CVPR Nashville JUNE 11-15, 2025

Topics
Multimodal Models
Personalization

Authors
Thao Nguyen[1], Krishna Kumar Singh[2], Jing Shi[2],
Trung Bui[2], Yong Jae Lee[1,*], Yuheng Li[2,*]

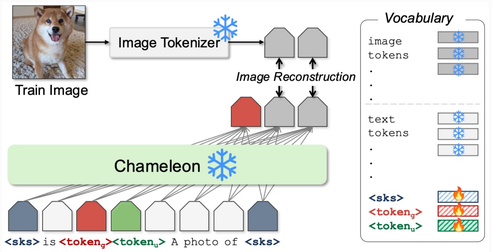
Affiliation
[1]UW–Madison
[2]Adobe Research

Your Personalized Vision & Language Assistant

APPROACH (YO'CHAMLEON)

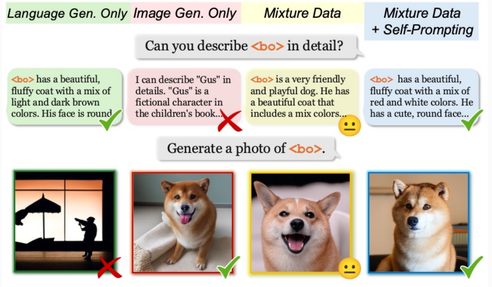
We represent personalized subjects as learnable prompts for LMMs:

"<sks> is <g-tokens><u-tokens>.



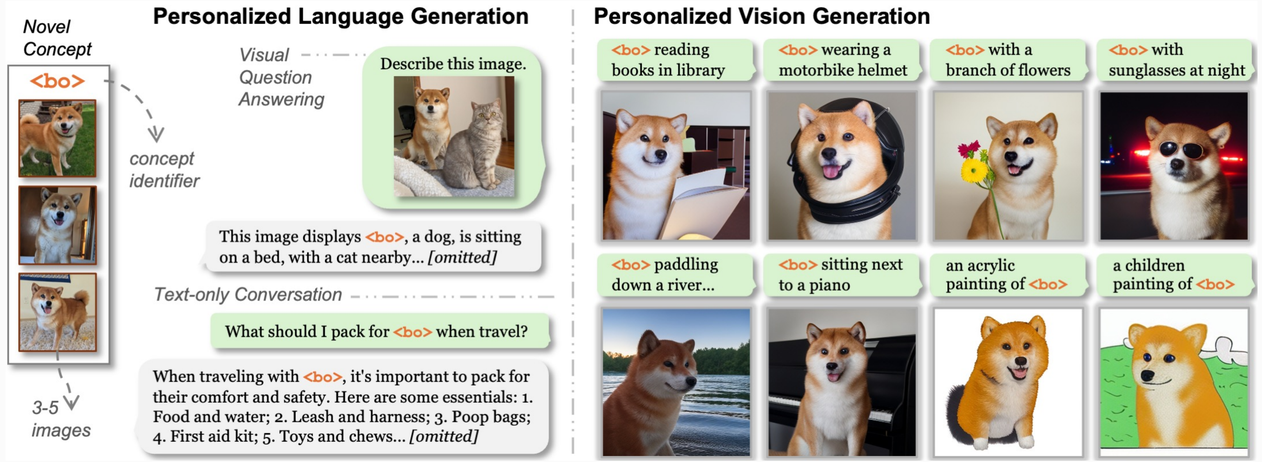
Training data:
(1) Understanding: Recognition & Question-Answering (Yo'LLaVA);
(2) Generation: Image Reconstruction.

WHY DIFFICULT?



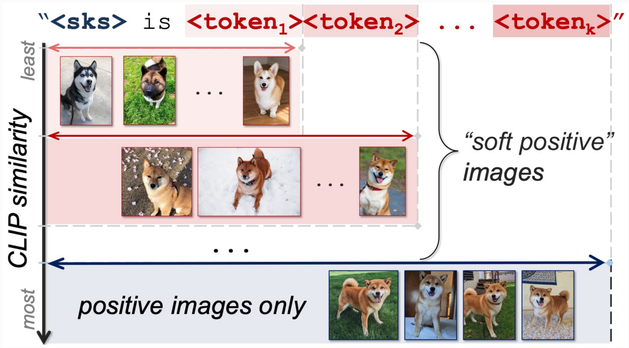
Optimized tokens for *one task* cannot *effectively perform another.*
And training on a mixture of data yields *suboptimal* performance across tasks.

How to make LMMs understand personalized concept?



3S: Soft-prompt + Soft-positive + Self-prompting

Soft-prompt: Personalized subject = Soft-prompt.



Soft-positive images.
Retrieved images are ranked according by similarity to positive images. Images that are more similar to the actual positive images are described with more latent tokens (i.e., more details).

Self-prompting mechanism.
Model first predicts which information should be used first, then preform the task.

Prompt	<sks> is <g-tokens><u-tokens>	
	Language Generation	Vision Generation
Question	What kind of subject is <sks>?	A photo of <sks>.
Answer	<u-tokens><sks> is a dog.	<g-tokens><image>

Personalized Large Multimodal Models

novel concept
3-5 images

✓ retain original capabilities
✓ personalized conversations

WHAT'S NEXT

Zero-shot personalization? Multiple concepts? etc.

"This is Bo  *- my dog. Bo is brother of Mắm*  *- my cat..."*

←← Scan QR for paper, page, & code!