Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu

Multi-Modal Retrieval

- Multi-modal retrieval focuses more on relevance modeling between queries and documents, single/cross modality matching, and modality routing.

Cross-Modality Retrieval （Text/Image Matching）

Query: A woman wearing a net on her head cutting a cake.

Retrieval Candidates:

Image Retrieval ⇧

Text Retrieval ⇩

Query:

Retrieval Candidates:

Text1: A woman wearing a net on head cutting a cake.

Text2: A baker woman preparing bread dough on a tray with wax paper.

Multi-Modal Retrieval （Relevance Modeling）

Query: What water-related object is sitting in front of the Torre del Reloj?

Retrieval Candidates:

Image1

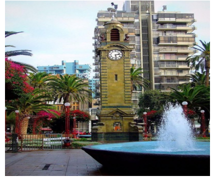


Image2

Image3

Text1: The Torre del Reloj Spanish is the main city gate of the historic center of Cartagena de Indias.

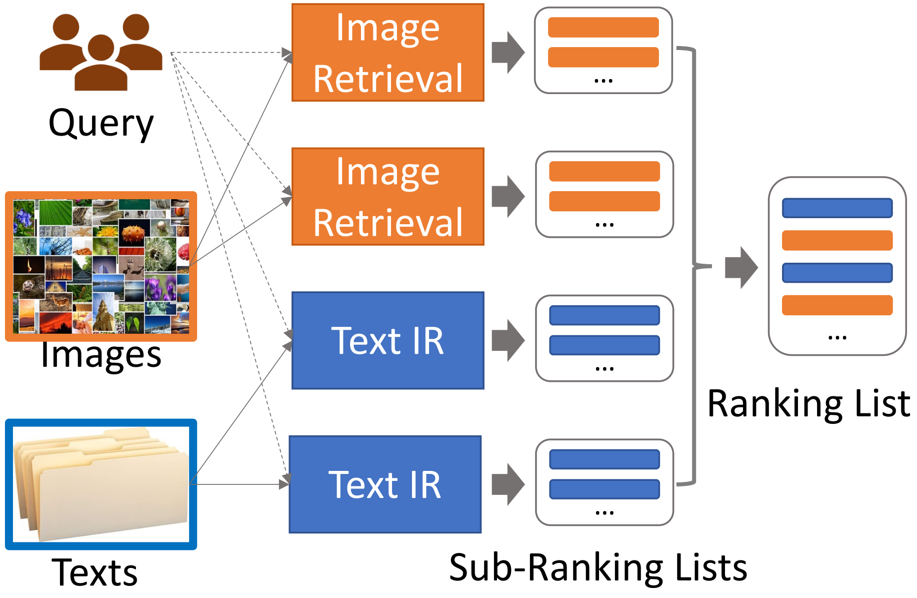
Text2: The Torre del Reloj is the clock tower, known as Arquillo Clock, and is one of the most emblematic buildings of Chiclana.

Text3: Other landmarks in the city include the Torre del Reloj (Clock Tower).

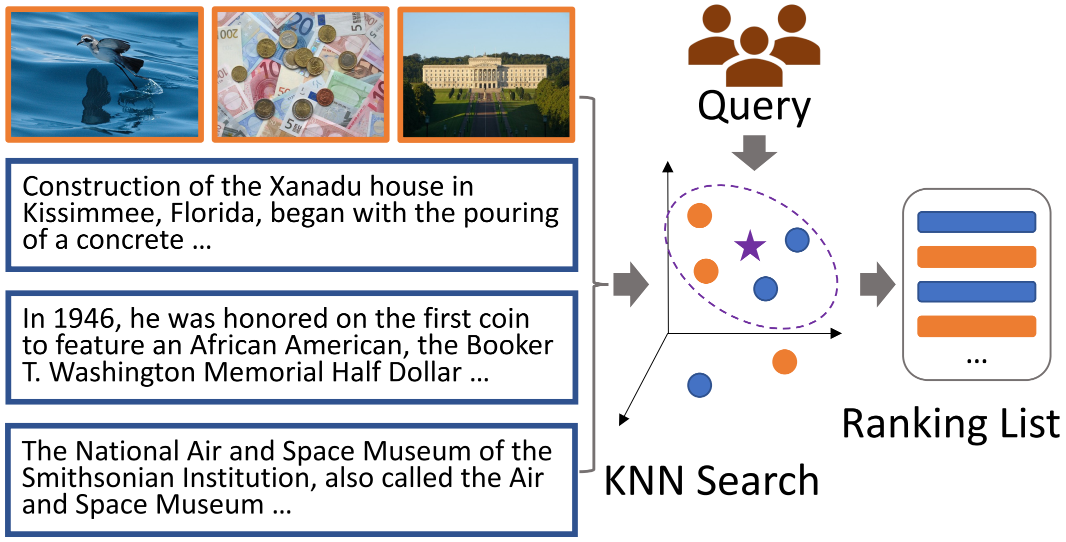
Multi-Modal Retrieval

- Divide-and-Conquer: Retrieving documents from multi-modalities and then fusion the retrieval results
- Universal Vision-Language Dense Retrieval: Leaning one universal embedding space for multi-modal retrieval, which unifies the relevance modeling and fusion stages

Divide-and-Conquer Retrieval



Universal Vision-Language Dense Retrieval

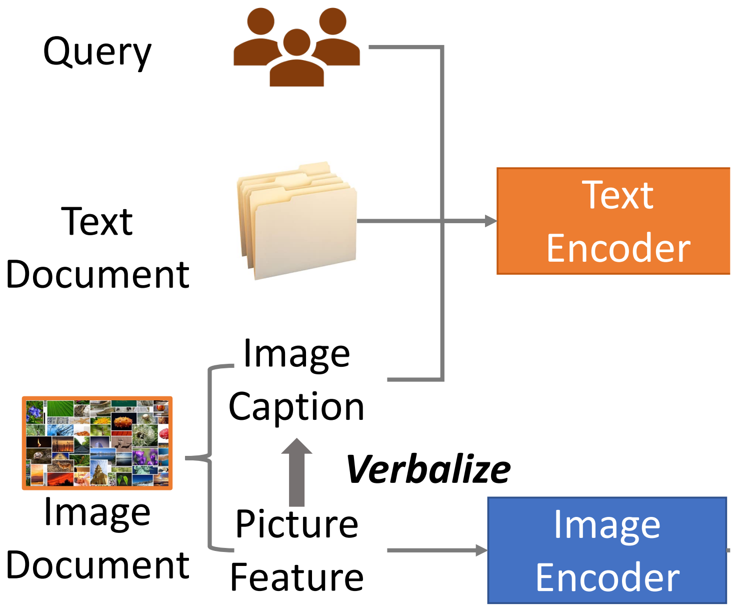

Construction of the Xanadu house in Kissimmee, Florida, began with the pouring of a concrete …

In 1946, he was honored on the first coin to feature an African American, the Booker T. Washington Memorial Half Dollar …

The National Air and Space Museum of the Smithsonian Institution, also called the Air and Space Museum …

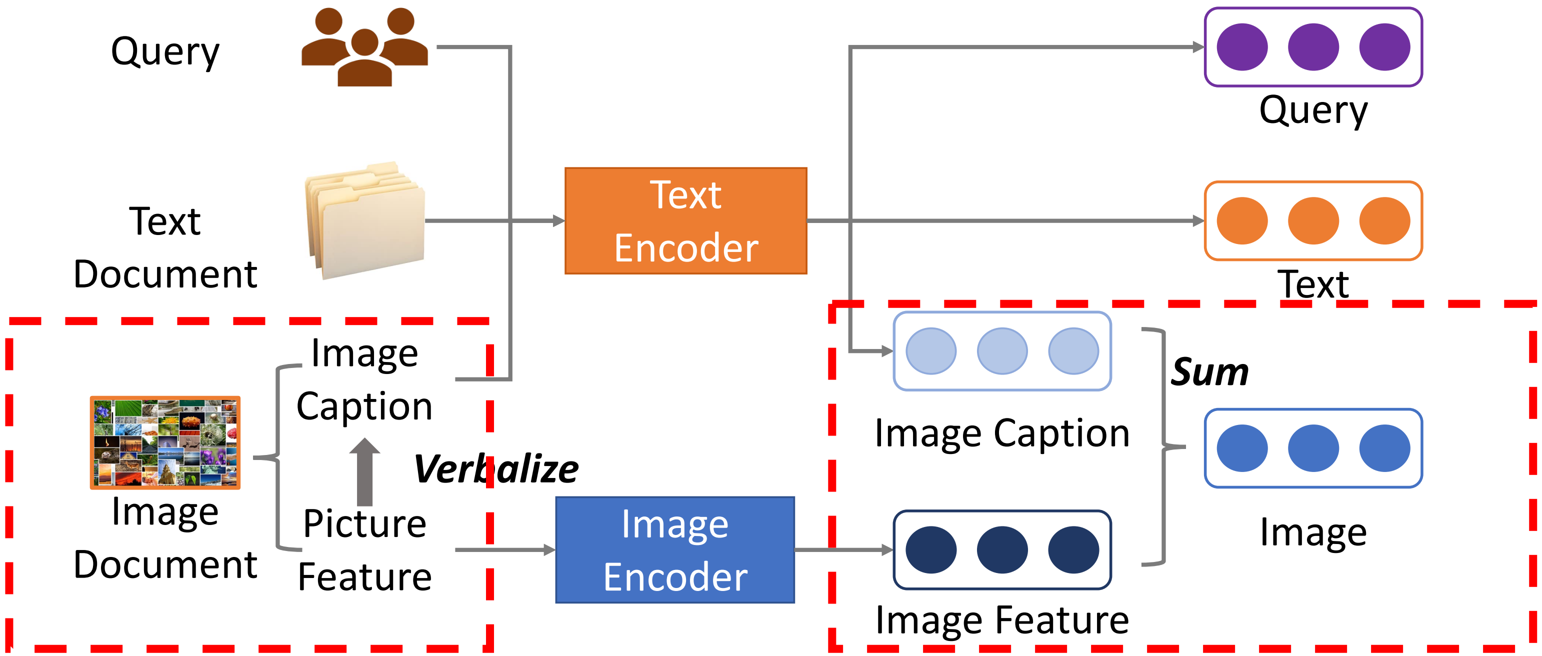
Universal Vision-Language Dense Retrieval (UniVL-DR)

- We use one Text Encoder to encode queries, text documents and image captions
- The Image Encoder is employed to encode image features as low-dimensional embeddings



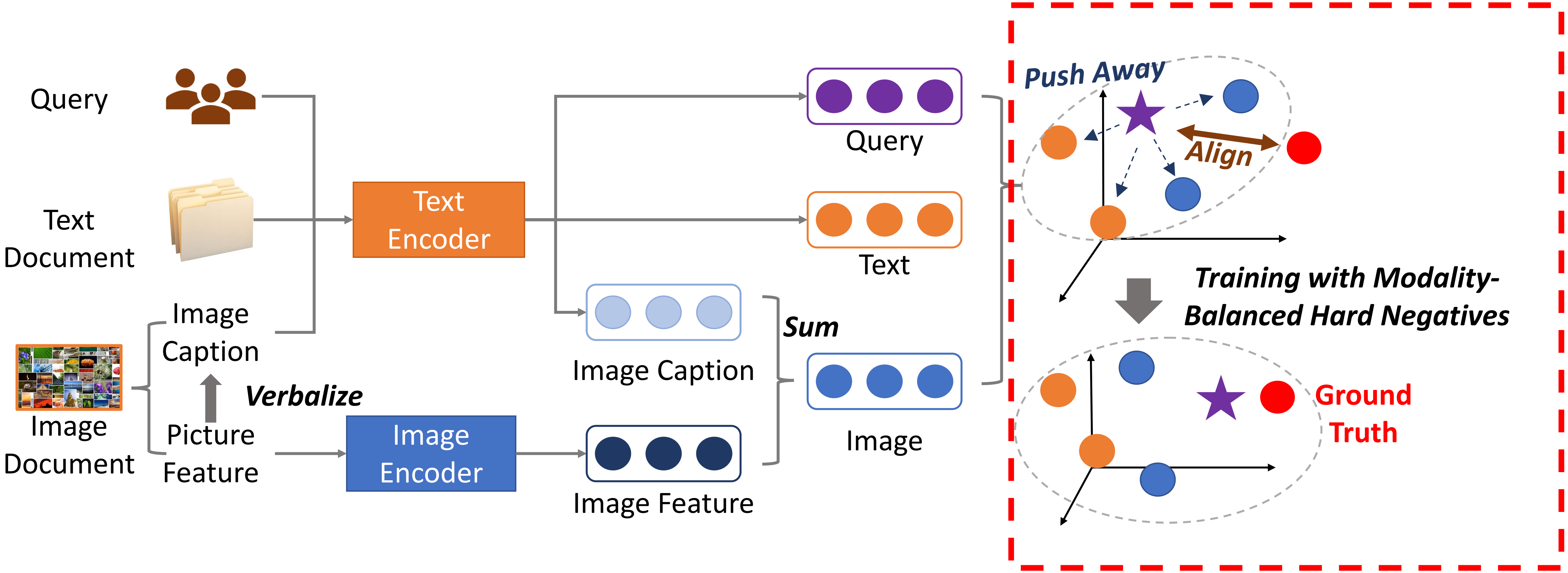
Universal Vision-Language Dense Retrieval (UniVL-DR)

- We enhance the representations of image documents by summing the representations of image captions and image features
- The image features are verbalized to enhance image documents in the raw text space



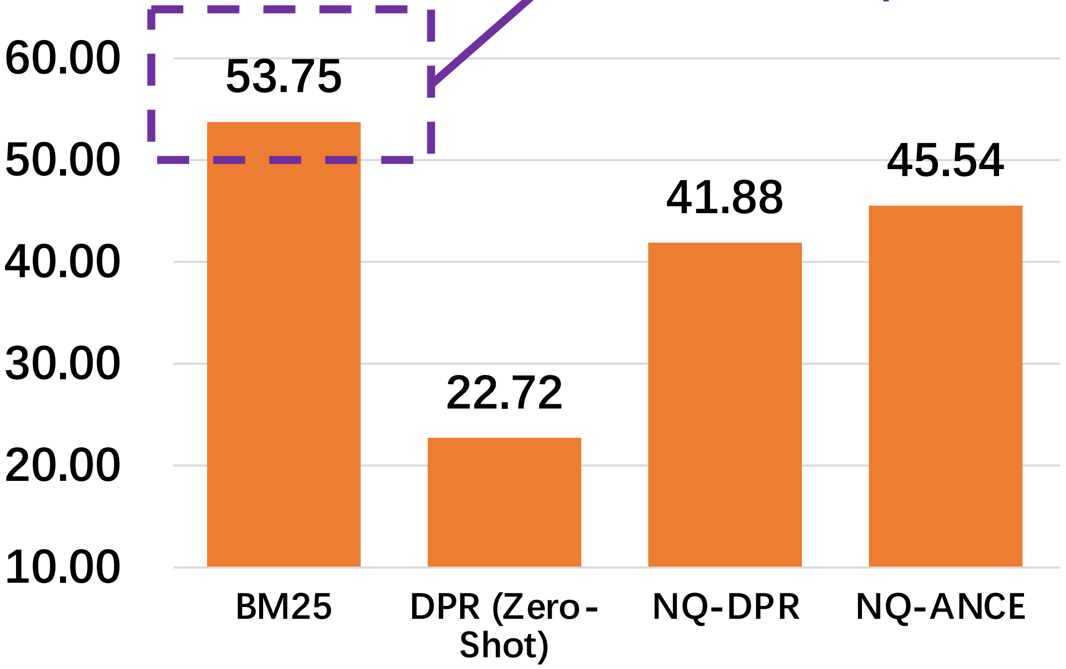
Universal Vision-Language Dense Retrieval (UniVL-DR)

- The queries, text documents and image documents are mapped in one universal embedding space
- UniVL-DR designs a modality-balanced hard negative training strategies to train retrieval models



Overall Performance

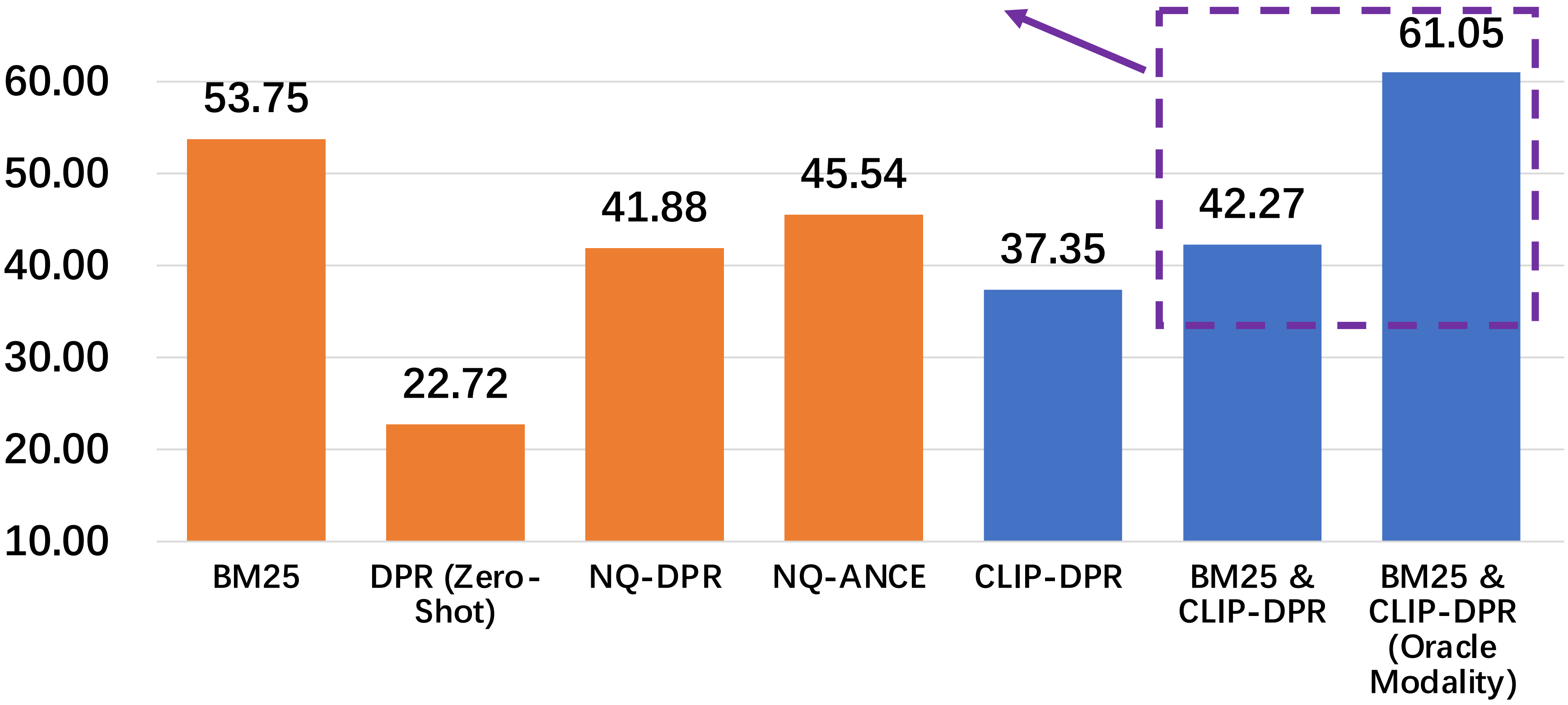
- We first conduct multi-modal retrieval using text retrieval models (***Multi-modal->Single modal***)
  - Encoding image captions as the representations of image documents



BM25 still outperforms existing dense retrievers

Overall Performance

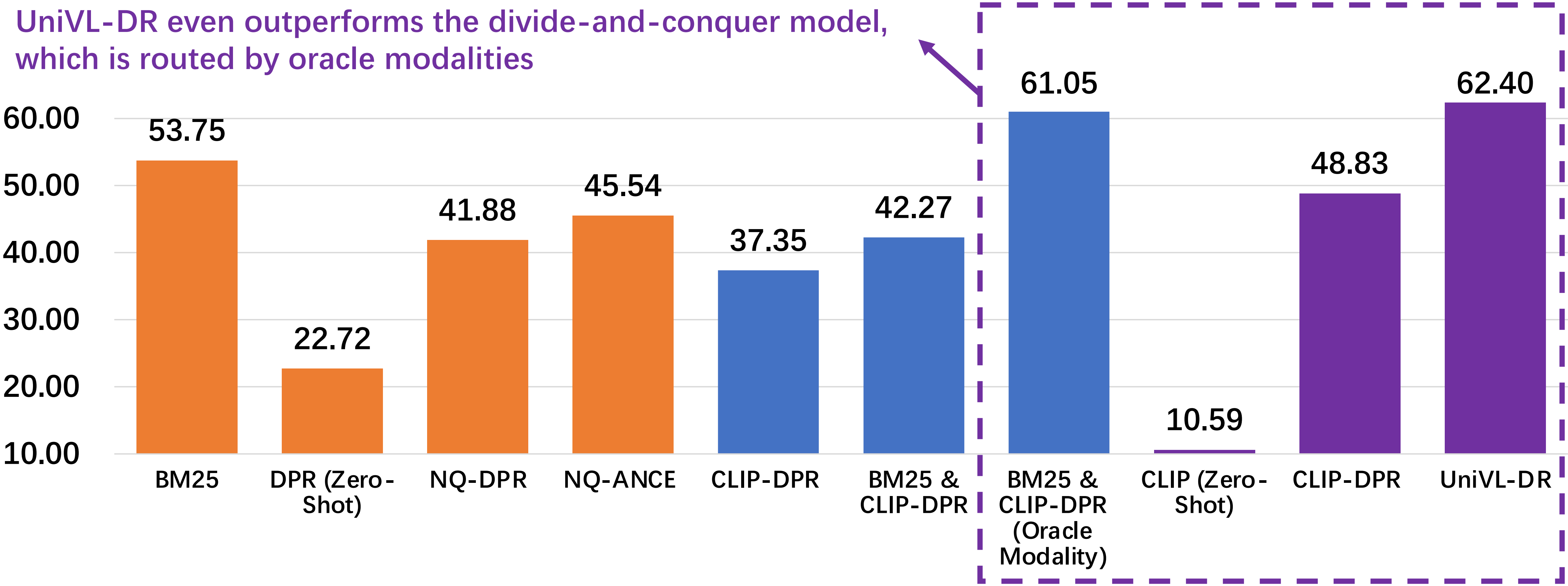
- Multi-modal retrieval can be implemented by divide-and-conquer models (***Multi-modal->Single/Cross modality retrieval & Fusion***)



Modality routing still challenges the multi-modal retrieval tasks

Overall Performance

- Learn one universal embedding space for queries, text documents and image documents (***Multi-modal -> Universal Dense Retrieval***)

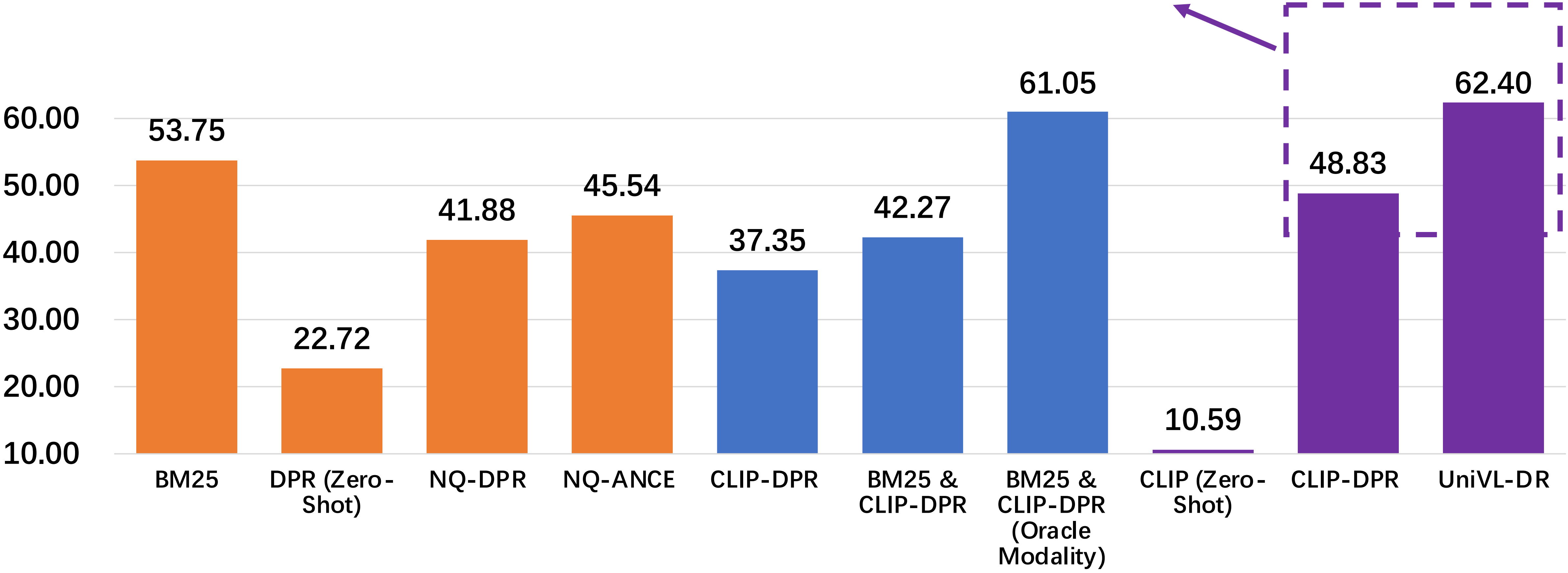


UniVL-DR even outperforms the divide-and-conquer model, which is routed by oracle modalities

Overall Performance

- Learn one universal embedding space for queries, text documents and image documents (***Multi-modal -> Universal Dense Retrieval***)

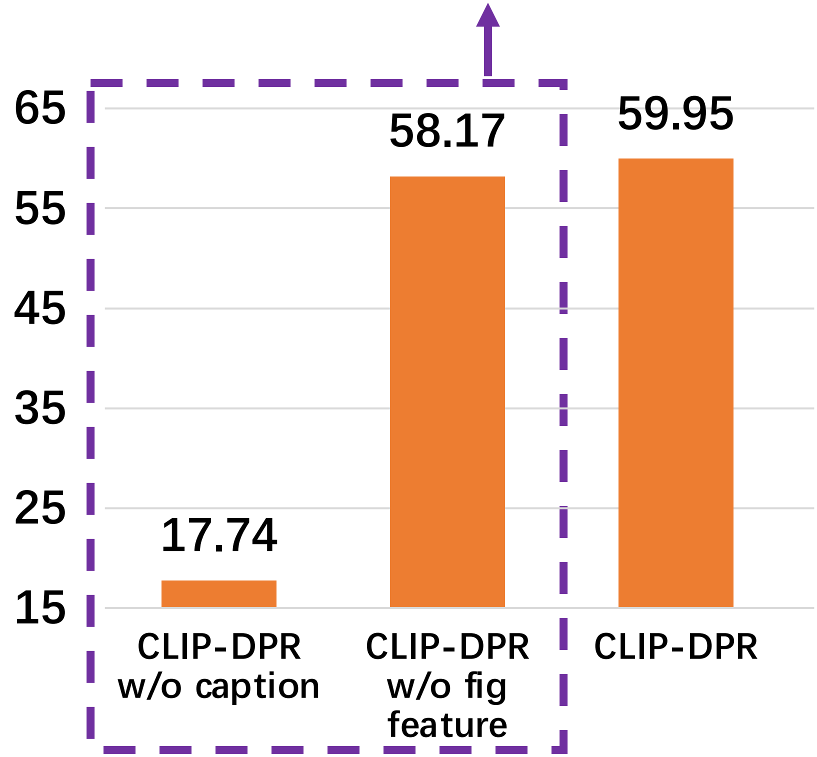
UniVL-DR outperforms CLIP-DPR



Ablation Studies on Image Retrieval

- We conduct experiments on the image retrieval task to show how to represent image documents

The image captions play a critical role in modeling relevance modeling between queries and image documents



Query: Does a Minnetonka Rhododendron flower have petals in a cup shape?

Figure Features:

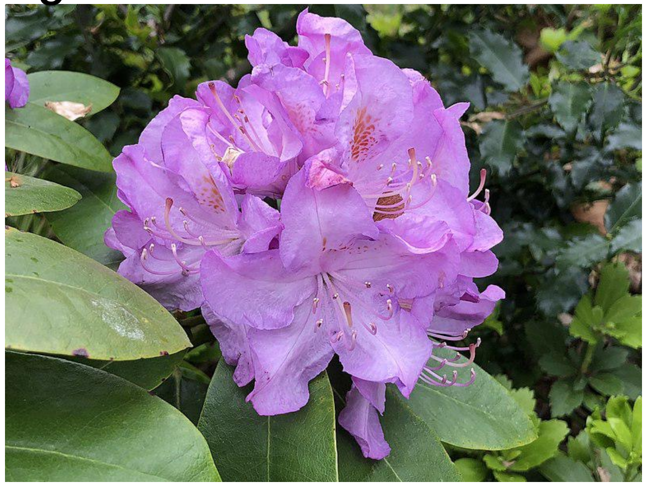
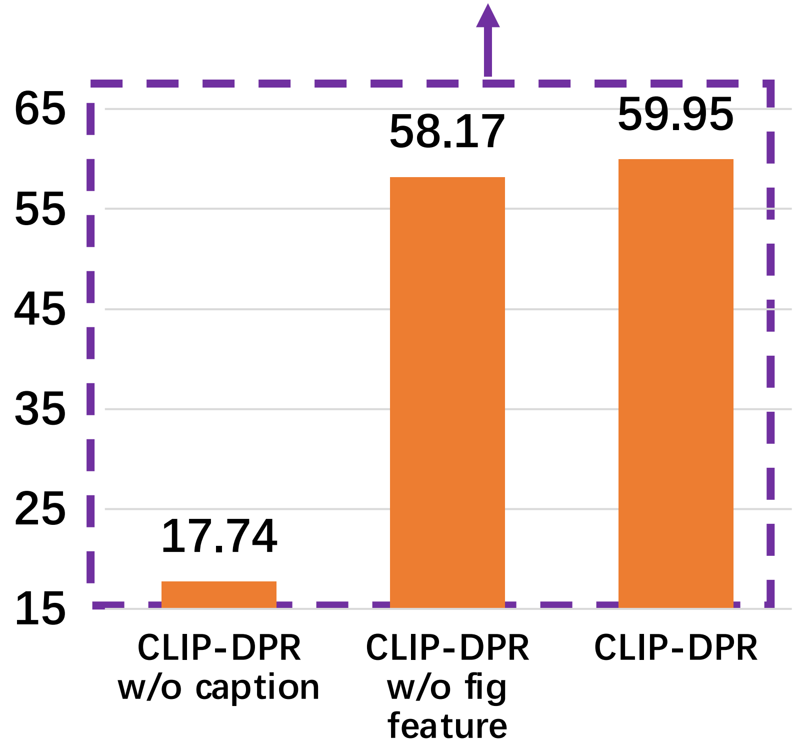


Image Captions:
Minnetonka Rhododendron flower along Tranquility Court …

Ablation Studies on Image Retrieval

- We conduct experiments on the image retrieval task to show how to represent image documents

The figure features can help better understand the semantics of image documents



Query: Does a Minnetonka Rhododendron flower have petals in a cup shape?

Figure Features:

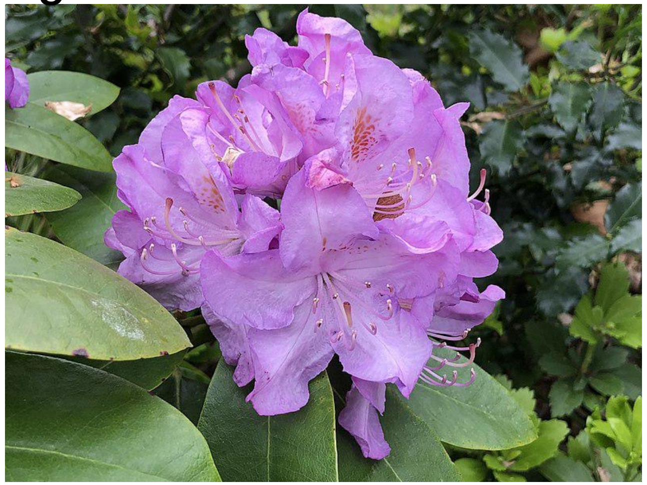
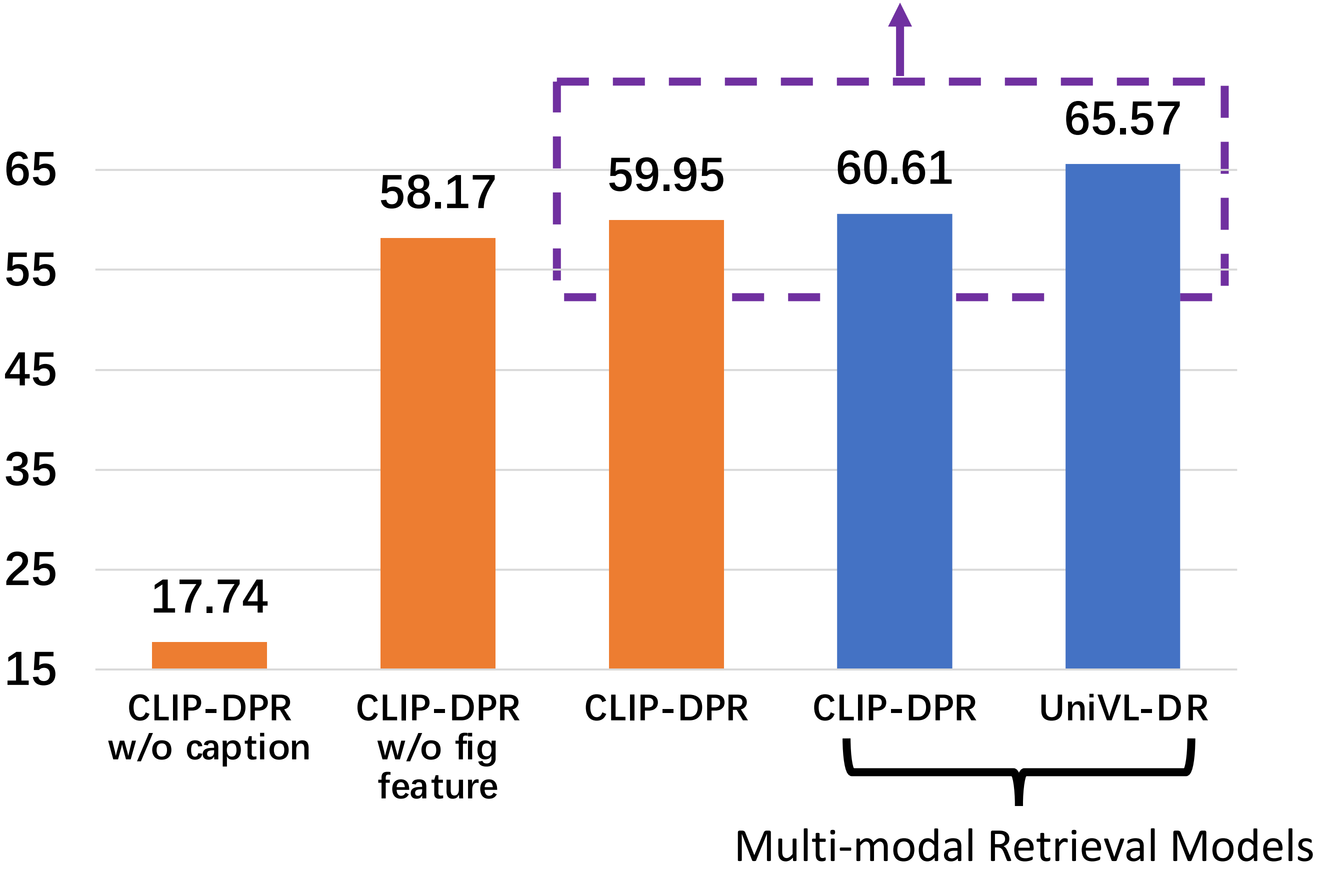


Image Captions:
Minnetonka Rhododendron flower along Tranquility Court …

Ablation Studies on Image Retrieval

- Then we further show the image retrieval performance of multi-modal retrieval models

The text document retrieval tasks can also benefit the image retrieval task



Query: Does a Minnetonka Rhododendron flower have petals in a cup shape?

Figure Features:

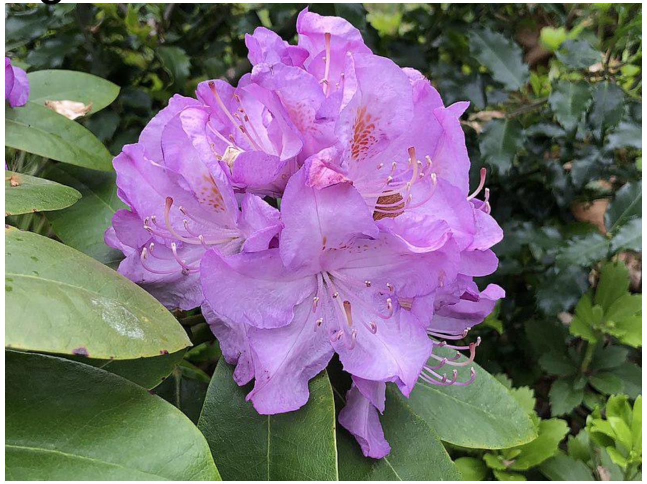
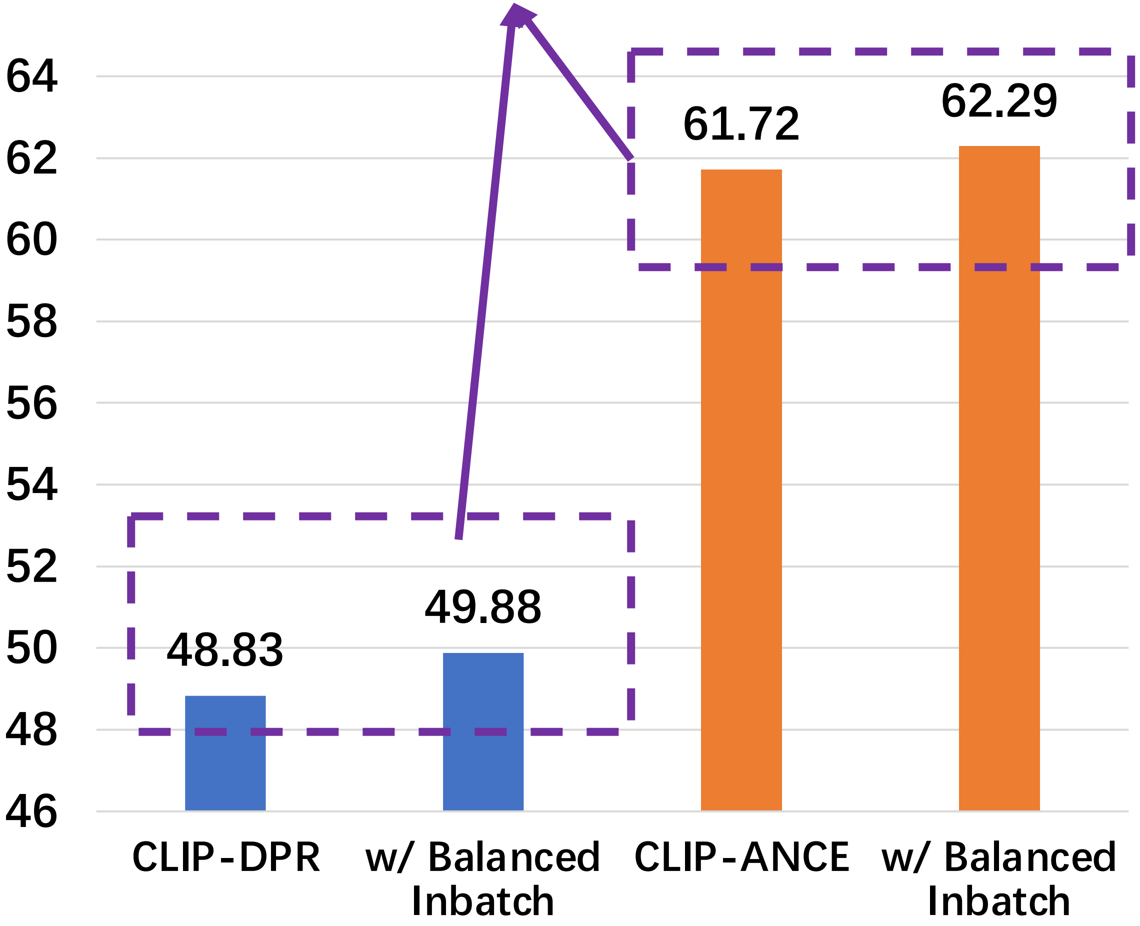


Image Captions:
Minnetonka Rhododendron flower along Tranquility Court …

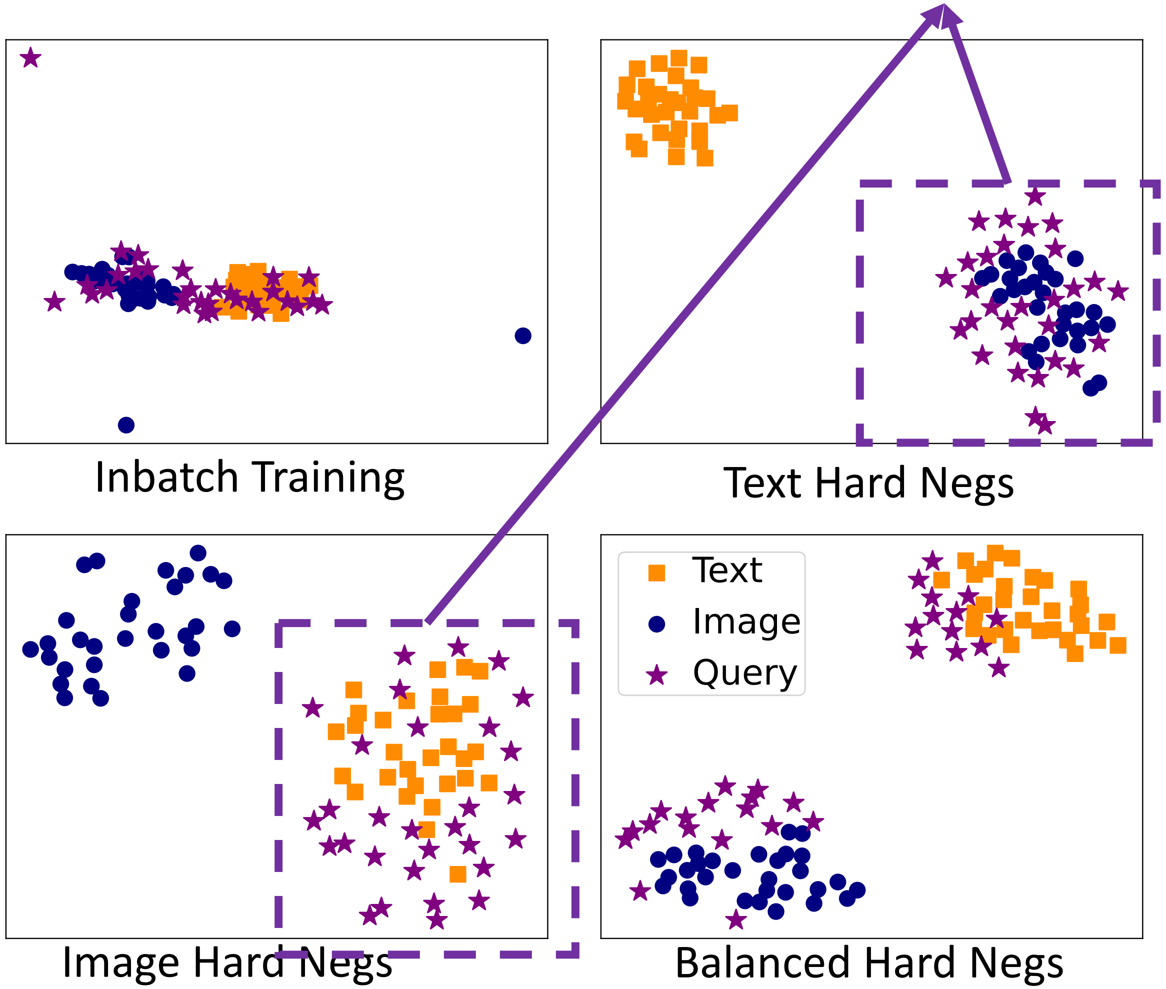
Effectiveness of Modality-Balanced Negative Training

Balancing the modality of negatives can improve the retrieval performance of inbatch and hard negative trained models



Effectiveness of Modality-Balanced Negative Training

Balancing the modality of hard negatives can alleviate the modality preference

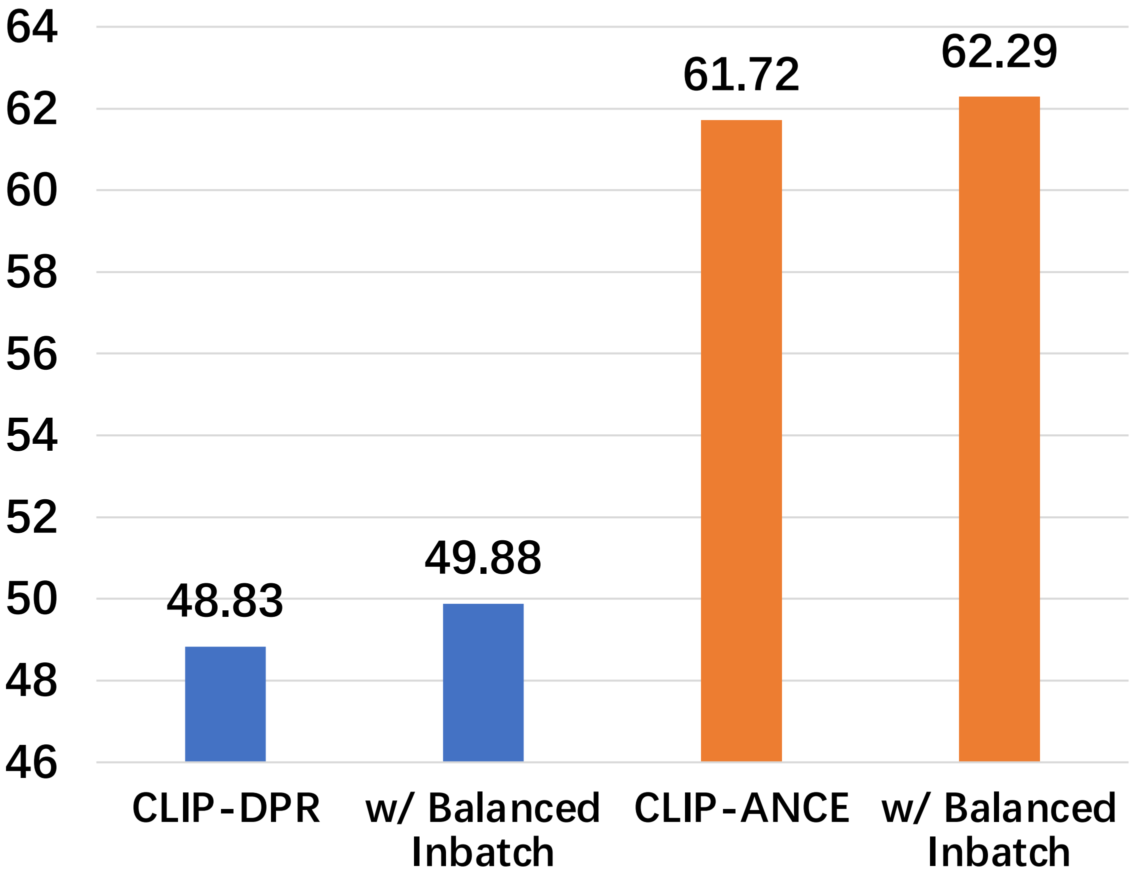


Inbatch Training

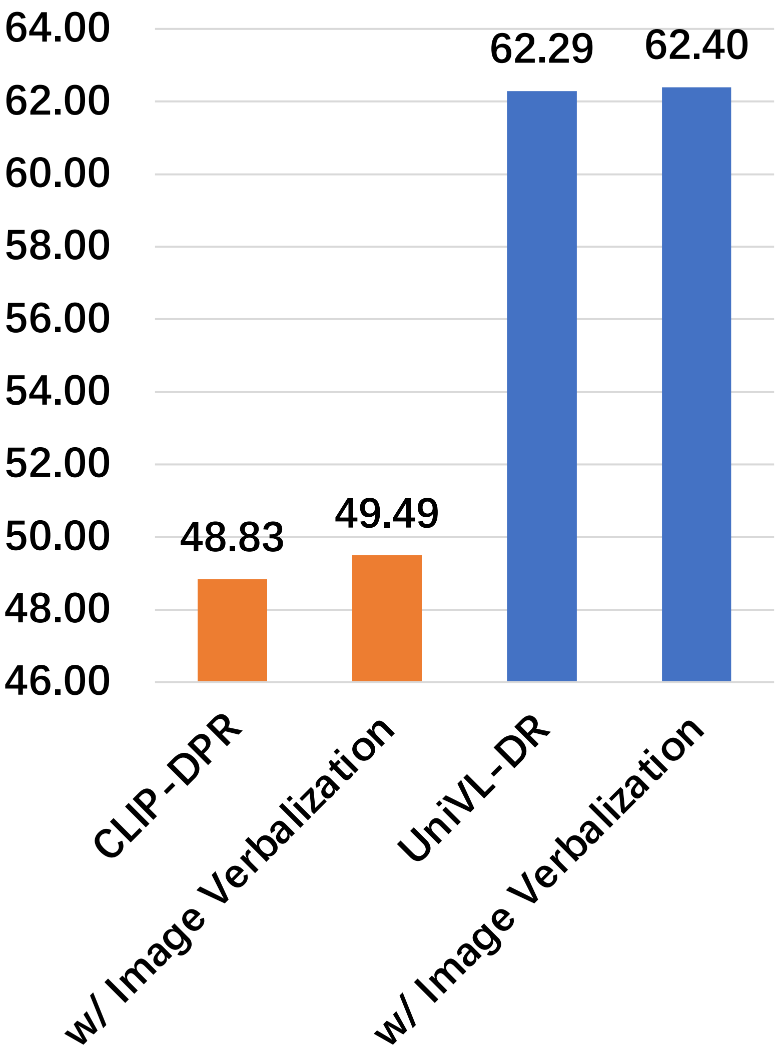
Text Hard Negs

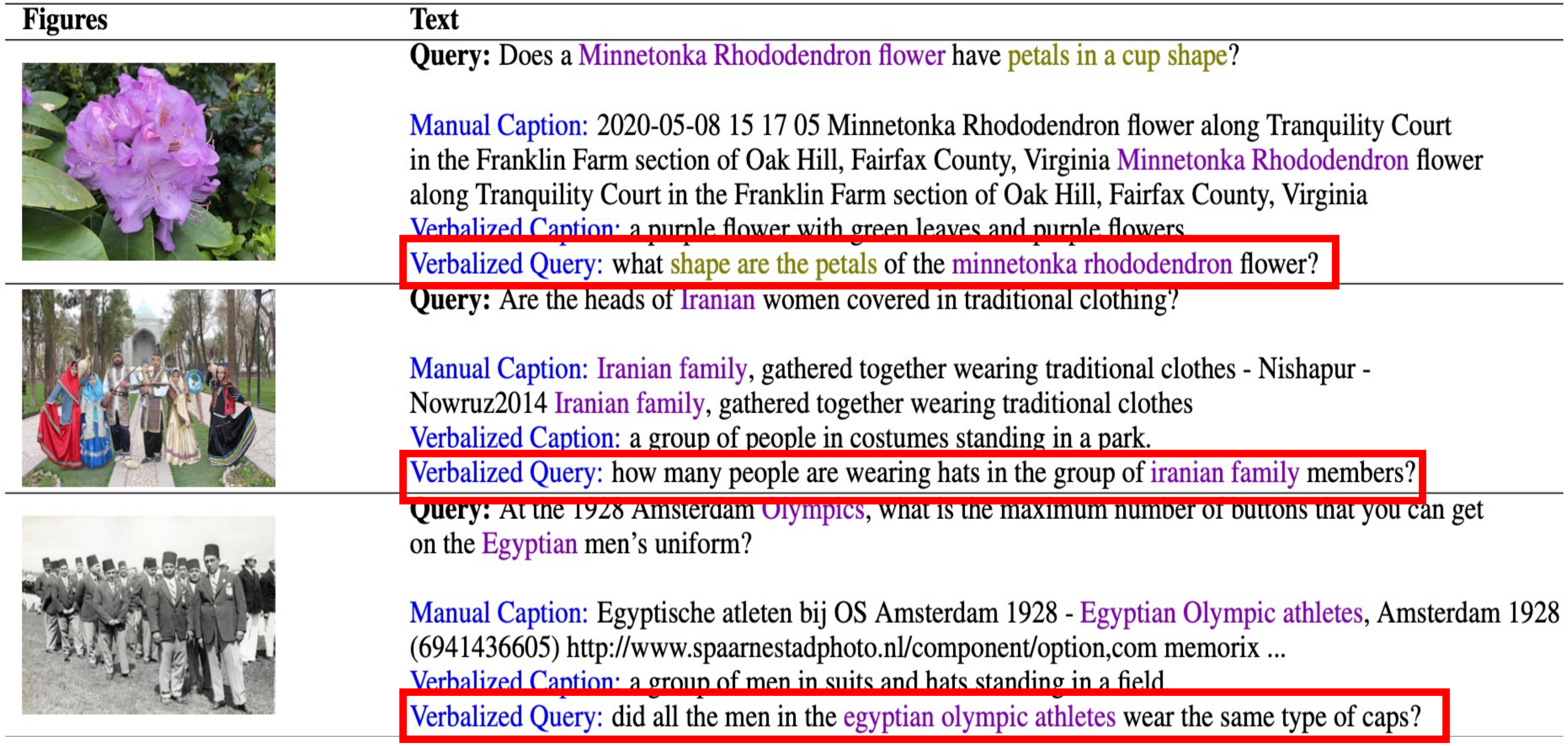
Image Hard Negs

Balanced Hard Negs



Effectiveness of Image Verbalization





Figures	Text
	Query: Does a Minnetonka Rhododendron flower have petals in a cup shape? Manual Caption: 2020-05-08 15 17 05 Minnetonka Rhododendron flower along Tranquility Court in the Franklin Farm section of Oak Hill, Fairfax County, Virginia Minnetonka Rhododendron flower along Tranquility Court in the Franklin Farm section of Oak Hill, Fairfax County, Virginia Verbalized Caption: a purple flower with green leaves and purple flowers Verbalized Query: what shape are the petals of the minnetonka rhododendron flower?
	Query: Are the heads of Iranian women covered in traditional clothing? Manual Caption: Iranian family, gathered together wearing traditional clothes - Nishapur - Nowruz2014 Iranian family, gathered together wearing traditional clothes Verbalized Caption: a group of people in costumes standing in a park. Verbalized Query: how many people are wearing hats in the group of iranian family members?
	Query: At the 1928 Amsterdam Olympics, what is the maximum number of buttons that you can get on the Egyptian men's uniform? Manual Caption: Egyptische atleten bij OS Amsterdam 1928 - Egyptian Olympic athletes, Amsterdam 1928 (6941436605) http://www.spaarnestadphoto.nl/component/option,com memorix ... Verbalized Caption: a group of men in suits and hats standing in a field Verbalized Query: did all the men in the egyptian olympic athletes wear the same type of caps?

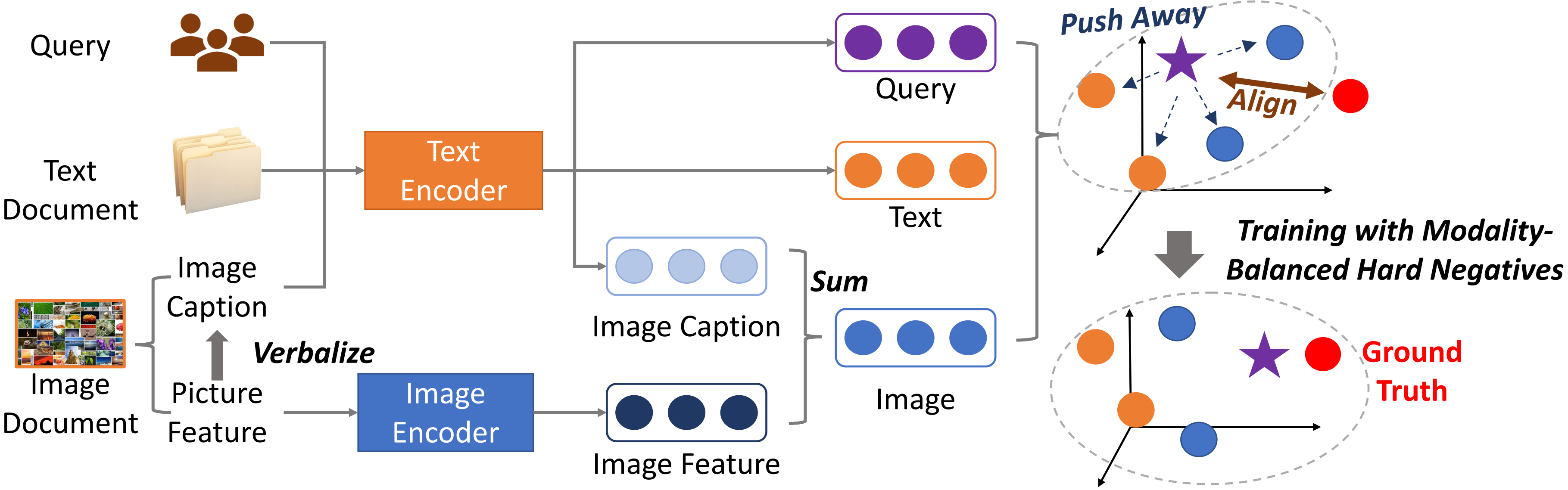
These verbalized queries usually contain matched entities and bridge the modality gap between images and texts

Reviewing UniVL-DR

Image Captions are crucial to ***understanding the image semantics***

Image Verbalization can ***bridge the modality gap*** *between images and texts in the raw text space*

Our Modality-Balanced Hard Negative training method is effective to ***alleviate the modality preference***



All codes and data are available at **https://github.com/OpenMatch/UniVL-DR**.

Thank you!

NEU CSE