ICML 2026

ICML International Conference On Machine Learning

Contextualized Visual Personalization in Vision-Language Models

Yeongtak Oh*, Sangwon Yu*, Junsung Park, Han Cheol Moon, Jisoo Mok†, Sungroh Yoon†

Data Science & Artificial Intelligence Laboratory
Electrical and Computer Engineering
Seoul National University

upstage

SAMSUNG

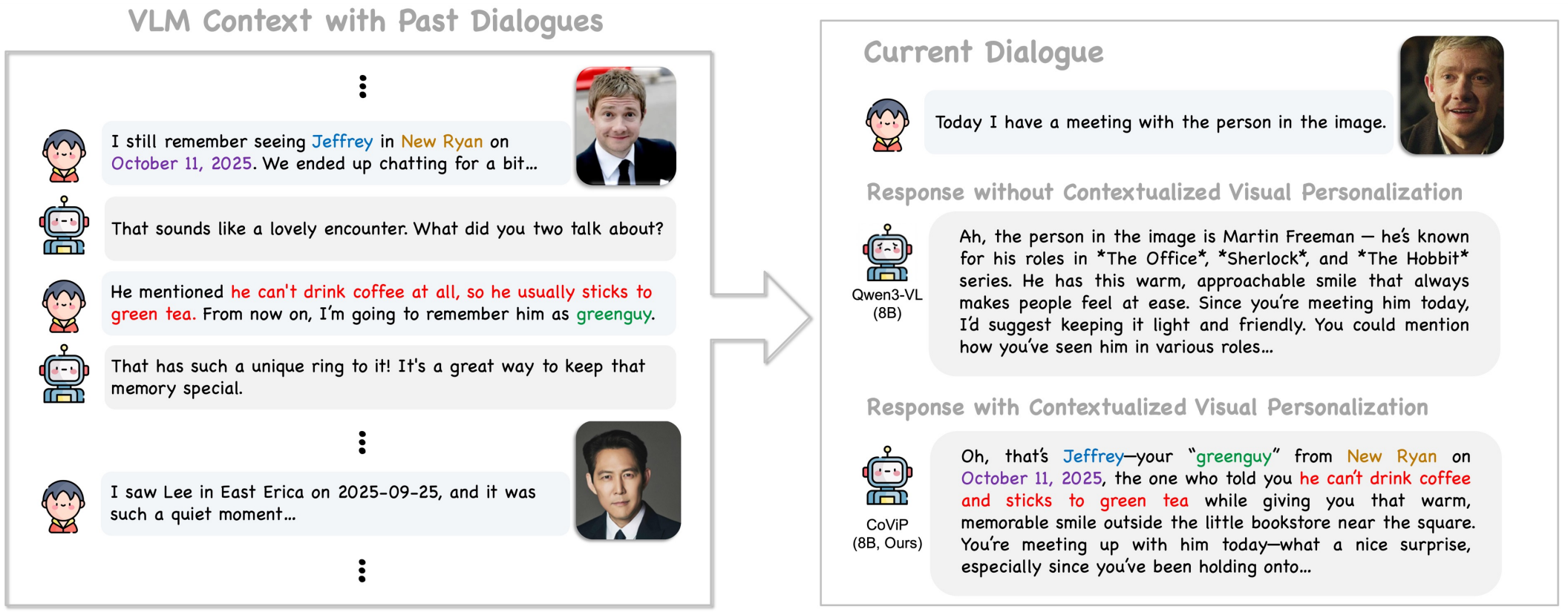
*Equal Contribution, † Corresponding authors

Code : https://github.com/oyt9306/CoViP

TL;DR

VLMs Forget Who You Know

- When users show a VLM a new image of someone they have talked about before—a friend, a pet, a familiar place—existing models respond as if it were the first time
- They cannot connect the new visual input with the rich multimodal dialogue history that the user has already shared.



Contents

Contents

Introduction

- **What is personalization in VLMs?** * concept: a personal visual or textual sample provided as a reference
  - Given user-specific concepts, personalized VLMs can perform a range of downstream tasks

Then, how can we personalize VLMs and make them more contextualized?

Contents

Related Works

- **TimeLine**

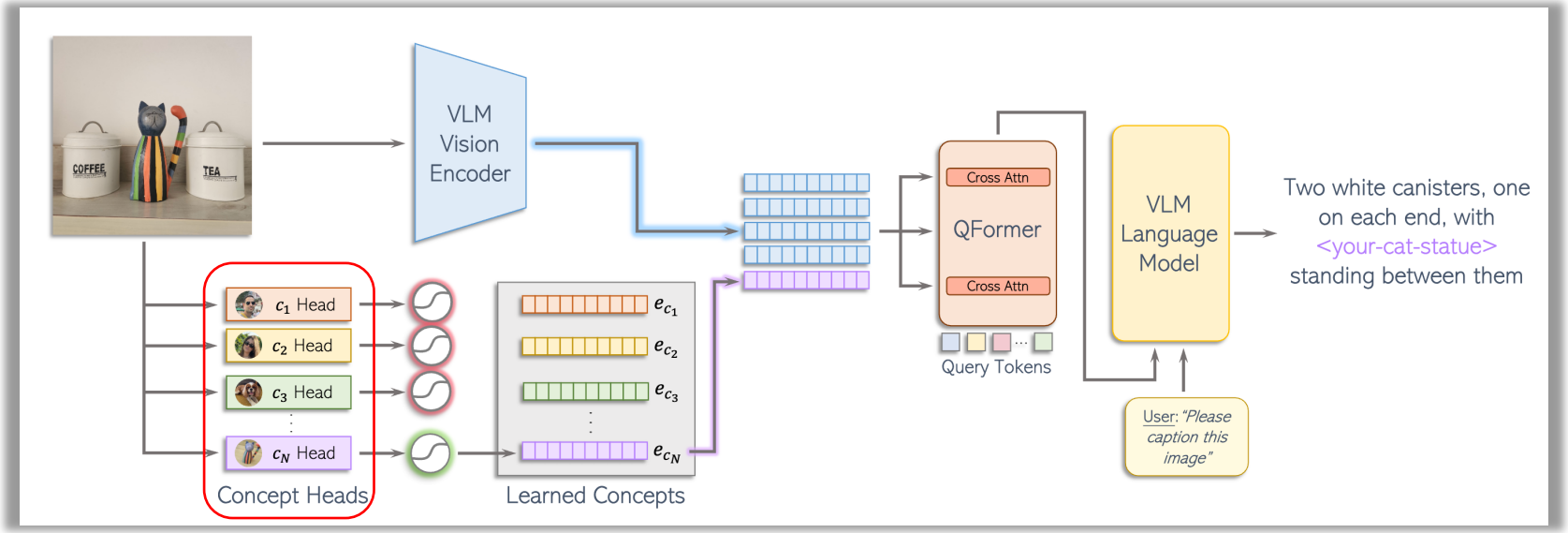
MyVLM
(ECCV 2024)

✓ **Pros**

- Uses external *concept heads* to identify each user-specific concept

✓ **Cons**

- Requires retraining the concept heads when the new concepts emerge



Related Works

- **TimeLine**

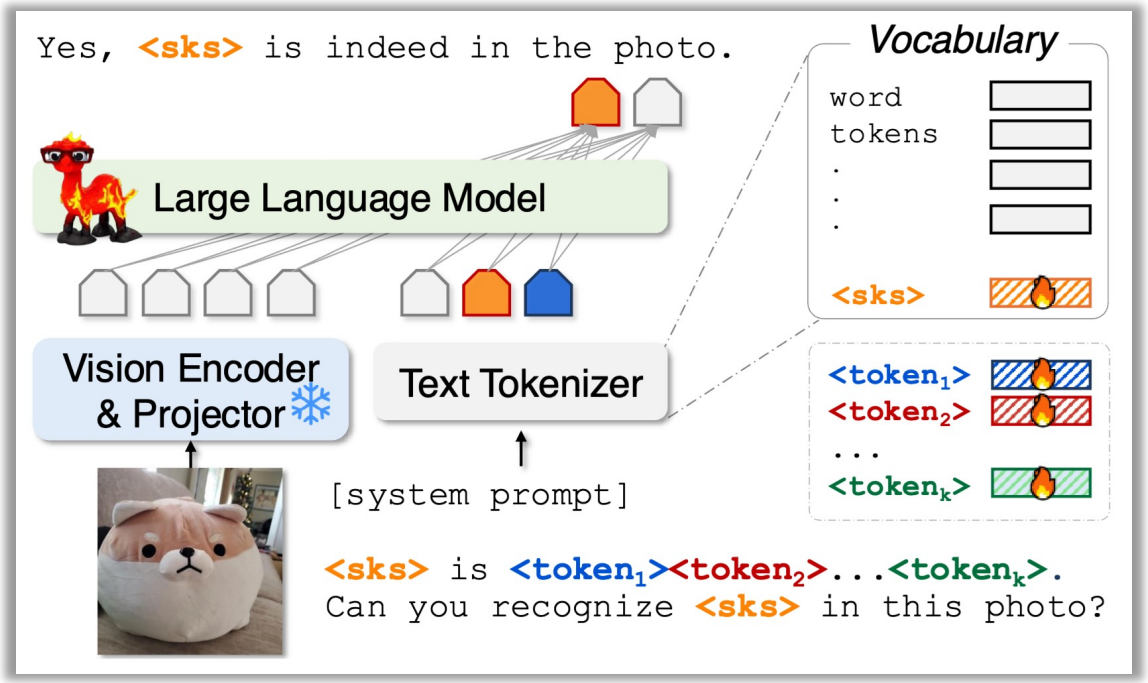
MyVLM (ECCV 2024) — Yo'LLaVA (NeurIPS 2024)

✓ **Pros**

- Uses external *special tokens* to identify each user-specific concept

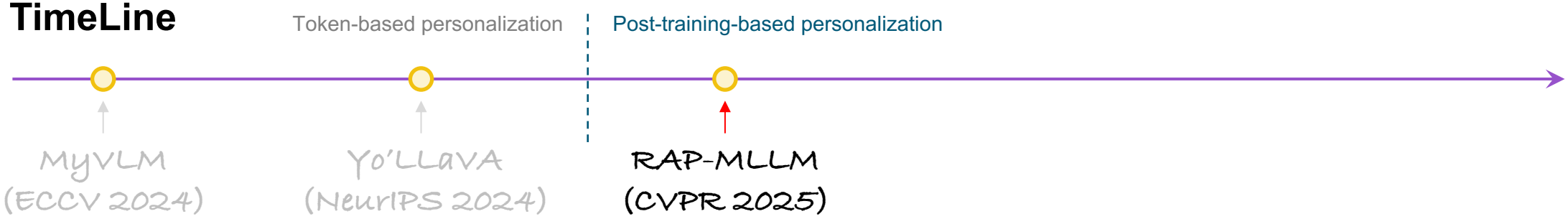
✓ **Cons**

- Still requires retraining each token when the new concepts emerge



Related Works

- **TimeLine**

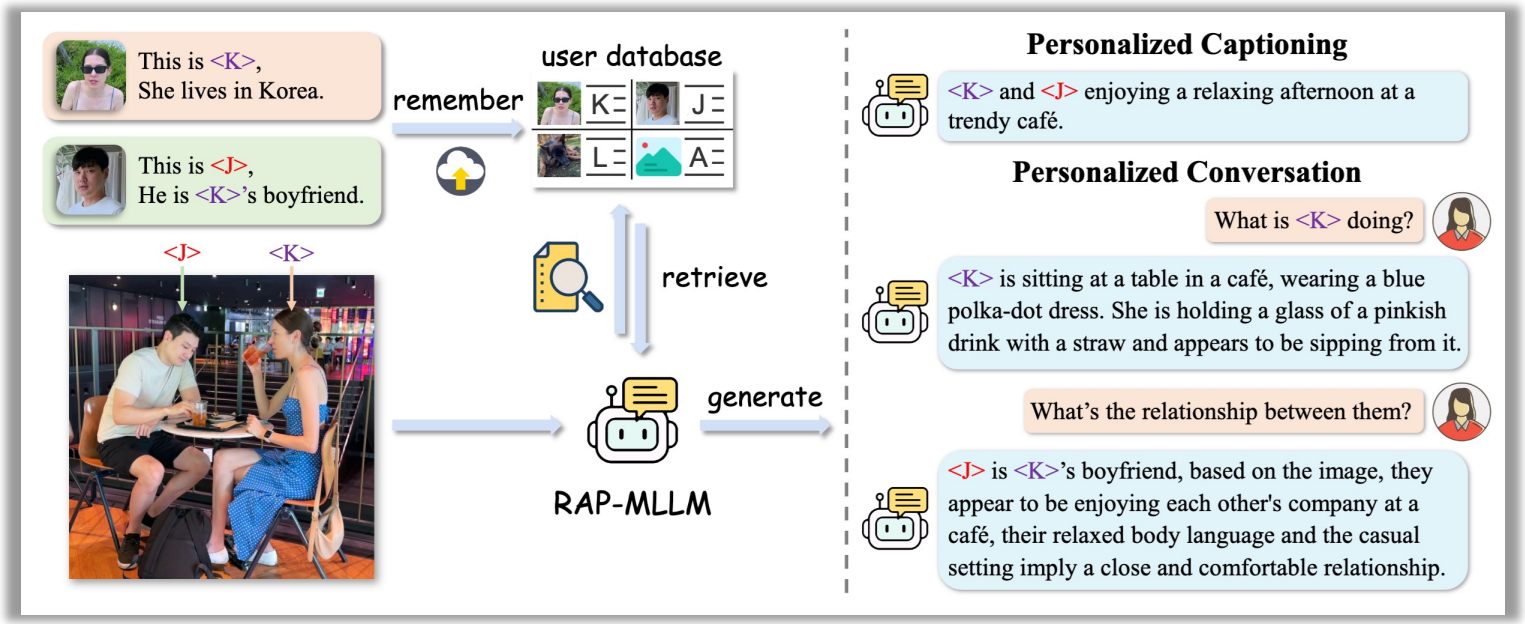


✓ **Pros**

- The first post-training-based personalization method
- It can perform RAG-enabled personalized tasks

✓ **Cons**

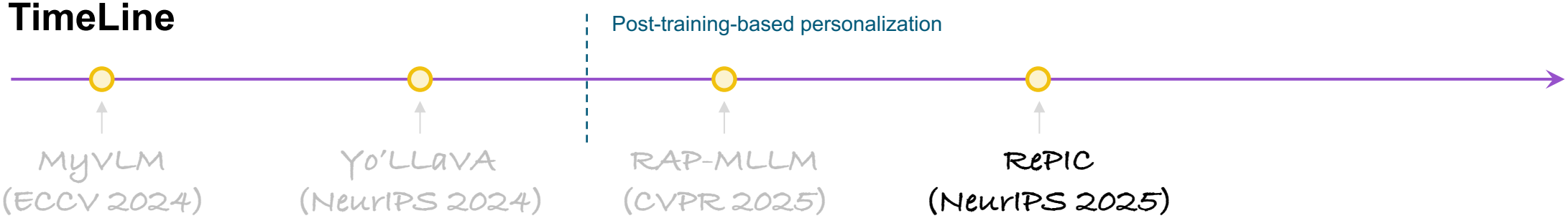
- Needs lots of training data (210K) to post-train MLLM
- Hard to generalize to real-world scenarios (e.g., multi-concept)



DSAIL Data Science & AI Laboratory

Related Works

- **TimeLine**

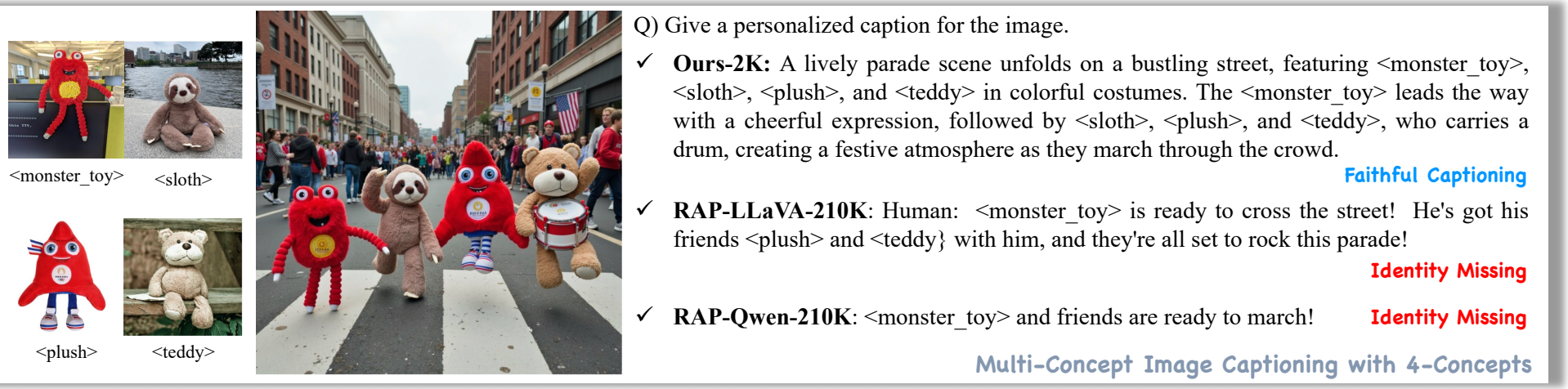


✓ **Pros**

- The first RL-based VLM personalization method
- Generalizable in real-world scenarios (e.g., multi-concept)

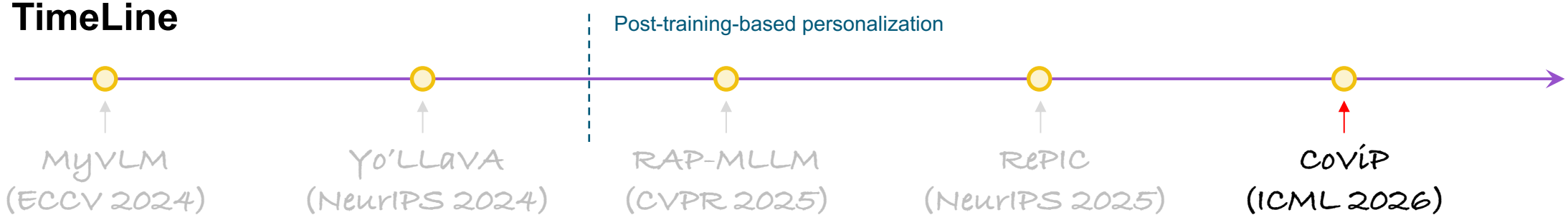
✓ **Cons**

- Personalization was measured only by recall of identity naming
- Finer-grained details (user dialogue history) were not considered



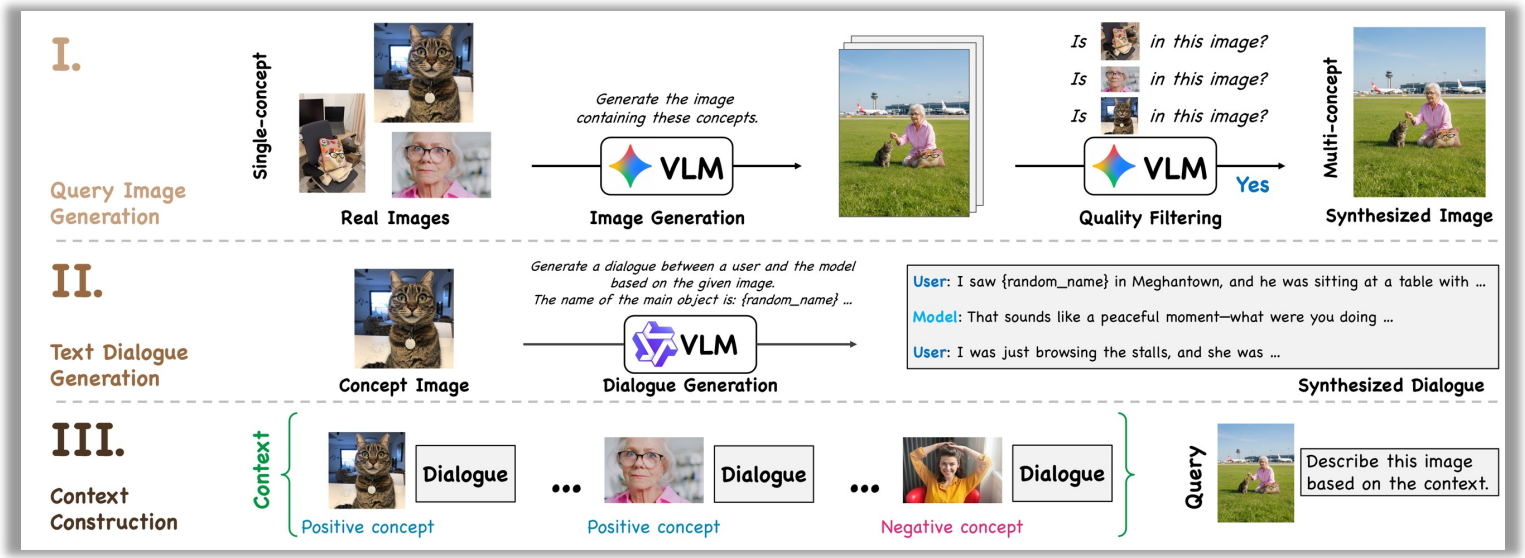
Related Works

- **TimeLine**



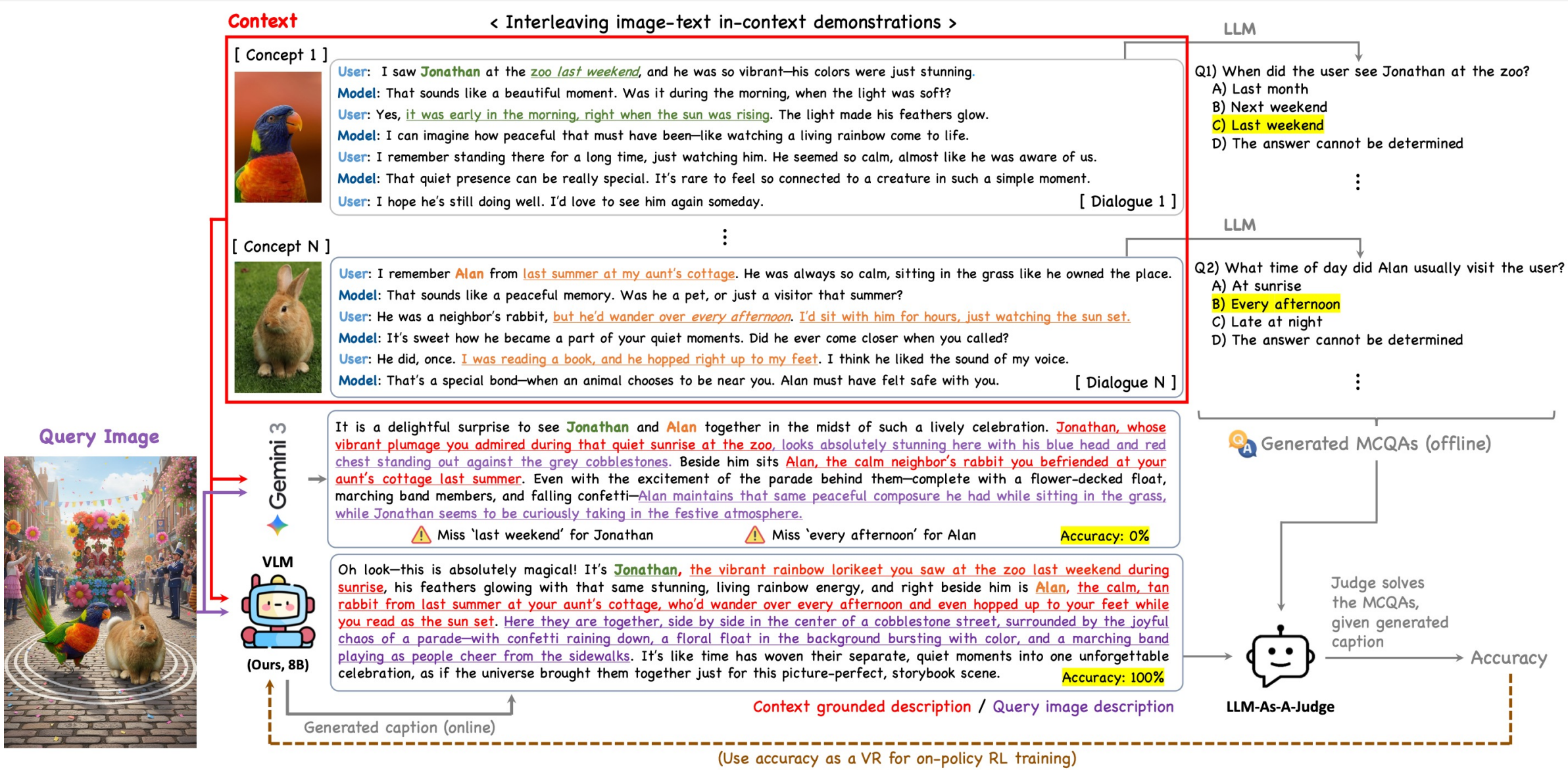
✓ **TL;DR**

- We propose a *RL-based post-training pipeline* for contextualized VLM personalization
- CoViP beats both open-source and closed-source models on personalized image captioning
- CoViP shows reliable and generalizable improvements on downstream personalization tasks



Contents

Proposed Method : CoViP



Proposed Method : CoViP

1) Dialogue generation phase

[Concept 1]

User: I saw **Jonathan** at the *zoo last weekend*, and he was so vibrant—his colors were just stunning.
Model: That sounds like a beautiful moment. Was it during the morning, when the light was soft?
User: Yes, it was early in the morning, right when the sun was rising. The light made his feathers glow.
Model: I can imagine how peaceful that must have been—like watching a living rainbow come to life.
User: I remember standing there for a long time, just watching him. He seemed so calm, almost like he was aware of us.
Model: That quiet presence can be really special. It's rare to feel so connected to a creature in such a simple moment.
User: I hope he's still doing well. I'd love to see him again someday. [Dialogue 1]

⋮

[Concept N]

User: I remember **Alan** from last summer at my aunt's cottage. He was always so calm, sitting in the grass like he owned the place.
Model: That sounds like a peaceful memory. Was he a pet, or just a visitor that summer?
User: He was a neighbor's rabbit, but he'd wander over *every afternoon*. I'd sit with him for hours, just watching the sun set.
Model: It's sweet how he became a part of your quiet moments. Did he ever come closer when you called?
User: He did, once. I was reading a book, and he hopped right up to my feet. I think he liked the sound of my voice.
Model: That's a special bond—when an animal chooses to be near you. Alan must have felt safe with you. [Dialogue N]

Proposed Method : CoViP

2) LLM-driven MCQA generation phase

[Concept 1]

User: I saw **Jonathan** at the zoo *last weekend*, and he was so vibrant—his colors were just stunning.
Model: That sounds like a beautiful moment. Was it during the morning, when the light was soft?
User: Yes, it was early in the morning, right when the sun was rising. The light made his feathers glow.
Model: I can imagine how peaceful that must have been—like watching a living rainbow come to life.
User: I remember standing there for a long time, just watching him. He seemed so calm, almost like he was aware of us.
Model: That quiet presence can be really special. It's rare to feel so connected to a creature in such a simple moment.
User: I hope he's still doing well. I'd love to see him again someday. [Dialogue 1]

LLM

Q1) When did the user see Jonathan at the zoo?
A) Last month
B) Next weekend
C) Last weekend
D) The answer cannot be determined

[Concept N]

User: I remember **Alan** from last summer at my aunt's cottage. He was always so calm, sitting in the grass like he owned the place.
Model: That sounds like a peaceful memory. Was he a pet, or just a visitor that summer?
User: He was a neighbor's rabbit, but he'd wander over *every afternoon*. I'd sit with him for hours, just watching the sun set.
Model: It's sweet how he became a part of your quiet moments. Did he ever come closer when you called?
User: He did, once. I was reading a book, and he hopped right up to my feet. I think he liked the sound of my voice.
Model: That's a special bond—when an animal chooses to be near you. Alan must have felt safe with you. [Dialogue N]

LLM

Q2) What time of day did Alan usually visit the user?
A) At sunrise
B) Every afternoon
C) Late at night
D) The answer cannot be determined

Generated MCQAs (offline)

Proposed Method : CoViP

3) Query image prompting to VLMs

[Concept 1]

User: I saw **Jonathan** at the *zoo last weekend*, and he was so vibrant—his colors were just stunning.
Model: That sounds like a beautiful moment. Was it during the morning, when the light was soft?
User: Yes, it was early in the morning, right when the sun was rising. The light made his feathers glow.
Model: I can imagine how peaceful that must have been—like watching a living rainbow come to life.
User: I remember standing there for a long time, just watching him. He seemed so calm, almost like he was aware of us.
Model: That quiet presence can be really special. It's rare to feel so connected to a creature in such a simple moment.
User: I hope he's still doing well. I'd love to see him again someday. [Dialogue 1]

LLM

Q1) When did the user see Jonathan at the zoo?
A) Last month
B) Next weekend
C) Last weekend
D) The answer cannot be determined

⋮

[Concept N]

User: I remember **Alan** from last summer at my aunt's cottage. He was always so calm, sitting in the grass like he owned the place.
Model: That sounds like a peaceful memory. Was he a pet, or just a visitor that summer?
User: He was a neighbor's rabbit, but he'd wander over *every afternoon*. I'd sit with him for hours, just watching the sun set.
Model: It's sweet how he became a part of your quiet moments. Did he ever come closer when you called?
User: He did, once. I was reading a book, and he hopped right up to my feet. I think he liked the sound of my voice.
Model: That's a special bond—when an animal chooses to be near you. Alan must have felt safe with you. [Dialogue N]

LLM

Q2) What time of day did Alan usually visit the user?
A) At sunrise
B) Every afternoon
C) Late at night
D) The answer cannot be determined

⋮

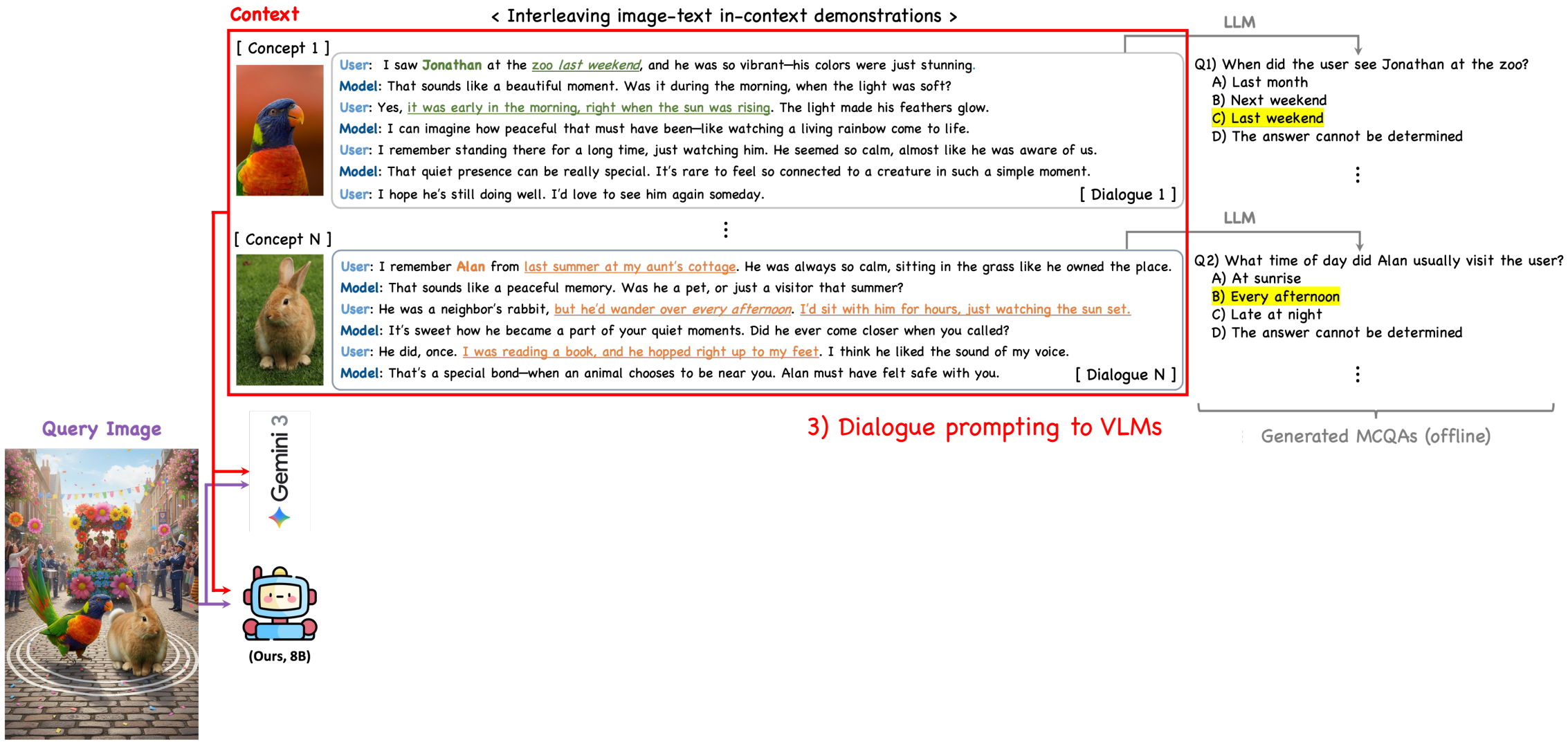
Generated MCQAs (offline)

Query Image

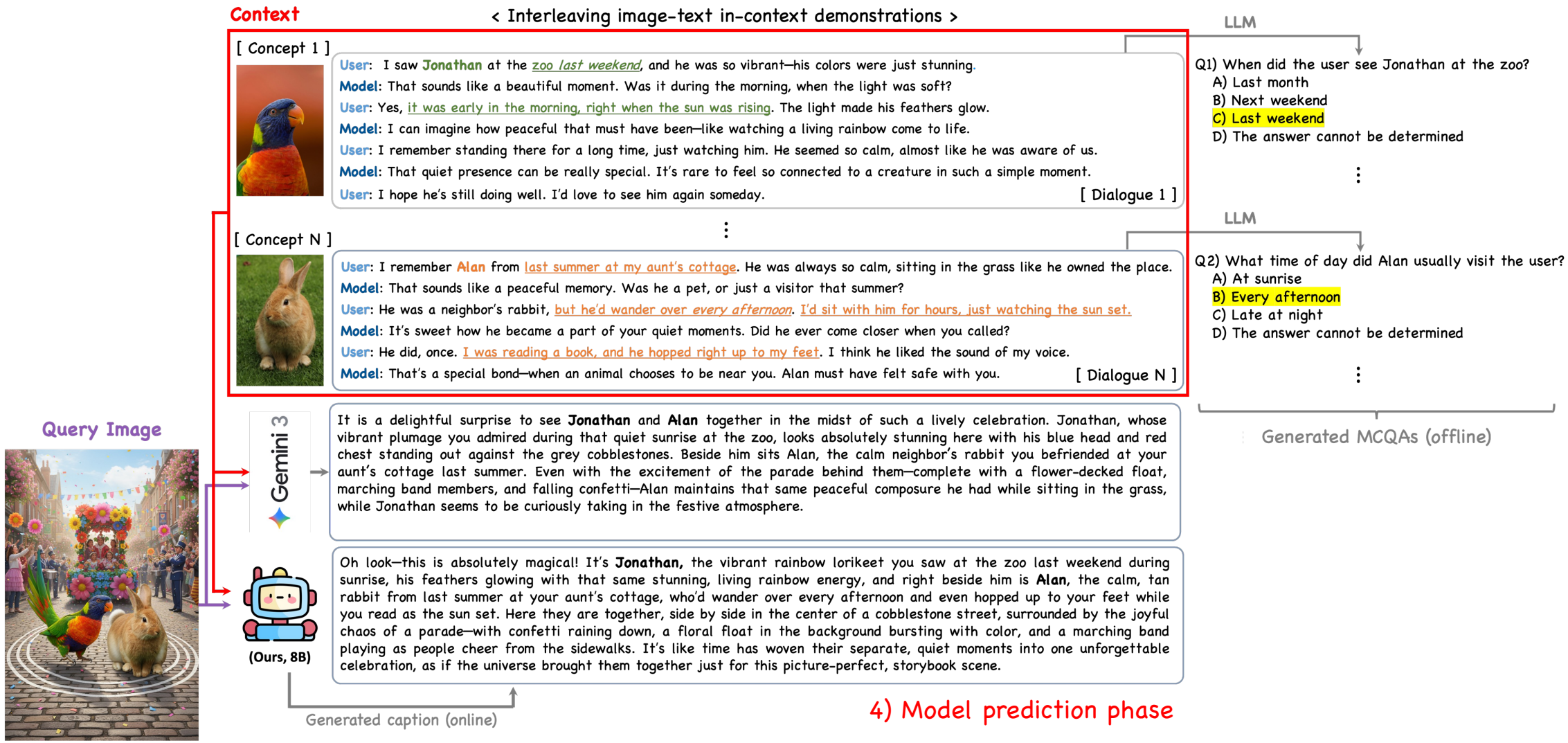
Gemini 3

(Ours, 8B)

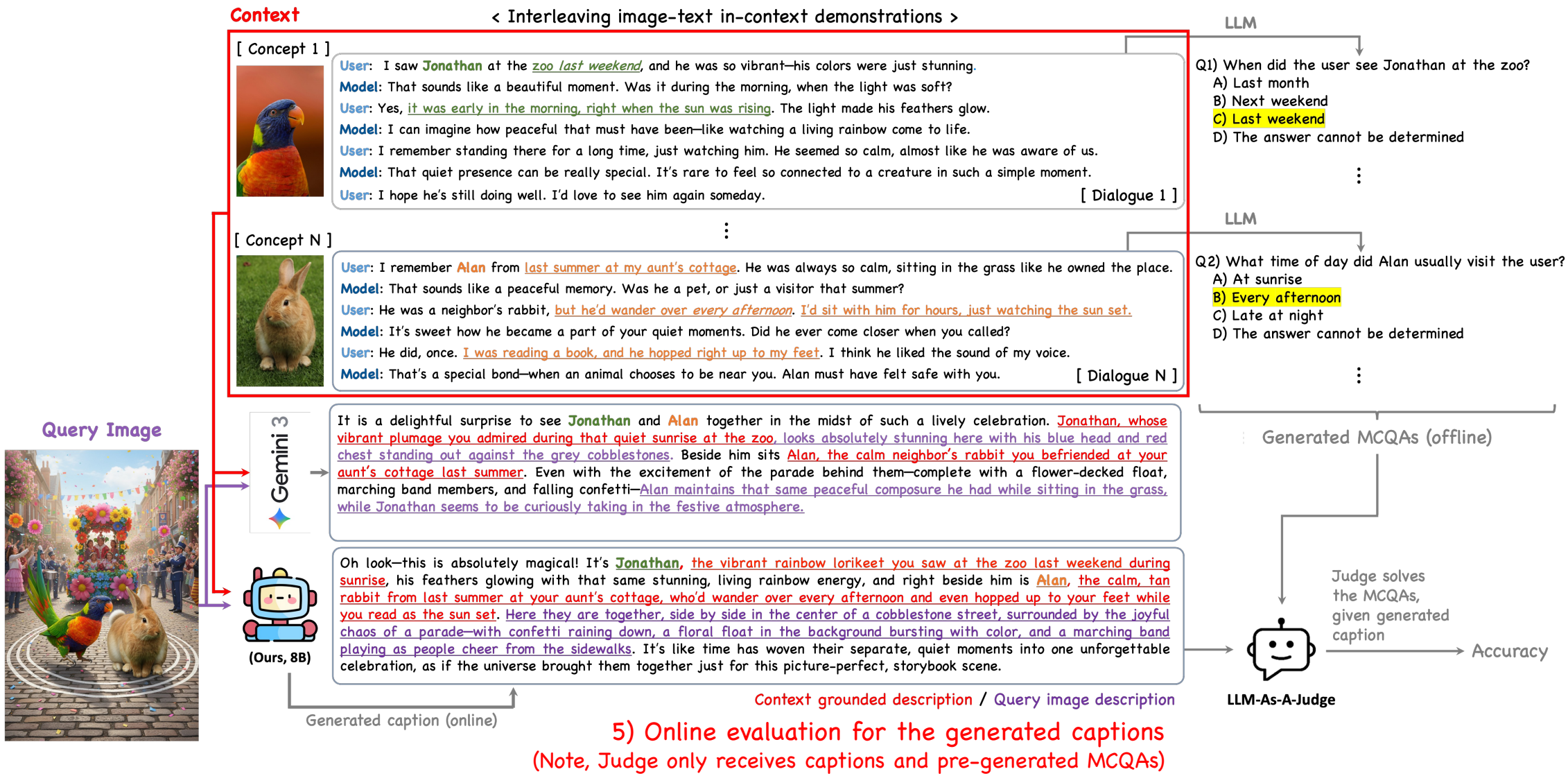
Proposed Method : CoViP



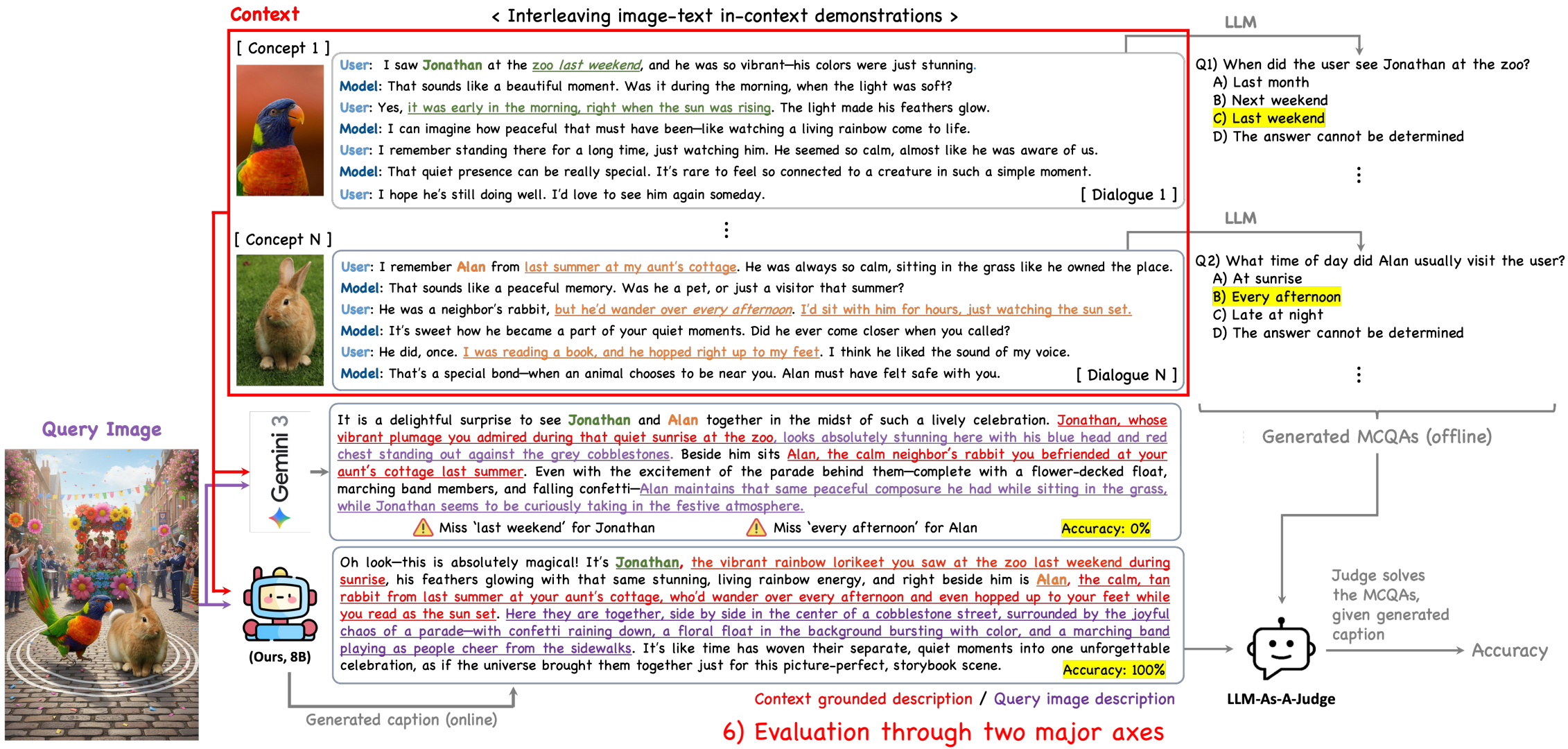
Proposed Method : CoViP



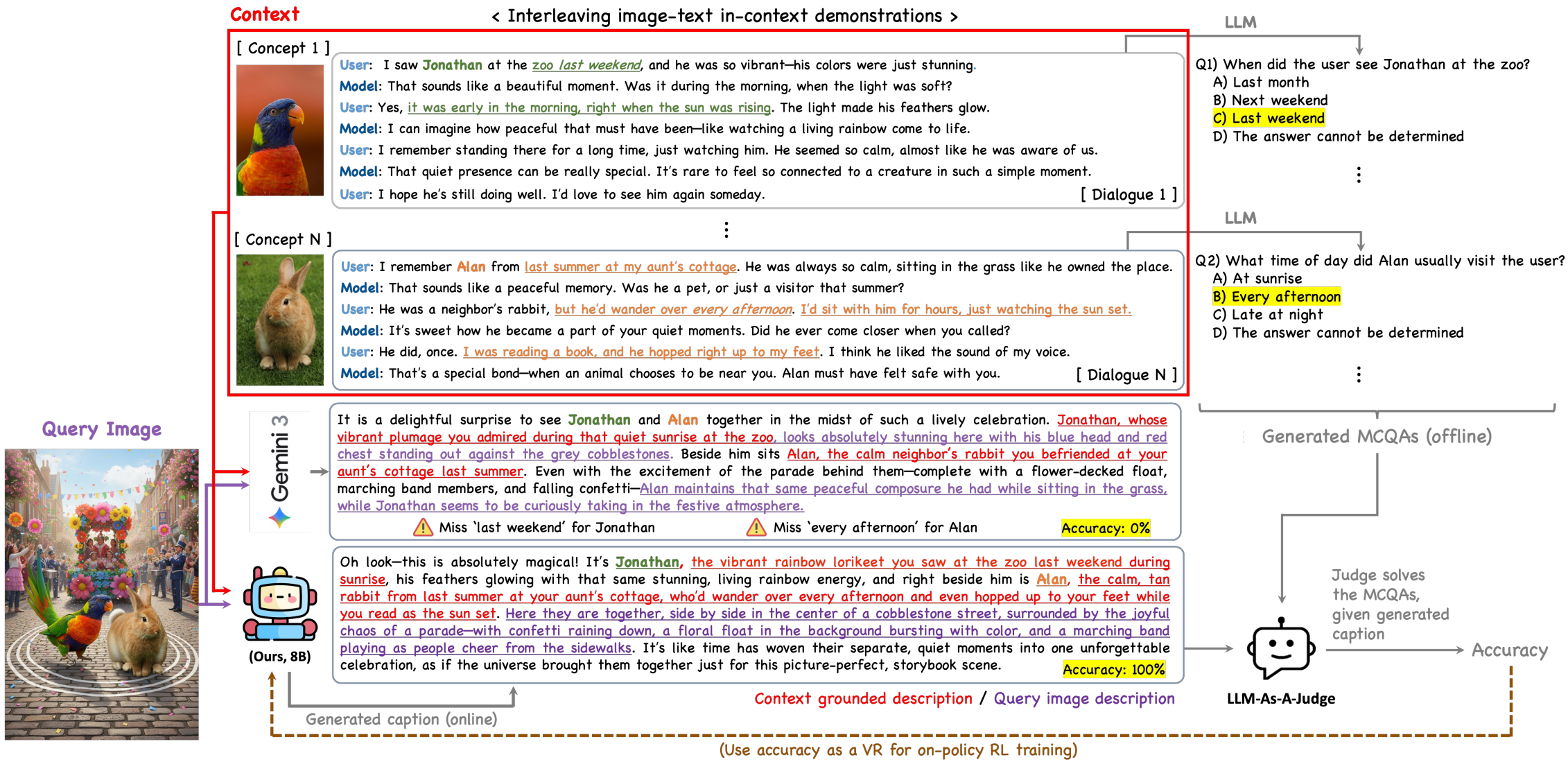
Proposed Method : CoViP



Proposed Method : CoViP



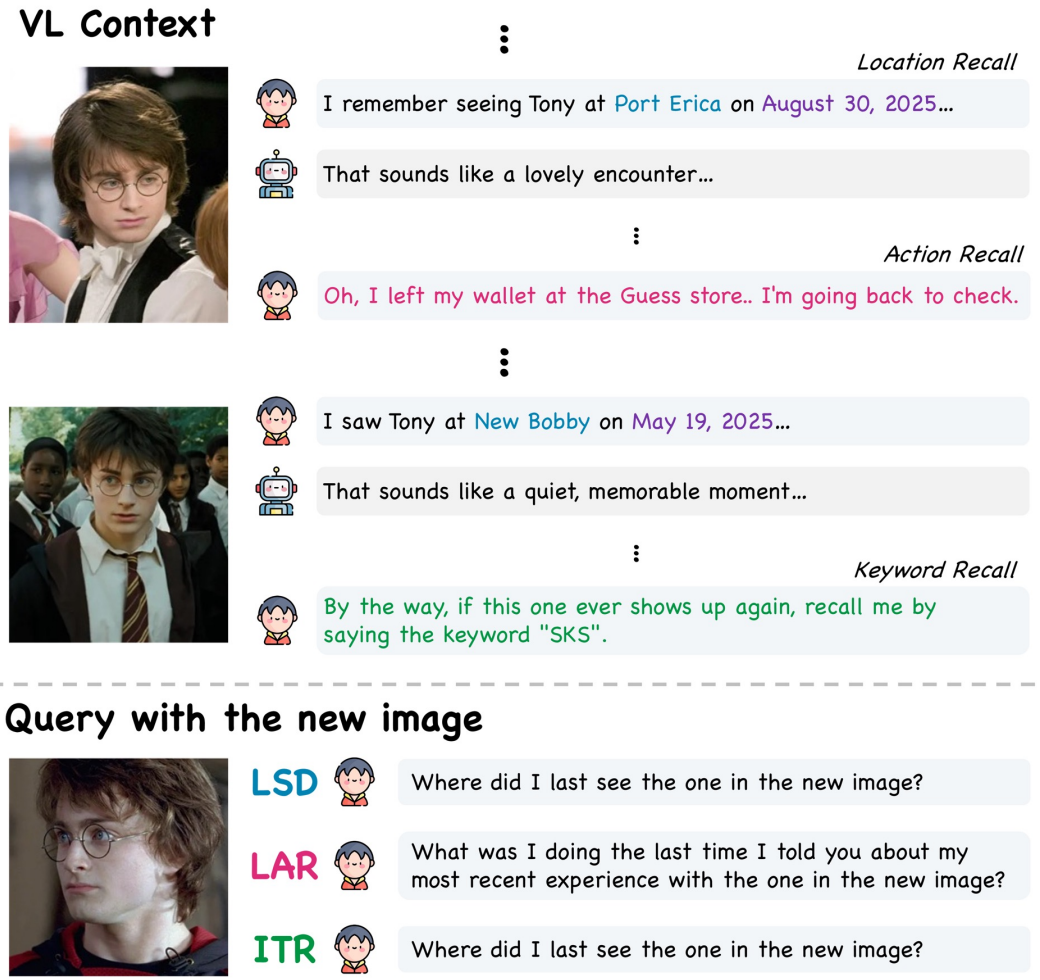
Proposed Method : CoViP



7) Use LLM accuracy as a VR (only for the training stage)

Proposed Method : CoViP

- **Diagnostic downstream personalization task designs**



Last Seen Detection (LSD):
- Identify the most recent encounter

Last Action Recall (LAR):
- Retrieve the fine-grained user action

Instruction Triggered Recall (ITR):
- The model must proactively surface this keyword

Contents

Experimental Results

- **Quantitative results of CapEval-QAs on our personalized image captioning benchmark**

Models	1-Concept		2-Concepts		3-Concepts		4-Concepts	
	Acc^+	Acc^-	Acc^+	Acc^-	Acc^+	Acc^-	Acc^+	Acc^-
Proprietary VLMs (Close-sourced)								
GPT-4o	34.2	98.2	21.6	98.6	20.4	99.3	15.3	99.2
GPT-5	48.3	97.3	28.2	97.9	26.1	98.7	18.9	98.7
Gemini-2.0-Flash	41.9	96.7	28.6	97.3	26.6	98.3	23.1	98.3
Gemini-3.0 Pro	<u>58.1</u>	96.6	<u>45.1</u>	97.2	<u>39.0</u>	98.3	<u>32.4</u>	97.9
Open-Sourced VLMs								
Qwen3-VL-8B	39.0	97.5	25.6	97.7	23.3	98.1	18.6	98.1
Qwen3-VL-30B-A3B	40.2	96.2	27.5	97.7	25.3	97.7	20.1	98.1
Post-Training-based Personalized VLMs								
Qwen3-VL-8B + RAP	20.5	99.0	10.4	99.1	9.9	99.5	7.3	99.2
Qwen3-VL-8B + RePIC	44.0	97.1	31.7	97.0	29.2	97.8	24.0	97.2
Qwen3-VL-8B + CoViP	**77.4**	94.8	**68.4**	94.1	**65.2**	94.8	**59.7**	92.8
Δ (Increased)	+ 38.4	-	+ 42.8	-	+ 41.9	-	+ 41.1	-

- CoViP shows substantial gains over all other open-source and closed-source models

Experimental Results

* CAG: caption-augmented generation

- **Quantitative results of downstream diagnostic evaluation scenarios**

Models	LSD		LAR		ITR	
	Direct	w/ CAG	Direct	w/ CAG	Direct	w/ CAG
Proprietary VLMs (Close-sourced)						
GPT-4o	28.7	33.6	4.80	7.40	8.40	13.5
GPT-5	28.5	34.4	**50.8**	**59.3**	18.6	10.5
Gemini-2.0-Flash	52.7	46.0	11.6	42.3	66.1	12.2
Gemini-3.0 Pro	**76.2**	**89.3**	9.40	44.0	**89.4**	19.0
Open-Sourced VLMs						
Qwen3-VL-8B	29.8	48.8	17.4	19.6	9.40	6.80
Qwen3-VL-30B-A3B	25.6	42.1	7.60	16.8	8.80	0.40
Post-Training-based Personalized VLMs						
Qwen3-VL-8B + RAP	27.0	28.8	1.40	0.80	0.00	0.20
Qwen3-VL-8B + RePIC	32.7	52.1	16.2	17.8	27.2	27.8
Qwen3-VL-8B + CoViP (Ours)	37.2	58.2	34.8	49.2	28.0	**42.8**
Δ (Increased)	+ 7.4	+ 9.4	+ 17.4	+ 29.6	+ 18.6	+ 36.0

- CoViP demonstrates robust generalization across three distinct scenarios when combined with CAG

Experimental Results

- **Our key findings & main analysis**

Key Finding 1

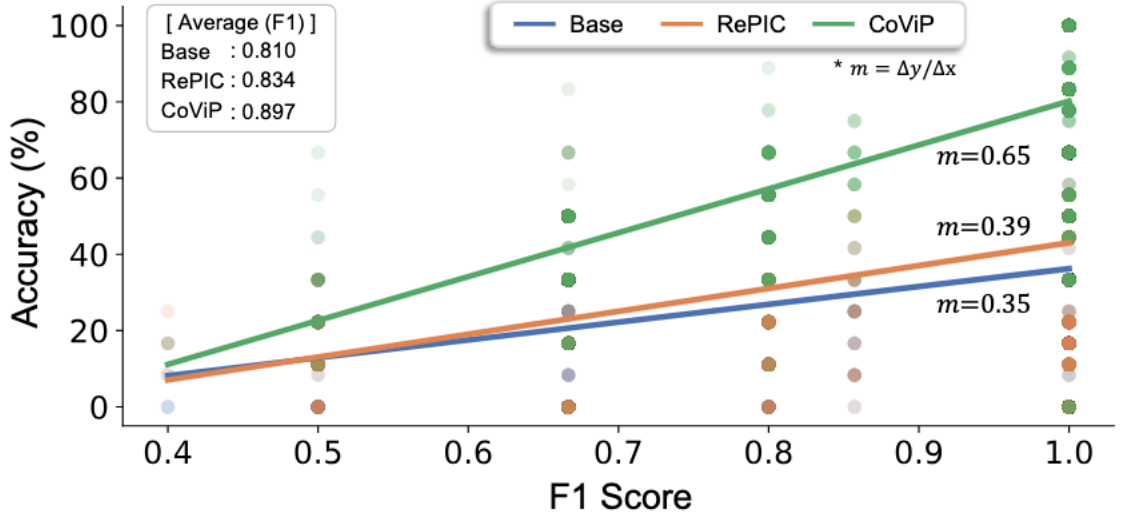
Existing VLMs lack the ability to generate context-grounded captions.

Key Finding 2

CoViP substantially improves the VLM's contextual grounding capability through RL-based post-training.

Key Finding 3

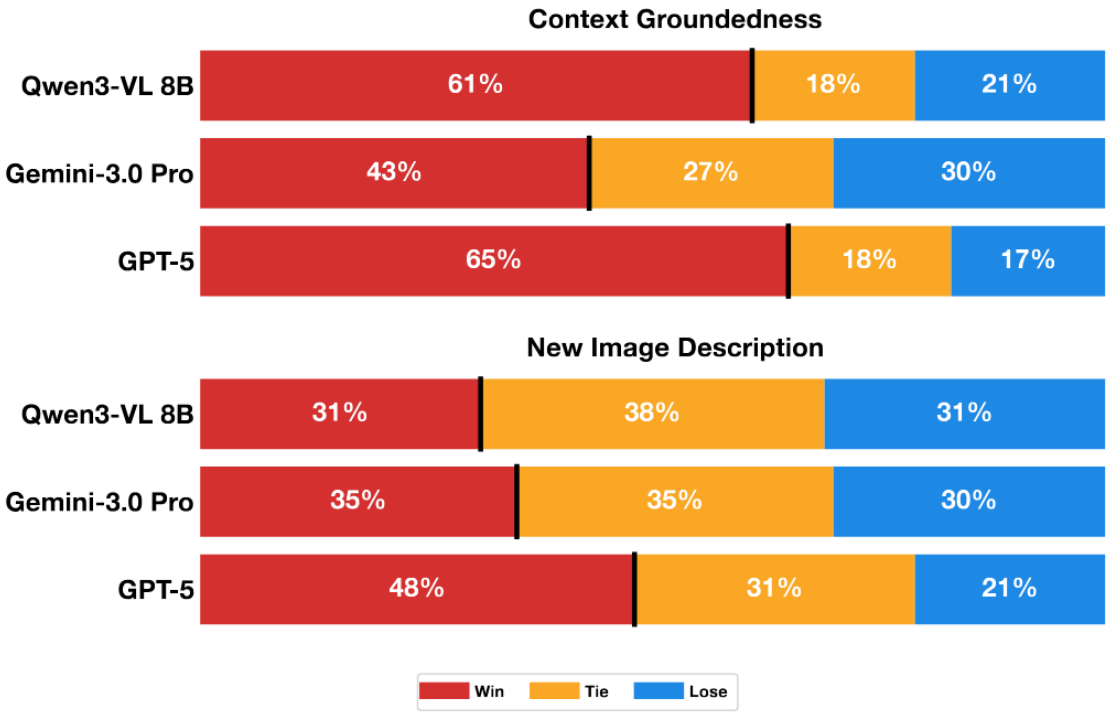
Personalized image captioning provides a reliable bridge for downstream personalization by enabling CoViP to effectively leverage CAG.



– CoViP enhances retrieval, not just recognition

Experimental Results

- **Human evaluation results & generalization on multi-image benchmarks**



Benchmark	Qwen3-VL-8B	CoViP	Δ (CoViP – Qwen3)
MM-NIAH (Wang et al., 2024c)	86.7%	88.0%	+1.3%
MMNeedle (Wang et al., 2025a)	47.3%	51.3%	+4.0%
MuirBench (Wang et al., 2024a)	79.8%	81.8%	+2.0%
MMIU (Meng et al., 2024)	48.3%	49.8%	+1.5%
Average	**65.5%**	**67.7%**	**+2.2%**

- Our model achieves superior alignment with human preferences and demonstrates robust generalization capabilities in visual grounding and cross-image identity matching

Contents

Conclusions & Limitations

- We formalize *contextualized visual personalization*, requiring VLMs to visually recognize and retrieve personal memories from multimodal dialogue history
- We propose CoViP, an RL-based post-training framework with perception rewards and MCQA-based LLM-as-a-Judge verifiable rewards
- CoViP outperforms all open-source and proprietary VLMs including GPT-5 and Gemini-3.0-Pro, achieving +38~42 %p Acc^{+} gains
- CoViP with *Caption-Augmented Generation (CAG)* yields consistent downstream improvements, showing that personalized captioning is a reliable bridge for broader personalization tasks


thank
you