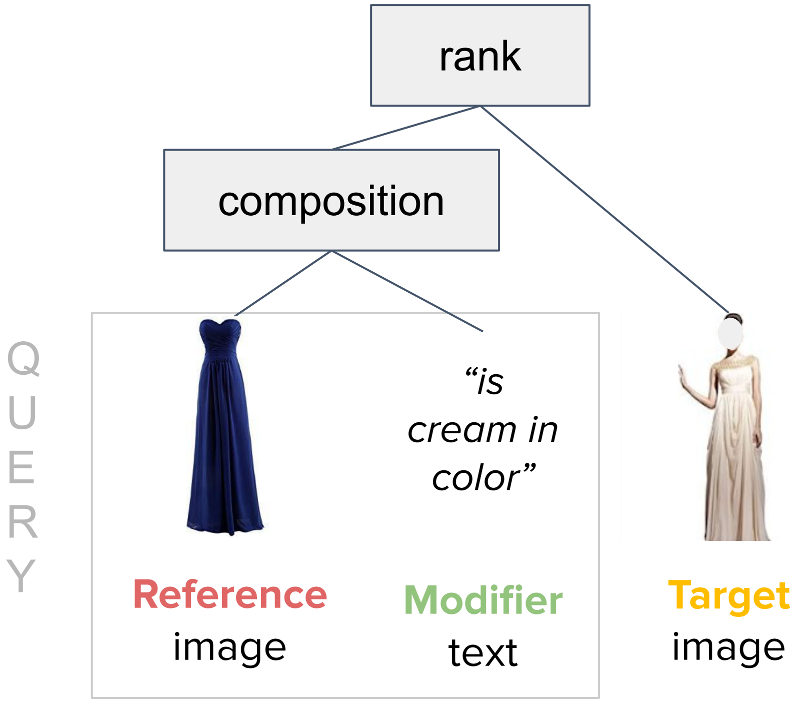
Prior work



Previous works doing composition

VAL, Chen et al. [CVPR 20]
CoSMo, Lee et al. [CVPR 21]
TIRG, Vo et al.[CVPR 19]

Prior work



rank

composition

Q U E R Y

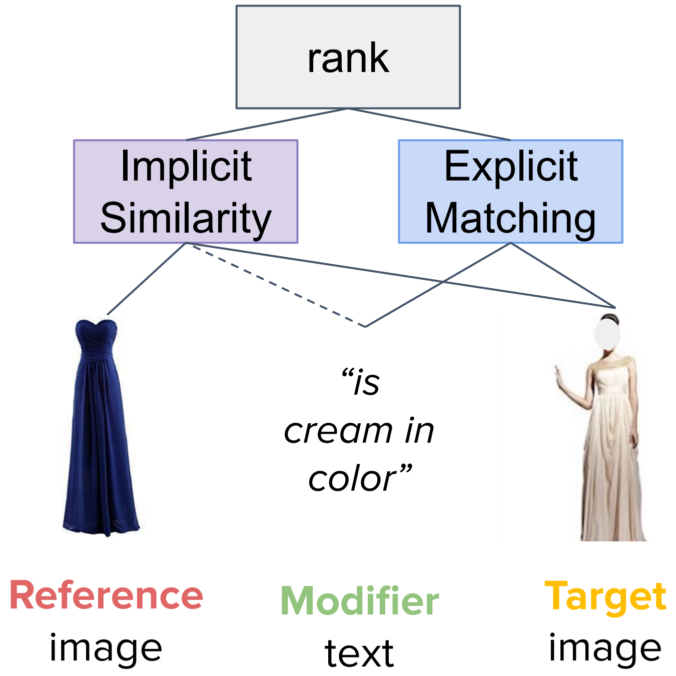
“is cream in color”

Reference image

Modifier text

Target image

Our approach



rank

Implicit Similarity

Explicit Matching

“is cream in color”

Reference image

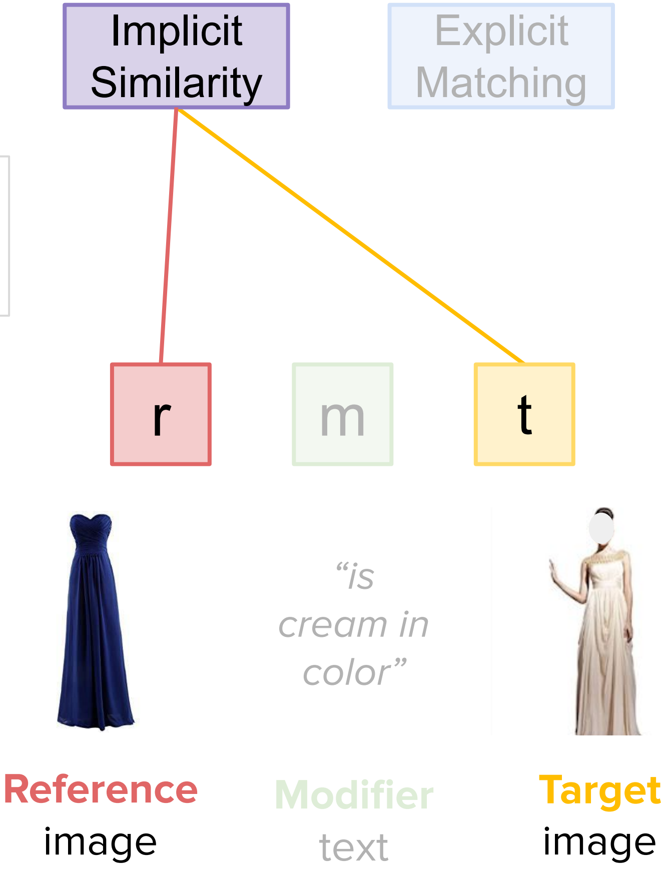
Modifier text

Target image

Implicit Similarity (IS)

Image similarity

$$r^{T} \qquad t$$



Implicit Similarity

Explicit Matching

r

m

t

"is cream in color"

Reference image

Modifier text

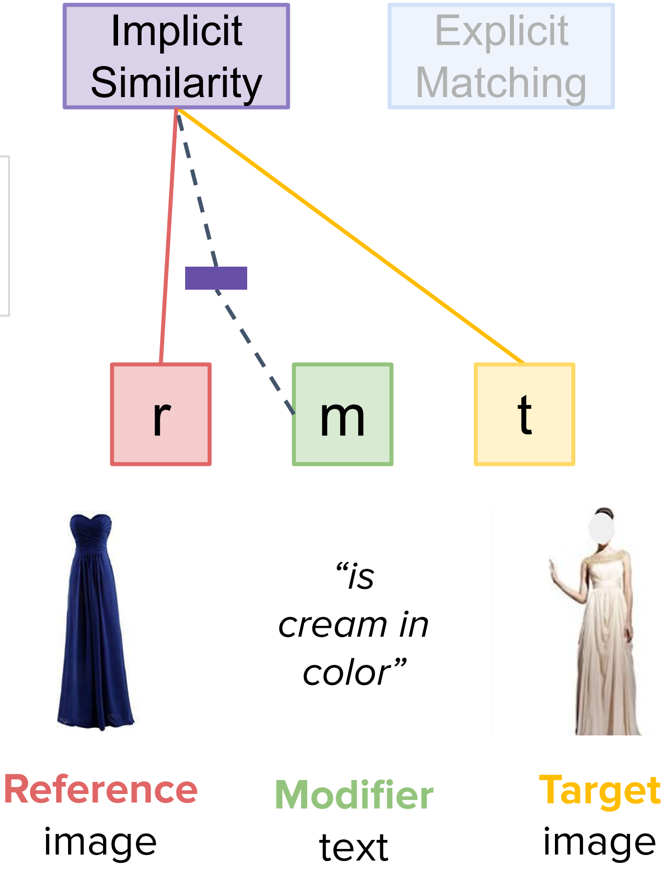
Target image

Implicit Similarity (IS)

Image similarity guided by text

*Lightweight attention
on visual-visual*

$$(A_{IS}[m]\odot r)^T\,(A_{IS}[m]\odot t)$$



Implicit Similarity

Explicit Matching

“is cream in color”

Reference image

Modifier text

Target image

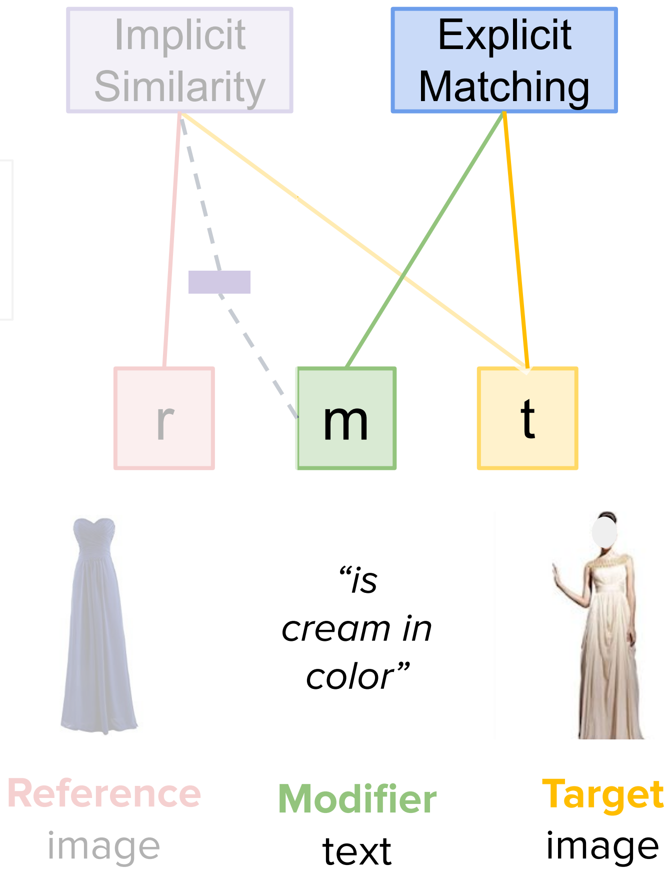
A_{IS} text-guided attention (MLP)

Explicit Matching (EM)

Image similarity guided by text

Lightweight attention on visual-visual

$$(A_{IS}[m] \odot r)^T (A_{IS}[m] \odot t)$$



cross-modal retrieval

$$m^T t$$

A_{IS} text-guided attention (MLP)

Explicit Matching (EM)

Image similarity guided by text

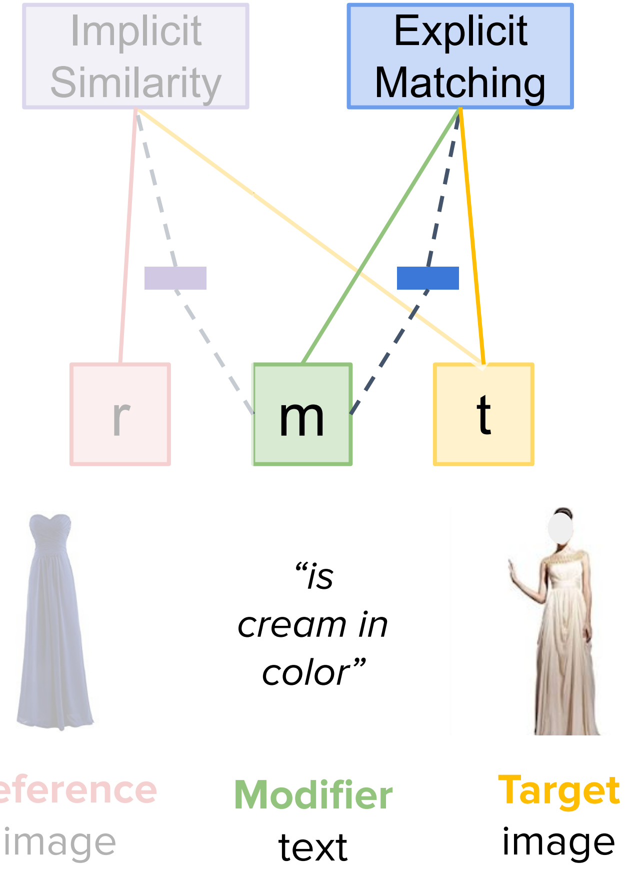
Lightweight attention on visual-visual

$(A_{IS}[m] \odot r)^T (A_{IS}[m] \odot t)$

Text-modified cross-modal retrieval

Lightweight attention on textual-visual

$m^T (A_{EM}[m] \odot t)$



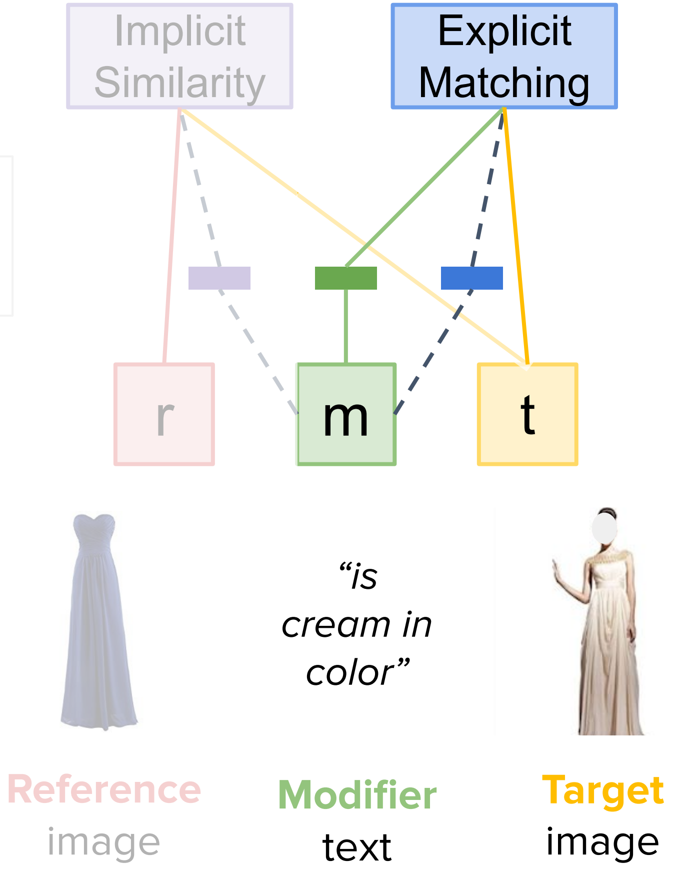
A_{IS}, A_{EM} text-guided attentions (MLP)

Explicit Matching (EM)

Image similarity guided by text

*Lightweight attention
on visual-visual*

$$(A_{IS}[m]\odot r)^T\ (A_{IS}[m]\odot t)$$



***Text-modified
cross-modal retrieval***

*Lightweight attention
on textual-visual*

$$(Tr[m])^T\ (A_{EM}[m]\odot t)$$

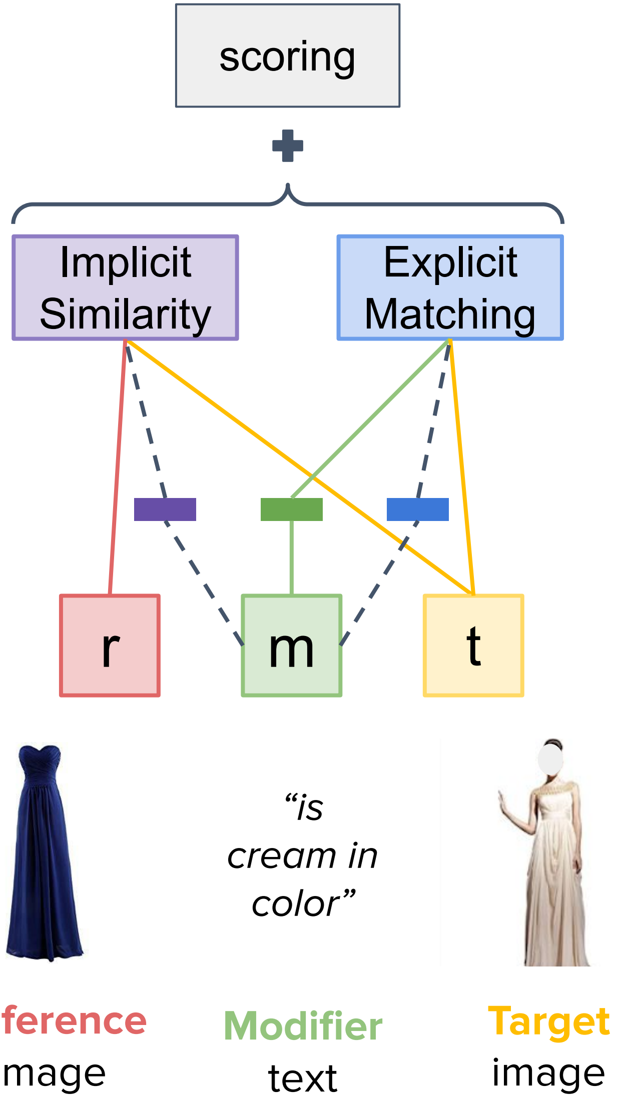
A_{IS}, A_{EM}	text-guided attentions (MLP)
Tr	linear transformation (FC)

ARTEMIS: EM + IS

Image similarity guided by text

*Lightweight attention
on visual-visual*

$$(A_{IS}[m]\odot r)^T\ (A_{IS}[m]\odot t)$$



scoring

Implicit Similarity

Explicit Matching

r

m

t

"is cream in color"

Reference image

Modifier text

Target image

***Text-modified
cross-modal retrieval***

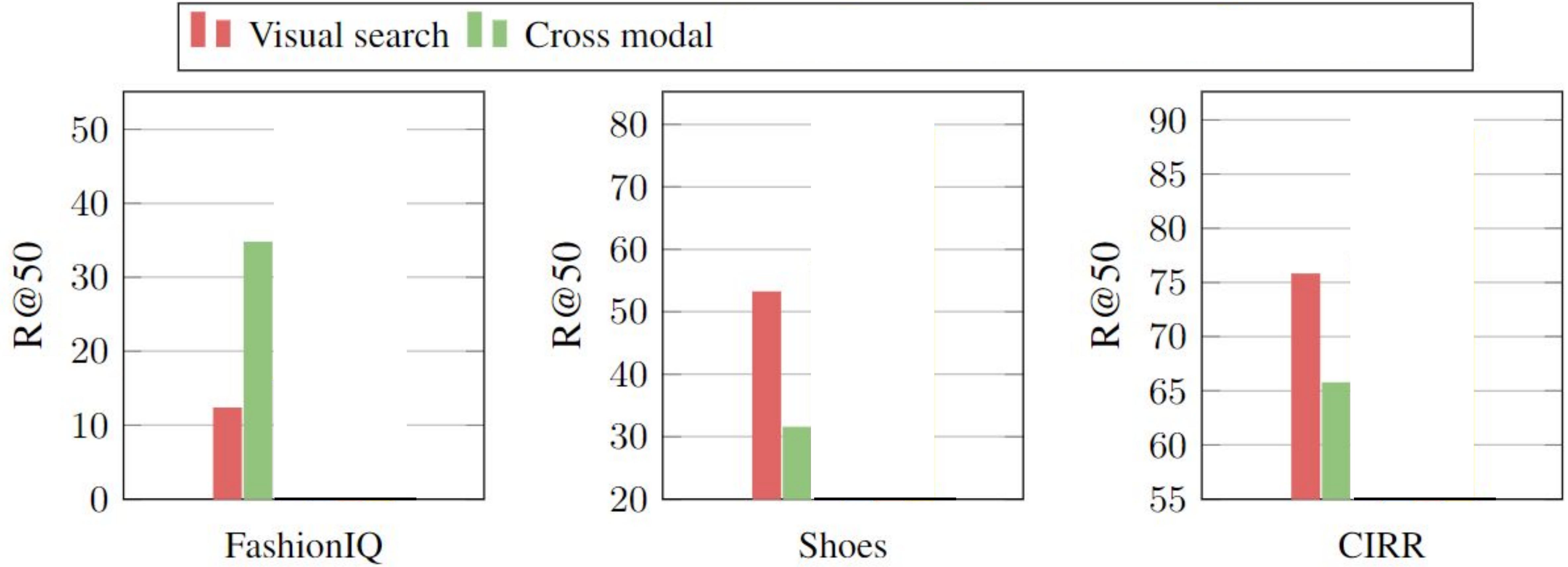
*Lightweight attention
on textual-visual*

$$(Tr[m])^T\ (A_{EM}[m]\odot t)$$

A_{IS}, A_{EM} text-guided attentions (MLP)
Tr linear transformation (FC)

Evaluation

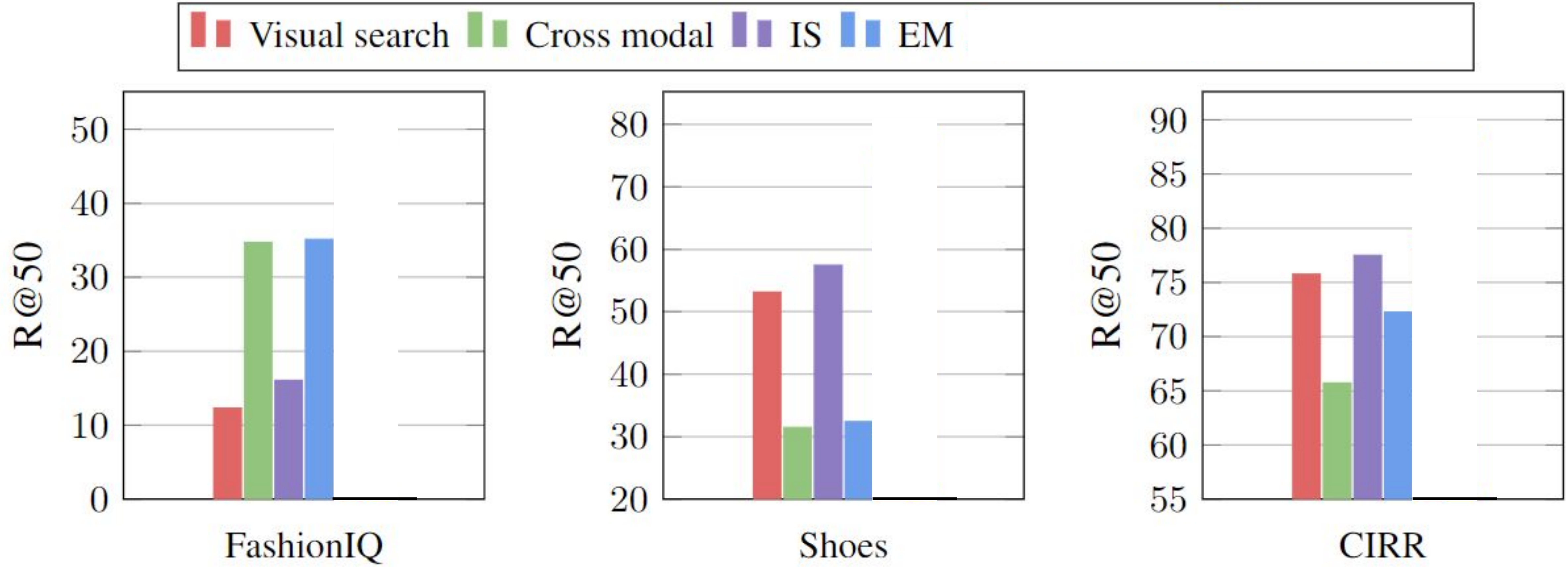
We evaluate on *text-centric* (FashionIQ) and *image-centric* (Shoes, CIRR) datasets.



Ablation

Implicit similarity > Visual search

Explicit matching > Cross modal

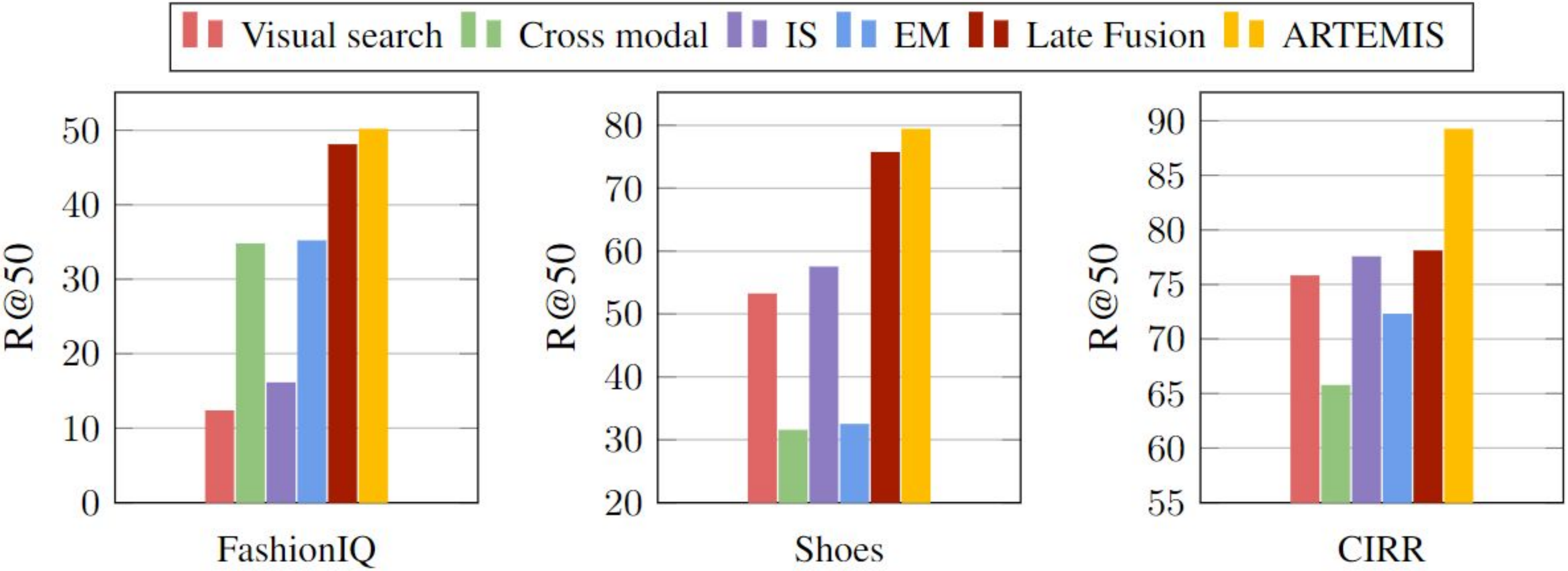


ARTEMIS evaluation

Implicit similarity > **Visual search**

Explicit matching > **Cross modal**

ARTEMIS $_{(IS\ +\ EM)}$ > **Late fusion** $_{(Visual\ search\ +Cross\ modal)}$ thanks to the attention modules.



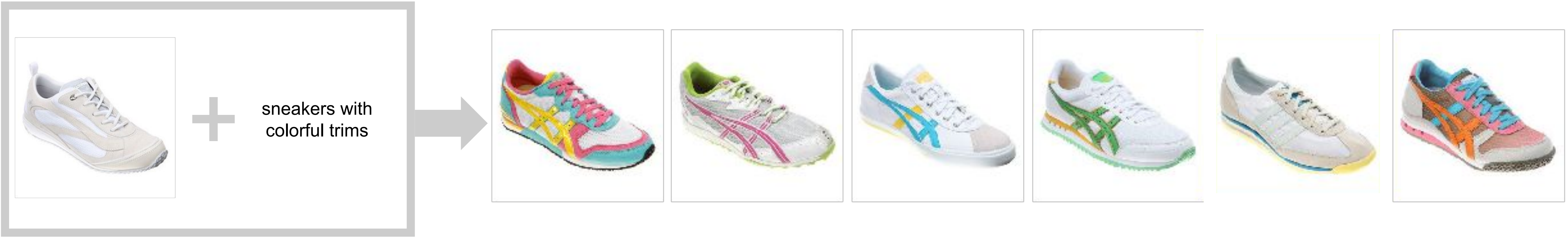
Qualitative results

Explicit Matching

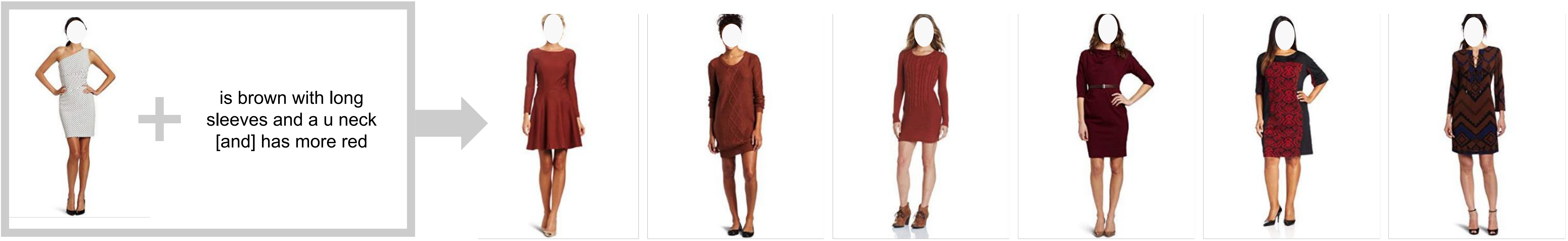
Color, length, form of the neck, color trims

Implicit Similarity

Dress length, style, shape, category, laces



sneakers with colorful trims



is brown with long sleeves and a u neck [and] has more red

Take-home message

- **ARTEMIS** makes **cross-modal and visual search** scoring strategies **compatible** for image search with text modifiers.

- **ARTEMIS** models all **pairwise interactions** including with the target image, **without large extra-cost**.

- **ARTEMIS** is **versatile**: it works with different visual and textual encoders and different domains.

Take-home message

- **ARTEMIS** makes **cross-modal and visual search** scoring strategies **compatible** for Image search with text modifiers.

- **ARTEMIS** models all **pairwise interactions** including with the target image, **without large extra-cost**.

- **ARTEMIS** is **versatile**: it works with different visual and textual encoders and different domains.

Take-home message

- **ARTEMIS** makes **cross-modal and visual search** scoring strategies **compatible** for Image search with text modifiers.

- **ARTEMIS** models all **pairwise interactions** including with the target image, **without large extra-cost**.

- **ARTEMIS** is **versatile**: it works with different visual and textual encoders, and different domains.

NAVER LABS
Europe

ARTEMIS

Attention-based **R**etrieval with **T**ext-**E**xplicit **M**atching and **I**mplicit **S**imilarity

Thank you!

Ginger Delmas, Rafael S. Rezende, Gabriela Csurka, Diane Larlus

ICLR