JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Meta-Personalizing Vision-Language Models to Find Named Instances in Video

Chun-Hsiao Yeh[1,3], Bryan Russell[3], Josef Sivic[2,3], Fabian Caba[3], Simon Jenni[3]

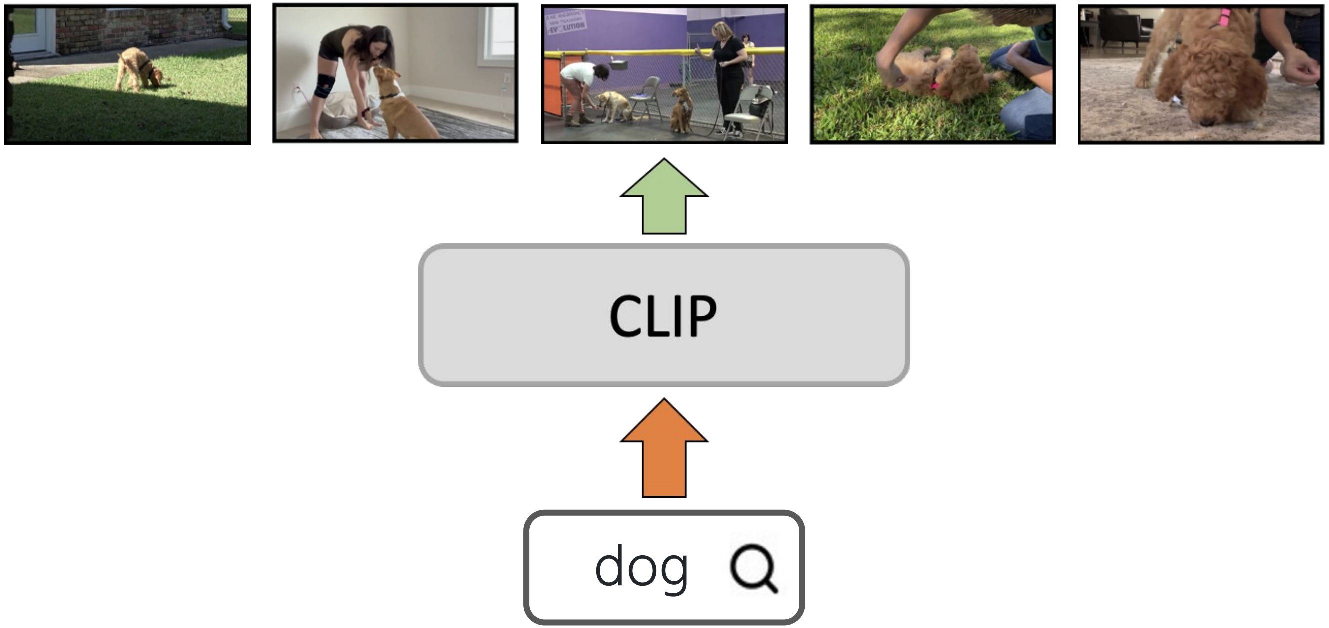
[1]UC Berkeley [2]CIIRC, CTU [3]Adobe Research

Poster Session:
THU-AM-252

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Can Vision-Language Models Perform Personal Instance Retrieval Tasks?

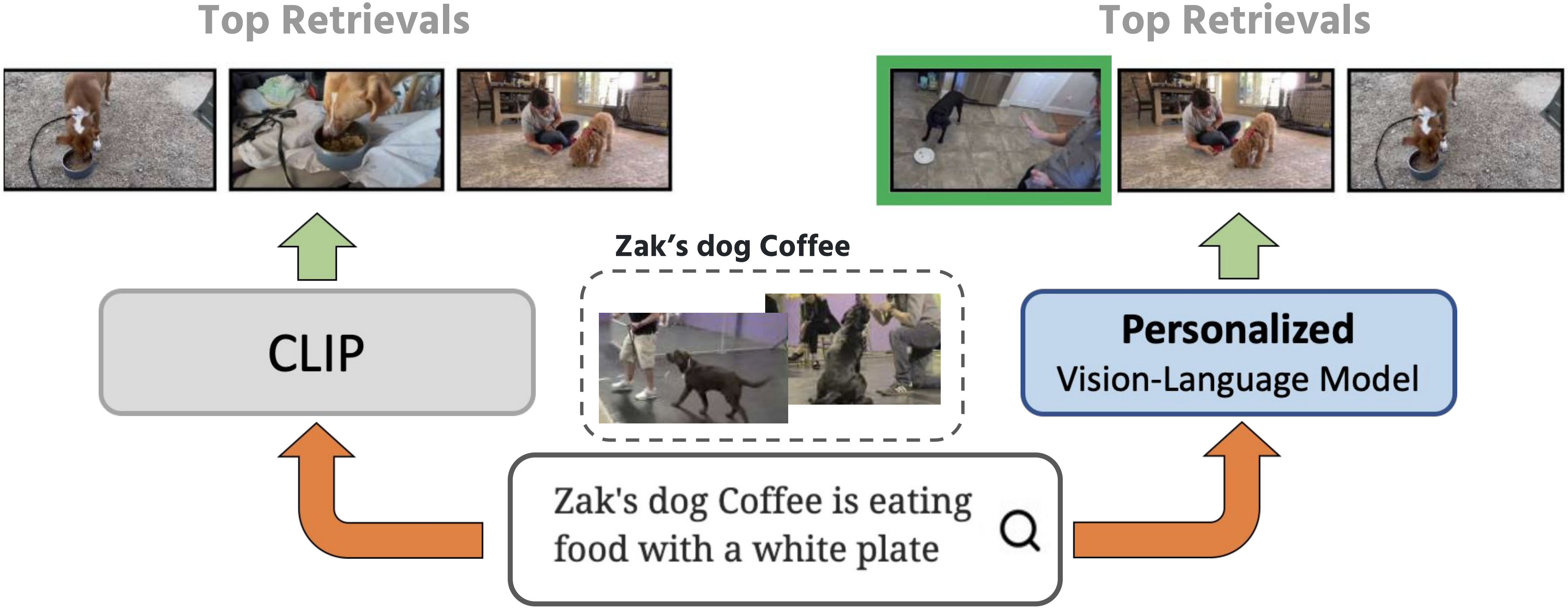
Retrievals from a Video Collection



CLIP

dog

Can Vision-Language Models Perform Personal Instance Retrieval Tasks?



Top Retrievals

Top Retrievals

CLIP

Zak's dog Coffee

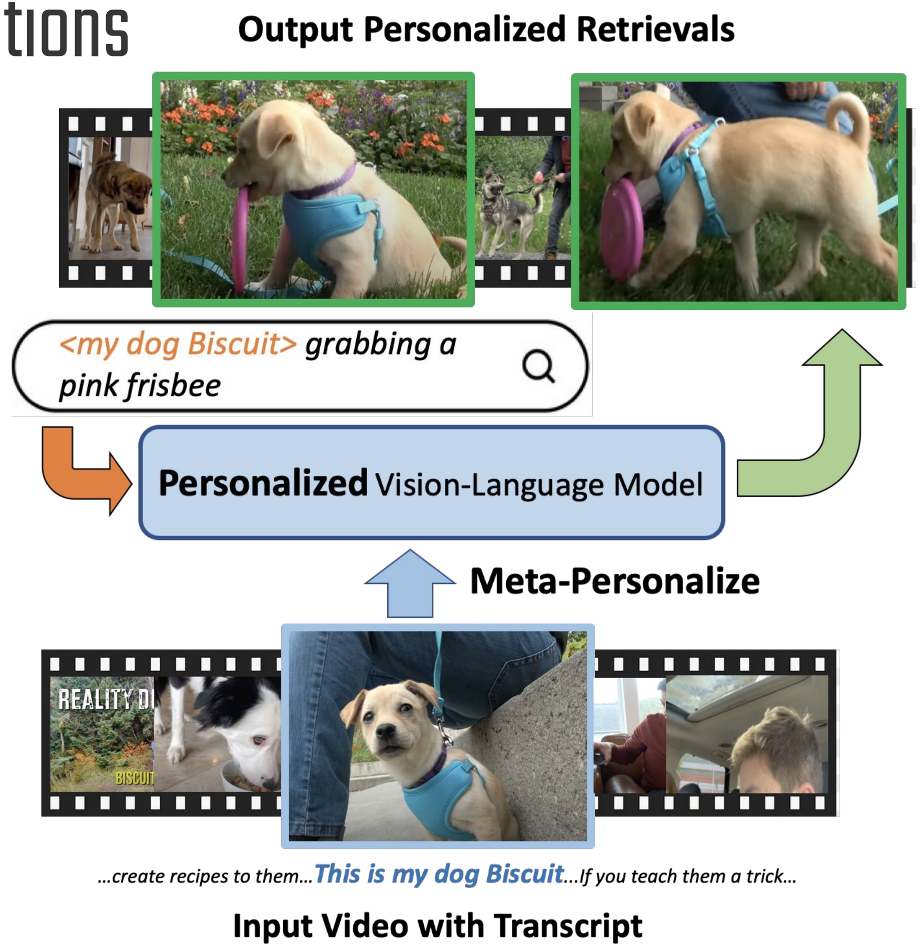
Personalized
Vision-Language Model

Zak's dog Coffee is eating food with a white plate

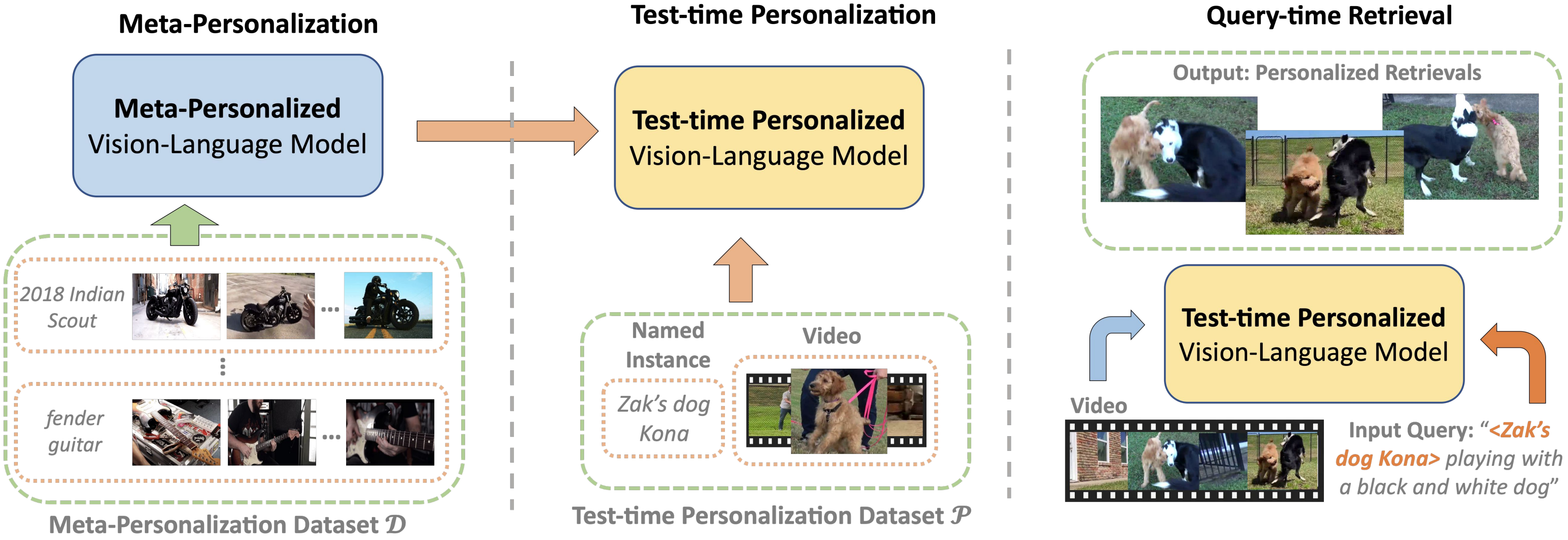
Contributions

Challenge #1:
How to adapt a vision-language model to learn a novel instance without overfitting?

→ **Meta-Personalization**


Output Personalized Retrievals
<my dog Biscuit> grabbing a pink frisbee
Personalized Vision-Language Model
Meta-Personalize
REALITY DI
BISCUIT
...create recipes to them...This is my dog Biscuit...If you teach them a trick...
Input Video with Transcript

Our Meta-Personalized VLM



Meta-Personalization

Meta-Personalized Vision-Language Model

2018 Indian Scout

fender guitar

Meta-Personalization Dataset $\mathcal{D}$

Test-time Personalization

Test-time Personalized Vision-Language Model

Named Instance

Video

Zak's dog Kona

Test-time Personalization Dataset $\mathcal{P}$

Query-time Retrieval

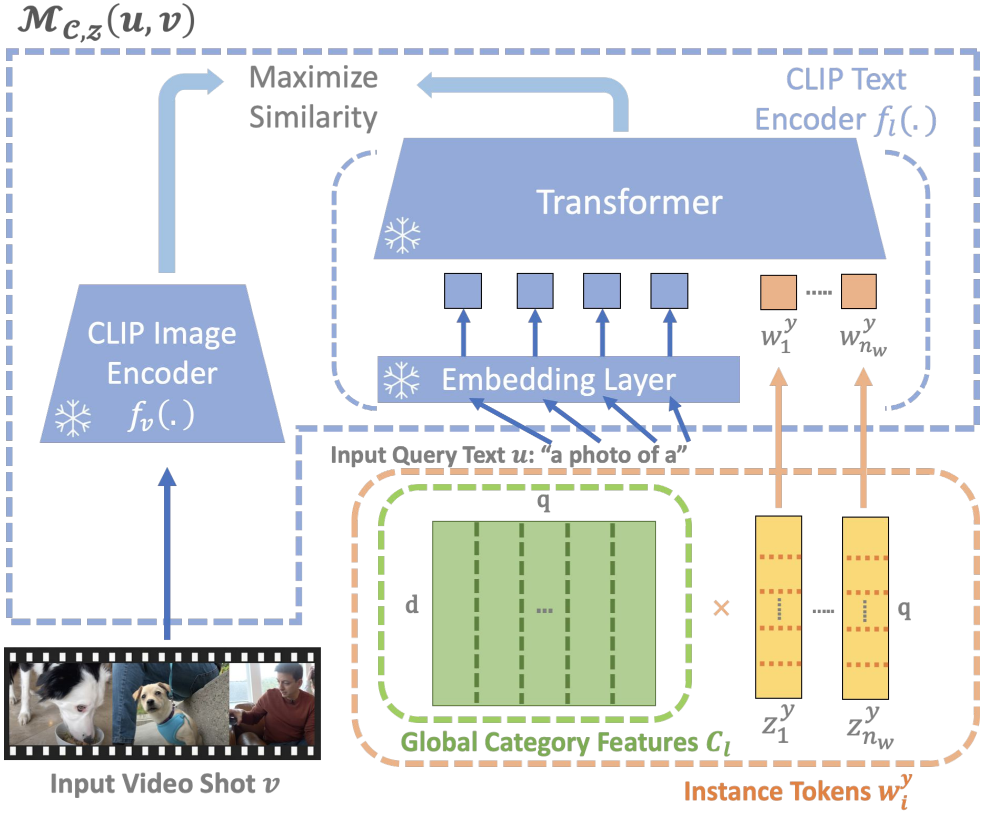
Output: Personalized Retrievals

Test-time Personalized Vision-Language Model

Video

Input Query: "***<Zak's dog Kona>*** *playing with a black and white dog*"

Our Meta-Personalized VLM


$\mathcal{M}_{\mathcal{C},z}(u,v)$
Maximize Similarity
CLIP Text Encoder $f_l(.)$
Transformer
CLIP Image Encoder $f_v(.)$
Embedding Layer
w_1^y
$w_{n_w}^y$
Input Query Text u: "a photo of a"
q
d
...
×
q
z_1^y
$z_{n_w}^y$
Global Category Features C_l
Input Video Shot v
Instance Tokens w_i^y

Highlights:

- **Personal instance tokens w** are a combination of:
  - **Global category features C**
  - **Instance-specific weights z**

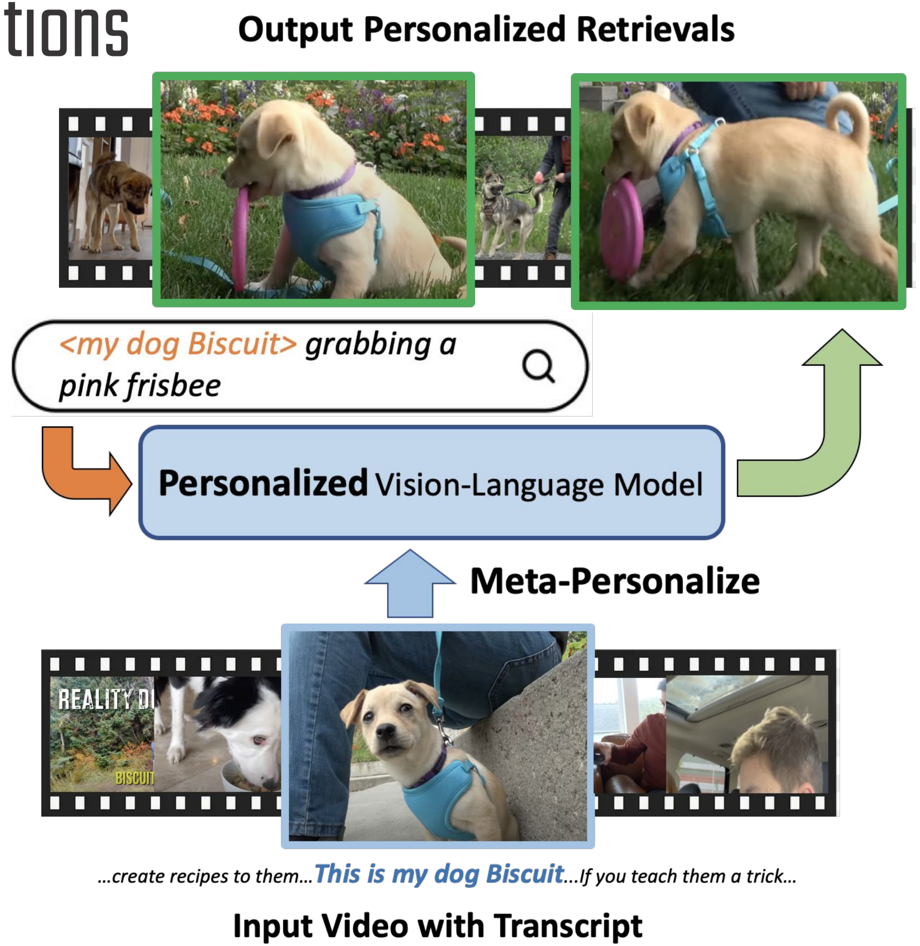
- The columns of C could correspond to attributes of an object category (e.g., color, brand, type of car)

Contributions

Challenge #2:
How to obtain data for training the meta-personalization approach?

→ **Automatic Mining of Named Instances**

→ ***This-Is-My* Benchmark**



Output Personalized Retrievals

<my dog Biscuit> grabbing a pink frisbee

Personalized Vision-Language Model

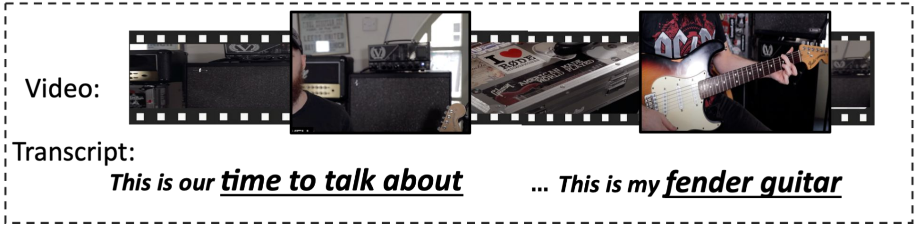
Meta-Personalize

*...create recipes to them...**This is my dog Biscuit**...If you teach them a trick...*

Input Video with Transcript

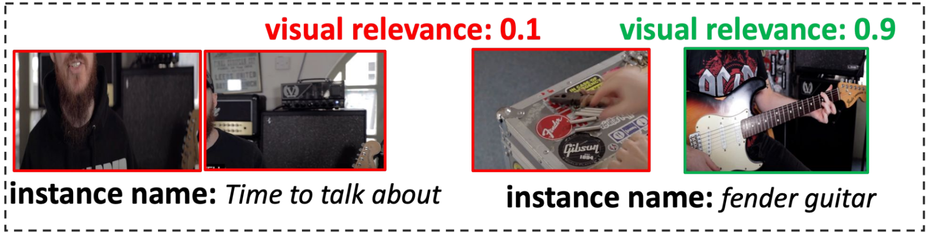
Automatic Mining of Named Instances

Spotting Named Instances (Step 1)



Step 1 finds named instances via possessive patterns (e.g., "This is my *") in video transcripts

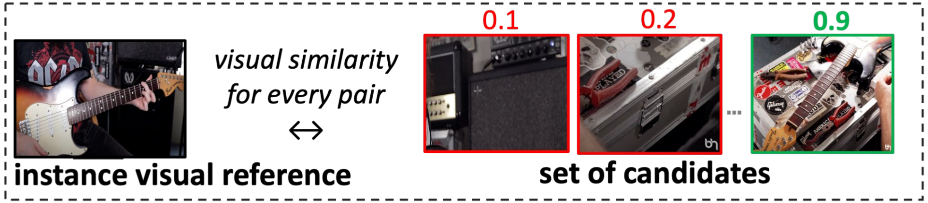
Filtering Non-visual Instances (Step 2)


visual relevance: 0.1
visual relevance: 0.9
instance name: Time to talk about
instance name: fender guitar

Step 2 filters non-visual instances using text-to-visual relevance between

- Instance name
- Video shots neighboring the named instance

Finding Additional Instance Examples (Step 3)



Step 3 retrieves additional shots with high visual similarity to the instance reference shot

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

This-Is-My Dataset

Meta-Personalization Dataset $\mathcal{D}$ (samples: 49256; instances: 2908)

Named Instance:
fender guitar

Named Instance:
first time coming dog

Test-time Personalization Dataset $\mathcal{P}$ (samples: 686; instances: 15)

Named Instance: *Alex's piano*

Named Instance: *Zak's dog Coffee*

Query-time Dataset $\mathcal{Q}$ (samples: 30; instances: 15)

Input Query: *"a woman is raising her hand in front of Alex's piano"*

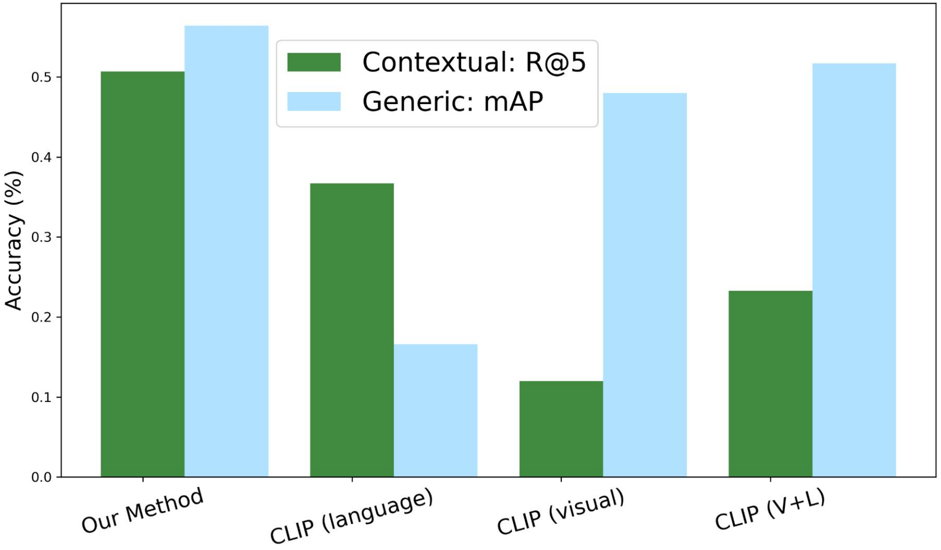
Input Query: *"a woman is standing in front of a birthday cake and Alex's piano"*

Input Query: *"Zak's dog Coffee is lying down around a man and women"*

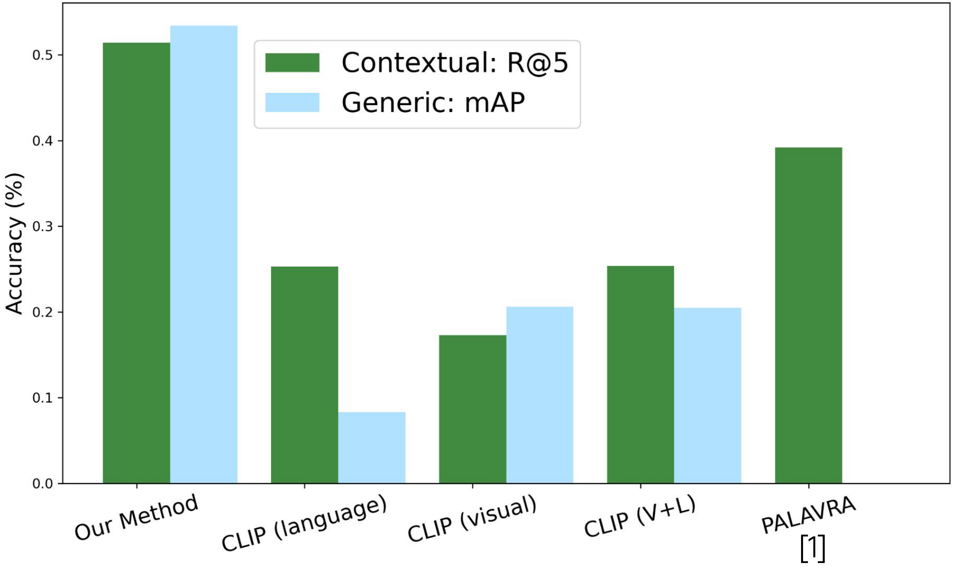
Input Query: *"a man is leading Zak's dog Coffee with a leash"*

Quantitative Retrieval Results

This-Is-My Video Retrieval



DeepFashion2 Fashion Item Retrieval



Contextualized Retrieval: Queries describe a specific context, e.g., "A photo of * lying on the beach." **(single correct match)**

Generic Retrieval: Queries correspond to the generic prompt "an image of * ". **(multiple correct matches)**

[1] Cohen, Niv, et al. ""This is my unicorn, Fluffy": Personalizing frozen vision-language representations." Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX. Cham: Springer Nature Switzerland, 2022.

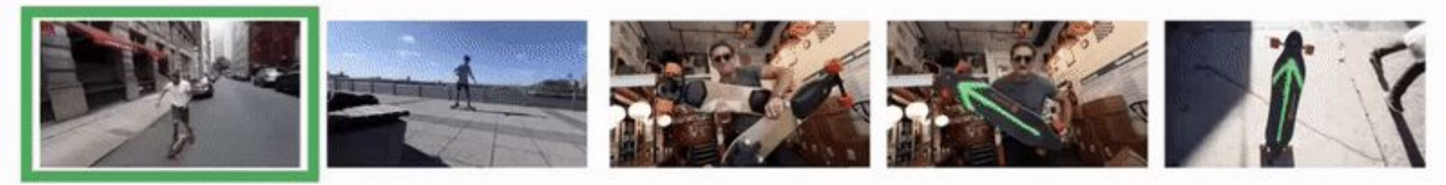
Qualitative Retrieval Results

Top-5 Personalized Retrievals

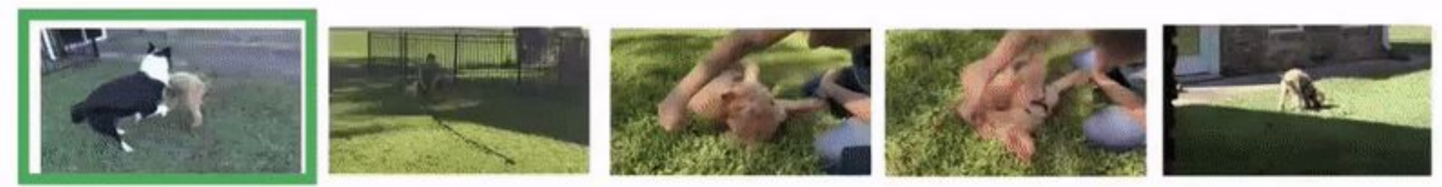
Language Queries

Success Retrieval

a man is riding ***<Casey's boosted board>*** and wearing white t-shirt and gray shorts



<Zak's dog Kona> is playing with a black and white dog on the grass



<Zak's dog Coffee> is lying down in front of a man and three women