“THIS IS MY UNICORN, FLUFFY”: PERSONALIZING FROZEN VISION-LANGUAGE REPRESENTATIONS

Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, Yuval Atzmon
CVPR Workshops, 2022

VISION & LANGUAGE PERSONALIZATION

Teach a pretrained model to recognize new objects from a few examples - allowing it to reason about them with free language.

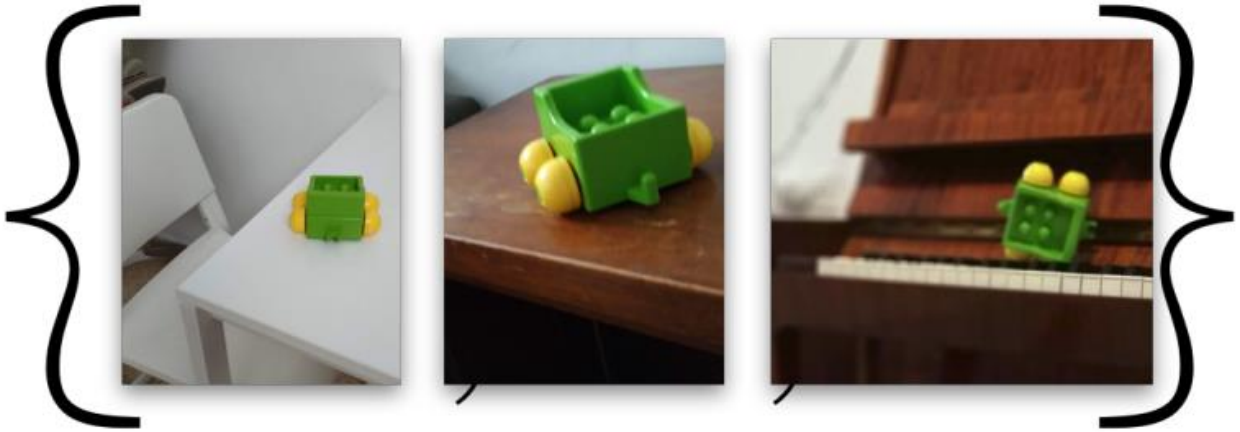
Challenges:

- Risk forgetting prior knowledge
- Accessing prior knowledge concurrently with newly learned concepts.

VISION & LANGUAGE PERSONALIZATION

Teach a pretrained model to recognize new objects from a few examples - allowing it to reason about them with free language.

This is "*my toy wagon*"



VISION & LANGUAGE PERSONALIZATION

Teach a pretrained model to recognize new objects from a few examples - allowing it to reason about them with free language.

Segment "*The **elephant** on my toy wagon*"

Frozen CLIP

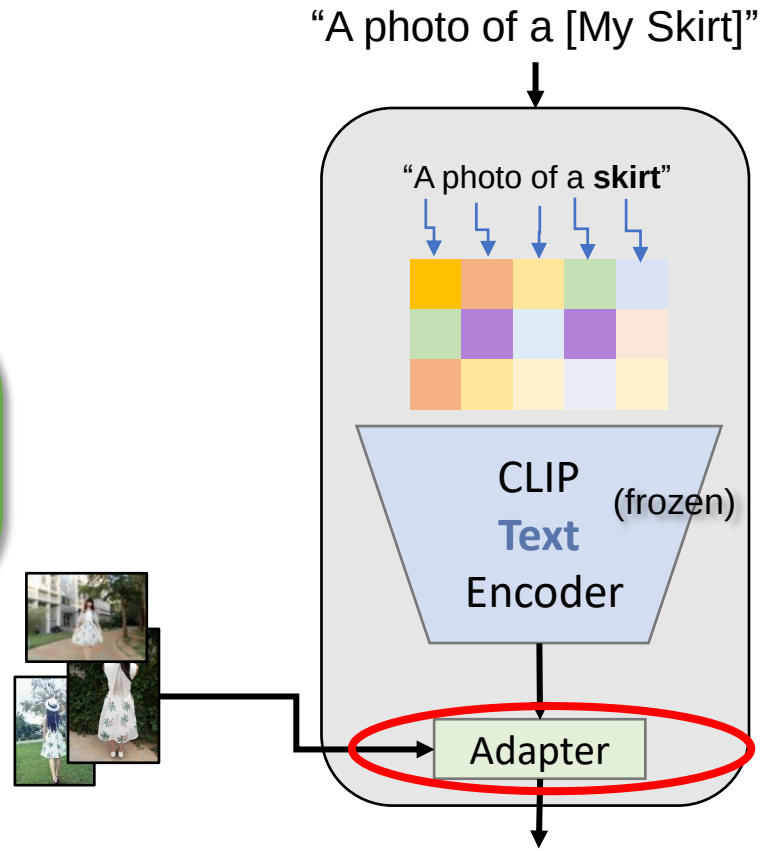
WHY IS IT IMPORTANT

In many domains collecting labelled data is costly and hard

Leverage the power of pretrained models to reason over a large body of prior knowledge jointly with the personalized concepts.

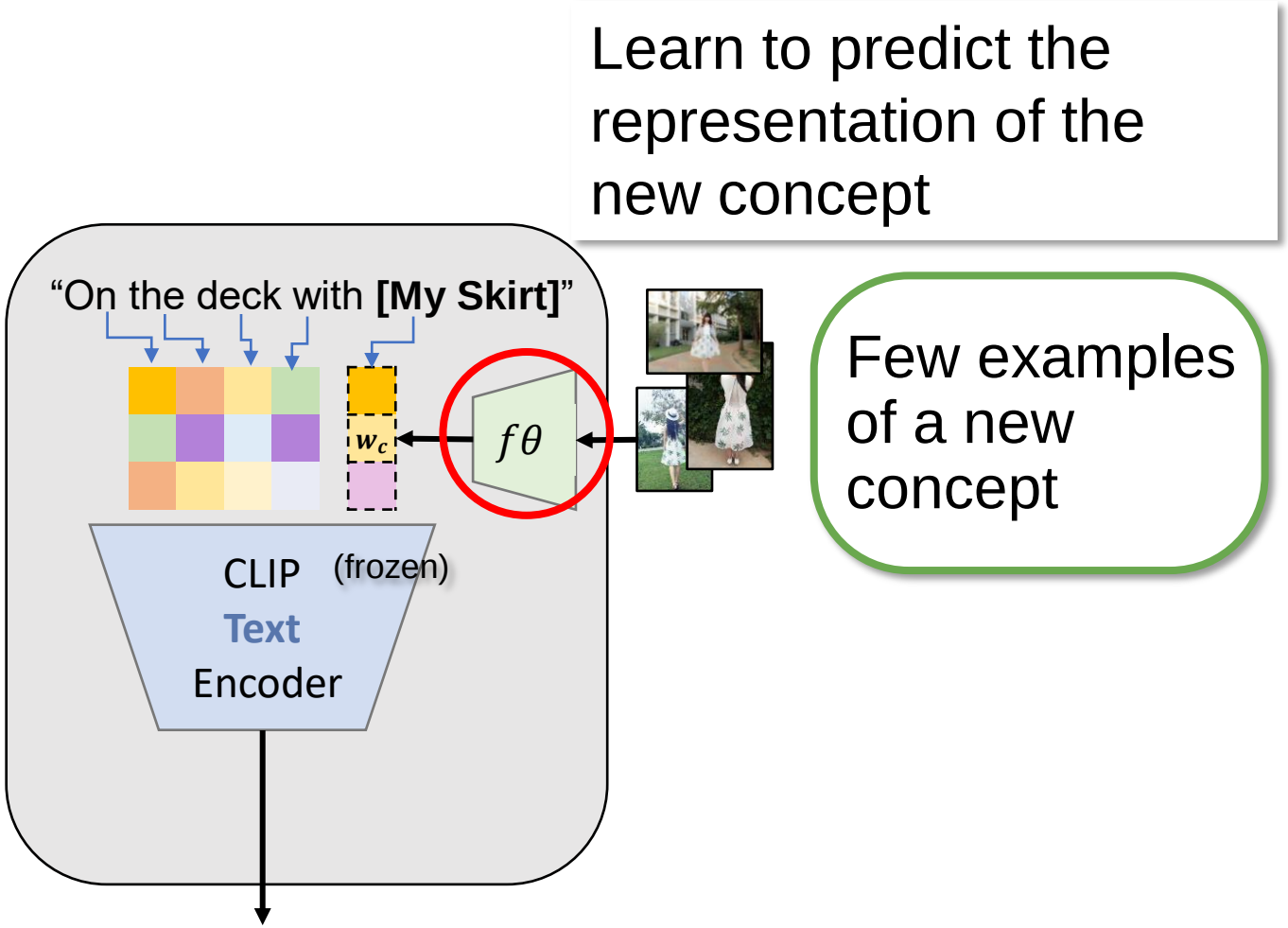
BASELINE (ADAPTER)

Few examples of a new concept



KEY IDEA - @LEARNING

Represent new concepts as *input representations*, such that they are correctly processed by the frozen model.



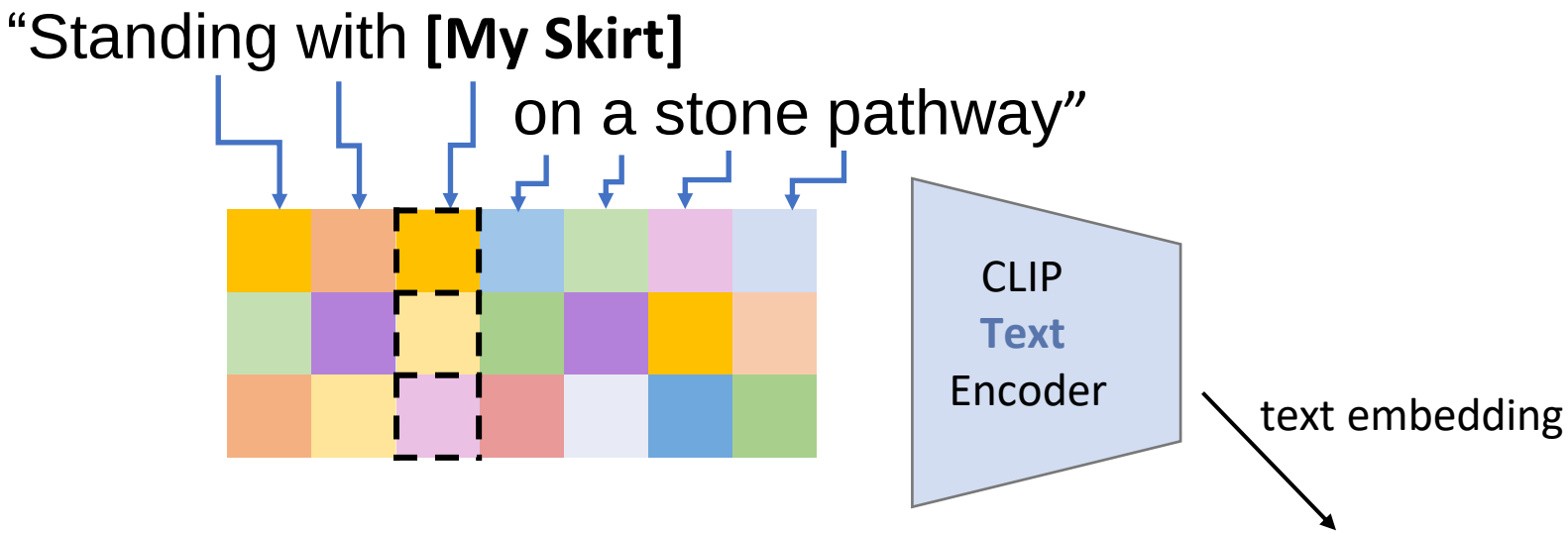
KEY IDEA - @INFERENCE

Example: Personalized album search

"Standing with **[My Skirt]**
on a stone pathway"

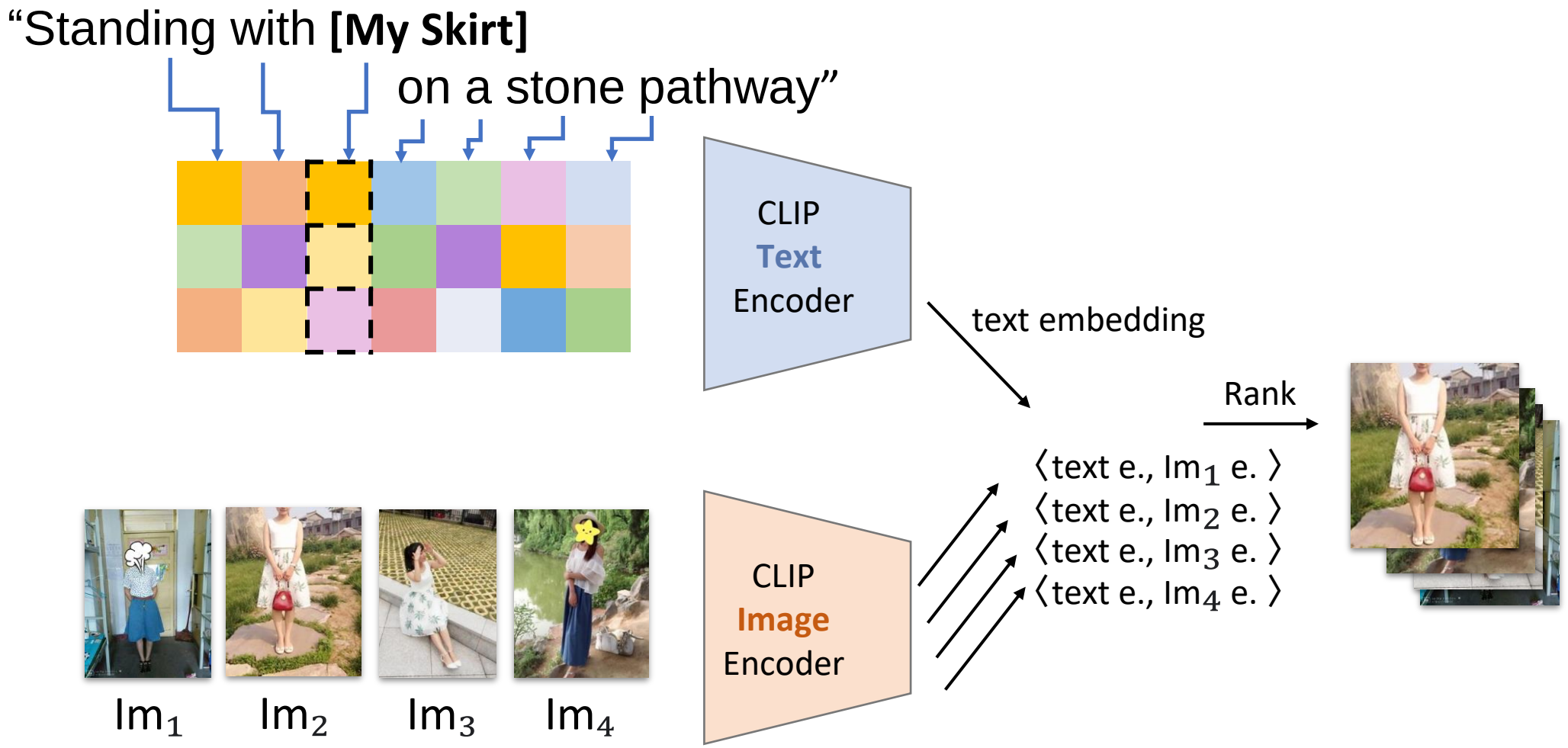
KEY IDEA - @INFERENCE

Example: Personalized album search



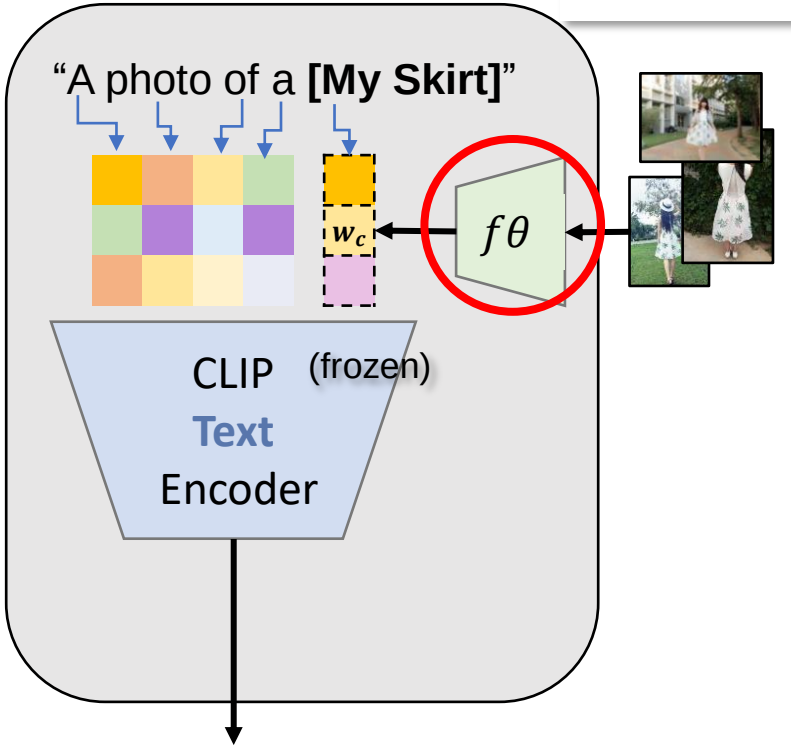
KEY IDEA - @INFERENCE

Example: Personalized album search

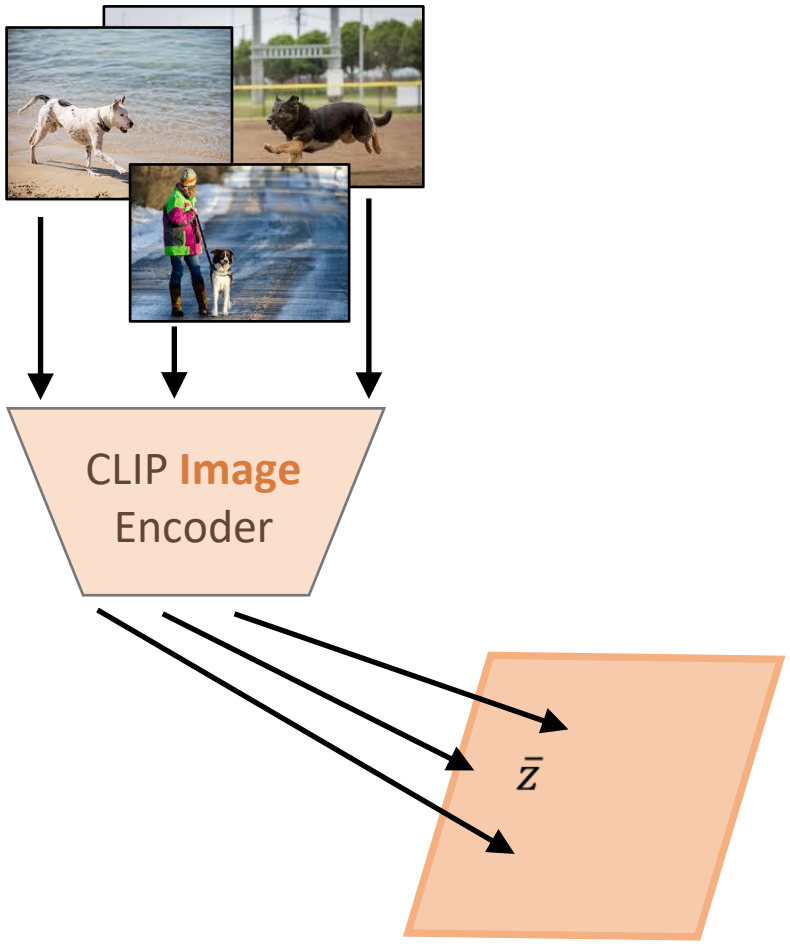


HOW TO LEARN f_θ

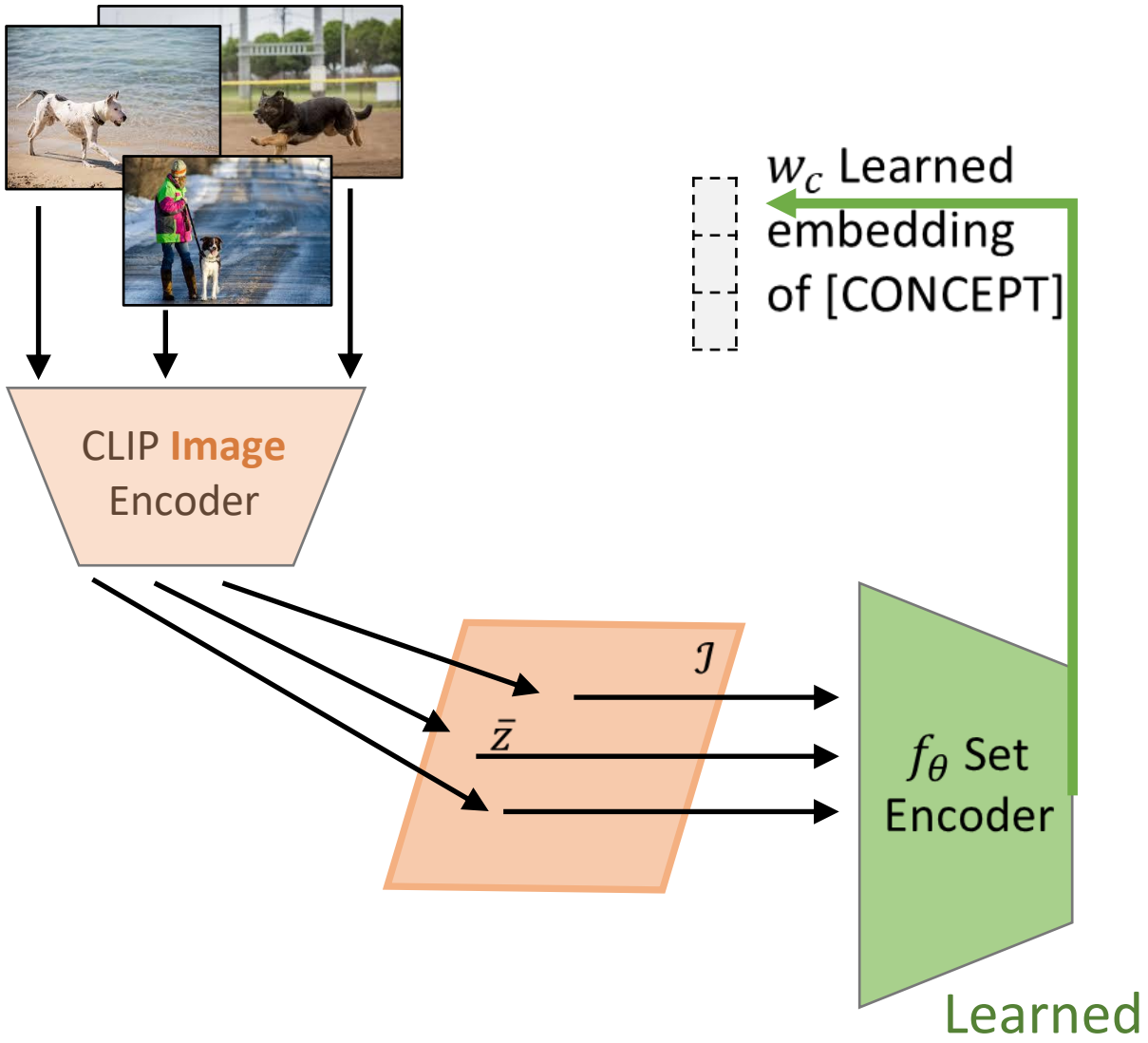
Learn to predict the representation of the new concept



HOW TO LEARN f_θ

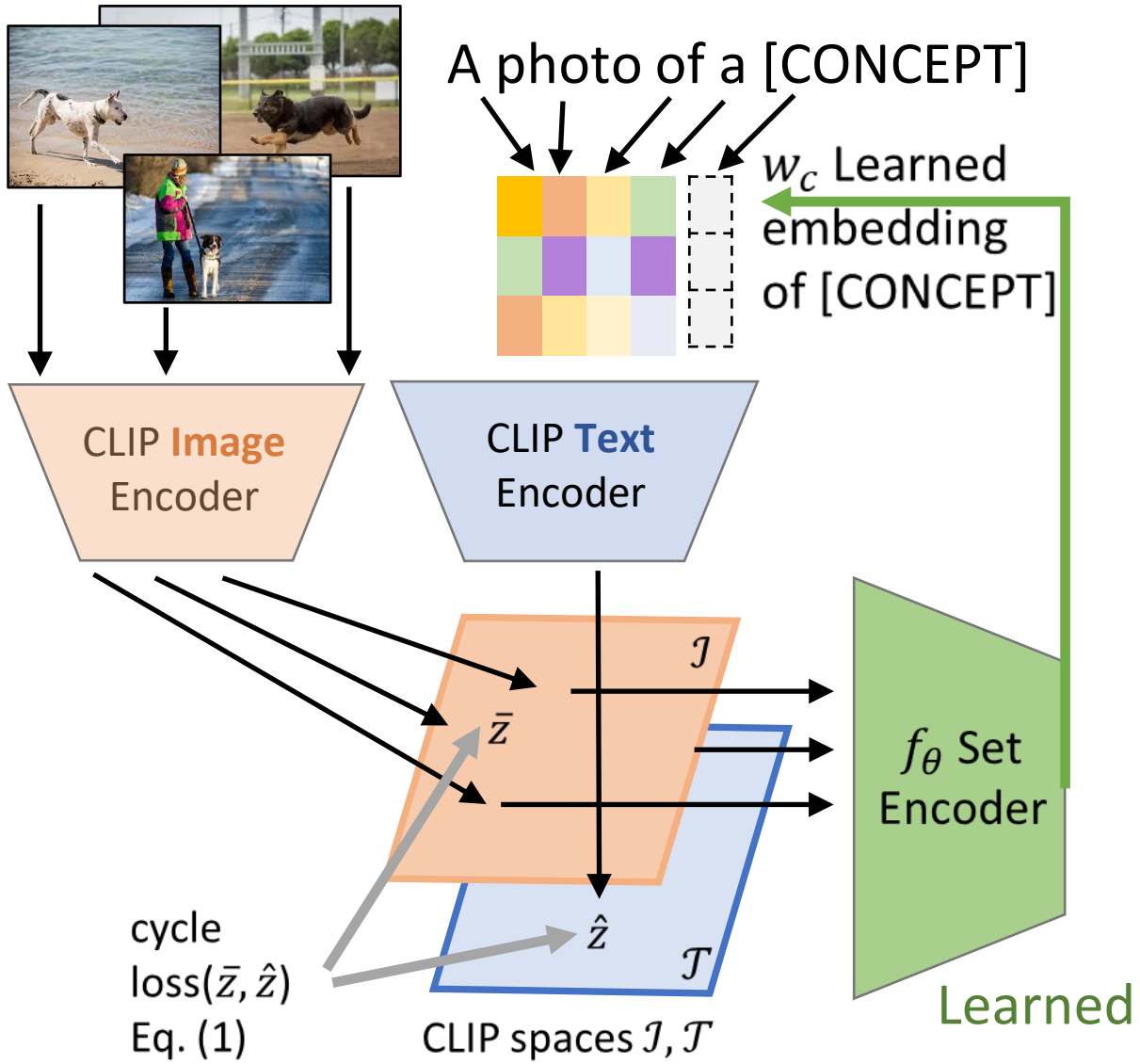


HOW TO LEARN f_θ

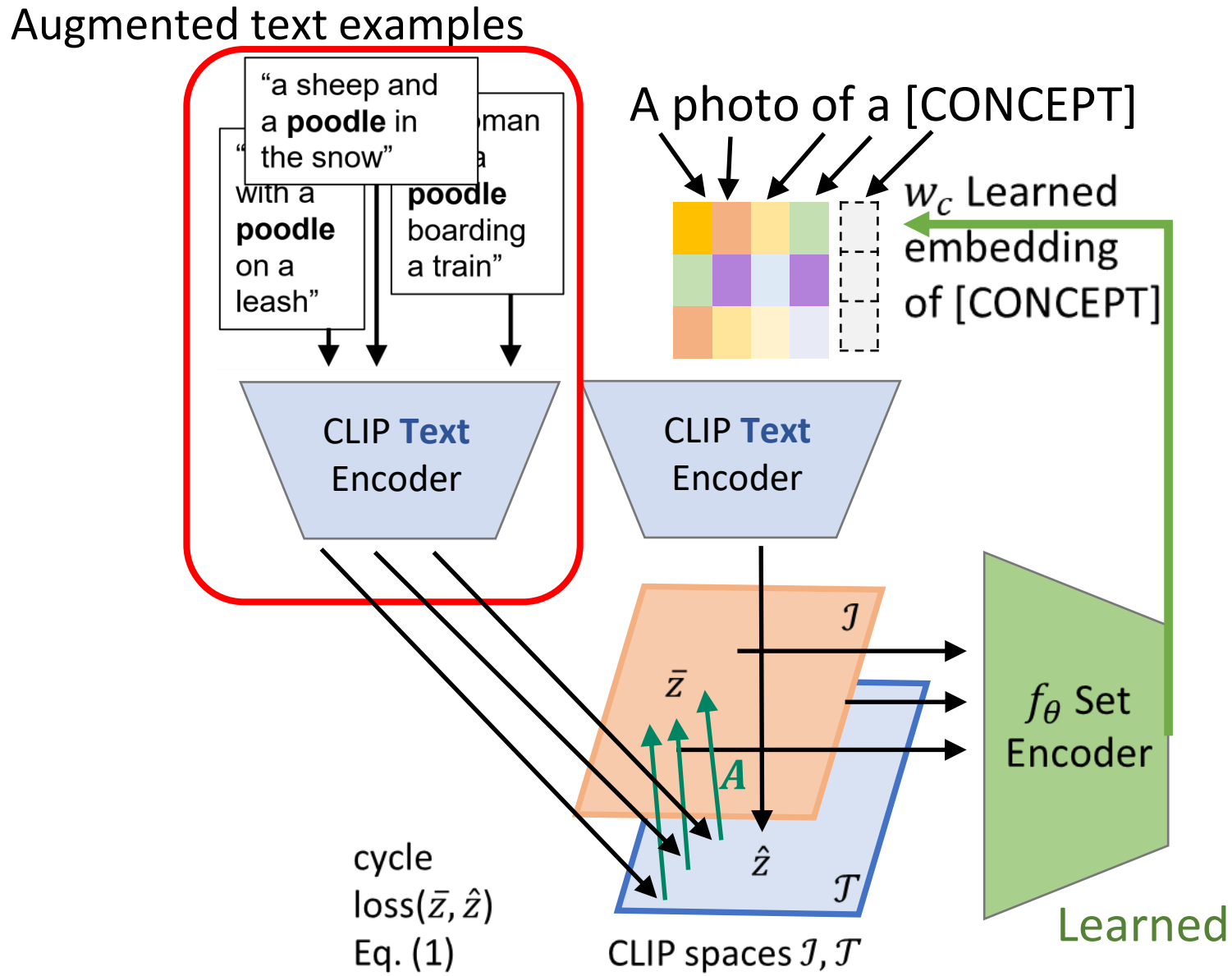


USE A CYCLE LOSS FOR LEARNING f_θ

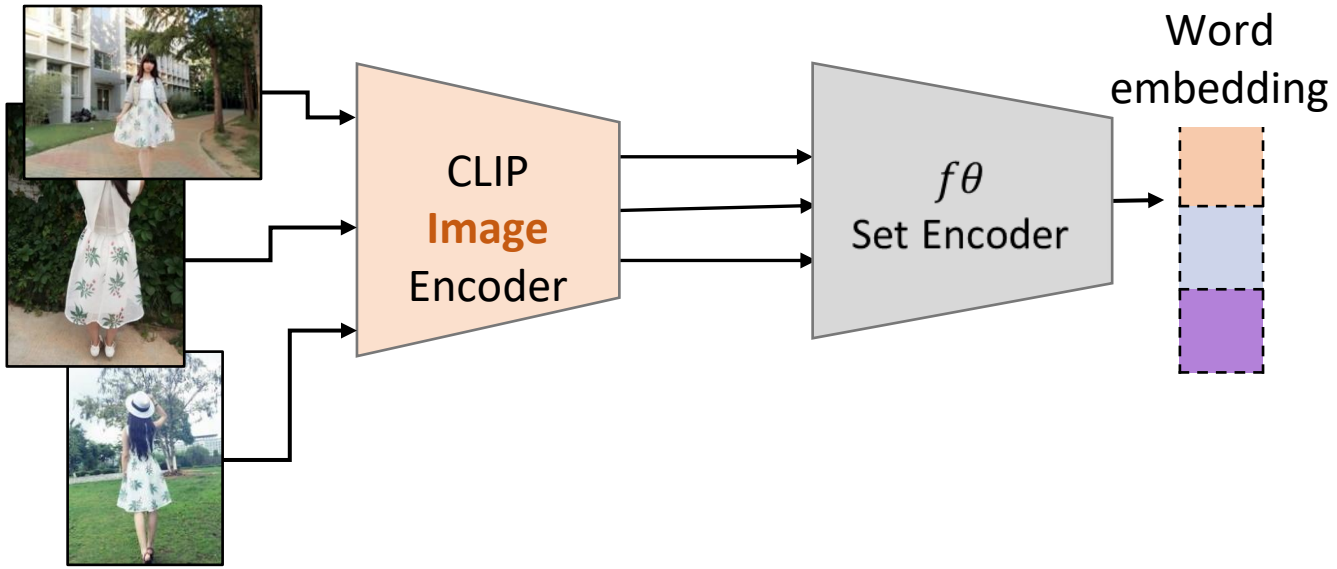
Images of a “dog” from COCO



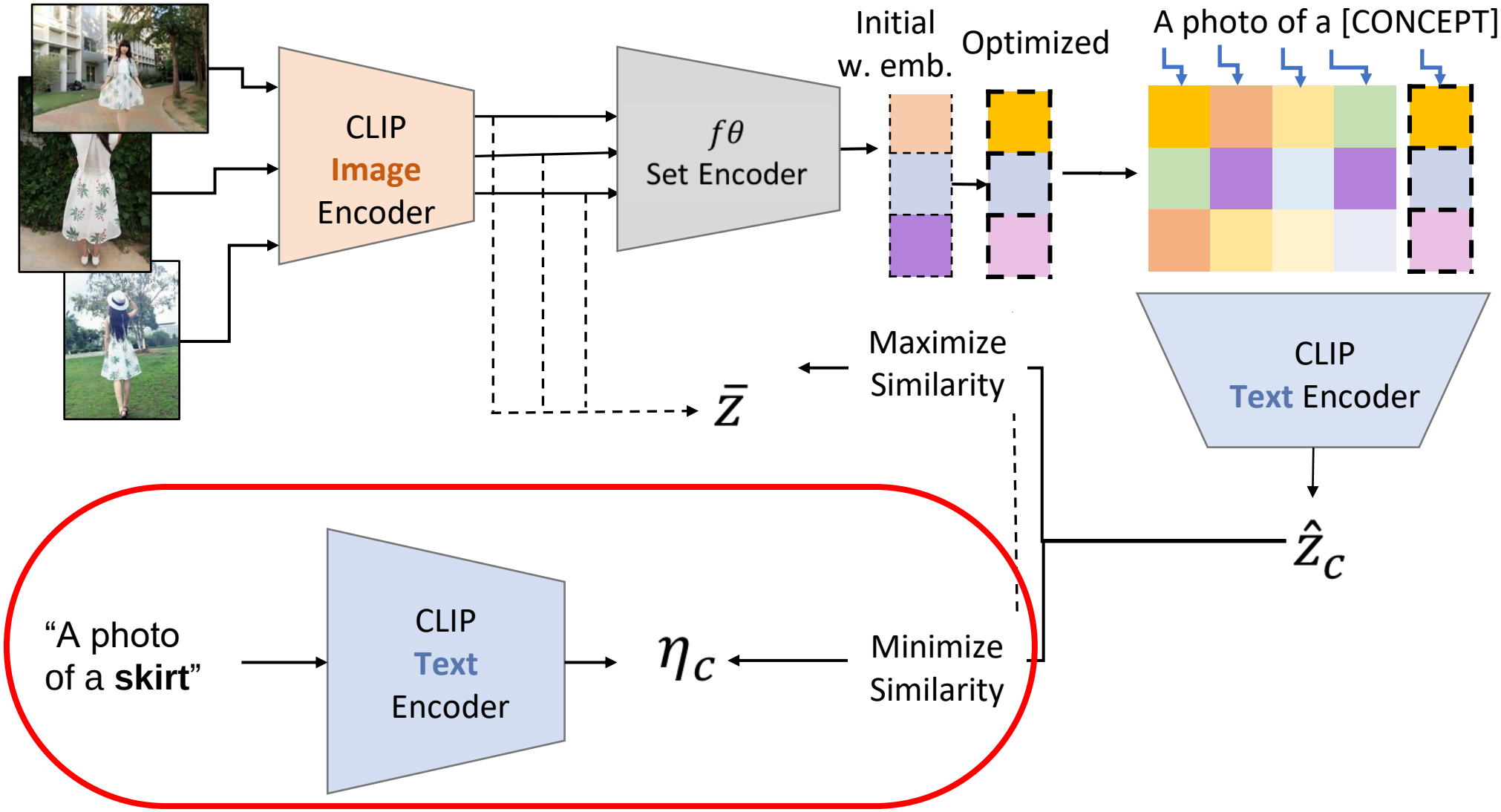
USE A CYCLE LOSS FOR LEARNING f_θ



LEARNING A PERSONALIZED CONCEPT



FURTHER TUNE THE PREDICTED CODE



DATASETS

YouTubeVOS
for Personalized instance segmentation

[CONCEPT] with a white nose patch.

[CONCEPT] is closest to the middle of the wire fence

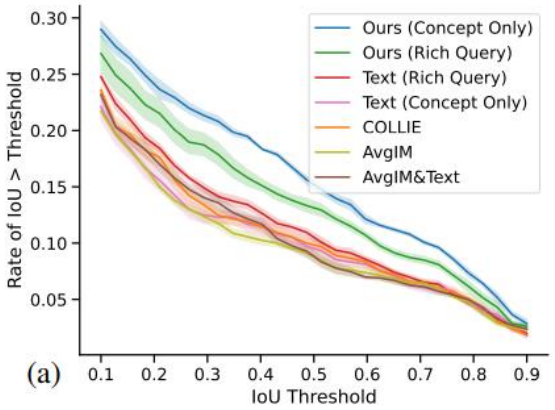
Deepfashion2
for Personalized image retrieval

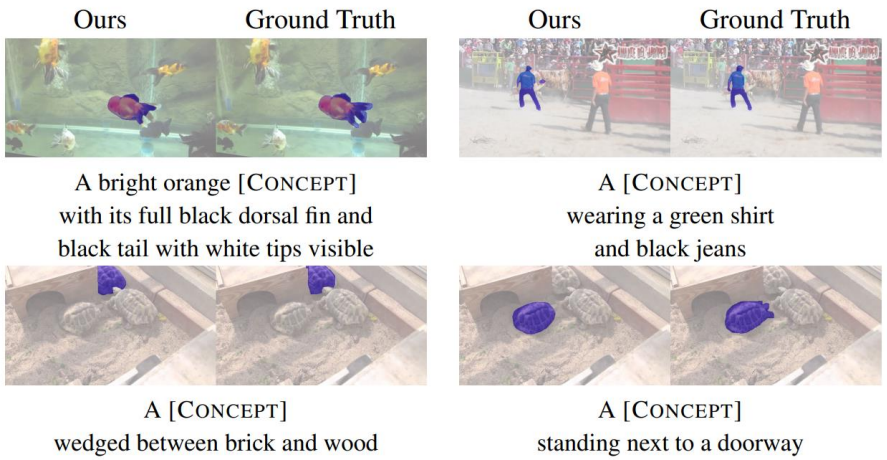
[CONCEPT] is leaning on a rock

A water dispenser is in front of [CONCEPT].

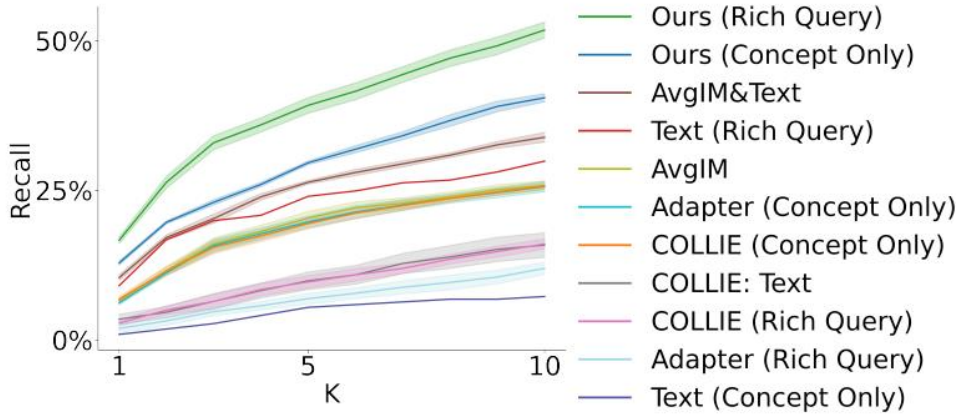
MAIN RESULT

Personalized instance segmentation (YouTubeVOS dataset)





Personalized image retrieval (Deepfashion2 dataset)



SUMMARY

Extend the vocabulary of a pretrained vision and language model, with novel personalized concepts.

Learn to map a set of images to word embeddings using a cycle loss, with either image or augmented text.

Further tune a word-embedding by distinguishing it from a “super-concept”

Inference: Simply use the word embedding, as just another word in the vocabulary of the pretrained model