

Enhancing Image Retrieval

RAMI BEN-ARI
OriginAI

Dvir
Samuel



Matan
Levy



Boaz
Lerner







Nir
Darshan



Outline

- Chat based Search
- Generative based Search

Types of Image Retrieval

Retrieval	Query	Target
Text → Image	Construction worker in orange safety vest is working on road	
Image → Image CBIR		
(Image, Text) → Image Composed Image Retrieval (CoIR)	<p>Data Roaming and Quality Assessment for Composed Image Retrieval AAAI 2024 Matan Levy¹, Rami Ben-Ari², Nir Darshan², Dani Lischinski¹</p>	
What 's Next?	Chat Image Retrieval 	?

Do you want to find an image? OK



Let's Chat



Chatting Makes Perfect - Chat-based Image Retrieval

Matan Levy¹ Rami Ben-Ari² Nir Darshan² Dani Lischinski¹

¹The Hebrew University of Jerusalem, Israel

²OriginAI, Israel


NeurIPS 2023



The target rank in the list (lower is better)



Rank #1 Rank #2 Rank #3 Rank #4 Rank #5

	Predicted rank:	Rank #1	Rank #2	Rank #3	Rank #4	Rank #5
	1149					
	198					
	48					
	1					

What location is the traffic light in?
a house

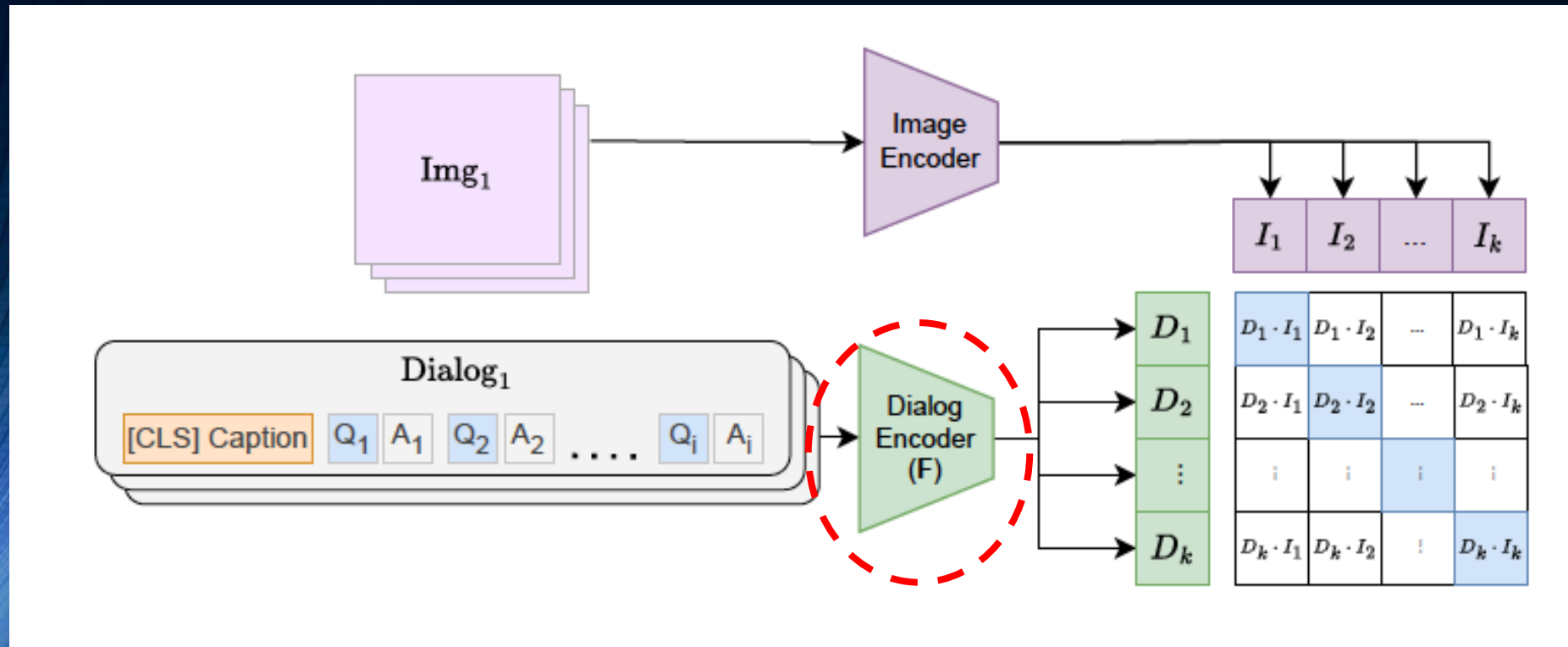
What color is the house?
white and brown

Is there any other object visible in the image apart from the traffic light?
a boat



How do we map a dialog to image representation?

Train with **dialog** and image data



Encoder Training Dataset → Visual Dialog



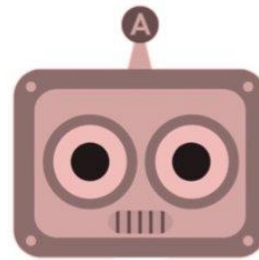
C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked
Q: What kind of dog is it?
A: Looks like beautiful pit bull mix

Q: What color is it?

Image

Dialog history

Question



Visual Dialog model

Answer

A: Light tan with white patch that runs up to bottom of his chin

Visual Dialog requires an AI agent to hold a meaningful dialog with humans about visual content. Specifically, given an image a dialog history, and a follow up question about the image the task is to answer the question.

We got the dialog encoder.

How do we generate the Questions?



Question Generation

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

Instruction:

“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:

Example:

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 there other animals in the scene?

To Complete:

Caption: a group of people standing on a snowy slope

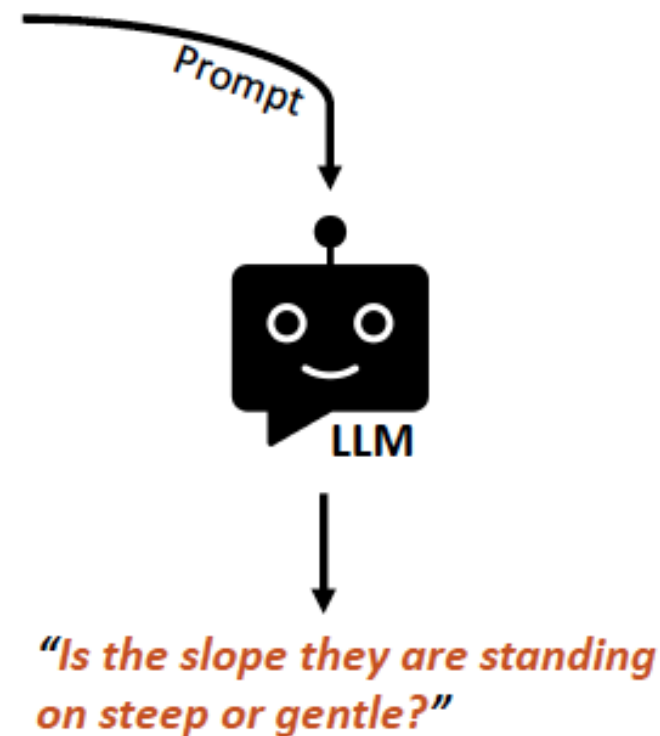
Question: Are there any trees visible in the background of the image?

Answer: no

Question: How many people are in the group?

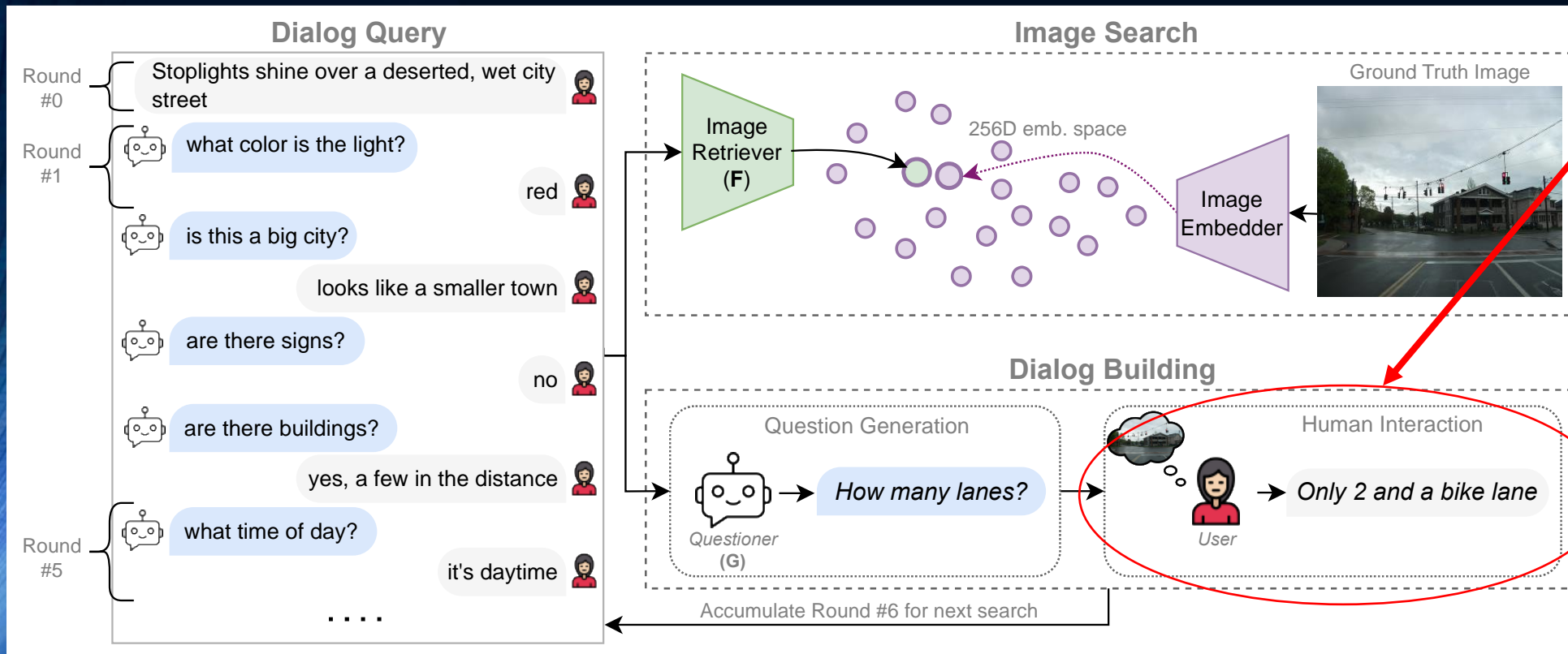
Answer: four

Question: ”



We have trained the pipeline
We have the question generator

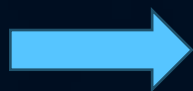
Evaluation Pipeline



Requires human in the loop!!!

Let's replace the human in the loop with an Answering Agent!

Let's use a VQA Engine



BLIP2 - 2023

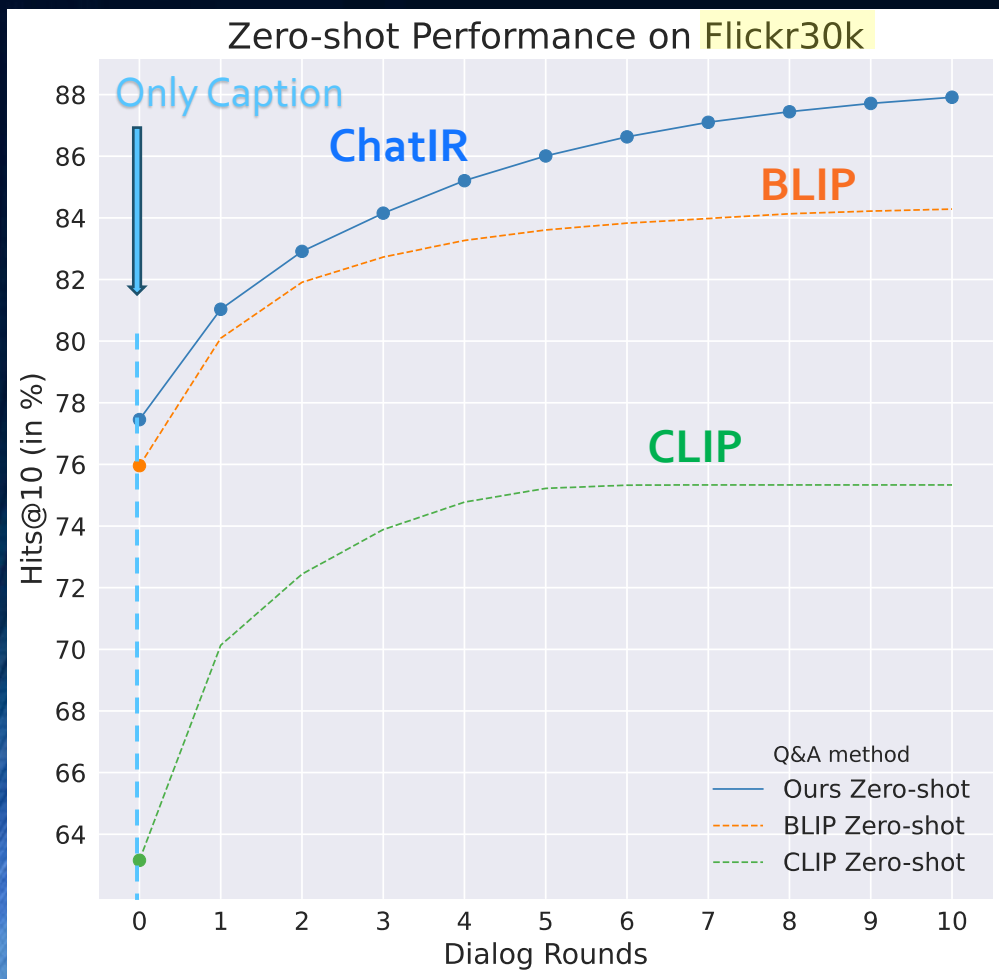
What are shown in the photo?

A man and a chicken.

What does the man feel and why?

He is scared of the chicken because it is flying at him.

Results – Zero Shot Text to Image Retrieval



- Chatting is beneficial – Dialog improves retrieval performance
- Learning to encode dialogs is better

Example



a airplane flying in the air on a sunny day

What type of airplane is it? a commercial airliner

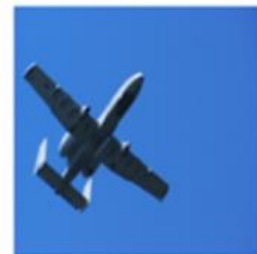
What is the color of the airplane? white

Is the airplane flying over land or water? land

Can you tell what type of terrain or cityscape the airplane is flying over? a forest

Predicted rank:

57



52



28



21



1



Back to Content Based Image Retrieval



Topic	Papers

HyperClass: A Hypernetwork for Few Shot One-Class Classification and Open Set Recognition

Boaz Lerner Nir Darshan Rami Ben-Ari
OriginAI, Israel
{boazl, nir, ramib}@originai.co

ICCVW-2023



Generating Images of Rare Concepts using Pre-trained Diffusion Models

Dvir Samuel^{1,2}, Rami Ben-Ari², Simon Raviv³, Nir Darshan², Gal Chechik^{1,3}

¹Bar-Ilan University, Ramat-Gan, Israel
²OriginAI, Tel-Aviv, Israel
³NVIDIA Research, Tel-Aviv, Israel

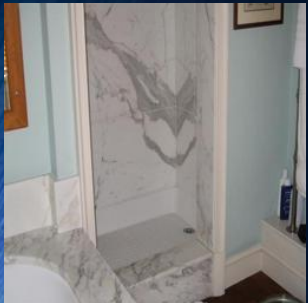
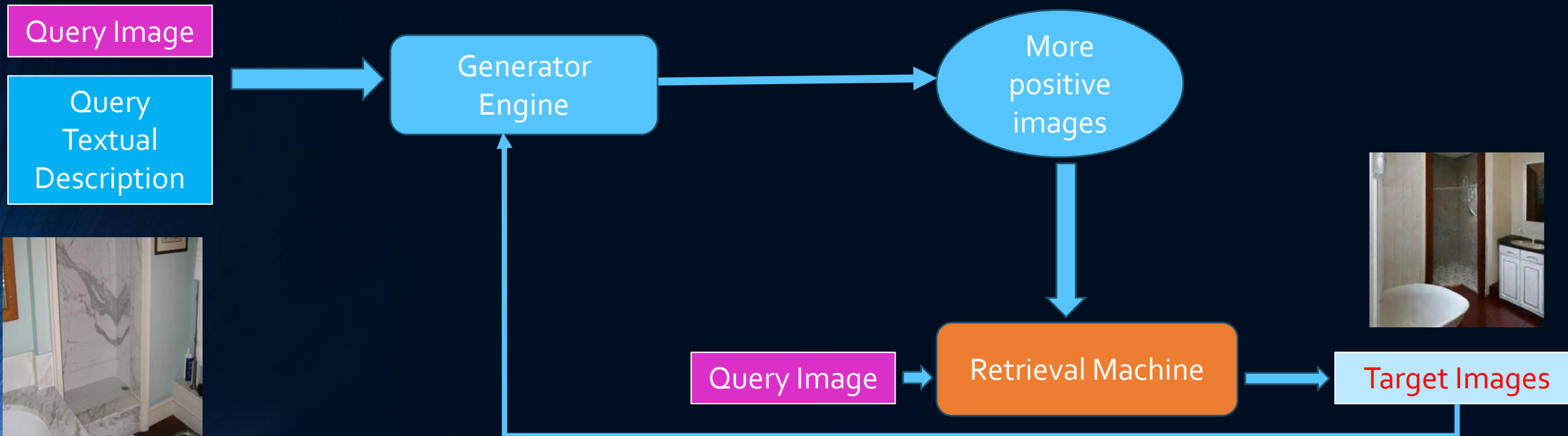


Norm-guided latent space exploration for text-to-image generation

Dvir Samuel^{1,2}, Rami Ben-Ari², Nir Darshan², Haggai Maron^{3,4}, Gal Chechik^{1,4}
¹Bar-Ilan University, ²OriginAI, ³Technion, ⁴NVIDIA Research, Israel



Retrieval with Image Generation



Bathroom



Stable Unclip (Image+Text)



positives

coffeemaker

Generate Images



Benchmarks

Places



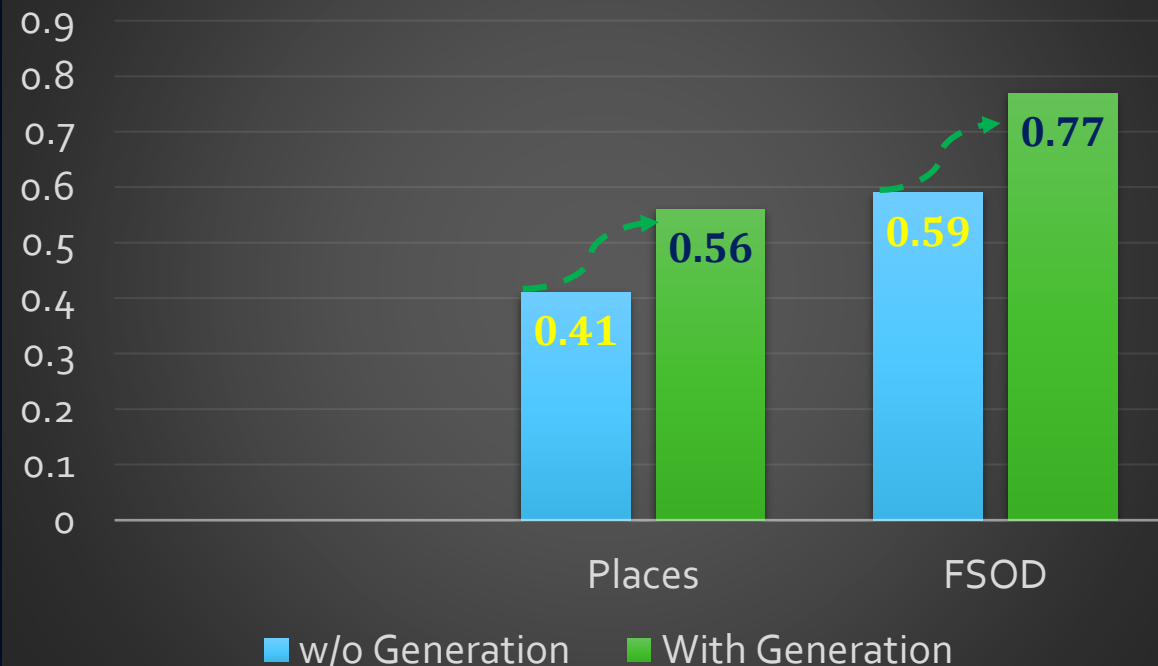
Few Shot Object Detection (FSOD)



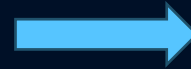
Retrieval - mAP



Pr@50



Q: Can a diffusion model generate any object?



A: No. Not "rare" objects

Stable Diffusion

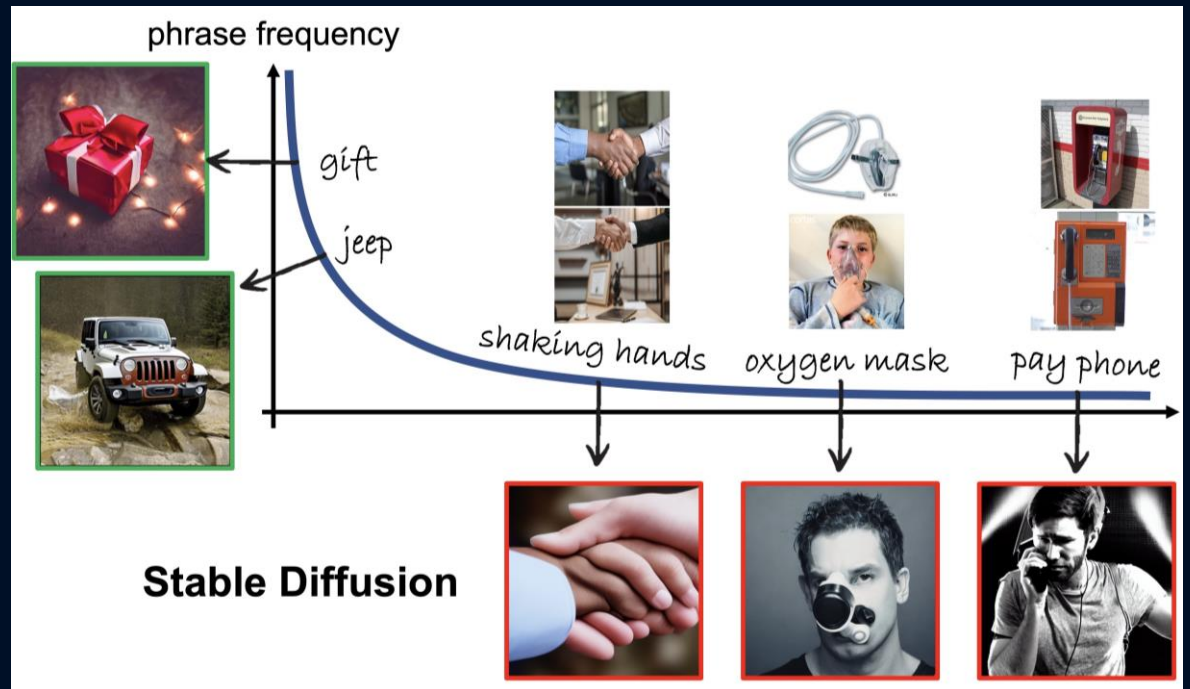
Oxygen Mask



Pay phone

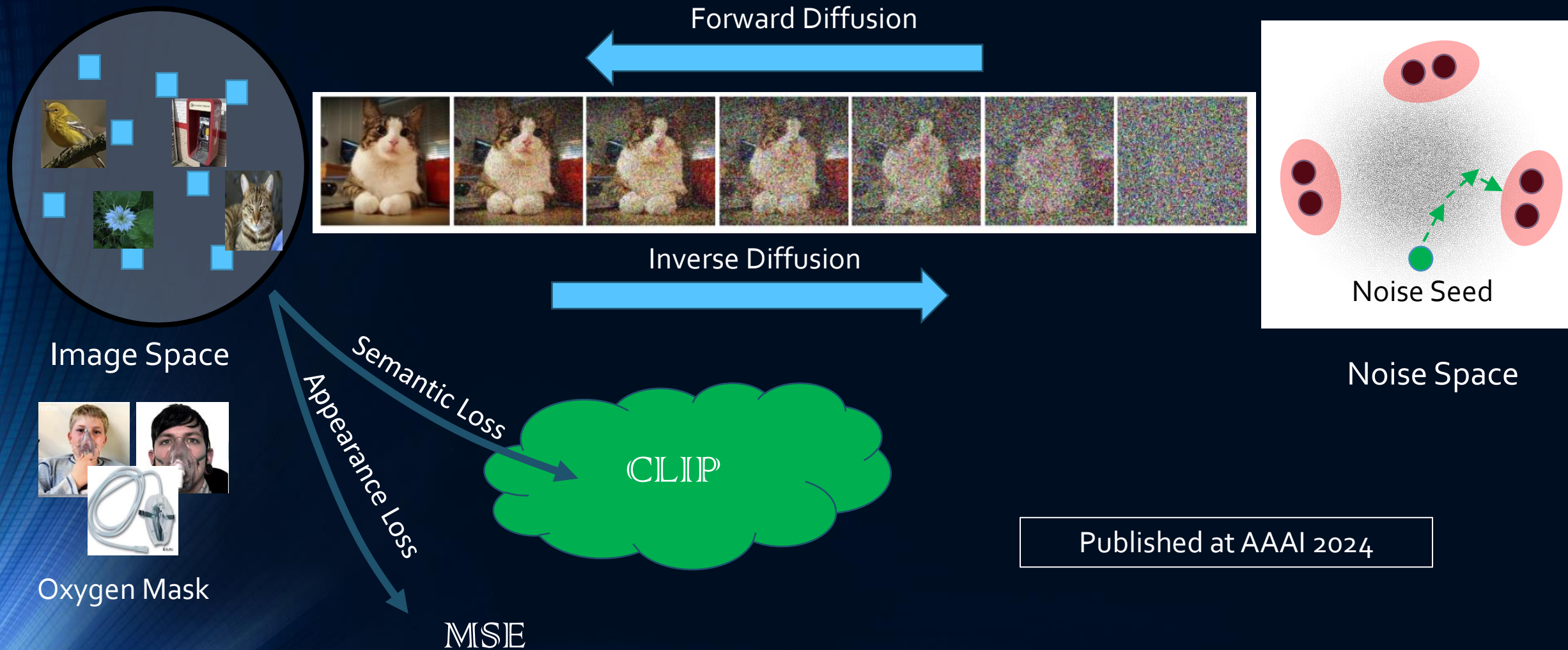


Shaking Hands

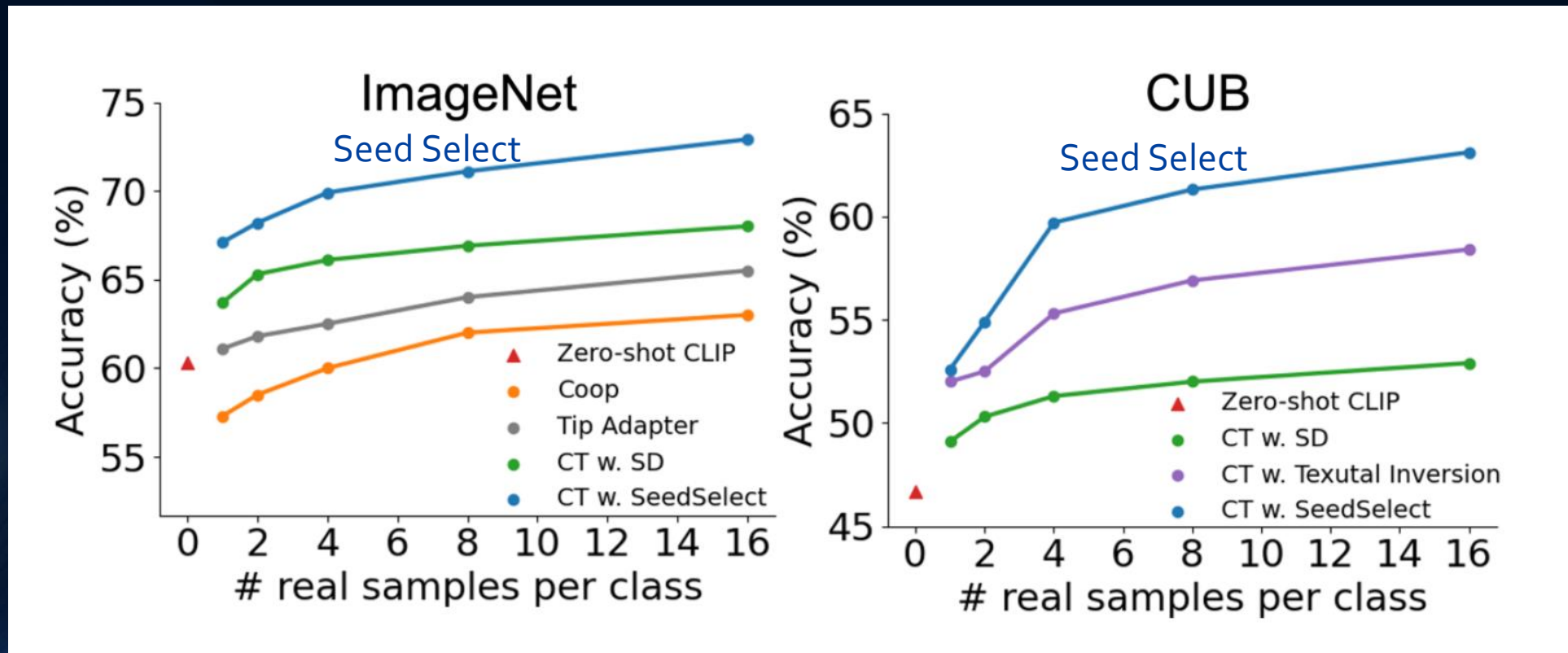


Object generation accuracy vs. prevalence of terms

Our approach – Optimize Seed (SeedSelect)



Performance for CLIP Few Shot Image Recognition



Finetune CLIP

Questions



*Thank you for
your attention!*