האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

OriginAI

NEURAL INFORMATION PROCESSING SYSTEMS

Chatting Makes Perfect: Chat-based Image Retrieval

Matan Levy[1] Rami Ben-Ari[2] Nir Darshan[2] Dani Lischinski[1]

[1]The Hebrew University of Jerusalem, Israel

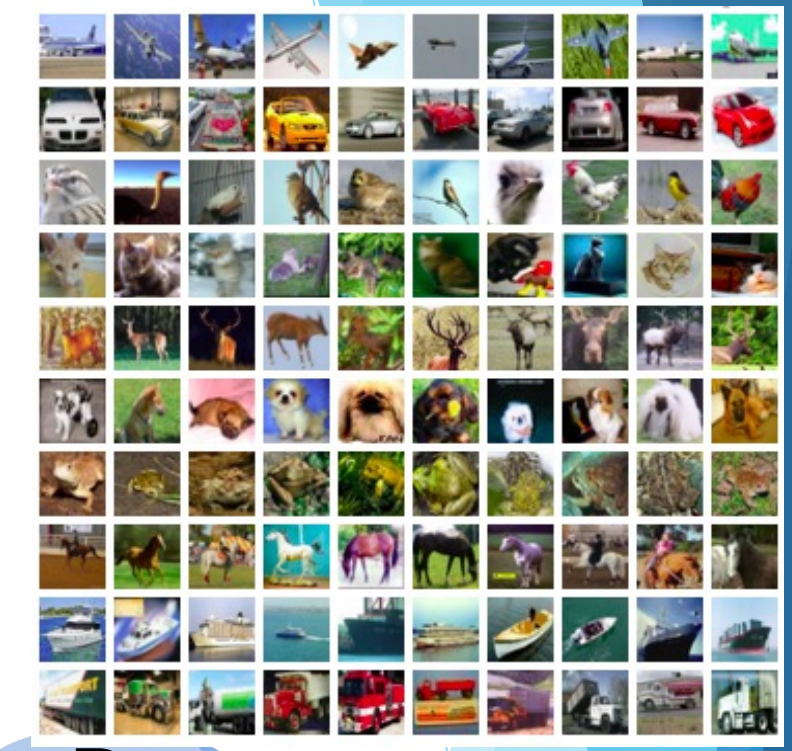
[2]OriginAI, Israel

Project Page:

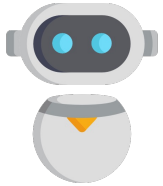
“Image Retrieval”

What is “Image Retrieval”?

- Let D be a large image corpus (1M, 10M, 100M images).
- The user wants to find a certain image in D.
- User provides a query **Q** to search for a desired image.

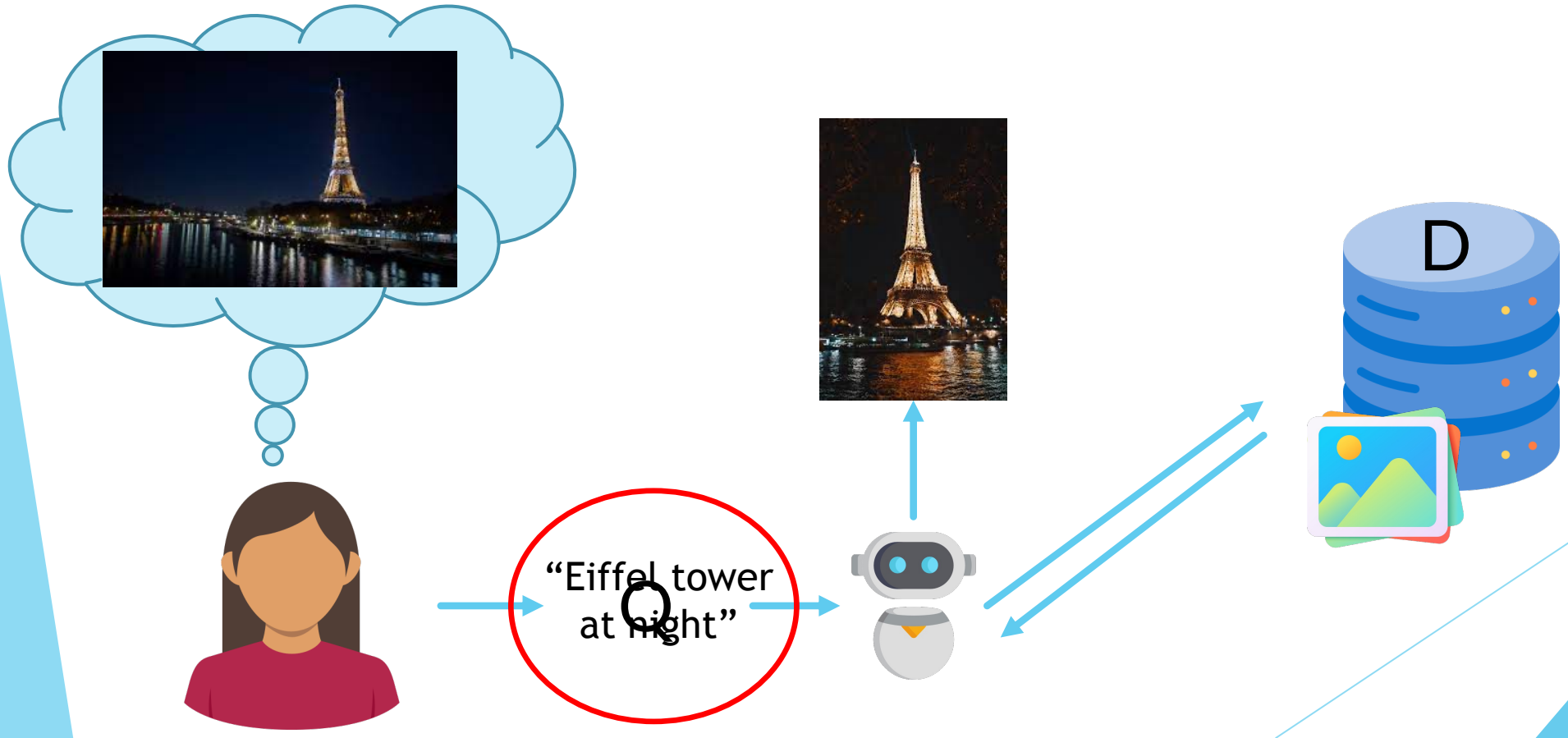


Q

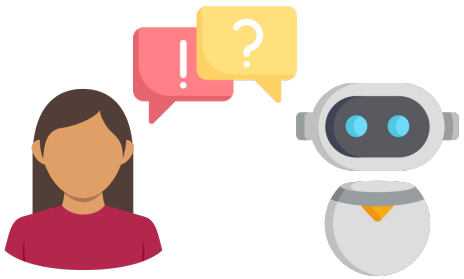


Text-To-Image Retrieval

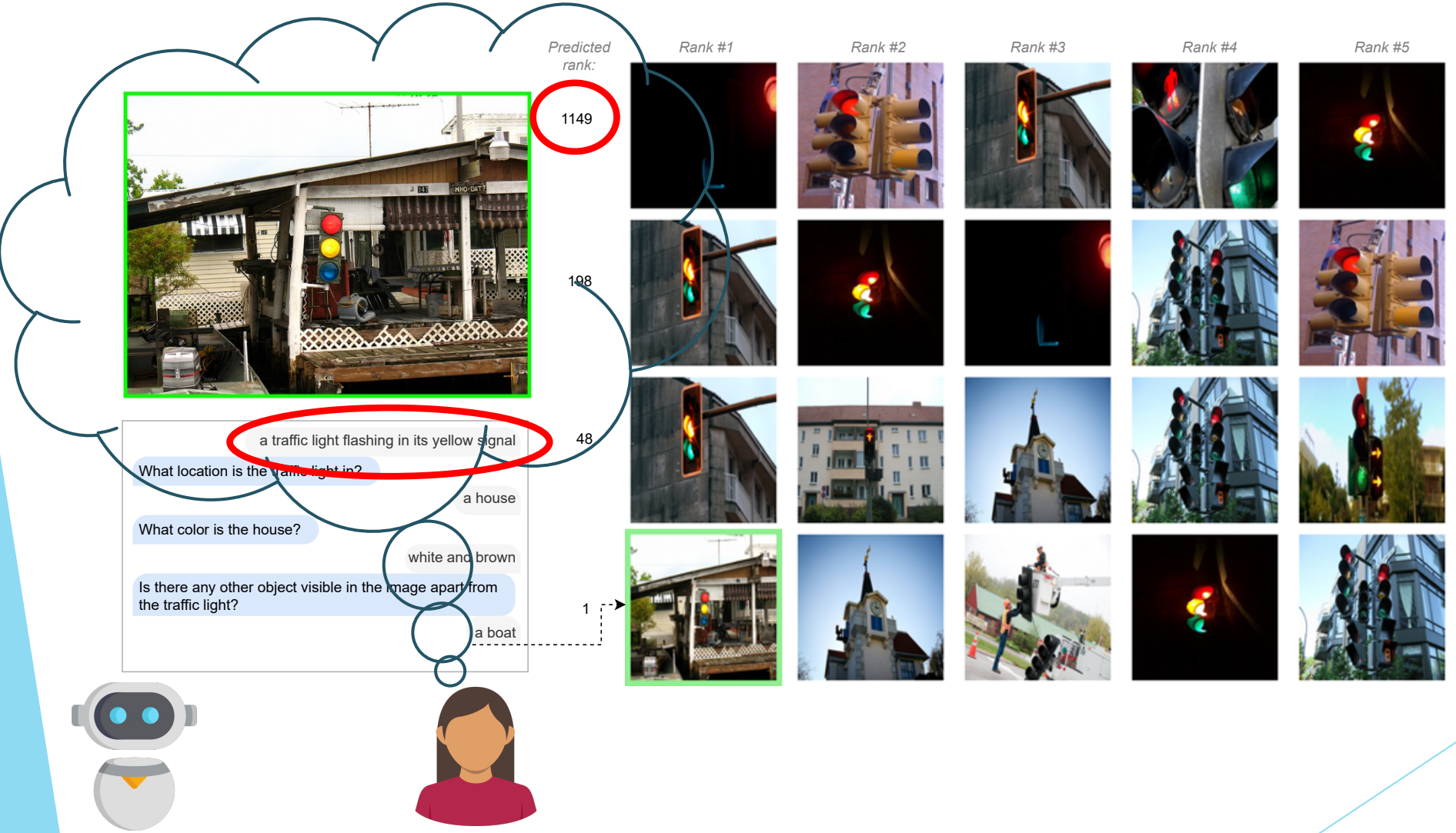
- In the Text-To-Image (TTI) Retrieval task, the query is a textual description of the target image.



Chat-Based Image Retrieval (ChatIR)

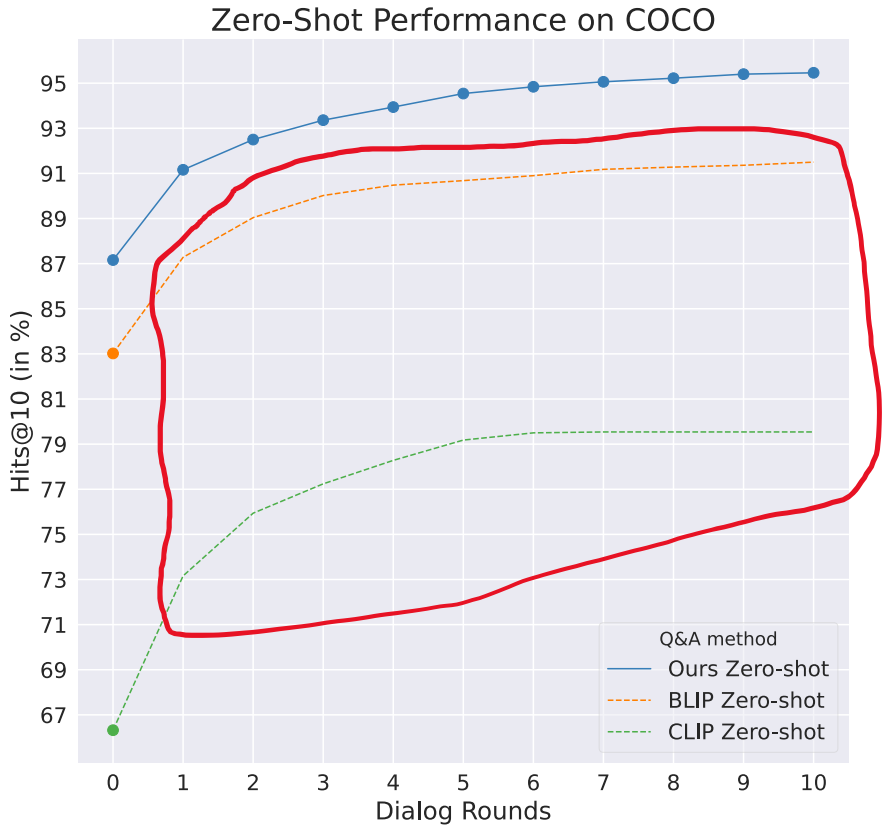


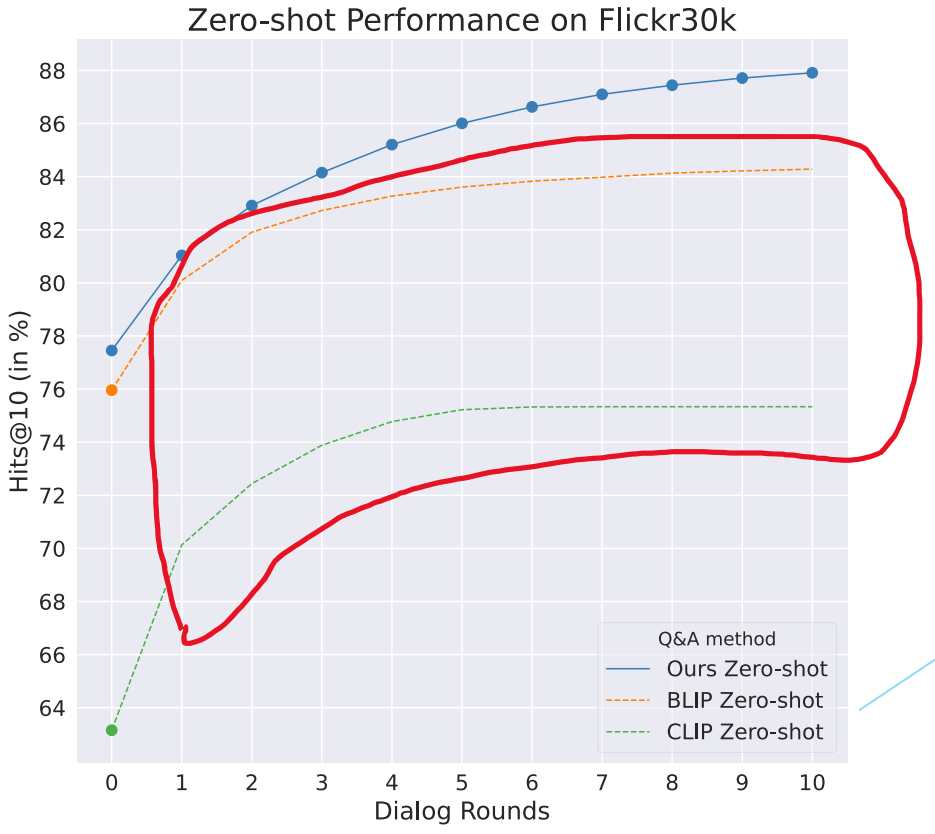
ChatIR Process



Comparison to Text-To-Image Retrieval

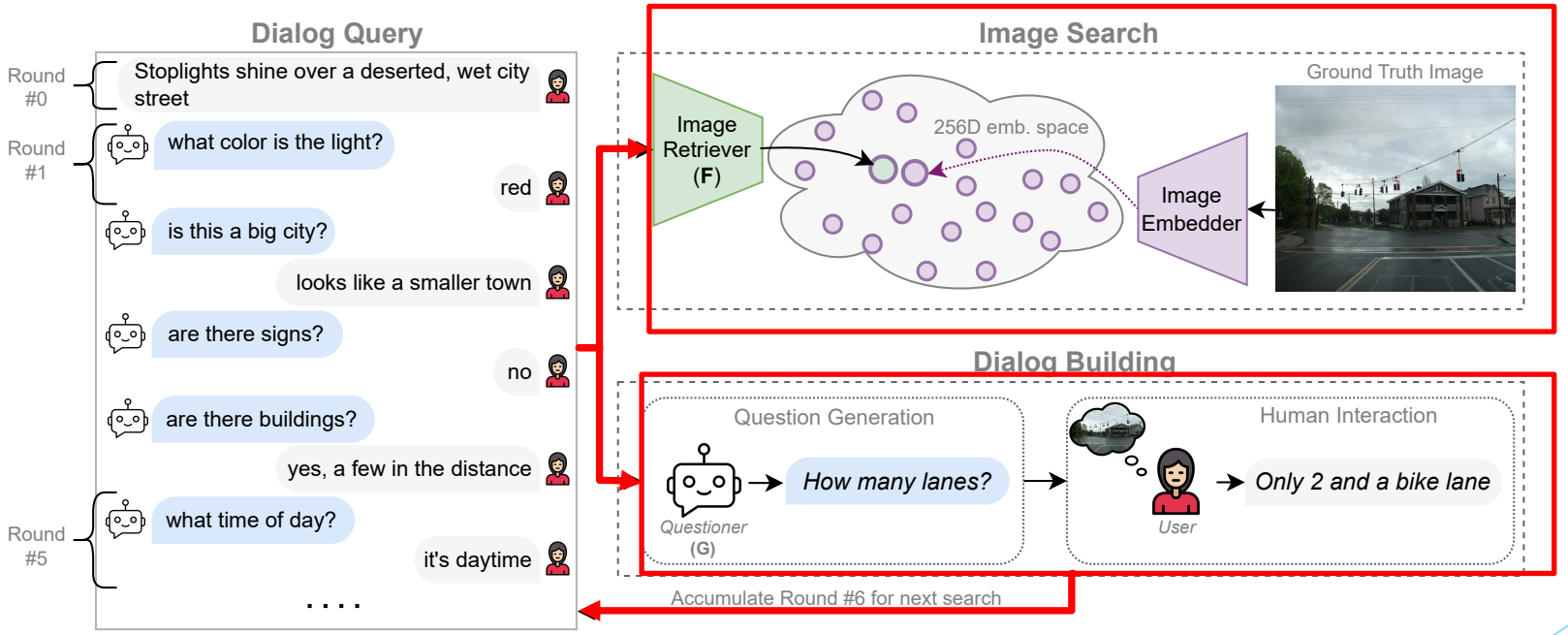
- Here we compare ChatIR to traditional SoTA TTI baselines, in **a zero-shot setting.**





ChatIR Pipeline

- Image Search: encode dialog to a shared space with image features.
- Dialog Building: add a new Q&A pair, interacting with the user.



Datasets

- We use the VisDial dataset, which contains natural images and human-annotated dialougues about each image.
- Answer “unseen” questions with VLM

a bride and groom cutting into their wedding cake together

ChatIR Pipeline

- Image Search: encode dialog to a shared space with image features.
- Dialog Building: add a new Q&A pair, interacting with the user.

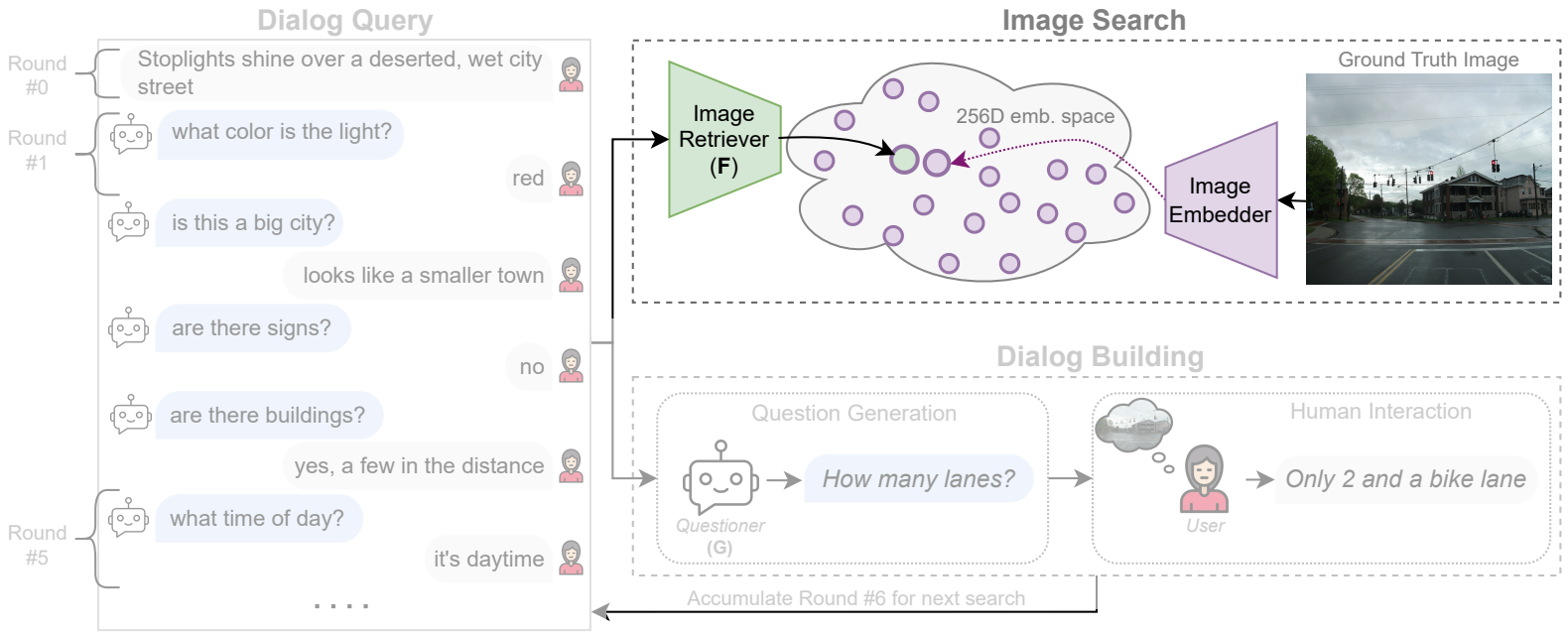
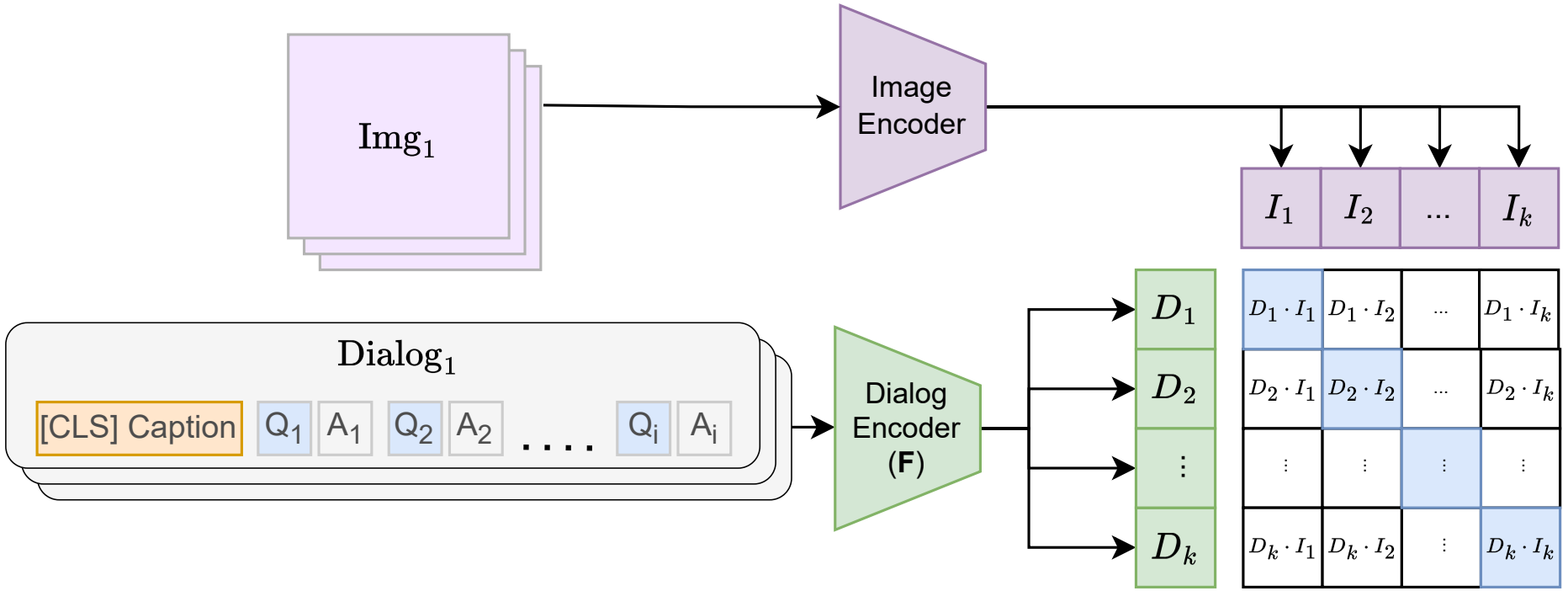


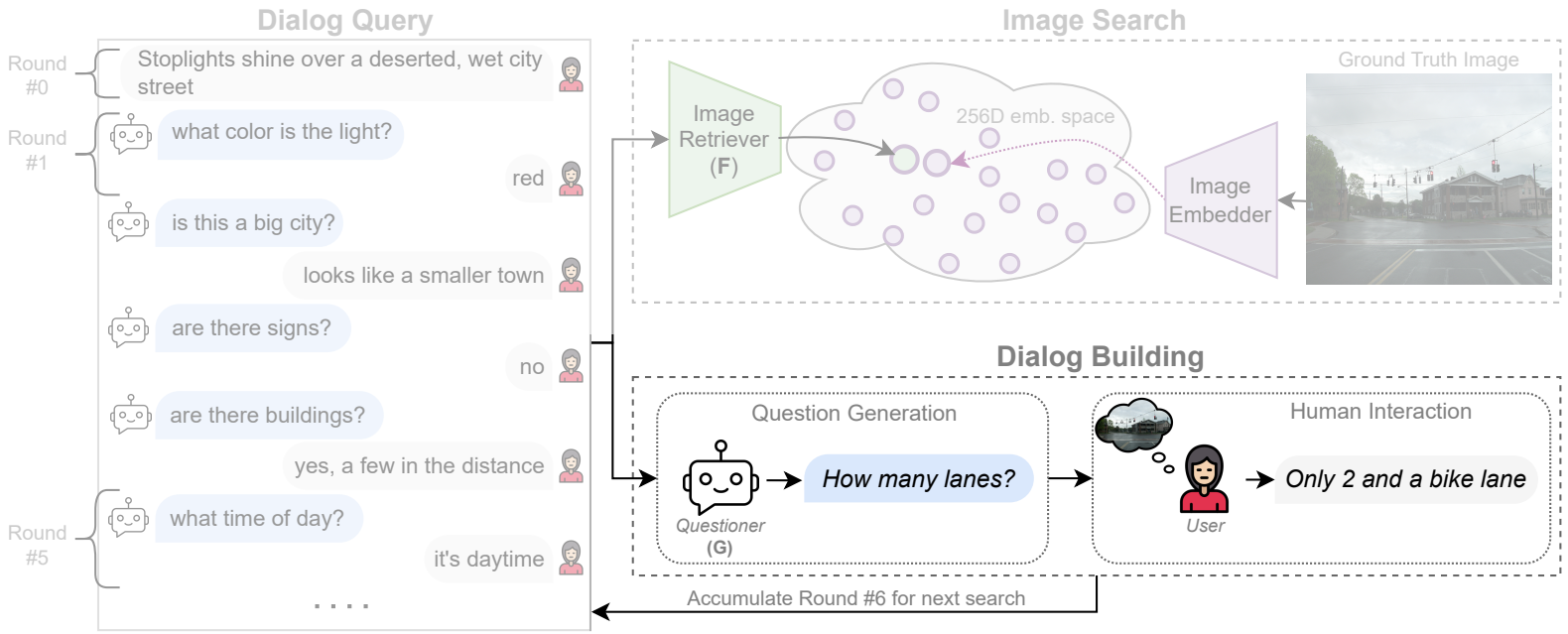
Image Search

- We train two encoders using contrastive loss to match dialogue-image features, similar to CLIP.



ChatIR Pipeline

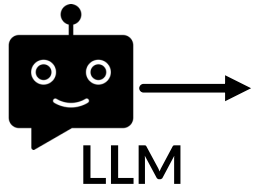
- Image Search: encode dialog to a shared space with image features.
- Dialog Building: add a new Q&A pair, interacting with the user.



Question Generation

- We leverage a pre-trained LLM to generate relevant questions, in the following few-shot setting:

Prompt: **Instruction** + **Example** + **To Complete**



Next Question

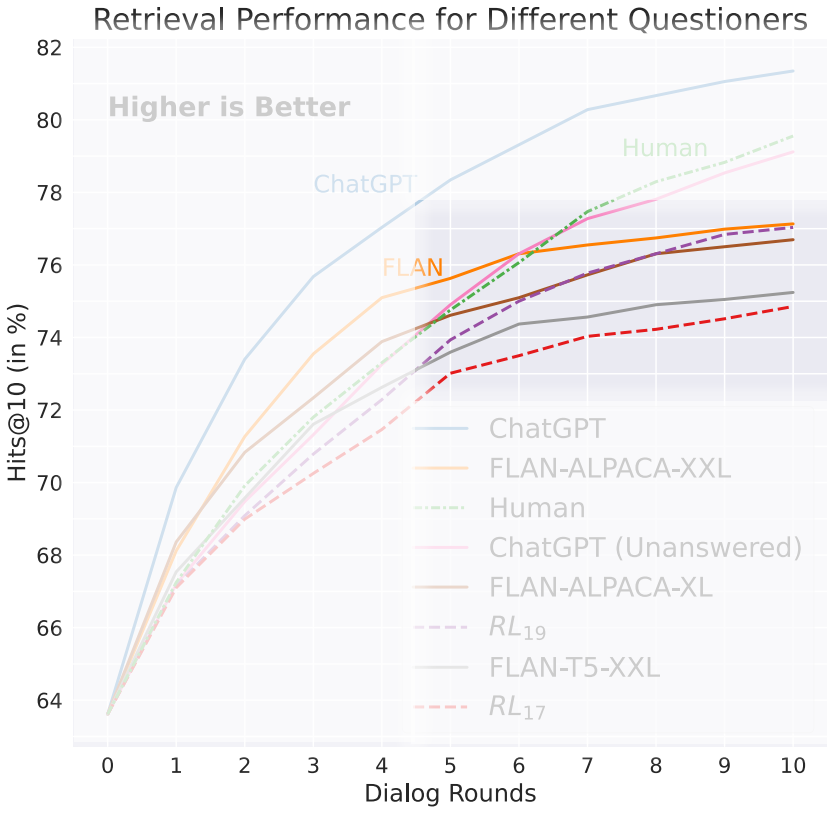
"Ask a new question in the following dialog, assume that the questions are designed to help us retrieve this image from a large collection of images:"

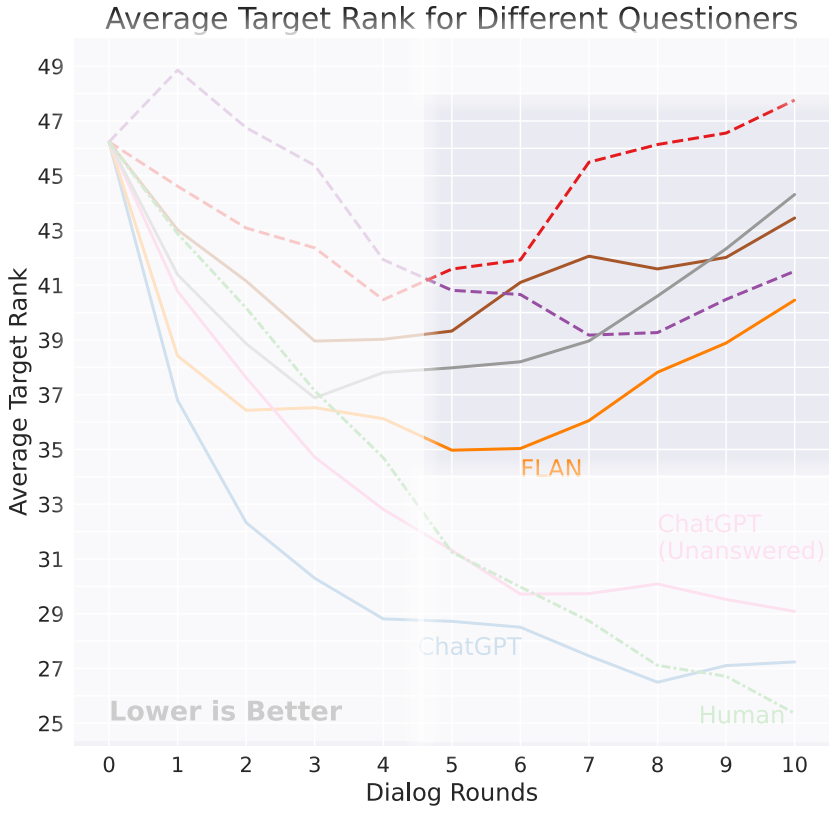
Caption: *2 full grown zebras standing by a brick building with a steel door*
Question: *is this picture in color?*
Answer: *yes*
Question: *do you see people?*
Answer: *no*
Question: *are the animals in a pen*

Caption: *a group of people standing on a snowy slope*
Question: *Are there any trees visible in the background of the image?*
Answer: *no*
Question: *How many people are in the group?*
Answer: *four*
Question:

"Is the slope they are standing on steep or gentle?"

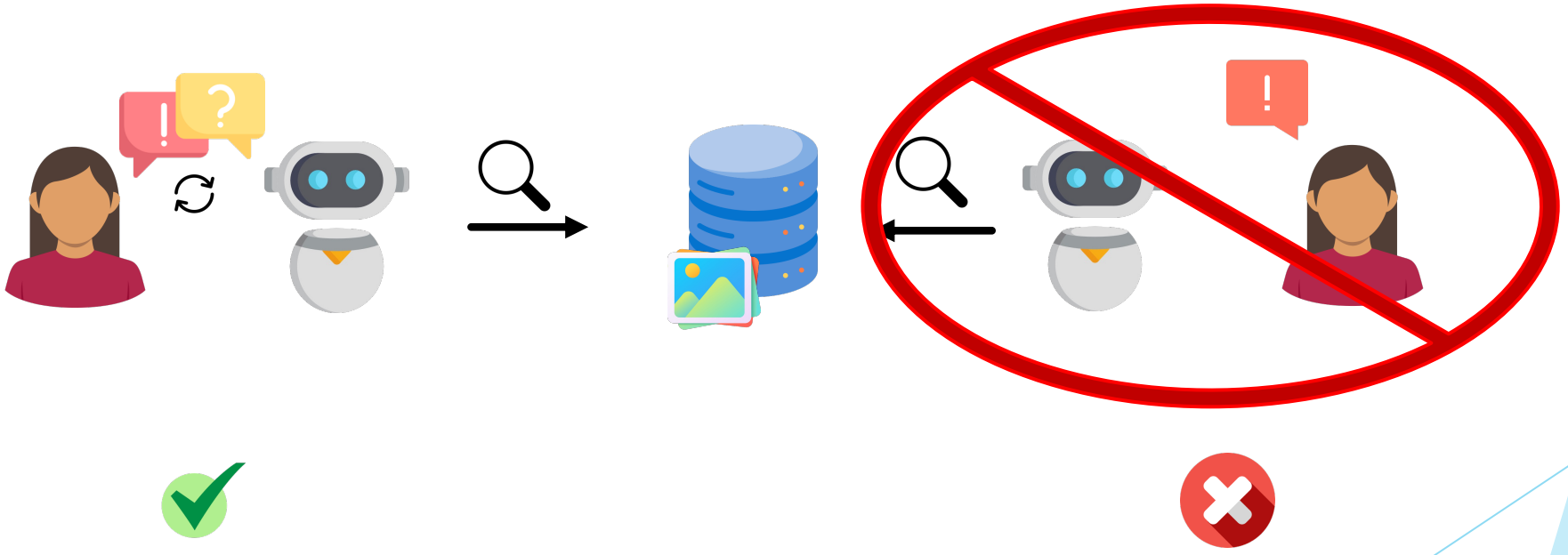
Quetioner Comparison





Left: Evaluation of different chat questioner methods. Note that dialog with 0 rounds is only the image caption, a special case of the text-to-image retrieval task.
Right: Average rank of target images after each round of dialog.

Remember: Use Chats!



Thank you

Project Page: https://vision.huji.ac.il/chatir

GitHub: https://github.com/levymsn/ChatIR